STATISTIKA

UP FAMNIT, Biopsihologija

Zapiski s predavanj

Martin Raič

Datum zadnje spremembe: 9. julij 2019

Kazalo

1.

Uvod

Statistika je veda, ki preučuje bolj ali manj množične podatke (pojave) ali pa tudi pojme, ki so motivirani z njimi. Med drugim zajema:

- *Zbiranje podatkov*, torej kako (pri določenih praktičnih, npr. finančnih omejitvah) pravilno zbrati podatke, od katerih lahko pričakujemo čim natančnejšo informacijo o zadevi, ki nas zanima. Pomemben del te veje je *teorija vzorčenja* (angl. *sampling*).

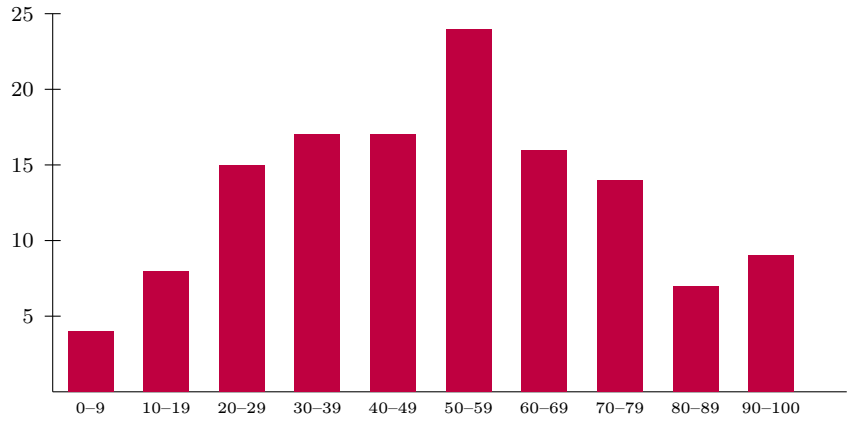
  Primer: Želimo vedeti, kdo bo zmagal na volitvah. Nikakor ne moremo povprašati vseh volivcev, a tudi če bi jih, ni rečeno, da nam bodo odgovorili enako, kot bodo volili, če nam bodo sploh dali odgovor. Ta motnja je toliko večja, kolikor več časa je še do volitev. Zato predvolilne ankete niso vedno zanesljive, zelo zanesljive pa so vzporedne volitve, če se prav izvedejo. Nekaj več o tem malo kasneje.

- *Povzemanje podatkov* – temu pravimo *opisna statistika* (angl. *descriptive statistics*).

  Primer: kaj vidimo iz naslednjih rezultatov kolokvija:

  50, 63, 52, 19, 69, 31, 40, 35, 47, 25, 35, 70, 99, 28, 52, 79, 68, 42, 55, 55, 0, 32, 58, 50, 28, 25, 67, 55, 60, 35, 27, 50, 55, 39, 75, 54, 75, 88, 60, 38, 64, 65, 53, 45, 29, 10, 55, 20, 27, 98, 85, 50, 55, 53, 74, 5, 50, 95, 49, 35, 23, 23, 72, 68, 30, 30, 80, 75, 47, 15, 88, 100, 60, 62, 17, 30, 100, 75, 40, 75, 78, 15, 90, 0, 25, 40, 68, 40, 55, 55, 55, 71, 45, 30, 85, 73, 33, 43, 41, 24, 37, 50, 85, 41, 48, 10, 35, 5, 40, 93, 33, 55, 20, 98, 56, 70, 25, 65, 68, 74, 80, 90, 57, 40, 15, 62, 37, 65, 25, 12, 49

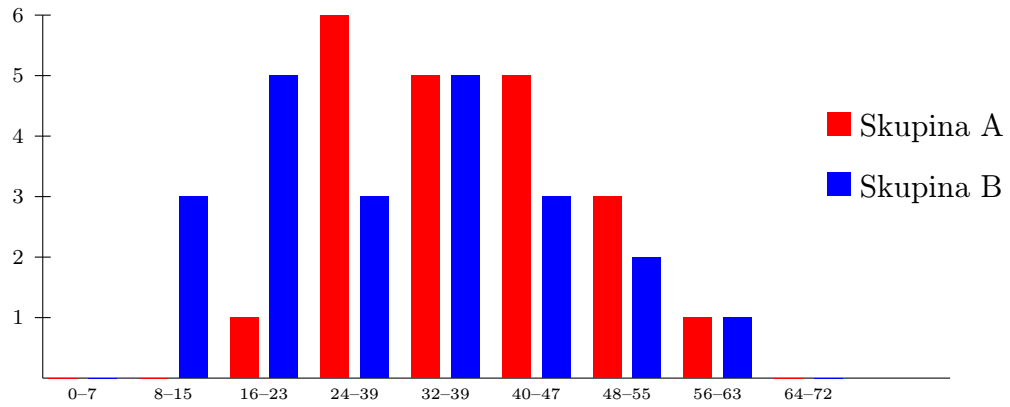
  Prav dosti že ne! Določen vtis nam da povprečni rezultat (aritmetična sredina) 50˙5, veliko pa pove tudi histogram:



Aritmetična sredina je primer *statistike*. Ta izraz ne pomeni samo vede, temveč pomeni tudi povzetek podatkov. Še en primer statistike je delež vseh študentov, ki so dosegli vsaj polovico vseh možnih točk.

- *Vrednotenje podatkov* – temu pravimo *inferenčna statistika*. Njeni najpomembnejši veji sta *statistično sklepanje* (dajanje sklepov – izjav na podlagi dobljenih podatkov, angl. *statistical inference*) in *statistično odločanje* (dajanje navodil, kako ravnati, da bomo v povprečju imeli največjo možno korist, angl. *decision theory*).

  Primer: neki drugi kolokvij so pisali v dveh skupinah: A in B. Skupina A je v povprečju zbrala 38˙38 točk od 72 možnih, skupina B pa 30˙73 točk. Ali lahko trdimo, da je skupina A dobila lažje naloge? Še histogram:



  Tukaj težko kar tako kaj trdimo. Možno je, da je bila skupina A res lažja, možno pa je tudi, da skupino so A pisali boljši študenti (tudi ne da bi izvajalec kolokvija to hotel). Tu nam inferenčna statistika lahko pomaga, a *nič ne moremo trditi z gotovostjo*. Lahko pa sklepanje nastavimo tako, da se zmotimo denimo največ v 5% primerov. To je osnovna filozofija inferenčne statistike.

Glede na zgoraj povedano je jasno, da kot matematično podlago za inferenčno statistiko potrebujemo *teorijo verjetnosti*. Le-to v zelo veliki meri potrebujemo tudi pri vzorčenju (nekaj več o tem malo kasneje).

1.1 Formalizacija podatkov

Podatki so v statistiki vrednosti ene ali več *statističnih spremenljivk* na *statistični množici*. Statistično množico sestavljajo *enote*. Primeri statističnih množic:

- če delamo anketo, množica anketirancev;
- če gledamo vreme po različnih krajih, nabor vremenskih postaj;
- če gledamo spreminjanje cen različnih artiklov s časom, nabor časov (enote so tako npr. 15. januar, 15. februar, 15. marec ...);

Število enot imenujemo *numerus* in ga navadno označujemo z n, tudi N.

Statistična spremenljivka je predpis, ki vsaki enoti (potem ko jo pomerimo) priredi določeno *vrednost*. Množico vrednosti posamezne spremenljivke imenujemo *zaloga vrednosti*. Statistične spremenljivke navadno označujemo z velikimi tiskanimi črkami s konca abecede, npr. X, Y, Z, njihove vrednosti pa z malimi črkami. Vrednosti statistične spremenljivke X tako navadno označimo z $x_1, x_2, \ldots, x_n$. Tako je x_i vrednost spremenljivke X na i-ti enoti statistične množice.

Primer: vreme po Sloveniji v torek, 19. februarja 2019, ob 9. uri:

Postaja	oblačnost	padavine	temperatura (°C)	smer vetra	hitrost vetra (km/h)
Kredarica	jasno	–	1	↘	7
Letališče Edvarda Rusjana Maribor	jasno	–	4	↘	4
Letališče Jožeta Pučnika Ljubljana	jasno	–	0	↙	0
Letališče Portorož	oblačno	–	4	↖	7
Ljubljana	jasno	–	1	↙	0

Tukaj postaja predstavlja enoto, padavine, oblačnost, temperatura ter smer in hitrost vetra pa so spremenljivke.

Včasih ni čisto nedvoumno, kaj je statistična množica oz. njene enote.

Primer: Recimo, da se vprašamo, koliko prebivalcev je imela v povprečju slovenska občina dne 1. 1. 2018. Izvleček iz podatkov:

Ljubljana 289.518, Maribor 110.871, Kranj 55.950, Koper 51.794, ..., Hodoš 362.
Vseh občin: 212.

V tem primeru je enota občina, spremenljivka pa število prebivalcev. Želeno povprečje izračunamo s formulo:

$$\frac{289.518 + 110.871 + 55.950 + 51.794 + \cdots + 362}{212} \doteq 9749\dot{}43\,.$$

Drugačno vprašanje pa je, v kako veliki občini je v povprečju živel prebivalec Slovenije. V tem primeru je statistična množica sestavljena iz prebivalcev Slovenije in ima 2.066.880 enot. Na njej lahko definiramo dve spremenljivki: občina, v kateri živi dani prebivalec, in število prebivalcev te občine. To je videti približno takole:

Enota	Občina	Št. prebivalcev
Zoran Janković	Ljubljana	289.518
... še 289.517 drugih	Ljubljana	289.518
Saša Arsenovič	Maribor	110.871
... še 110.870 drugih	Maribor	110.871
Matjaž Rakovec	Kranj	55.950
... še 55.949 drugih	Kranj	55.950
Aleš Bržan	Koper	51.794
... še 51.793 drugih	Koper	51.794
⋮	⋮	⋮
Ludvik Orban	Hodoš	362
... še 361 drugih	Hodoš	362

Želeno povprečje je zdaj enako:

$$\frac{289.518 \cdot 289.518 + 110.871 \cdot 110.871 + 55.950 \cdot 55.960 + 51.794 \cdot 51.794 + \cdots + 362 \cdot 362}{2.066.880}$$

$$\doteq 60.615\dot{}93\,.$$

Opazimo, da je višje kot prej, saj imajo zdaj občine z več prebivalci večjo težo.

Primer: V spodnji tabeli je prikazano število stanovanj glede na velikost v dveh mestnih območjih v Sloveniji:[1]

	do 20	21–40	41–60	61–80	81–100	101 +
Žalec	29	442	788	351	158	324
Žiri	5	61	184	197	169	559

Zanima nas, ali so se velikosti stanovanj v obeh območjih določale po istem ključu. V tem primeru je enota stanovanje. Populacijo sestavlja 3267 enot, od tega 2092 enot iz Žalca in 1175 enot iz Žirov. Na njej lahko definiramo dve smiselni spremenljivki: velikost

[1] Vir: Statistični letopis Republike Slovenije 2012

stanovanja, ki ima 6 možnih vrednosti in je urejenostna spremenljivka, in mestno območje, ki ima dve možni vrednosti: Žalec in Žiri. Slednja spremenljivka je imenska, a je, ker ima le dve možni vrednosti, tudi dihotomna.

Narobe pa bi bilo to interpretirati kot množico 12 enot, 6 iz Žalca in 6 iz Žirov, na katerih bi bila definirana razmernostna spremenljivka, ki bi imela na naši množici vrednosti 29, 442, 788, 351, 158, 324, 5, 61, 184, 197, 169 in 559.

1.2 Merske lestvice

Merska lestvica se v statistiki nanaša na statistično spremenljivko in pomeni, kakšno strukturo imajo vrednosti meritev spremenljivke oz. katere operacije lahko delamo s temi vrednostmi. Ločimo *opisne (kvalitativne, atributivne)* in *številske (kvantitativne, numerične)* merske lestvice. Opisne lestvice se nadalje delijo na:

- *Imenske (nominalne)*, pri katerih gledamo le gole vrednosti, na katerih ni definiranih nobenih operacij. Primeri: barva, politična stranka, rasa, pasma, skupina. Včasih lahko povemo, katere vrednosti so si blizu oz. sosedne – to je pomembno pri združevanju vrednosti v razrede.
- *Urejenostne (ordinalne)*, pri katerih lahko povemo, katera vrednost je večja in katera manjša. Primeri: kakovost razpoloženja, čin, stopnja izobrazbe, trdota minerala, odtenek sivine (kadar ocenjujemo na oko). Še zlasti pogosto se urejenostne lestvice pojavljajo pri raznih vprašalnikih, recimo ko imajo anketiranci na vprašanje, kako so razpoloženi, naslednje možne odgovore:

  • mizerno • slabo • srednje • dobro • izvrstno

Številske lestvice pa se delijo na:

- *Intervalske*, pri katerih lahko definiramo *razlike* med posameznima vrednostma in jih seštevamo, odštevamo ali delimo, medtem ko seštevanje, množenje in deljenje samih vrednosti *a priori* ni definirano oz. smiselno. Intervalske spremenljivke nimajo vnaprej določenega izhodišča (ničle). Lahko torej recimo povemo, da je razlika med vrednostma a in b dvakratnik razlike med b in c, ni pa recimo smiselno reči, da je vrednost b dvakratnik vrednosti a. Primeri intervalskih spremenljivk: letnica, nadmorska višina (v običajnih okoliščinah), temperatura (v običajnih okoliščinah, ko jo je smiselno gledati v Celzijevih stopinjah, recimo ko je ne povezujemo z energijo molekul).
- *Razmernostne*, pri katerih lahko vrednosti same seštevamo, odštevamo in delimo. Le-te imajo naravno izhodišče (ničlo) in lahko recimo povemo, da je vrednost b dvakratnik vredosti a. Primeri: moč motorja, dohodek, odtenek sivine (če jo merimo z instrumentom ali določimo računalniško) in tudi temperatura, kadar jo je smiselno gledati v kelvinih, recimo pri fiziki nizkih temperatur (blizu absolutne ničle), pri kinetični teoriji plinov in pri preučevanju zvezd.

Krajše pravimo, da je statistična spremenljivka imenska, urejenostna, intervalska oz. razmernostna, če je izmerjena na imenski, urejenostni, intervalski oz. razmernostni merski lestvici.

Vsako razmernostno spremenljivko lahko gledamo tudi kot intervalsko, vsako intervalsko kot urejenostno in vsako urejenostno kot imensko. Vendar pa pri tem vedno izgubimo nekaj informacije. Posebej veliko jo izgubimo, če urejenostno spremenljivko degradiramo v imensko, zato tega navadno ne počnemo.

Poseben primer merskih lestvic so *dihotomne* ali tudi *binarne*, to so take, ki lahko zavzemajo le dve vrednosti, recimo:

- da/ne;
- za/proti;
- pravilno/nepravilno;
- kontrolna/eksperimentalna skupina.

Tudi če je dihotomna lestvica opisna, jo lahko včasih naravno obravnavamo kot številsko, navadno tako, da vrednostma priredimo števili 0 in 1.

Pri primeru z vremenom so padavine, oblačnost in smer vetra imenske spremenljivke, pri katerih lahko povemo, katere vrednosti so si blizu. Temperatura je intervalska, hitrost vetra pa je razmernostna spremenljivka.

Smer in hitrost vetra lahko združimo v razmernostno vektorsko spremenljivko. Tudi te so pomembne, a se z njimi ne bomo ukvarjali.

Če bi pri oblačnosti gledali le, koliko neba ni vidnega (meglo bi torej izenačili z oblačnostjo) in tega ne bi kvantitativno merili (recimo v odstotkih), temveč bi le ločili npr. med jasnim, delno oblačnim, pretežno oblačnim in oblačnim vremenom, bi bila oblačnost urejenostna spremenljivka. Delež neba v odstotkih, ki ga zakrivajo oblaki, pa bi bil razmernostna spremenljivka.

Iz padavin je malo težje narediti urejenostno spremenljivko, ki ne bi mogla biti tudi razmernostna: težko je namreč primerjati dež in sneg. Najbolj objektivno bi bilo meriti, koliko milimetrov padavin pade recimo na uro: to bi bila razmernostna spremenljivka.

Glavna razlika med urejenostnimi in intervalskimi lestvicami je ta, da ne moremo primerjati razkorakov med posameznimi vrednostmi. Zato tudi ne moremo računati povprečij. Dostikrat sicer urejenostno spremenljivko ‘povišamo’ v intervalsko, tako da vrednostim priredimo številske vrednosti. Rezultati nadaljnje obdelave pa so lahko zavajajoči. V nekem podjetju bi lahko imeli naslednjo strukturo izobrazbe:

Nedokončana osnovna šola	70
Osnovna šola	5
Poklicna srednja šola	2
Gimnazija	1
Fakulteta	22

in lahko bi izračunali, da je povprečna izobrazba osnovnošolska. Če pa bi šli fakultetno izobrazbo podrobneje razčlenjevat, bi recimo dobili:

Nedokončana osnovna šola	70
Osnovna šola	5
Poklicna srednja šola	2
Gimnazija	1
Visoka strokovna izobrazba	2
Univerzitetna diploma bolonjske 1. stopnje	0
Univerzitetna diploma po starih programih	0
Univerzitetna diploma bolonjske 2. stopnje	0
Magisterij po starem programu	0
Doktorat	20

in izračunali, da je povprečna izobrazba poklicna srednja šola.

1.3 Nekaj več o vzorčenju

Često ne moremo zbrati podatkov na vsej statistični množici. Včasih si tega ne moremo finančno privoščiti, včasih (denimo pri testih trkov avtomobilov) pa merjenje pomeni tudi fizično uničenje enote in pač ne moremo uničiti vse statistične množice. Pač pa imamo možnost to narediti na določeni podmnožici.

Celotni statistični množici pravimo *populacija*, podmnožici, na kateri zberemo podatke, pa *vzorec*. Seveda vzorec ne bo dal popolne informacije o populaciji, lahko pa bo dal približno informacijo o določeni statistiki na populaciji. Iskani statistiki na populaciji bomo rekli *karakteristika*. Cilj vzorčenja je dobiti vzorec, iz katerega se bo dalo dobro oceniti vrednost izbrane karakteristike na populaciji.

Karakteristika navadno temelji na eni ali več statističnih spremenljivkah. Če se omejimo le na eno spremenljivko, to pomeni, da je odvisna le od *porazdelitve* te spremenljivke na populaciji. Porazdelitev pa pomeni, kolikšni so deleži posameznih vrednosti – recimo da je 2/3 ljudi za reformo, 1/6 ljudi proti njej, 1/6 pa je nepredeljenih. Če imamo več spremenljivk, pa mora karakteristika temelji na njih, če je odvisna od *skupne* ali *navzkrižne porazdelitve*. Ta pa pomeni, kolikšni so deleži *kombinacij* možnih vrednosti, npr. 5/12 ljudi je žensk in je za reformo.

Primer. Če nas zanima, katera stranka bo zmagala na volitvah, populacijo tvorijo vsi volivci, spremenljivka pa je stranka, za katero je opredeljen volivec. Karakteristika, ki nas zanima, je stranka, za katero je opredeljenih največ volivcev. To lahko povemo tudi tako, da je zanjo opredeljen največji delež volivcev, torej je odvisna samo od porazdelitve spremenljivke – deležev posameznih strank.

Nemogoče je povprašati vse volivce, a pogosto lahko zmagovalca volitev dokaj zanesljivo napovemo že iz ankete – vzamemo ustrezen vzorec recimo 1000 volivcev.

Vzorec omogoča dobro ocenjevanje karakteristike na populaciji, če je *reprezentativen*. Reprezentativen pomeni, da dobro odslikuje populacijo. Če ocenjujemo vrednost karakteristike, ki temelji na določeni spremenljivki, je dovolj, da dobro odslikuje populacijo, kar zadeva tisto spremenljivko. Vzorec je za določeno spremenljivko popolnoma reprezentativen, če se porazdelitev spremenljivke na vzorcu (*empirična porazdelitev*) ujema s porazdelitvijo na populaciji. V tem primeru se vrednost karakteristike na populaciji ujema z vrednostjo ustrezne statistike na vzorcu. Na podlagi predvolilne ankete bomo recimo lahko napovedali delež glasov za določeno stranko: to je karakteristika populacije, ustrezna statistika na vzorcu pa je vzorčni delež volivcev, opredeljenih za to stranko. Seveda pa popolne reprezentativnosti tipično ne moremo doseči, lahko pa dosežemo približno reprezentativnost. Statistika na vzorcu je ocena iskane karakteristike na populaciji – bolj ali manj natančna.

Reprezentativnost je težko doseči, če je vzorec zelo majhen, medtem ko je cela populacija popolnoma reprezentativen vzorec. Pri dobro izvedenem postopku vzorčenja je vzorec tipično dovolj reprezentativen, brž ko je dovolj velik. Za takšno vzorčenje bomo rekli, da je *asimptotično reprezentativno* (za določeno spremenljivko ali nabor spremenljivk). V tem primeru bo vzorčna statistika za dovolj velik vzorec tipično dober približek iskane populacijske karakteristike – večji kot je vzorec, natančnejši približek lahko pričakujemo. Pravimo, da je statistika *dosledna* cenilka karakteristike.

Ta definicija asimptotične reprezentativnosti je sicer matematično nenatančna. Prvič bi bilo treba opredeliti, kaj pomeni, da je vzorec dovolj reprezentativen, za ta namen pa bi bilo treba opredeliti, kdaj je empirična porazdelitev blizu porazdelitve na populaciji. Poleg tega bi bilo treba opredeliti, kaj pomeni dovolj velik vzorec: cela populacija je vsekakor dovolj velik vzorec, a to ni tisto, kar želimo. Niti enega niti drugega tu ne bomo precizirali, pomembno pa je, da 'dovolj velik vzorec' ne pomeni, da mora zajemati velik del populacije: glavna ideja vzorčenja je, da o populaciji sklepamo na osnovi njenega *majhnega* dela.

Primer: *predsedniške volitve v ZDA l. 1936.* [19, 21] Pomerila sta se Alfred Landon in Franklin Delano Roosevelt. Pred volitvami je revija *Literary Digest* izvedla obsežno javnomnenjsko raziskavo, ki je zajela 10 milijonov volivcev. Odgovorilo je več kot 2.300.000 volivcev. Šlo je za verjetno največji vzorec v zgodovini. Rezultat je bil 57% za Landona in 43% za Roosevelta.[2] Kdo pa je dejansko zmagal, se ve: v resnici je bilo za Landona 38%, za Roosevelta pa 62% veljavnih glasovnic. [12, 24][3]

V istem času pa je mladi statistik George Gallup povprašal le 50.000 volivcev in dobil rezultat 44% za Landona in 56% za Roosevelta.[4] Kljub veliko manjšemu vzorcu je dobil dosti natančnejšo oceno, ki je pravilno napovedala zmagovalca. Gallup pa je napovedal tudi, kakšen bo izid raziskave Literary Digesta: iz vzorca 3.000 anketirancev je napovedal, da bo rezultat 56% za Landona in 44% za Roosevelta. [21] Zmotil se je torej le za odstotek!

[2]Natančneje, prišlo je 2.376.523 odgovorov, od katerih jih je bilo 1.293.669 za Landona in 972.897 za Roosevelta, preostalih 109.957 pa jih ni bilo niti za enega niti za drugega. [21, 23]

[3]Natančneje, za Landona ali Roosevelta je glasovalo 44.434.510 volivcev (približno 2˙7% vseh glasov je bilo za druge kandidate), od tega 16.681.862 (37˙54%) za Landona in 27.752.648 (62˙46%) za Roosevelta.

[4]Natančneje, 44˙3% za Landona in 55˙7% za Roosevelta.

Za dobro oceno je torej bolj od velikosti vzorca pomembno, na kakšen način, po kakšnem protokolu ga vzamemo. Kaže, da je Gallup vzorčil bolje kot Literary Digest.

Preprosta metoda, s katero se da doseči dosti natančno statistično sklepanje, je *sistematično vzorčenje*, kjer enote oštevilčimo, nakar v vzorec vzamemo recimo vsako deseto enoto. A tudi tu se lahko skrivajo pasti: če želimo recimo oceniti, koliko ljudi v povprečju v eni uri prečka Titov trg v Kopru in to naredimo tako, da jih štejemo 24 nedelj zapored med 6. in 7. uro zjutraj, ocena ne bo dobra.[5]

Če želimo, da je sistematično vzorčenje učinkovito, torej štetvilčenje ne sme biti povezano z merjenimi spremenljivkami. Splošneje lahko rečemo, da, v kolikor ne poznamo dodatnih povezav, učinkovito vzorčenje dosežemo, če je izbor enot v vzorec čimbolj nepovezan s spremenljivkami, ki jih merimo. Preizkušen način, s katerim tovrstno povezanost odpravimo, je vpeljava *slučaja* v vzorčni načrt: odločitev, katere enote vzeti v vzorec, je slučajna. Temu pravimo *verjetnostno vzorčenje*. Povedano slikovito, pri tem vzorčenju mečemo kocko. Če je pravilno izvedeno, se vzorčni delež določene vrednosti določene statistične spremenljivke *v povprečju*[6] ujema s populacijskim. Podobno se vzorčno povprečje spremenljivke v povprečju ujema s populacijskim. Pravimo, da je vzorčni delež oz. povprečje *nepristranska* cenilka populacijskega. To je že dober korak do majhne napake.

Najpreprostejši primer verjetnostnega vzorčenja, kjer je prej omenjena neodvisnost skupaj z dodatnimi pogoji izpolnjena, je *enostavno slučajno vzorčenje*. To pomeni, da so vsi možni vzorci predpisane velikosti enako verjetni. Na populaciji velikosti 6 npr. obstaja 20 vzorcev velikosti 3:

$$\begin{array}{lllll} \{1,2,3\} & \{1,2,4\} & \{1,2,5\} & \{1,2,6\} & \{1,3,4\} \\ \{1,3,5\} & \{1,3,6\} & \{1,4,5\} & \{1,4,6\} & \{1,5,6\} \\ \{2,3,4\} & \{2,3,5\} & \{2,3,6\} & \{2,4,5\} & \{2,4,6\} \\ \{2,5,6\} & \{3,4,5\} & \{3,4,6\} & \{3,5,6\} & \{4,5,6\} \end{array}$$

Če so res vsi možni vzorci enako verjetni, torej če je "steklenica dobro pretresena", je vzorčni delež nepristranska ocena populacijskega, enako pa velja tudi za povprečje. Seveda pa bo ocena toliko točnejša, kolikor večji bo vzorec in kolikor manj raznolika bo merjena spremenljivka: enostavno slučajno vzorčenje je asimptotično reprezentativno. Če ocenjujemo delež enot z določeno lastnostjo in ta delež ni preblizu 0 ali 1, je tipična napaka, ki jo naredimo, *reda velikosti* $1/\sqrt{n}$. *Le-ta ni odvisen od velikosti populacije.*

Opomba. Za uporabnika, ki želi zelo natančno oceno, je to slaba novica: ena decimalka več, torej 10-krat natančnejšo ocena, zahteva 100-krat večji vzorec.

Primer. Za vzorec velikosti 2˙3 milijona, kot ga je vzela revija Literary Digest, predvideni red velikosti napake znaša približno 0˙00066. To je znatno manj od dejanske napake, ki

[5]Ta raziskava sicer ne paše čisto v paradigmo populacija – vzorec, a v dovolj dobrem približku lahko vzamemo, da so enote enourni intervali v določenem obdobju, spremenljivka pa je število ljudi, ki v posameznem intervalu prečkajo Titov trg. Če štejemo vsako nedeljo ob isti uri, to pomeni, da v vzorec vzamemo vsako 168. enoto, se pravi, da gre za sistematično vzorčenje.

[6]Raba besede povprečje je tu nekoliko nenatančna in ustreza pojmu *pričakovane vrednosti* iz teorije verjetnosti.

je znašala 0˙19. Vzorec, ki ga je pri predsedniških volitvah v ZDA leta 1936 vzela revija Literary Digest, je bil torej zelo pristranski.

A tudi pri Gallupovem vzorcu velikosti 50.000 predvideni red velikosti napake znaša približno 0˙0045, kar je kar nekajkrat manj od dejanske napake 0˙06. Torej se je tudi Gallup odrezal slabše kot pri enostavnem slučajnem vzorčenju. Vsekakor pa je bil manj pristranski kot Literary Digest. Ta revija se je kmalu po omenjenih volitvah znašla v stečaju.

Pač pa se predvideni red velikosti napake sklada z dejansko napako pri Gallupovem vzorcu ankatirancev Literary Digesta: za vzorec velikosti 3000 dobimo napako reda velikosti 0˙018, dejanska napaka pa je bila manj kot 0˙01.

Enostavno slučajno vzorčenje je torej učinkovito, zahteva pa *popoln pregled nad celotno populacijo in popolno dostopnost do nje*. Na voljo moramo imeti npr. register prebivalstva, poleg tega pa tudi zagotovljen odziv. To velja tudi za sistematično vzorčenje.

Dostikrat pa nad celotno populacijo nimamo pregleda, lahko pa populacijo razdelimo na več delov in dosežemo popoln pregled nad poljubnim njenim delom (storiti to za vse njene dele pa bi bilo predrago). Če npr. izvajamo raziskavo med oskrbovanci domov za starejše občane, bomo morda lahko v vsakem dobili seznam oskrbovancev, prav tako tudi seznam vseh domov starejših občanov v Sloveniji, skupaj s števili oskrbovancev. Nimamo pa dovolj sredstev, da bi se odpravili v vse domove, temveč se odpravimo le v nekaj domov. V tem primeru bomo vzorčenje izvedli v dveh fazah: najprej bomo izbrali seznam domov, ki jih bomo obiskali (morda bomo pri tem upoštevali tudi, koliko oskrbovancev imajo), nato pa bomo v vsakem od domov, ki ga bomo zajeli v raziskavo, poizvedeli po seznamu oskrbovancev in vzeli enostavni slučajni vzorec. Morda bomo oskrbovance pred tem razdelili še glede na spol, psihofizično stanje in podobno. Pri tem mora biti *vnaprej znano*, kako bomo ravnali v vsaki situaciji, na katero naletimo (toliko in toliko dementnih, neodziv itd.) Temu pravimo *vzorčni načrt*.

Zgoraj opisanemu postopku pravimo *stratificirano vzorčenje*. Gre za to, da populacijo razdelimo na več podpopulacij, ki jim pravimo *stratumi*. Za vsak stratum predpišemo, kakšno nadaljnje vzorčenje bomo izvedli na njem, med drugim tudi, koliko enot bo obsegal ustrezni podvzorec.

Stratificirano vzorčenje se izvede tudi pri vzporednih volitvah, in sicer v v kombinaciji s sistematičnim: najprej se izbere vzorec volišč, nato pa na izbranih voliščih izvedejo sistematično vzorčenje.

Kot merilo za delitev v stratume izberemo dejavnike, ki v kar se da veliki meri vplivajo na merjene spremenljivke in nad katerimi imamo pregled. Te dejavnike lahko formaliziramo kot statistične spremenljivke. Pogoste spremenljivke, ki služijo kot kriterij za formiranje stratumov, so regija, spol, starost, stopnja izobrazbe, sorta itd. Spremenljivka spol tako določa dva stratuma. Vzorčenje tipično izvedemo tako, da so spremenljivke, ki služijo za formiranje stratumov, na vzorcu porazdeljene čimbolj enako kot na populaciji (temu pravimo *proporcionalna alokacija*). Če je npr. v populaciji 10% visoko izobraženih, poskusimo doseči, da je tako tudi na vzorcu. Tak vzorec je reprezentativen glede

na izobrazbo. Obetamo si, da je, če je vzorec reprezentativen glede na spremenljivke, za katere to lahko dosežemo, približno reprezentativen tudi glede na spremenljivke, ki nas zanimajo. To se lahko zgodi v primeru, če so prve povezane z drugimi.

Vzorec Literary Digesta je bil zelo daleč od reprezentativnosti. Zakaj? Literary Digest je ankete po pošti pošiljal svojim naročnikom, telefonskim naročnikom, imetnikom avtomobilov, članom raznih elitnih klubov in podobno, skratka volivcem, ki jih je bilo lahko izbrskati. Toda biti naročen na Literary Digest, imeti telefon ali avto ali biti član elitnega kluba je v tistem času pomenilo biti dobro situiran, politična opredelitev pa je lahko zelo *odvisna* od socialnega položaja. Spomnimo, da je bil to čas velike gospodarske krize, ko je bilo biti dobro situiran še težje kot sicer. Huda težava raziskave Literary Digesta je bila tudi velika neodzivnost, saj je na anketo odgovorilo le 23% vprašanih. Tudi dejstvo, ali se je kandidat odzval na anketo ali ne, je lahko zelo povezano z vrednostjo merjene spremenljivke, zato je neodzivnost lahko znaten vir pristranskosti.

1.4 Nekaj več o statističnem sklepanju

Pri opisni statistiki se osredotočimo le na podatke, ki jih imamo (na to, kar opazimo) in poskusimo narediti smiseln povzetek. Pri inferenčni statistiki pa gledamo podatke kot del nečesa večjega, česar ne poznamo v celoti. Tipičen primer je vzorec iz populacije: vrednosti statistične spremenljivke na vzorcu poznamo, na celotni populaciji pa ne. To pa ni edina možnost. Regresijska analiza se npr. ukvarja z napovedjo dogajanja v prihodnosti na podlagi podatkov iz preteklosti.

V pričujoči publikaciji se bomo posvetili še naslednji pogosti situaciji: nekajkrat izvedemo poskus, ki se izide na slučajen način, in na podlagi opaženih izidov poskusimo sklepati o verjetnosti, da se poskus izide na določen način. Primer: pri kovancu, ki ni nujno pošten, nas zanima verjetnost, da pade grb, zato ga nekajkrat vržemo in opaženi delež grbov je ocena za verjetnost, da na kovancu pade grb.

V splošnem gre pri inferenčni statistiki za to, da opazimo X, želeli pa bi povedati kaj o Y (*statistično sklepati*).[7] Omenili bomo tri vrste sklepanja:

- *Točkasto ocenjevanje*, pri katerem sestavimo algoritem, ki nam za vsako opažanje X vrne oceno $Y \approx \hat{Y}$. Pri tem mora biti količina $\hat{Y}$ *opazljiva* (deterministično določena z opažanjem X), želeli pa bi narediti čim manjšo napako. Količini $\hat{Y}$ pravimo *cenilka* oziroma *ocena* za Y. Izraz cenilka se nanaša bolj na načrtovanje poskusa, torej na formulo, po kateri iz X izračunamo Y. Izraz ocena pa se nanaša bolj na konkretno vrednost.

  Primer. V drugem krogu predsedniških volitev v Sloveniji 2. 12. 2012 sta se pomerila Borut Pahor in Danilo Türk. V dneh od 27. do 29. 11. je agencija Delo Stik izvedla

[7] V teoriji X in Y predstavimo kot slučajni spremenljivki na istem verjetnostnem prostoru, ki pa nima nujno znane verjetnostne mere. Temu pravimo *statistični model*. Če je porazdelitev vendarle znana, modelu pravimo *bayesovski*. S takimi modeli se ukvarja *bayesovska statistika*. Tako jo imenujemo zato, ker temelji na Bayesovi formuli.

anketo, v kateri se je 55% vprašanjih opredelilo za Pahorja, 24% za Türka, preostalih 21% pa je bilo neopredeljenih.

Postavimo se v čas pred volitvami, torej nas zanima, koliko glasov bosta dobila kandidata. Na podlagi opažanja $X = (55\%, 21\%, 11\%)$ želimo recimo oceniti dejanski delež glasov za Pahorja, ki ga označimo z Y. Smiselno je oceniti:

$$Y \approx \hat{Y} := \frac{55\%}{55\% + 21\% + 11\%} \doteq 0{\cdot}70 = 70\% .$$

Ocenimo torej, da se bo približno 70% volivcev opredelilo za Pahorja.

- *Intervalsko ocenjevanje*, pri katerem poskusimo Y umestiti v opazljiv interval, npr. $Y_{\min} < Y < Y_{\max}$. Intervalu $(Y_{\min}, Y_{\max})$ pravimo *interval zaupanja*. Seveda morata biti meji intervala $Y_{\min}$ in $Y_{\max}$ opazljivi. Če o Y nimamo popolne informacije, izjava $Y_{\min} < Y < Y_{\max}$ tipično ni vedno pravilna, da pa se kontrolirati verjetnost tega statističnega sklepa. Želimo doseči dvoje:
  - Verjetnost, da je res $Y_{\min} < Y < Y_{\max}$ (verjetnost pokritosti), naj bo v vsakem primeru vsaj β.
  - Širina intervala naj bo čim manjša.

  Parametru β pravimo *stopnja zaupanja*. Tipični stopnji zaupanja sta $\beta = 0{\cdot}95$ in $\beta = 0{\cdot}99$.

  Namesto stopnje zaupanja lahko povemo tudi *stopnjo tveganja* $\alpha = 1 - \beta$. Prej omenjenima tipičnima stopnjama zaupanja torej ustrezata stopnji tveganja $\alpha = 0{\cdot}05$ in $\alpha = 0{\cdot}01$. Stopnja tveganja torej pomeni verjetnost, da se bomo pri sklepanju zmotili.

  Primer. Spet vzemimo prejšnjo anketo Dela Stik. Za interval zaupanja ni dovolj poznati samo deležev vprašanih volivcev, temveč moramo poznati tudi *numerus*, torej število vprašanih. Le-to je $n = 786$. Informacija, ki jo imamo na voljo, je torej $X = (55\%, 21\%, 11\%, 786)$ (ali pa kar vektor iz števil vprašanih, ki so se opredelili za posameznega kandidata ali pa se niso opredelili). Za 95% interval zaupanja za delež volivcev, ki se bodo opredelili za Pahorja, se izkaže, da pride od $66{\cdot}3\%$ do $73{\cdot}6\%$ (kako se ga izračuna, glej podrazdelek 2.1.2).

  V resnici je Pahor dobil $67{\cdot}37\%$ glasov, kar je znotraj intervala zaupanja.

- *Testi značilnosti*, pri katerem o Y postavimo neko *hipotezo* (domnevo), npr. $Y = y^*$. Tej hipotezi navadno pravimo *ničelna hipoteza* in jo označimo s H_0. Nasprotje ničelne hipoteze je *alternativna hipoteza* in jo navadno označimo s H_1. Ničelna in alternativna hipoteza skupaj sestavljata *statistični model*.

  Primer: Loterija Slovenije trdi, da je polovica srečk dobitnih. Kupimo 8 srečk in samo dve zadeneta. V tem primeru ima smisel testirati ničelno hipotezo, da je res polovica vseh srečk dobitnih, proti alternativni, da je dobitnih manj kot polovica

srečk. Statistični model torej predvideva le, da je Loterija bodisi poštena bodisi nepoštena v svojo korist. Količina Y je torej v tem primeru delež dobitnih srečk v celi seriji, vrednost y^* pa je enaka $0{\cdot}5 = 50\%$.

Test značilnosti je postopek, ki za vsako opažanje za vsako opažanje izda enega od naslednjih sklepov:

- H_0 zavrnemo, torej sprejmemo alternativno hipotezo H_1.
- Ne rečemo nič (naredimo prazen sklep).

Pri testih značilnosti torej ničelne hipoteze ne sprejmemo – slednje je namreč zelo tvegano! Dobitnost srečk se namreč ne spremeni bistveno, če je delež dobitnih v celi seriji enak recimo $49{\cdot}99\%$, a ničelna hipoteza v tem primeru ne velja. Želimo doseči dvoje:

- H_0 naj se zavrne v čimveč primerih, ko ne velja.
- Verjetnost dogodka, da ničelno hipotezo zavrnemo, ko velja, naj bo v vsakem primeru največ α. Omenjenemu dogodku pravimo *napaka prve vrste* (napako druge vrste bomo definirali malo kasneje).

Parametru α pravimo *stopnja značilnosti*. Če ničelno hipotezo zavrnemo pri stopnji značilnosti $\alpha = 0{\cdot}05$, pravimo, da so odstopanja *statistično značilna*. Če pa jo zavrnemo pri $\alpha = 0{\cdot}01$, pa, da so *statistično zelo značilna*.

Stopnja značilnosti α pove, koliko smo pogumni pri zavračanju ničelne hipoteze. To je torej *stopnja tveganja* za napako prve vrste. Pri istem opažanju bomo pri velikih α ničelno hipotezo zavrnili, pri majhnih pa ne. Mejni stopnji značilnosti, ki loči zavrnitev ničelne hipoteze od nezavrnitve, pravimo p-vrednost. Ta je definirana za vsako opažanje pri določenem testu.

Odločanje, ali ničelno hipotezo zavrnemo ali ne, lahko zastavimo tako, da vsa možna opažanja razvrstimo glede na to, katera kažejo bolj in katera manj zoper ničelno hipotezo, t. j. v prid alternativni hipotezi. Nato izračunamo verjetnost, da, če velja H_0, opažanje kaže vsaj toliko zoper H_0 kot aktualno opažanje. Če je ta verjetnost manjša ali enaka α, ničelno hipotezo zavrnemo. Pri takem protokolu je omenjena verjetnost p-vrednost opažanja.

Če je p-vrednost manjša ali enaka $0{\cdot}05$, pravimo, da opažanje *statistično značilno* kaže zoper H_0, če je manjša ali enaka $0{\cdot}01$, pa pravimo, da zoper H_0 kaže *statistično zelo značilno*.

Primer: Izračunamo p-vrednost pri primeru z Loterijo. Odločiti se moramo, kdaj opažanje kaže bolj zoper ničelno hipotezo. V statistiki obstajajo za to metode, a v tem primeru je zelo očitno, da opažanje kaže toliko bolj zoper ničelno hipotezo, kolikor manj srečk je bilo dobitnih. p-vrednost je torej verjetnost, da izmed 8 srečk

zadeneta dve ali manj, seveda ob predpostavki, da je verjetnost, da je srečka dobitna, enaka 1/2. Ta verjetnost je enaka:

$$p = \frac{1 + \binom{8}{1} + \binom{8}{2}}{2^8} \doteq 14{\cdot}5\%\,,$$

torej odstopanja niso statistično značilna. Ne moremo torej sklepati, da Loterija goljufa.

Primer: Testiramo ničelno hipotezo, da je bilo vzorčenje pri anketi Dela Stik nepristransko, kar v tem primeru pomeni, da so vzorčili iz baze, v kateri je bil delež glasov za oba predsedniška kandidata enak (če se omejimo samo na opredeljene). Alternativna hipoteza trdi, da to ni res, torej da je bilo vzorčenje pristransko. Podobno lahko storimo za vzporedne volitve. Vzorec iz ankete Dela Stik ima glede na dejanske rezultate volitev p-vrednost $0{\cdot}232$. Pri Gallupovi anketi o izidu predsedniških volitev v ZDA leta 1936 pa p-vrednost pride manj kot 10^{-100}. Torej je tudi Gallup vzorčil pristransko (statistično zelo značilno), čeprav je pravilno napovedal zmagovalca. No, tudi Gallup ni v vseh svojih raziskavah napovedal prav.

Pomembno: *ničelne hipoteze nikoli ne sprejmemo*! Pri primeru z Loterijo nismo rekli, da je Loterija poštena, rekli smo le, da ne moremo reči, da ni poštena. Opažanje, da sta zadeli 2 srečki od 16, namreč podobno kot od situacije, ko je Loterija poštena, "odstopa" tudi od situacije, ko srečka zadene z verjetnostjo $49{\cdot}9\%$. V slednji situaciji pa ničelna hipoteza ne velja.

Dogodku, da ničelno hipotezo sprejmemo, čeprav ne velja, pravimo *napaka druge vrste*. To napako je težko ali celo nemogoče kontrolirati, zato pri testih značilnosti raje pravimo, da ne moremo reči ničesar, kot pa da storimo napako druge vrste.

Kako izračunati vse navedeno, bomo spoznali v naslednjem poglavju.

2.

