Kazalo

Poglavje 1

Vzorčenje, Rice 7 + nekaj delov 8

Začenjamo s poglavjem o vzorčenju, govorimo o tem, kako izbrati vzorec, osnovne predpostavke, ki naj bi sledile iz vzorčenja. Ogledali si bomo, kaj sploh so količine, ki jih želimo ocenjevati.

Cilji poglavja:

- Populacija vs vzorec, ocenjevanje populacijskih količin s pomočjo vzorca.
- Cenilka, ocena
- Različne sheme vzorčenja
- Lastnosti: pristranskost, varianca
- Porazdelitev vzorčnega povprečja, intervali zaupanja

1.1 Populacijske vrednosti (Rice, 7.2)

Začnemo s populacijskimi vrednostmi - najprej moramo v populaciji dobro definirati, kaj bomo ocenjevali s pomočjo vzorca.

1.1.1 Povprečje in varianca končne populacije

Naj N označuje število enot v končni populaciji. N je tipično zelo velika številka, populacija je velika. Ker je veliko število podatkov težko pregledovati, si želimo neke mere, neke številke, s katerimi bi to populacijo opisali

('povzetek'). Govorimo npr. o merah središčnosti, o merah variabilnosti. Vrednosti posameznih enot naj bodo označene z $x_1, \ldots, x_N$. Povprečje teh enot je enako $\bar{x} = \mu = 1/N \sum_{i=1}^{N} x_i$. (Oznak za povprečje je veliko, mi bomo za povprečje enot v populaciji uporabljali bodisi μ bodisi $\bar{x}$. Če je le mogoče, bomo za populacijske vrednosti uporabljali grške črke.)

V kakšnem smislu je **povprečje mera središčnosti**?

Poglejmo si vsoto vseh odmikov:

$$\sum_{i=1}^{N}(x_i - \mu) = \sum_{i=1}^{N}(x_i) - N\mu = N\mu - N\mu = 0$$

Povprečje je tudi tista vrednost, za katero je **vsota kvadriranih odmikov minimalna**.

Pokažimo še to, naj bo a poljubna vrednost, pokazati želimo, da je

$$\min_a \sum_{i=1}^{N}(x_i - a)^2 = \sum_{i=1}^{N}(x_i - \mu)^2$$

Napišemo

$$\begin{aligned}
\sum_{i=1}^{N}(x_i - a)^2 &= \sum_{i=1}^{N}(x_i - \mu + \mu - a)^2 \\
&= \sum_{i=1}^{N}[(x_i - \mu)^2 + 2(x_i - \mu)(\mu - a) + (\mu - a)^2] \\
&= \sum_{i=1}^{N}(x_i - \mu)^2 + 2(\mu - a)\sum_{i=1}^{N}(x_i - \mu) + N(\mu - a)^2
\end{aligned}$$

Vsota $\sum_{i=1}^{N}(x_i - \mu) = 0$, zato se gornja izpeljava poenostavi v

$$\sum_{i=1}^{N}(x_i - a)^2 = \sum_{i=1}^{N}(x_i - \mu)^2 + N(\mu - a)^2 \tag{1.1}$$

Oba izraza na desni sta pozitivna, torej dobimo najmanjšo vrednost pri tistem a, pri katerem ja najbolj desni izraz enak 0, torej $a = \mu$.

Gornjo izpeljavo bomo v podobni obliki srečali zelo pogosto in sicer predvsem dva njena dela:

- Trik, da prištejemo in odštejemo vrednost, za katero odmike že poznamo, oz. ki nas zanima $(-\mu + \mu)$
- Dejstvo, da je vsota srednjih členov enaka 0

Vpeljimo še eno oznako:

$$\sigma^2 = 1/N \sum_{i=1}^{N} (x_i - \mu)^2.$$

Ponovitev:

Ponovimo pravila za računanje s pričakovano vrednostjo in varianco:

$$\begin{aligned} E(aX) &= aE(X) \\ E(X+Y) &= E(X) + E(Y) \\ var(aX) &= a^2 var(X) \\ var(X+Y) &= var(X) + var(Y) + 2cov(X,Y) \end{aligned}$$

Definicija kovariance:

$$\text{cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

Pravila za kovarianco:

$$\text{cov}(\sum X_i, \sum X_j) = \sum \sum cov(X_i, X_j)$$

Vemo, da je kovarianca enaka 0, kadar sta spremenljivki neodvisni (obratno ni nujno res) in da nam predznak kovariance pove, na kakšen način sta povezani spremenljivki (večja vrednost X poveča verjetnost za večjo vrednost Y ali obratno). Tako kot standardni odklon (oz. varianca), je tudi kovarianca odvisna od enot, v katerih merimo. Zato vrednost kovariance ne pove ničesar o moči povezanosti med spremenljivkama. Vrednost 3 npr. le pove, da gre za pozitivno povezanost, da bi vedeli, kako močna je, pa bi morali poznati tudi varianci spremenljivke.
Namesto kovariance zato pogosto interpretiramo korelacijski koeficient, ki je definiran kot:

$$\text{cor}(X,Y) = \frac{cov(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}.$$

Takoj lahko vidimo, da je korelacijski koeficient brez enote, pri predmetu 'Verjetnost' ste dokazali, da je vedno med -1 in 1, predznak in vrednost 0 imajo enak pomen kot pri kovarianci.

.---

Naj bo X_i slučajna vrednost iz populacije. Ker je vrednosti N in imajo vse enako verjetnost izbora, je verjetnost, da je $X_i = x_1$, enaka $1/N$. Pričakovana vrednost je definirana kot $E(X) = \sum_x xp(x)$. Pokažimo, da velja

$$E(X_i) = \mu,$$

torej, da je na končni populaciji **pričakovana vrednost enaka povprečju**.

Vzemimo sedaj naključno enoto iz vzorca, njeno vrednost označimo z X_1.

$$E(X_1) = \sum_x xp(x) = \sum_{i=1}^{N} x_i 1/N = \mu.$$

Podobno pokažimo, da je σ^2, ki smo jo definirali zgoraj, ravno enaka **varianci**.

$$\text{var}(X_1) = E[(X_1 - E(X_1))^2] = \sum_x (x-\mu)^2 p(x) = \sum_{i=1}^{N} (x_i - \mu)^2 1/N = \sigma^2.$$

Iz izpeljave (1.1) lahko povzamemo izraz za varianco, ki bo pogosto prišel prav (vzamemo $a = 0$ in obrnemo izraz):

$$\sigma^2 = 1/N \sum_{i=1}^{N} (x_i - \mu)^2 = 1/N \sum_{i=1}^{N} x_i^2 - \mu^2 = E(X^2) - E(X)^2 \qquad (1.2)$$

Mimogrede: Standardni odklon označujemo s $\sigma = \sqrt{\sigma^2}$.

Primer:
Recimo, da imamo le dve vrsti vrednosti: vsaka enota i je lahko enaka le 0 ali 1, torej $x_i = 0$ ali $x_i = 1$ za vsak i. Povprečje vrednosti je potem enako

$$\mu = 1/N \sum_{i=1}^{N} x_i = 1/N \sum_{x_i=1} x_i + 1/N \sum_{x_i=0} x_i = 1/N \sum_{x_i=1} x_i = 1/N \sum_{i=1}^{N} I[x_i = 1] = \pi.$$

Povprečje je torej enako deležu enic v populaciji, tega označimo s π (za populacijske vrednosti vrednosti bomo uporabljali grške črke).
Kaj pa varianca? Uporabimo zapis (1.2) in upoštevamo, da je $1^2 = 1$ in $0^2 = 0$, torej je $x_i^2 = x_i$ ne glede na to, kakšna je vrednost x_i.

$$\sigma^2 = 1/N \sum_{i=1}^{N} x_i^2 - \mu^2 = 1/N \sum_{i=1}^{N} x_i - \pi^2 = \pi - \pi^2 = \pi(1 - \pi).$$

V primeru zvezne porazdelitve v populaciji lahko populacijo npr. opišemo s povprečjem in standardnim odklonom. V primeru binarne (dihotomne) spremenljivke je primernejši opis z deležem. Standardnega odklona nam niti ni potrebno poročati, saj ga lahko izračunamo iz deleža (je vezan na delež).

Na končni populaciji torej pričakovane vrednosti za lažjo predstavo vedno lahko zamenjamo z vsotami. V neskončni populaciji vrednosti ne moremo prešteti, zato bomo govorili o njihovi pogostosti - o njihovi verjetnosti.

1.1.2 Neskončna populacija

Če je populacija neskončna (ima neskončno enot), ne moremo zapisati vrednosti x_i za vse enote. Da bi populacijo lahko opisali, zato namesto vrednosti podamo delež (verjetnost $p(x)$) enot z neko vrednostjo. Če je možnih vrednosti neskončno (in gre za zvezno spremenljivko) uporabljamo gostoto $f(x)$.
V neskončni populaciji je govoriti o povprečju vrednosti nekoliko težje, pričakovana vrednost je definirana kot $E(X) = \int_{-\infty}^{\infty} xf(x)dx$ za zvezne spremenljivke oziroma $E(X) = \sum_x xp(x)$ za diskretne.

Za vajo za zvezne spremenljivke izpeljimo še rezultat, ki je ekvivalenten

izrazu (1.2):

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\
&= \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f(x)dx \\
&= \int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu \int_{-\infty}^{\infty} x f(x)dx + \mu^2 \int_{-\infty}^{\infty} f(x)dx \\
&= \int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu\mu + \mu^2 \\
&= \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 \\
&= E(X^2) - E(X)^2
\end{aligned}$$

1.2 Vzorec (Rice 7.3, 7.5)

Iz populacije naključno izberimo **eno enoto**. Njeno vrednost označimo z X - uporabimo veliko tiskano črko, kar označuje, da gre za slučajno spremenljivko. Porazdelitev te slučajne spremenljivke je enaka porazdelitvi populacije, verjetnosti v povezavi z vrednostjo X računamo s pomočjo gostote $f(x)$.

Predstavljajmo si, da populacijo predstavlja škatla z listki. Želimo oceniti neko količino povezano s številkami, ki so na teh listkih, npr μ. Izvleči smemo n listkov. Kaj lahko povemo o μ?

Ponavadi nas bo zanimal primer, ko bomo iz populacije vzeli $n > 1$ vrednosti $X_1, \ldots, X_n$.
Pravimo, da slučajne spremenljivke $X_1, \ldots, X_n$ sestavljajo **naključni vzorec** in neke populacije, kadar so enako porazdeljene in med seboj paroma neodvisne. Pogosto to zapišemo tudi kot *i.i.d. spremenljivke* (independent identically distributed variables).

Predstavljajmo si, da naključno vzorčenje iz **neskončne populacije** poteka zaporedoma. Vrednost v prvem poskusu označuje slučajna spremenljivka X_1. Njena porazdelitev je enaka porazdelitvi populacije ($f(x)$ oz. $p(x)$). Poznamo torej verjetnost $P(X_1 = x_1)$ za poljubno vrednost x_1 (oz. poljuben interval). Sedaj poskus ponovimo, naslednjo vrednost označimo z X_2. Dogodka $\{X_1 = x_1\}$ in $\{X_2 = x_2\}$ sta neodvisna, torej verjetnosti za spremenljivko X_2 niso nič drugačne od verjetnosti za spremenljivko X_1. Dejstvo, da smo x_1 'odstranili' iz populacije prav nič ne vpliva na porazdelitev spremenljivke X_2.