Obravnava ene statistične spremenljivke: univariatna analiza

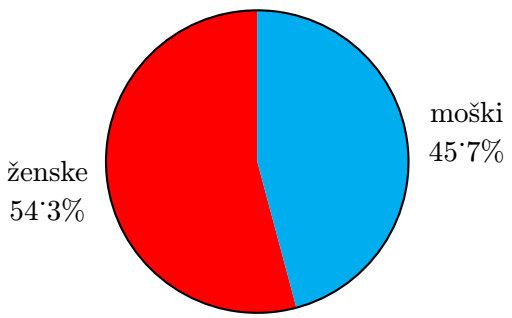
2.1 Dihotomne spremenljivke

2.1.1 Povzemanje

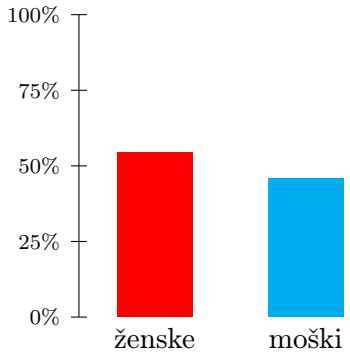
Pri dihotomnih spremenljivkah lahko podatke v glavnem povzamemo na dva načina:

- S *frekvencama*, ki povesta, koliko enot v statistični množici ima eno in koliko drugo vrednost. Denimo, v predavalnici je 35 slušateljev, od tega 19 žensk in 16 moških. Formalno bomo število enot z določeno lastnostjo označevali z znakom $\sharp$. Če spremenljivko označimo z X, vrednosti pa z a in b, sta frekvenci $\sharp(X = a)$ in $\sharp(X = b)$.
- Z *relativnima frekvencama* oz. *deležema* $\sharp(X = a)/n$ in $\sharp(X = b)/n$. Deleže često povemo v odstotkih. V prejšnjem primeru je bilo med slušatelji v predavalnici (do zaokrožitve natančno) $54{,}3\%$ žensk in $45{,}7\%$ moških.

Grafično podatke najpogosteje prikažemo s *tortnim grafikonom* (če ga narišemo na roko, potrebujemo kotomer – delež p ustreza kotu $p \cdot 360{,}0°$):



Lahko pa jih prikažemo tudi s *histogramom*:



2.1.2 Točkasto in intervalsko ocenjevanje

Recimo, da se podatki, ki smo jih dobili, nanašajo na vzorec iz neke populacije. Oceniti želimo delež enot v populaciji, ki imajo dano lastnost. To količino bomo označili s θ. Na voljo imamo vzorec iz n enot, od katerih jih ima našo lastnost natanko f. Vrednost f je *opažena oz. vzorčna frekvenca*.

Smiselna ocena za populacijski delež θ je *opažena relativna frekvenca*. Ocenimo torej:

$$\theta \approx \hat{\theta} = f^\circ = \frac{f}{n}\,.$$

Strešica nad črko, ki označuje določeno neznano količino, torej pomeni njeno *cenilko* ali *oceno*. Izraz *cenilka* uporabljamo, ko statistično sklepanje šele načrtujemo, torej preden dobimo podatke. Potem ko podatke že imamo, pa uporabljamo izraz *ocena*.

Če je vzorčenje asimptotično reprezentativno, se, ko večamo vzorec, opažena relativna frekvenca bliža populacijskemu deležu.

Primer. Za oddajo Moja Slovenija, ki je bila dne 2. marca 2013 na sporedu na RTV Slovenija, so 100 Slovencev (moških) med 15. in 75. letom starosti povprašali, ali nameravajo za 8. marec ženski, ki jim je blizu (ženi, partnerici, materi), podariti cvet. Pritrdilno jim je odgovorilo 85. Torej je $n = 100$, $f = 85$ in izračunamo $\hat{\theta} = f/n = 0{\cdot}85 = 85\%$. Na podlagi ankete torej ocenimo, da približno 85% vseh moških v Sloveniji med 15. in 75. letom starosti ženski, ki jim je blizu, podari cvet. Bolj formalno, če je θ delež vseh Slovencev med 15. in 75. letom starosti, ki namerava ženski, ki jim je blizu, podariti cvet, ocenimo $\theta \approx 0{\cdot}85$.

Še en primer: anketa Dela Stik v zvezi z drugim krogom predsedniških volitev v Sloveniji dne 2. 12. 2012, ki je bila izvedena v dneh 27.–29. 11. 2012: za Pahorja se je opredelilo 55%, za Türka 24% vprašanih, 21% pa jih je bilo neopredeljenih. Tedaj za n postavimo število opredeljenih, za f pa število tistih, ki so se opredelili za Pahorja. Točnih podatkov žal ni na voljo, glede na razpoložljivo pa je približno $n \doteq (0{\cdot}55 + 0{\cdot}25)m = 0{\cdot}79\,m$ in $f \doteq 0{\cdot}55\,m$, kjer je m število vprašanih. Torej delež glasov na volitvah za Pahorja, ki ga označimo s θ, ocenimo z:

$$\hat{\theta} \doteq \frac{f}{n} \doteq \frac{0{\cdot}55\,m}{0{\cdot}79\,m} = \frac{55}{79} \doteq 0{\cdot}70\,.$$

Kot smo že omenili, lahko statistično sklepamo ne le o deležu v populaciji, temveč tudi o verjetnosti, da se določen slučajni poskus izide na določen način. Temu pravimo *dogodek*. V tem primeru je to količina θ. Na voljo imamo n realizacij tega poskusa, pri čemer privzamemo, da so bile izvedene na naslednji način:

- Pri vsaki izvedbi je verjetnost, da se izbrani dogodek zgodi, enaka θ.
- Izvedbe so med seboj verjetnostno neodvisne.

Takemu zaporedju pravimo *Bernoullijevo*[1] *zaporedje*.

Opažena frekvenca f je zdaj število realizacij, ki so se izšle na izbrani način, t. j. pri katerih se je zgodil izbrani dogodek. Spet ocenimo $\theta \approx \hat{\theta} = f^\circ = f/n$. Če so izvedbe poskusa tvorijo Bernoullijevo zaporedje, se z večanjem njihovega števila opaženi delež f° določenega dogodka bliža verjetnosti θ tega dogodka. To dejstvo je poseben primer *zakona velikih števil* v teoriji verjetnosti.

Primer. Niso vsi kovanci pošteni: to je odvisno tudi od načina metanja. Če 50-krat vržemo kovanec in 38-krat pade grb, bomo ocenili, da na tem kovancu pri tem načinu metanja grb pade z verjetnostjo približno $38/50 = 76\%$.

Intervalsko ocenjevanje pa je nekoliko bolj zapleteno in zahteva še določene dodatne pogoje. Konstrukcij intervalov zaupanja je celo več in ne odgovarjajo vse glavni zahtevi po *pokritosti*, t. j. da je verjetnost, da je populacijsko povprečje res v intervalu zaupanja, enaka (najmanj) stopnji zaupanja β.

Preprosta konstrukcija je *Waldov*[2] *interval zaupanja*. Le-ta zahteva naslednje dodatne pogoje:

- Gre za enostavni slučajni vzorec iz velike populacije ali za Bernoullijevo zaporedje.
- Vzorec oz. število izvedb poskusa ni premajhno. Za tipično dogovorjeno natančnost se zahteva, da je $n \geq 30$.
- Opažena frekvenca ni preveč skrajna. Za tipično dogovorjeno natančnost se zahteva, da je $f > 5$ in $n - f > 5$. Malo kasneje bomo povedali, kaj se da narediti pri majhnih frekvencah.

[1] Jakob Bernoulli (1655–1705), švicarski matematik
[2] Abraham Wald (1902–1950), transilvanski matematik judovskega rodu

Širina Waldovega intervala zaupanja temelji na *standardni napaki*:

$$\mathrm{SE} = \sqrt{\frac{\hat\theta(1-\hat\theta)}{n}}\,,$$

ki predstavlja tipičen red velikosti razlike med opaženim in dejanskim deležem oz. opaženim deležem in verjetnostjo. Standardno napako pomnožimo z določenim koeficientom c, ki je odvisen od stopnje zaupanja oz. tveganja. V statistiki uporabljamo predvsem stopnji zaupanja 95% in 99% in iskana koeficienta za ti dve stopnji sta:

$$\begin{aligned}\beta = 95\%: &\qquad c \doteq 1{\cdot}96\,,\\ \beta = 99\%: &\qquad c \doteq 2{\cdot}58\,.\end{aligned}$$

Vrednosti temeljita na *normalni oz. Gaussovi porazdelitvi*. Natančneje, koeficient c je kvantil standardne normalne porazdelitve za verjetnost $\frac{1+\beta}{2} = 1 - \frac{\alpha}{2}$ – pišemo $c = z_{(1+\beta)/2} = z_{1-\alpha/2}$.

Waldov interval zaupanja je $\theta_{\min} < \theta < \theta_{\max}$, kjer je:

$$\theta_{\min} = \hat\theta - c\,\mathrm{SE}\,, \quad \theta_{\max} = \hat\theta + c\,\mathrm{SE}\,.$$

Žal se pri Waldovem intervalu zaupanja izkaže, da je dejanska pokritost slabša od deklarirane. To lahko izboljšamo s prisilnim zaokroževanjem, in sicer:

- Spodnjo mejo zaokrožimo navzdol, zgornjo pa navzgor.
- Meji zaokrožimo na toliko decimalk, kolikor mest ima velikost vzorca. A če je prva števka 1, le-te ne štejemo. Če ima torej vzorec 200 enot, zaokrožimo na tri, če ima 199 enot, pa le na dve decimalki.

Primer: anketa iz oddaje Moja Slovenija, kjer je 85 od 100 moških odgovorilo, da namerava ženski, ki jim je blizu, podariti cvet. Če postavimo $\beta = 95\%$, dobimo:

$$\begin{aligned}\mathrm{SE} &\doteq \sqrt{\frac{0{\cdot}85\cdot 0{\cdot}15}{100}} \doteq 0{\cdot}03570714\,,\\ \theta_{\min} &\doteq 0{\cdot}85 - 1{\cdot}96\cdot 0{\cdot}03570714 \doteq 0{\cdot}780014\,, \quad \text{kar zaokrožimo na } 0{\cdot}78,\\ \theta_{\max} &\doteq 0{\cdot}85 + 1{\cdot}96\cdot 0{\cdot}03570714 \doteq 0{\cdot}919986\,, \quad \text{kar zaokrožimo na } 0{\cdot}92.\end{aligned}$$

Pri stopnji zaupanja 95% torej ocenimo, da namerava med 78% in 92% moških v dani kategoriji ženski, ki jim je blizu, podariti cvet. Če bi postavili $\beta = 99\%$, pa bi dobili:

$$\begin{aligned}\theta_{\min} &\doteq 0{\cdot}85 - 2{\cdot}58\cdot 0{\cdot}03570714 \doteq 0{\cdot}757876\,, \quad \text{kar zaokrožimo na } 0{\cdot}75,\\ \theta_{\max} &\doteq 0{\cdot}85 + 2{\cdot}58\cdot 0{\cdot}03570714 \doteq 0{\cdot}942124\,, \quad \text{kar zaokrožimo na } 0{\cdot}95.\end{aligned}$$

Opomba. *Višja kot je stopnja zaupanja, širši mora biti interval zaupanja*: če želimo, da bo naša napoved z večjo verjetnostjo pravilna, moramo biti bolj ohlapni. Edini interval zaupanja s stopnjo zaupanja 100% je interval $[0, 1]$, to pa je seveda neuporabno. Sprejeti moramo torej kompromis med natančnostjo in zanesljivostjo.

Obstaja konstrukcija, ki zagotavlja deklarirano pokritost in pri kateri se širina intervala bliža optimalni, ko večamo vzorec. To je *Clopper–Pearsonov interval zaupanja*, ki pa je malo težje izračunljiv. Dober kompromis med pokritostjo, optimalnostjo širine in izračunljivostjo je *Agresti–Coullov* interval zaupanja.[16] Pri njem izračunamo:

$$\tilde{n} = n + c^2\,, \quad \tilde{\theta} = \frac{f + \frac{c^2}{2}}{\tilde{n}}\,, \quad \widetilde{\mathrm{SE}} = \sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{\tilde{n}}}\,, \quad \theta_{\min} = \tilde{\theta} - c\,\widetilde{\mathrm{SE}}\,, \quad \theta_{\max} = \tilde{\theta} + c\,\widetilde{\mathrm{SE}}\,.$$

Primer: spet anketa iz oddaje Moja Slovenija, kjer je 85 od 100 moških odgovorilo, da namerava ženski, ki jim je blizu, podariti cvet. Če postavimo $\beta = 95\%$, dobimo:

$$\tilde{n} = 100 + 1{\cdot}96^2 \doteq 103{\cdot}84\,, \quad \tilde{\theta} \doteq \frac{85 + 1{\cdot}96^2/2}{103{\cdot}84} \doteq 0{\cdot}83705\,,$$
$$\widetilde{\mathrm{SE}} \doteq \sqrt{\frac{0{\cdot}83705 \cdot 0{\cdot}16295}{103{\cdot}84}} \doteq 0{\cdot}03624\,,$$
$$\theta_{\min} \doteq 0{\cdot}83705 - 1{\cdot}96 \cdot 0{\cdot}03624 \doteq 0{\cdot}766\,, \quad \theta_{\max} \doteq 0{\cdot}83705 + 1{\cdot}96 \cdot 0{\cdot}03624 \doteq 0{\cdot}909\,.$$

Pri stopnji zaupanja 95% torej zdaj ocenimo, da namerava med 76·6% in 90·9% moških v dani kategoriji ženski, ki jim je blizu, podariti cvet. *Spodnjo mejo smo zaokrožili navzdol, zgornjo pa navzgor.*

Če pa bi vzeli $\beta = 99\%$, bi dobili:

$$\tilde{n} = 100 + 2{\cdot}58^2 \doteq 106{\cdot}66\,, \quad \tilde{\theta} \doteq \frac{85 + 2{\cdot}58^2/2}{106{\cdot}66} \doteq 0{\cdot}82813\,,$$
$$\widetilde{\mathrm{SE}} \doteq \sqrt{\frac{0{\cdot}82813 \cdot 0{\cdot}17187}{106{\cdot}66}} \doteq 0{\cdot}03653\,,$$
$$\theta_{\min} \doteq 0{\cdot}82813 - 2{\cdot}58 \cdot 0{\cdot}03653 \doteq 0{\cdot}733\,, \quad \theta_{\max} \doteq 0{\cdot}82813 + 2{\cdot}58 \cdot 0{\cdot}03653 \doteq 0{\cdot}923\,.$$

Opomba. Količina $\tilde{\theta}$ ni čisto enaka $\hat{\theta}$: Agresti–Coullov interval zaupanja je pomaknjen nekoliko stran od krajišč 0 in 1. Prav tako se modificirana standardna napaka $\widetilde{\mathrm{SE}}$ malenkost spreminja z β, medtem ko je nemodificirana standardna napaka SE neodvisna od β.

Še en primer: prej omenjena anketa Dela Stik v zvezi z drugim krogom predsedniških volitev v Sloveniji. Točni podatki sicer niso na voljo, a v okviru danih (poleg že omenjenih deležev potrebujemo še, da je bilo vprašanih $m = 786$ volivcev) bo smiselno postaviti $0{\cdot}79m \doteq 621 =: n$ in $0{\cdot}55m \doteq 432 =: f$. Pri $\beta = 95\%$ dobimo:

$$\tilde{n} = 621 + 1{\cdot}96^2 \doteq 624{\cdot}84\,, \quad \tilde{\theta} \doteq \frac{432 + 1{\cdot}96^2/2}{624{\cdot}84} \doteq 0{\cdot}69445\,,$$
$$\widetilde{\mathrm{SE}} \doteq \sqrt{\frac{0{\cdot}69445 \cdot 0{\cdot}30555}{624{\cdot}84}} \doteq 0{\cdot}01843\,,$$
$$\theta_{\min} \doteq 0{\cdot}69445 - 1{\cdot}96 \cdot 0{\cdot}01843 \doteq 0{\cdot}658\,, \quad \theta_{\max} \doteq 0{\cdot}69445 + 1{\cdot}96 \cdot 0{\cdot}01843 \doteq 0{\cdot}731\,.$$

Pri $\beta = 99\%$ pa dobimo:

$$\tilde{n} = 621 + 2{\cdot}58^2 \doteq 627{\cdot}66\,, \quad \tilde{\theta} \doteq \frac{432 + 2{\cdot}58^2/2}{627{\cdot}66} \doteq 0{\cdot}69358\,,$$
$$\widetilde{\mathrm{SE}} \doteq \sqrt{\frac{0{\cdot}69358 \cdot 0{\cdot}30642}{627{\cdot}66}} \doteq 0{\cdot}01840\,,$$
$$\theta_{\min} \doteq 0{\cdot}69358 - 2{\cdot}58 \cdot 0{\cdot}01840 \doteq 0{\cdot}646\,, \quad \theta_{\max} \doteq 0{\cdot}69358 + 2{\cdot}58 \cdot 0{\cdot}01840 \doteq 0{\cdot}742\,.$$

Torej bi na podlagi ankete pri stopnji zaupanja 95% napovedali, da bo za Pahorja glasovalo med 65·8% in 73·1% volivcev. Pri stopnji zaupanja 99% pa bi bila ta napoved med 64·6% in 74·2%.

Primer: vzporedne volitve pri drugem krogu predsedniških volitev v Sloveniji. Čisto točni podatki spet niso na voljo, a vemo, da so vprašali $n = 11.629$ volivcev, med katerimi se jih je 67·03% opredelilo za Pahorja in 32·97% za Türka. Postavimo $0{\cdot}6703 \cdot 11629 \doteq 7795 =: f$. Pri $\beta = 99\%$ dobimo:

$$\tilde{n} = 11629 + 2{\cdot}58^2 \doteq 11635{\cdot}7\,, \quad \tilde{\theta} \doteq \frac{7795 + 2{\cdot}58^2/2}{11635{\cdot}7} \doteq 0{\cdot}670210\,,$$
$$\widetilde{\mathrm{SE}} \doteq \sqrt{\frac{0{\cdot}670210 \cdot 0{\cdot}329790}{11635{\cdot}7}} \doteq 0{\cdot}004358\,,$$
$$\theta_{\min} \doteq 0{\cdot}670210 - 2{\cdot}58 \cdot 0{\cdot}004358 \doteq 0{\cdot}6589\,,$$
$$\theta_{\max} \doteq 0{\cdot}670210 + 2{\cdot}58 \cdot 0{\cdot}004358 \doteq 0{\cdot}6815\,.$$

Na podlagi vzporednih volitev bi torej pri stopnji zaupanja 99% napovedali, da bo za Pahorja glasovalo med 65·89% in 68·15% volivcev.

V resnici je na volitvah za Pahorja glasovalo 67˙37% volivcev, kar je v vseh intervalih zaupanja, ki smo jih obravnavali.

Tudi Agresti–Coullov interval nam sicer ne zagotavlja v vsakem primeru verjetnosti pokritosti vsaj β, toda verjetnost pokritosti se pri vsakem θ bliža β, ko se n veča. Bližanje je še hitrejše, če gledamo *povprečno* verjetnost pokritosti, ko θ preteče določen interval. Le-ta je zelo blizu β že za majhne n.

Če so frekvence zunaj postavljenega okvira legitimnosti Waldovega intervala zaupanja (torej zelo majhne ali zelo velike), si lahko navadno še vseeno pomagamo na preprost način. Če je $n \geq 10$ in $n \geq f^2$, se lahko namreč poslužimo naslednje tabele (vedno gledamo vrednost z majhno frekvenco):

$\beta = 0.95$

f	$\theta_{\min}$	$\theta_{\max}$
0	0	$3.45/n$
1	$0.025/n$	$5.58/n$
2	$0.242/n$	$7.24/n$
3	$0.618/n$	$8.77/n$
4	$1.08/n$	$10.3/n$
5	$1.62/n$	$11.7/n$

$\beta = 0.99$

f	$\theta_{\min}$	$\theta_{\max}$
0	0	$4.94/n$
1	$0.005/n$	$7.43/n$
2	$0.103/n$	$9.28/n$
3	$0.337/n$	$11.0/n$
4	$0.672/n$	$12.6/n$
5	$1.07/n$	$15.7/n$

Pojasnilo. Številke so dobljene iz kvantilov porazdelitve hi kvadrat: če je $\chi^2_p(m)$ kvantil porazdelitve hi kvadrat z m prostostnimi stopnjami za verjetnost p, za velike n interval zaupanja pride $\left(\frac{1}{2n}\chi^2_{(1-\beta)/2}(2f), \frac{1}{2n}\chi^2_{(1+\beta)/2}(2f+2)\right)$, če je $f > 0$. Pri $f = 0$ pa se izkaže, da je potrebno vzeti $\left[0, \frac{1}{2n}\chi^2_{0.025}(16)\right)$ pri $\beta = 0.95$ in $\left[0, \frac{1}{2n}\chi^2_{0.005}(24)\right)$ pri $\beta = 0.99$.

Primer. V vzorcu je 33 žensk in 2 moška. Recimo, da želimo oceniti delež žensk v populaciji. Legitimnost Waldovega intervala zaupanja spodbije frekvenca moških. A ker je $35 \geq 10$ in $35 \geq 2^2$, se lahko poslužimo tabele, iz katere odčitamo 95% interval zaupanja za delež moških, ki znaša od 0˙0069 do 0˙207. Ustrezni interval zaupanja za ženske znaša od 0˙793 do 0˙9931.

2.1.3 Testiranje deleža

Tako kot v prejšnjem razdelku naj bo tudi tu θ delež enot v populaciji z določeno lastnostjo ali verjetnost, da se določen poskus izide na določen način. Testiramo ničelno hipotezo H_0, da je ta delež oz. verjetnost enaka *hipotetičnemu deležu oz. verjetnosti* θ^*, na voljo pa imamo vzorec iz n enot, od katerih jih ima našo lastnost natanko f, oziroma n realizacij poskusa, od katerih se jih f izide na izbrani način.

Podobno kot pri intervalskem ocenjevanju je treba tudi pri testiranju privzeti določene dodatne pogoje. Ti so odvisni od izbire testa. Lahko gledamo tudi obratno – pri določenih pogojih je primeren določen test. Omenili bomo *Z-test*, ki sicer *ni eksakten* (kar pomeni, da se dejanska stopnja tveganja ne ujema z deklarirano), se pa preprosto izračuna. Ta zahteva naslednje dodatne pogoje:

- Gre za enostavni slučajni vzorec iz velike populacije ali za Bernoullijevo zaporedje.
- Vzorec oz. število izvedb poskusa ni premajhno. Za tipično dogovorjeno natančnost se zahteva, da je $n \geq 30$.
- *Pričakovana frekvenca* $n\theta^*$ ni preveč skrajna. Za tipično dogovorjeno natančnost se zahteva, da je $n\theta^* \geq 5$ in $n(1-\theta^*) \geq 5$.

Obravnavali bomo tri alternativne hipoteze: $H_1^{\pm}$, da je $\theta \neq \theta^*$, H_1^-, da je $\theta > \theta^*$, in H_1^-, da je $\theta < \theta^*$. Prva alternativna hipoteza je *dvostranska*, drugi dve pa sta *enostranski*. Zato v prvem primeru tudi za test pravimo, da je dvostranski, v drugih dveh primerih pa, da je enostranski.

Tako kot v prejšnjem razdelku naj ima vzorec velikost n in naj bo v njem f enot z dano lastnostjo. Spet označimo $\hat{\theta} := f/n$. Ključ do statističnega sklepanja je *testna statistika*, po kateri se ta test imenuje:

$$Z := \frac{\hat{\theta} - \theta^*}{\mathrm{SE}}\,, \quad \text{kjer je} \quad \mathrm{SE} = \sqrt{\frac{\theta^*(1-\theta^*)}{n}}\,.$$

Statistika Z je torej *razmerje med opaženo razliko in standardno napako*.

Spomnimo se, da ničelno hipotezo zavrnemo, če je p-vrednost manjša od stopnje značilnosti α. p-vrednost pa je odvisna od alternativne hipoteze:

- če je to $H_1^{\pm}$, je $p = 1 - 2\Phi(|Z|)$;
- če je to H_1^+, je $p = \frac{1}{2} - \Phi(Z)$;
- če je to H_1^-, je $p = \frac{1}{2} + \Phi(Z)$.

Tu je Φ *Gaussov verjetnostni integral*:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-t^2/2}\,\mathrm{d}t\,,$$

njegovo vrednost pa lahko odčitamo iz tabele 1.

Če pa nas ne zanima p-vrednost, temveč izberemo stopnjo značilnosti α, pri kateri bomo ničelno hipotezo zavrnili ali pa nič sklepali, pa potrebujemo le kvantile normalne porazdelitve (več o normalni porazdelitvi v razdelku 2.4.7):

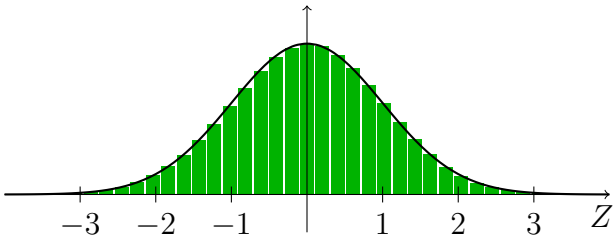
$$z_{0.95} \doteq 1{\cdot}645\,, \quad z_{0.975} \doteq 1{\cdot}960\,, \quad z_{0.99} \doteq 2{\cdot}326\,, \quad z_{0.995} \doteq 2{\cdot}576\,.$$

Splošneje, za poljuben $0 < \gamma < 1$ velja $z_\gamma = \Phi^{-1}\bigl(\gamma - \frac{1}{2}\bigr)$ oziroma $\gamma = \Phi(z_\gamma) + \frac{1}{2}$. Zaradi narave stvari, ki jih računamo, so vsi kvantili zaokroženi navzgor.

Ničelno hipotezo zavrnemo:

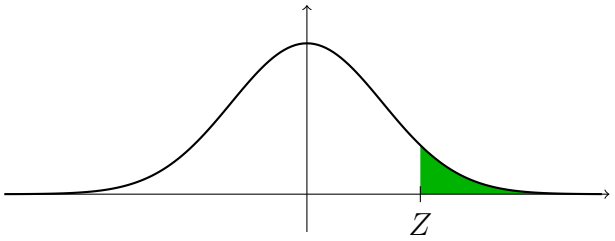
- proti $H_1^{\pm}$, če je $|Z| > z_{1-\alpha/2}$;
- proti H_1^+, če je $Z > z_{1-\alpha}$;
- proti H_1^-, če je $Z < -z_{1-\alpha}$.

Množici Z-vrednosti, kjer ničelno hipotezo zavrnemo, imenujemo *kritično območje*. Odvisna je od stopnje značilnosti α in različice testa (enostranski, dvostranski). Test se imenuje Z-test, ker ima testna statistika v primeru veljavnosti ničelne hipoteze približno standardno normalno porazdelitev in zato izračunano vrednost primerjamo s kvantili te porazdelitve, ki jih označujemo s črko z. Koliko je porazdelitev blizu od standardni normalni, ilustrirajmo s stolpčnim grafikonom dejanske porazdelitve in krivuljo standardne normalne porazdelitve pri $n = 100$ in $\theta^* = 0{\cdot}6$:

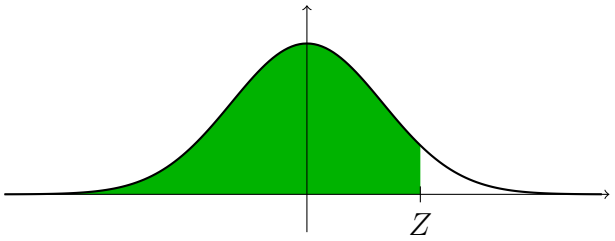


Ilustracija p-vrednosti pri isti opaženi testni statistiki Z za različne različice Z-testa – p-vrednost je ploščina osenčenega dela:

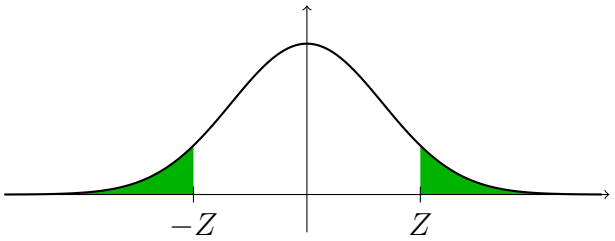
Enostranski test v desno (alternativna hipoteza je H_1^+):



Enostranski test v levo (alternativna hipoteza je H_1^-):

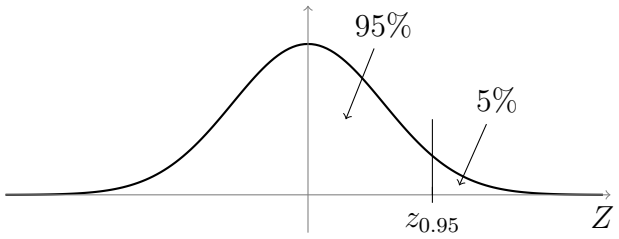


Dvostranski test (alternativna hipoteza je $H_1^\pm$):

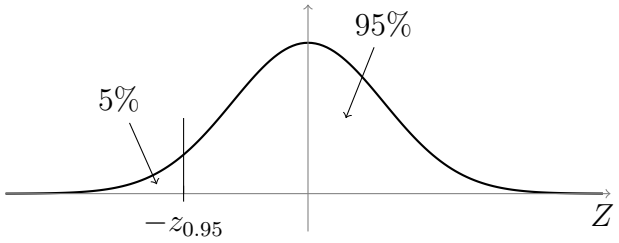


Ilustracija kritičnega območja za $\alpha = 0{\cdot}05$:

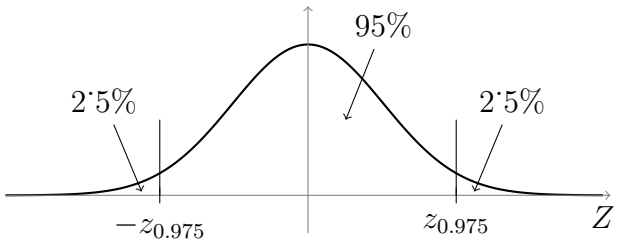
Enostranski test v desno
(alternativna hipoteza je H_1^+):



Enostranski test v levo
(alternativna hipoteza je H_1^-):



Dvostranski test
(alternativna hipoteza je $H_1^\pm$):



Primer: anketa Dela Stik v zvezi s predsedniškimi volitvami. Ob predpostavki, da je Delo Stik vzelo enostavni slučajni vzorec, testiramo ničelno hipotezo, da je bila opredeljenost volivcev ob anketiranju enaka kot opredeljenost na volitvah, proti alternativni hipotezi, da temu ni bilo tako: izvedemo torej dvostranski test. Opredeljenost lahko opišemo z deležem volivcev, ki so glasovali za Pahorja. V tem primeru je torej:

θ = delež vseh volivcev, ki so bili ob izvedbi ankete opredeljeni za Pahorja
Tega ne poznamo.

θ^* = delež volivcev, ki so na volitvah glasovali za Pahorja
$= 0{\cdot}6737\,,$

$\hat{\theta}$ = delež anketirancev, ki so se opredelili za Pahorja, med tistimi, ki so se opredelili
$= 0{\cdot}6962\,.$

Če se je opredelilo 621 anketirancev, izračunamo:

$$\mathrm{SE} \doteq \sqrt{\frac{0{\cdot}6737 \cdot 0{\cdot}3263}{621}} \doteq 0{\cdot}0188\,, \quad Z \doteq \frac{0{\cdot}6963 - 0{\cdot}6737}{0{\cdot}0188} \doteq 1{\cdot}201\,.$$

Tako dobimo $p \doteq 1 - 2 \cdot \Phi(1{\cdot}201) \doteq 0{\cdot}23$. Ker je $p \geq 0{\cdot}05$, ničelne hipoteze pri $\alpha = 0{\cdot}05$ ne moremo zavrniti, kaj šele, da bi jo zavrnili pri $\alpha = 0{\cdot}05$. Odstopanja torej niso statistično značilna. To se vidi tudi iz tega, da je $Z < z_{0.975} \doteq 1{\cdot}960$.

Primer. Recimo spet, da Loterija trdi, da je polovica srečk dobitnih. Kupimo določeno število srečk, med katerimi je spet določeno število dobitnih. Ali lahko trdimo, da Loterija

laže? V skladu s trditvijo Loterije bomo postavili $\theta^* = 1/2$, za alternativno hipotezo pa bomo postavili, da je $\theta < 1/2$, kjer je θ verjetnost, da je posamezna srečka dobitna: primer, ko je ta verjetnost večja od $1/2$, nas ne skrbi, zato ga v alternativno hipotezo ne vključimo (izpeljava pokaže, da dobimo isti kriterij odločanja tudi, če za ničelno hipotezo postavimo $\theta \geq \theta^*$). Izvedemo torej enostranski test v levo.

Denimo, da smo kupili 100 srečk in je dobitnih le 41. Izračunamo:

$$\mathrm{SE} = \sqrt{\frac{0{\cdot}5 \cdot 0{\cdot}5}{100}} = 0{\cdot}05\,, \quad Z = \frac{0{\cdot}41 - 0{\cdot}5}{0{\cdot}05} = -1{\cdot}8\,,$$

od koder sledi $p = \frac{1}{2} + \Phi(1{\cdot}8) \doteq 0{\cdot}036$. Ničelno hipotezo torej pri $\alpha = 0{\cdot}05$ zavrnemo, pri $\alpha = 0{\cdot}01$ pa tega ne moremo storiti. Z drugimi besedami, velja $Z < -1{\cdot}65$, toda $Z > -2{\cdot}33$, zato so odstopanja statistično značilna, niso pa statistično zelo značilna. Še drugače povedano, če smo pripravljeni sprejeti 5-odstotno tveganje, da Loterijo obtožimo po krivici, bomo rekli, da Loterija laže, če pa smo pripravljeni sprejeti le 1-odstotno tveganje, bomo molčali.

Primer. Kdaj lahko na podlagi določenega števila metov kovanca trdimo, da ni pošten? Tu spet postavimo $\theta^* = 1/2$, toda zdaj moramo biti občutljivi na obe strani: za alternativno hipotezo postavimo $\theta \neq 1/2$. Recimo, da 100-krat vržemo kovanec in 41-krat pade grb. Tedaj je še vedno $Z = -1{\cdot}8$ (če delamo s popravkom za zveznost, pa pride $Z = 1{\cdot}7$), toda p-vrednost je zdaj enaka $1 - 2\Phi(1{\cdot}8) \doteq 0{\cdot}072$ (ravno dvakratnik prejšnje, to pa je zato, ker smo občutljivi na dve strani). Z drugimi besedami, velja $|Z| < 1{\cdot}96$. To pomeni, da odstopanja niso statistično značilna.

Če pa bi kovanec vrgli 1000-krat in bi 410-krat padel grb, bi bilo $\mathrm{SE} = 0{\cdot}05/\sqrt{10} \doteq 0{\cdot}0158$ in $Z = -1{\cdot}8 \cdot \sqrt{10} \doteq 5{\cdot}59$. V tem primeru bi bila odstopanja zalo značilna. Iz tabele 1 se da razbrati, da pride p-vrednost manj kot $0{\cdot}00005$ (v resnici pride $6{\cdot}3 \cdot 10^{-9}$).

Primer: Gallupova napoved volilnega izida predsedniških volitev v ZDA leta 1936. V skladu z uradnim izidom postavimo $\theta^* = 0{\cdot}38$ (gledamo delež tistih, ki so glasovali za Landona) in v skladu z Gallupovo napovedjo postavimo $\hat{\theta} = 0{\cdot}44$. Izvedemo dvostranski test. Spomnimo se, da je Gallup povprašal $n = 50.000$ volivcev. Izračunamo:

$$\mathrm{SE} = \sqrt{\frac{0{\cdot}38 \cdot 0{\cdot}62}{50000}} \doteq 0{\cdot}00217\,, \quad Z = \frac{0{\cdot}44 - 0{\cdot}38}{0{\cdot}00217} \doteq 27{\cdot}64\,,$$

Pogled v tabelo 1 pove, da je p-vrednost manjša od $0{\cdot}00005$ (v resnici je celo manjša od 10^{-100}). Odstopanja so torej statistično več kot zelo značilna, torej lahko tudi za Gallupa rečemo, da je bil njegov vzorec pristranski. Tudi Gallupov inštitut ni pravilno napovedal izidov vseh predsedniških volitev v ZDA.

Za θ smo tu vzeli delež tistih, ki so glasovali za Landona. Enako bi dobili, tudi če bi gledali delež tistih, ki so glasovali za Roosevelta.

Primer: Gallupova napoved izida ankete revije Literary Digest. V skladu z izidom ankete postavimo $\theta^* = 0{\cdot}571$. Za Gallupovo napoved 56% avtor žal ni našel natančnejših

podatkov, torej vemo le, da je bil Gallupov delež za Landona med 55˙5% in 56˙5%. Izvedemo dvostranski test, pri čemer se spomnimo, da je Gallup povprašal $n = 3.000$ volivcev. Najprej velja:

$$\mathrm{SE} = \sqrt{\frac{0{\cdot}571 \cdot 0{\cdot}429}{3000}} \doteq 0{\cdot}009036\,.$$

Nadalje za $\hat{\theta} \doteq 0{\cdot}555$ dobimo:

$$Z \doteq \frac{0{\cdot}555 - 0{\cdot}571}{0{\cdot}00936} \doteq -1{\cdot}77\,, \quad p \doteq 0{\cdot}077\,,$$

za $\hat{\theta} \doteq 0{\cdot}565$ dobimo:

$$Z \doteq \frac{0{\cdot}555 - 0{\cdot}571}{0{\cdot}00936} \doteq -0{\cdot}66\,, \quad p \doteq 0{\cdot}51\,.$$

Vidimo, da v okviru razpoložljivih podatkov p-vrednost močno variira, vendar odstopanja v nebenem primeru niso statistično značilna.

2.2 Imenske spremenljivke

2.2.1 Frekvenčna porazdelitev

Če ima imenska spremenljivka, ki jo gledamo, fiksen končen nabor možnih vrednosti, je podobno kot pri dihotomni vrednosti smiselno govoriti o frekvencah, torej kolikokrat se je pojavila določena vrednost. Namesto tega lahko povemo tudi relativne frekvence (deleže), torej frekvence, deljene s številom enot. Zapis vseh vrednosti skupaj z (relativnimi) frekvencami imenujemo *frekvenčna porazdelitev*, ki jo lahko predstavimo v obliki tabele:

vrednosti	frekvence	relativne frekvence
a_1	f_1	f_1°
a_2	f_2	f_2°
$\vdots$	$\vdots$	$\vdots$
a_k	f_k	f_k°

Frekvenca f_i je torej število enot, na katerih ima spremenljivka vrednost a_i. Število enot z določeno lastnostjo bomo označevali z znakom $\sharp$. Tako lahko s formulo zapišemo:

$$f_i = \sharp(X = a_i)\,; \quad i = 1, 2, \ldots, k\,.$$

Velja še:

$$f_1 + f_2 + \cdots + f_k = n\,, \quad f_i^\circ = \frac{f_i}{n}\,, \quad f_1^\circ + f_2^\circ + \cdots + f_k^\circ = 1\,.$$

Frekvenčno porazdelitev imenskih spremenljivk grafično predstavimo s *tortnim diagramom* (angl. *pie chart* ali *circle graph*) ali s *histogramom*.

Če se naši podatki nanašajo na enostavni slučajni vzorec iz neke populacije, so relativne frekvence tudi točkaste ocene populacijskih deležev. Če so torej $\theta_1, \theta_2, \ldots, \theta_k$ deleži enot, na katerih ima spremenljivka vrednost $a_1, a_2, \ldots, a_k$, so njihove ocene kar $\hat{\theta}_i = f_i^\circ$.

Modus je vrednost z najvišjo frekvenco. Označevali ga bomo z M, pogosta oznaka pa je tudi M_o ali Mo. Modusov je lahko več.

Modus je ena od *mer centralne tendence*.

Primer: 32 ljudi so vprašali, kaj v življenju jim največ pomeni.[3] Možni odgovori so bili:

(D) Družina, otroci, starši.
(F) Denar, finančna neodvisnost.
(Z) Zabava, sprostitev.
(H) Hiša, avto, dobre obleke.
(U) Ugled, spoštovanje.

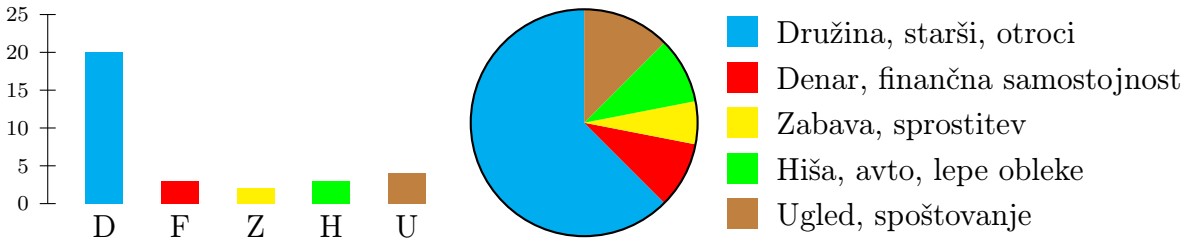
Odgovori, ki so jih dobili:[4]

F, D, D, U, Z, D, D, D, U, D, D, D, H, D, D, D, F, F, D, U, D, D, H, H, D, D, D, D, D, D, U, Z.

Frekvenčna porazdelitev:

vrednosti	frekvence	relativne frekvence
Družina, otroci, starši	20	0·625 = 62·5%
Denar, finančna samostojnost	3	0·094 = 9·4%
Zabava, sprostitev	2	0·063 = 6·3%
Hiša, avto, dobre obleke	3	0·094 = 9·4%
Ugled, spoštovanje	4	0·125 = 12·5%

Histogram in tortni grafikon:



Modus je ‘družina, starši, otroci’.

[3] `www.anketnik.net`, 9. 9. 2010–9. 3. 2011
[4] Vrstni red je izmišljen.

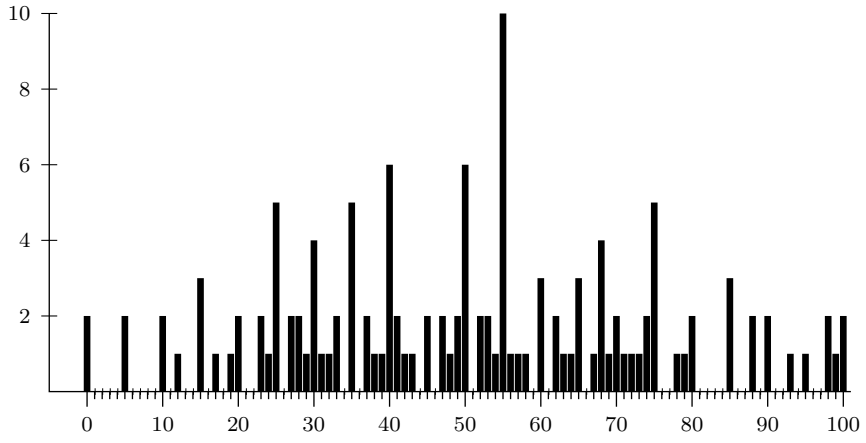
Če vemo, katere vrednosti so si blizu oz. sosedne, lahko definiramo tudi *lokalne moduse*. Porazdelitev je *bimodalna*, če ima dva izrazita lokalna modusa, pri čemer enake frekvence na sosednih vrednostih obravnavamo kot *en* lokalni modus. Porazdelitev je *multimodalna*, če ima več kot dva izrazita lokalna modusa v prejšnjem smislu.

Včasih modusi, posebej lokalni, za prvotne vrednosti ne odražajo realne slike. To se zgodi takrat, ko je vrednosti veliko, frekvence pa majhne.

Primer: rezultati kolokvija iz uvoda (urejeni po velikosti):

0, 0, 5, 5, 10, 10, 12, 15, 15, 15, 17, 19, 20, 20, 23, 23, 24, 25, 25, 25, 25, 25, 27, 27, 28, 28, 29, 30, 30, 30, 30, 31, 32, 33, 33, 35, 35, 35, 35, 35, 37, 37, 38, 39, 40, 40, 40, 40, 40, 40, 41, 41, 42, 43, 45, 45, 47, 47, 48, 49, 49, 50, 50, 50, 50, 50, 50, 52, 52, 53, 53, 54, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 56, 57, 58, 60, 60, 60, 62, 62, 63, 64, 65, 65, 65, 67, 68, 68, 68, 68, 69, 70, 70, 71, 72, 73, 74, 74, 75, 75, 75, 75, 75, 78, 79, 80, 80, 85, 85, 85, 88, 88, 90, 90, 93, 95, 98, 98, 99, 100, 100

Pisalo je 131 študentov, možno pa je bilo zbrati od 0 do 100 točk. Histogram po rezultatih je videti takole:

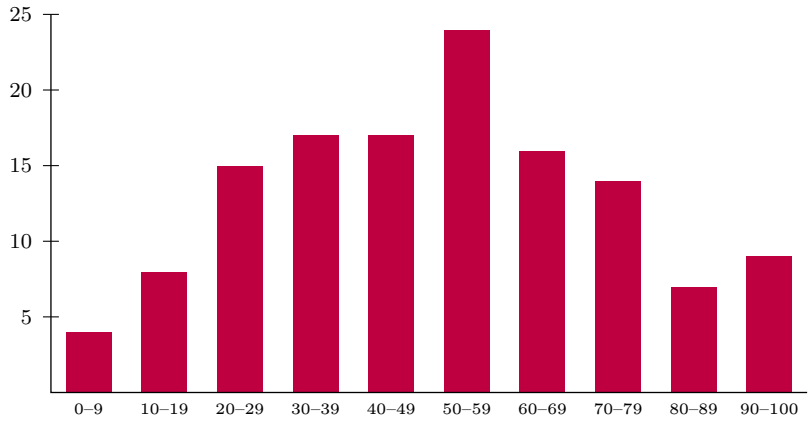


in ni pretirano ilustrativen – je zelo "nažagan". Tudi modus (55) je lahko zavajajoč.