Kadar vzorčimo iz **končne populacije**, pa je potrebno biti bolj pozoren. Predstavljajmo si, da našo populacijo predstavlja N listkov z vrednostmi v škatli in da v vsakem poskusu izvlečemo en listek ter zabeležimo vrednost, dogodek $X_1 = x_1$ pomeni, da smo v prvem poskusu izvlekli listek z vrednostjo x_1. Pri tem naj imajo vsi listki enako verjetnost, da bodo izbrani, torej $P(X_1 = x_1) = 1/N$. Nato poskus ponovimo, vrednost naslednjega listka označimo z X_2. Kakšna je porazdelitev vrednosti X_2? Ali je nam to, kaj smo izvlekli v prvem poskusu, da kako informacijo o vrednosti v drugem poskusu?

Odgovor na gornji vprašanji je odvisen od **načina vzorčenja**. Da bi zagotovili naključni vzorec po gornji definiciji in torej medsebojno neodvisne slučajne spremenljivke, moramo izžrebani listek vsakokrat vrniti v škatlo.

Tako vzorčenje imenujemo vzorčenje s ponavljanjem. V nadaljevanju bomo posebej obravnavali obe možnosti - vzorčenje brez ali z ponavljanjem (=neskončna populacija), začeli bomo s slednjo, saj so izpeljave zaradi neodvisnosti preprostejše.

Zanimalo nas bo, kako s pomočjo vzorca povemo nekaj o populaciji. Vrednosti, ki jo izračunamo iz vzorca in ki naj bi ocenjevala neko količino v populaciji pravimo **cenilka** oziroma **ocena**. Pri tem je cenilka je neka funkcija slučajnih spremenljivk $f(X_1, \ldots, X_n)$, ocena pa je njena vrednost na dejanskih vrednostih nekega vzorca $f(x_1, \ldots, x_n)$. V tem razdelku se bomo osredotočili na ocenjevanje populacijskega povprečja in populacijske variance.

1.2.1 Vzorčenje iz neskončne populacije

Imamo torej $X_1, \ldots, X_n$ i.i.d. slučajnih spremenljivk, naj bo $E(X_i) = \mu$, $\text{var}(X_i) = \sigma^2$ za vsak i. Začnimo z ocenjevanjem populacijskega povprečja.

Želimo oceniti **populacijsko povprečje**. Najprej nam pride na misel, da ga ocenimo s pomočjo vzorčnega povprečja, preprosto torej izračunamo

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Preprosto vprašanje **'Ali $\bar{X}$ ocenjuje μ?'** je nekoliko slabo definirano:

- Skoraj vedno bo veljalo $\bar{X} \neq \mu$ (če je X zvezna slučajna spremenljivka, velja $P(\bar{X} = \mu) = 0$, dokaz malce kasneje).
- Če bi vzeli več vzorcev, bi bil torej $\bar{X}$ včasih manj, drugič več od μ. Kaj lahko torem rečemo o variabilnosti $\bar{X}$ okrog μ?
- Ali vsaj 'v povprečju' dobimo pravo vrednost? Zanima nas pričakovana vrednost $E(\bar{X})$.

Najprej torej ugotovimo, da je $\bar{X}$ **slučajna spremenljivka** - na vsakem vzorcu bomo dobili drugačno vrednost. Da bi vedeli 'vse´ o vzorčnem povprečju, nas torej zanima njegova porazdelitev. A o tem kasneje, za začetek bomo nekoliko manj zahtevni in nas bo zanimala pričakovana vrednost vzorčnega povprečja.

Nepristranskost povprečja

Pri izpeljavi uporabimo, da so vse slučajne spremenljivke enako porazdeljene, in je zato $E(X_i) = E(X) = \mu$ za vsak i:

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu.$$

Pričakovana vrednost vzorčnega povprečja je torej enaka populacijskemu povprečju, $E(\bar{X}) = \mu$, pravimo, da je $\bar{X}$ *nepristranska cenilka* μ.

Primer:

Vzemimo Bernoullijevo porazdeljeno slučajno spremenljivko, verjetnost 0 in 1 je enaka 0,5. Pokažimo, da vzorčno povprečje za $n = 3$ iz te porazdelitve nepristransko ocenjuje populacijsko povprečje:

$$\mu = \pi = 0{,}5$$
$$E(\bar{X}) = \frac{1}{3}\frac{3}{8} + \frac{2}{3}\frac{3}{8} + 1\frac{1}{8} = \frac{4}{8} = 0{,}5$$

Pokazali smo, da je v povprečju naša cenilka 'dobra' - je nepristranska. Hkrati vidimo, da na nobenem vzorcu sploh ne moremo dobiti prave vrednosti. Kaj lahko rečemo o variabilnosti ocene?

Tabela 1.1

	0,0,0	1,0,0	1,1,0	1,1,1
P	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$
$\bar{X}$	0	$\frac{1}{3}$	$\frac{2}{3}$	1
$(\bar{X}-\mu)$	$-\frac{1}{2}$	$-\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$
$\sum(X_i-\bar{X})^2$	0	$(\frac{2}{3})^2+2(\frac{-1}{3}^2)$	$2(\frac{1}{3})^2+(\frac{2}{3})^2$	0
$\sum(X_i-\bar{X})^2$	0	$\frac{6}{9}$	$\frac{6}{9}$	0
$\frac{1}{n}\sum(X_i-\bar{X})^2$	0	$\frac{2}{9}$	$\frac{2}{9}$	0
$\frac{1}{n-1}\sum(X_i-\bar{X})^2$	0	$\frac{1}{3}$	$\frac{1}{3}$	0
$\sqrt{\frac{1}{n-1}\sum(X_i-\bar{X})^2}$	0	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	0

Napaka vzorčnega povprečja

Naslednji korak je izračun variabilnosti okrog tega povprečja. Označimo varianco v populaciji s σ^2, uporabimo, da so vrednosti med seboj neodvisne, torej da je $\text{cov}[X_i, X_j] = 0$ za vsak $i \neq j$. V tem primeru je varianca vsote enaka vsoti varianc.

$$\begin{aligned}
\text{var}[\bar{X}] &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \text{var}X_i\right) \\
&= \frac{1}{n^2} n\text{var}(X) = \frac{\sigma^2}{n}
\end{aligned}$$

Standardni odklon vzorčnega povprečja imenujemo **standardna napaka**. Z vsakim vzorcem ocenjujemo μ, vsakič se nekaj zmotimo. Koliko se motimo v 'povprečju' nam pove standardni odklon teh vzorčnih povprečij okrog pravega, populacijskega' povprečja. Temu standardnemu odklonu pravimo standardna napaka. je enaka $SE = \frac{\sigma}{\sqrt{n}}$.

Interpretacija formule za standardno napako: standardna napaka je odvisna od dveh količin.

- Velikosti vzorca n - večji kot je vzorec, manj se bomo zmotili pri oceni populacijske vrednosti; n je zato v imenovalcu. Standardna napaka se zmanjšuje s korenom velikosti vzorca: nekaj dodanih enot bo pri majhnih vzorcih imelo precejšen vpliv na standardno napako, medtem ko se pri velikih vzorcih praktično ne bo poznalo.

- Velikosti populacijske variance - če so vrednosti v populaciji med seboj precej različne (velika varianca), bomo v vzorec dobivali vrednosti, ki precej odstopajo od populacijskega povprečja, zato bo naša ocena lahko precej oddaljena od prave vrednosti. Če pa se vrednosti že v populaciji ne razlikujejo dosti, bo vzorčno povprečje hitro precej natančno (primer višin študentov in robotov).

Primer:

Ilustrirajmo formulo na našem primeru:

$$\sigma^2 = \pi(1-\pi) = \frac{1}{4} \;\rightarrow\; \frac{\sigma^2}{n} = \frac{1}{12}$$

$$var(\bar{X}) = E[(\bar{X} - 0{,}5)^2] = 2\frac{1}{4}\frac{1}{8} + 2\frac{1}{36}\frac{3}{8} = \frac{1}{16} + \frac{1}{48} = \frac{4}{48} = \frac{1}{12}$$

.

Standardna napaka vs. standardni odklon

Standardni odklon opisuje variabilnost posameznih enot populacije (ali vzorca) - koliko se razlikujejo od povprečja, kakšni so odmiki. Standardna napaka opisuje variabilnost vzorčnega povprečja. Na vsakem vzorcu je seveda samo eno povprečje, sprašujemo se, koliko bi pri enakem načinu vzorčenja (povsem enakih pogojih) dobili drugačne vrednosti, če bi zadevo ponavljali.

Primer: standardni odklon višin študentov (predstavljamo si z normalno porazdelitvijo), standardna napaka povprečne višine.

Primer:

Radi bi dokazali, da se po dveh mesecih jemanja nekega zdravila bolnikom zniža sistolični tlak. Standardni odklon te razlike je $\sigma = 25$, v vzorec vzamemo $n = 100$ bolnikov. Kakšna je verjetnost, da na vzorcu opazimo povprečno spremembo za vsaj 10 mmHg, čeprav v populaciji ni sprememb po jemanju zdravila?

Zanima nas verjetnost, da na vzorcu dobimo $|\bar{X}| \geq 10$, čeprav je populacijsko povprečje enako $\mu = 0$. Vzorec je velikosti 100, zato je standardna napaka enaka $SE = \frac{25}{10} = 2{,}5$. Opaženo znižanje je 4 standardne napake proč od pričakovane vrednosti. Da bi izračunali, kakšna je verjetnost te razlike, bi morali poznati porazdelitev vzorčnega povprečja. Ker te ne poznamo, se lahko zatečemo k neenačbi Čebiševa (kasneje bomo poizkusili še drugače - uporabili bomo centralni limitni izrek): $P(|X - \mu| \geq b) \leq \frac{\sigma^2}{b^2}$. V našem primeru pišemo

$$P(|\bar{X} - \mu| \geq 4SE) \leq \frac{(SE)^2}{(4SE)^2} = \frac{(SE)^2}{16\sigma^2/n} = \frac{1}{16}.$$

Verjetnost, da do take oz. še večje razlike pride po naključju, je manjša ali enaka 0,0625 ne glede na porazdelitev populacije oziroma vzorčnega povprečja.

.

Primer:

Denimo, da smo iz neskončne populacije vzeli vzorec velikosti 100. Naj bo $\widetilde{\mu}$ cenilka za populacijsko povprečje, ki jo izračunamo le na podlagi prvih petih vrednosti vzorca: $\widetilde{\mu} = \frac{1}{5}\sum_{i=1}^{5} X_i$. Ali je ta cenilka nepristranska? Na kakšen način je 'slabša' od vzorčnega povprečja?

Taka cenilka je nepristranska, vendar pa da večjo standardno napako - standardna napaka je enaka $\sigma/(\sqrt{5})$ namesto $\sigma/(\sqrt{100})$, standardna napaka je torej cca 4x večja.

.

Kadar bomo imeli na voljo več cenilk za isto količino, bomo primerjali njihovo *standardno napako*. Ko se bomo pogovarjali o lastnostih neke cenilke, nas bosta tako zanimala tako *nepristranskost* kot tudi *standardna napaka*.

Ocena variance

Ocenjujemo μ v populaciji. Zanima nas, koliko se motimo. V standardni napaki nastopa σ^2, tega seveda v praksi skoraj nikoli ne bomo poznali (saj ne poznamo niti μ), kako bi σ^2 ocenili iz podatkov?

Primer:

Pokazali smo že, da povprečje vzorca nepristransko oceni populacijsko povprečje. Ali lahko kaj takega rečemo tudi za oceno variance?
Varianca je enaka $\pi(1-\pi) = \frac{1}{4}$.
Oglejmo si cenilko $\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Pričakovana vrednost te cenilke je

$$E(\widetilde{\sigma}^2) = 2\frac{2}{9}\frac{3}{8} = \frac{1}{6}$$

Vidimo, da ta cenilka ni nepristranska. Še več - prav na vsakem možnem vzorcu ta cenilka podceni dejansko varianco.

.

Na podlagi našega vzorca želimo oceniti σ^2. Videli smo, da morda najbolj intuitivno smiselna ocena ni tudi nepristranska.

Naj bo naša cenilka $\widehat{\sigma}^2 = c\sum_{i=1}^{n}(X_i - \bar{X})^2$. Vemo že, da $c = \frac{1}{n}$ ne da nepristranske ocene - kakšna mora biti vrednost konstante c, da bo naša ocena nepristranska?

S formulo: poiščimo c tako, da bo veljalo

$$E[c\sum_{i=1}^{n}(X_i - \bar{X})^2] = \sigma^2$$

Iz pričakovane vrednosti lahko izpostavimo c, saj je konstanta. Zanima nas pričakovana vrednost kvadriranih odmikov na vzorcu, radi bi povedali, kako daleč od željene vrednosti smo.

Vemo, da velja $E(Y^2) = E(Y)^2 + \mathrm{var}(Y)$ (glej (1.2).
Zanima nas pričakovana vrednost kvadriranega odmika, vzemimo to za spremenljivko Y v prejšnjem izrazu in dobimo

$$\begin{aligned}
E[(X_i - \bar{X})^2] &= [E(X_i - \bar{X})]^2 + \text{var}(X_i - \bar{X}) \\
&= [E(X_i) - E(\bar{X})]^2 + \text{var}(X_i) + \text{var}(\bar{X}) - 2cov(X_i, \bar{X}) \\
&= \text{var}(X_i) + \text{var}(\bar{X}) - 2cov(X_i, \frac{1}{n}\sum_j X_j) \\
&= \sigma^2 + \frac{\sigma^2}{n} - 2[cov(X_i, \frac{1}{n}X_i) + (n-1)cov(X_i, \frac{1}{n}X_j)] \\
&= \sigma^2 + \frac{\sigma^2}{n} - 2\frac{1}{n}\text{var}(X_i) \\
&= \sigma^2\left[1 + \frac{1}{n} - 2\frac{1}{n}\right] \\
&= \sigma^2\left[1 - \frac{1}{n}\right] = \frac{n-1}{n}\sigma^2
\end{aligned}$$

Vidimo, da pričakovana vrednost kvadriranega odmika vsekakor je povezana s populacijsko varianco, poiščimo torej konstanto c, za katero bo $E[c\sum_i (X_i - \bar{X})^2] = \sigma^2$:

$$E[c\sum_i (X_i - \bar{X})^2] = c(n-1)\sigma^2$$

Ker želimo, da velja $E(\widehat{\sigma}^2) = \sigma^2$, mora biti $c = \frac{1}{n-1}$.
Nepristranska cenilka populacijske variance je enaka $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$.

Povzemimo gornjo izpeljavo v pomožnem izreku (Rice, str. 480, Lemma A):

Naj bodo X_i neodvisne slučajne spremenljivke, tako da velja $E(X_i) = \mu_i$, $E(\bar{X}) = \mu$ in $\text{var}X_i = \sigma^2$. Potem velja:

$$E[(X_i - \bar{X})^2] = (\mu_i - \mu)^2 + \frac{n-1}{n}\sigma^2$$

Oglejmo si še enkrat naš rezultat. Kako bi intuitivno razložili, zakaj je potrebno deljenje z $n-1$? Oglejmo si pričakovano vrednost izraza $\frac{1}{n}\sum(X_i - \mu)^2$:

$$E[\frac{1}{n}\sum_i (X_i - \mu)^2] = \frac{1}{n}\sum_i \left\{[E(X_i - \mu)]^2 + \text{var}(X_i - \mu)\right\} = \text{var}(X_i) = \sigma^2$$

Če bi poznali vrednost μ, bi s kvadriranimi odmiki od te vrednosti tvorili nepristransko cenilko. Mi pa te vrednosti ne poznamo, zato jo nadomeščamo z vzorčnim povprečjem. Vzorčno povprečje je seveda nekoliko bližje vzorčnim vrednostim kot populacijsko (povprečje je tista vrednost, pri kateri je vsota kvadriranih odmikov minimalna). Zato je jasno, da bomo z ocenjenim vzorčnim povprečjem podcenili dejansko varianco (pri deljenju z n).

Primer:

Preverimo na našem primeru. Oglejmo si cenilko $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Pričakovana vrednost te cenilke je

$$E(\widetilde{\sigma}^2) = 2\frac{1}{3}\frac{3}{8} = \frac{1}{4}.$$

Ta cenilka je nepristranska.

.

Primer:

Za konec si oglejmo še cenilko $\widetilde{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$. Ali je ta cenilka nepristranska?
Pa poizkusimo: Če to velja, potem je

$$var(\widetilde{\sigma}) = E(\widetilde{\sigma}^2) - E(\widetilde{\sigma})^2 = \sigma^2 - \sigma^2 = 0$$

Seveda ne velja $E(X^2) = E(X)^2$ (če bi bilo to res, bi bila varianca vsake slučajne spremenljivke enaka 0). Oglejmo si to na našem primeru, vemo, da je $\sigma = 0{,}5$.

$$E(\widetilde{\sigma}) = 2\frac{1}{\sqrt{(3)}}\frac{3}{8} = \frac{\sqrt{3}}{4}.$$

Ta cenilka ni nepristranska.

.

Doslednost

V tem razdelku smo se precej ukvarjali z nepristranskostjo. Na primerih smo vzeli precej majhen vzorec in to iz dveh razlogov: prvi je bil, da je na manjšem vzorcu manj možnih izidov za $\bar{X}$ in zato manj računanja, drugi razlog pa je bil, da do tako opazne pristranskosti pride le na majhnih vzorcih. Marsikdaj namreč ocena ni nepristranska, je pa dosledna. Doslednost cenilke je definirana takole:

Naj bo $\hat{\theta}$ cenilka parametra θ na vzorcu velikosti n enot. Tedaj pravimo, da je cenilka dosledna, če za vsak $\epsilon > 0$ velja[1]

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0.$$

Oglejmo si dva izreka, ki bosta zelo koristna pri dokazovanju doslednosti.

Šibki zakon velikih števil:
Naj bodo X_i i.i.d., $E(X_i) = \mu$, $\mathrm{var}(X_i) = \sigma^2 < \infty$. Označimo $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Potem za vsak $\epsilon > 0$ velja

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1.$$

Izrek:
Denimo, da zaporedje $X_1, X_2, \ldots$ konvergira v verjetnosti proti slučajni spremenljivki X in da je h zvezna funkcija. Potem zaporedje $h(X_1), h(X_2), \ldots$ v verjetnosti konvergira proti $h(X)$.

Z opisanima izrekoma lahko hitro dokažemo doslednost cenilke $\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. Neformalno si poglejmo glavne korake dokaza

- Zaradi zakona velikih števil vsota $\frac{1}{n}\sum_{i=1}^n X_i^2$ konvergira proti $E(X_i^2) = \sigma^2 + \mu^2$, vsota $\frac{1}{n}\sum_{i=1}^n X_i$ pa proti $E(X)$.
- Ker je funkcija $h(x) = x^2$ zvezna funkcija, vemo da tudi $[\frac{1}{n}\sum_{i=1}^n X_i]^2$ konvergira proti μ^2.
- Ker velja
$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i^2) - (\frac{1}{n}\sum_{i=1}^n X_i)^2,$$
vemo, da $\frac{1}{n}\sum_{i=1}^n X_i^2$ konvergira proti $\sigma^2 + \mu^2 - \mu^2 = \sigma^2$.

[1] Zaporedje slučajnih spremenljivk Y_n konvergira v verjetnosti proti neki konstanti λ, če za vsak $\epsilon > 0$, velja $\lim_{n\to\infty} P(|Y_n - \lambda| < \epsilon) = 1$

Podobno bi lahko doslednost pokazali tudi za cenilko za standardni odklon. Ker je koren zvezna funkcija, to velja po drugem izreku.

Kaj pa nasprotno - ali velja, da je vsaka nepristranska cenilka tudi dosledna?

Ne, to ni nujno. Kot protiprimer vzemimo X_1 kot cenilko za povprečje - populacijsko povprečje vedno ocenimo kar s prvim elementom vzorca. Ta cenilka je nepristranska ($E(X_1) = \mu$), ni pa dosledna, saj ne konvergira nikamor. (enako bi dobili s povprečjem prvih petih enot)
Pripomnimo še, da je tale primer precej umeten - vsako zaporedje cenilk (v n), ki je nepristransko in nekam konvergira, bo tudi dosledno. In to praktično vedno velja za cenilke.

Povzemimo:

- Doslednost je lastnost, ki opisuje, kaj velja, ko gre n proti neskončnosti. Oziroma v praksi - kaj skoraj velja za zelo velike vzorce. Dosledna cenilka z večanjem vzorca postaja vse bolj podobna vrednosti, ki jo ocenjuje. In to na kateremkoli vzorcu - verjetnost, da se vzorčna vrednost kaj dosti razlikuje od populacijske gre proti 0.

- Cenilka je nepristranska, če je njena pričakovana vrednost enaka populacijski vrednosti parametra, ki ga ocenjujemo.

- Doslednost še ne zagotavlja nepristranskosti, zato pa so nepristranske cenilke praktično vedno dosledne.

- Doslednost je osnovna lastnost cenilke, ki jo zahtevamo na zelo velikih vzorcih (ko gre n proti neskončno), na manjših vzorcih nas bo zanimalo, ali je cenilka tudi nepristranska.

- Lastnosti cenilke, ki opisujejo njeno obnašanje, ko gre n proti neskončno, imenujemo asimptotske lastnosti. Poleg asimptotskih lastnosti cenilk nas bo vedno zanimalo tudi njeno obnašanje na majhnih vzorcih. Pogosto se bo izkazalo, da v teoriji lahko dokažemo asimptotske lastnosti, medtem ko bo dokazovanje lastnosti na majhnih vzorcih velikokrat teoretično prezahtevno in se bomo zatekli k simulacijam.

Povzetek

- Nepristranskost: pričakovana vrednost cenilke je enaka populacijski vrednosti (primer: populacijsko povprečje lahko nepristransko ocenimo z vzorčnim)
- Standardna napaka (standardni odklon cenilke): da neko informacijo o tem, koliko se motimo na vzorcu. Primerjamo lahko standardne napake različnih nepristranskih cenilk. Najboljša bo seveda tista, ki bo imela najmanjši SE
- Ocena variance: potrebujemo za oceno standardne napake ali samo zase (populacijski parameter, ki nas zanima)

Povzetek primerov:

- Bernoullijeva slučajna spremenljivka, vzorec velikosti $n = 3$. Nepristranska ocena povprečja, nepristranska ocena variance, dosledna ocena variance in standardnega odklona. Izračunali smo vse količine, za katere smo izpeljali cenilke.
- Primer s pritiskom: na podlagi podatkov smo ocenili SE, da bi dobili občutek, kako natančen je naš rezultat. Da bi nekaj povedali o verjetnosti, bi morali imeti še kak podatek o porazdelitvi. Tu ga nismo imeli, zato smo uporabili neenakost Čebiševa.
- Primer cenilke kot povprečje prvih petih meritev: namerno smo si ogledali manj smiselno cenilko, da bi na tem očitnem primeru znali formalno zapisati lastnost, zaradi katere nam ta cenilka ni vseč. Enako (podobno) cenilko smo uporabili tudi kot protiprimer, s katerim smo pokazali, da nepristranskost še ne pomeni doslednosti.

1.2.2 Vzorčenje iz končne populacije

Sedaj si predstavljajmo, da listke iz škatle vzorčimo brez vračanja - vsaka enota je lahko izbrana največ enkrat. Takemu vzorčenju pravimo *enostavno slučajno vzorčenje* (simple random sampling).
Slučajne spremenljivke X_i, ki predstavljajo vrednost enote izbrane na i-tem koraku, sedaj postanejo medsebojno odvisne.