A če vemo, katere vrednosti so si blizu oz. sosedne, jih lahko združimo v *razrede* in na njih gledamo frekvenčno porazdelitev. Na ta način navadno dobimo dosti ilustrativnejši histogram. Obstaja več kriterijev, kako veliki naj bodo razredi.

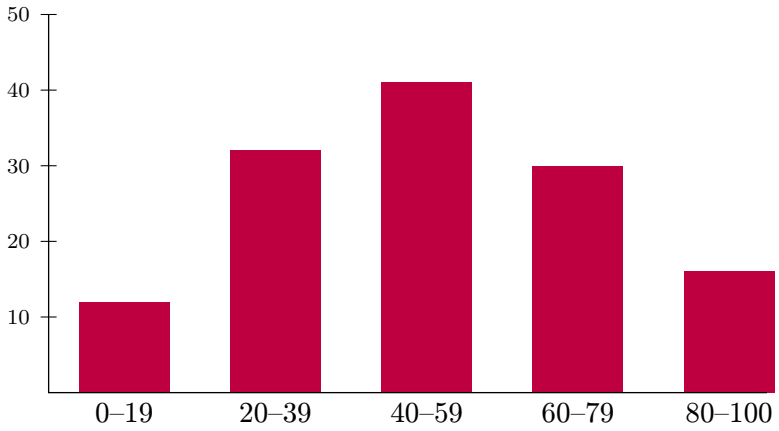
- Najbolj grobo je *korensko pravilo*, po katerem se naredi približno $\sqrt{n}$ razredov po približno $\sqrt{n}$ enot. Izkaže se, da pride pri velikih statističnih množicah histogram preveč "nažagan" – delitev je prefina.
- Če želimo manj "nažagan" histogram, uporabimo *pravilo tretjega korena*, po katerem se naredi približno $n^{1/3}$ razredov po približno $n^{2/3}$ enot. Pri tem pravilu pride histogram približno enako "nažagan" ne glede na število enot.
- Za višje merske lestvice obstajajo še bolj sofisticirana pravila – glej razdelek o intervalskih spremenljivkah.

Primer: razdelimo v razrede podatke iz prejšnjega primera. Če uporabimo korensko pravilo, izračunamo $\sqrt{131} \doteq 11\dot{}44$. Še malo zaokrožimo in razdelimo podatke na 10 razredov v razponu po 10 točk. Dobimo:



Slika je mnogo boljša. Namesto modusa ima pomen *modalni razred* od 50 do 59 točk.

Oglejmo si še, kaj dobimo, če uporabimo pravilo tretjega korena. Izračunamo $131^{1/3} \doteq 5\dot{}08$ in se odločimo, da podatke razdelimo v 5 razredov v razponu po 20 točk. Dobimo:



Histogram ima pravilnejšo obliko, a je tudi bolj grob in morda skrije kakšno podrobnost.

2.2.2 Točkasto ocenjevanje in test skladnosti

Recimo, da se podatki, ki smo jih dobili, nanašajo na vzorec iz neke populacije, ki je približno reprezentativen. Ocenjujemo količine $\theta_1, \theta_2, \ldots, \theta_k$, kjer je θ_i populacijski delež vrednosti i. Tako kot pri dihotomnih spremenljivkah to ocenimo z *opaženim* oz. *vzorčnim* deležem oz. relativno frekvenco f_i°:

$$\theta_i \approx \hat{\theta}_i = f_i^\circ = \frac{f_i}{n}\,.$$

Tako npr. na podlagi ankete iz prejšnjega primera ocenimo, da približno $62{\cdot}5\%$ največ pomeni družina.

Če je vzorčenje asimptotično reprezentativno, se, ko večamo vzorec, opažene relativne frekvence f_i° bližajo populacijskim deležem θ_i.

Podobno je, če so opažene vrednosti dobljene iz realizacij določenega slučajnega poskusa. Tedaj spremenljivki, ki jo opazujemo, pravimo *slučajna spremenljivka*. Poskus mora vsakič potekati po istih verjetnostnih zakonitostih. Tak slučajni poskus je recimo met kocke, slučajna spremenljivka pa je recimo število pik, ki padejo na tej kocki. V tem primeru je θ_i *verjetnost*, da je slučajna spremenljivka enaka i-ti vrednosti. Naboru verjetnosti za vse možne vrednosti pravimo *verjetnostna porazdelitev* te slučajne spremenljivke, porazdelitvi, ki jo dobimo iz opaženih vrednosti, pa pravimo *opažena* ali *empirična* porazdelitev. Podobno kot pri vzorčenju bomo tudi za zaporedje izvedb poskusa rekli, da je *asimptotično reprezentativno*, če se empirična porazdelitev z večanjem števila izvedb bliža verjetnostni. To se zgodi, če je možnih vrednosti končno mnogo in so izvedbe poskusa med seboj *verjetnostno neodvisne* – to je spet zakon velikih števil.

Tako kot pri dihotomnih spremenljivkah se da tudi tu pod istimi dodatnimi konstruirati tudi intervale zaupanja, vendar v tem primeru verjetnost pokritosti velja le za *posamezno* vrednost, ne pa za vse hkrati. Da se sicer konstruirati splošnejše *množice zaupanja*, katerih elementi so vektorji deležev (torej porazdelitve na populaciji) in s tem doseči pravo verjetnost pokritosti, vendar se tu s tem ne bomo ukvarjali.

Da pa se testirati ničelno hipotezo o določeni porazdelitvi na populaciji oz. o verjetnostih možnih izidov določenega slučajnega poskusa. Natančneje, testiramo hipotezo, da je $\theta_1 = \theta_1^*, \theta_2 = \theta_2^*, \ldots, \theta_k = \theta_k^*$, alternativna hipoteza pa trdi, da temu ni tako. Ničelna hipoteza torej trdi, da so *dejanski populacijski deleži ali verjetnosti* $\theta_1, \theta_2, \ldots, \theta_n$ enaki *hipotetičnim populacijskim deležem ali verjetnostim* $\theta_1^*, \theta_2^*, \ldots, \theta_n^*$.

Napačno bi bilo testirati vsako vrednost posebej, saj verjetnost, da ničelno hipotezo zavrnemo, čeprav velja, velja samo za izvedbo posameznega testa, ne pa tudi za izvedbo vseh testov hkrati. Če bi ničelno hipotezo o celotni porazdelitvi zavrnili, brž ko bi zavrnili vsaj eno hipotezo o posamezni vrednosti, bi bila stopnja tveganja, t. j. verjetnost, da ničelno hipotezo zavrnemo, čeprav velja, višja od stopnje tveganja za posamezen test.

Ena rešitev bi bila sicer, da bi stopnje tveganja pri testih za posamezne vrednosti ustrezno prilagodili; kako prilagoditi, bi morali natančno izračunati, kar pa ni prav lahko. Boljša rešitev je enoten, *omnibusni* test. Za to ničelno hipotezo je primeren *Pearsonov*[5]

[5]Karl Pearson (1857–1936), angleški matematik in biostatistik

test skladnosti, ki je prav tako kot Z-test deleža preprost, a žal ni eksakten. Zahteva tudi podobne pogoje kot test deleža:

- Gre za enostavni slučajni vzorec iz velike populacije oz. za neodvisne izvedbe poskusa, pri čemer le-ta vsakič sledi istim verjetnostnim zakonitostim.
- Vzorec oz. število izvedb poskusa ni premajhno. Za tipično dogovorjeno natančnost se zahteva, da je $n \geq 30$.
- *Pričakovane frekvence* $n\theta_1^*, n\theta_2^*, \ldots, n\theta_2^*$ niso premajhne. Za tipično dogovorjeno natančnost se zahteva, da je $n\theta_i^* \geq 5$ za vse i. Če to ni res, si lahko pomegamo tako, da združimo bližnje vrednosti.

Pearsonov test skladnosti temelji na testni statistiki *hi kvadrat* (angl. *chi-squared*):

$$\chi^2 = n \sum_{i=1}^{k} \frac{(\hat{\theta}_i - \theta_i^*)^2}{\theta_i^*} = n \sum_{i=1}^{k} \frac{(f_i^\circ - \theta_i^*)^2}{\theta_i^*} = \sum_{i=1}^{k} \frac{(f_i - n\theta_i^*)^2}{n\theta_i^*} \,.$$

Zgornje izraze lahko razumemo na naslednji način:

- Prvi izraz primerja *ocenjene deleže oz. verjetnosti* $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ s *hipotetičnimi deleži oz. verjetnostmi* $\theta_1^*, \theta_2^*, \ldots, \theta_k^*$.
- Drugi izraz dobimo iz prvega, če upoštevamo, da so ocenjeni deleži oz. verjetnosti v resnici kar *opažene relativne frekvence* $f_1^\circ, f_2^\circ, \ldots, f_k^\circ$.
- Tretji izraz primerja *opažene frekvence* $f_1, f_2, \ldots, f_k$ s *pričakovanimi frekvencami* $n\theta_1^*, n\theta_2^*, \ldots, n\theta_k^*$.

Ničelno hipotezo zavrnemo, če je $\chi^2 > \chi^2_{1-\alpha}(k-1)$. Na desni je kvantil porazdelitve hi kvadrat s $k-1$ prostostnimi stopnjami za verjetnost $1-\alpha$ in je torej *kritična vrednost*. Kvantile lahko odčitamo iz tabele 3.

V ozadju kritičnih vrednosti je, da ima, če velja ničelna hipoteza, testna statistika približno porazdelitev hi kvadrat s $k-1$ prostostnimi stopnjami. Zato takemu testu pravimo tudi test hi kvadrat. S testi hi kvadrat se bomo še srečali.

V primeru, ko imamo le dve možni vrednosti (t. j. dihotomno spremenljivko), je Pearsonov test skladnosti ekvivalenten dvostranskemu Z-testu deleža (če hipotezo zavrnemo pri enem testu, jo zavrnemo tudi pri drugem).

Primer: predčasne volitve v Sloveniji dne 4. 12. 2011. Agencija Mediana je izvedla vzporedne volitve, na katerih je povprašala $n = 16.200$ volivcev. Rezultati ankete skupaj z uradnimi rezultati volitev so prikazani spodaj.

Stranka	Vzporedne volitve	Uradni rezultat
Lista Zorana Jankovića – Pozitivna Slovenija	29˙08%	28˙51%
Slovenska demokratska stranka	26˙54%	26˙19%
Socialni demokrati	10˙79%	10˙52%
Lista Gregorja Viranta	8˙66%	8˙37%
Demokratična stranka upokojencev Slovenije	6˙70%	6˙97%
Slovenska ljudska stranka	6˙38%	6˙83%
Nova Slovenija	4˙70%	4˙88%
Drugi	7˙15%	7˙73%

Na Medianini spletni strani[6] piše, da je bila njihova napoved NATANČNA. Ujemanje je res precejšnje, a tudi število vprašanih volivcev je bilo veliko. Je napoved res natančna v smislu inferenčne statistike?

Privzemimo, da je Mediana vzorčila korektno, da pa morda volivci pri odgovorih niso bili iskreni. Na volivcih gledamo dve statistični spremenljivki: ena je glas na volitvah, druga pa je odgovor pri anketi. Ničelna hipoteza bo trdila, da sta obe spremenljivki enaki. V tem primeru bo:

θ_i = delež kandidatov, ki bi se, če bi bili vprašani, izrekli za i-to listo
Tega ne poznamo!

$\hat{\theta}_i$ = delež anketirancev, ki so se v anketi izrekli za i-to listo

θ_i^* = delež volivcev, ki so na volitvah glasovali za i-to listo

Izračunajmo:

$$\chi^2 = 16200\left[\frac{(0{\cdot}2908-0{\cdot}2851)^2}{0{\cdot}2851}+\frac{(0{\cdot}2654-0{\cdot}2619)^2}{0{\cdot}2619}+\frac{(0{\cdot}1079-0{\cdot}1052)^2}{0{\cdot}1052}+\right.$$
$$+\frac{(0{\cdot}0866-0{\cdot}0837)^2}{0{\cdot}0837}+\frac{(0{\cdot}0670-0{\cdot}0697)^2}{0{\cdot}0697}+\frac{(0{\cdot}0638-0{\cdot}0683)^2}{0{\cdot}0683}+$$
$$\left.+\frac{(0{\cdot}0470-0{\cdot}0488)^2}{0{\cdot}0488}+\frac{(0{\cdot}0715-0{\cdot}0773)^2}{0{\cdot}0773}\right] \doteq$$
$$\doteq 19{\cdot}98\,.$$

Ker je 8 skupin, imamo $df = 7$ prostostnih stopenj. Kvantil porazdelitve hi kvadrat za verjetnost 0˙99 je $\chi^2_{0.99}(7) \doteq 18{\cdot}48$. To pomeni, da hipotezo, da je Medianin vzorec nepristranski, zavrnemo tudi pri stopnji značilnosti $\alpha = 0{\cdot}01$: odstopanja so zelo značilna. Mediana se torej v tem primeru ne bi smela preveč hvaliti z natančnostjo (tudi če ni sama kriva, da volivci niso odgovarjali iskreno).

[6] `http://www.mediana.si/novice/?stran=1#CmsC58E1C449E3`, presneto dne 7. 3. 2013

2.3 Urejenostne spremenljivke

2.3.1 Ranžirna vrsta, rangi

Vrednosti urejenostne spremenljivke lahko uredimo po velikosti – razvrstimo v *ranžirno vrsto*:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}\,.$$

Ranžirna vrsta je natančno določena z zgornjim pogojem in s tem, da se njena frekvenčna porazdelitev ujema s frekvenčno porazdelitvijo statistične spremenljivke. Elementu $x_{(i)}$ pravimo *i-ta vrstilna statistika* (angl. *order statistic*).

Rang dane vrednosti je njen položaj v ranžirni vrsti: rang vrednosti x je enak i, če je $x = x_{(i)}$.

Če so vse vrednosti spremenljivke X različne in je vrednost x zavzeta, je njen rang natančno določen. Označimo ga z $R(x)$. V tem primeru velja:

$$R(x) = \sharp(X \leq x) = \sharp(X < x) + 1\,.$$

Primer: če izmerjene vrednosti:

$$x_1 = 4,\ \ x_2 = 2,\ \ x_3 = 75,\ \ x_4 = 42,\ \ x_5 = 15,\ \ x_6 = 63$$

razvrstimo v ranžirno vrsto, dobimo:

$$x_{(1)} = 2,\ \ x_{(2)} = 4,\ \ x_{(3)} = 15,\ \ x_{(4)} = 42,\ \ x_{(5)} = 63,\ \ x_{(6)} = 75$$

in velja:

$$R(2) = 1\,,\quad R(4) = 2\,,\quad R(15) = 3\,,\quad R(42) = 4\,,\quad R(63) = 5\,,\quad R(75) = 6\,.$$

Rangi ostalih vrednosti (še) niso definirani.

Če vrednosti spremenljivke niso vse različne, govorimo o *vezeh* (angl. *ties*): vez je skupek dveh ali več enot, na katerih ima spremenljivka enako vrednost. Če so prisotne vezi, rang ni nujno natančno določen.

Primer: naj bo $A < B < C < D < E$ in naj bo dana ranžirna vrsta podatkov:

$$A,\ B,\ B,\ B,\ B,\ C,\ D,\ D,\ D,\ E\,.$$

Očitno je $R(A) = 1$, $R(C) = 6$ in $R(E) = 10$. Rang vrednosti B je lahko 2, 3, 4 ali 5, rang vrednosti D pa 7, 8 ali 9.

Vsem možnim rangom vrednosti x pravimo *surovi rangi*. *Spodnji rang* je najnižji, *zgornji rang* pa najvišji možni surovi rang. Velja:

$$\begin{aligned}\text{spodnji rang} &= \sharp(X < x) + 1\,,\\ \text{zgornji rang} &= \sharp(X \leq x)\,.\end{aligned}$$

Spodnji in zgornji rang lahko definiramo za poljubno, ne le zavzeto vrednost. *Vezani rang* je aritmetična sredina spodnjega in zgornjega ranga in oznaka $R(x)$ bo zadevala to število:

$$R(x) = \frac{\text{spodnji rang} + \text{zgornji rang}}{2} = \frac{\sharp(X < x) + \sharp(X \leq x) + 1}{2}\,.$$

Tako je v zgornjem primeru $R(A) = 1$, $R(B) = 3{\cdot}5$, $R(C) = 6$, $R(D) = 8$ in $R(E) = 10$. Če bi namesto $A, \ldots, E$ imeli števila, npr.:

$$21,\ 27,\ 27,\ 27,\ 27,\ 28,\ 29,\ 29,\ 29,\ 32\,,$$

bi veljalo npr. $R(27) = 3{\cdot}5$, $R(30) = R(31) = 9{\cdot}5$, $R(20) = 0{\cdot}5$ in $R(40) = 10{\cdot}5$.

Relativni ali tudi *kvantilni rang* je definiran po predpisu:

$$r(x) = \frac{R(x) - \frac{1}{2}}{n}\,.$$

in ne glede na vezi velja:

$$r(x) = \frac{\sharp(X < x) + \sharp(X \leq x)}{2n}\,.$$

V prejšnjem primeru bi tako veljalo $r(27) = 0{\cdot}3$, $r(30) = r(31) = 0{\cdot}9$, $r(20) = 0$ in $r(40) = 1$.

Relativni rang pove položaj posamezne vrednosti glede na skupino.

Primer: oglejmo si rezultate dveh kolokvijev:

Ambrož	83
Blaž	22
Cvetka	61
Darja	45
Emil	49

Florjan	84
Gal	86
Helena	71
Iva	67
Jana	67
Karmen	88
Lev	89
Mojca	64

in se vprašajmo, kdo je glede na svoje kolege pisal bolje: Cvetka ali Gal?

Cvetka ima rang 4 in relativni rang $3{\cdot}5/5 = 0{\cdot}7$, Gal pa ima rang 6 in relativni rang $5{\cdot}5/8 = 0{\cdot}6875$, kar je skoraj enako.

2.3.2 Kumulativne frekvence

Če ima urejenostna spremenljivka, ki jo gledamo, fiksen končen nabor možnih vrednosti, lahko spet gledamo frekvenčno porazdelitev. Vrednosti uredimo po velikosti:

$$a_1 < a_2 < \cdots < a_k$$

ter dodamo še *kumulativne frekvence* in *relativne kumulativne frekvence*:

$$F_i = \sharp(X \le a_i) = f_1 + f_2 + \cdots + f_i\,, \qquad F_i^\circ = \frac{F_i}{n} = f_1^\circ + f_2^\circ + \cdots + f_i^\circ\,.$$

To lahko spet predstavimo v tabeli:

vrednosti	frekvence	relativne frekvence	kumulativne frekvence	relativne kumulativne frekvence
			$F_0 = 0$	$F_0^\circ = 0$
a_1	f_1	f_1°	$F_1 = f_1$	$F_1^\circ = f_1^\circ$
a_2	f_2	f_2°	$F_2 = F_1 + f_2$	$F_2^\circ = F_1^\circ + f_2^\circ$
a_3	f_3	f_3°	$F_3 = F_2 + f_3$	$F_3^\circ = F_2^\circ + f_3^\circ$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	f_k	f_k°	$F_k = F_{k-1} + f_k = n$	$F_k^\circ = F_{k-1}^\circ + f_k^\circ = 1$

Primer: ocene s kolokvijev pri predmetu Verjetnost in statistika na univerzitetnem študiju matematike na UL FMF v študijskem letu 2010/11:

ocena	f_i	F_i	f_i°	F_i°
neg.	25	25	0.391	0.391
6	13	38	0.203	0.594
7	12	50	0.188	0.781
8	7	57	0.109	0.891
9	3	60	0.047	0.938
10	4	64	0.063	1

Iz frekvenčne porazdelitve lahko odčitamo vrstilne statistike, in sicer velja:

$$x_{(i)} = a_j\,, \quad \text{če je} \quad 1 + F_{j-1} \le i \le F_j\,.$$

Pri določanju i-te vrstilne statistike moramo torej pogledati prvo kumulativno frekvenco, ki je enaka vsaj i.

Nekaj vrstilnih statistik iz prejšnjega primera: $x_{(40)} = 7$, $x_{(60)} = 9$, $x_{(61)} = 10$.

Iz kumulativnih frekvenc lahko odčitamo tudi range: vrednost a_j ima surove range od $1 + F_{j-1}$ do F_j in vezani rang:

$$R(a_j) = \frac{F_{j-1} + F_j + 1}{2} = F_{j-1} + \frac{f_j + 1}{2}\,.$$

Seveda so vezani rangi definirani tudi za vrednosti, ki niso zavzete: če je $a < a_1$, je $R(a) = 1/2$; če je $a > a_k$, je $R(a) = n + 1/2$. Za $a_{j-1} < a < a_j$ pa je $R(a) = F_{j-1} + 1/2$.

Rangi ocen pri prejšnjem primeru:

$$R(\text{neg.}) = 13\,,\ R(6) = 32\,,\ R(7) = 44{\cdot}5\,,\ R(8) = 54\,,\ R(9) = 59\,,\ R(10) = 62{\cdot}5\,.$$

Podobno lahko iz (relativnih) kumulativnih frekvenc odčitamo tudi relativne range:

$$r(a_j) = \frac{F_{j-1} + F_j}{2n} = \frac{F^\circ_{j-1} + F^\circ_j}{2} = F^\circ_{j-1} + \frac{f^\circ_j}{2}\,.$$

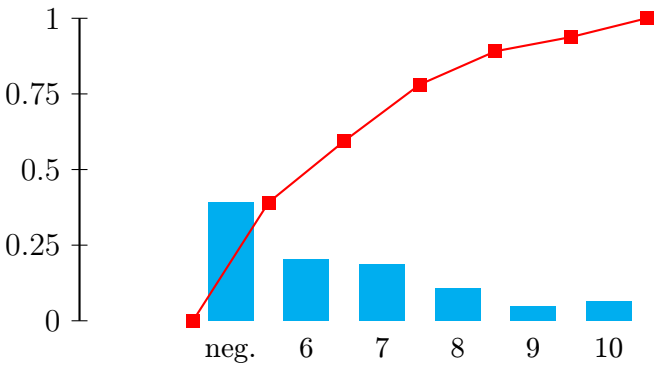
Poleg tega za $a < a_1$ velja $r(a) = 0$, za $a > a_k$ velja $r(a) = 1$, za $a_{j-1} < a < a_j$ pa je $r(a) = F^\circ_{j-1}$.

Relativni rangi ocen pri prejšnjem primeru:

$$\begin{aligned} r(\text{neg.}) &\doteq 0{\cdot}195\,, & r(6) &\doteq 0{\cdot}492\,, & r(7) &\doteq 0{\cdot}688\,,\\ r(8) &\doteq 0{\cdot}844\,, & r(9) &\doteq 0{\cdot}914\,, & r(10) &\doteq 0{\cdot}969\,. \end{aligned}$$

Tako kot pri imenskih spremenljivkah lahko tudi tu porazdelitev prikažemo grafično. Tortni grafikon je za urejenostne spremenljivke manj primeren, saj se iz njega ne vidi urejenost. Primerna pa sta histogram in črtni grafikon (angl. *line chart, line graph*). Prikažemo lahko razredne in kumulativne frekvence (absolutne ali relativne). Kadar kumulativne frekvence prikazujemo s črtnim grafikonom, vozle postavimo *vmes* med vrednosti. Takemu črtnemu grafikonu pravimo pravimo *ogiva*, tudi *oživa* (angl., fr. *ogive*, v prvotnem pomenu gotski lok). Če so vozli točno na sredini med vrednostmi in so prikazane relativne frekvence, višina črte nad posamezno vrednostjo ustreza relativnemu rangu.

Histogram iz razrednih relativnih frekvenc in ogiva pri prejšnjem primeru:



2.3.3 Kvantili

Kvantil je površno povedano meja med ustreznim zgodnjim in spodnjim delom statistične množice glede na dano statistično spremenljivko. Malo natančneje, kvantil statistične spremenljivke za določen delež je vrednost, pod katero leži približno dani delež podatkov.

- Kvantilu za delež 1/2 pravimo *mediana* in jo bomo označevali z m. Pogosta oznaka je tudi M_e ali Me. Mediana torej razdeli statistično množico na dve polovici: približno polovica vrednosti leži pod njo, približno polovica pa nad njo. Zato ji pravimo tudi *srednja vrednost* in je *mera centralne tendence*. Pri dihotomnih spremenljivkah je mediana enaka modusu.

- Kvantila za deleža $1/3$ in $2/3$ sta prvi in drugi *tercil*. Tercila torej razdelitev statistično množico na tri približno enako velike dele: približno tretjina vrednosti leži pod prvim tercilom, približno tretjina med prvim in drugim tercilom in približno tretjina nad drugim tercilom.
- Kvantili za deleže $1/4$, $1/2$ in $3/4$ so *kvartili*. Drugi kvartil je torej mediana.
- Kvantilom za deleže $0{\cdot}1, 0{\cdot}2, \ldots, 0{\cdot}9$ pravimo *decili*.
- Kvantilom za deleže $0{\cdot}01, 0{\cdot}02, \ldots, 0{\cdot}99$ pravimo *centili* ali tudi *percentili*. 1., 5., 95. in 99. percentil so pomembni v inferenčni statistiki, ker na njih temeljijo dogovorjeni pojmi. Pomembni so tudi $q_{0.005}$, $q_{0.025}$, $q_{0.975}$ in $q_{0.995}$.

Precizna definicija kvantila pa žal ni enotna. Tu bomo podali matematično definicijo kvantila.

Vrednost q_γ je kvantil statistične spremenljivke X za delež γ, če velja:

$$\frac{\sharp(X < q_\gamma)}{n} \leq \gamma \quad \text{in} \quad \frac{\sharp(X \leq q_\gamma)}{n} \geq \gamma\,.$$

Primer: dana je ranžirna vrsta:

$$10,\ 10,\ 20,\ 30,\ 50,\ 80,\ 130,\ 210,\ 340,\ 550\,.$$

Kvantil $q_{0.49}$ mora izpolnjevati pogoja:

$$\sharp(X < q_{0.49}) \leq 4{\cdot}9 \quad \text{in} \quad \sharp(X \leq q_{0.49}) \geq 4{\cdot}9\,.$$

Prvi pogoj izpolnjujejo vrednosti do vključno 50, drugega pa vrednosti od vključno 50 naprej. Torej je 50 edini možni kvantil za delež $0{\cdot}49$.

Kvantil $q_{0.5}$, torej mediana, pa mora izpolnjevati pogoja:

$$\sharp(X < q_{0.5}) \leq 5 \quad \text{in} \quad \sharp(X \leq q_{0.5}) \geq 5\,.$$

Prvi pogoj izpolnjujejo vrednosti do vključno 80, drugega pa vrednosti od vključno 50 naprej. Torej je vsako število iz intervala $[50, 80]$ lahko kvantil za delež $0{\cdot}5$. To je *kvantilni interval* za ta delež.

Kvantil $q_{0.1}$ mora izpolnjevati pogoja:

$$\sharp(X < q_{0.1}) \leq 1 \quad \text{in} \quad \sharp(X \leq q_{0.1}) \geq 1\,.$$

Prvi pogoj izpolnjujejo vrednosti do vključno 10, drugega pa vrednosti od vključno 10 naprej. Torej je 10 edini možni kvantil za delež $0{\cdot}1$.

Lastnosti kvantilov:

- Za vsak $\gamma \in [0, 1]$ obstaja kvantil dane spremenljivke za delež γ.

- Kvantili niso nujno enolično določeni.
- Če sta q' in q'' kvantila za isti delež γ ter velja $q' \leq q \leq q''$, je tudi q kvantil za ta delež.

Kvantil za delež γ je vrednost s kvantilnim rangom *približno* γ. Velja tudi, da je vrednost, ki ima kvantilni rang γ, kvantil za delež γ. Sicer pa lahko kvantile (za deleže, ki niso nujno kvantilni rangi) dobimo iz vrstilnih statistik, in sicer:

- Kvantil za delež 0 je katero koli število iz $(-\infty, x_{(1)}]$.
- Kvantil za delež 1 je katero koli število iz $[x_{(n)}, \infty)$.
- Če je $0 < \gamma < 1$ in je $n\gamma$ celo število, je kvantil za delež γ katero koli število iz intervala $[x_{(n\gamma)}, x_{(n\gamma+1)}]$. Dobljeni kvantilni interval bomo pisali tudi kot $[q_\gamma^-, q_\gamma^+]$, krajišči pa imenovali *spodnji* in *zgornji* kvantil.
- Če je $0 < \gamma < 1$ in $n\gamma$ ni celo število, je kvantil za delež γ enolično določen, in sicer je enak $x_{(\lceil n\gamma \rceil)}$ (oznaka $\lceil h \rceil$ tukaj pomeni h, zaokrožen navzgor). V tem primeru bomo postavili $q_\gamma^- = q_\gamma^+ = q_\gamma$.

Primer: pri ranžirni vrsti:

$$10,\ 10,\ 20,\ 30,\ 50,\ 80,\ 130,\ 210,\ 340,\ 550$$

je mediana kar koli iz $[x_{(5)}, x_{(6)}] = [50, 80]$ (kar smo že ugotovili), tretji kvartil pa je $x_{(8)} = 210$. Prav tako je natančno določen prvi decil, saj je $x_{(1)} = x_{(2)} = 10$.

Vrednost 20 ima kvantilni rang $0{\cdot}25$ in je zato tudi kvantil za delež $0{\cdot}25$; kvantil za ta delež je enolično določen. Prav tako pa je enolično določen tudi kvantil za delež $0{\cdot}26$, prav tako je enak 20, vendar $0{\cdot}26$ *ni* kvantilni rang vrednosti 20.

Pri sodem številu podatkov mediana tipično ni natančno določena:

pri lihem pa je:

Še en primer z rezultati 50 meritev, kjer je s sivo označen interval za 9. decil:

Primer: pri ocenah s kolokvijev so vsi kvartili natančno določeni. Prvi kvartil je sicer res na intervalu $[x_{(16)}, x_{(17)}]$, mediana na $[x_{(32)}, x_{(33)}]$ in tretji kvartil $[x_{(48)}, x_{(49)}]$, toda $x_{(16)} = x_{(17)} =$ neg., $x_{(32)} = x_{(33)} = 6$ in $x_{(48)} = x_{(49)} = 7$, zato lahko zapišemo $q_{1/4} =$ neg., $m = 6$ in $q_{3/4} = 7$.

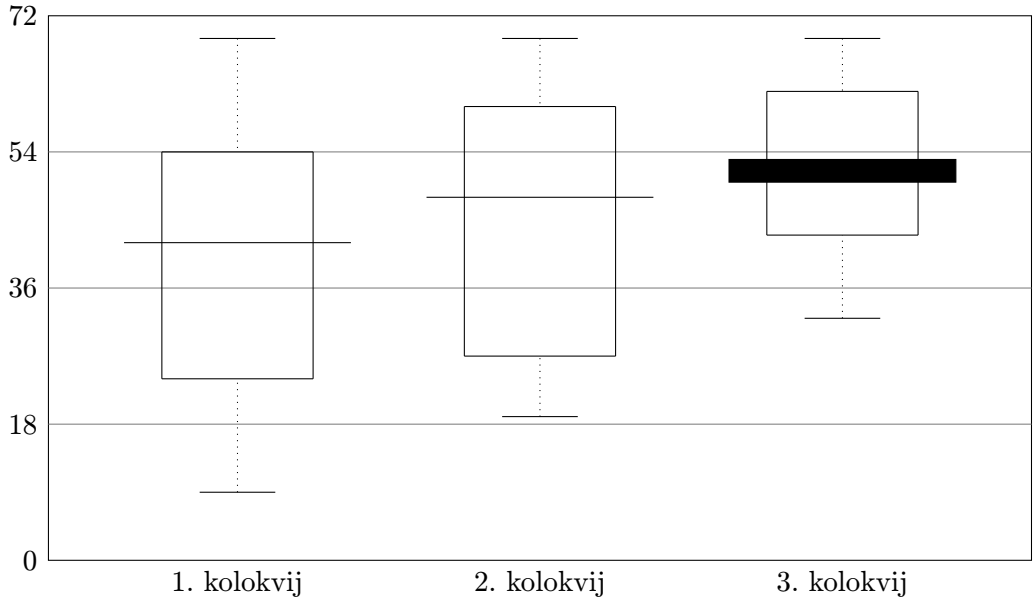
Statistike urejenostnih spremenljivk lahko grafično predstavimo s *škatlo z brki* (angl. *box plot*). Navadno narišemo minimalno vrednost, kvartile in maksimalno vrednost, lahko pa tudi kakšne druge statistike.

Primer: rezultati kolokvijev iz matematike na univerzitetnem študiju gozdarstva na UL BTF v študijskem letu 2004/05 (prikazani so minimalna vrednost, kvartili in maksimalna vrednost):

1. kolokvij: 9, 11, 12, 14, 17, 17, 24, 24, 26, 30, 34, 35, 36, 37, 42, 42, 44, 45, 49, 50, 51, 54, 57, 62, 63, 65, 65, 68, 69

2. kolokvij: 19, 19, 20, 24, 27, 27, 36, 45, 47, 47, 48, 48, 49, 57, 57, 60, 61, 63, 64, 65, 69

3. kolokvij: 32, 32, 39, 42, 43, 47, 49, 50, 50, 53, 53, 56, 60, 62, 68, 68, 69, 69



2.3.4 Točkasto ocenjevanje karakteristik

Denimo zdaj, da se podatki, ki smo jih dobili, nanašajo na vzorec iz določene populacije. Želeli bi oceniti vse karakteristike, ki smo jih obravnavali do sedaj in ki so "stabilne", ko se populacija veča: populacijske relativne range, relativne kumulativne frekvence in kvantile. Vrednosti ustreznih statistik *na vzorcu* bomo označevali kot doslej, za vrednosti karakteristik *na populaciji* pa bomo uporabili nove oznake v skladu z naslednjo tabelo:

	na vzorcu	na populaciji
relativni rang, ki pripada vrednosti x	$r(x)$	$\rho(x)$
relativna kumulativna frekvenca, ki pripada vrednosti a_i	F_i°	Φ_i
kvantil za delež γ	q_γ	Q_γ

Pri kvantilih oznaka velja za kateri koli kvantil za delež γ, a če je populacija velika, so kvantili navadno zelo natančno določeni.

Na enak način obravnavamo tudi primer, ko so podatki dobljeni kot realizacije določenega slučajnega poskusa. Za točno definicijo je treba poznati osnove teorije verjetnosti. Količina $\rho(x)$ je v tem primeru srednja kumulativna porazdelitvena funkcija v x, t. j. $\rho(x) = \frac{1}{2}\big[\mathbb{P}(X < x) + \mathbb{P}(X \leq x)\big]$, količina Φ_i je zgornja kumulativna porazdelitvena funkcija za vrednost a_i, t. j. $\Phi_i = \mathbb{P}(X \leq a_i)$, Q_γ pa je kvantil ustrezne verjetnostne porazdelitve za verjetnost γ.

Če so opažene vrednosti dobljene iz vsaj približno reprezentativnega vzorca, je prvi dve karakteristiki na populaciji smiselno oceniti z ustreznima statistikama na vzorcu:

$$\hat{\rho}(x) = r(x)\,, \quad \hat{\Phi}_i = F_i^\circ\,.$$

To lahko storimo tudi za (zgornje in spodnje) kvantile, vendar za določene posebne primere obstajajo tudi druge smiselne cenilke, ki imajo prednosti pred vzorčnimi kvantili samimi po sebi. *Omejili se bomo na intervalske spremenljivke, katerih vrednosti na populaciji oz. verjetnostni porazdelitvi so ustrezno razpršene* (natančneje, pomembno je, da so zagotovljene vrednosti na dovolj majhnih intervalih, od koder sledi tudi, da so kvantili zelo natančno določeni). Kvantil za delež γ bomo tedaj ocenili na naslednji način, ki ga uporablja tudi excel:

- Izračunamo $h = (n-1)\gamma + 1$.
- Naj bo k celi del števila h.
- Cenilka za Q_γ je $\hat{Q}_\gamma = x_{(k)} + (h-k)(x_{(k+1)} - x_{(k)})$.

Točkasta ocena za mediano po zgornji metodi je natančno $\frac{1}{2}(m^- + m^+)$, torej sredina medianskega intervala.

To, kar smo tukaj definirali kot oceno za populacijski kvantil na podlagi opaženih podatkov, se često šteje kar kot (pravi) kvantil spremenljivke na danih podatkih. Eden od razlogov je gotovo ta, da je ocena neka točno določena vrednost, medtem ko kvantil po matematični definiciji ni nujno natančno določen.

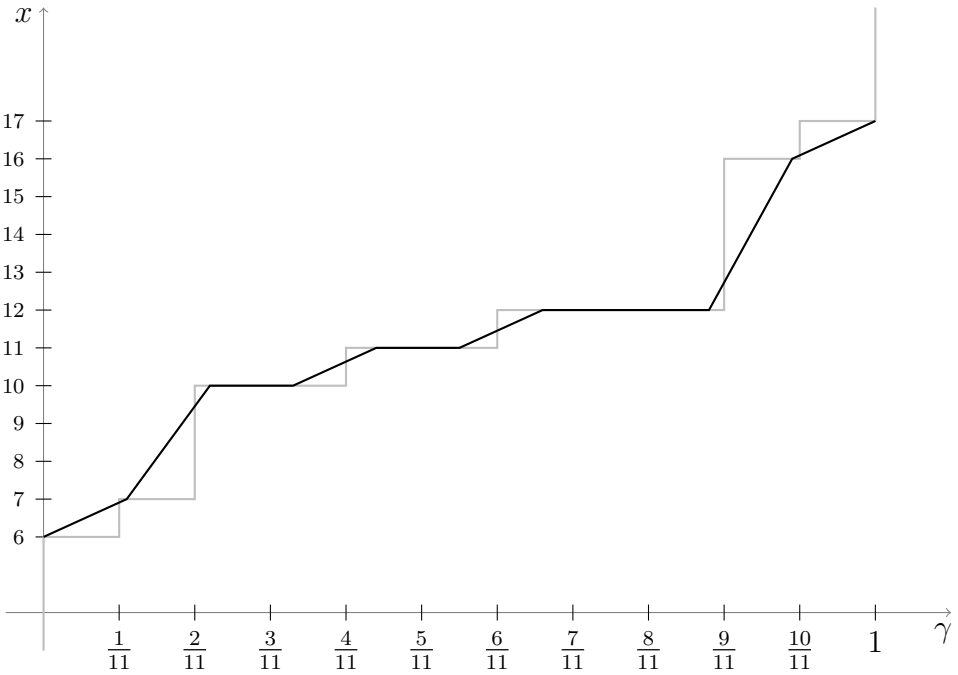
Če so opažene vrednosti dobljene iz vzorca in če je vzorčenje asimptotično reprezentativno, se, ko večamo velikost vzorca, ocene omenjenih populacijskih karakteristik bližajo dejanskim vrednostim. Podobno velja, če so opažene vrednosti dobljene iz izvedb slučajnega poskusa, ki so med seboj verjetnostno neodvisne. V tem primeru se tudi empirična porazdelitev bliža verjetnostni porazdelitvi. Natančneje, empirična kumulativna porazdelitvena funkcija se bliža kumulativni porazdelitveni funkciji verjetnostne porazdelitve. To dejstvo se imenuje *Glivenko*[7]*–Cantellijev*[8] *izrek* in sodi med zakone velikih števil.

Primer: vzemimo ranžirno vrsto iz vzorca velikosti 11:

$$6,\ 7,\ 10,\ 10,\ 11,\ 11,\ 12,\ 12,\ 12,\ 16,\ 17.$$

[7]Valerij Ivanovič Glivenko (1897–1940), ukrajinski matematik
[8]Francesco Paolo Cantelli (1875–1966), italijanski matematik

Oglejmo si npr. 17. percentil. Vrednost na vzorcu je enolično določena: iz $\lceil 11 \cdot 0{\cdot}17 \rceil = 2$ dobimo $q_{0.17} = x_{(2)} = 7$. Za oceno populacijskega prvega kvartila pa dobimo $h = 2{\cdot}7$, $k = 2$ in $\hat{Q}_{0{\cdot}17} = x_{(2)} + 0{\cdot}7 \cdot (x_{(3)} - x_{(2)}) = 9{\cdot}1$. Na spodnjem grafu je s sivo prikazana vzorčna kvantilna funkcija, s črno pa ocena populacijske kvantilne funkcije:



2.3.5 Intervalsko ocenjevanje karakteristik

Omejili se bomo na predpostavke, kot smo jih navedli že za dihotomne in imenske spremenljivke. Če so torej podatki dobljeni kot vzorec, privzamemo, da gre za enostavni slučajni vzorec iz velike populacije. Če pa so bili dobljeni iz realizacij slučajnega poskusa, privzamemo, da so izvedbe verjetnostno neodvisne in da poskus vsakič sledi istim verjetnostnim zakonitostim. Obakrat privzamemo, da število podatkov ni premajhno (za tipično dogovorjeno natančnost to pomeni $n \geq 30$).

Kumulativna frekvenca je v resnici delež enot, na katerih statistična spremenljivka ne presega dane vrednosti, zato jo tudi ocenjujemo tako kot delež. Relativni rang pa je povprečje dveh deležev in tudi pri njem se lahko poslužimo istih metod. Vzeli bomo torej Waldov interval zaupanja.

Primer: ponovno ocene s kolokvijev pri predmetu Verjetnost in statistika na univerzitetnem študiju matematike na UL FMF v študijskem letu 2010/11:

ocena	f_i	F_i	f_i°	F_i°
neg.	25	25	0.391	0.391
6	13	38	0.203	0.594
7	12	50	0.188	0.781
8	7	57	0.109	0.891
9	3	60	0.047	0.938
10	4	64	0.063	1

Ocena 7 ima vzorčno relativno kumulativno frekvenco $F_3^\circ \doteq 0{\cdot}781$ in vzorčni relativni rang $r(7) = (0{\cdot}594 + 0{\cdot}781)/2 \doteq 0{\cdot}688$. Recimo sedaj, da bi bil to vzorec univerzitetnih študentov matematike, ki so kdaj koli pisali kolokvije iz verjetnosti in statistike (čeprav je pri vzorcu študentov, ki so pisali v določenem letu, reprezentativnost močno vprašljiva). Določimo 95% interval zaupanja za Φ_3 in $\rho(7)$ za primer, ko *bi bil* to enostavni slučajni vzorec iz velike populacije. Pri Φ_3 izračunamo:

$$\mathrm{SE} \doteq \sqrt{\frac{0{\cdot}781 \cdot 0{\cdot}219}{64}} \doteq 0{\cdot}0517\,, \quad F_3^\circ - z_{0.975} \cdot \mathrm{SE} \doteq 0{\cdot}6797\,, \quad F_3^\circ + z_{0.975} \cdot \mathrm{SE} \doteq 0{\cdot}8823$$

in po zaokrožitvi dobimo interval zaupanja:

$$0{\cdot}67 < \Phi_3 < 0{\cdot}89\,.$$

Pri $\rho(7)$ pa izračunamo:

$$\mathrm{SE} \doteq \sqrt{\frac{0{\cdot}688 \cdot 0{\cdot}312}{64}} \doteq 0{\cdot}0579\,, \quad r(7) - z_{0.975} \cdot \mathrm{SE} \doteq 0{\cdot}5745\,, \quad r(7) + z_{0.975} \cdot \mathrm{SE} \doteq 0{\cdot}8015$$

in po zaokrožitvi dobimo interval zaupanja:

$$0{\cdot}57 < \rho(7) < 0{\cdot}81\,.$$

Intervalsko ocenjevanje kvantilov pa je malo drugačno. Tu ne bomo privzeli predpostavke iz točkastega ocenjevanja, t. j. da gre za intervalsko spremenljivko z razpršenimi vrednostmi. Tako ni nujno, da so kvantili na populaciji oz. verjetnostni porazdelitvi natančno določeni. Intervali zaupanja bodo veljali za cel kvantilni interval, torej za zgornji in spodnji kvantil.

Če želimo poiskati interval zaupanja za $Q_\gamma^\pm$, najprej izračunamo:

$$\mathrm{SE} = \sqrt{\frac{\gamma(1-\gamma)}{n}}\,, \qquad \gamma_{\min} = \gamma - c \cdot \mathrm{SE}\,, \qquad \gamma_{\max} = \gamma - c \cdot \mathrm{SE}\,,$$

kjer je $c = z_{(1+\beta)/2}$ že dobro znani kvantil normalne porazdelitve. Interval zaupanja ima obliko:

$$q^-_{\gamma_{\min}} \le Q^-_\gamma \le Q^+_\gamma \le q^+_{\gamma_{\max}}\,.$$

Opomba. V nasprotju s točkastim ocenjevanjem tu nismo privzeli, da je spremenljivka intervalska.

Primer: izračunajmo 95% interval zaupanja za 6. decil populacije, iz katere dobimo enostavni slučajni vzorec:

$$10,\ 10,\ 20,\ 30,\ 50,\ 80,\ 130,\ 210,\ 340,\ 550\,.$$

Velja:

$$c \doteq 1{\cdot}96\,, \qquad \mathrm{SE} = \sqrt{\frac{0{\cdot}6 \cdot 0{\cdot}4}{10}} \doteq 0{\cdot}15492\,,$$
$$\gamma_{\min} = 0{\cdot}6 - 1{\cdot}96 \cdot 0{\cdot}15492 \doteq 0{\cdot}296\,, \qquad \gamma_{\max} = 0{\cdot}6 + 1{\cdot}96 \cdot 0{\cdot}15492 \doteq 0{\cdot}904\,,$$
$$q^-_{0.296} = x_{(3)} = 20\,, \qquad q^+_{0.904} = x_{(10)} = 550\,,$$

od koder sklepamo, da je:

$$20 \le Q^-_{0.6} \le Q^+_{0.6} \le 550\,.$$

2.3.6 Testiranje karakteristik

Tudi pri testiranju bomo privzeli iste pogoje kot pri intervalskem ocenjevanju. Podobno kot tam velja, da z relativnimi frekvencami in relativnimi rangi ravnamo tako kot z deleži – testiramo jih z Z-testom. Če testiramo ničelno hipotezo, da je $\Phi_i = \Phi_i^*$, izračunamo:

$$\mathrm{SE} = \sqrt{\frac{\Phi_i^*(1-\Phi_i^*)}{n}}\,, \qquad Z = \frac{F_i^\circ - \Phi_i^*}{\mathrm{SE}}\,.$$

Če testiramo ničelno hipotezo, da je $\rho(x) = \rho^*(x)$, izračunamo:

$$\mathrm{SE} = \sqrt{\frac{\rho^*(x)\big(1-\rho^*(x)\big)}{n}}\,, \qquad Z = \frac{r(x) - \rho^*(x)}{\mathrm{SE}}\,.$$

Nato testiramo tako, kot smo testirali delež θ.

Primer: če imamo dan vzorec rezultatov:

ocena	f_i	F_i	f_i°	F_i°
neg.	25	25	0.391	0.391
6	13	38	0.203	0.594
7	12	50	0.188	0.781
8	7	57	0.109	0.891
9	3	60	0.047	0.938
10	4	64	0.063	1

testiramo ničelno hipotezo, da je študentov, ki so pisali 6 ali manj, v celotni populaciji točno polovica, proti alternativni hipotezi, da jih je več kot pol. Izračunajmo:

$$\mathrm{SE} = \sqrt{\frac{0{\cdot}5 \cdot 0{\cdot}5}{64}} \doteq 0{\cdot}0625\,, \qquad Z = \frac{0{\cdot}594 - 0{\cdot}5}{0{\cdot}0625} \doteq 1{\cdot}52\,.$$

Ker je to manj od $z_{0.95} \doteq 1{\cdot}645$, odstopanja niso statistično značilna: ničelne hipoteze ne moremo zavrniti niti pri $\alpha = 0{\cdot}05$ niti pri $\alpha = 0{\cdot}01$.

Če pa testiramo ničelno hipotezo, da je relativni rang ocene 7 na populaciji enak točno $0{\cdot}5$, proti alternativni hipotezi, da je različen od $0{\cdot}5$, izračunamo:

$$\mathrm{SE} = \sqrt{\frac{0{\cdot}5 \cdot 0{\cdot}5}{64}} \doteq 0{\cdot}0625\,, \qquad Z = \frac{0{\cdot}688 - 0{\cdot}5}{0{\cdot}0625} \doteq 3$$

in dobimo, da so odstopanja statistično zelo značilna.

Pri testiranju hipotez o kvantilih pa je treba biti previdnejši. Če kvantil ni natančno določen, je treba sploh premisliti, kaj je ničelna in kaj alternativna hipoteza. Če gre za vzorec, ničelna hipoteza H_0 trdi, da je Q^*_γ populacijski kvantil za delež γ (če slednji ni natančno določen, je torej Q^*_γ eden od možnih populacijskih kvantilov za ta delež). Alternativna hipoteza pa je lahko:

- H_1^+, da je *vsak* populacijski kvantil za delež γ večji od Q^*_γ;
- H_1^-, da je *vsak* populacijski kvantil za delež γ manjši od Q^*_γ;
- $H_1^\pm$, da je *vsak* populacijski kvantil za delež γ bodisi večji bodisi manjši od Q^*_γ (kar je ekvivalentno izjavi, da je bodisi vsak populacijski kvantil za delež γ večji od Q^*_γ bodisi vsak populacijski kvantil za delež γ manjši od Q^*_γ).

V primeru, ko so podatki dobljeni iz izvedb slučajnega poskusa, namesto populacijskega kvantila pride kvantil verjetnostne porazdelitve, ki izhaja iz poskusa.

Za testiranje spet izračunamo $\mathrm{SE} = \sqrt{\gamma(1-\gamma)/n}$.

- Pri alternativni hipotezi H_1^+ postavimo $c = z_{1-\alpha}$ in ničelno hipotezo zavrnemo, če je $\frac{\sharp(X \le Q^*_\gamma)}{n} < \gamma - c \cdot \mathrm{SE}$ ali, ekvivalentno, če je $q^-_{\gamma - c\cdot\mathrm{SE}} > Q^*_\gamma$.
- Pri alternativni hipotezi H_1^- postavimo $c = z_{1-\alpha}$ in ničelno hipotezo zavrnemo, če je $\frac{\sharp(X < Q^*_\gamma)}{n} > \gamma + c \cdot \mathrm{SE}$ ali, ekvivalentno, če je $q^+_{\gamma + c\cdot\mathrm{SE}} < Q^*_\gamma$.
- Pri alternativni hipotezi $H_1^\pm$ pa postavimo $c = z_{1-\alpha/2}$ in ničelno hipotezo zavrnemo, če je $\frac{\sharp(X < Q^*_\gamma)}{n} > \gamma + c \cdot \mathrm{SE}$ ali $\frac{\sharp(X \le Q^*_\gamma)}{n} < \gamma - c \cdot \mathrm{SE}$. To je ekvivalentno dejstvu, da je $q^-_{\gamma - c\cdot\mathrm{SE}} > Q^*_\gamma$ ali pa $q^+_{\gamma + c\cdot\mathrm{SE}} < Q^*_\gamma$.

Pri tem se spomnimo na kvantile standardne normalne porazdelitve:

$$z_{0.95} \doteq 1{\cdot}645\,, \quad z_{0.975} \doteq 1{\cdot}960\,, \quad z_{0.99} \doteq 2{\cdot}326\,, \quad z_{0.995} \doteq 2{\cdot}576\,.$$

Primer: Pri prej omenjenih rezultatih kolokvijev pri stopnji značilnosti $\alpha = 0{\cdot}01$ testiramo hipotezo, da je mediana na populaciji enaka 8 (natančneje, da obstaja mediana, ki je enaka 8), proti alternativni hipotezi, da je manjša od 8. Velja $c = z_{0.99} \doteq 2{\cdot}326$ in $\mathrm{SE} = \sqrt{0{\cdot}5 \cdot 0{\cdot}5/64} = 0{\cdot}0625$. Nadalje je:

$$\frac{\sharp(X < 8)}{64} \doteq 0{\cdot}781 > 0{\cdot}5 + 2{\cdot}33 \cdot 0{\cdot}0625 \doteq 0{\cdot}646\,,$$

zato ničelno hipotezo zavrnemo. Odstopanja so torej statistično zelo značilna. To se vidi tudi iz dejstva, da je $q^+_{0.646} = 7 < 8$.

2.3.7 Primerjava parov: test z znaki

Naj bosta na vsaki enoti populacije definirani dve urejenostni spremenljivki: X in Y. Ničelna in alternativne hipoteze so v tem primeru zapletenejše. Lahko si jih predstavljamo tako, da so prisotni dejavniki, ki večajo X na račun Y, in dejavniki, ki večajo Y na račun X. Ničelna hipoteza H_0 pravi, da so ti dejavniki *uravnovešeni*. Če so podatki dobljeni kot vzorec iz populacije, to pomeni, da je delež enot na populaciji, kjer je $X > Y$, enak deležu enot, kjer je $Y > X$. Spet bomo obravnavali tri alternativne hipoteze.

- Alternativna hipoteza v korist spremenljivke X, H_1^X, pravi, da dejavniki, ki večajo X na račun Y, prevladujejo nad dejavniki, ki delajo nasprotno. Če so podatki dobljeni kot vzorec iz populacije, to pomeni, da je delež enot na populaciji, kjer je $X > Y$, večji od deleža enot, kjer je $Y > X$.
- Alternativna hipoteza v korist spremenljivke Y, H_1^Y, pravi, da dejavniki, ki večajo Y na račun X, prevladujejo nad dejavniki, ki delajo nasprotno. Če so podatki dobljeni kot vzorec iz populacije, to pomeni, da je delež enot na populaciji, kjer je $Y > X$, večji od deleža enot, kjer je $X > Y$.
- Dvostranska alternativna hipoteza $H_1^\pm$ pravi, da velja bodisi H_1^X bodisi H_1^Y.

Podobno formuliramo hipoteze tudi, če so podatki dobljeni iz realizacij slučajnega poskusa: namesto deležev nastopajo verjetnosti.

Test z znaki gleda stvari, ki v formulaciji ničelne in alternativnih hipotez nastopajo na populaciji oz. verjetnostni porazdelitvi, na vzorcu: na vsaki enot pogledamo, ali je spremenljivka X večja od Y, manjša od Y ali pa sta enaki. Od tod tudi ime testa.

Tudi tu se bomo omejili na predpostavke, kot smo jih navedli že v prejšnjem razdelku, ko smo testirali kvantile. Če so torej podatki dobljeni kot vzorec, privzamemo, da gre za enostavni slučajni vzorec iz velike populacije. Če pa so bili dobljeni iz izvedb slučajnega poskusa, privzamemo, da so le-te verjetnostno neodvisne in da poskus vsakič sledi istim verjetnostnim zakonitostim. Obakrat privzamemo, da število podatkov ni premajhno (za tipično dogovorjeno natančnost to pomeni $n \geq 30$).

Za ta primer lahko test z znaki kot Z-test: naj bo S_X število enot, za katere je $X > Y$, S_Y pa število enot, za katere je $X < Y$. Testna statistika je:

$$Z := \frac{S_X - S_Y}{\sqrt{S_X + S_Y}}$$

in ničelno hipotezo zavrnemo:

- proti H_1^X, če je $Z > z_{1-\alpha}$;
- proti H_1^Y, če je $Z < -z_{1-\alpha}$;

- proti $H_1^{\pm}$, če je $|Z| > z_{1-\alpha/2}$.

Primer: 50 ljudi so pred ogledom in po ogledu filma povprašali, kako se počutijo: zelo slabo, slabo, srednje, dobro ali zelo dobro. Rezultati so naslednji:[9]

pred	po
srednje	srednje
dobro	zelo dobro
srednje	zelo dobro
dobro	srednje
srednje	zelo dobro
dobro	dobro
srednje	dobro
dobro	dobro
dobro	zelo dobro
zelo dobro	zelo dobro
dobro	zelo dobro
zelo dobro	dobro
dobro	srednje
zelo dobro	srednje
srednje	dobro
srednje	dobro
dobro	zelo dobro
srednje	dobro
dobro	zelo dobro
zelo dobro	dobro
zelo dobro	zelo dobro
zelo dobro	dobro
slabo	dobro
dobro	srednje
srednje	zelo dobro

pred	po
dobro	zelo dobro
dobro	dobro
zelo dobro	zelo dobro
dobro	dobro
srednje	zelo slabo
srednje	zelo dobro
zelo dobro	srednje
dobro	dobro
dobro	dobro
srednje	slabo
slabo	srednje
srednje	srednje
zelo slabo	slabo
slabo	srednje
slabo	srednje
slabo	zelo dobro
zelo slabo	srednje
srednje	slabo
srednje	slabo
zelo slabo	srednje
srednje	dobro
slabo	zelo dobro
slabo	slabo
slabo	slabo
zelo slabo	srednje

Testirajmo ničelno hipotezo, da ogled filma ne spremeni počutja, proti alternativni hipotezi, da ga spremeni. Ko preštejemo, dobimo, da se je 12 ljudi pred ogledom počutilo boljše kot po ogledu, 25 ljudi pa po ogledu boljše kot pred ogledom; 13 ljudi se je pred in po ogledu počutilo enako. Testna statistika pride:

$$Z = \frac{12 - 25}{\sqrt{37}} \doteq -2{\cdot}14\,.$$

Če testiramo pri stopnji značilnosti $\alpha = 0{\cdot}05$, moramo $|Z|$ primerjati s kritično vrednostjo $z_{0.975} \doteq 1{\cdot}960$. Hipotezo zavrnemo, torej je ogled filma na naši skupini statistično značilno vplival na počutje. Če pa testiramo pri stopnji značilnosti $\alpha = 0{\cdot}01$, je kritična vrednost $z_{0.995} \doteq 2{\cdot}576$ in hipoteze ne zavrnemo: ogled ni vplival statistično zelo značilno.

[9]Dejansko so izmišljeni, dobljeni pa so s simulacijo mešanice dveh parov porazdelitev.

2.4 Intervalske spremenljivke

2.4.1 Mere centralne tendence

Mera centralne tendence za dano statistično spremenljivko nam grobo povedano da vrednost, proti kateri se nagibajo vrednosti te spremenljivke na statistični množici.

Dve meri centralne tendence smo že spoznali: pri imenskih spremenljivkah je bil to modus, pri urejenostnih pa mediana. Pri intervalskih spremenljivkah pa kot mero centralne tendence najpogosteje gledamo *aritmetično sredino* (angl. *arithmetic mean*):

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}\,.$$

Primer: temperature po Sloveniji v ponedeljek, 20. februarja 2012, ob 17. uri (v Celzijevih stopinjah):

$$-13,\ 2,\ 1,\ 5,\ 2\,.$$

Aritmetična sredina ali povprečje:

$$\bar{x} = \frac{-13 + 2 + 1 + 5 + 2}{5} = -0\dot{6}$$

To seveda ni verodostojna ocena za povprečno temperaturo vseh naseljenih krajev v Sloveniji, ker je reprezentativnost tega vzorca močno vprašljiva, a s tem se ne bomo ukvarjali. Kot zanimivost pa naj povemo, da bi bila verodostojnejša ocena za mediano: $m = 2$. Če bi namreč skrajno temperaturo -13 stopinj, ki je bila izmerjena na Kredarici, zamenjali npr. s temperaturo 0 stopinj, izmerjeno v kakšnem nižje ležečem kraju, bi za povprečje dobili 2 stopinji, mediana pa se ne bi spremenila. Več o skrajnih vrednostih pa malo kasneje.

Včasih je aritmetično sredino lažje izračunati po *u-metodi*: za poljubno izhodišče u velja:

$$\bar{x} = u + \frac{(x_1 - u) + (x_2 - u) + \cdots + (x_n - u)}{n}\,.$$

Razlikam $x_i - u$ pravimo tudi *odkloni* (angl. *deviations*).

u-metoda izkorišča dejstvo, da se, če vsem podatkom prištejemo neko število, tudi njihova aritmetična sredina poveča za to število. To velja tudi za modus in mediano.

Primer:

$$876,\ 879,\ 878,\ 878,\ 877\,.$$

Če za izhodišče vzamemo $u = 876$, dobimo:

$$\bar{x} = 876 + \frac{0 + 3 + 2 + 2 + 1}{5} = 876 + 1\dot{6} = 877\dot{6}\,.$$

2.4.2 Mere razpršenosti

Mere razpršenosti povedo, za koliko se posamezne vrednosti med seboj razlikujejo. Verjetno najpreprostejša izmed njih je kar razlika med največjo in najmanjšo vrednostjo. Tej pravimo *variacijski razmik* (angl. *range*):

$$\mathrm{VR} = \max - \min\,.$$

Variacijski razmik pa navadno ni najbolj verodostojna mera razpršenosti, saj ga lahko že ena skrajna vrednost znatno spremeni. Verodostojnejša in robustnejša mera je variacijski razmik *srednje polovice* podatkov, natančneje razlika med tretjim in prvim kvartilom, natančneje med zgornjim tretjim in spodnjim prvim kvartilom. Tej pravimo *interkvartilni razmik* (angl. *interquartile range, midspread, middle fifty*):

$$\mathrm{IQR} = q^+_{3/4} - q^-_{1/4}\,.$$

Lahko gledamo tudi *povprečni absolutni odklon* (*average absolute deviation*) od primerne referenčne vrednosti. Če le-to začasno označimo z u, dobimo količino:

$$\mathrm{AAD}_u = \frac{|x_1-u| + |x_2-u| + \cdots + |x_n-u|}{n}\,.$$

Ta količina je najmanjša, če za referenčno vrednost u vzamemo mediano. Zato je smiselno gledati povprečni absolutni odmik od mediane:

$$\mathrm{AAD}_m = \frac{|x_1-m| + |x_2-m| + \cdots + |x_n-m|}{n}\,.$$

Čeprav mediana ni natančno določena, je zgornja količina vedno natančno določena. Dostikrat za referenčno vrednost vzame tudi aritmetična sredina – dobimo:

$$\mathrm{AAD}_{\bar x} = \frac{|x_1-\bar x| + |x_2-\bar x| + \cdots + |x_n-\bar x|}{n}\,.$$

Najlepše računske lastnosti pa ima *standardni odklon*:

$$s = \sqrt{\frac{(x_1-\bar x)^2 + (x_2-\bar x)^2 + \cdots + (x_n-\bar x)^2}{n}}\,,$$

ki ga lahko izračunamo tudi po u-metodi:

$$s = \sqrt{\frac{(x_1-u)^2 + (x_2-u)^2 + \cdots + (x_n-u)^2}{n} - (\bar x-u)^2}\,.$$

Kvadratu standardnega odklona pravimo *varianca* ali *disperzija*.

Vse omenjene mere razpršenosti (VR, IQR, AAD_m, $\mathrm{AAD}_{\bar x}$ in s) ostanejo nespremenjene, če vsem podatkom prištejemo isto število.

Primer:

$$80,\ 80,\ 90,\ 110,\ 110,\ 130,\ 140,\ 140,\ 140,\ 370\,.$$

Prvi in tretji kvartil sta natančno določena: $q_{1/4} = 90$, $q_{3/4} = 140$, torej je $\mathrm{IQR} = 50$.

Mediana ni natančno določena – je namreč kar koli iz intervala $[110, 130]$. Toda kar koli iz tega intervala vzamemo za izračun AAD_m, dobimo isto:

$$\begin{aligned}
&\frac{1}{10}\Big[2\cdot|80-110|+|90-110|+2\cdot|110-110|+|130-110|+3\cdot|140-110|+\\
&\qquad\qquad +|370-110|\Big] = 45\,,\\
&\frac{1}{10}\Big[2\cdot|80-130|+|90-130|+2\cdot|110-130|+|130-130|+3\cdot|140-130|+\\
&\qquad\qquad +|370-130|\Big] = 45\,.
\end{aligned}$$

Nadalje velja:

$$\begin{aligned}
\bar{x} &= 140 + \frac{2\cdot(-60)+(-50)+2\cdot(-30)+(-10)+230}{10} = 139\,,\\
\mathrm{AAD}_{\bar{x}} &= \frac{1}{10}\Big[2\cdot|80-139|+|90-139|+2\cdot|110-139|+|130-139|+\\
&\qquad\qquad +3\cdot|140-139|+|370-139|\Big] = 46\dot{8}\,.
\end{aligned}$$

Končno je še:

$$\begin{aligned}
s &= \left[\frac{2\cdot(80-139)^2+(90-139)^2+2\cdot(110-139)^2+(130-139)^2+(370-139)^2}{10}\right]^{1/2} =\\
&= \Bigg[\frac{2\cdot(80-140)^2+(90-140)^2+2\cdot(110-140)^2+(130-140)^2+(370-140)^2}{10} -\\
&\qquad -(139-140)^2\Bigg]^{1/2} =\\
&= \sqrt{6449} \doteq 80\dot{3}\,.
\end{aligned}$$

Različne mere razpršenosti so različno občutljive za skrajne vrednosti. Manj občutljivim pravimo, da so *robustne*. Na naslednjem diagramu so mere razpršenosti razvrščene po robustnosti oz. občutljivosti:

$$\underset{\text{robustna}}{\mathrm{IQR}} \qquad \mathrm{AAD} \qquad s \qquad \underset{\text{občutljiva}}{\mathrm{VR}}$$

2.4.3 Izračun karakteristik iz frekvenčnih porazdelitev

Vse zgoraj omenjene količine preprosto dobimo iz frekvenčnih porazdelitev. Omenimo le izražavo aritmetične sredine:

$$\begin{aligned}\bar{x} &= \frac{1}{n}\big(f_1a_1 + f_2a_2 + \cdots + f_ka_k\big) = \\ &= f_1^\circ a_1 + f_2^\circ a_2 + \cdots + f_k^\circ a_k = \\ &= u + \frac{1}{n}\big(f_1(a_1-u) + f_2(a_2-u) + \cdots + f_k(a_k-u)\big) = \\ &= u + f_1^\circ(a_1-u) + f_2^\circ(a_2-u) + \cdots + f_k^\circ(a_k-u)\end{aligned}$$

in standardnega odklona:

$$\begin{aligned}s &= \sqrt{\frac{1}{n}\big(f_1(a_1-\bar{x})^2 + f_2(a_2-\bar{x})^2 + \cdots + f_k(a_k-\bar{x})^2\big)} = \\ &= \sqrt{f_1^\circ(a_1-\bar{x})^2 + f_2^\circ(a_2-\bar{x})^2 + \cdots + f_k^\circ(a_k-\bar{x})^2} = \\ &= \sqrt{\frac{1}{n}\big(f_1(a_1-u)^2 + f_2(a_2-u)^2 + \cdots + f_k(a_k-u)^2\big) - (u-\bar{x})^2} = \\ &= \sqrt{f_1^\circ(a_1-u)^2 + f_2^\circ(a_2-u)^2 + \cdots + f_k^\circ(a_k-u)^2 - (\bar{x}-u)^2}\,.\end{aligned}$$

Pravimo, da je $\bar{x}$ *tehtana sredina* vrednosti $a_1, a_2, \ldots, a_k$ z *utežmi* $f_1^\circ, f_2^\circ, \ldots, f_k^\circ$. V splošnem je tehtana sredina vsak izraz zgornje oblike, pri katerem so uteži nenegativne, njihova vsota pa je 1.

Primer: pozitivne ocene s kolokvijev pri predmetu Verjetnost in statistika na univerzitetnem študiju matematike na UL FMF v študijskem letu 2010/11:

ocena	f_i
6	13
7	12
8	7
9	3
10	4

Velja:

$$\bar{x} = \frac{13\cdot 6 + 12\cdot 7 + 7\cdot 8 + 3\cdot 9 + 4\cdot 10}{39} = \frac{285}{39} \doteq 7{\cdot}31\,.$$

Lahko računamo tudi po u-metodi:

$$\bar{x} = 8 + \frac{13\cdot(-2) + 12\cdot(-1) + 7\cdot 0 + 3\cdot 1 + 4\cdot 2}{39} = 8 - \frac{27}{39} \doteq 7{\cdot}31\,.$$

Za izračun standardnega odklona je navadno potreben kalkulator. Lahko računamo tako, da damo $\bar{x}$ v spomin in vtipkamo:

$$\begin{aligned}s &= \sqrt{\frac{13\cdot(6-\bar{x})^2 + 12\cdot(7-\bar{x})^2 + 7\cdot(8-\bar{x})^2 + 3\cdot(9-\bar{x})^2 + 4\cdot(10-\bar{x})^2}{39}} = \\ &\doteq \sqrt{1{\cdot}649} \doteq 1{\cdot}284\,,\end{aligned}$$

lahko pa računamo tudi po u-metodi:

$$s = \sqrt{\frac{13 \cdot (-2)^2 + 12 \cdot 1^2 + 7 \cdot 0^2 + 3 \cdot 1^2 + 42^2}{39} - (8 - \bar{x})^2} \doteq 1{\cdot}284\,.$$

Posebej preprosti so izračuni za dihotomne spremenljivke:

- Če spremenljivka zavzame le vrednosti 0 in 1, je aritmetična sredina enaka kar relativni frekvenci vrednosti 1.
- Če spremenljivka zavzame vrednost a z relativno frekvenco q, vrednost b pa z relativno frekvenco p, velja:

$$\mathrm{AAD}_m = \min\{p, q\}|b - a|\,, \quad \mathrm{AAD}_{\bar{x}} = 2pq|b - a|\,, \quad s = |b - a|\sqrt{pq}\,.$$

2.4.4 Standardizacija

Standardizacija je postopek, pri katerem od vrednosti odštejemo aritmetične sredine in jih delimo s standardnim odklonom. Dobimo *standardizirane vrednosti* ali vrednosti v *standardnih enotah* ali *z-vrednosti*:

$$z_i = \frac{x_i - \bar{x}}{s}\,.$$

Standardizirana vrednost ima podobno vlogo kot kvantilni rang, pove nam položaj posamezne vrednosti glede na skupino. Negativna standardizirana vrednost nam pove, da je vrednost pod povprečjem, pozitivna pa, da je nad povprečjem.

Standardizirane vrednosti nam omogočajo primerjavo različnih spremenljivk, recimo na isti enoti.

Primer: spet si oglejmo rezultate dveh kolokvijev:

Ambrož	83
Blaž	22
Cvetka	61
Darja	45
Emil	49

Florjan	84
Gal	86
Helena	71
Iva	67
Jana	67
Karmen	88
Lev	89
Mojca	64

in se vprašajmo, kdo je glede na svoje kolege pisal bolje: Cvetka ali Gal?

Če spremenljivko, ki predstavlja rezultat na prvem kolokviju, označimo z X, spremenljivko, ki predstavlja rezultat na drugem kolokviju, pa z Y, je pri prvem kolokviju $\bar{x} = 52$ in $s_X = 20$, torej je Cvetkina standardizirana vrednost:

$$\frac{61 - 52}{20} = 0{\cdot}45\,.$$

Cvetkin rezultat je bil torej za slabo polovico standardnega odklona nad povprečjem.

Pri drugem kolokviju pa je $\bar{y} = 77$ in $s_Y = 10$, torej je Galova standardizirana vrednost:

$$\frac{86 - 77}{10} = 0{\cdot}9\,.$$

Njegov rezultat je bil skoraj za cel standardni odklon nad povprečjem, torej je v smislu standardiziranih vrednosti glede na svojo skupino pisal bolje.

2.4.5 Skrajne vrednosti

Včasih določene vrednosti izstopajo – so *skrajne*, angl. *outliers*. Te vrednosti moramo včasih izločiti iz obravnave, saj lahko povzemanje in statistično sklepanje znatno popačijo. Tudi pri grafični predstavitvi podatkov jih je smiselno prikazati posebej. Za skrajne vrednosti bomo vzeli tiste, nižje od $q^-_{1/4} - 1{\cdot}5 \cdot \mathrm{IQR}$, in tiste, višje od $q^+_{3/4} + 1{\cdot}5 \cdot \mathrm{IQR}$.

Primer:

$$80,\ 80,\ 90,\ 110,\ 110,\ 130,\ 140,\ 140,\ 140,\ 370\,.$$

Izračunali smo že, da je $q_{1/4} = 90$, $q_{3/4} = 140$, torej $\mathrm{IQR} = 50$. Iz:

$$q^-_{1/4} - 1{\cdot}5 \cdot \mathrm{IQR} = 15 \quad \text{in} \quad q^+_{3/4} + 1{\cdot}5 \cdot \mathrm{IQR} = 215$$

dobimo, da spodaj skrajnih vrednosti ni, medtem ko imamo zgoraj eno skrajno vrednost 370.

Primer: razporeditev točk v svetovnem pokalu v alpskem smučanju za ženske, sezona 2012/13.[10] Tekmovalo je 116 smučark. Točke:

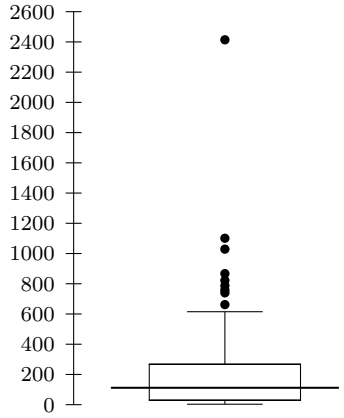
2414, 1101, 1029, 867, 822, 787, 759, 740, 662, 615, 512, 500, 460, 448, 435, 423, 406, 395, 381, 359, 349, 323, 323, 314, 310, 292, 273, 269, 269, 266, 264, 263, 261, 251, 236, 219, 215, 212, 209, 203, 198, 192, 180, 180, 172, 170, 162, 157, 156, 150, 148, 134, 127, 127, 127, 127, 125, 124, 115, 109, 109, 106, 104, 100, 95, 91, 80, 78, 74, 72, 69, 66, 60, 58, 53, 50, 44, 43, 39, 38, 36, 36, 33, 32, 32, 31, 30, 29, 28, 26, 24, 24, 22, 22, 21, 17, 16, 16, 15, 15, 15, 14, 13, 11, 10, 10, 9, 9, 8, 8, 6, 6, 6, 5, 3, 3

Iz:

$$q^-_{1/4} = x_{(29)} = 29\,, \quad q^+_{3/4} = x_{(88)} = 269\,, \quad \mathrm{IQR} = 240\,,$$
$$q^-_{1/4} - 1{\cdot}5 \cdot \mathrm{IQR} = -331\,, \quad q^+_{3/4} + 1{\cdot}5 \cdot \mathrm{IQR} = 629$$

dobimo, da pri dnu ni skrajnih vrednosti, medtem ko je pri vrhu devet skrajnih vrednosti. Če podatke prikažemo npr. s škatlo z brki, jih prikažemo posebej:

[10] Presneto 27. 3. 2013 s strani
`http://www.fis-ski.com/uk/disciplines/alpine-skiing/cupstandings.html`.



2.4.6 Združevanje vrednosti v razrede

Kadar je vrednosti veliko, frekvence pa so majhne (če so vrednosti zelo natančno izmerjene, se vsaka od njih tipično pojavi le enkrat), se splača vrednosti združevati v *razrede*. Pri tem obstajajo določena pravila:

- Razredi se ne smejo prekrivati.
- Pri imenskih spremenljivkah, pri katerih vemo, katere vrednosti so bližnje, morajo razredi obsegati bližnje vrednosti.
- Pri urejenostnih spremenljivkah mora vsak razred zajemati vrednosti iz določenega *intervala*. Paziti moramo na enoten dogovor, katera krajišča intervalov (spodnja, zgornja) razred vključuje in katerih ne.
- Pri intervalskih spremenljivkah lahko določamo *širine razredov*. Kadar to delamo, se morajo sosedni intervali stikati: zgornja meja prejšnjega razreda se mora ujemati s spodnjo mejo naslednjega. Meje so pomembne za določanje širine razredov (glej spodaj). Izbiramo čimbolj realistične meje: če so podatki, ki so na voljo, zaokroženi, poskusimo predvideti, iz katerih realnih vrednosti je lahko bila dobljena posamezna zaokrožena vrednost. Ne gre vedno za najbližjo vrednost – starost se zaokrožuje navzdol.
- Skrajne vrednosti prikažemo posebej.

Ni enotnega pravila, koliko razredov narediti oziroma kako široki naj bodo.

- V splošnem se lahko držimo že omenjenega *pravila tretjega korena*, po katerem podatke razdelimo na približno $n^{1/3}$ razredov po približno $n^{2/3}$ enot.
- Če želimo dobiti enako široke razrede, *Freedman*[11]*–Diaconisovo*[12] *pravilo* [20] pravi, naj bo širina posameznega razreda približno $2 \cdot \mathrm{IQR}/\sqrt[3]{n}$.

[11] David Amiel Freedman (1838–2008), ameriški statistik
[12] Persi Diaconis (1945), ameriški matematik in iluzionist grškega rodu

Če so razredi in z njimi tudi stolpci v histogramu različno široki, je pomembno, da so (relativne) frekvence sorazmerne *ploščinam* in ne širinam stolpcev. Višine stolpcev pa so sorazmerne *gostotam frekvenc* (angl. *frequency densities*). Če je f_i frekvenca, d_i pa širina i-tega razreda, je gostota frekvence njun kvocient, lahko pa definiramo tudi *relativno gostoto*:

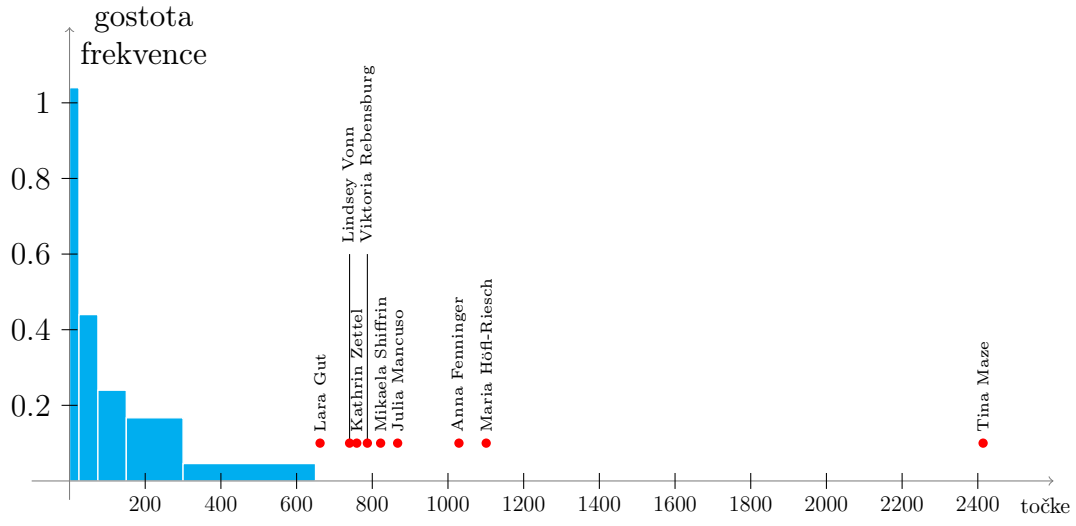
$$g_i = \frac{f_i}{d_i}\,, \quad g_i^\circ = \frac{f_i^\circ}{d_i}\,.$$

Zaradi lažje berljivosti gostote frekvenc često preračunamo na določeno širino razreda.

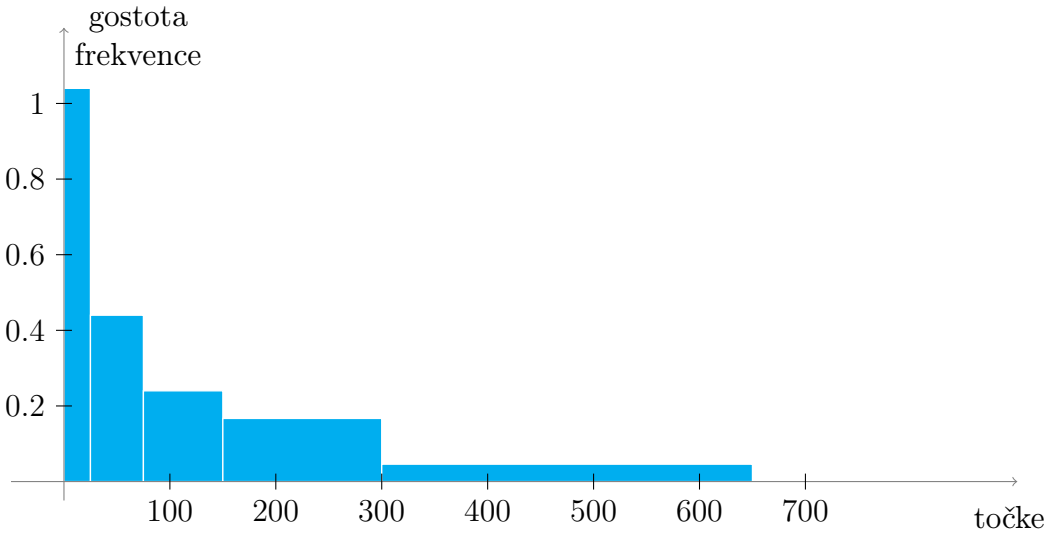
Primer: spet vzemimo razporeditev točk v svetovnem pokalu v alpskem smučanju za ženske, sezona 2012/13. Tekmovalo je 116 smučark. Po pravilu tretjega korena mora biti v posameznem razredu približno $\sqrt[3]{116^2} \doteq 23{,}8$ smučark. Izločimo pa skrajne vrednosti, za katere smo že izračunali, da so tiste nad 629. Temu sledi naslednja frekvenčna tabela:

Točke	Frekvenca	Gostota	Rel. gostota	Rel. gostota na 1000 točk
0 – manj kot 25	26	1,040	0,00897	8,97
25 – manj kot 75	22	0,440	0,00379	3,79
75 – manj kot 150	18	0,240	0,00207	2,07
150 – manj kot 300	25	0,167	0,00144	1,44
300 – manj kot 650	16	0,046	0,00039	0,39
650 –	9	skrajne vrednosti		

iz katere dobimo naslednji histogram:



Jasnejši prikaz glavnine smučark:



Za širino razredov po Freedman–Diaconisovem pravilu pa se spomnimo na vrstilne statistike:

$$q^{-}_{1/4} = x_{(29)} = 29\,, \quad q^{+}_{3/4} = x_{(88)} = 269\,, \quad \mathrm{IQR} = 240\,,$$

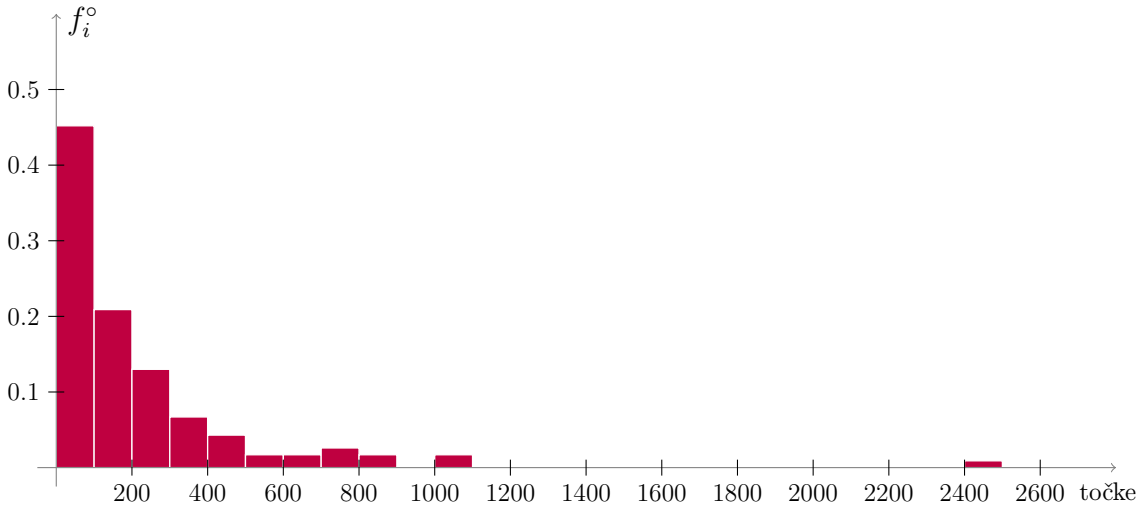
od koder dobimo, da mora biti širina razreda približno:

$$\frac{2 \cdot 240}{\sqrt[3]{116}} \doteq 98{\cdot}4 \approx 100\,.$$

Dobimo naslednjo frekvenčno tabelo:

Točke	Frekvenca	Rel. frekvenca
0 – manj kot 100	52	0·452
100 – manj kot 200	24	0·209
200 – manj kot 300	15	0·130
300 – manj kot 400	8	0·067
400 – manj kot 500	5	0·043
500 – manj kot 600	2	0·017
600 – manj kot 700	2	0·017
700 – manj kot 800	3	0·026
800 – manj kot 900	2	0·017
900 – manj kot 1000	0	0·000
1000 – manj kot 1100	2	0·017
⋮	⋮	⋮
2400 – manj kot 2500	1	0·009

Histogram:

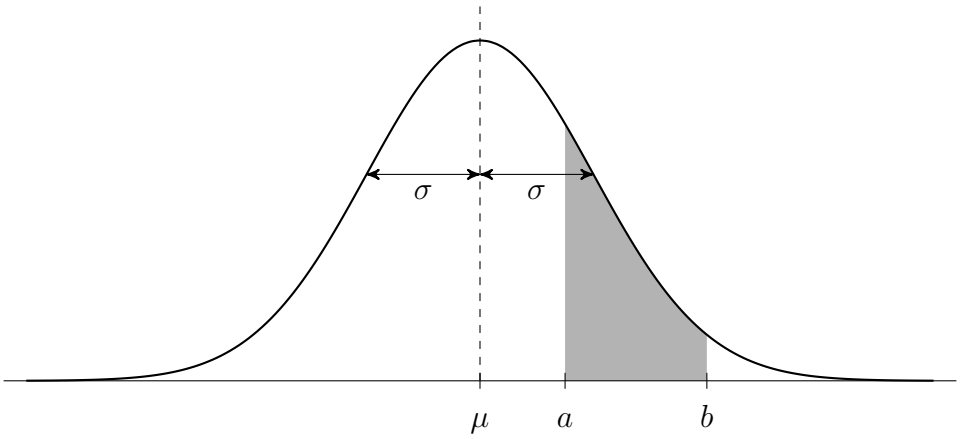


2.4.7 Normalna (Gaussova) porazdelitev

Normalna ali Gaussova porazdelitev je v statistiki zgolj idealizacija – v resnici je to *verjetnostna porazdelitev* oz. *verjetnostni zakon*. Statistična spremenljivka X je porazdeljena približno normalno s sredino oz. povprečjem μ in standardnim odklonom σ, če je delež enot, za katere X leži med a in b, kjer je $a < b$, približno:

$$\frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}x = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) .$$

Drugače prikazano, histogram porazdelitve sledi *Gaussovi krivulji*:



in delež enot, za katere X leži med a in b, je približno ploščina osenčenega območja, deljena s ploščino pod celotno krivuljo.

Normalna porazdelitev bi pomenila, da:

- znotraj intervala od $\mu - \sigma$ do $\mu + \sigma$ leži približno $68{\cdot}3\%$ enot;

- znotraj intervala od $\mu - 2\sigma$ do $\mu + 2\sigma$ leži približno $95{\cdot}5\%$ enot;
- znotraj intervala od $\mu - 3\sigma$ do $\mu + 3\sigma$ leži približno $99{\cdot}7\%$ enot.
- je skrajnih vrednosti približno $0{\cdot}7\%$.

Na nobeni statistični množici (iz končno mnogo, četudi veliko enot) to ne more veljati za *vse* a in b. Res pa je, da več kot je enot, natančneje je to lahko doseženo. Res pa je tudi, da veliko enot še malo ni jamstvo za normalno porazdelitev. Tako je npr. porazdelitev točk v svetovnem pokalu iz alpskega smučanja iz prejšnjega razdelka daleč stran od normalne.