Primer:
Denimo, da imamo v škatli pet listkov z vrednostmi 1,3,5,7,9, ven vzamemo vzorec velikosti 2. Če je vrednost $X_1 = 5$, vemo da na drugem koraku listka s številko 5 ne moremo izbrati. Vrednost izbrana na prvem koraku nam torej pove nekaj o vrednosti na drugem koraku - spremenljivki nista neodvisni.

.

Vzorčenje iz končne populacije si lahko predstavljamo tudi kot **permutacijo enot populacije**, pri čemer v vzorec zajamemo prvih n permutiranih enot: enote od 1 do N preuredimo, X_1 naj bo vrednost prve enote v novem vrstnem redu, X_N vrednost zadnje enote. V vzorec nato vzamemo le prvih n enot: $X_1, X_2, \ldots X_n$.

Primer:
Populacija, ki nas zanima, so študenti prisotni na predavanjih. Izbrati želimo vzorec velikosti 3, slučajna spremenljivka naj bo višina študenta. Vse študente postavimo v vrsto v naključnem vrstnem redu. V vzorec zajamemo prve tri študente v vrsti.

.

Porazdelitev

Slučajne spremenljivke sedaj niso med seboj neodvisne. Kaj pa lahko rečemo o njihovi porazdelitvi? Če naključno izbiramo iz populacije velikosti N, je $P(X_1 = x_j) = 1/N$ za vsak j. Kaj pa lahko rečemo o verjetnostni porazdelitvi X_2? Naj bo na prvem koraku izbrana enota j, na drugem koraku torej izbiramo le še med $N - 1$ vrednostmi. Ali to pomeni, da je porazdelitev X_2 drugačna?

Odgovor na to vprašanje lahko utemeljimo na več načinov:

- Poizkusimo najprej 'intuitivno': zamislimo si, da nažrebamo veliko permutacij N enot, nato pa si ogledamo le vrednosti, ki so na drugem mestu. Nobenega razloga ni, da bi drugo mesto imelo drugačno porazdelitev kot na primer prvo ali tretje. Slučajne spremenljivke $X_1, X_2, \ldots X_n$ imajo torej vse enako porazdelitev, $P(X_i = x_j) = 1/N$ za vsak i in j.
- Sedaj preverimo še z izračunom (uporabimo formulo za popolno verjetnost):

$$\begin{aligned} P(X_2 = x_j) &= \sum_{i=1}^{N} P(X_2 = x_j | X_1 = x_i) P(X_1 = x_i) \\ &= \sum_{i=1; i \neq j}^{N} P(X_2 = x_j | X_1 = x_i) P(X_1 = x_i) + \\ &\quad + P(X_2 = x_i | X_1 = x_i) P(X_1 = x_i) \\ &= \sum_{i=1; i \neq j}^{N} \frac{1}{(N-1)} \cdot \frac{1}{N} + 0 \cdot \frac{1}{N} = \frac{N-1}{(N-1)N} \\ &= \frac{1}{N} \end{aligned}$$

Primer:
Študente postavljamo v vrsto v različnih vrstnih redih, vsakič nam vzorec predstavljajo študenti na prvih petih mestih. Če poznam velikost prvega študenta, vem nekaj o velikosti drugega. Brez te informacije pa so vsa mesta enaka - enaka verjetnost je, da se največji postavi na prvo kot na drugo mesto. Pričakovana vrednost je na vseh mestih enaka.
..

Torej: pogojna verjetnost $P(X_2 = x_j | X_1)$ zaradi odvisnosti ni enaka robni porazdelitvi $P(X_2 = x_j)$ (kot je bilo to res pri neskončni populaciji). Robne porazdelitve spremeljivk X_i so vse enake. Spremenljivke X_i torej niso i.i.d., so enako porazdeljene, niso pa neodvisne.

Odvisnost

Sedaj skušajmo z izračunom opredeliti še odvisnost - kaj lahko rečemo o $\text{cov}(X_1, X_2)$? Kaj pa za splošen $i \neq j$?

Uporabimo trik. Vrednosti X_j si znova predstavljamo kot permutacijo, vemo da velja:

$$\mathrm{cov}(X_i, \sum_{j=1}^{N} X_j) = \mathrm{cov}(X_i, \sum_{k=1}^{N} x_k) = 0$$

Ker je vsota vseh vrednosti (od 1 do N) konstanta, je zgornji izraz enak 0, torej

$$\mathrm{cov}(X_i, \sum_{j=1}^{N} X_j) = \mathrm{cov}(X_i, X_i) + (N-1)\mathrm{cov}(X_i, X_j) = 0$$

In zato (za $i \neq j$)

$$\mathrm{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

Spremenljivki sta torej negativno povezani, povezanost je enaka za vsak $i \neq j$. Ker je absolutno vrednost kovariance nemogoče interpretirati (odvisna je od enot, v katerih merimo), izpeljemo še formulo za korelacijo:

$$\mathrm{cor}(X_i, X_j) = -\frac{\sigma^2}{(N-1)\sigma^2} = -\frac{1}{N-1}.$$

Korelacija je torej odvisna izključno od velikosti populacije in z večanjem populacije hitro postane zelo majhna. Večja kot je populacija, manj pomembna bo povezanost - v veliki populaciji se ob odvzetju ene enote le malo spremeni.

Pričakovana vrednost in varianca vzorčnega povprečja

Sedaj si poglejmo, kako dejstvo, da spremenljivke niso več neodvisne (so pa še vedno enako porazdeljene), vpliva na pričakovano vrednost in varianco vzorčnega povprečja.

Najprej se spomnimo izpeljave pričakovane vrednosti pri neskončni populaciji: pri izpeljavi smo uporabili, da so vse slučajne spremenljivke enako porazdeljene, neodvisnost za izpeljavo ni bila potrebna. Vzorčno povprečje je tudi sedaj nepristranska ocena populacijskega povprečja.

Spremeni pa se izpeljava standardne napake:

$$\begin{aligned}
\text{var}\left(\bar{X}\right) &= \text{cov}\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{cov}\left(X_i, \frac{1}{n}\sum_{j=1}^{n} X_j\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left\{\text{cov}\left(X_i, X_i\right) + (n-1)\text{cov}\left(X_i, X_j\right)\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left\{\sigma^2 - (n-1)\frac{\sigma^2}{N-1}\right\} = \frac{\sigma^2}{n}\frac{N-n}{(N-1)}
\end{aligned}$$

Standardna napaka je torej enaka $SE = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

Primer:

Oceniti želimo povprečno število zaposlenih v podjetjih neke panoge ob začetku letošnjega leta. Panoga je razdeljena na podskupine, v eni izmed skupin je le 11 podjetij. Uspeli smo pridobiti podatke za naključen vzorec šestih izmed teh podjetij. Naj bo X_i število zaposlenih v i-tem podjetju našega vzorca, μ naj označuje njihovo povprečje, σ pa standardni odklon. V drugi panogi imamo 100 podjetij. Kako velik vzorec moramo vzeti iz te panoge, da bomo dobili približno enako veliko standardno napako (privzemimo da je varianca tudi v tej panogi enaka σ^2)? Kaj pa če bi imeli panogo z zelo velikim številom podjetij?

Za $N = 11$ in $n = 6$ dobimo $SE^2 = \frac{\sigma^2}{6}\frac{11-6}{(10)} = \frac{\sigma^2}{12}$. Pri populaciji velikosti 100 nam vzorec velikosti 10 da standardno napako $SE^2 = \frac{\sigma^2}{11}$, vzorec velikosti 11 pa standardno napako $SE^2 = \frac{\sigma^2}{12,2}$.
Če je populacija večja, bomo za enako standardno napako torej potrebovali več enot. Če bi imeli skupino z zelo velikim številom podjetij, bi bil izraz $\frac{N-n}{(N-1)}$ približno enak 1, zato bi potrebovali 12 podjetij. Velikost potrebnega vzorca se z večanjem populacije veča, vendar pa to večanje kmalu ni več bistveno. .

Ocena variance

Oglejmo si še oceno variance. Kakšna mora biti vrednost konstante c, da bo cenilka $\widehat{\sigma}^2 = c\sum_{i=1}^{n}(X_i - \bar{X})^2$ nepristransko ocenila vrednost σ^2?

Ponovimo izračun za neskončno populacijo, upoštevamo, da je $var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$:

$$
\begin{aligned}
&E[c\sum_{i=1}^{n}(X_i - \bar{X})^2] = \\
&= c\sum_{i=1}^{n}\left[E(X_i^2 - \bar{X}^2)\right] \\
&= c\sum_{i=1}^{n}\left[\mathrm{var}(X_i) + \mathrm{var}(\bar{X}) - 2cov(X_i, \bar{X})\right] \\
&= c\sum_{i=1}^{n}\left\{\sigma^2 + \frac{\sigma^2}{n}\frac{N-n}{N-1} - 2\left[cov\left(X_i, \frac{1}{n}X_i\right) + (n-1)cov\left(X_i, \frac{1}{n}X_j\right)\right]\right\} \\
&= c\sum_{i=1}^{n}\left\{\sigma^2 + \frac{\sigma^2}{n}\frac{N-n}{N-1} - 2\left[\frac{1}{n}\sigma^2 - \frac{(n-1)}{n}\frac{\sigma^2}{N-1}\right]\right\} \\
&= \frac{cn\sigma^2}{n(N-1)}\left[n(N-1) + N - n - 2(N-1) + 2(n-1)\right] \\
&= \frac{cn\sigma^2}{n(N-1)}\left[nN - n + N - n - 2N + 2 + 2n - 2\right] \\
&= \frac{cn\sigma^2}{n(N-1)}(nN - N) \\
&= c\sigma^2\frac{N(n-1)}{N-1}
\end{aligned}
$$

Ker želimo, da velja $E(\widehat{\sigma}^2) = \sigma^2$, mora biti $c = \frac{1}{n-1}\frac{N-1}{N}$, torej

$$\widehat{\sigma}^2 = \frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Za konec zapišimo še nepristransko cenilko za varianco povprečja:

Združimo dosedanje rezultate in dobimo

$$\begin{aligned}\widehat{SE}^2 &= \frac{\widehat{\sigma}^2(N-n)}{N-1} = \frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i-\bar{X})^2\frac{(N-n)}{N-1} \\ &= \frac{s^2}{n}\frac{N-n}{N},\end{aligned}$$

kjer je $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$.

Primer, Rice 7.3.3., Example D:

V nekem naselju je 8000 stanovanj. Anketa na naključnem vzorcu 100 prebivalcev poda oceno: povprečno število motornih vozil je 1,6, standardni odklon na vzorcu 0,8.

- Ocenjena standardna napaka je 0,08.
- Pokažimo, da lahko popravek za končno populacijo pri oceni standardne napake zanemarimo: $\frac{N-n}{N} = 7900/8000 = 0{,}9875$, ocenjena std. napaka bi bila 0,079.

Povzetek

- Če je populacija končna, vrednosti enot izbranih v vzorec navkljub naključnemu izbiranju med seboj niso neodvisne, kovarianca med enotami je enaka:

$$\text{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}.$$

- Vzorčno povprečje je nepristranska cenilka populacijskega povprečja ne glede na to, ali imamo končno ali neskončno populacijo.

- Varianca vzorčnega povprečja in nepristranska cenilka za varianco v populaciji sta enaki

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{(N-1)};\ \widehat{\sigma}^2 = \frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

  za končno in

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n};\ \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

  za neskončno populacijo (ko gre N proti neskončnosti formuli sovpadeta).