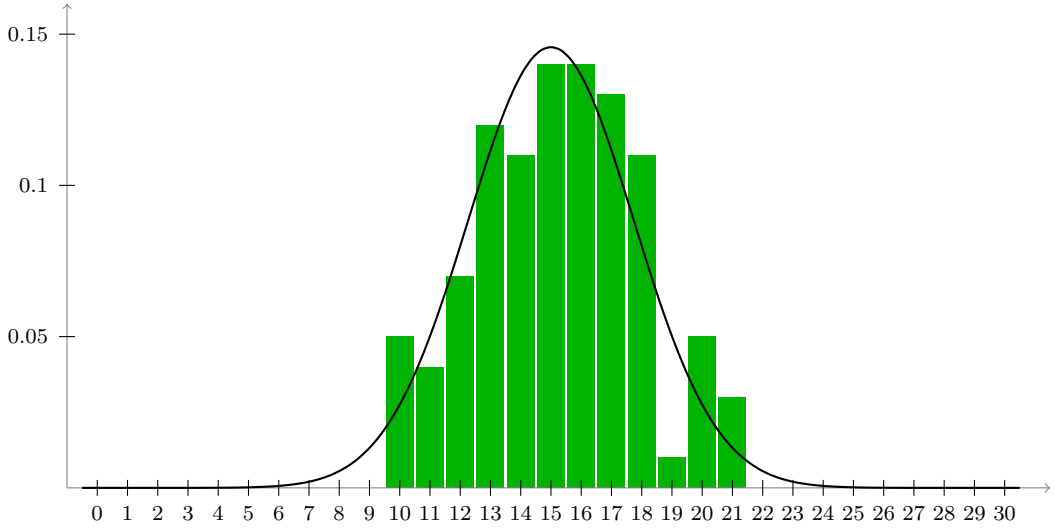
Normalna porazdelitev bi pomenila tudi neskončen variacijski razmik, kar je še en dokaz za to, da nobena končna statistična množica ne more imeti točno normalne porazdelitve. Na končni statistični množici je variacijski razmik vedno končen; a če želimo vedeti, koliko približno bo enak, predpostavka o približni normalnosti ni dovolj, potrebujemo še velikost množice. Večja kot je množica, večji variacijski razmik lahko pričakujemo. Ob približni normalnosti je pri statistični množici iz 100 enot variacijski razmik enak približno 5σ, pri množici iz 10000 enot pa približno 8σ. Precej natančneje pa je določen interkvartilni razmik: ta je ne glede na velikost statistične množice enak približno $1{\cdot}35\,\sigma$.

Porazdelitev intervalske statistične spremenljivke na določeni statistični množici je približno normalna, če sta izpolnjena naslednja dva pogoja:

- Statistična množica je velika, vrednosti spremenljivke na posameznih enotah pa so slučajne, neodvisne in sledijo istemu verjetnostnemu zakonu.
- Mehanizem, ki narekuje verjetnostni zakon, deluje tako, da je vrednost spremenljivke rezultat velikega števila med seboj neodvisnih slučajnih vplivov, ki se med seboj seštevajo, nimajo prevelikih ekscesov in med katerimi nobeden posebej ne izstopa.

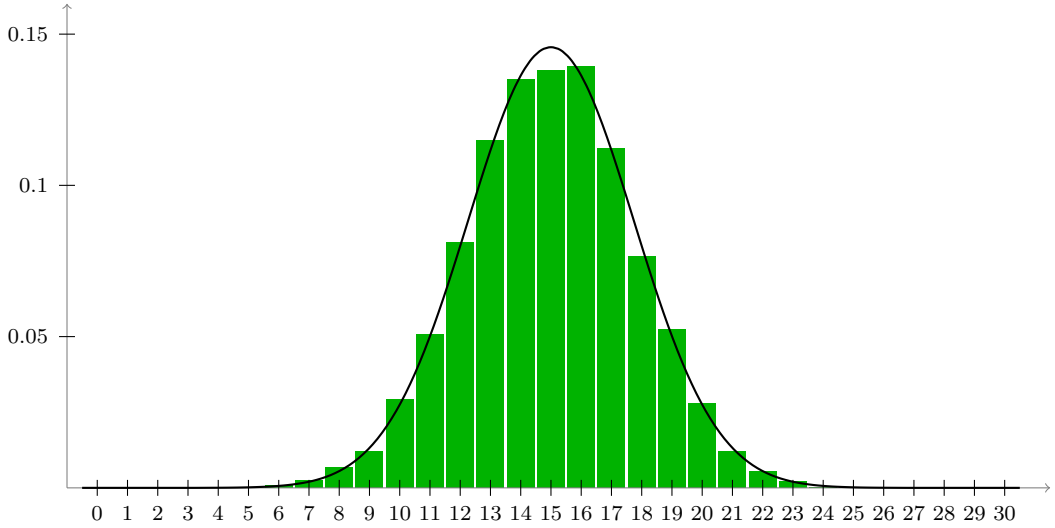
Prvi pogoj pride iz *zakonov velikih števil*, natančneje Glivenko–Cantellijevega izreka, omenjenega že pri urejenostnih spremenljivkah. Drugi pogoj pa pride iz *centralnega limitnega izreka*, ki pravi, da je slučajna spremenljivka, ki je rezultat veliko neodvisnih vplivov, ki se med seboj seštevajo, nimajo prevelikih ekscesov in med katerimi nobeden posebej ne izstopa, porazdeljena približno normalno.

Primer: simulacija 100 metov 30 poštenih kovancev. Vsak met predstavlja enoto, statistična spremenljivka pa je skupno število cifer na vseh kovancih v posameznem metu. Histogram skupaj s pripadajočim histogramom verjetnostne porazdelitve (ki bi jo dobili pri veliko metih) in Gaussovo krivuljo, ki predstavlja idealizacijo:



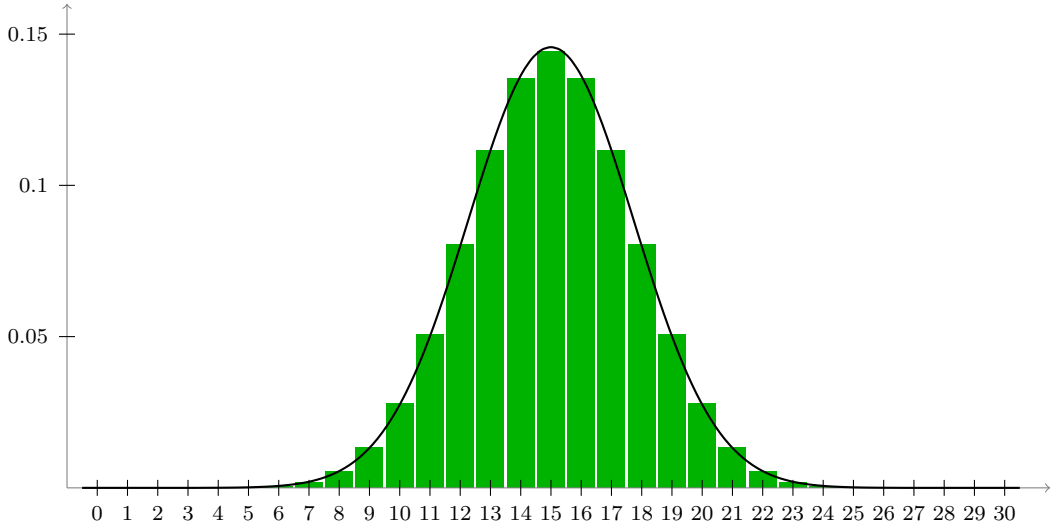
Razkorak med histogramom in Gaussovo krivuljo lahko nastopi tako zaradi napake v centralnem limitnem izreku kot tudi zaradi napake v centralnem limitnem izreku (oba govorita o *približni* enakosti porazdelitev, torej dopuščata določeno napako). Razkorak v zgornjem primeru nastopi predvsem zaradi približnosti v Glivenko–Cantellijevem izreku in manj zaradi približnosti v centralnem limitnem izreku. Napaka v Glivenko–Cantellijevem izreku se zmanjša, če povečamo število metov.

Primer: simulacija 10.000 metov 30 poštenih kovancev (ostalo isto kot pri prejšnjem primeru):



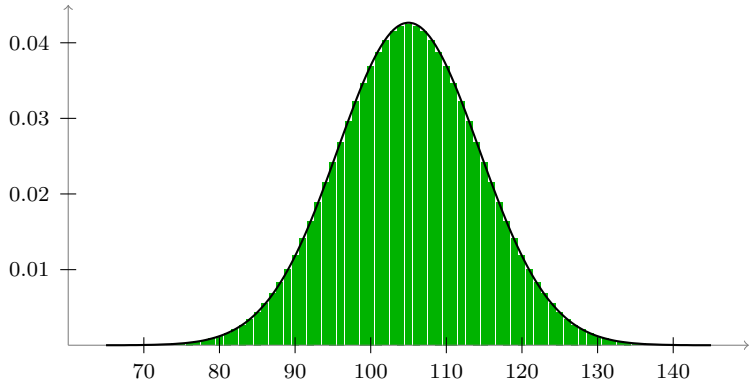
Vidimo, da je razkorak drastično manjši. Razkorak v naslednjem primeru pa je le ‘prispevek’ napake v centralnem limitnem izreku.

Primer: verjetnostna porazdelitev metov 30 poštenih kovancev (ki bi jo dobili, če bi poskus ponovili velikokrat):



Pri prej prikazanih metih kovanca statistična spremenljivka *prešteva* cifre. Centralni limitni izrek pa dopušča tudi *seštevanje*, ki je posplošitev preštevanja.

Primer: verjetnostna porazdelitev pri metih 30 poštenih kock, statistična spremenljivka je skupno število pik na vseh 30 kockah pri posameznem metu:



2.4.8 Točkasto ocenjevanje

Spet privzemimo, da se opažene vrednosti nanašajo na vzorec iz populacije ali pa na realizacije določenega slučajnega poskusa. Če gre za vzorec, bomo ocenjevali aritmetično sredino na populaciji, ki jo bomo označili z μ, in standardni odklon na populaciji, ki ga bomo označili s σ. Če pa gre za realizacije poskusa, pa bosta μ in σ pričakovana vrednost in standardni odklon verjetnostne porazdelitve, ki izhaja iz poskusa.

Označimo opažene vrednosti z $x_1, x_2, \ldots, x_n$. Cenilka za aritmetično sredino na populaciji oz. pričakovano vrednost verjetnostne porazdelitve je aritmetična sredina na vzorcu:

$$\hat{\mu} = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}\,.$$

Pri ocenjevanju standardnega odklona pa naredimo manjši popravek: za oceno populacijskega standardnega odklona vzamemo:

$$\hat{\sigma} = s_+ = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}\,.$$

Razlog za ta popravek je, da je potem s_+^2 *nepristranska* cenilka za σ^2, kar pomeni, da bi se, če bi jemali vedno več neodvisnih vzorcev iste velikosti in vsakič ocenili standardni odklon, povprečje ocen za kvadrat standardnega odklona bližalo dejanskemu kvadratu standardnega odklona populacije σ^2, medtem ko se, če bi vzeli nepopravljen standardni odklon, to ne bi zgodilo.

Če je vzorčenje ali izvajanje poskusov asimptotično reprezentativno, se, ko večamo velikost vzorca oz. število izvedb, ocenjeni vrednosti $\hat{\mu}$ in $\hat{\sigma}$ bližata pravima vrednostma μ in σ.

Primer: oglejmo si vzorec, na katerem ima statistična spremenljivka vrednosti:

$$101,\ 91,\ 93,\ 103,\ 91,\ 101,\ 103,\ 95,\ 95\,.$$

Dobimo:

$$\hat{\mu} = \frac{101 + 91 + 93 + 103 + 91 + 101 + 103 + 95 + 95}{9} = 97$$

in:

$$\hat{\sigma} = \left[\frac{1}{8}\Big((101-97)^2 + (91-97)^2 + (93-97)^2 + (103-97)^2 + (91-97)^2 + \right.$$
$$\left. + (101-97)^2 + (103-97)^2 + (95-97)^2 + (95-97)^2\Big)\right]^{1/2} =$$
$$= 5\,.$$

2.4.9 Intervalsko ocenjevanje in testiranje

Tu bomo obravnavali intervalsko ocenjevanje in testiranje populacijskega povprečja μ in standardnega odklona σ, a ob naslednjih dodatnih predpostavkah:

- Gre za enostavni slučajni vzorec iz velike populacije oz. za verjetnostno neodvisne izvedbe poskusa, pri čemer le-ta vsakič sledi istim verjetnostnim zakonitostim.
- Populacijska oz. verjetnostna porazdelitev izbrane statistične oz. slučajne spremenljivke je približno normalna (Gaussova).

Predpostavka o normalni porazdelitvi je močna, a za primer, ko ocenjujemo ali testiramo povprečje, so metode do določene mere robustne: če je vzorec dovolj velik, še vedno delujejo že, če je porazdelitev spremenljivke dovolj lepa, a ne nujno normalna – predvsem morata obstajati matematično upanje in varianca. To sledi iz centralnega limitnega izreka. Drugače pa je pri standardnem odklonu: predpostavka o normalni porazdelitvi je tu ključna. Obstajajo pa bolj zapletene konstrukcije, ki za velike n približno delujejo tudi pri porazdelitvah, ki niso normalne, so pa dovolj "lepe".

Povprečje pri znanem standardnem odklonu

Privzemimo, da nas zanima μ, pri čemer σ poznamo. V tem primeru poznamo tudi standardno napako:

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}\,.$$

Potrebovali bomo še kvantil normalne porazdelitve $c = z_{(1+\beta)/2}$. Spomnimo se:

$$z_{0.95} \doteq 1{\cdot}645\,, \qquad z_{0.975} \doteq 1{\cdot}960\,, \qquad z_{0.99} \doteq 2{\cdot}326\,, \qquad z_{0.995} \doteq 2{\cdot}576\,.$$

Spodnja in zgornja meja intervala zaupanja za μ sta:

$$\mu_{\min} = \bar{x} - c \cdot \mathrm{SE}\,, \qquad \mu_{\max} = \bar{x} + c \cdot \mathrm{SE}\,.$$

Primer: če bi pri vzorcu iz prejšnjega primera vedeli, da je $\sigma = 5$, bi pri $\beta = 95\%$ izračunali:

$$\mathrm{SE} = \frac{5}{\sqrt{9}} \doteq 1{\cdot}667\,, \quad c \doteq 1{\cdot}96\,,$$
$$\mu_{\min} \doteq 97 - 1{\cdot}96 \cdot 1{\cdot}667 \doteq 93{\cdot}73\,, \quad \mu_{\max} \doteq 97 + 1{\cdot}96 \cdot 1{\cdot}667 \doteq 100{\cdot}27\,.$$

Opomba. Če povečamo velikost vzorca, se standardna napaka zmanjša. Z drugimi besedami, več kot imamo na voljo podatkov, natančnejše so naše ocene.

Zdaj pa si oglejmo še testiranje ničelne hipoteze, da je $\mu = \mu^*$. Tako kot pri testiranju deleža bomo obravnavali tri alternativne hipoteze: $H_1^{\pm}$, da je $\mu \neq \mu^*$, H_1^{+}, da je $\mu > \mu^*$, in H_1^{-}, da je $\mu > \mu^*$. Pri alternativni hipotezi $H_1^{\pm}$ gre torej za dvostranski, pri H_1^{+} in H_1^{-} pa za enostranski test. Testiramo z Z-testom na testni statistiki, ki je razmerje med opaženo razliko in standardno napako:

$$Z = \frac{\bar{x} - \mu^*}{\mathrm{SE}}\,,$$

kar pomeni, da ničelno hipotezo zavrnemo:

- proti $H_1^{\pm}$, če je $|Z| > z_{1-\alpha/2}$;
- proti H_1^{+}, če je $Z > z_{1-\alpha}$;
- proti H_1^{-}, če je $Z < -z_{1-\alpha}$.

Primer. Meritve neke količine, porazdeljene normalno $\mathrm{N}(\mu, 5)$, dajo naslednje vrednosti:

$$101,\ 91,\ 93,\ 103,\ 91,\ 101,\ 103,\ 95,\ 95$$

Ta vzorec ima $\bar{x} = 97$ in $\mathrm{SE} \doteq 1{\cdot}667$.

Testirajmo ničelno hipotezo, da je $\mu = 100$. V tem primeru testna statistika pride $Z = -1{\cdot}8$. Sicer pa moramo test še doreči. Ogledali si bomo več različic.

- Pri stopnji značilnosti $\alpha = 0{\cdot}05$ testirajmo ničelno hipotezo proti alternativni hipotezi, da je $\mu \neq 100$. To pomeni, da moramo absolutno vrednost testne statistike, $|Z| = 1{\cdot}8$, primerjati z $z_{0.975} \doteq 1{\cdot}960$. Vidimo, da ničelne hipoteze ne moremo zavrniti. Z drugimi besedami, odstopanja niso statistično značilna.

- Še vedno pri stopnji značilnosti $\alpha = 0{\cdot}05$ testirajmo ničelno hipotezo proti alternativni hipotezi, da je $\mu < 100$. Testno statistiko $Z = -1{\cdot}8$ moramo zdaj primerjati z $-z_{0.95} \doteq -1{\cdot}645$. To pomeni, da ničelno hipotezo zdaj zavrnemo. Z drugimi besedami, odstopanja v levo so statistično značilna. Če smo občutljivi le na eno stran, smo lahko tam bolj restriktivni.

- Še vedno pri stopnji značilnosti $\alpha = 0{\cdot}05$ testirajmo ničelno hipotezo proti alternativni hipotezi, da je $\mu > 100$. Testno statistiko $Z = -1{\cdot}8$ moramo zdaj primerjati z $z_{0.95} \doteq 1{\cdot}645$. Ničelne hipoteze seveda ne zavrnemo. Odstopanja v desno ne morejo biti statistično značilna, če povprečje od ničelne hipoteze odstopa v levo.

- Tokrat pri stopnji značilnosti $\alpha = 0{\cdot}01$ testirajmo ničelno hipotezo proti alternativni hipotezi, da je $\mu < 100$. Testno statistiko $Z = -1{\cdot}8$ moramo zdaj primerjati z $-z_{0.99} \doteq -2{\cdot}326$ in vidimo, da ničelne hipoteze zdaj ne moremo zavrniti. Odstopanja v levo so torej sicer statistično značilna, niso pa zelo značilna.

Povprečje pri neznanem standardnem odklonu

Če standardni odklon ni znan, se da metode iz prejšnjega podrazdelka prilagoditi tako, da standardni odklon σ nadomestimo z njegovo oceno. Tako je standardna napaka zdaj enaka:

$$\mathrm{SE} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s_+}{\sqrt{n}}\,.$$

Kvantile standardne normalne porazdelitve pa moramo nadomestiti s kvantili *Studentove*[13] porazdelitve. Studentova porazdelitev je v resnici cela družina porazdelitev, ki se razlikujejo glede na število *prostostnih stopenj df* (angl. *degrees of freedom*). Intuitivno lahko število prostostnih stopenj pri Studentovi porazdelitvi gledamo kot količino informacije, ki jo imamo na voljo za oceno standardnega odklona. V našem primeru je $df = n-1$, to pa zato, ker smo eno enoto informacije že porabili za ocenjevanje povprečja: če bi povprečje poznali, bi bilo $df = n$. Kvantil Studentove porazdelitve z df prostostnimi stopnjami za verjetnost p označimo s $t_p(df)$. Kvantile lahko odčitamo iz tabele 2 ali pa izračunamo s pomočjo ustrezne programske opreme.

Spodnja in zgornja meja intervala zaupanja za povprečje imata enako obliko kot prej:

$$\mu_{\min} = \bar{x} - c \cdot \mathrm{SE}\,, \quad \mu_{\max} = \bar{x} + c \cdot \mathrm{SE}\,,$$

le da je zdaj $c = t_{(1+\beta)/2}(n-1)$.

[13]Ime ji je dal angleški statistik William Sealy Gosset (1876–1937), ki je pisal pod psevdonimom Student.

Primer: če pri vzorcu iz prejšnjega primera populacijskega standardnega odklona ne bi poznali, bi pri $\beta = 95\%$ izračunali:

$$\mathrm{SE} = \frac{5}{\sqrt{9}} \doteq 1{\cdot}667\,, \quad c = t_{0.975}(8) \doteq 2{\cdot}306\,,$$
$$\mu_{\min} \doteq 97 - 2{\cdot}31 \cdot 1{\cdot}667 \doteq 93{\cdot}14\,, \quad \mu_{\max} \doteq 97 + 2{\cdot}31 \cdot 1{\cdot}667 \doteq 100{\cdot}86\,.$$

Interval zaupanja je zdaj malo širši: ker zdaj manj vemo, je tudi naša ocena manj natančna.

Podobno modificiramo tudi testiranje. Ničelno hipotezo H_0, da je $\mu = \mu^*$, testiramo s pomočjo testne statistike, ki je spet razmerje med opaženo razliko in standardno napako:

$$T := \frac{\bar{x} - \mu^*}{\mathrm{SE}}\,,$$

kjer je spet $\mathrm{SE} = \dfrac{s_+}{\sqrt{n}}$. Ničelno hipotezo zavrnemo:

- proti $H_1^{\pm}$: $\mu \neq \mu^*$, če je $|T| > t_{1-\alpha/2}(n-1)$;
- proti H_1^{+}: $\mu > \mu^*$, če je $T > t_{1-\alpha}(n-1)$;
- proti H_1^{-}: $\mu < \mu^*$, če je $T < -t_{1-\alpha}(n-1)$.

V ozadju tega je seveda, da ima testna statistika T pri veljavnosti ničelne hipoteze Studentovo porazdelitev z $n-1$ prostostnimi stopnjami. Testu, kjer to velja, pravimo T-test.

Primer. Isti vzorec kot pri prejšnjem primeru, le da ne vemo, da je $\sigma = 5$. Pri $\alpha = 0{\cdot}05$ testiramo ničelno hipotezo, da je $\mu = 100$, proti alternativni hipotezi, da je $\mu < 100$.

Spomnimo se, da je $\bar{x} = 97$. Izračunajmo še $s_+ = 5$, $\mathrm{SE} \doteq 1{\cdot}667$ in od tod $T = -1{\cdot}8$, kar primerjamo z $-t_{0.95}(8) \doteq -2{\cdot}306$. Tokrat ničelne hipoteze ne moremo zavrniti: odstopanja v levo niso statistično značilna. Nauk: če določene reči (recimo standardnega odklona) ne poznamo v popolnosti, moramo biti bolj previdni – tako kot smo bili tudi pri intervalskem ocenjevanju.

Standardni odklon pri neznanem povprečju

Pri standardnem odklonu bomo potrebovali porazdelitev hi kvadrat. Spomnimo se, da $\chi^2_p(df)$ označuje kvantil porazdelitve hi kvadrat z df prostostnimi stopnjami za verjetnost p. V našem primeru bo spet $df = n - 1$.

Pri konstrukciji intervalov zaupanja bomo potrebovali kvantila:

$$c_1 = \chi^2_{(1-\beta)/2}(n-1)\,, \qquad c_2 = \chi^2_{(1+\beta)/2}(n-1)\,.$$

Spodnja in zgornja meja bosta enaki:

$$\sigma_{\min} = s_+ \sqrt{\frac{n-1}{c_2}}\,, \quad \sigma_{\max} = s_+ \sqrt{\frac{n-1}{c_1}}\,.$$

Pri vzorcu iz prejšnjega primera bi pri $\beta = 95\%$ izračunali:

$$c_1 = \chi^2_{0.025}(8) \doteq 2{\cdot}180\,, \quad c_2 = \chi^2_{0.975}(8) \doteq 17{\cdot}53\,,$$
$$\sigma_{\min} = 5 \cdot \sqrt{\frac{8}{17{\cdot}53}} \doteq 3{\cdot}38\,, \quad \sigma_{\max} = 5 \cdot \sqrt{\frac{8}{2{\cdot}180}} \doteq 9{\cdot}58\,.$$

Oglejmo si še testiranje ničelne hipoteze H_0, da je $\sigma = \sigma^*$. Le-to testiramo s pomočjo testne statistike:

$$\chi^2 := (n-1)\,\frac{s_+^2}{(\sigma^*)^2}\,.$$

Spet bomo gledali tri alternativne hipoteze in temu ustrezno postavili kritična območja. Ničelno hipotezo zavrnemo:

- proti $H_1^{\pm}\colon\ \sigma \neq \sigma^*$, če je $\chi^2 < \chi^2_{\alpha/2}(n-1)$ ali $\chi^2 > \chi^2_{1-\alpha/2}(n-1)$;
- proti $H_1^{+}\colon\ \sigma > \sigma^*$, če je $\chi^2 > \chi^2_{1-\alpha}(n-1)$;
- proti $H_1^{-}\colon\ \sigma < \sigma^*$, če je $\chi^2 < \chi^2_{1-\alpha}(n-1)$.

Primer. Meritve neke količine, porazdeljene normalno $\mathrm{N}(\mu, \sigma)$, dajo naslednje vrednosti:

$$99,\ 90,\ 108,\ 111,\ 97,\ 93,\ 90,\ 106,\ 104,\ 102$$

Pri $\alpha = 0{\cdot}05$ testirajmo ničelno hipotezo, da je $\sigma = 5$, proti alternativni hipotezi, da je $\sigma \neq 5$. Izračunajmo $s_+ \doteq 7{\cdot}45$ in $\chi^2 = 20$, kar moramo primerjati s kritičnima vrednostma $\chi^2_{0.025}(9) \doteq 2{\cdot}700$ in $\chi^2_{0.975}(9) \doteq 19{\cdot}02$. Torej hipotezo zavrnemo, odstopanja so statistično značilna.

Primerjava povprečij dveh spremenljivk na istih enotah

Denimo, da imamo za vsako enoto dani dve intervalski spremenljivki, X in Y. Označimo z μ_X aritmetično sredino prve, z μ_Y pa aritmetično sredino druge spremenljivke na celotni populaciji. Testiramo ničelno hipotezo $H_0\colon\ \mu_X = \mu_Y$, alternativno hipotezo pa lahko postavimo na tri načine: dvostransko $H_1^{\pm}\colon\ \mu_X \neq \mu_Y$, enostransko v korist X, ki je $H_1^X\colon\ \mu_X > \mu_Y$ in enostransko v korist Y, ki je $H_1^Y\colon\ \mu_X < \mu_Y$.

Ta test se prevede na običajni T-test za eno spremenljivko, ki je kar razlika $X - Y$. Če so torej $x_1, \ldots, x_n$ vrednosti prve, $y_1, \ldots, y_n$ pa vrednosti druge spremenljivke na vzorcu, izračunamo:

$$s_+ = \sqrt{\frac{\big((x_1 - y_1) - (\bar{x} - \bar{y})\big)^2 + \cdots + \big((x_n - y_n) - (\bar{x} - \bar{y})\big)^2}{n-1}}\,,$$
$$\mathrm{SE} = \frac{s_+}{\sqrt{n}}\,, \qquad T = \frac{\bar{x} - \bar{y}}{\mathrm{SE}}$$

in ničelno hipotezo zavrnemo:

- proti $H_1^{\pm}$, če je $|T| > t_{1-\alpha/2}(n-1)$;
- proti H_1^X, če je $T > t_{1-\alpha}(n-1)$;
- proti H_1^Y, če je $T < -t_{1-\alpha}(n-1)$.

Primer. Pri predmetu Analiza III na Interdisciplinarnem študiju računalništva in matematike na Univerzi v Ljubljani se pišeta dva kolokvija. Rezultati študentov, ki so v študijskem letu 2008/09 pisali oba kolokvija, so zbrani v naslednji tabeli:

	1. kolokvij (X)	2. kolokvij (Y)	$X-Y$
	89	96	-7
	59	65	-6
	51	79	-28
	98	99	-1
	46	68	-22
	79	60	19
	68	65	3
	63	85	-22
	73	65	8
	52	73	-21
	82	97	-15
	82	100	-18
	50	95	-45
	46	80	-34
Povprečje	67·0	80·5	$-13{\cdot}5$

Od tod izračunamo:

$$s_+ \doteq 17{\cdot}30\,, \qquad \mathrm{SE} \doteq 4{\cdot}624\,, \qquad T \doteq -2{\cdot}92\,.$$

Izvedemo dvostranski test. Podatkov je 14, torej je prostostnih stopenj 13. Ker je $t_{0.975}(13) \doteq 2{\cdot}160$, ničelno hipotezo pri $\alpha = 0{\cdot}05$ zavrnemo – razlika med kolokvijema je statistično značilna. Pri $\alpha = 0{\cdot}01$ pa vrednost testne statistike primerjamo s $t_{0.995}(13) \doteq 3{\cdot}012$ in dobimo, da razlika ni statistično zelo značilna.

Zgoraj opisani test velja ob predpostavki, da je porazdelitev normalna ali dihotomna ali pa da je vzorec dovolj velik. Če temu ni tako, lahko namesto T-testa izvedemo test z znaki. V našem primeru pride:

$$S_+ = 3\,, \quad S_- = 11\,, \quad Z = -\frac{8}{\sqrt{14}} \doteq -2{\cdot}14\,.$$

Pri $\alpha = 0{\cdot}05$ to primerjamo z $z_{0.975} \doteq 1{\cdot}960$, pri $\alpha = 0{\cdot}01$ pa z $z_{0.995} \doteq 2{\cdot}576$. Spet dobimo, da je razlika statistično značilna, ni pa zelo značilna.

Testiranje normalne porazdelitve

Orodja iz inferenčne statistike (intervalskega ocenjevanja in testiranja hipotez) za intervalske spremenljivke, opisana v tem razdelku, so zasnovana ob predpostavki, da ima statistična spremenljivka normalno (Gaussovo) porazdelitev. Ta predpostavka ni vedno izpolnjena in zastavi se vprašanje, ali jo lahko preverimo. Kot nasploh v inferenčni statistiki ne more obstajati algoritem, ki bi na podlagi vzorca dal odgovor da ali ne glede porazdelitve na populaciji, še zlasti pa ne z gotovostjo. Lahko pa hipotezo o normalnosti testiramo.

Testov normalne porazdelitve je veliko. V večini primerov so najustreznejši t. i. *prilagoditveni testi* (angl. *goodness of fit*), ki merijo, koliko empirična (vzročna) porazdelitev odstopa od normalne (ali nasprotno, kako tesno se ji prilega). Tu bomo spoznali *Anderson*[14]*–Darlingov*[15] *test*, natančneje, D'Agostinovo[16] modifikacijo tega testa [17]. Podatke najprej uredimo po velikosti – naredimo ranžirno vrsto:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}\,.$$

Spomnimo se: $X_{(i)}$ je i-ta vrstilna statistika. Podatke standardiziramo – za ta namen izračunamo aritmetično sredino in popravljeni vzorčni standardni odklon:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}\,, \qquad s_+ = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\,.$$

Standardizirane vrednosti so $Z_i := (X_i - \bar{X})/s_+$. Potrebovali bomo standardizirane vrstilne statistike $Z_{(i)} := (X_{(i)} - \bar{X})/s_+$. Iz njih izračunamo Anderson–Darlingovo testno statistiko:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\ln\left(\frac{1}{2} + \Phi(Z_{(i)})\right) + (2n-2i+1)\ln\left(\frac{1}{2} - \Phi(Z_{(i)})\right)\right]\,.$$

Tu je Φ Gaussov verjetnostni integral:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-z^2/2}\,\mathrm{d}z\,.$$

Vrednosti te funkcije lahko odčitamo iz tabele 1. Kot smo že omenili, bomo uporabili D'Agostinovo modifikacijo testa, ki temelji na naslenjem popravku Anderson–Darlingove statistike:

$$A^{*2} = A^2\left(1 + \frac{0{\cdot}75}{n} + \frac{2{\cdot}25}{n^2}\right)$$

Ničelno hipotezo o normalnosti zavrnemo:

- pri stopnji značilnosti $\alpha = 0{\cdot}05$, če je $A^{*2} > 0{\cdot}752$;

[14]Theodore Wilbur Anderson (1918–2016), ameriški matematik in statistik
[15]Donald Allan Darling (1915–2014), ameriški statistik
[16]Ralph B D'Agostino, ameriški statistik

- pri stopnji značilnosti $\alpha = 0{\cdot}01$, če je $A^{*2} > 1{\cdot}035$.

Primer. Oglejmo si vzorec, na katerem ima statistična spremenljivka vrednosti:

$$101,\ 91,\ 93,\ 103,\ 91,\ 101,\ 103,\ 95,\ 95\,.$$

Vrednosti uredimo po velikosti:

$$91,\ 91,\ 93,\ 95,\ 95,\ 101,\ 101,\ 103,\ 103\,.$$

Iz $\bar{X} = 97$ in $s_+ = 5$ dobimo standardizirane vrednosti:

$$-1{\cdot}2,\ -1{\cdot}2,\ -0{\cdot}8,\ -0{\cdot}4,\ -0{\cdot}4,\ 0{\cdot}8,\ 0{\cdot}8,\ 1{\cdot}2,\ 1{\cdot}2\,.$$

Anderson–Darlingova statistika je enaka:

$$\begin{aligned} A^2 \doteq -9 - \frac{1}{9}\Big[& 1 \cdot \ln 0{\cdot}1151 + 17 \ln 0{\cdot}8849 + 3 \ln 0{\cdot}1151 + 15 \ln 0{\cdot}8849 + \\ & + 5 \ln 0{\cdot}2119 + 13 \ln 0{\cdot}7881 + 7 \ln 0{\cdot}3446 + 11 \ln 0{\cdot}6554 + \\ & + 9 \ln 0{\cdot}3446 + 9 \ln 0{\cdot}6554 + 11 \ln 0{\cdot}7881 + 7 \ln 0{\cdot}2119 + \\ & + 13 \ln 0{\cdot}7881 + 5 \ln 0{\cdot}2119 + 15 \ln 0{\cdot}8849 + 3 \ln 0{\cdot}1151 + \\ & + 17 \ln 0{\cdot}8849 + 1 \cdot \ln 0{\cdot}1151 \Big] \doteq \\ \doteq 0{\cdot}5342\,, & \end{aligned}$$

modificirana vrednost pa je enaka:

$$A^{*2} \doteq 0{\cdot}5342 \left(1 + \frac{0{\cdot}75}{9} + \frac{2{\cdot}25}{81}\right) \doteq 0{\cdot}5936\,.$$

Ker je $0{\cdot}5936 < 0{\cdot}752$, odstopanja od normalne porazdelitve niso statistično značilna.

Če so vrednosti podane v frekvenčni tabeli:

vrednosti	frekvence	kumulativne frekvence
a_1	f_1	F_1
a_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$
a_k	f_k	F_k

jih najprej spet standardiziramo – izračunamo:

$$n = \sum_{j=1}^{k} f_j = F_k\,, \quad \bar{X} = \frac{1}{n} \sum_{j=1}^{k} f_j a_j\,, \quad s_+ = \sqrt{\frac{1}{n-1} \sum_{j=1}^{k} f_j (a_j - \bar{X})^2}\,, \quad b_j = \frac{a_j - \bar{X}}{s_+}\,.$$

Anderson–Darlingova testna statistika je enaka:

$$A^2 = -n - \frac{1}{n} \sum_{j=1}^{k} f_j \left[(F_{j-1} + F_j) \ln \left(\frac{1}{2} + \Phi(b_j) \right) + (2n - F_{j-1} - F_j) \ln \left(\frac{1}{2} - \Phi(b_j) \right) \right]\,,$$

modificirana statistika A^{*2} pa se iz A^2 seveda dobi na isti način kot prej.

Primer. Za 860 žensk poizvemo, koliko otrok imajo. Dobimo naslednjo frekvenčno porazdelitev:

število otrok	0	1	2	3	4	5	6	7	8	9	10
število žensk	227	168	320	96	29	11	4	2	2	0	1

Povprečno število otrok na žensko je $1{\cdot}548$, popravljeni standardni odklon pa je $1{\cdot}304$. Standardiziramo in izračunamo kumulativne frekvence (razred, ki ni zastopan, pa lahko izpustimo):

b_j	$-1{\cdot}187$	$-0{\cdot}420$	$0{\cdot}347$	$1{\cdot}114$	$1{\cdot}880$	$2{\cdot}647$	$3{\cdot}414$	$4{\cdot}181$	$4{\cdot}947$	$6{\cdot}481$
f_j	227	168	320	96	29	11	4	2	2	1
F_j	227	395	715	811	840	851	855	857	859	860
$F_{j-1}+F_j$	227	622	1110	1526	1651	1691	1706	1712	1716	1719

Anderson–Darlingova statistika je enaka:

$$\begin{aligned} A^2 \doteq -860 - \frac{1}{860}\Big[& 227\Big(227\ln 0{\cdot}1177 + 1493\ln 0{\cdot}8823\Big) + \\ & + 168\Big(622\ln 0{\cdot}3373 + 1098\ln 0{\cdot}6627\Big) + \\ & + 320\Big(1110\ln 0{\cdot}6356 + 610\ln 0{\cdot}3654\Big) + \\ & + 96\Big(1526\ln 0{\cdot}8672 + 194\ln 0{\cdot}1328\Big) + \\ & + 29\Big(1651\ln 0{\cdot}96997 + 69\ln 0{\cdot}03003\Big) + \\ & + 11\Big(1691\ln 0{\cdot}995941 + 29\ln 0{\cdot}004059\Big) + \\ & + 4\Big(1706\ln\big(1-3{\cdot}202\cdot 10^{-4}\big) + 14\ln\big(3{\cdot}202\cdot 10^{-4}\big)\Big) + \\ & + 2\Big(1712\ln\big(1-1{\cdot}452\cdot 10^{-5}\big) + 8\ln\big(1{\cdot}452\cdot 10^{-5}\big)\Big) + \\ & + 2\Big(1716\ln\big(1-3{\cdot}760\cdot 10^{-7}\big) + 4\ln\big(3{\cdot}760\cdot 10^{-7}\big)\Big) + \\ & + 1\cdot\Big(1719\ln\big(1-4{\cdot}556\cdot 10^{-11}\big) + 1\cdot\ln\big(4{\cdot}556\cdot 10^{-11}\big)\Big)\Big] \\ \doteq\ & 35{\cdot}036\,. \end{aligned}$$

Modificirana vrednost pa je $35{\cdot}036\cdot\left(1+\dfrac{0{\cdot}75}{860}+\dfrac{225}{860^2}\right)\doteq 35{\cdot}066$. Ker je to večje od $1{\cdot}035$, je odstopanje od normalne porazdelitve tokrat statistično zelo značilno.

Pomembno: če vrednosti Gaussovega verjetnostnega integrala odčitavamo iz tabele 1 in pademo izven tabeliranega območja, ne smemo vedno vzeti kar $\Phi(z)\approx\pm 1/2$. To vedno

privede do $\ln 0$ ali $\ln 1$. Približek $\ln 1 = 0$ je dober, medtem ko vrednost $\ln 0$ ni definirana. Namesto približka $\Phi(z) \approx \pm 1/2$ pa lahko vzamemo približni obrazec:

$$\frac{1}{2} - \Phi(z) \sim \frac{1}{z\sqrt{2\pi}}\, e^{-z^2/2}\,,$$

ki velja, ko gre z proti plus neskončno.

3.

Povezanost dveh statističnih spremenljivk – bivariatna analiza

V tem poglavju se bomo ukvarjali z dvema statističnima spremenljivkama, definiranima na *isti statistični množici*. Naučili se bomo dve stvari:

- Za vsak par merskih lestvic bomo poiskali statistiko, ki bo vrednotila stopnjo povezanosti med spremenljivkama. Čeprav vse statistike niso neposredno primerljive, bomo povedali, katera vrednost določene statistike ustreza določeni vrednosti druge statistike. Uvedli pa bomo tudi opisno (kvalitativno) lestvico povezanosti: neznatna, nizka, zmerna, visoka in zelo visoka. To bo olajšalo primerjavo statistik.
- Če se podatki nanašajo na vzorec iz določene populacije, lahko izbrana statistika na podatkih služi kot ocena ustrezne karakteristike na populaciji. Podobno, če so podatki dobljeni kot neodvisne realizacije določenega slučajnega poskusa, ki vsakič sledi istim verjetnostnim zakonitostim, lahko izbrana statistika služi kot ocena ustrezne karakteristike skupne verjetnostne porazdelitve izbranih slučajnih spremenljivk. Če je vzorčenje oz. izvajanje poskusov asimptotično reprezentativno (glede na obe spremenljivki, gledani skupaj), se, ko večamo velikost vzorca oz. število izvedb poskusa, vrednost statistike na vzorcu bliža vrednosti karakteristike na populaciji oz. verjetnostni porazdelitvi.
- Za primer, ko je naša statistična množica enostavni slučajni vzorec iz velike populacije, pa bomo za vsak par merskih lestvic konstruirali tudi test hipoteze, da sta statistični spremenljivki na celi populaciji nepovezani (neodvisni). Podobno, če so podatki dobljeni iz verjetnostno neodvisnih izvedb določenega slučajnega poskusa, ki vsakič sledi istim verjetnostnim zakonitostim, lahko testiramo hipotezo, da sta spremenljivki verjetnostno neodvisni.

  Tako bo povezanost lahko statistično neznačilna, značilna ali zelo značilna. Statistična značilnost (p-vrednost) je drug pojem kot stopnja povezanosti na vzorcu: pri majhnih vzorcih se lahko zgodi, da je povezanost visoka, a statistično neznačilna. Pri velikih vzorcih pa se lahko zgodi celo, da je povezanost kvalitativno ovrednotena kot neznatna, a je statistično zelo značilna.

POZOR! Povezanost dveh statističnih spremenljivk še ne pomeni, da ena od njiju neposredno vpliva na drugo – povezanost ne implicira vzročnosti. Navadno povezanost nastane zaradi tega, ker na obe spremenljivki vpliva neka tretja spremenljivka (lahko zelo posredno), le-to pa je dostikrat težko določiti.

Primer: Če bi raziskovali povezavo med številom smučarjev in številom primerov gripe, bi bila ta verjetno visoka. To pa ne pomeni, da smučanje povzroča dovzetnost za gripo: oboje se znatno pogosteje pojavlja pozimi.

3.1 Povezanost dveh imenskih spremenljivk: asociiranost

3.1.1 Vrednotenje asociiranosti

Asociiranost ugotavljamo na podlagi *kontingenčne tabele*, kjer so podane frekvence za vse možne kombinacije vrednosti prve in druge spremenljivke:

	b_1	b_2	$\cdots$	b_l	
a_1	f_{11}	f_{12}	$\cdots$	f_{1l}	$f_{1\cdot}$
a_2	f_{21}	f_{22}	$\cdots$	f_{2l}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_k	f_{k1}	f_{k2}	$\cdots$	f_{kl}	$f_{k\cdot}$
	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot l}$	n

Pri tem so:

- $a_1, a_2, \ldots, a_k$ vrednosti prve spremenljivke, npr. X;
- $b_1, b_2, \ldots, b_l$ vrednosti druge spremenljivke, npr. Y;
- f_{ij} *navzkrižne frekvence* (angl. *joint frequencies, cross-frequencies*): navzkrižna frekvenca f_{ij} pove, na koliko enotah je $X = a_i$ in hkrati $Y = b_j$ ali s formulo:

$$f_{ij} = \sharp(X = a_i, Y = b_j)\,;$$

- $f_{i\cdot}$ in $f_{\cdot j}$ *robne frekvence* (angl. *marginal frequencies*): robne frekvence:

$$f_{i\cdot} = f_{i1} + f_{i2} + \cdots + f_{il} = \sharp(X = a_i)$$

tvorijo frekvenčno porazdelitev spremenljivke X, robne frekvence:

$$f_{\cdot j} = f_{1j} + f_{2j} + \cdots + f_{lj} = \sharp(Y = b_j)$$

pa tvorijo frekvenčno porazdelitev spremenljivke Y. Seveda velja:

$$n = f_{1\cdot} + f_{2\cdot} + \cdots + f_{k\cdot} = f_{\cdot 1} + f_{\cdot 2} + \cdots + f_{\cdot l}\,.$$

Definiramo lahko tudi *relativne navzkrižne frekvence* in relativne robne frekvence:

$$f^\circ_{ij} = \frac{f_{ij}}{n}\,, \quad f^\circ_{i\cdot} = \frac{f_{i\cdot}}{n} = \sum_{j=1}^{l} f_{ij}\,, \quad f^\circ_{\cdot j} = \frac{f_{\cdot j}}{n} = \sum_{i=k}^{l} f_{ij}\,.$$

Primer: barva oči in barva las neke skupine ljudi

Absolutne frekvence:

oči\lasje	rdeči	blond	rjavi, črni	Skupaj
modre	1	11	1	13
zelene	0	14	9	23
rjave	2	2	22	26
Skupaj	3	27	32	62

Relativne frekvence:

oči\lasje	rdeči	blond	rjavi, črni	Skupaj
modre	0.016	0.177	0.016	0.210
zelene	0.000	0.226	0.145	0.371
rjave	0.032	0.032	0.355	0.419
Skupaj	0.048	0.435	0.516	1.000

Obstaja veliko številskih mer (pokazateljev) povezanosti. Eden najboljših je *Cramérjev*[1] *koeficient asociiranosti*, osnova za njegov izračun pa so *pričakovane relativne navzkrižne frekvence*:

$$\tilde{f}^\circ_{ij} = f^\circ_{i\cdot}\, f^\circ_{\cdot j}\,,$$

Pričakovane relativne navzkrižne frekvence bi se pri danih relativnih robnih frekvencah pojavile, če bi bili spremenljivki neasociirani, t. j. neodvisni v smislu teorije verjetnosti, če bi statistično množico obravnavali kot verjetnostni prostor, na katerem bi bile vse enote enako verjetne. Če bi bilo npr. v populaciji 20% oseb z modrimi očmi in 50% oseb z rjavimi ali črnimi lasmi ter barva oči in barva las ne bi bili povezani, bi bilo oseb, ki imajo tako modre oči kot tudi rjave ali črne lase, 20% od 50%, kar znaša 10%. Z drugimi besedami, delež bi bil $0.2 \cdot 0.5 = 0.1$.