- V končni populaciji bo na izbiro velikosti vzorca vplivala tudi velikost populacije. Vendar pa bo ta vpliv pomemben le pri sorazmerno majhnih populacijah, oziroma takrat, ko velikost vzorca predstavlja nezanemarljiv delež velikosti populacije. Taka situacija pa bo redka - zakaj bi namreč vzorčili, če prav tako lahko zberemo podatke za celo populacijo. Taka situacija je pogosto tudi 'sumljiva' - če podatki niso zbrani za populacijo 11 podjetij, lahko pomislimo, da vzorec ni bil zbran naključno, temveč imamo podatke za tistih 6 podjetij, ki smo jih uspeli dobiti, veliko vprašanje pa je ali je porazdelitev manjkajočih podatkov zares enaka porazdelitvi zbranih.

Povzetek primerov:

- Primera z višinami študentov se spomnimo takrat, ko nas zanima, zakaj je porazdelitev X_2 enaka porazdelitvi X_1. V splošnem torej velja, da je $E(X_i)$ enak ne glede na i, prav tako tudi $\text{var}(X_i)$. To nam poenostavi vse vsote. Enaka je tudi $\text{cov}(X_i, X_j)$ - vedno je ista vrednost, razen seveda, kadar je $i = j$.

- Primer s podjetji nam nazorno pokaže, da je pri majhnih populacijah velikost populacije pomembna. Manjša kot je velikost populacije, manjši vzorec bo potreben za enako standardno napako. Ko pa velikost populacije postane velika glede na velikost vzorca, velikost populacije postane nepomembna.

Opombe glede vzorčenja, motivacija za naslednje poglavje:

- Mariskateri dokaz v statistiki se bo začel z 'naj bodo X_i i.i.d. spremenljivke, ...'.
- Vsakršna odvisnost prinese precej komplikacij. Enote med seboj odvisne (dvojčki), ponovljene meritve na istih posameznikih ...
- Naključnost = enaka porazdelitev kot v populaciji. So posamezniki res vsi enako porazdeljeni. So, zame, ki ne vem nič dodatnega o njih. Paziti je potrebno, da podskupina podatkov nima neke skupne lastnosti, ki jo razločuje od drugih podskupin. Če jo ima, bo na to potrebno paziti pri načrtu vzorčenja.
- Naključnosti še zdaleč ni preprosto zagotoviti. Ločiti je potrebno med slučajnim in priložnostnim vzorcem
- Primeri: ali je pričujoči vzorec naključen po velikosti za celotno populacijo študentov? Primeri enega zdravnika. Telefonske, internetne ankete. Vzporedne volitve. Manjkajoče vrednosti splošno (bolnik ne pride na kontrolo ali medicinska sestra na dopustu)
- Problemi z naključnostjo vzorca: Vedeti moramo, kaj so naše predpostavke, za kaj jih potrebujemo. Razumeti moramo, kje se zaplete, če predpostavke niso izpolnjene. Vedeti moramo, kakšna vprašanja moramo zastaviti raziskovalcu. Obstajajo cela področja statistike, ki se borijo s temi problemi - v nekaterih primerih je možno kaj storiti (informacije o vzorcih manjkajočih vrednostih, lastnostih podskupine ...), včasih se ne da prav dosti (primer internetne ankete na občini).
- Reprezentativnost vzorca: kaj si mislimo o izjavi: 'vzorec ima 300 enot, zato je reprezentativen'. Je bolje zbrati večji vzorec ali ostati pri manjšem, ki je reprezentativen? Uteževanje ali sklepanje o neki 'pod-populaciji'.

1.3 Porazdelitev vzorčnega povprečja

Porazdelitev igra osrednjo vlogo v statistiki. Poznati porazdelitev pomeni biti zmožen izračunati poljubno verjetnost (da se vrednosti nahajajo v nekem intervalu, etc). Porazdelitev: skupek vseh možnih izidov in njihovih verjetnosti (v primeru zvezne spremenljivke gostota).

Ogledali smo si, kako ocenimo vzorčno povprečje in varianco ter kako vzorčno povprečje variira, vendar bi radi naredili še korak več: jasno je, da samo poročanje ocene povprečja ne zadostuje, populacijsko povprečje ne bo praktično nikoli identično enako vzorčnemu (če je porazdelitev vzorčnega povprečja zvezna, je verjetnost, da bo $\bar{X} = \mu$ enaka 0). Standardna napaka nam daje neko predstavo o velikosti odstopanja, vendar pa vemo, da je verjetnost odmika za eno (ali več) standardnih napak odvisna od porazdelitve. Želimo torej nekaj povedati o porazdelitvi vzorčnega povprečja. Zaradi preprostosti bomo v vseh nadaljnjih razdelkih predpostavili, da gre za neskončno populacijo, oz. da je populacija dovolj velika, da lahko popravke za končnost zanemarimo.

1.3.1 Normalno porazdeljena populacija

Vzemimo najprej, da je populacija normalno porazdeljena. Vemo, da je vsota neodvisnih normalno porazdeljenih spremenljivk zopet normalna in ker je povprečje le s konstanto $(1/n)$ pomnožena vsota, vemo, da je tudi povprečje normalno porazdeljeno. Poznamo tudi oba parametra te porazdelitve, torej $\bar{X} \sim N(\mu, \sigma^2/n)$.

Če torej poznamo porazdelitev cenilke na vzorcu, lahko naša pričakovanja na vzorcu opišemo s pomočjo verjetnosti. Ta pričakovanja pogosto opišemo s pomočjo intervala zaupanja. Ta nam poda neko predstavo o natančnosti naše ocene. Poglejmo, kako ga tvorimo.
Vemo, da je vzorčno povprečje $\bar{X}$ normalno porazdeljeno okrog μ. Poznamo tudi varianco in tako lahko npr. izračunamo, da je verjetnost, da vzorčno povprečje iz te populacije od μ ni oddaljeno za več kot $1{,}96\sigma/\sqrt{n}$, enaka 0,95. V praksi nas seveda zanima obrnjeno vprašanje - s podatkov lahko izračunamo $\bar{X}$, zanima pa nas, kaj lahko rečemo o μ. Najprej moramo razumeti dve dejstvi:

- O vrednosti populacijskega povprečja s pomočjo vzorca z gotovostjo ne

moremo trditi prav ničesar. Res je, da je $\bar{X}$ večinoma blizu μ, a vedno obstaja neka (četudi majhna) verjetnost, da je naša vzorčna ocena daleč proč od μ.

- Populacijsko povprečje μ ni slučajna spremenljivka. Ne glede na to, kakšen vzorec vzamemo, bo vedno na istem mestu. Zato torej ni smiselno govoriti o tem, kakšna je na podlagi našega vzorca verjetnost, da je μ na nekem intervalu.

Kaj torej lahko trdimo? Vemo da je verjetnost, da je $|\bar{X} - \mu| < 1{,}96\sigma/\sqrt{n}$ enaka 0,95, v 95% primerov bo torej vzorčno povprečje manj kot toliko oddaljeno od populacijskega. Če torej okrog $\bar{X}$ narišemo interval širine $1{,}96\sigma/\sqrt{n}$, bo v 95% primerov vseboval populacijsko povprečje. Ko narišemo interval z našimi konkretnimi podatki seveda ne vemo, ali smo med 95% tistih, ki so populacijsko povprečje uspeli zajeti, ali ne. Zato namesto o 'verjetnostnem' intervalu, govorimo o intervalu 'zaupanja'. Za naš interval imamo 95% zaupanje, da vsebuje populacijsko povprečje, če ga res vsebuje, pa na podlagi vzorca ne moremo vedeti.

Primer:
Vrnimo se k primeru zniževanja sistoličnega tlaka: radi bi dokazali, da se po dveh tednih jemanja nekega zdravila bolnikom spremeni sistolični tlak. Na vzorcu $n = 100$ smo dobili $\bar{X} = 20$, vemo, da je $\sigma = 50$, torej $SE = 5$. Predpostavimo, da je populacija sprememb pri posameznih bolnikih normalno porazdeljena. Kaj lahko trdimo glede na vzorčno povprečje?

95 % interval zaupanja izračunamo kot $\bar{X} \pm 1{,}96SE$, dobimo interval $[10{,}2, 29{,}8]$. S 95% zaupanjem torej lahko trdimo, da je populacijsko povprečje na tem intervalu. Ker so v intervalu same pozitivne vrednosti, lahko s 95% zaupanjem trdimo, da se tlak po jemanju zdravila zniža. S 95% zaupanjem lahko tudi trdimo, da se ne zmanjša za več kot 30 mmHg.

Primer (Rice, 22):
Raziskovalec vzorči iz normalne porazdelitve in oceno populacijskega povprečja poda kot $\bar{X} \pm SE_{\bar{X}}$. Kakšna je velikost tega intervala zaupanja?

Primer (Rice, 22):
Za koliko moramo povečati vzorec, da razpolovimo 95% interval zaupanja?

1.3.2 Porazdelitev t

Vrnimo se k normalno porazdeljeni populaciji - v tem razdelku vedno predpostavimo, da so vrednosti X_i v vzorcu neodvisne normalno porazdeljene slučajne spremenljivke.
Do sedaj smo se spraševali, kakšna bo SE na vzorcu, če populacijsko vrednost σ poznamo oziroma smo računali intervale zaupanja pri znani populacijski vrednosti σ. V praksi bo naše razmišljanje največkrat obrnjeno. O populaciji ne bomo vedli prav dosti, naš cilj bo sklepanje o populacijskem povprečju, če poznamo vzorčno povprečje. Ker populacijskega povprečja ne bomo poznali, je le težko pričakovati, da bomo vedeli, koliko so vrednosti oddaljene od populacijskega povprečja, torej da bomo poznali populacijsko varianco.
Do sedaj smo standardno napako izračunali kot $SE = \sigma/\sqrt{n}$ in uporabili, da je linearna transformacija normalne porazdelitve zopet normalna, torej da je $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
Kaj torej naredimo, če SE ni znan? Seveda ga lahko ocenimo iz podatkov. To že znamo, vemo, da je $\hat{\sigma}^2 = 1/(n-1)\sum(X_i - \bar{X})^2$ nepristranska cenilka σ^2 in zato $\hat{\sigma}^2/n$ nepristranska cenilka za SE^2. Vendar pa s pomočjo te ocene ne moremo 'standardizirati' vzorčnega povprečja: izraz $\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}}$ ni linearna transformacija normalne porazdelitve, saj je $\widehat{SE}$ slučajna spremenljivka in ne konstanta. Zanima nas torej porazdelitev kvocienta dveh slučajnih spremenljivk, pri čemer je v številcu normalna spremenljivka. Nobenega razloga seveda ni, da bi bila ta porazdelitev spet normalna, če bomo hoteli računati intervale zaupanja tudi v tem primeru, bomo torej morali poiskati to porazdelitev.
Najprej nekaj intuitivnih ugotovitev:

- Kadar vrednosti σ ne poznamo, imamo na voljo nekoliko manj informacije - pričakujemo torej, da bo interval zaupanja pri ocenjeni standardni napaki nekoliko širši. Večji kot bo vzorec, manjša bo negotovost pri ocenjevanju SE, zato pričakujemo, da bo interval zaupanja vse bolj podoben tistemu, ki bi ga dobili ob znani vrednosti σ.

- Varianca izraza v imenovalcu ima spredaj faktor $1/n$. Z večanjem velikosti vzorca se bo ta varianca manjšala bistveno hitreje kot varianca v številcu, zato bo postala skoraj konstantna v primerjavi z variabilnostjo izraza v številcu. Pričakujemo torej, da porazdelitev tega kvocienta postaja čedalje bolj normalna.

Sedaj izpeljimo porazdelitev ulomka $\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}}$, zopet predpostavljamo, da spremenljivke X_i izhajajo iz normalne porazdelitve. Iz Uvoda v statistiko že vemo, da gre za porazdelitev t in sicer t_{n-1}. Pokažimo torej, zakaj to velja.