V našem prejšnjem primeru so pričakovane relativne navzkrižne frekvence enake:

oči\lasje	rdeči	blond	rjavi, črni	Skupaj
modre	0.010	0.091	0.108	0.210
zelene	0.018	0.162	0.191	0.371
rjave	0.020	0.183	0.216	0.419
Skupaj	0.048	0.435	0.516	1.000

[1] Carl Harald Cramér (1893–1985), švedski matematik, aktuar in statistik

Cramérjev koeficient asociiranosti je zasnovan na razkoraku med opaženimi in pričakovanimi relativnimi frekvencami in je definiran s formulo:

$$V := \sqrt{\frac{1}{\min\{k,l\}-1}\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\left(f_{ij}^{\circ}-\tilde{f}_{ij}^{\circ}\right)^2}{\tilde{f}_{ij}^{\circ}}}\,.$$

Pri našem primeru izračunamo:

$$\begin{gathered}\frac{(0{\cdot}016-0{\cdot}010)^2}{0{\cdot}010}+\frac{(0{\cdot}177-0{\cdot}091)^2}{0{\cdot}091}+\frac{(0{\cdot}016-0{\cdot}108)^2}{0{\cdot}108}+\\+\frac{(0{\cdot}000-0{\cdot}018)^2}{0{\cdot}018}+\frac{(0{\cdot}226-0{\cdot}162)^2}{0{\cdot}162}+\frac{(0{\cdot}145-0{\cdot}191)^2}{0{\cdot}191}+\\+\frac{(0{\cdot}032-0{\cdot}020)^2}{0{\cdot}020}+\frac{(0{\cdot}032-0{\cdot}183)^2}{0{\cdot}183}+\frac{(0{\cdot}355-0{\cdot}216)^2}{0{\cdot}216}\doteq 0{\cdot}43716\,,\\ V\doteq\sqrt{\frac{1}{3-1}\cdot 0{\cdot}43716}\doteq 0{\cdot}47\,.\end{gathered}$$

Lastnosti Cramérjevega koeficienta asociiranosti:

- Velja $0 \le V \le 1$.
- Koeficient je minimalen (enak 0) natanko tedaj, ko sta spremenljivki neasociirani. Še več, če so podatki dobljeni kot dovolj velik enostavni slučajni vzorec iz populacije, na kateri sta spremenljivki X in Y neasociirani, je V blizu 0: večji kot je vzorec, bolj natančno to velja. Malo kasneje se bomo naučili, kako testirati neasociiranost na populaciji.
- Koeficient V je maksimalen (enak 1) natanko tedaj, ko spremenljivki natančno določata druga drugo (sta *popolnoma asociirani*).

Kvalitativno opredeljevanje koeficienta asociiranosti je precej subjektivne narave. Tu se bomo držali naslednjih dogovorov:

- do 0·2: neznatna asociiranost;
- od 0·2 do 0·4: rahla asociiranost;
- od 0·4 do 0·7: zmerna asociiranost;
- od 0·7 do 0·9: močna asociiranost;
- od 0·9 do 1: zelo močna asociiranost.

Pri primeru z barvo oči in las sta torej spremenljivki *zmerno* asociirani.

Če sta obe spremenljivki *dihotomni* (lahko zavzameta le dve vrednosti) in je njuna porazdelitev podana s kontingenčno tabelo:

	b_1	b_2
a_1	A	B
a_2	C	D

,

je izražava Cramérjevega koeficienta asociiranosti preprostejša:

$$V = \frac{|AD - BC|}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}\,.$$

Več informacije pa nam da predznačena vrednost:

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}\,.$$

- Če je $\phi > 0$, to pomeni, da se enote, na katerih prva spremenljivka zavzame prvo vrednost, nagibajo k temu, da tudi druga spremenljivka zavzame prvo vrednost; nasprotno se enote, na katerih prva spremenljivka zavzame drugo vrednost, nagibajo k temu, da tudi druga spremenljivka zavzame drugo vrednost.
- Če je $\phi < 0$, to pomeni, da se enote, na katerih prva spremenljivka zavzame prvo vrednost, nagibajo k temu, da druga spremenljivka zavzame drugo vrednost; nasprotno se enote, na katerih prva spremenljivka zavzame drugo vrednost, nagibajo k temu, da tudi druga spremenljivka zavzame prvo vrednost.

Primer: rezultati ankete z dvema vprašanjema:

1. Ali verjamete v horoskop?
2. Ali verjamete v NLP-je?

so zbrani v naslednji tabeli:

Horoskop\NLP	vsaj malo	ne	Skupaj
vsaj malo	5	7	12
ne	6	9	15
Skupaj	11	16	27

Velja $\phi = \dfrac{5 \cdot 9 - 7 \cdot 6}{\sqrt{12 \cdot 15 \cdot 11 \cdot 16}} \doteq 0{\cdot}017$.
Gre torej za neznatno pozitivno povezanost.

3.1.2 Testiranje neasociiranosti

Ničelno hipotezo, da spremenljivki na celotni populaciji nista asociirani, testiramo s *kontingenčnim testom*. To je test hi kvadrat, in sicer z enostransko različico v desno in $(k-1)(l-1)$ prostostnimi stopnjami, kjer je kot prej k število možnih vrednosti prve, l pa število možnih vrednosti druge spremenljivke. Testna statistika hi kvadrat pa se izraža s formulo:

$$\chi^2 = n \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(f_{ij}^{\circ} - \tilde{f}_{ij}^{\circ}\right)^2}{\tilde{f}_{ij}^{\circ}} = n\big(\min\{k,l\} - 1\big)V^2\,.$$

Ničelno hipotezo torej zavrnemo, če je $\chi^2 > \chi^2_{1-\alpha}\big((k-1)(l-1)\big)$. Če jo zavrnemo pri $\alpha = 0{\cdot}05$, pravimo, da sta spremenljivki *značilno asociirani*, če jo zavrnemo pri $\alpha = 0{\cdot}01$, pa pravimo, da sta *zelo značilno asociirani*.

Opomba. Tudi kontingenčni test hi kvadrat ni eksakten, je zgolj približen. Dovolj natančen je pri naslednjih predpostavkah:

- Gre za enostavni slučajni vzorec iz velike populacije ali pa neodvisne izvedbe slučajnega poskusa, ki vsakič sledi istim verjetnostnim zakonitostim.
- *Pričakovane absolutne frekvence* so najmanj 5: $\tilde{f}_{ij} = n\tilde{f}^{\circ}_{ij} \geq 5$ za vse i in j. Ekvivalentno, veljati mora $\tilde{f}^{\circ}_{ij} \geq 5/n$. Sicer združimo bližnje razrede.

Primer: recimo, da prejšnja tabela barv las in oči pripada enostavnemu slučajnemu vzorcu iz velike populacije. Pri stopnji značilnosti $\alpha = 0{\cdot}01$ testiramo hipotezo, da sta barva oči in barva las na populaciji neasociirani. Najprej poglejmo, ali so sploh izpolnjeni pogoji za izvedbo testa. Za ta namen morajo biti pričakovane relativne navzkrižne frekvence vsaj $5/62 \doteq 0{\cdot}0806$. To pa ni res, zato združimo rdečelasce in blondince. Dobimo:

Opažene absolutne frekvence:

oči\lasje	rdeči, blond	rjavi, črni	Skupaj
modre	12	1	13
zelene	14	9	23
rjave	4	22	26
Skupaj	30	32	62

Opažene relativne frekvence:

oči\lasje	rdeči, blond	rjavi, črni	Skupaj
modre	0·194	0·016	0·210
zelene	0·226	0·145	0·371
rjave	0·065	0·355	0·419
Skupaj	0·484	0·516	1·000

Pričakovane absolutne frekvence:

oči\lasje	rdeči, blond	rjavi, črni	Skupaj
modre	6·29	6·71	13
zelene	11·13	11·87	23
rjave	12·58	13·42	26
Skupaj	30	32	62

Pričakovane relativne frekvence:

oči\lasje	rdeči, blond	rjavi, črni	Skupaj
modre	0·101	0·108	0·210
zelene	0·180	0·191	0·371
rjave	0·203	0·216	0·419
Skupaj	0·484	0·516	1·000

Cramérjev koeficient asociiranosti bo za združene razrede drugačen:

$$\frac{(0{\cdot}194-0{\cdot}101)^2}{0{\cdot}101}+\frac{(0{\cdot}016-0{\cdot}108)^2}{0{\cdot}108}+$$
$$+\frac{(0{\cdot}226-0{\cdot}180)^2}{0{\cdot}180}+\frac{(0{\cdot}145-0{\cdot}191)^2}{0{\cdot}191}+$$
$$+\frac{(0{\cdot}065-0{\cdot}203)^2}{0{\cdot}203}+\frac{(0{\cdot}355-0{\cdot}216)^2}{0{\cdot}216}\doteq 0{\cdot}36799\,,$$
$$V\doteq\sqrt{\frac{0{\cdot}36799}{2-1}}\doteq 0{\cdot}61\,.$$

Povezanost torej še vedno pride zmerna. Testna statistika pa pride:

$$\chi^2\doteq 62\cdot 0{\cdot}36799\doteq 22{\cdot}82\,.$$

Ker kritična vrednost pride $\chi^2_{0.99}(2)\doteq 9{\cdot}210$, ničelno hipotezo, da barva oči in barva las nista asociirani, zavrnemo: na našem vzorcu sta barvi statistično zelo značilno asociirani.

3.2 Povezanost dveh intervalskih spremenljivk: koreliranost

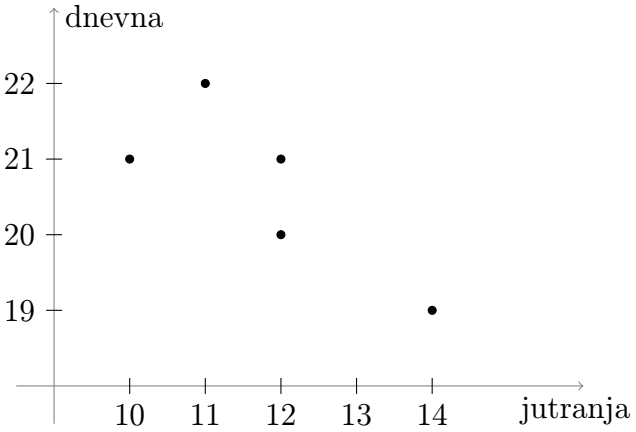
Koreliranost pove, v kolikšni meri sta spremenljivki povezani glede na naraščanje in padanje: če se ena od spremenljivk poveča, ali se druga v povprečju poveča, zmanjša ali nič od tega. Zato je koreliranost predznačena količina: možna je pozitivna ali negativna koreliranost.

Pri preučevanju koreliranosti nam pride prav *diagramom razpršenosti* (tudi *razsevni diagram*, angl. *scatter plot, scattergraph*), kjer podatke predstavimo kot pike v ravnini, pri čemer koordinata x pove vrednost prve, koordinata y pa vrednost druge spremenljivke.

Primer: vremenska napoved temperatur za naslednjih nekaj dni

dan	jutranja	dnevna
petek	14	19
sobota	12	20
nedelja	10	21
ponedeljek	11	22
torek	12	21

Pripadajoči diagram razpršenosti:



Več diagramov razpršenosti pride kasneje, prej pa bomo spoznali, kako koreliranost kvantitativno in kvalitativno opredelimo.

3.2.1 Kovarianca

Kovarianca je prvi korak do opredeljevanja povezanosti dveh intervalskih spremenljivk. Če vrednosti prve označimo z $x_1, \ldots, x_n$, vrednosti druge pa z $y_1, \ldots, y_n$, je kovarianca enaka:

$$K = K_{X,Y} := \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n},$$

kjer sta $\bar{x}$ in $\bar{y}$ aritmetični sredini naših spremenljivk:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}, \quad \bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

Če je kovarianca pozitivna, pravimo, da sta spremenljivki *pozitivno korelirani*. Če je negativna, sta *negativno korelirani*. Če je enaka nič, sta *nekorelirani*.

Kovarianca spremenljivke same s seboj je kvadrat standardnega odklona: $K_{X,X} = s_X^2$.

Kovarianco lahko računamo tudi po u-metodi: za poljubna u in v velja:

$$K_{X,Y} = \frac{(x_1 - u)(y_1 - v) + (x_2 - u)(y_2 - v) + \cdots + (x_n - u)(y_n - v)}{n} - (\bar{x} - u)(\bar{y} - v).$$

Primer: Izračun kovariance jutranjih (x_i) in dnevnih (y_i) temperatur (vzamemo $u = 10$ in $v = 20$):

dan	x_i	y_i	$x_i - 10$	$y_i - 20$	$(x_i - 10)(y_i - 20)$
	14	19	4	-1	-4
	12	20	2	0	2
	10	21	0	1	0
	11	22	1	2	0
	12	21	2	1	2
Vsota			9	3	2
Povprečje			1.8	0.6	0

Kovarianca: $K_{X,Y} \doteq 0 - 1.8 \cdot 0.6 = -1.08$.

Kovarianco lahko računamo tudi iz kontingenčne tabele:

$$\begin{aligned} K_{X,Y} &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij}(a_i - \bar{x})(b_j - \bar{y}) = \\ &= \sum_{i=1}^{k} \sum_{j=1}^{l} f^{\circ}_{ij}(a_i - \bar{x})(b_j - \bar{y}) = \\ &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij}(a_i - u)(b_j - v) - (\bar{x} - u)(\bar{y} - v) = \\ &= \sum_{i=1}^{k} \sum_{j=1}^{l} f^{\circ}_{ij}(a_i - u)(b_j - v) - (\bar{x} - u)(\bar{y} - v)\,. \end{aligned}$$

Še en primer: med 20 študenti izvedemo anketo z dvema vprašanjema:

1. Koliko ur na dan preživiš na računalniku?
2. Koliko ur na dan preživiš zunaj s prijatelji?

Rezultati ankete so zbrani v naslednji kontingenčni tabeli (vrstice so ure na računalniku, stolpci pa ure s prijatelji):

$X \setminus Y$	1	2	3	4	5	$f_{i\cdot}$
1	1	1	0	0	1	3
2	2	2	2	0	1	7
3	3	1	1	0	0	5
4	0	1	1	1	0	3
5	1	1	0	0	0	2
$f_{\cdot j}$	7	6	4	1	2	20

Najprej izračunamo povprečja:

$$\bar{x} = \frac{1\cdot 3 + 2\cdot 7 + 3\cdot 5 + 4\cdot 3 + 5\cdot 2}{20} = 2{\cdot}7\,,$$
$$\bar{y} = \frac{1\cdot 7 + 2\cdot 6 + 3\cdot 4 + 4\cdot 1 + 5\cdot 2}{20} = 2{\cdot}25\,.$$

Anketiranci so, kot trdijo, torej v povprečju na dan preživeli $2{\cdot}7$ ure za računalnikom in $2{\cdot}25$ ure s prijatelji.

Izračun kovariance po u-metodi pri $u = 3$ (za x) in $v = 2$ (za y):

$$\begin{aligned} K_{x,y} = \frac{1}{20}\Big[&(-2)\cdot(-1)\cdot 1 + (-2)\cdot 0\cdot 1 + (-2)\cdot 1\cdot 0 + (-2)\cdot 2\cdot 0 + (-2)\cdot 3\cdot 1 + \\ &+ (-1)\cdot(-1)\cdot 2 + (-1)\cdot 0\cdot 2 + (-1)\cdot 1\cdot 2 + (-1)\cdot 2\cdot 0 + (-1)\cdot 3\cdot 1 + \\ &+ 0\cdot(-1)\cdot 3 + 0\cdot 0\cdot 1 + 0\cdot 1\cdot 1 + 0\cdot 2\cdot 0 + 0\cdot 2\cdot 0 + \\ &+ 1\cdot(-1)\cdot 0 + 1\cdot 0\cdot 1 + 1\cdot 1\cdot 1 + 1\cdot 2\cdot 1 + 1\cdot 2\cdot 0 + \\ &+ 2\cdot(-1)\cdot 1 + 2\cdot 0\cdot 1 + 2\cdot 1\cdot 0 + 2\cdot 2\cdot 0 + 2\cdot 2\cdot 0\Big] - \\ &- (2{\cdot}7 - 3)\cdot(2{\cdot}25 - 2) = \\ = {}& -0{\cdot}225\,. \end{aligned}$$

S pomočjo kovariance na vzorcu lahko točkasto ocenimo kovarianco na celotni populaciji. Vendar pa moramo podobno kot pri standardnem odklonu za nepristransko oceno deliti z $n-1$ namesto z n. Če je torej $\sigma_{X,Y}$ populacijska kovarianca, je njena cenilka:

$$\hat{\sigma}[X,Y] = K_{X,Y+} := \frac{(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + \cdots + (x_n-\bar{x})(y_n-\bar{y})}{n-1}\,.$$

Primer: kovarianca na vzorcu 20 študentov je prišla $-0{\cdot}225$. Ocena za kovarianco na celotni populaciji pa je:

$$\hat{\sigma}[X,Y] = K_{X,Y+} = -\frac{20}{19}\cdot 0{\cdot}225 \doteq 0{\cdot}237\,.$$

Če preučujemo več spremenljivk hkrati (*multivariatna analiza*), je pomembna *kovariančna matrika*:

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1r} \\ K_{21} & K_{22} & \dots & K_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ K_{r1} & K_{r2} & \dots & K_{rr} \end{bmatrix}$$

S K_{ij} smo tu označili kovarianco i-te in j-te spremenljivke. Tako na primer v psihometriji (in tudi drugje) pomembno vlogo igra *Cronbachov*[2] α, ki je razmerje med vsoto kovarianc

[2]Lee Joseph Cronbach (1916–2001), ameriški psiholog

parov različnih spremenljivk in vsoto vseh kovarianc (t. j. vključno z variancami), vse skupaj pomnoženo z $r/(r-1)$:

$$\alpha = \frac{r}{r-1}\,\frac{\sum_{i,j;i\neq j} K_{ij}}{\sum_{i,j} K_{ij}} = \frac{r}{r-1}\left(1 - \frac{\sum_i K_{ii}}{\sum_{i,j} K_{ij}}\right).$$

Če so komponente nekorelirane, je $\alpha = 0$.

Večina metod v multivariatni analizi zahteva matematično analizo matrik, ki temelji na *linearni algebri*.

3.2.2 Pearsonov korelacijski koeficient

Kovarianca sama po sebi ni dobro merilo za stopnjo povezanosti, saj je odvisna od merskih enot: če npr. eno od spremenljivk pomnožimo s 100 (recimo če jo podamo v centimetrih namesto v metrih), se tudi kovarianca pomnoži s 100. Pearsonov[3] korelacijski koeficient to pomanjkljivost odpravi tako, da kovarianco deli s produktom standardnih odklonov:

$$r = r_{X,Y} = \frac{K_{X,Y}}{s_X\, s_Y},$$

kjer je:

$$s_X = \sqrt{\frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n}},$$
$$s_Y = \sqrt{\frac{(y_1-\bar{y})^2 + (y_2-\bar{y})^2 + \cdots + (y_n-\bar{y})^2}{n}}.$$

Lastnosti Pearsonovega korelacijskega koeficienta:

- Definiran je, če nobena od spremenljivk ni konstantna.
- Velja $-1 \leq r \leq 1$.
- Če sta X in Y neodvisni (na statistični množici, iz katere so podatki), je $r = 0$. Velja tudi, da, če podatki temeljijo na velikem enostavnem slučajnem vzorcu iz velike populacije, na kateri sta X in Y neodvisni, je r blizu 0 (malo kasneje pri testiranju se bomo naučili, kako postaviti mejo).
- Pearsonov korelacijski koeficient je maksimalen (enak 1), če je katera koli od spremenljivk naraščajoča linearna funkcija druge.
- Pearsonov korelacijski koeficient je minimalen (enak -1), če je katera koli od spremenljivk padajoča linearna funkcija druge.

[3]Karl Pearson (1857–1936), angleški matematik in biostatistik

Pearsonov korelacijski koeficient meri stopnjo *linearne* povezanosti med statističnima spremenljivkama.

Absolutna vrednost Pearsonovega korelacijskega koeficienta ($|\,r\,|$) je v grobem primerljiva s Cramérjevim koeficientom asociiranosti. Še več, za par dihotomnih spremenljivk se to dvoje celo ujema, ne glede na to, kateri dve (različni) števili priredimo vrednostma posamezne spremenljivke (ki nista nujno številski – lahko sta le imenski). Če gledamo par urejenostnih dihotomnih spremenljivk in če številske vrednosti priredimo v skladu z urejenostjo, velja $r = \phi$.

Zato je smiselno, če tudi Pearsonov korelacijski koeficient enako kvalitativno opredeljujemo kot pri Cramérjev koeficient, s tem da lahko sedaj povemo tudi smer povezanosti: vrednost $r = -0{\cdot}6$ torej pomeni zmerno negativno koreliranost.

Kvadratu korelacijskega koeficienta (r^2) pravimo *determinacijski koeficient*. Njegovo kvalitativno opredeljevanje je torej naslednje:

- do $0{\cdot}04$: neznatna povezanost;
- od $0{\cdot}04$ do $0{\cdot}16$: rahla povezanost;
- od $0{\cdot}16$ do $0{\cdot}49$: zmerna povezanost;
- od $0{\cdot}49$ do $0{\cdot}81$: močna povezanost;
- od $0{\cdot}81$ do 1: zelo močna povezanost.

Primer: pri vremenski napovedi temperatur:

dan	jutranja	dnevna
petek	14	19
sobota	12	20
nedelja	10	21
ponedeljek	11	22
torek	12	21

pride:

$$s_X \doteq 1{\cdot}327\,, \quad s_Y \doteq 1{\cdot}020\,; \qquad r_{X,Y} \doteq \frac{-1{\cdot}08}{1{\cdot}327 \cdot 1{\cdot}020} \doteq -0{\cdot}80\,.$$

Determinacijski koeficient: $0{\cdot}64$.
Jutranja in dnevna temperatura sta torej visoko *negativno* povezani: pri višji jutranji temperaturi lahko pričakujemo nižjo dnevno.

Pri takšni napovedi, kot je ta (za nekaj zaporednih dni) ima pri korelaciji verjetno največjo težo vpliv oblačnosti, ki viša jutranjo, a niža dnevno temperaturo. Pri napovedi za daljše obdobje bi bila korelacija bistveno drugačna.

Primer: pri kontingenčni tabeli, ki se nanaša na vprašanji, koliko ur anketirani študent preživi na računalniku in koliko s prijatelji:

$X \setminus Y$	1	2	3	4	5	$f_{i\cdot}$
1	1	1	0	0	1	3
2	2	2	2	0	1	7
3	3	1	1	0	0	5
4	0	1	1	1	0	3
5	1	1	0	0	0	2
$f_{\cdot j}$	7	6	4	1	2	20

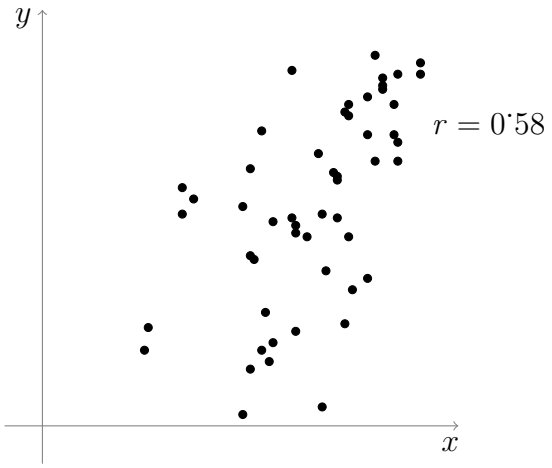
pride:

$$s_X \doteq 1{\cdot}1874\,, \quad s_Y \doteq 1{\cdot}2600\,, \qquad r_{X,Y} \doteq \frac{-0{\cdot}225}{1{\cdot}1874 \cdot 1{\cdot}2600} \doteq -0{\cdot}15\,.$$

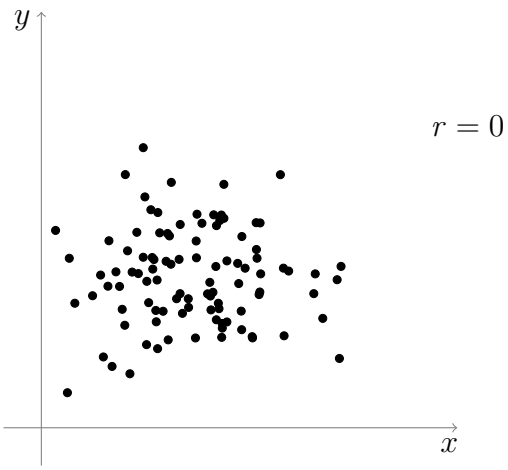
Gre torej za neznatno negativno povezanost.

Še nekaj primerov diagramov razpršenosti z različnimi korelacijskimi koeficienti:

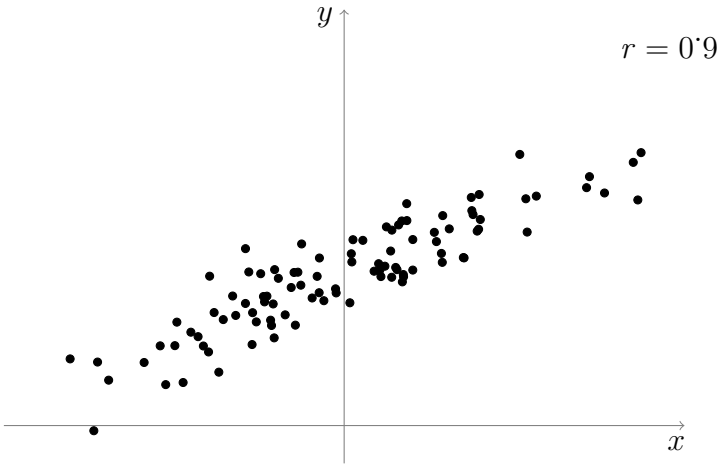
To so rezultati tistih 52 študentov, ki so v študijskem letu 2010/11 na biopsihologiji pisali oba kolokvija iz statistike:



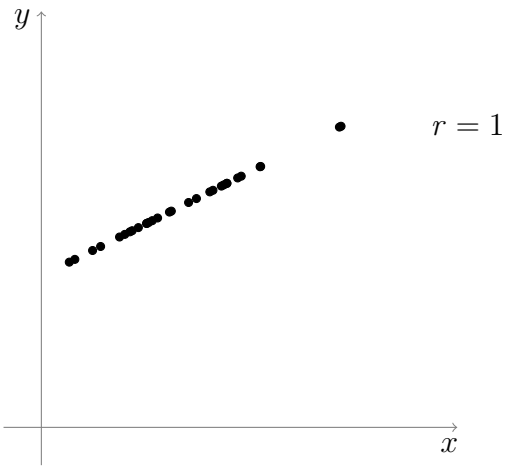
To so naključno generirani podatki, a naključnost je nastavljena tako, da so nekorelirani.



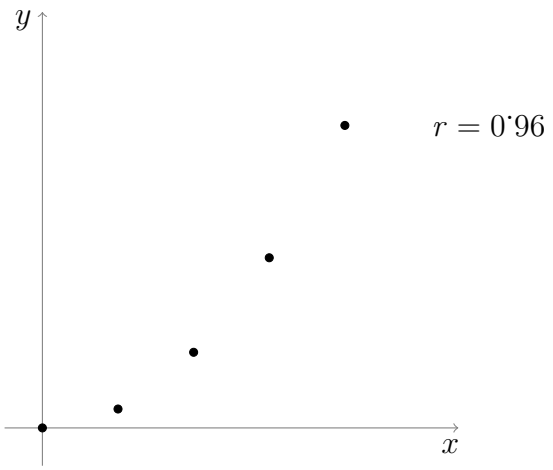
Ti podatki so visoko pozitivno korelirani.



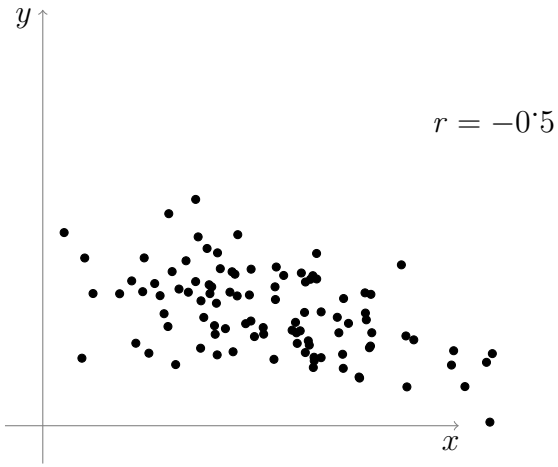
Skrajni primer je korelacija 1, ko gre za linearno odvisnost.



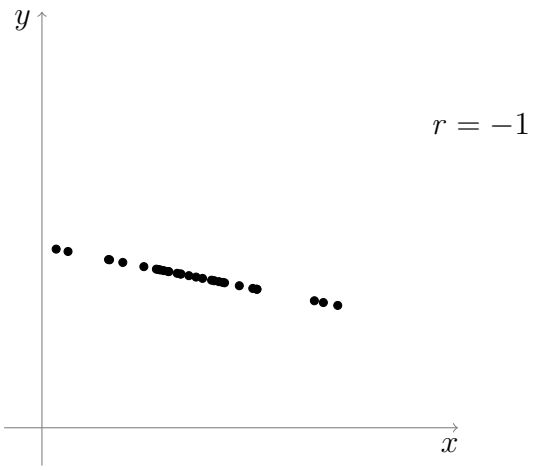
Tu korelacija ni enaka 1, čeprav sta spremenljivki v deterministični strogo naraščajoči povezavi:



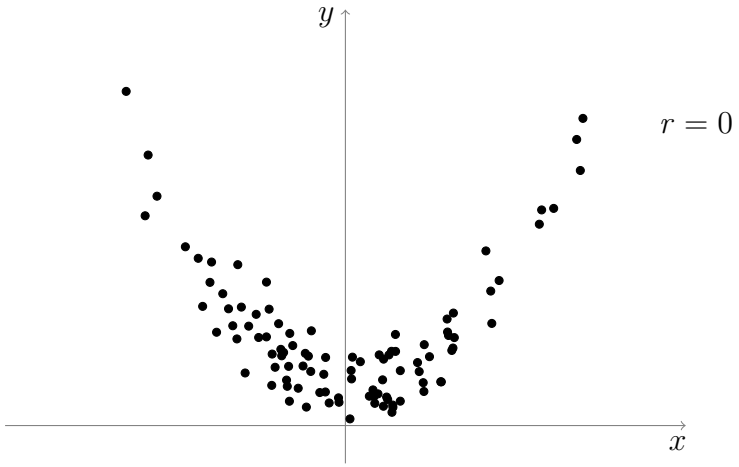
Korelacija je lahko tudi negativna:



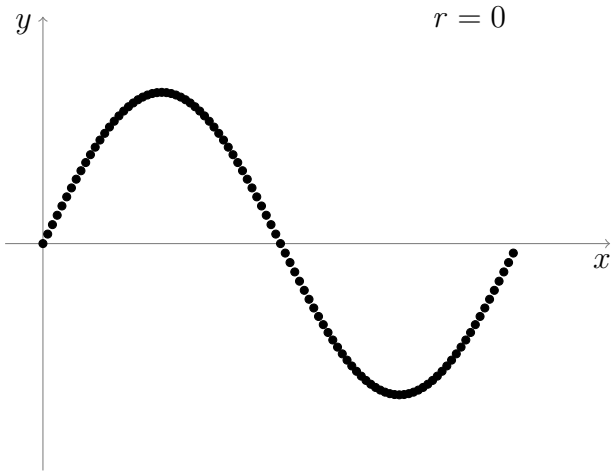
Tukaj je korelacija enaka -1:



Še en primer nekoreliranih podatkov:



Tukaj se y deterministično izraža z x, podatki pa so nekorelirani.



3.2.3 Testiranje nekoreliranosti

Ničelno hipotezo H_0, da sta spremenljivki nekorelirani, testiramo s T-testom, in sicer s pomočjo testne statistike:

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}\,,$$

kjer je $r = r_{X,Y}$ Pearsonov korelacijski koeficient. Število prostostnih stopenj je $n-2$. Aktualne so tri alternativne hipoteze. Ničelno hipotezo zavrnemo:

- proti $H_1^{\pm}$, da sta X in Y korelirani, če je $|T| > t_{1-\alpha/2}(n-2)$;
- proti H_1^{+}, da sta X in Y pozitivno korelirani, če je $T > t_{1-\alpha}(n-2)$;
- proti H_1^{-}, da sta X in Y negativno korelirani, če je $T < -t_{1-\alpha}(n-2)$.

Test je zasnovan ob predpostavki, da sta X in Y na populaciji porazdeljeni normalno.

Primer: pri temperaturah za 5 dni je korelacijski koeficient prišel $-0{\cdot}80$, torej je bila koreliranost močna. Recimo, da bi šlo za enostavni slučajni vzorec in da bi testirali ničelno hipotezo, da sta jutranja in dnevna temperatura nekorelirani, proti alternativni hipotezi, da sta korelirani. Testna statistika pride:

$$T = -\frac{0{\cdot}80}{\sqrt{1-0{\cdot}80^2}}\sqrt{3} \doteq -2{\cdot}31\,.$$

Pri stopnji značilnosti $\alpha = 0{\cdot}05$ to primerjamo s $t_{0.975}(3) \doteq 3{\cdot}182$ in hipoteze ne zavrnemo: koreliranost ni statistično značilna.

Primer: pri 52 študentih, ki so v študijskem letu 2010/11 na biopsihologiji pisali oba kolokvija iz statistike, korelacijski koeficient med obema kolokvijema pride $0{\cdot}58$, torej je koreliranost zmerna. Pa recimo, da bi bil to spet enostavni slučajni vzorec in da bi testirali ničelno hipotezo, prvi in drugi kolokvij nekorelirani, proti alternativni hipotezi, da sta korelirana. Testna statistika pride:

$$T = \frac{0{\cdot}58}{\sqrt{1-0{\cdot}58^2}}\sqrt{50} \doteq 5{\cdot}03\,.$$

Ker je $t_{0.995}(50) \doteq 2{\cdot}678$, je koreliranost statistično zelo značilna, čeprav je "le" zmerna, medtem ko je bila prej koreliranost visoka, a statistično neznačilna. Toda zdaj smo imeli na voljo precej več podatkov.

3.3 Povezanost intervalske in dihotomne spremenljivke: primerjava povprečij

Podatke, kjer sta na isti statistični množici definirani intervalska spremenljivka (recimo U) in dihotomna spremenljivka (recimo G), lahko predstavimo bodisi kot:

$$u_1, u_2, \ldots, u_N$$
$$g_1, g_2, \ldots, g_N$$

bodisi podatke razdelimo glede na vrednost dihotomne spremenljivke. Če le-ta zavzame vrednosti a in b, lahko podatke, na katerih druga spremenljivka zavzame vrednost a, predstavimo z:

$$x_1, x_2, \ldots, x_m\,,$$

podatke, na katerih druga spremenljivka zavzame vrednost b, pa z:

$$y_1, y_2, \ldots, y_n\,.$$

Še drugače, gledati dve spremenljivki, od katerih je druga dihotomna, na eni statistični množici, je ekvivalentno gledanju prve spremenljivke na dveh različnih statističnih množicah (dihotomna spremenljivka nam statistično množico razdeli na dve skupini).

Primer: pri nekem izpitu gledamo rezultat in spol:

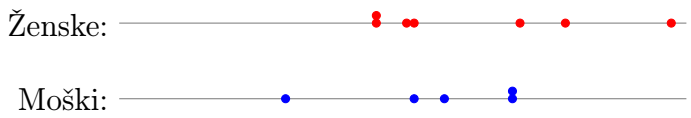
Ime	Rezultat (u_i)	Spol (g_i)
Jan	22	M
Karmen	39	Ž
Barbara	73	Ž
Kristina	34	Ž
Domen	52	M
Katja	34	Ž
Aljaž	39	M
Rok	52	M
Sabina	38	Ž
Diana	53	Ž
Jerica	59	Ž
Tilen	43	M

Rezultate lahko ločimo po spolih:

Ženske: $x_1 = 39\,,\ x_2 = 73\,,\ x_3 = 34\,,\ x_4 = 34\,,\ x_5 = 38\,,\ x_6 = 53\,,\ x_7 = 59\,.$

Moški: $y_1 = 22\,,\ y_2 = 52\,,\ y_3 = 39\,,\ y_4 = 52\,,\ y_5 = 43\,.$

Prikaz s pikami:



3.3.1 Točkovni biserialni korelacijski koeficient

Če sta a in b numerični vrednosti, lahko izračunamo Pearsonov korelacijski koeficient. Brž ko je $a > b$, je le-ta enak:

$$r_{\text{pb}} = \frac{\bar{x} - \bar{y}}{s}\,\frac{\sqrt{mn}}{m+n}\,,$$

ne glede na to, koliko sta vrednosti a in b dejansko enaki. Zato ni nujno, da je dihotomna spremenljivka numerična, lahko je le imenska. Koeficient za ta primer imenujemo *točkovni biserialni korelacijski koeficient* (angl. *point biserial correlation coefficient*).

Oznaka s se tu nanaša na *skupni* standardni odklon, t. j. standardni odklon spremenljivke U:

$$s = \sqrt{\frac{(u_1-\bar{u})^2+(u_2-\bar{u})^2+\cdots+(u_N-\bar{u})^2}{N}} = \\ = \sqrt{\frac{(x_1-\bar{u})^2+\cdots+(x_m-\bar{u})^2+(y_1-\bar{u})^2+\cdots+(y_n-\bar{u})^2}{N}}\,,$$

$\bar{u}$ pa je aritmetična sredina vseh podatkov:

$$\bar{u} = \frac{u_1+u_2+\cdots+u_N}{N}\,,$$

Pri našem primeru je:

$$\bar{x} \doteq 47{\cdot}14\,,\ \ \bar{y} = 41{\cdot}6\,,\ \ \bar{u} \doteq 44{\cdot}83\,,\ \ s^2 \doteq 169{\cdot}8\,,\ \ s \doteq 13{\cdot}03\,,$$

torej je:

$$r_{\mathrm{pb}} \doteq \frac{47{\cdot}14-41{\cdot}6}{13{\cdot}03}\,\frac{\sqrt{7\cdot 5}}{7+5} \doteq 0{\cdot}2097\,.$$

Točkovni biserialni koeficient je vedno med -1 in 1. Skrajni vrednosti sta doseženi takrat, ko so vse vrednosti na posamezni skupini enake. Vrednost 1 je dosežena takrat, ko so vrednosti na prvi skupini strogo večje od vrednosti na drugi skupini, vrednost -1 pa je dosežena takrat, ko so vrednosti na prvi skupini strogo manjše od vrednosti na drugi skupini. Če so kar vse vrednosti enake, pa koeficient ni definiran.

Kvalitativno opredeljevanje točkovnega biserialnega koeficienta je enako kot pri Pearsonovem: pri prejšnjem primeru gre torej za rahlo povezanost v korist žensk. Ali drugače, ženske so pisale *malo* boljše kot moški.

Aritmetična sredina vseh podatkov je enaka tehtani sredini aritmetičnih sredin posameznih skupin:

$$\bar{u} = \frac{m}{m+n}\,\bar{x} + \frac{n}{m+n}\,\bar{y}\,.$$

z utežema, ki sta sorazmerni z velikostma skupin, ki ju predstavljata.

Kvadrat skupnega standardnega odklona, torej skupno varianco, pa lahko zapišemo kot vsoto:

$$s^2 = s_{\mathrm{W}}^2 + s_{\mathrm{B}}^2\,,$$

kjer je:

$$s_{\mathrm{W}}^2 = \frac{m}{m+n}\,s_X^2 + \frac{n}{m+n}\,s_Y^2$$

varianca *znotraj skupin* (angl. *within groups*) ali tudi *nepojasnjena varianca* (angl. *unexplained variance, pooled variance*) in:

$$s_{\mathrm{B}}^2 = \frac{mn}{(m+n)^2}\,(\bar{x}-\bar{y})^2$$

varianca *med skupinama* (angl. *between groups* ali tudi *pojasnjena varianca* (angl. *explained variance*). To je tisti del variance, ki jo pojasnjuje skupina, v kateri je podatek.

Na zgornji in splošnejših razčlenitvah variance temelji *analiza variance* (angl. *analysis of variance, ANOVA*), ki je pomemben del inferenčne statistike. Malo kasneje bomo omenili posplošitev na več skupin.

Kvadrat točkovnega biserialnega korelacijskega koeficienta (točkovni biserialni determinacijski koeficient) predstavlja *delež pojasnjene variance* ali tudi *moč učinka* (angl. *strength of effect, effect size*), saj velja:

$$r_{\mathrm{pb}}^2 = \frac{s_{\mathrm{B}}^2}{s^2} = \frac{s_{\mathrm{B}}^2}{s_{\mathrm{W}}^2 + s_{\mathrm{B}}^2}\,.$$

Njegovo kvalitativno opredeljevanje je torej enako kot pri determinacijskem koeficientu.

Pri našem primeru je recimo:

$$\bar{u} \doteq 44{\cdot}83 \doteq \frac{7}{12} \cdot 47{\cdot}14 + \frac{5}{12} \cdot 41{\cdot}6\,.$$

Varianca med skupinama (varianca, pojasnjena s spolom), je enaka:

$$s_{\mathrm{B}}^2 = \frac{7 \cdot 5}{(7+5)^2}\,(47{\cdot}14 - 41{\cdot}6)^2 \doteq 7{\cdot}5\,.$$

Nadalje je:

$$s_X^2 \doteq 191{\cdot}3\,,\ s_X \doteq 13{\cdot}83\,, \quad s_Y^2 \doteq 121{\cdot}8\,,\ s_Y \doteq 11{\cdot}04\,.$$

in varianca znotraj skupin (nepojasnjena varianca) je enaka:

$$s_{\mathrm{W}}^2 \doteq \frac{7}{12} \cdot 191{\cdot}3 + \frac{5}{12} \cdot 121{\cdot}8 \doteq 162{\cdot}3\,.$$

Opazimo, da je res $s^2 \doteq 169{\cdot}8 = 7{\cdot}5 + 162{\cdot}3 \doteq s_{\mathrm{B}}^2 + s_{\mathrm{W}}^2$. Delež pojasnjene variance je enak:

$$\frac{7{\cdot}5}{169{\cdot}8} \doteq 0{\cdot}04398 \doteq 0{\cdot}2097^2\,.$$

Če sta obe spremenljivki dihotomni, velja $r_{\mathrm{pb}} = \phi$.

3.3.2 Standardizirana razlika povprečij

Kot mero za povezanost intervalske in dihotomne spremenljivke lahko gledamo tudi *standardizirano razliko povprečij* (angl. *standardized mean difference*) ali tudi *Cohenov*[4] *koeficient*:

$$d = \frac{\bar{x} - \bar{y}}{s_{\mathrm{W}}}$$

[4]Jacob Cohen (1923–1998), ameriški statistik in psiholog

Točkovni biserialni in Cohenov koeficient nam dajeta isto informacijo, saj se izražata drug z drugim:

$$r_{\mathrm{pb}} = \frac{d}{\sqrt{d^2 + \frac{(m+n)^2}{mn}}}\,, \qquad d = \frac{m+n}{\sqrt{mn}}\,\frac{r_{\mathrm{pb}}}{\sqrt{1-r_{\mathrm{pb}}^2}}\,.$$

Nudita pa dva različna pogleda: točkovni biserialni korelacijski koeficient je osredotočen bolj na povezanost, Cohenov koeficient pa bolj na razliko.

V našem primeru je:

$$d \doteq \frac{47{\cdot}14 - 41{\cdot}6}{\sqrt{162{\cdot}3}} \doteq 0{\cdot}435\,.$$

3.3.3 Testiranje enakosti povprečij

Recimo, da sta na določeni populaciji definirani intervalska in dihotomna spremenljivka. Dihotomna spremenljivka razdeli populacijo na dve podpopulaciji. Označimo z μ_X povprečje spremenljivke na prvi, z μ_Y pa na drugi podpopulaciji. Testiramo ničelno hipotezo H_0, da je $\mu_X = \mu_Y$.

Opažene enote, na podlagi katerih opravimo test, so lahko dobljene kot enostavni slučajni vzorec velikosti vsaj 3, pri čemer mora biti iz vsake podpopulacije vsaj ena enota; če se to ne zgodi (kar je pri običajnih velikostih vzorcev in običajnih deležih podpopulacij zelo malo verjetno), vzorčenje ponovimo. Lahko pa tudi predpišemo minimalno ali točno število enot iz posamezne podpopulacije. V slednjem primeru iz vsake vzamemo enostavni slučajni vzorec, vzorca pa morata biti med seboj neodvisna. Na populaciji torej izvedemo preprosto *stratificirano vzorčenje* – stratificiramo glede na dihotomno spremenljivko. Za ta namen moramo seveda imeti popoln pregled nad njo.

Podobno lahko test uporabimo tudi, če so opažene vrednosti dobljene iz izvedb dveh slučajnih poskusov, za katere privzamemo, da so vse med seboj neodvisne. Števili izvedb posameznega poskusa sta lahko različni, morata pa biti predpisani. Poskusa imata lahko različne verjetnostne zakonitosti. V tem primeru je μ_X pričakovana vrednost slučajne spremenljivke pri prvem, μ_Y pa pri drugem poskusu.

V vsakem primeru je vzorec razdeljen na dve skupini. Označimo z m velikost prve, z n pa velikost druge skupine. Ničelno hipotezo testiramo s T-testom, in sicer s pomočjo testne statistike:

$$T = \frac{\bar{x} - \bar{y}}{\mathrm{SE}} = d\,\frac{\sqrt{(N-2)mn}}{N}\,,$$

kjer je:

$$\mathrm{SE} = s_{\mathrm{W}+}\sqrt{\frac{N}{mn}}\,,$$
$$s_{\mathrm{W}+} = \sqrt{\frac{(x_1-\bar{x})^2 + \cdots + (x_m-\bar{x})^2 + (y_1-\bar{y})^2 + \cdots + (y_n-\bar{y})^2}{N-2}}\,.$$

Število prostostnih stopenj je $N-2$, kar pomeni, da ničelno hipotezo zavrnemo:

- proti $H_1^{\pm}$: $\mu_X \neq \mu_Y$ korelirani, če je $|T| > t_{1-\alpha/2}(n-2)$;
- proti H_1^X: $\mu_X > \mu_Y$, če je $T > t_{1-\alpha}(n-2)$;
- proti H_1^Y: $\mu_Y > \mu_X$, če je $T < -t_{1-\alpha}(n-2)$.

Test je eksakten pod predpostavko, da je spremenljivka na obeh podpopulacijah porazdeljena normalno in da sta varianci na obeh podpopulacijah enaki (temu pravimo *homoskedastičnost*). Če se podatki nanašajo na izvedbe poskusov, predpostavka trdi, da je spremenljivka pri obeh poskusih porazdeljena normalno in da ima pri obeh poskusih enako varianco. Je pa test pri velikih vzorcih v precejšnji meri robusten, kar pomeni, da je stopnja značilnosti še vedno približno drži, tudi če ti dve predpostavki nista tako natančno izpolnjeni.

Primer: spol in rezultat kolokvija. Pri stopnji značilnosti $\alpha \doteq 0{\cdot}05$ dvostransko testiramo, ali sta spol in rezultat nedvisna. Spomnimo se:

$$m = 7\,, \quad n = 5\,, \quad \bar{x} \doteq 47{\cdot}14\,, \quad \bar{y} \doteq 41{\cdot}60\,, \quad s_{\mathrm{W}}^2 \doteq 162{\cdot}3\,.$$

Od tod dobimo:

$$s_{\mathrm{W}+} \doteq \sqrt{\frac{12}{10} \cdot 162{\cdot}3} \doteq 13{\cdot}96\,, \quad \mathrm{SE} \doteq \sqrt{\frac{12}{35}} \cdot 13{\cdot}96 \doteq 8{\cdot}173\,.$$

Testna statistika pride:

$$\frac{47{\cdot}14 - 41{\cdot}60}{8{\cdot}173} \doteq 0{\cdot}678\,.$$

Glede na test moramo to primerjati s $t_{0.975}(10) \doteq 2{\cdot}228$, torej hipoteze ne zavrnemo: razlike med spoloma niso statistično značilne.

3.4 Povezanost intervalske in imenske spremenljivke: analiza variance z enojno klasifikacijo

3.4.1 Pojasnjena in nepojasnjena varianca

Podatke, kjer sta na isti statistični množici definirani intervalska spremenljivka (recimo X) in imenska spremenljivka (recimo G), lahko spet predstavimo na dva načina. Tako, kot je prej opisano, bomo vrednosti intervalske in imenske spremenljivke tokrat označevali z:

$$x_1, x_2, \ldots, x_n$$
$$g_1, g_2, \ldots, g_n$$

Lahko pa spet podatke razdelimo glede na vrednost imenske spremenljivke – preindeksiramo jih na naslednji način:

$$\begin{aligned} x_{11}, x_{12}, \ldots, x_{1n_1} &: \text{vrednosti spremenljivke } X, \text{ kjer je } G = g_1 \\ x_{21}, x_{22}, \ldots, x_{2n_2} &: \text{vrednosti spremenljivke } X, \text{ kjer je } G = g_2 \\ &\vdots \\ x_{k1}, x_{k2}, \ldots, x_{kn_2} &: \text{vrednosti spremenljivke } X, \text{ kjer je } G = g_k \end{aligned}$$

Seveda velja $n_1 + n_2 + \cdots + n_k = n$.

Še drugače, na eni statistični množici gledati dve spremenljivki, od katerih je druga imenska, ki zavzame k vrednosti, je ekvivalentno gledanju prve spremenljivke na k različnih statističnih množicah (imenska spremenljivka nam statistično množico razdeli na k skupin).

Merjenje povezanosti med intervalsko in imensko spremenljivko temelji na analizi variance (natančneje, v našem kontekstu je to analiza variance z enojno klasifikacijo). Označimo z $\bar{x}$ aritmetično sredino na celotni statistični množici:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}\,,$$

z $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ pa aritmetične sredine na posameznih skupinah:

$$\bar{x}_i = \frac{x_{i1} + x_{i2} + \cdots + x_{in_i}}{n_i}\,,$$

Tedaj je $\bar{x}$ tehtana sredina aritmetičnih sredin μ_i:

$$\bar{x} = \frac{n_1}{n}\,\bar{x}_1 + \frac{n_2}{n}\,\bar{x}_2 + \cdots + \frac{n_k}{n}\,\bar{x}_k\,.$$

Skupna varianca:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

spet razpade na nepojasnjeno in pojasnjeno varianco:

$$s^2 = s_{\mathrm{W}}^2 + s_{\mathrm{B}}^2\,,$$

kjer je nepojasnjena varianca ali varianca znotraj skupin tehtana sredina posameznih varianc v skupinah:

$$\begin{aligned} s_i^2 &= \frac{(x_{i1} - \bar{x}_i)^2 + (x_{i2} - \bar{x}_i)^2 + \cdots + (x_{in_i} - \bar{x}_i)^2}{n_i}\,, \\ s_{\mathrm{W}}^2 &= \frac{n_1}{n}\,s_1^2 + \frac{n_2}{n}\,s_2^2 + \cdots + \frac{n_k}{n}\,s_k^2\,, \end{aligned}$$

pojasnjena varianca ali varianca med skupinami pa je tehtana sredina kvadratov odklonov aritmetičnih sredin posameznih skupin od *skupne* aritmetične sredine:

$$s_{\mathrm{B}}^2 = \frac{n_1}{n}\,(\bar{x}_1 - \bar{x})^2 + \frac{n_2}{n}\,(\bar{x}_2 - \bar{x})^2 + \cdots + \frac{n_k}{n}\,(\bar{x}_k - \bar{x})^2\,.$$

Lahko jo izračunamo tudi po u-metodi:

$$s_{\mathrm{B}}^2 = \frac{n_1}{n}\,(\bar{x}_1 - u)^2 + \frac{n_2}{n}\,(\bar{x}_2 - u)^2 + \cdots + \frac{n_k}{n}\,(\bar{x}_k - u)^2 - (\bar{x} - u)^2\,.$$

Zgoraj definirane variance so posplošitve varianc, ki smo jih gledali pri povezanosti intervalske in dihotomne spremenljivke. Tako povezanost intervalske in imenske spremenljivke spet merimo z deležem pojasnjene variance oz. močjo učinka:

$$\eta^2 = \frac{s_{\mathrm{B}}^2}{s^2}\,,$$

Delež pojasnjene variance kvalitativno opredeljujemo enako kot determinacijski koeficient: če je pojasnjenih 25% variance ($\eta^2 = 0{\cdot}25$), to pomeni zmerno povezanost.

Opombe:

- Delež pojasnjene variance je enak nič natanko tedaj, ko so povprečja na vseh skupinah enaka (niso pa enake vse vrednosti spremenljivke).
- Delež pojasnjene variance je enak ena natanko tedaj, ko so na vsaki skupini vse vrednosti spremenljivke enake (med skupinami pa niso vse enake).
- Če je intervalska spremenljivka dihotomna, se delež pojasnjene variance ujema s kvadratom Cramérjevega koeficienta asociiranosti: $\eta^2 = V^2$. Če sta torej obe spremenljivki dihotomni, velja $\eta^2 = \phi^2$.

Primer: primerjava rezultatov kolokvijev v študijskem letu 2010/11 med študenti biopsihologije pri predmetu Statistika, univerzitetnimi študenti matematike pri predmetu Verjetnost in statistika in študenti praktične matematike pri predmetu Matematika 2. Šteti so le študenti, ki so pisali polno število zahtevanih kolokvijev.

Biopsihologi (52):

Univerzitetni matematiki (48):

Praktični matematiki (22):

Za delež pojasnjene variance poleg števila študentov v posamezni skupini potrebujemo še aritmetično sredino in standardni odklon. Za posamezne skupine je to enako:

Biopsihologi:	$\bar{x}_1 \doteq 64{\cdot}885\,,$	$s_1 \doteq 19{\cdot}603\,.$
Univerzitetni matematiki:	$\bar{x}_2 \doteq 63{\cdot}056\,,$	$s_2 \doteq 14{\cdot}891\,.$
Praktični matematiki:	$\bar{x}_3 \doteq 54{\cdot}318\,,$	$s_3 \doteq 11{\cdot}279\,.$

Lotimo se računanja. Najprej seštejmo, koliko je skupaj študentov:

$$n = 52 + 48 + 22 = 122\,.$$

Nato izračunamo skupno povprečje:

$$\bar{x} \doteq \frac{52}{122} \cdot 64{\cdot}885 + \frac{48}{122} \cdot 63{\cdot}056 + \frac{22}{122} \cdot 54{\cdot}318 \doteq 62{\cdot}260\,.$$

Podobno dobimo varianco znotraj skupin:

$$s_{\mathrm{W}}^2 \doteq \frac{52}{122} \cdot 19{\cdot}603^2 + \frac{48}{122} \cdot 14{\cdot}891^2 + \frac{22}{122} \cdot 11{\cdot}279^2 \doteq 273{\cdot}99\,.$$

Varianca med skupinami pa je enaka:

$$s_{\mathrm{B}}^2 \doteq \frac{52}{122}(64{\cdot}885 - 62{\cdot}260)^2 + \frac{48}{122}(63{\cdot}056 - 62{\cdot}260)^2 + \frac{22}{122} \cdot (54{\cdot}318 - 62{\cdot}260)^2 \doteq 14{\cdot}56\,.$$

Torej je skupna varianca enaka:

$$s^2 \doteq 273{\cdot}99 + 14{\cdot}56 \doteq 288{\cdot}55\,,$$

delež pojasnjene variance pa je:

$$\eta^2 \doteq \frac{14{\cdot}56}{288{\cdot}55} \doteq 0{\cdot}0505\,.$$

Z drugimi besedami, študijski program pojasni dobrih 5% variance, ostalih slabih 95% variance pa nastane zaradi drugih vplivov. Kvalitativno gre za *rahlo* povezanost.

Primer: želimo izmeriti povezavo med pogostnostjo zahajanja v kino in najbolj priljubljeno zvrstjo filma. Za ta namen izvedemo anketo z dvema vprašanjema:

1. Kolikokrat na mesec greš v kino?
2. Katera zvrst filma ti je najbolj všeč?
   (a) Komedija.
   (b) Akcija.

(c) Romantični film.

(d) Drama.

(e) Grozljivka.

Pogostnost zahajanja v kino je intervalska, zvrst filma pa imenska spremenljivka. Rezultati ankete so naslednji:

zvrst filma\št. obiskov kina	0	1	2	Skupaj	Povprečje
komedija	4	2	2	8	0·75
akcija	0	1	0	1	1
romantični	0	3	1	4	1·25
drama	4	1	2	7	0·7143
grozljivka	0	0	0	0	–
Skupaj	8	7	5	20	0·85

Skupna varianca:

$$s^2 = \frac{8 \cdot (0 - 0{\cdot}85)^2 + 7 \cdot (1 - 0{\cdot}85)^2 + 5 \cdot (2 - 0{\cdot}85)^2}{20} = 0{\cdot}6275\,.$$

Varianca med skupinami (pojasnjena varianca):

$$s_{\mathrm{B}}^2 = \frac{8 \cdot (0{\cdot}75 - 0{\cdot}85)^2 + 1 \cdot (1 - 0{\cdot}85)^2 + 4 \cdot (1{\cdot}25 - 0{\cdot}85)^2 + 7 \cdot (0{\cdot}7143 - 0{\cdot}85)^2}{20} \doteq$$
$$\doteq 0{\cdot}0436\,.$$

Delež pojasnjene variance:

$$\eta^2 \doteq \frac{0{\cdot}0436}{0{\cdot}6275} \doteq 0{\cdot}069\,.$$

Različnost najljubših zvrsti filma torej pojasni 6·9% variance števila obiskov kina. To pomeni *rahlo* povezanost.

Recimo zdaj, da so naši podatki dobljeni vzorec iz določene populacije. Želeli bi testirati ničelno hipotezo, da med intervalsko in imensko spremenljivko (skupino) ni povezave, proti alternativni hipotezi, da povezava je. To izvedemo z *F-testom* na testni statistiki:

$$F = \frac{n-k}{k-1}\,\frac{s_{\mathrm{B}}^2}{s_{\mathrm{W}}^2} = \frac{n-k}{k-1}\,\frac{\eta^2}{1-\eta^2}$$

s $(k-1, n-k)$ prostostnimi stopnjami, in sicer uporabimo enostransko različico v desno. To pomeni, da ničelno hipotezo zavrnemo, če je $F > F_{1-\alpha}(k-1, n-k)$ kjer je $F_p(r,s)$ kvantil *Fisher*[5]*–Snedecorjeve*[6] *porazdelitve* z (r,s) prostostnimi stopnjami. Omenjeni test je eksakten pod naslednjimi predpostavkami:

[5]Sir Ronald Aymler Fisher (1899–1962), angleški statistik in biolog

[6]George Waddel Snedecor (1882–1974), ameriški matematik in statistik

- Na vsaki skupini, t. j. za vsako vrednost imenske spremenljivke, vzamemo enostavni slučajni vzorec predpisane velikosti – torej stratificiramo glede na imensko spremenljivko.
- Vzorci morajo biti med seboj neodvisni.
- Na vsaki skupini je spremenljivka porazdeljena normalno in vse variance po skupinah so enake (*homoskedastičnost*).

Test v resnici testira enakost *povprečij* na vseh skupinah – to je prava ničelna hipoteza. Alternativna hipoteza trdi nasprotno – da sta vsaj dve povprečji med seboj različni.

Podobno lahko test uporabimo tudi, če so opažene vrednosti dobljene iz izvedb več slučajnih poskusov, za katere privzamemo, da so vse med seboj neodvisne. Privzamemo, da je slučajna spremenljivka pri vsakem poskusu porazdeljena normalno in da je varianca za vse poskuse enaka. Števila izvedb posameznih poskusov so lahko različna, morata pa biti predpisana. Testira se ničelna hipoteza, da je pričakovana vrednost slučajne spremenljivke pri vseh poskusih enaka.

Primer: primerjava rezultatov kolokvijev med prej omenjenimi tremi skupinami študentov:

$$F \doteq 3{\cdot}162\,.$$

Če testiramo pri stopnji značilnosti $\alpha = 0{\cdot}05$, to primerjamo z $F_{0.95}(2, 119) \doteq 3{\cdot}072$ in dobimo, da je povezava med rezultatom in predmetom, ki ga je študent delal, statistično značilna. Pri tem se pretvarjamo, da gre za enostavni slučajni vzorec.

3.5 Povezanost dveh urejenostnih spremenljivk: Spearmanova koreliranost

Povezanost dveh urejenostnih spremenljivk merimo s *Spearmanovim*[7], *Kendallovim*[8] ali *Goodman*[9]*–Kruskalovim*[10] *korelacijskim koeficientom*. Najenostavnejši za računanje je prvi, a druga dva imata lepše statistične lastnosti – sta zanesljivejša za statistično sklepanje. Goodman–Kruskalov koeficient je še posebej primeren za podatke iz kontingenčnih tabel, ko je veliko vezi. A zaradi enostavnosti se bomo tu posvetili le Spearmanovemu koeficientu.

Statistična obravnava urejenostnih spremenljivk često poteka tako, da iz konstruiramo intervalske spremenljivke in nato na njih uporabimo znane metode. Načini so bolj ali manj uspešni in matematično upravičeni. Eden od načinov, ki je v veliko situacijah matematično dobro utemeljen, je, da je ustrezna intervalska spremenljivka kar *rang*. Tako dobimo tudi Spearmanov koeficient: to je Pearsonov koeficient, izračunan za vezane range. Če z

[7]Charles Edward Spearman (1863–1945), angleški psiholog
[8]Sir Maurice George Kendall (1907–1983), angleški statistik
[9]Leo A. Goodman (1928), ameriški statistik
[10]William Henry Kruskal (1919–2005), ameriški matematik in statistik

$R_1^{(X)}, R_2^{(X)}, \ldots, R_n^{(X)}$ označimo vezane range spremenljivke X, z $R_1^{(Y)}, R_2^{(Y)}, \ldots, R_n^{(Y)}$ pa vezane range spremenljivke Y po enotah, se kovarianca rangov izraža s formulo:

$$K_{R^{(X)},R^{(Y)}} = \\ = \frac{\big(R_1^{(X)} - \bar{R}\big)\big(R_1^{(Y)} - \bar{R}\big) + \big(R_2^{(X)} - \bar{R}\big)\big(R_2^{(Y)} - \bar{R}\big) + \cdots + \big(R_n^{(X)} - \bar{R}\big)\big(R_n^{(Y)} - \bar{R}\big)}{n},$$

kjer je:

$$\bar{R} = \frac{n+1}{2}$$

povprečni rang (ker je le-ta celo število ali pa celo število in pol, u-metoda tu ni toliko smiselna). Nato izračunamo še standardna odklona rangov:

$$s_{R^{(X)}} = \sqrt{\frac{\big(R_1^{(X)} - \bar{R}\big)^2 + \big(R_2^{(X)} - \bar{R}\big)^2 + \cdots + \big(R_n^{(X)} - \bar{R}\big)^2}{n}},$$

$$s_{R^{(Y)}} = \sqrt{\frac{\big(R_1^{(Y)} - \bar{R}\big)^2 + \big(R_2^{(Y)} - \bar{R}\big)^2 + \cdots + \big(R_n^{(Y)} - \bar{R}\big)^2}{n}},$$

Če ni vezi, velja kar:

$$s_{R^{(X)}} = s_{R^{(Y)}} = \sqrt{\frac{n^2-1}{12}},$$

sicer pa sta standardna odklona manjša. Spearmanov korelacijski koeficient definiramo po formuli:

$$\rho = \rho_{X,Y} := \frac{K_{R^{(X)},R^{(Y)}}}{s_{R^{(X)}}\, s_{R^{(Y)}}}.$$

Primer: želimo izmeriti povezavo med zadovoljstvom s telesno težo in subjektivnim vplivom medijev na samopodobo. Za ta namen izvedemo anketo z dvema vprašanjema, pri katerih imamo naslednje izbire:

1. Ali ste zadovoljni s svojo težo?
   (a) Da.
   (b) Srednje.
   (c) Ne.
2. V kolikšni meri mediji vplivajo na vašo samopodobo?
   (a) Sploh ne vplivajo.
   (b) Srednje vplivajo.
   (c) Močno vplivajo.

Obe spremenljivki (zadovoljstvo s telesno težo in vpliv medijev) sta tako urejenostni. Dogovorimo se za naslednjo smer urejenosti: pri zadovoljstvu s telesno težo postavimo:

$$\text{da} < \text{srednje} < \text{ne}\,,$$

(torej v resnici gledamo nezadovoljstvo), vpliv medijev pa uredimo takole:

$$\text{nič} < \text{srednje} < \text{močno}\,.$$

Denimo, da povprašamo štiri študente in dobimo naslednje odgovore:

zadovoljen/a s težo	srednje	srednje	ne	da
mediji vplivajo	srednje	nič	močno	nič

Pri zadovoljstvu s težo ima odgovor 'da' rang 1, odgovor 'srednje' rang 2·5, odgovor 'ne' pa rang 4. Pri vplivu medijev pa ima odgovor 'nič' rang 1·5, odgovor 'srednje' rang 3, odgovor 'močno' pa rang 4. Torej bo:

$$\begin{aligned} &R_1^{(X)} = 2{\cdot}5\,, && R_2^{(X)} = 2{\cdot}5\,, && R_3^{(X)} = 4\,, && R_4^{(X)} = 1\,,\\ &R_1^{(Y)} = 3\,, && R_2^{(X)} = 1{\cdot}5\,, && R_3^{(X)} = 4\,, && R_4^{(X)} = 1{\cdot}5\,. \end{aligned}$$

Povprečni rang je enak $\frac{4+1}{2} = 2{\cdot}5$. Kovarianca rangov:

$$\begin{aligned} K_{R^{(X)},R^{(Y)}} = \frac{1}{4}\Big[&(2{\cdot}5-2{\cdot}5)(3-2{\cdot}5)+(2{\cdot}5-2{\cdot}5)(1{\cdot}5-2{\cdot}5)+\\ &+(4-2{\cdot}5)(4-2{\cdot}5)+(1-2{\cdot}5)(1{\cdot}5-2{\cdot}5)\Big] =\\ &= 0{\cdot}9375\,. \end{aligned}$$

Standardna odklona rangov:

$$\begin{aligned} s_{R^{(X)}} &= \sqrt{\frac{(2{\cdot}5-2{\cdot}5)^2+(2{\cdot}5-2{\cdot}5)^2+(4-2{\cdot}5)^2+(1-2{\cdot}5)^2}{4}} \doteq 1{\cdot}0607\,,\\ s_{R^{(Y)}} &= \sqrt{\frac{(3-2{\cdot}5)^2+(1{\cdot}5-2{\cdot}5)^2+(4-2{\cdot}5)^2+(1{\cdot}5-2{\cdot}5)^2}{4}} \doteq 1{\cdot}0607\,. \end{aligned}$$

sta le malo manjša od maksimalne vrednosti $\sqrt{(4^2-1)/12} \doteq 1{\cdot}118$. Spearmanov korelacijski koeficient pride:

$$\rho \doteq \frac{0{\cdot}9375}{1{\cdot}0607\cdot 1{\cdot}0607} \doteq 0{\cdot}833\,.$$

in je pozitiven, kar pomeni, da ljudje, ki mislijo, da imajo mediji večji vpliv na njihovo samopodobo, nagibajo k večjemu nezadovoljstvu s telesno težo in obratno. To je tudi neposredno razvidno iz podatkov.

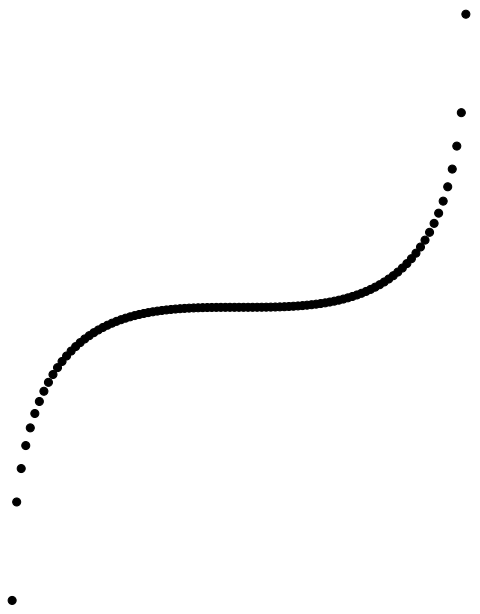
Spearmanov korelacijski koeficient je primerljiv s Pearsonovim in ga tudi enako kvalitativno opredeljujemo. Pri primeru s štirimi študenti je bila torej povezanost *visoka*.

Če sta obe spremenljivki dihotomni, se Spearmanov in Pearsonov koeficient ujemata. Nasploh ima Spearmanov korelacijski koeficient podobne lastnosti kot Pearsonov:

- Definiran je, če nobena od spremenljivk ni konstantna.
- Velja $-1 \leq \rho \leq 1$.
- Če sta X in Y neodvisni (na statistični množici, iz katere so podatki), je $\rho = 0$. Velja tudi, da, če podatki temeljijo na velikem enostavnem slučajnem vzorcu iz velike populacije, na kateri sta X in Y neodvisni, je ρ blizu 0 (malo kasneje pri testiranju se bomo naučili, kako postaviti mejo).
- Spearmanov korelacijski koeficient je maksimalen (enak 1), če je katera koli od spremenljivk strogo naraščajoča (a ne nujno linearna) funkcija druge.
- Spearmanov korelacijski koeficient je minimalen (enak -1), če je katera koli od spremenljivk strogo padajoča (a ne nujno linearna) funkcija druge.

Spearmanov korelacijski koeficient torej meri stopnjo *monotone* povezanosti. Podobne lastnosti ima tudi Kendallov korelacijski koeficient (τ).

Primer: pri naslednjih podatkih:



večja vrednost koordinate x pomeni tudi večjo vrednost koordinate y, zato je $\rho = 1$. Povezava med x in y je deterministična, ni pa linearna, zato $r \neq 1$: pride $r \doteq 0{\cdot}792$. Koreliranost po Pearsonu je torej zgolj visoka, niti ne zelo visoka.

Spearmanov korelacijski koeficient je preprosto izračunati tudi za podatke iz kontingenčne tabele. Če z $R^{(x)}(a)$ označimo vezani rang vrednosti a glede na spremenljivko X, z $R^{(y)}(b)$ označimo vezani rang vrednosti b glede na spremenljivko Y, velja:

$$K_{R^{(x)},R^{(y)}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}\big(R^{(x)}(a_i)-\bar{R}\big)\big(R^{(y)}(b_j)-\bar{R}\big)\,,$$

$$s_{R^{(x)}} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} f_{i\cdot}\big(R^{(x)}(a_i)-\bar{R}\big)^2}\,,$$

$$s_{R^{(y)}} = \sqrt{\frac{1}{n}\sum_{j=1}^{l} f_{\cdot j}\big(R^{(y)}(b_j)-\bar{R}\big)^2}\,.$$

Primer: vrnimo se k povezavi med zadovoljstvom s telesno težo in subjektivnim vplivom medijev na samopodobo. Zdaj povprašamo 20 študentov in rezultate zberemo v naslednji kontingenčni tabeli:

Zad. s težo\Vpliv medijev	ga ni	srednji	močan	Skupaj
da	8	1	0	9
srednje	2	1	0	3
ne	2	4	2	8
Skupaj	12	6	2	20

Povprečni rang: $\bar{R} = 10{\cdot}5$.

Rangi posameznih odgovorov in njihovi odmiki od povprečnega:

a_i	$R^{(x)}(a_i)$	$R^{(x)}(a_i)-\bar{R}$
da	5	$-5{\cdot}5$
srednje	11	$0{\cdot}5$
ne	$16{\cdot}5$	6

b_j	$R^{(y)}(b_j)$	$R^{(y)}(b_j)-\bar{R}$
ga ni	$6{\cdot}5$	-4
srednji	$15{\cdot}5$	5
močan	$19{\cdot}5$	9

Standardna odklona rangov:

$$s_{R^{(x)}} = \sqrt{\frac{1}{20}\Big(9\cdot(-5{\cdot}5)^2+3\cdot 0{\cdot}5^2+8\cdot 6^2\Big)} = 5{\cdot}296\,,$$

$$s_{R^{(y)}} = \sqrt{\frac{1}{20}\Big(12\cdot(-4)^2+6\cdot 5^2+2\cdot 9^2\Big)} = 5{\cdot}020\,.$$

Kovarianca rangov:

$$\begin{aligned} K_{R^{(x)},R^{(y)}} = \frac{1}{20}\Big[&8\cdot(-5{\cdot}5)\cdot(-4)+1\cdot(-5{\cdot}5)\cdot 5+0\cdot(-5{\cdot}5)\cdot 9+\\ &+2\cdot 0{\cdot}5\cdot(-4)+1\cdot 0{\cdot}5\cdot 5+0\cdot 0{\cdot}5\cdot 9+\\ &+2\cdot 6\cdot(-4)+4\cdot 6\cdot 5+2\cdot 6\cdot 9\Big] = \\ = 16{\cdot}35\,. \end{aligned}$$

Spearmanov korelacijski koeficient:

$$\rho \doteq \frac{16{\cdot}35}{5{\cdot}296\cdot 5{\cdot}020} \doteq 0{\cdot}615\,.$$

Tokrat torej dobimo *zmerno* povezanost, a v isto smer kot prej pri štirih študentih.

Spearmanovo koreliranost testiramo tako kot Pearsonovo, s T-testom na temelju testne statistike:

$$T = \frac{\rho}{\sqrt{1-\rho^2}}\sqrt{n-2}\,.$$

Ničelno hipotezo zavrnemo:

- proti $H_1^{\pm}$, da sta X in Y korelirani, če je $|T| > t_{1-\alpha/2}(n-2)$;
- proti H_1^{+}, da sta X in Y pozitivno korelirani, če je $T > t_{1-\alpha}(n-2)$;
- proti H_1^{-}, da sta X in Y negativno korelirani, če je $T < -t_{1-\alpha}(n-2)$.

Testiranje Spearmanove koreliranosti je dobra alternativa testiranju Pearsonove koreliranosti, če sumimo, da porazdelitev katere od spremenljivk ni normalna, saj je test Pearsonove koreliranosti zasnovan na predpostavki normalnosti.

Kot zgled testirajmo pri primeru z 20 študenti hipotezo, da nezadovoljstvo s telesno težo in vpliv medijev na samopodobo nista povezana, proti alternativni hipotezi, da se ljudje, ki mislijo, da imajo mediji večji vpliv na njihovo samopodobo, nagibajo k večjemu nezadovoljstvu s telesno težo in obratno. Postavimo $\alpha = 0{\cdot}01$. Testna statistika pride $T \doteq 3{\cdot}31$, kar primerjamo s $t_{0.99}(18) \doteq 2{\cdot}552$. Odstopanja so torej statistično zelo značilna. Z drugimi besedami, ljudje, ki mislijo, da imajo mediji večji vpliv na njihovo samopodobo, se statistično zelo značilno nagibajo k večjemu nezadovoljstvu s telesno težo (in obratno).

3.6 Povezanost urejenostne in dihotomne spremenljivke

Povezanost urejenostne in dihotomne spremenljivke bi se dalo meriti s Spearmanovim korelacijskim koeficientom (izbrani vrstni red vrednosti dihotomne spremenljivke vpliva le na predznak). Vendar pa se navadno uporablja *rangovni biserialni koeficient* – glej [18, 22].

Za izračun tega koeficienta potrebujemo ranžirno vrsto vseh vrednosti prve (urejenostne) spremenljivke. Označimo z $R(a)$ vezani rang vrednosti a glede na to ranžirno vrsto. Nato podatke razdelimo glede na vrednosti druge (dihotomne spremenljivke), nastaneta dve skupini. Naj bodo:

$$x_1, x_2, \ldots, x_m$$

vrednosti urejenostne spremenljivke v prvi skupini,

$$y_1, y_2, \ldots, y_n$$

pa vrednosti urejenostne spremenljivke v drugi skupini. Tedaj lahko izračunamo povprečna ranga obeh skupin:

$$\bar{R}_X = \frac{R(x_1) + R(x_2) + \cdots + R(x_m)}{m}\,, \qquad \bar{R}_Y = \frac{R(y_1) + R(y_2) + \cdots + R(y_n)}{n}\,.$$

Rangovni biserialni koeficient je definiran kot:

$$r_{\mathrm{rb}} = \frac{2(\bar{R}_X - \bar{R}_Y)}{m+n} = \frac{2\bar{R}_X - (m+n+1)}{n} = \frac{m+n+1-2\bar{R}_Y}{m}\,.$$

To je modifikacija točkovnega biserialnega koeficienta, izračunanega na rangih. Lastnosti koeficienta:

- Definiran je vedno.
- Velja $-1 \leq r_{\mathrm{rb}} \leq 1$.
- Če sta vrednost urejenostne spremenljivke in skupina neodvisni (na statistični množici, iz katere so podatki), je $r_{\mathrm{rb}} = 0$. Velja tudi, da, če podatki temeljijo na velikem enostavnem slučajnem vzorcu iz velike populacije, na kateri sta X in Y neodvisni, je r_{rb} blizu 0 (malo kasneje pri testiranju se bomo naučili, kako postaviti mejo).
- Koeficient r_{rb} je minimalen (enak -1), če so vse vrednosti iz prve skupine $(x_1, \ldots, x_m)$ strogo manjše od vseh vrednosti iz druge skupine $(y_1, \ldots, y_n)$.
- Koeficient r_{rb} je maksimalen (enak 1), če so vse vrednosti iz prve skupine $(x_1, \ldots, x_m)$ strogo večje od vseh vrednosti iz druge skupine $(y_1, \ldots, y_n)$.

Rangovni biserialni koeficient kvalitativno opredeljujemo enako kot vse ostale podobne koeficiente (točkovni biserialni korelacijski, Pearsonov, Spearmanov).

Primer: Med 17 študenti so izvedli anketo z naslednjima vprašanjema:

1. Ocenite stopnjo stresa pri vas v zadnjih dveh tednih.
   (zelo majhna/majhna/srednja/velika/zelo velika)
2. Ali ste se v zadnjih dveh tednih posvečali študiju bolj kot ponavadi?
   (da/ne)

Rezultati ankete so naslednji:

st. stresa\študij	da	ne	Skupaj	Kumulativno	Vezani rang
zelo majhna	0	0	0	0	–
majhna	2	5	7	7	4
srednja	1	2	3	10	9
velika	5	0	5	15	13
zelo velika	2	0	2	17	16·5
Skupaj	10	7	17		

Povprečna ranga sta enaka:

$$\bar{R}_{\mathrm{da}} = \frac{2\cdot 4 + 1\cdot 9 + 5\cdot 13 + 2\cdot 16{\cdot}5}{10} = 11{\cdot}5\,, \qquad \bar{R}_{\mathrm{ne}} = \frac{5\cdot 4 + 2\cdot 9}{7} \doteq 5{\cdot}429\,,$$

rangovni biserialni koeficient pa je enak:

$$r_{\mathrm{rb}} \doteq \frac{2(11{\cdot}5 - 5{\cdot}429)}{17} = \frac{2\cdot 11{\cdot}5 - 18}{7} \doteq \frac{18 - 2\cdot 5{\cdot}429}{10} \doteq 0{\cdot}714\,.$$

Povezanost je torej visoka. Z drugimi besedami, študenti, ki so se posvečali študiju, so bili torej *precej bolj* pod stresom od tistih, ki se študiju niso posvečali.

Povezavo med skupino in vrednostjo spremenljivke testiramo z *inverzijskim testom*, ki mu pravimo tudi *Wilcoxon*[11]*–Mann*[12]*–Whitneyjev*[13] *test*. Za opažene enote, na podlagi katerih opravimo test, predpostavimo podobno kot pri testiranju enakosti povprečij: lahko privzamemo, da so dobljene kot enostavni slučajni vzorec, pri čemer mora biti iz vsake podpopulacije vsaj ena enota; če se to ne zgodi (kar je pri običajnih velikostih vzorcev in običajnih deležih podpopulacij zelo malo verjetno), vzorčenje ponovimo. Lahko pa tudi predpišemo minimalno ali točno število enot iz posamezne podpopulacije. V slednjem primeru iz vsake vzamemo enostavni slučajni vzorec, vzorca pa morata biti med seboj neodvisna. Na populaciji torej izvedemo preprosto stratificirano vzorčenje.

Podobno lahko test uporabimo tudi, če so opažene vrednosti dobljene iz izvedb dveh slučajnih poskusov, za katere privzamemo, da so vse med seboj neodvisne. Števili izvedb posameznega poskusa sta lahko različni.

Testiramo ničelno hipotezo, da med skupino in spremenljivko ni povezave, ali natančneje, da je spremenljivka na obeh skupinah oz. pri obeh poskusih enako porazdeljena.

Za formulacijo alternativne hipoteze pa moramo razumeti stohastično primerjavo porazdelitev. Podobno kot pri testu z znaki je ideja, da je X *stohastično večja* od Y, če vplivi, ki večajo X na račun Y, prevladujejo nad vplivi, ki delujejo obratno. Podobno bi bila X *stohastično manjša* od Y, če vplivi, ki večajo Y na račun X, prevladujejo nad

[11] Frank Wilcoxon (1892–1965), ameriški kemik in statistik

[12] Henry Berthold Mann, rojen kot Heinrich Mann (1905–2000), avstrijski matematik judovskega rodu, deloval v ZDA

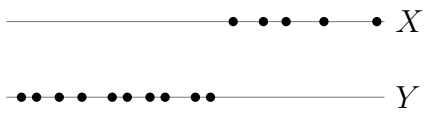
[13] Donald Ransom Whitney (1915–2007), ameriški statistik

vplivi, ki delujejo obratno. A to je treba zdaj formulirati za primer, ko sta spremenljivki X in Y definirani na *različnih* statističnih množicah.

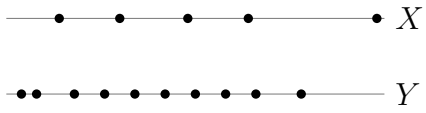
Tako npr. pri študentih iz zgornjega primera lahko dejstvo, da so tisti, ki intenzivneje študirajo, bolj pod stresom, izrazimo tudi tako, da je stopnja stresa pri tistih, ki študirajo intenzivneje, stohastično večja od stopnje stresa pri tistih, ki študirajo enako intenzivno.

Precizna definicija je naslednja: X je stohastično večja od Y, če je za vsako fiksno vrednost u delež enot, za katere je $X \geq u$, večji ali enak deležu enot, za katere je $Y \geq u$. Delež vselej gledamo v okviru statistične množice, na kateri je definirana *posamezna* spremenljivka. Nadalje je X stohastično *strogo* večja od Y, če je stohastično večja in če obstaja tudi tak u, da je delež enot, za katere je $X \geq u$, strogo večji od deleža enot, za katere je $Y \geq u$. Slučajna spremenljivka X je stohastično (strogo) manjša od Y, če je Y stohastično (strogo) manjša od X.

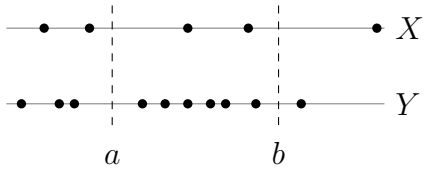
Primer: če so vse vrednosti X strogo večje od vseh vrednosti Y, je X stohastično strogo večja od Y:



Primer: tudi pri teh podatkih je X stohastično strogo večja od Y.



Primer: podatki, kjer niti X ni niti stohastično večja niti stohastično manjša od Y.



Delež enot, kjer je $X \geq a$, $0{\cdot}6$, je strogo manjši od deleža enot, kjer je $Y \geq a$, $0{\cdot}7$.

Delež enot, kjer je $X \geq b$, $0{\cdot}2$, je strogo večji od deleža enot, kjer je $Y \geq b$, $0{\cdot}1$.

Zdaj ko smo pojasnili stohastično primerjavo spremenljivk, lahko formuliramo alternativne hipoteze. Spet bomo gledali tri različice:

- Enostranska alternativna hipoteza H_1^X v korist prve skupine trdi, da je porazdelitev na prvi skupini stohastično strogo večja od porazdelitve na drugi skupini.
- Enostranska alternativna hipoteza H_1^Y v korist druge skupine trdi, da je porazdelitev na drugi skupini stohastično strogo večja od porazdelitve na prvi skupini.

- Dvostranska alternativna hipoteza $H_1^{\pm}$ trdi, da velja ena izmed prej omenjenih enostranskih hipotez.

Za dovolj velike vzorce je inverzijski test lahko kar Z-test na testni statistiki:

$$Z := r_{\mathrm{rb}} \sqrt{\frac{3mn}{m+n+1}}\,,$$

kar pomeni, da ničelno hipotezo zavrnemo:

- proti $H_1^{\pm}$, če je $|Z| > z_{1-\alpha/2}$;
- proti H_1^X, če je $Z > z_{1-\alpha}$;
- proti H_1^Y, če je $Z < -z_{1-\alpha}$.

Ta test žal ni eksakten: minimalni pogoj za njegovo legitimnost, da je v vsaki skupini vsaj 5 enot, torej $m, n \geq 5$.

Inverzijski test je dobra alternativa T-testu, če sumimo, da porazdelitev katere od spremenljivk zelo odstopa od normalne, saj je T-test zasnovan na predpostavki normalnosti (čeprav je do neke mere robusten). Predvsem se inverzijski test bolje obnese, če je veliko skrajnih vrednosti.

Primer. Na podlagi ankete iz prejšnjega primera bi želeli testirati ničelno hipotezo, da posvečanje študiju ne vpliva na stopnjo stresa, proti alternativni hipotezi, da so študenti, ki se posvečajo študiju, bolj pod stresom od tistih, ki se ne. Dobimo:

$$Z \doteq 0{\cdot}714 \sqrt{\frac{3 \cdot 70}{18}} \doteq 2{\cdot}44\,,$$

kar je večje od $z_{0.99} \doteq 2{\cdot}326$, torej ničelno hipotezo zavrnemo tudi pri $\alpha = 0{\cdot}01$. Z drugimi besedami, študenti, ki so se posvečali študiju, so bili statistično zelo značilno bolj pod stresom od tistih, ki se študiju niso posvečali.

3.7 Povezanost urejenostne in imenske spremenljivke: Kruskal–Wallisova analiza variance

Povezanost urejenostne in imenske spremenljivke merimo s *Kruskal*[14]*–Wallisovim*[15] *deležem pojasnjene variance*. V skladu s splošno filozofijo obravnave urejenostnih spremenljivk je to delež pojasnjene variance za vezane range. Gre torej za vrsto analize variance.

[14] William Henry Kruskal (1919–2005), ameriški matematik in statistik
[15] Wilson Allen Wallis (1912–1998), ameriški ekonomist in statistik

Če torej imenska spremenljivka G zavzame vrednosti $g_1, g_2, \ldots, g_k$, lahko range urejenostne spremenljivke indeksiramo takole:

$$\begin{aligned} R_{11}, R_{12}, \ldots, R_{1n_1} &: \text{rangi na enotah, kjer je } G = g_1 \\ R_{21}, R_{22}, \ldots, R_{2n_2} &: \text{rangi na enotah, kjer je } G = g_2 \\ &\vdots \\ R_{k1}, R_{k2}, \ldots, R_{kn_2} &: \text{rangi na enotah, kjer je } G = g_k \end{aligned}$$

Seveda velja $n_1 + n_2 + \cdots + n_k = n$.

Spet je povprečni rang vedno enak $\bar{R} = \dfrac{n+1}{2}$, skupna varianca pa je enaka:

$$s^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (R_{ij} - \bar{R})^2$$

in če ni vezi, je $s^2 = \dfrac{n^2 - 1}{12}$. Če zdaj z $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_k$ označimo povprečne range na posameznih skupinah:

$$\bar{R}_i = \frac{R_{i1} + R_{i2} + \cdots + R_{in_i}}{n_i}\,,$$

je pojasnjena varianca rangov enaka:

$$s_{\mathrm{B}}^2 = \frac{n_1}{n} (\bar{R}_1 - \bar{R})^2 + \frac{n_2}{n} (\bar{R}_2 - \bar{R})^2 + \cdots + \frac{n_k}{n} (\bar{R}_k - \bar{R})^2 .$$

Kruskal–Wallisov delež pojasnjene variance (moč učinka) pa je enak:

$$\eta_{\mathrm{KW}}^2 = \frac{s_{\mathrm{B}}^2}{s^2}\,,$$

Doseže lahko vrednosti od 0 do 1. Vrednost 0 je dosežena takrat, ko so povprečni rangi na vseh skupinah enaki, vrednost 1 pa je dosežena takrat, ko so vse vrednosti na posamezni skupini enake. Kvalitativno ga interpretiramo enako kot običajni delež pojasnjene variance.