- Definicija porazdelitve t: naj bo $Z \sim N(0,1)$ in $U \sim \chi^2_n$ ter Z in U neodvisni. Potem je kvocient $Z/\sqrt{U/n}$ porazdeljen po porazdelitvi t_n
- Kvocient torej prepišemo v

$$\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\frac{\hat{\sigma}/\sqrt{n}}{\sigma/\sqrt{n}}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\hat{\sigma}^2}{\sigma^2}}/\sqrt{n-1}}$$

Sedaj imamo torej števec, ki je porazdeljen standardno normalno, pokazati moramo še, da je spremenljivka $U = \frac{(n-1)\hat{\sigma}^2}{\sigma^2}$ porazdeljena po χ^2_{n-1} porazdelitvi in da sta števec in imenovalec neodvisni slučajni spremenljivki.

Do neke mere ste oboje že dokazali pri predemtu Verjetnost, podroben dokaz je tudi v Rice-u (poglavje 6), povzemimo rezultate:

- Oglejmo si odmike $X_i - \bar{X}$ neodvisni od $\bar{X}$. Preprosto je pokazati, da je kovarianca enaka 0:

$$\begin{aligned} cov(\bar{X}, X_i - \bar{X}) &= cov(\bar{X}, X_i) - \mathrm{var}\bar{X} \\ &= \frac{1}{n} cov(X_i, X_i) + \frac{n-1}{n} cov(X_i, X_j) - \mathrm{var}\bar{X} = \sigma^2/n + 0 - \sigma^2/n. \end{aligned}$$

  V gornji izpeljavi smo uporabili, da so X_i med seboj neodvisne.

- Ali nekoreliranost že pomeni tudi neodvisnost? V splošnem ne, je pa to res, pri spremenljivkah, ki jih lahko zapišemo kot robne spremenljivke iz neke večrazsežne normalne porazdelitve. Vprašanje torej je, ali lahko vektor $[\bar{X}, X_i - \bar{X}]^T$ zapišemo kot AX, kjer je A neka matrika. To bi namreč pomenilo, da je iskani vektor linearna kombinacija neodvisnih normalno porazdeljenih spremenljivk in zato porazdeljen normalno.
  V našem primeru je to mogoče, zato sta $X_i - \bar{X}$ in $\bar{X}$ neodvisni. Ker je vzorčna varianca vsota $(X_i - \bar{X})^2$ (pomnožena s konstanto), je vzorčna varianca neodvisna od vzorčnega povprečja.

- Potrebujemo še porazdelitev $\frac{(n-1)\hat{\sigma}^2}{\sigma^2}$, iščemo torej porazdelitev $\hat{\sigma}^2$, ki je edina slučajna spremenljivka v tem izrazu. Zapišemo lahko

$$\sum(X_i-\mu)^2 \;=\; \sum(X_i-\bar{X}+\bar{X}-\mu)^2 = \sum(X_i-\bar{X})^2 + \sum(\mu-\bar{X})^2$$

Vemo, da je $\sum \frac{(X_i-\mu)^2}{\sigma^2}$ porazdeljen kot χ^2_n, $\frac{(\bar{X}-\mu)^2}{\sigma^2/n}$ pa kot χ^2. Gornjo vsoto torej lahko prepišemo v vsoto

$$\sigma^2 Y = \sigma^2 Z + (n-1)\hat{\sigma}^2 \;\rightarrow\; Y = Z + (n-1)\hat{\sigma}^2/\sigma^2.$$

Velja naslednje:
Naj bo $X+Z=Y$, kjer je $Y \sim \chi^2_m$ in $Z \sim \chi^2_n$ $(m>n)$, X in Z sta neodvisni slučajni spremenljivki. Potem je $X \sim \chi^2_{m-n}$ (dokaz z momentno rodovnimi funkcijami).

V našem primeru je Y porazdeljena kot χ^2_n, Z pa kot χ^2_1, pokazali smo tudi neodvisnost X in Z. Velja torej, da je porazdelitev $(n-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$.

Povzemimo - poznamo porazdelitev $\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}}$, torej znamo interval zaupanja izračunati tudi tedaj, ko populacijske variance ne poznamo.

Opomba:

Denimo, da velja $X \sim N(0,1)$, $Y \sim N(0,1)$, spremenljivki sta neodvisni. Vemo, da je vsota $Z = X+Y$ porazdeljena kot $N(0,2)$. Vzemimo sedaj, da je $Z \sim N(0,2)$ in $Y \sim N(0,1)$, spremenljivki sta neodvisni. Kako je porazdeljena razlika $Z-Y$?

Razlika je zagotovo porazdeljena normalno (linearna kombinacija normalnih), poglejmo še parametra:

$$\begin{aligned} E(Z-Y) &= E(Z)-E(Y)=0 \\ var(Z-Y) &= var(Z)+var(Y)=3 \end{aligned}$$

Velja torej: spremenljivka $X = Z-Y$ je porazdeljena kot $Y \sim N(0,3)$.
Vidimo, da moramo v enačbah, v katerih nastopajo slučajne spremenljivke, biti zelo pozorni: če sta Z vsota X in Y, potem Z in Y nista neodvisni

spremenljivki, kar seveda vpliva na porazdelitev njune razlike. Pri ugotavljanju porazdelitve vsote ali razlike je ključno vprašanje, ali sta spremenljivki neodvisni.

Primer:
Zopet se vrnimo k primeru zniževanja sistoličnega tlaka: radi bi dokazali, da se po dveh tednih jemanja nekega zdravila bolnikom spremeni sistolični tlak. Ko smo ta primer obiskali prejšnjič smo predpostavili, da poznamo populacijsko varianco. Sedaj že vemo, da nam je ni treba poznati, recimo, da smo jo ocenili z vzorca $\hat{\sigma}^2 = 2600$. Izračunajmo $\hat{SE} = \hat{\sigma}/\sqrt{n} = 5{,}1$. Še vedno pa potrebujemo predpostavko, da je populacija sprememb pri posameznih bolnikih normalno porazdeljena. Poiščemo meje za 95% interval zaupanja iz porazdelitve t_{99}, dobimo $-1{,}98$, 95% interval zaupanja je torej $\bar{X} \pm 1{,}98\hat{SE}$, dobimo interval $[9{,}9, 30{,}1]$. S 95% zaupanjem torej lahko trdimo, da je populacijsko povprečje na tem intervalu. Ker so v intervalu same pozitivne vrednosti, lahko s 95% zaupanjem trdimo, da se tlak po jemanju zdravila zniža.

Primer:
Ali je interval zaupanja, kadar ne poznamo σ vedno širši kot če bi vrednost σ poznali?

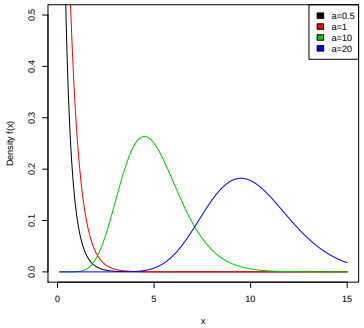
Če vrednost σ poznam, lahko uporabim normalno porazdelitev, ki vedno da manjši faktor od porazdelitve t. Če bi torej ocena populacijske variance bila enaka populacijski varianci, bi bil interval zaupanja z ocenjeno σ vedno širši. Vendar pa je ocena seveda lahko tudi manjša in tako naš interval zaupanja ožji.
Medtem ko so bili 95% intervali zaupanja pri vseh vzorcih iz neke normalne porazdelitve katere varianco poznamo enako široki, to ne velja za intervale zaupanja, kadar varianco ocenimo iz podatkov.

Toda pozor - v tem razdelku je še vedno ostajala pomembna predpostavka o populaciji in sicer, da poznamo populacijsko porazdelitev in da je ta normalna. To je seveda še ena zahteva, ki je v praksi ne bomo mogli izpolnjevati. Tu pa v igro stopi centralni limitni izrek, ki omogoča, da porazdelitev vzorčnega povprečja poznamo tudi takrat, ko o populacijski porazdelitvi ne vemo praktično ničesar.

1.3.3 Ostale porazdelitve populacije

V prejšnjem razdelku smo videli, da se povprečja vzorcev iz normalne populacije porazdeljuje normalno. Ali tudi pri drugih porazdelitvah ostajamo znotraj iste porazdelitve?

Iz predmeta Verjetnost vemo, da je vsota gama porazdeljenih spremenljivk z enakim parametrom λ spet gama porazdeljena in da je vsota χ^2 porazdeljenih spremenljivk zopet χ^2 (χ^2 je poseben primer gama porazdelitve). Gostota porazdelitve gama je prikazana na sliki 1.1.



Slika 1.1: *Porazdelitev gama za različne vrednosti parametra a pri* $\lambda = 0{,}5$ *(*χ^2_{2a}*).*

Nasprotno vemo, da vsota Bernoullijevih spremenljivk ni Bernoullijevo porazdeljena (vsekakor lahko zavzame več kot 2 različni vrednosti). Vsota Bernoullijevih spremenljivk je porazdeljena z binomsko porazdelitvijo.
Kaj pa vsota enakomerno porazdeljenih spremenljivk? Tudi tu je očitno, da ne bomo dobili enakomerne porazdelitve.

Zanima nas torej porazdelitev vzorčnega povprečja. Za računanje je bistveni korak predvsem vprašanje, kako se porazdeljuje vsota spremenljivk iz neke porazdelitve, množenje s konstanto za izračun verjetnosti ne bo huda ovira. Vidimo, da pri nekaterih porazdelitvah z vsoto ostajamo znotraj iste porazdelitve, pri drugih dobimo novo, znano porazdelitev, pri tretjih pa je potrebno še precej računanja in izpeljevanja, da bi lahko nazadnje izračunali interval zaupanja.

Več o tem v razdelku o centralnem limitnem izreku.

1.3.4 Centralni limitni izrek

Centralni limitni izrek je bil omenjen pri predmetu verjetnost in pravi (tega ne bomo dokazovali, dokaz ne bo prispeval k našemu razumevanju o uporabnosti izreka):

Centralni limitni izrek
Naj bo $X_1, X_2, \ldots, X_n$ zaporedje neodvisnih enako porazdeljenih slučajnih spremenljivk s pričakovano vrednostjo μ in varianco σ^2, $0 < \sigma^2 < \infty$. Potem za vsoto $S_n = X_1 + X_2 + \ldots + X_n$ velja

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x),$$

kjer je $\Phi(x)$ porazdelitvena funkcija standardizirane normalne porazdelitve.

Oziroma drugače:

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq x\right) = P(Z \leq x) = \Phi(x).$$

Če je torej $\bar{X}$ vzorčno povprečje, kumulativna porazdelitvena funkcija slučajne spremenljivke $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ konvergira proti standardni normalni porazdelitvi.

Nekaj opomb:

- Gornji izrek velja za vzorčno povprečje iz katerekoli porazdelitve. Torej - o porazdelitvi X nam ni treba predpostaviti skoraj ničesar (le da je varianca končna), pa vedno dobimo normalno porazdelitev!
- Normalna porazdelitev je posledica vsote mnogih naključnih vplivov.
- Pomembno, je, da so X_i neodvisne spremenljivke, torej vzorčimo iz neskončne porazdelitve oz. s ponavljanjem.
- Praktično edini pogoj je, da je varianca končna. Obstaja več verzij - večrazsežnostni CLT, CLT, ki dovoljuje določeno vrsto odvisnosti med spremenljivkami, itd.
- Poznamo spremenljivke, katerih vsota ni normalno porazdeljena (gama, binomska), ker tudi za njih velja CLT, je to možno le pri porazdelitvah,

ki konvergirajo (postajajo čedalje bolj podobne) k normalni porazdelitvi. To dejstvo lahko tudi uporabimo kot pripomoček za lažje računanje - pogosto se npr. binomsko porazdelitev aproksimira z normalno (treba je le izračunati povprečje in std. odklon, torej parametra ustrezne normalne porazdelitve). Vendar je ta uporaba danes zaradi vsesplošne prisotnosti hitrih računalnikov manj pomembna kot nekoč. Je pa tako razumevanje ključno za občutek, kako velika je npr. varianca neke binomske porazdelitve.