Primer: želimo izmeriti povezavo med počutjem in barvo zgornjega dela oblačila. Za ta namen vzamemo 20 anketirancev in:

- Jih povprašamo, kako se počutijo, pri čemer jim damo na voljo 5-stopenjsko lestvico. To je urejenostna spremenljivka.
- Si ogledamo barvo njihovega zgornjega oblačila. Barve razdelimo v štiri kategorije: temne (črna, rjava, siva, temno modra, vijoličasta), bela, svetle (rumena, rdeča, roza, oranžna, zelena, svetlo modra), pisane. To je imenska spremenljivka.

Rezultati ankete:

počutje\barva	temna	bela	svetla	pisana	Skupaj	Kum.	Rang
zelo slabo	0	0	0	0	0	0	–
slabo	2	0	1	0	3	3	2
nevtralno	3	2	1	4	10	13	8·5
kar dobro	4	1	0	1	6	19	16·5
odlično	1	0	0	0	1	20	20
Skupaj	10	3	2	5	20		
Povprečni rang	11·55	11·17	5·25	10·10	10·5		

Skupni povprečni rang: $\dfrac{20+1}{2} = 10{\cdot}5.$

Povprečni rangi po skupinah v zadnji vrstici tabele so dobljeni na naslednji način:

$$\begin{aligned}
\bar{R}_1 &= \frac{1}{10}\Big[2\cdot 2 + 3\cdot 8{\cdot}5 + 4\cdot 16{\cdot}5 + 1\cdot 20\Big] \doteq 11{\cdot}55\,,\\
\bar{R}_2 &= \frac{1}{3}\Big[2\cdot 8{\cdot}5 + 1\cdot 16{\cdot}5\Big] \doteq 11{\cdot}17\,,\\
\bar{R}_3 &= \frac{1}{2}\Big[1\cdot 2 + 1\cdot 8{\cdot}5\Big] \doteq 5{\cdot}25\,,\\
\bar{R}_4 &= \frac{1}{5}\Big[4\cdot 8{\cdot}5 + 1\cdot 16{\cdot}5\Big] \doteq 10{\cdot}1\,.
\end{aligned}$$

Skupna varianca ranga:

$$s^2 = \frac{1}{20}\Big[3\cdot(2-10{\cdot}5)^2 + 10\cdot(8{\cdot}5-10{\cdot}5)^2 + 6\cdot(16{\cdot}5-10{\cdot}5)^2 + 1\cdot(20-10{\cdot}5)^2\Big] = 28{\cdot}15\,.$$

Varianca ranga med skupinami (pojasnjena varianca):

$$s_{\mathrm{B}}^2 \doteq \frac{1}{20}\Big[10\cdot(11{\cdot}55-10{\cdot}5)^2+3\cdot(11{\cdot}17-10{\cdot}5)^2+2\cdot(5{\cdot}25-10{\cdot}5)^2+5\cdot(10{\cdot}10-10{\cdot}5)^2\Big] \doteq 3{\cdot}414\,.$$

Kruskal–Wallisov delež pojasnjene variance:

$$\eta^2_{\mathrm{KW}} \doteq \frac{3{\cdot}414}{28{\cdot}15} \doteq 0{\cdot}121\,.$$

Gre torej za *rahlo* povezanost.

Pojasnjeno varianco rangov približno testiramo s testom hi kvadrat na testni statistiki:

$$K = \frac{12}{n+1}\,s_{\mathrm{B}}^2\,,$$

in sicer s $k-1$ prostostnimi stopnjami: ničelno hipotezo zavrnemo, če je $K > \chi^2_{1-\alpha}(k-1)$.

Če se opažene vrednosti nanašajo na vzorec iz določene populacije, sta osnovni predpostavki testa spet:

- Na vsaki skupini, t. j. za vsako vrednost imenske spremenljivke, vzamemo enostavni slučajni vzorec predpisane velikosti – torej stratificiramo glede na imensko spremenljivko.
- Vzorci sta med seboj neodvisna.

Podobno lahko test uporabimo tudi, če so opažene vrednosti dobljene iz izvedb več slučajnih poskusov, za katere privzamemo, da so vse med seboj neodvisne. Števila izvedb posameznih poskusov so lahko različni, morajo pa biti predpisana.

Žal pa test ni eksakten, niti če sta zgornji predpostavki izpolnjeni. Minimalni pogoj za njegovo legitimnost je, da je v vsaki skupini vsaj 5 enot.

Pri prejšnjem primeru slednji pogoj ni izpolnjen, ni pa izpolnjen tudi pogoj o predpisanem številu enot v posamezni skupini oz. stratifikaciji. Če to odmislimo, si lahko pomagamo tako, da združimo bela in svetla oblačila. Dobimo:

počutje\barva	temna	svetla	pisana	Skupaj	Kum.	Rang
zelo slabo	0	0	0	0	0	–
slabo	2	1	0	3	3	2
nevtralno	3	3	4	10	13	8·5
kar dobro	4	1	1	6	19	16·5
odlično	1	0	0	1	20	20
Skupaj	10	5	5	20		
Povprečni rang	11·55	8·80	10·10	10·5		

Skupna varianca ranga se ne spremeni, nova pojasnjena varianca pa je:

$$s_{\mathrm{B}}^2 \doteq \frac{1}{20}\Big[10 \cdot (11{\cdot}55 - 10{\cdot}5)^2 + 5 \cdot (8{\cdot}80 - 10{\cdot}5)^2 + 5 \cdot (10{\cdot}10 - 10{\cdot}5)^2\Big] \doteq 1{\cdot}314\,.$$

Testna statistika pride:

$$K \doteq \frac{12}{21} \cdot 1{\cdot}314 \doteq 0{\cdot}75\,.$$

Pri stopnji značilnosti $\alpha = 0{\cdot}05$ moramo to primerjati s $\chi^2_{0.95}(3) \doteq 7{\cdot}815$, torej povezava med počutjem in barvo zgornjega oblačila ni bila statistično značilna.

3.8 Povzetek bivariatne analize

	dihotomna	imenska	urejenostna	intervalska
dihotomna	Cramérjev V, test hi kvadrat			
imenska	Cramérjev V, test hi kvadrat	Cramérjev V, test hi kvadrat		
urejenostna	r_{rb}, inverzijski test	Kruskal–Wallis, test hi hvadrat	Spearmanov ρ, T-test	
intervalska	r_{pb}, T-test	ANOVA, F-test	Spearmanov ρ, T-test	Pearsonov r, T-test

Tabele

V tem dodatku so prikazane tabele porazdelitev, ki jih najpogosteje srečamo v statistiki. Vse vrednosti so bile izračunane s programom R.

TABELA 1:
GAUSSOV VERJETNOSTNI INTEGRAL

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\,\mathrm{d}t\,; \qquad \Phi(-x) = -\Phi(x)\,; \qquad x \text{ velik} \Longrightarrow \frac{1}{2} - \Phi(x) \sim \frac{1}{x\sqrt{2\pi}}\, e^{-x^2/2}$$

	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

TABELA 2:
KVANTILI STUDENTOVE PORAZDELITVE

$T \sim \text{Student}(df)\colon\ P(T < t_p(df)) = p\,; \quad t_p(\infty) = z_p\,; \quad df \text{ velik} \Longrightarrow t_p(df) \approx z_p + \dfrac{z_p^3 + z_p}{4\,df}$

$df \backslash p$	0.9	0.95	0.975	0.99	0.995
1	3.078	6.314	12.71	31.82	63.66
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
32	1.309	1.694	2.037	2.449	2.738
34	1.307	1.691	2.032	2.441	2.728
36	1.306	1.688	2.028	2.434	2.719
38	1.304	1.686	2.024	2.429	2.712
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
80	1.292	1.664	1.990	2.374	2.639
100	1.290	1.660	1.984	2.364	2.626
200	1.286	1.653	1.972	2.345	2.601
∞	1.282	1.645	1.960	2.326	2.576

TABELA 3:
KVANTILI PORAZDELITVE HI KVADRAT

$$\chi^2 \sim \chi^2(df)\colon\ P(\chi^2 < \chi^2_p(df)) = p$$

$$df \text{ velik} \Longrightarrow \chi^2_p \approx df\left(1 - \frac{2}{9\,df} + \frac{\sqrt{2}}{3\sqrt{df}}\,z_p\right)^3 = df\left(1 - \frac{2}{9\,df} - \frac{\sqrt{2}}{3\sqrt{df}}\,z_{1-p}\right)^3$$

$df\backslash p$	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.00003927	0.0001571	0.0009821	0.003932	0.01579	2.706	3.841	5.024	6.635	7.879
2	0.01003	0.02010	0.05064	0.1026	0.2107	4.605	5.991	7.378	9.210	10.60
3	0.07172	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.34	12.84
4	0.2070	0.2971	0.4844	0.7107	1.064	7.779	9.488	11.14	13.28	14.86
5	0.4117	0.5543	0.8312	1.145	1.610	9.236	11.07	12.83	15.09	16.75
6	0.6757	0.8721	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55
7	0.9893	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28
8	1.344	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	21.95
9	1.735	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21	25.19
11	2.603	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72	26.76
12	3.074	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22	28.30
13	3.565	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69	29.82
14	4.075	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14	31.32
15	4.601	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58	32.80
16	5.142	5.812	6.908	7.962	9.312	23.54	26.30	28.85	32.00	34.27
17	5.697	6.408	7.564	8.672	10.09	24.77	27.59	30.19	33.41	35.72
18	6.265	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.81	37.16
19	6.844	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.434	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.034	8.897	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.643	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.260	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.886	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

TABELA 4: KVANTILI FISHER–SNEDECORJEVE PORAZDELITVE

$X \sim \mathrm{F}(m,n)\colon\ P(X < F_p(m,n)) = p$ $\qquad$ $F_p(m,n) = 1/F_{1-p}(n,m)$

$p = 0.9$

$m\backslash n$	1	2	3	4	5	6	7	8	9	10	11	12
1	39.86	8.526	5.538	4.545	4.06	3.776	3.589	3.458	3.36	3.285	3.225	3.177
2	49.5	9	5.462	4.325	3.78	3.463	3.257	3.113	3.006	2.924	2.86	2.807
3	53.59	9.162	5.391	4.191	3.619	3.289	3.074	2.924	2.813	2.728	2.66	2.606
4	55.83	9.243	5.343	4.107	3.52	3.181	2.961	2.806	2.693	2.605	2.536	2.48
5	57.24	9.293	5.309	4.051	3.453	3.108	2.883	2.726	2.611	2.522	2.451	2.394
6	58.2	9.326	5.285	4.01	3.405	3.055	2.827	2.668	2.551	2.461	2.389	2.331
7	58.91	9.349	5.266	3.979	3.368	3.014	2.785	2.624	2.505	2.414	2.342	2.283
8	59.44	9.367	5.252	3.955	3.339	2.983	2.752	2.589	2.469	2.377	2.304	2.245
9	59.86	9.381	5.24	3.936	3.316	2.958	2.725	2.561	2.44	2.347	2.274	2.214
10	60.19	9.392	5.23	3.92	3.297	2.937	2.703	2.538	2.416	2.323	2.248	2.188
11	60.47	9.401	5.222	3.907	3.282	2.92	2.684	2.519	2.396	2.302	2.227	2.166
12	60.71	9.408	5.216	3.896	3.268	2.905	2.668	2.502	2.379	2.284	2.209	2.147
14	61.07	9.42	5.205	3.878	3.247	2.881	2.643	2.475	2.351	2.255	2.179	2.117
16	61.35	9.429	5.196	3.864	3.23	2.863	2.623	2.455	2.329	2.233	2.156	2.094
20	61.74	9.441	5.184	3.844	3.207	2.836	2.595	2.425	2.298	2.201	2.123	2.06
24	62	9.45	5.176	3.831	3.191	2.818	2.575	2.404	2.277	2.178	2.1	2.036
30	62.26	9.458	5.168	3.817	3.174	2.8	2.555	2.383	2.255	2.155	2.076	2.011
40	62.53	9.466	5.16	3.804	3.157	2.781	2.535	2.361	2.232	2.132	2.052	1.986
50	62.69	9.471	5.155	3.795	3.147	2.77	2.523	2.348	2.218	2.117	2.036	1.97
75	62.9	9.478	5.148	3.784	3.133	2.754	2.506	2.33	2.199	2.097	2.016	1.949
100	63.01	9.481	5.144	3.778	3.126	2.746	2.497	2.321	2.189	2.087	2.005	1.938
200	63.17	9.486	5.139	3.769	3.116	2.734	2.484	2.307	2.174	2.071	1.989	1.921
500	63.26	9.489	5.136	3.764	3.109	2.727	2.476	2.298	2.165	2.062	1.979	1.911
∞	63.33	9.491	5.134	3.761	3.105	2.722	2.471	2.293	2.159	2.055	1.972	1.904

TABELA 4: KVANTILI FISHER–SNEDECORJEVE PORAZDELITVE

$X \sim \mathrm{F}(m,n)\colon\ P(X < F_p(m,n)) = p$ $\qquad$ $F_p(m,n) = 1/F_{1-p}(n,m)$

$p = 0.9$

$m\backslash n$	14	16	20	24	30	40	50	75	100	200	500	∞
1	3.102	3.048	2.975	2.927	2.881	2.835	2.809	2.774	2.756	2.731	2.716	2.706
2	2.726	2.668	2.589	2.538	2.489	2.44	2.412	2.375	2.356	2.329	2.313	2.303
3	2.522	2.462	2.38	2.327	2.276	2.226	2.197	2.158	2.139	2.111	2.095	2.084
4	2.395	2.333	2.249	2.195	2.142	2.091	2.061	2.021	2.002	1.973	1.956	1.945
5	2.307	2.244	2.158	2.103	2.049	1.997	1.966	1.926	1.906	1.876	1.859	1.847
6	2.243	2.178	2.091	2.035	1.98	1.927	1.895	1.854	1.834	1.804	1.786	1.774
7	2.193	2.128	2.04	1.983	1.927	1.873	1.84	1.798	1.778	1.747	1.729	1.717
8	2.154	2.088	1.999	1.941	1.884	1.829	1.796	1.754	1.732	1.701	1.683	1.67
9	2.122	2.055	1.965	1.906	1.849	1.793	1.76	1.716	1.695	1.663	1.644	1.632
10	2.095	2.028	1.937	1.877	1.819	1.763	1.729	1.685	1.663	1.631	1.612	1.599
11	2.073	2.005	1.913	1.853	1.794	1.737	1.703	1.658	1.636	1.603	1.583	1.57
12	2.054	1.985	1.892	1.832	1.773	1.715	1.68	1.635	1.612	1.579	1.559	1.546
14	2.022	1.953	1.859	1.797	1.737	1.678	1.643	1.596	1.573	1.539	1.518	1.505
16	1.998	1.928	1.833	1.77	1.709	1.649	1.613	1.565	1.542	1.507	1.485	1.471
20	1.962	1.891	1.794	1.73	1.667	1.605	1.568	1.519	1.494	1.458	1.435	1.421
24	1.938	1.866	1.767	1.702	1.638	1.574	1.536	1.485	1.46	1.422	1.399	1.383
30	1.912	1.839	1.738	1.672	1.606	1.541	1.502	1.449	1.423	1.383	1.358	1.342
40	1.885	1.811	1.708	1.641	1.573	1.506	1.465	1.41	1.382	1.339	1.313	1.295
50	1.869	1.793	1.69	1.621	1.552	1.483	1.441	1.384	1.355	1.31	1.282	1.263
75	1.846	1.769	1.664	1.593	1.523	1.451	1.407	1.346	1.315	1.266	1.236	1.214
100	1.834	1.757	1.65	1.579	1.507	1.434	1.388	1.326	1.293	1.242	1.209	1.185
200	1.816	1.738	1.629	1.556	1.482	1.406	1.359	1.293	1.257	1.199	1.16	1.13
500	1.805	1.726	1.616	1.542	1.467	1.389	1.34	1.27	1.232	1.168	1.122	1.082
∞	1.797	1.718	1.607	1.533	1.456	1.377	1.327	1.254	1.214	1.144	1.087	1

TABELA 4: KVANTILI FISHER–SNEDECORJEVE PORAZDELITVE

$X \sim \mathrm{F}(m,n)\colon\ P(X < F_p(m,n)) = p$ $\qquad$ $F_p(m,n) = 1/F_{1-p}(n,m)$

$p = 0.95$

$m\backslash n$	1	2	3	4	5	6	7	8	9	10	11	12
1	161.4	18.51	10.13	7.709	6.608	5.987	5.591	5.318	5.117	4.965	4.844	4.747
2	199.5	19	9.552	6.944	5.786	5.143	4.737	4.459	4.256	4.103	3.982	3.885
3	215.7	19.16	9.277	6.591	5.409	4.757	4.347	4.066	3.863	3.708	3.587	3.49
4	224.6	19.25	9.117	6.388	5.192	4.534	4.12	3.838	3.633	3.478	3.357	3.259
5	230.2	19.3	9.013	6.256	5.05	4.387	3.972	3.687	3.482	3.326	3.204	3.106
6	234	19.33	8.941	6.163	4.95	4.284	3.866	3.581	3.374	3.217	3.095	2.996
7	236.8	19.35	8.887	6.094	4.876	4.207	3.787	3.5	3.293	3.135	3.012	2.913
8	238.9	19.37	8.845	6.041	4.818	4.147	3.726	3.438	3.23	3.072	2.948	2.849
9	240.5	19.38	8.812	5.999	4.772	4.099	3.677	3.388	3.179	3.02	2.896	2.796
10	241.9	19.4	8.786	5.964	4.735	4.06	3.637	3.347	3.137	2.978	2.854	2.753
11	243	19.4	8.763	5.936	4.704	4.027	3.603	3.313	3.102	2.943	2.818	2.717
12	243.9	19.41	8.745	5.912	4.678	4	3.575	3.284	3.073	2.913	2.788	2.687
14	245.4	19.42	8.715	5.873	4.636	3.956	3.529	3.237	3.025	2.865	2.739	2.637
16	246.5	19.43	8.692	5.844	4.604	3.922	3.494	3.202	2.989	2.828	2.701	2.599
20	248	19.45	8.66	5.803	4.558	3.874	3.445	3.15	2.936	2.774	2.646	2.544
24	249.1	19.45	8.639	5.774	4.527	3.841	3.41	3.115	2.9	2.737	2.609	2.505
30	250.1	19.46	8.617	5.746	4.496	3.808	3.376	3.079	2.864	2.7	2.57	2.466
40	251.1	19.47	8.594	5.717	4.464	3.774	3.34	3.043	2.826	2.661	2.531	2.426
50	251.8	19.48	8.581	5.699	4.444	3.754	3.319	3.02	2.803	2.637	2.507	2.401
75	252.6	19.48	8.563	5.676	4.418	3.726	3.29	2.99	2.771	2.605	2.473	2.367
100	253	19.49	8.554	5.664	4.405	3.712	3.275	2.975	2.756	2.588	2.457	2.35
200	253.7	19.49	8.54	5.646	4.385	3.69	3.252	2.951	2.731	2.563	2.431	2.323
500	254.1	19.49	8.532	5.635	4.373	3.678	3.239	2.937	2.717	2.548	2.415	2.307
∞	254.3	19.5	8.526	5.628	4.365	3.669	3.23	2.928	2.707	2.538	2.404	2.296

TABELA 4: KVANTILI FISHER–SNEDECORJEVE PORAZDELITVE

$X \sim \mathrm{F}(m,n)\colon\ P(X < F_p(m,n)) = p$ $\qquad$ $F_p(m,n) = 1/F_{1-p}(n,m)$

$p = 0.95$

$m\backslash n$	14	16	20	24	30	40	50	75	100	200	500	∞
1	4.6	4.494	4.351	4.26	4.171	4.085	4.034	3.968	3.936	3.888	3.86	3.841
2	3.739	3.634	3.493	3.403	3.316	3.232	3.183	3.119	3.087	3.041	3.014	2.996
3	3.344	3.239	3.098	3.009	2.922	2.839	2.79	2.727	2.696	2.65	2.623	2.605
4	3.112	3.007	2.866	2.776	2.69	2.606	2.557	2.494	2.463	2.417	2.39	2.372
5	2.958	2.852	2.711	2.621	2.534	2.449	2.4	2.337	2.305	2.259	2.232	2.214
6	2.848	2.741	2.599	2.508	2.421	2.336	2.286	2.222	2.191	2.144	2.117	2.099
7	2.764	2.657	2.514	2.423	2.334	2.249	2.199	2.134	2.103	2.056	2.028	2.01
8	2.699	2.591	2.447	2.355	2.266	2.18	2.13	2.064	2.032	1.985	1.957	1.938
9	2.646	2.538	2.393	2.3	2.211	2.124	2.073	2.007	1.975	1.927	1.899	1.88
10	2.602	2.494	2.348	2.255	2.165	2.077	2.026	1.959	1.927	1.878	1.85	1.831
11	2.565	2.456	2.31	2.216	2.126	2.038	1.986	1.919	1.886	1.837	1.808	1.789
12	2.534	2.425	2.278	2.183	2.092	2.003	1.952	1.884	1.85	1.801	1.772	1.752
14	2.484	2.373	2.225	2.13	2.037	1.948	1.895	1.826	1.792	1.742	1.712	1.692
16	2.445	2.333	2.184	2.088	1.995	1.904	1.85	1.78	1.746	1.694	1.664	1.644
20	2.388	2.276	2.124	2.027	1.932	1.839	1.784	1.712	1.676	1.623	1.592	1.571
24	2.349	2.235	2.082	1.984	1.887	1.793	1.737	1.663	1.627	1.572	1.539	1.517
30	2.308	2.194	2.039	1.939	1.841	1.744	1.687	1.611	1.573	1.516	1.482	1.459
40	2.266	2.151	1.994	1.892	1.792	1.693	1.634	1.555	1.515	1.455	1.419	1.394
50	2.241	2.124	1.966	1.863	1.761	1.66	1.599	1.518	1.477	1.415	1.376	1.35
75	2.205	2.087	1.927	1.822	1.718	1.614	1.551	1.466	1.422	1.354	1.312	1.283
100	2.187	2.068	1.907	1.8	1.695	1.589	1.525	1.437	1.392	1.321	1.275	1.243
200	2.159	2.039	1.875	1.768	1.66	1.551	1.484	1.391	1.342	1.263	1.21	1.17
500	2.142	2.022	1.856	1.747	1.637	1.526	1.457	1.36	1.308	1.221	1.159	1.106
∞	2.131	2.01	1.843	1.733	1.622	1.509	1.438	1.338	1.283	1.189	1.113	1

TABELA 4: KVANTILI FISHER–SNEDECORJEVE PORAZDELITVE

$X \sim \mathrm{F}(m,n)\colon\ P(X < F_p(m,n)) = p$ $\qquad$ $F_p(m,n) = 1/F_{1-p}(n,m)$

$p = 0.975$

$m\backslash n$	1	2	3	4	5	6	7	8	9	10	11	12
1	647.8	38.51	17.44	12.22	10.01	8.813	8.073	7.571	7.209	6.937	6.724	6.554
2	799.5	39	16.04	10.65	8.434	7.26	6.542	6.059	5.715	5.456	5.256	5.096
3	864.2	39.17	15.44	9.979	7.764	6.599	5.89	5.416	5.078	4.826	4.63	4.474
4	899.6	39.25	15.1	9.605	7.388	6.227	5.523	5.053	4.718	4.468	4.275	4.121
5	921.8	39.3	14.88	9.364	7.146	5.988	5.285	4.817	4.484	4.236	4.044	3.891
6	937.1	39.33	14.73	9.197	6.978	5.82	5.119	4.652	4.32	4.072	3.881	3.728
7	948.2	39.36	14.62	9.074	6.853	5.695	4.995	4.529	4.197	3.95	3.759	3.607
8	956.7	39.37	14.54	8.98	6.757	5.6	4.899	4.433	4.102	3.855	3.664	3.512
9	963.3	39.39	14.47	8.905	6.681	5.523	4.823	4.357	4.026	3.779	3.588	3.436
10	968.6	39.4	14.42	8.844	6.619	5.461	4.761	4.295	3.964	3.717	3.526	3.374
11	973	39.41	14.37	8.794	6.568	5.41	4.709	4.243	3.912	3.665	3.474	3.321
12	976.7	39.41	14.34	8.751	6.525	5.366	4.666	4.2	3.868	3.621	3.43	3.277
14	982.5	39.43	14.28	8.684	6.456	5.297	4.596	4.13	3.798	3.55	3.359	3.206
16	986.9	39.44	14.23	8.633	6.403	5.244	4.543	4.076	3.744	3.496	3.304	3.152
20	993.1	39.45	14.17	8.56	6.329	5.168	4.467	3.999	3.667	3.419	3.226	3.073
24	997.2	39.46	14.12	8.511	6.278	5.117	4.415	3.947	3.614	3.365	3.173	3.019
30	1001	39.46	14.08	8.461	6.227	5.065	4.362	3.894	3.56	3.311	3.118	2.963
40	1006	39.47	14.04	8.411	6.175	5.012	4.309	3.84	3.505	3.255	3.061	2.906
50	1008	39.48	14.01	8.381	6.144	4.98	4.276	3.807	3.472	3.221	3.027	2.871
75	1011	39.48	13.97	8.34	6.101	4.937	4.232	3.762	3.426	3.175	2.98	2.824
100	1013	39.49	13.96	8.319	6.08	4.915	4.21	3.739	3.403	3.152	2.956	2.8
200	1016	39.49	13.93	8.289	6.048	4.882	4.176	3.705	3.368	3.116	2.92	2.763
500	1017	39.5	13.91	8.27	6.028	4.862	4.156	3.684	3.347	3.094	2.898	2.74
∞	1018	39.5	13.9	8.257	6.015	4.849	4.142	3.67	3.333	3.08	2.883	2.725

TABELA 4: KVANTILI FISHER–SNEDECORJEVE PORAZDELITVE

$X \sim \mathrm{F}(m,n)\colon\ P(X < F_p(m,n)) = p$ $\qquad$ $F_p(m,n) = 1/F_{1-p}(n,m)$

$p = 0.975$

$m\backslash n$	14	16	20	24	30	40	50	75	100	200	500	∞
1	6.298	6.115	5.871	5.717	5.568	5.424	5.34	5.232	5.179	5.1	5.054	5.024
2	4.857	4.687	4.461	4.319	4.182	4.051	3.975	3.876	3.828	3.758	3.716	3.689
3	4.242	4.077	3.859	3.721	3.589	3.463	3.39	3.296	3.25	3.182	3.142	3.116
4	3.892	3.729	3.515	3.379	3.25	3.126	3.054	2.962	2.917	2.85	2.811	2.786
5	3.663	3.502	3.289	3.155	3.026	2.904	2.833	2.741	2.696	2.63	2.592	2.567
6	3.501	3.341	3.128	2.995	2.867	2.744	2.674	2.582	2.537	2.472	2.434	2.408
7	3.38	3.219	3.007	2.874	2.746	2.624	2.553	2.461	2.417	2.351	2.313	2.288
8	3.285	3.125	2.913	2.779	2.651	2.529	2.458	2.366	2.321	2.256	2.217	2.192
9	3.209	3.049	2.837	2.703	2.575	2.452	2.381	2.289	2.244	2.178	2.139	2.114
10	3.147	2.986	2.774	2.64	2.511	2.388	2.317	2.224	2.179	2.113	2.074	2.048
11	3.095	2.934	2.721	2.586	2.458	2.334	2.263	2.17	2.124	2.058	2.019	1.993
12	3.05	2.889	2.676	2.541	2.412	2.288	2.216	2.123	2.077	2.01	1.971	1.945
14	2.979	2.817	2.603	2.468	2.338	2.213	2.14	2.046	2	1.932	1.892	1.866
16	2.923	2.761	2.547	2.411	2.28	2.154	2.081	1.986	1.939	1.87	1.83	1.803
20	2.844	2.681	2.464	2.327	2.195	2.068	1.993	1.896	1.849	1.778	1.736	1.708
24	2.789	2.625	2.408	2.269	2.136	2.007	1.931	1.833	1.784	1.712	1.669	1.64
30	2.732	2.568	2.349	2.209	2.074	1.943	1.866	1.765	1.715	1.64	1.596	1.566
40	2.674	2.509	2.287	2.146	2.009	1.875	1.796	1.692	1.64	1.562	1.515	1.484
50	2.638	2.472	2.249	2.107	1.968	1.832	1.752	1.645	1.592	1.511	1.462	1.428
75	2.59	2.422	2.197	2.052	1.911	1.772	1.689	1.578	1.522	1.435	1.381	1.345
100	2.565	2.396	2.17	2.024	1.882	1.741	1.656	1.542	1.483	1.393	1.336	1.296
200	2.526	2.357	2.128	1.981	1.835	1.691	1.603	1.483	1.42	1.32	1.254	1.205
500	2.503	2.333	2.103	1.954	1.806	1.659	1.569	1.444	1.378	1.269	1.192	1.128
∞	2.487	2.316	2.085	1.935	1.787	1.637	1.545	1.417	1.347	1.229	1.137	1

TABELA 4: KVANTILI FISHER–SNEDECORJEVE PORAZDELITVE

$X \sim \mathrm{F}(m,n)\colon\ P(X < F_p(m,n)) = p$ $\qquad$ $F_p(m,n) = 1/F_{1-p}(n,m)$

$p = 0.99$

$m\backslash n$	1	2	3	4	5	6	7	8	9	10	11	12
1	4052	98.5	34.12	21.2	16.26	13.75	12.25	11.26	10.56	10.04	9.646	9.33
2	4999	99	30.82	18	13.27	10.92	9.547	8.649	8.022	7.559	7.206	6.927
3	5403	99.17	29.46	16.69	12.06	9.78	8.451	7.591	6.992	6.552	6.217	5.953
4	5625	99.25	28.71	15.98	11.39	9.148	7.847	7.006	6.422	5.994	5.668	5.412
5	5764	99.3	28.24	15.52	10.97	8.746	7.46	6.632	6.057	5.636	5.316	5.064
6	5859	99.33	27.91	15.21	10.67	8.466	7.191	6.371	5.802	5.386	5.069	4.821
7	5928	99.36	27.67	14.98	10.46	8.26	6.993	6.178	5.613	5.2	4.886	4.64
8	5981	99.37	27.49	14.8	10.29	8.102	6.84	6.029	5.467	5.057	4.744	4.499
9	6022	99.39	27.35	14.66	10.16	7.976	6.719	5.911	5.351	4.942	4.632	4.388
10	6056	99.4	27.23	14.55	10.05	7.874	6.62	5.814	5.257	4.849	4.539	4.296
11	6083	99.41	27.13	14.45	9.963	7.79	6.538	5.734	5.178	4.772	4.462	4.22
12	6106	99.42	27.05	14.37	9.888	7.718	6.469	5.667	5.111	4.706	4.397	4.155
14	6143	99.43	26.92	14.25	9.77	7.605	6.359	5.559	5.005	4.601	4.293	4.052
16	6170	99.44	26.83	14.15	9.68	7.519	6.275	5.477	4.924	4.52	4.213	3.972
20	6209	99.45	26.69	14.02	9.553	7.396	6.155	5.359	4.808	4.405	4.099	3.858
24	6235	99.46	26.6	13.93	9.466	7.313	6.074	5.279	4.729	4.327	4.021	3.78
30	6261	99.47	26.5	13.84	9.379	7.229	5.992	5.198	4.649	4.247	3.941	3.701
40	6287	99.47	26.41	13.75	9.291	7.143	5.908	5.116	4.567	4.165	3.86	3.619
50	6303	99.48	26.35	13.69	9.238	7.091	5.858	5.065	4.517	4.115	3.81	3.569
75	6324	99.49	26.28	13.61	9.166	7.022	5.789	4.998	4.449	4.048	3.742	3.501
100	6334	99.49	26.24	13.58	9.13	6.987	5.755	4.963	4.415	4.014	3.708	3.467
200	6350	99.49	26.18	13.52	9.075	6.934	5.702	4.911	4.363	3.962	3.656	3.414
500	6360	99.5	26.15	13.49	9.042	6.902	5.671	4.88	4.332	3.93	3.624	3.382
∞	6366	99.5	26.13	13.46	9.02	6.88	5.65	4.859	4.311	3.909	3.602	3.361

TABELA 4: KVANTILI FISHER–SNEDECORJEVE PORAZDELITVE

$X \sim \mathrm{F}(m,n)\colon\ P(X < F_p(m,n)) = p$ $\qquad$ $F_p(m,n) = 1/F_{1-p}(n,m)$

$p = 0.99$

$m\backslash n$	14	16	20	24	30	40	50	75	100	200	500	∞
1	8.862	8.531	8.096	7.823	7.562	7.314	7.171	6.985	6.895	6.763	6.686	6.635
2	6.515	6.226	5.849	5.614	5.39	5.179	5.057	4.9	4.824	4.713	4.648	4.605
3	5.564	5.292	4.938	4.718	4.51	4.313	4.199	4.054	3.984	3.881	3.821	3.782
4	5.035	4.773	4.431	4.218	4.018	3.828	3.72	3.58	3.513	3.414	3.357	3.319
5	4.695	4.437	4.103	3.895	3.699	3.514	3.408	3.272	3.206	3.11	3.054	3.017
6	4.456	4.202	3.871	3.667	3.473	3.291	3.186	3.052	2.988	2.893	2.838	2.802
7	4.278	4.026	3.699	3.496	3.304	3.124	3.02	2.887	2.823	2.73	2.675	2.639
8	4.14	3.89	3.564	3.363	3.173	2.993	2.89	2.758	2.694	2.601	2.547	2.511
9	4.03	3.78	3.457	3.256	3.067	2.888	2.785	2.653	2.59	2.497	2.443	2.407
10	3.939	3.691	3.368	3.168	2.979	2.801	2.698	2.567	2.503	2.411	2.356	2.321
11	3.864	3.616	3.294	3.094	2.906	2.727	2.625	2.494	2.43	2.338	2.283	2.248
12	3.8	3.553	3.231	3.032	2.843	2.665	2.562	2.431	2.368	2.275	2.22	2.185
14	3.698	3.451	3.13	2.93	2.742	2.563	2.461	2.329	2.265	2.172	2.117	2.082
16	3.619	3.372	3.051	2.852	2.663	2.484	2.382	2.249	2.185	2.091	2.036	2
20	3.505	3.259	2.938	2.738	2.549	2.369	2.265	2.132	2.067	1.971	1.915	1.878
24	3.427	3.181	2.859	2.659	2.469	2.288	2.183	2.048	1.983	1.886	1.829	1.791
30	3.348	3.101	2.778	2.577	2.386	2.203	2.098	1.96	1.893	1.794	1.735	1.696
40	3.266	3.018	2.695	2.492	2.299	2.114	2.007	1.866	1.797	1.694	1.633	1.592
50	3.215	2.967	2.643	2.44	2.245	2.058	1.949	1.806	1.735	1.629	1.566	1.523
75	3.147	2.898	2.572	2.367	2.17	1.98	1.868	1.72	1.646	1.534	1.465	1.419
100	3.112	2.863	2.535	2.329	2.131	1.938	1.825	1.674	1.598	1.481	1.408	1.358
200	3.059	2.808	2.479	2.271	2.07	1.874	1.757	1.599	1.518	1.391	1.308	1.247
500	3.026	2.775	2.445	2.235	2.032	1.833	1.713	1.551	1.466	1.328	1.232	1.153
∞	3.004	2.753	2.421	2.211	2.006	1.805	1.683	1.516	1.427	1.279	1.164	1

TABELA 4: KVANTILI FISHER–SNEDECORJEVE PORAZDELITVE

$X \sim \mathrm{F}(m,n)\colon\ P(X < F_p(m,n)) = p$ $\qquad$ $F_p(m,n) = 1/F_{1-p}(n,m)$

$p = 0.995$

$m\backslash n$	1	2	3	4	5	6	7	8	9	10	11	12
1	16211	198.5	55.55	31.33	22.78	18.63	16.24	14.69	13.61	12.83	12.23	11.75
2	19999	199	49.8	26.28	18.31	14.54	12.4	11.04	10.11	9.427	8.912	8.51
3	21615	199.2	47.47	24.26	16.53	12.92	10.88	9.596	8.717	8.081	7.6	7.226
4	22500	199.2	46.19	23.15	15.56	12.03	10.05	8.805	7.956	7.343	6.881	6.521
5	23056	199.3	45.39	22.46	14.94	11.46	9.522	8.302	7.471	6.872	6.422	6.071
6	23437	199.3	44.84	21.97	14.51	11.07	9.155	7.952	7.134	6.545	6.102	5.757
7	23715	199.4	44.43	21.62	14.2	10.79	8.885	7.694	6.885	6.302	5.865	5.525
8	23925	199.4	44.13	21.35	13.96	10.57	8.678	7.496	6.693	6.116	5.682	5.345
9	24091	199.4	43.88	21.14	13.77	10.39	8.514	7.339	6.541	5.968	5.537	5.202
10	24224	199.4	43.69	20.97	13.62	10.25	8.38	7.211	6.417	5.847	5.418	5.085
11	24334	199.4	43.52	20.82	13.49	10.13	8.27	7.104	6.314	5.746	5.32	4.988
12	24426	199.4	43.39	20.7	13.38	10.03	8.176	7.015	6.227	5.661	5.236	4.906
14	24572	199.4	43.17	20.51	13.21	9.877	8.028	6.872	6.089	5.526	5.103	4.775
16	24681	199.4	43.01	20.37	13.09	9.758	7.915	6.763	5.983	5.422	5.001	4.674
20	24836	199.4	42.78	20.17	12.9	9.589	7.754	6.608	5.832	5.274	4.855	4.53
24	24940	199.5	42.62	20.03	12.78	9.474	7.645	6.503	5.729	5.173	4.756	4.431
30	25044	199.5	42.47	19.89	12.66	9.358	7.534	6.396	5.625	5.071	4.654	4.331
40	25148	199.5	42.31	19.75	12.53	9.241	7.422	6.288	5.519	4.966	4.551	4.228
50	25211	199.5	42.21	19.67	12.45	9.17	7.354	6.222	5.454	4.902	4.488	4.165
75	25295	199.5	42.09	19.55	12.35	9.074	7.263	6.133	5.367	4.816	4.402	4.08
100	25337	199.5	42.02	19.5	12.3	9.026	7.217	6.088	5.322	4.772	4.359	4.037
200	25401	199.5	41.93	19.41	12.22	8.953	7.147	6.019	5.255	4.706	4.293	3.971
500	25439	199.5	41.87	19.36	12.17	8.909	7.104	5.978	5.215	4.666	4.252	3.931
∞	25464	199.5	41.83	19.32	12.14	8.879	7.076	5.951	5.188	4.639	4.226	3.904

TABELA 4:
KVANTILI FISHER–SNEDECORJEVE PORAZDELITVE

$X \sim \mathrm{F}(m,n)\colon\ P(X < F_p(m,n)) = p$ $\qquad$ $F_p(m,n) = 1/F_{1-p}(n,m)$

$p = 0.995$

$m\backslash n$	14	16	20	24	30	40	50	75	100	200	500	∞
1	11.06	10.58	9.944	9.551	9.18	8.828	8.626	8.366	8.241	8.057	7.95	7.879
2	7.922	7.514	6.986	6.661	6.355	6.066	5.902	5.691	5.589	5.441	5.355	5.298
3	6.68	6.303	5.818	5.519	5.239	4.976	4.826	4.635	4.542	4.408	4.33	4.279
4	5.998	5.638	5.174	4.89	4.623	4.374	4.232	4.05	3.963	3.837	3.763	3.715
5	5.562	5.212	4.762	4.486	4.228	3.986	3.849	3.674	3.589	3.467	3.396	3.35
6	5.257	4.913	4.472	4.202	3.949	3.713	3.579	3.407	3.325	3.206	3.137	3.091
7	5.031	4.692	4.257	3.991	3.742	3.509	3.376	3.208	3.127	3.01	2.941	2.897
8	4.857	4.521	4.09	3.826	3.58	3.35	3.219	3.052	2.972	2.856	2.789	2.744
9	4.717	4.384	3.956	3.695	3.45	3.222	3.092	2.927	2.847	2.732	2.665	2.621
10	4.603	4.272	3.847	3.587	3.344	3.117	2.988	2.823	2.744	2.629	2.562	2.519
11	4.508	4.179	3.756	3.497	3.255	3.028	2.9	2.736	2.657	2.543	2.476	2.432
12	4.428	4.099	3.678	3.42	3.179	2.953	2.825	2.661	2.583	2.468	2.402	2.358
14	4.299	3.972	3.553	3.296	3.056	2.831	2.703	2.54	2.461	2.347	2.281	2.237
16	4.2	3.875	3.457	3.201	2.961	2.737	2.609	2.445	2.367	2.252	2.185	2.142
20	4.059	3.734	3.318	3.062	2.823	2.598	2.47	2.306	2.227	2.112	2.044	2
24	3.961	3.638	3.222	2.967	2.727	2.502	2.373	2.208	2.128	2.012	1.943	1.898
30	3.862	3.539	3.123	2.868	2.628	2.401	2.272	2.105	2.024	1.905	1.835	1.789
40	3.76	3.437	3.022	2.765	2.524	2.296	2.164	1.995	1.912	1.79	1.717	1.669
50	3.698	3.375	2.959	2.702	2.459	2.23	2.097	1.925	1.84	1.715	1.64	1.59
75	3.612	3.29	2.872	2.614	2.37	2.137	2.001	1.824	1.737	1.605	1.525	1.47
100	3.569	3.246	2.828	2.569	2.323	2.088	1.951	1.771	1.681	1.544	1.46	1.402
200	3.503	3.18	2.76	2.5	2.251	2.012	1.872	1.685	1.59	1.442	1.346	1.276
500	3.463	3.139	2.719	2.457	2.207	1.965	1.821	1.629	1.529	1.369	1.26	1.17
∞	3.436	3.112	2.69	2.428	2.176	1.932	1.786	1.589	1.485	1.314	1.184	1

Literatura

[1] A. Ferligoj: *Osnove statistike na prosojnicah.* Ljubljana, 1997.

[2] R. Jamnik: *Matematična statistika.* DZS, Ljubljana, 1980.

[3] J. A. Čibej: *Matematika: kombinatorika, verjetnostni račun, statistika.* DZS, Ljubljana, 1994.

[4] J. Sagadin: *Osnovne statistične metode za pedagoge.* FF, Ljubljana, 1992.

[5] M. Blejec: *Uvod v statistiko.* EF, Ljubljana, 1996.

[6] L. Pfajfar: *Statistika 1.* EF, Ljubljana, 2005.

[7] F. Arh, L. Pfajfar: *Statistika 1 z zgledi.* EF, Ljubljana, 2005.

[8] M. Blejec, M. Lovrečič–Saražin, M. Perman, M. Štraus: *Statistika.* Visoka šola za podjetništvo Piran, 2003. Dosegljivo na:
`http://valjhun.fmf.uni-lj.si/~mihael/ul/vs/pdfpredavanja/gradiva.pdf`

[9] A. Jurišić: *Verjetnostni račun in statistika.* Dosegljivo na:
`http://lkrv.fri.uni-lj.si/~ajurisic/stat10/`

[10] B. Petz: *Osnovne statističke metode.* Liber, Zagreb, 1985.

[11] J. A. Rice: *Mathematical Statistics and Data Analysis.* Thomson/Brooks/Cole, Belmont, 2007.

[12] D. Freedman, R. Pisani, R. Purves: *Statistics.* Norton&Company, New York, 1998.

[13] A. Ferligoj: *Naloge iz statističnih metod.* Ljubljana, 1981.

[14] F. Arh, L. Pfajfar: *Statistika 1. Zbirka rešenih izpitnih nalog.* EF, Ljubljana, 2002.

[15] M. R. Spiegel: *Schaum's outline of theory and problems of statistics.* New York, 1999.

Viri

[16] A. Agresti, B. A. Coull: Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician* **52** (1998), 119–126.

[17] R. B. D'Agostino: Tests for the Normal Distribution. V knjigi: R. B. D'Agostino, M. A. Stephens: *Goodness-of-Fit Techniques.* Marcel Dekker, New York, 1986.

[18] E. E. Cureton: Rank-biserial correlation. *Psychometrika* **21** (1956), 287–290.

[19] B. Z. Doktorov: *George Gallup: Biography and Destiny.* Poligraf-Inform, Kaluga, 2011.

[20] D. Freedman, P. Diaconis: On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeit und verwandte Gebiete* **57** (1981), 453–476.

[21] G. Gallup: *The Sophisticated Poll Watcher's Guide.* Princeton Opinion, Princeton, 1972.

[22] G. V. Glass: Note on rank biserial correlation. *Educational and Psychological Measurement* **26** (1966), 623–631.

[23] Landon, 1,293,669; Roosevelt, 972,897. *Literary Digest* 31. 10. 1936, 5–6.

[24] P. Squire: Why the 1936 Literary Digest Poll failed. *The Public Opinion Quarterly* **52** (1988), 125–133.