- Čeprav izrek velja ne glede na porazdelitev X, pa se hitrost konvergence precej razlikuje glede na porazdelitev spremenljivke X. Kako dobra bo aproksimacija, torej ni odvisno le od n temveč tudi od oblike porazdelitve. Hitrost konvergence je potrebno preveriti za vsako porazdelitev posebej

Vidimo torej, da nam ni treba poznati porazdelitve X, pa vseeno poznamo (približno) porazdelitev izraza $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. Kaj pa lahko rečemo o izrazu $\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}}$?

- Pokazati se da (Slutskyjev izrek), da izraz zopet konvergira proti normalni porazdelitvi. Normalna porazdelitev torej aproksimira ta izraz tudi, če populacijske variance ne poznam.

- Kaj pa za manjše vzorce? Nam bo spet v pomoč porazdelitev t? Poglejmo nazaj v dokaz porazdelitve t, v številcu je vse OK, za imenovalec pa v splošnem ne moremo trditi, prav tako ne moremo na enak način kot prej pokazati neodvisnosti. V splošnem ne moremo trditi, da bo t boljša aproksimacija od normalne. Če je populacija približno normalna bo verjetno OK. Za kako 'gršo' porazdelitev, bi bilo treba preveriti s simulacijami (obstajajo izreki, zahtevajo npr. znane momente, etc).

Primer:
Zavarovalnica želi oceniti povprečni strošek obiska patronažne sestre, pri tem so se obrnili na pomoč k statistiku. V Sloveniji je 715 patronažnih sester, zanima jih, kako velik vzorec morajo zbrati (in jih povprašati po njihovem strošku), da bodo lahko ocenili to povprečje.

Statistik na to vprašanje še ne more odgovoriti - potrebuje še dodatne informacije. Odgovor bomo podali s 95% intervalom zaupanja, torej nas zanima, kako natančno oceno želijo. Drugi podatek, ki ga potrebujemo za

izračun intervala zaupanja pa je, kakšna menijo, da je variabilnost v populaciji.

Upamo, da bo vzorec dovolj velik, da bi približek s centralnim limitnim izrekom moral biti dovolj dober, za vsak primer se vseeno lahko pozanimamo, ali je populacija morda zelo asimetrično porazdeljena (nekaj posameznikov z zelo velikimi stroški). Če je, morda povprečni strošek ni najboljša mera. Prav tako lahko naročnika vprašamo, ali je 95% zaupanje zanj sprejemljiva raven (ne)gotovosti.

Denimo, da smo dobili naslednje odgovore: ocena naj bi bila natančna na cca 10 EUR, stroški v populaciji variirajo s standardnim odklonom cca 50 EUR, populacija je približno normalno porazdeljena.

Ker so podane številke približne, je tak lahko tudi naš izračun: popravek za končnost populacije zanemarimo, uporabimo lahko, da je 95% interval zaupanja širok približno 4 standardne napake. Ker naj bi bila naša ocena natančna na 10 EUR, mora biti standardna napaka torej enaka 2,5, vzorec mora torej biti velik cca 400.

Opomba: Dobili smo precej velik vzorec. Ali popravek za končnost v tem primeru res lahko zanemarimo?

Primer, Rice 7.3.3., primer D:

V nekem naselju je 8000 stanovanj. Anketa na naključnem vzorcu 100 prebivalcev poda oceno: povprečno število motornih vozil je 1,6, standardni odklon na vzorcu 0,8.

- Ocenjena standardna napaka je 0,08.
- Popravek za končno populacijo lahko zanemarimo: $\frac{N-n}{N-1} = 7900/7999 = 0{,}9877$, ocenjena std. napaka bi bila 0,0795.
- 95% interval zaupanja izračunajmo najprej s pomočjo normalne porazdelitve: $\bar{X} \pm 1{,}96\hat{SE} = [1{,}44, 1{,}76]$. Če namesto normalne porazdelitve vzamemo porazdelitev t, bo ustrezni faktor 1,98, zaokroženo na dve decimalki dobimo enak rezultat.
- Če nas zanima število vseh avtomobilov v naselju: 95% interval zaupanja pomnožimo z 8000, dobimo $[11520, 14080]$. Uporaba t oz. normalne porazdelitve bo dala nekaj razlike (cca 50 avtomobilov).

- Interpretacija intervala zaupanja!

Delež prebivalcev v vzorcu, ki bodo naslednje leto prodajali stanovanje je 0,12. Ocenite št. prodaj naslednje leto.

Ocenili smo delež v vzorcu, zanima nas delež (oz. število) v populaciji. Delež na vzorcu je enak vzorčnemu povprečju, ker je vzorčno povprečje nepristranska cenilka, to lahko trdimo tudi za delež. Vemo, da vzorčni ni kar enak populacijskemu deležu, zanima nas koliko se motimo. Uporabimo formuli, ki smo ju izpeljali za poljubno porazdelitev, torej veljata tudi za to

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{(N-1)};\ \widehat{\sigma}^2 = \frac{\sum(X_i-\bar{X})^2}{n-1}\frac{N-1}{N}$$

Da bi izračunali interval zaupanja, potrebujemo oceno standardne napake, še prej pa oceno populacijske variance. Izračunajmo izraz $\frac{\sum(X_i-\bar{X})}{n-1}$ za naš primer (uporabimo, da velja $X_i^2 = X_i$):

$$\begin{aligned}
\frac{1}{n-1}\sum(X_i-\bar{X})^2 &= \frac{1}{n-1}\sum[X_i^2 - 2X_i\bar{X} + \bar{X}^2] = \frac{1}{n-1}[\sum X_i^2 - 2\sum X_i\bar{X} + \sum \bar{X}^2] \\
&= \frac{1}{n-1}[\sum X_i^2 - n\bar{X}^2] \\
&= \frac{1}{n-1}[\sum X_i - n\bar{X}^2] = \frac{1}{n-1}[n\bar{X} - n\bar{X}^2] \\
&= \frac{n}{n-1}[\hat{\pi} - \hat{\pi}^2] = \frac{n}{n-1}\hat{\pi}(1-\hat{\pi})
\end{aligned}$$

Mimogrede - zapomnimo si, da vedno velja

$$\sum(X_i-\bar{X})^2 = \sum X_i^2 - n\bar{X}^2$$

Če zanemarimo končnost populacije, je naša ocena populacijske variance torej enaka

$$\widehat{\sigma}^2 = \frac{n}{n-1}\hat{\pi}(1-\hat{\pi})$$

Ocena standardne napake je potem

$$\hat{SE}^2 = \frac{n}{n-1}\hat{\pi}(1-\hat{\pi})/n = \frac{1}{n-1}\hat{\pi}(1-\hat{\pi})$$

- $\hat{SE} = \sqrt{\frac{1}{n-1}\hat{\pi}(1-\hat{\pi})} = 0{,}03266$
- 95% interval zaupanja za populacijski delež je torej $0{,}12 \pm 1{,}96 \cdot 0{,}03266 = [0{,}056, 0{,}184]$.
- Naša ocena za število prodaj je torej $N\hat{\pi} = 960$, interval zaupanja pa [448,1472].
- Če izračune primerjamo s tistimi v prejšnji nalogi, se nam dobljeni interval zaupanja lahko zazdi precej velik.

Primer:

Kako velik vzorec potrebujemo, da z neko natančnostjo ocenimo delež v neskončni populaciji?

Kot vidimo, bo to odvisno od deleža, $SE = \frac{1}{n}\pi(1-\pi)$. Da bi našli maksimum te funkcije, odvajajmo po π, dobimo 0,5. Najtežje bo torej ocenjevati delež, kadar bo približno enak 0,5, bolj kot je delež blizu 0 oz. 1, manjšo standardno napako bomo imeli (manjši vzorec potrebujemo za enako natančno oceno).

1.4 Vzorčenje po skupinah - Rice, 7.6.

Populacija je pogosto razdeljena na podpopulacije (stratume) in temu želimo prirediti tudi načrt vzorčenja. Država je razdeljena na regije, podjetja so razdeljena na različne panoge, medicinske sestre lahko delimo na različna področja dela. Zanima nas lahko vsaka podpopulacija posebej, nato pa še celota. Tu se lahko pojavi več vprašanj:

- Do sedaj smo se ukvarjali s primerom, ko nas je zanimala vsaka podpopulacija posebej. Na primer, želeli smo oceniti povprečni strošek bolnika pri nekem zapletu v vseh slovenskih bolnišnicah, ki zdravijo tovrstne zaplete. Ker so bolnišnice seveda različnih velikosti in sprejmejo zelo različno število bolnikov letno, smo se vprašali ali je smiselno v vseh bolnišnicah zbrati enako velik vzorec, oziroma ali je potrebno v velikih bolnišnicah zbrati proporcionalno večji vzorec.

- Zdaj nas bo zanimalo nas bo tudi, kako povprečja podpopulacij združiti pri oceni skupnega povprečja in kakšno standardno napako dobimo v tem primeru.

- Če podatkov o vsaki podpopulaciji posebej ne potrebujemo, ni potrebno vzorčiti iz vsake in lahko se odločimo, da bomo najprej naključno izbrali nekaj podpopulacij, nato pa vzorčili iz vsake izmed njih. Na primer: zanima nas sposobnost logičnega mišljenja dijakov slovenskih osnovnih šol, radi bi jo preverjali z mednarodno primerljivim testom. Na prvi stopnji želimo izbrati nekaj šol, ki bodo sodelovale, nato pa na vsaki šoli še vzorčiti določeno število učencev. Tu nas bo spet zanimalo, koliko učencev naj vzamemo na posamezni šoli in seveda, koliko šol naj sploh izberemo, da bomo dobili željeno natančnost končnega odgovora.

V vseh primerih nas bodo torej zopet zanimala tri osnovna vprašanja:

- Kako združiti ocenjena povprečja v skupno, nepristransko oceno populacijskega povprečja.

- Kako razporediti vzorec oziroma kakšno standardno napako bomo dobili pri nekem načrtu vzorčenja. Ali je tak načrt vzorčenja boljši od naključnega vzorčenja?

- Kako oceniti standardno napako s pomočjo naših podatkov

Zamislimo si lahko več situacij (vzorčenje po nivojih je med seboj vedno neodvisno):

- Izberemo naključen vzorec iz vsakega stratuma (temu pravimo stratificirano vzorčenje, Rice 7.6.)
- Naključno izberemo le nekaj stratumov, populacija stratumov je lahko končna ali neskončna
- V vsakem stratumu je lahko končno ali neskončno enot

Vemo že, da nam končnost populacije prinese kar nekaj dodatnega dela, hkrati pa tudi, da popravki za končnost pri proporcionalno majhnemu vzorcu hitro postanejo zanemarljivi. Zato se bomo osredotočili na tri situacije:

- Stratificirano vzorčenje, v vsakem stratumu je končno enot
- Stratificirano vzorčenje, v vsakem stratumu je neskončno enot
- Naključen vzorec stratumov, stratumov je neskončno, enot v posameznem stratumu je neskončno

Primer

Začnimo s preprostim primerom. Zanima nas povprečni dosežek na mednarodnem testu matematike 12-letnikov v neki slovenski regiji. V regiji imamo 1000 12-letnikov, od tega jih je 700 v sedmem razredu, 300 pa v osmem.

Dosežek želimo oceniti s pomočjo vzorca, denimo, da smo testirali 50 učencev, od tega 40 sedmošolcev in 10 osmošolcev, povprečje sedmošolcev označimo z $\bar{X}_1$, povprečje osmošolcev pa z $\bar{X}_2$.

Kako bi s pomočjo povprečij obeh skupin zapisali nepristransko cenilko populacijskega povprečja?

$$\bar{X}_s = w_1\bar{X}_1 + w_2\bar{X}_2$$

Kakšne morajo biti uteži, da bo cenilka nepristranska? Izračunajmo njeno pričakovano vrednost, označimo populacijsko povprečno vrednost v sedmem razredu z μ_1, povprečno vrednost v osmem pa z μ_2:

$$E(\bar{X}_s) = w_1 E(\bar{X}_1) + w_2 E(\bar{X}_2) = w_1\mu_1 + w_2\mu_2$$

Da bi zagotovili nepristransko cenilko morata uteži biti torej enaki kot v populaciji, v našem primeru je $w_1 = 0{,}7$, $w_2 = 0{,}3$, dejansko število testiranem v vsakem stratumu za nepristranskost ocene ni pomembno.

. ---

Primer

Nadaljujmo s primerom, denimo, da se na nas (statistike) izvajalci testa obrnejo že pred izvedbo. Ker je strošek izvedbe preizkusa odvisen od velikosti vzorca, želijo izbrati najmanjši možni vzorec, ki bo zagotovil, da bo standardna napaka ocenjenega povprečja manj kot 2 točki (na preizkusu je možno zbrati od 0 do 100 točk).
Pričakujejo, da je povprečje v sedmem in osmem razredu nekoliko različno, prav tako so mnenja, da so razlike med učenci v sedmem razredu precej večje kot v osmem. Zanima jih, kako velik vzorec morajo vzeti in kako ga naj razporedijo vzorec, da bodo zadostili kriterijem glede standardne napake.

Intuitivno smiselno se zdi, da bi v vzorec izbrali proporcionalno enak delež učencev iz vsakega razreda kot je v populaciji. Kaj pa odvisnost od variabilnosti? Zamislimo si skrajni primer - da v osmem razredu variabilnosti ni (recimo, vsi pišejo 100%). Če je tako, je vsekakor nesmiselno izbrati več kot enega učenca iz osmega razreda, saj nam ne bo prinesel nobene nove informacije. Optimalni načrt vzorčenja je torej zagotovo odvisen tudi od variabilnosti.

. ---

Stratificirano vzorčenje, v vsakem stratumu je končno enot, Rice 7.6

Označimo velikost populacije z N, število stratumov pa z L. Število enot znotraj vsakega stratuma označimo z N_1, N_2, ... N_L. Povprečje v vsakem stratumu označimo z μ_i, $i = 1, \dots L$, varianco v vsakem stratumu pa z σ^2_{wi}. Posamezne populacijske vrednosti označimo z x_{ij}, pri čemer indeks i označuje stratum, indeks j pa posamezne enote iz tega stratuma.

Nepristranska cenilka: Najprej zapišimo populacijsko povprečje s pomočjo povprečij posameznih skupin:

$$\mu = \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N_i} x_{ij} = \frac{1}{N}\sum_{i=1}^{L}\mu_i N_i = \sum_{i=1}^{L}\mu_i w_i,$$

kjer je $w_i = \frac{N_i}{N}$.

Označimo z $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ vzorčno povprečje stratuma i. Pokažimo, da je

$$\bar{X}_s = \sum_{i=1}^{L} w_i \bar{X}_i$$

nepristranska cenilka populacijskega povprečja:

$$E(\bar{X}_s) = \sum_{i=1}^{L} w_i E(\bar{X}_i) = \sum_{i=1}^{L} w_i \mu_i = \mu.$$

Varianca cenilke: Sedaj izračunajmo varianco te cenilke:

$$\mathrm{var}(\bar{X}_s) = \mathrm{var}\left(\sum_{i=1}^{L} w_i \bar{X}_i\right) = \sum_{i=1}^{L} w_i^2 \mathrm{var}\left(\bar{X}_i\right)$$

Ker so ocene povprečij v posameznih skupinah med seboj neodvisne, smo lahko varianco vsote zapisali kot vsoto varianc. Znotraj vsake skupine imamo opravka z naključnim vzorčenjem (končna populacija, brez ponavljanja), zato že poznamo varianco posameznih vzorčnih povprečij:

$$\mathrm{var}(\bar{X}_s) = \sum_{i=1}^{L} w_i^2 \frac{\sigma^2_{wi}}{n_i}\frac{N_i - n_i}{N_i - 1}$$

Stratificirano vzorčenje, v vsakem stratumu je neskončno enot (ni v Rice-u)

Vzemimo, da poznamo delež enot v vsakem stratumu v populaciji (če ga ne poznamo in ga ocenjujemo iz podatkov, je to enostavno vzorčenje, saj nimamo nobene dodatne informacije - ne moremo zmanjšati SE).

Ker je enot neskončno, moramo zadeve nekoliko drugače zapisati. Na primer takole, vsak element iz populacije zapišemo kot:

$$Y = \mu + X + \epsilon$$

Pri tem naj μ označuje celotno povprečje, X je slučajna spremenljivka, ki ima toliko različnih vrednosti x_i, $i = 1, \ldots, L$, kolikor je stratumov, vrednosti $\mu + x_i$ so povprečja posameznih stratumov. Ker smo povprečje vseh enot označili z μ, mora veljati $E(X) = E(\epsilon) = 0$. Elementi skupine i naj se okrog svojega povprečja porazdeljujejo z varianco $\text{var}(\epsilon|X = x_i) = \sigma^2_{wi}$, ki je za vsak i lahko drugačna. Vrednosti x_i so realizacije neke slučajne spremenljivke X, označimo $P(X = x_i) = \pi_i$, pri tem je število različnih vrednosti x_i enako L, posamezna odstopanja pa naj bodo taka, da je $E(X) = \sum \pi_i x_i = 0$.

Posamezne enote v vzorcu so slučajne spremenljivke, zapišemo jih kot

$$Y_{ij} = \mu + x_i + \epsilon_{ij},$$

pri tem indeks i označuje stratum $i = 1, \ldots, L$, indeks j pa število enot v vzorcu iz vsakega stratuma, $j = 1, \ldots, n_i$.

Nepristranska cenilka Na analogen način kot pri končni populaciji definiramo cenilko povprečja:

$$\bar{Y}_s = \sum_{i=1}^{L} \pi_i \bar{Y}_{i\cdot},$$

pri čemer naj nam $\bar{Y}_{i\cdot}$ označuje vzorčno povprečje v i skupini (povprečje vseh n_i enot vzorčenih iz te skupine). Pričakovana vrednost enot v tej skupini je $E(Y_{ij}) = \mu + x_i$. (Pri tem poudarimo, da indeks i vedno označuje isto, i-to skupino.)

Da bi pokazali, da je cenilka nepristranska, moramo najprej vedeti, kako se s povprečji skupin zapiše populacijsko povprečje:

- Uporabimo enakost $E(Y) = E(E(Y|X))$

- Najprej izračunajmo vrednost znotraj oklepajev. $E(Y|X) = \mu + X$
- $\mu = E(Y) = E(\mu + X) = \sum \pi_i(\mu + x_i)$

Pokažimo, da je ta cenilka nepristranska:

$$E(\bar{Y}_s) = \sum_{i=1}^{L} \pi_i E(\bar{Y}_{i\cdot}) = \sum_{i=1}^{L} \pi_i(\mu + x_i).$$

Napaka cenilke Da bi izračunali varianco te cenilke, tako kot prej potrebujemo varianco v posamezni skupini. Znotraj neke skupine so slučajne le vrednosti ϵ, njihovo varianco v posamezni skupini smo označili s σ^2_{wi}. Zapišimo še s formulo:

$$\text{var}(Y_{ij}|X_i = x_i) = \text{var}(\mu + x_i + \epsilon|X = x_i) = \text{var}(\epsilon|X = x_i) = \sigma^2_{wi}$$

in zato $\text{var}(\bar{Y}_{i\cdot}) = \frac{\sigma^2_{wi}}{n_i}$. Zato velja

$$\text{var}(\bar{Y}_s) = \text{var}\left(\sum_{i=1}^{L} \pi_i \bar{Y}_{i\cdot}\right) = \sum_{i=1}^{L} \pi_i^2 \frac{\sigma^2_{wi}}{n_i}$$

Ocena standardne napake: V izrazu za standardno napako nastopa σ_{wi}, če je ne poznamo, jo moramo za izračun intervala zaupanja oceniti iz podatkov. Ker v vsakem stratumu jemljemo naključni vzorec, jo seveda lahko ocenimo na enak način, kot smo že pokazali za neskončno populacijo.

Vzemimo sedaj, da so variance znotraj vseh stratumov enake ($\sigma^2_{wi} = \sigma^2_w$ za vsak i). Intuitivno se zdi smiselno, da σ^2_{wi} ocenimo v vsakem stratumu posebej in nato vzamemo povprečje teh ocen.

Zapišimo cenilko

$$\widehat{\sigma}^2_w = \frac{1}{L} \sum_{i=1}^{L} \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$$

in pokažimo, da je nepristranska. Uporabimo izrek , ki smo si ga pripravili v razdelku 1.2.1, tako da nadomestimo X_i z Y_{ij}:
Izrek: X_i neodv, $E(X_i) = \mu_i$, $E(\bar{X}) = \mu$, $\text{var}(X_i) = \sigma^2$

Mi: Y_{ij} neodv, $E(Y_{ij}) = \mu + x_i$, $E(\bar{Y}_{i.}) = \mu x_i$, $\mathrm{var}(Y_{ij}) = \sigma_w^2$

Dobimo

$$E[(Y_{ij} - \bar{Y}_{i.})^2] = 0 + \sigma_w^2 \frac{n_i - 1}{n_i}$$

$$\begin{aligned} E(\hat{\sigma}_w^2) &= \frac{1}{L} \sum_{i=1}^{L} \frac{1}{n_i - 1} \sum_{j=1}^{n_i} E\left[\left(Y_{ij} - \bar{Y}_{i.} \right)^2 \right] \\ &= \frac{1}{L} \sum_{i=1}^{L} \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \sigma_w^2 \frac{n_i - 1}{n_i} \\ &= \frac{1}{L} \sum_{i=1}^{L} \frac{1}{n_i - 1} n_i \sigma_w^2 \frac{n_i - 1}{n_i} \\ &= \frac{1}{L} \sum_{i=1}^{L} \sigma_w^2 = \sigma_w^2 \end{aligned}$$

Če to cenilko vstavimo v formulo za standardno napako, dobimo nepristransko oceno standardne napake.

Naključen vzorec stratumov, stratumov je neskončno (ni v Rice-u)

Sedaj vzemimo še, da je stratumov neskončno. Pri tem bomo ponovili postopke iz primera s stratificiranim vzorčenjem, z eno veliko razliko. Skupin (stratume) v populaciji ne moremo več preprosto prešteti oz. označiti z $1, \ldots, L$. i-ta skupina v vzorcu torej ni tudi i-ta skupina v populaciji - vsakič ko vzorčimo bo i-ta skupina pomenila nekaj drugega.

Primer

Za lažjo predstavo začnimo s primerom. Ocenili bi radi vrednost hemoglobina pri (nedopingiranih) vzdržljivostnih športnikih. Vemo, da hemoglobin pri športnikih ni ves čas enak, temveč niha okrog nekega posameznikovega povprečja. Prav tako povprečja posameznikov niso enaka, nekateri posamezniki imajo nižji, drugi višji povprečni hemoglobin. Oceniti želimo povprečno vrednost v populaciji, v vzorec vzamemo 10 športnikov, pri vsakem opravimo 5 meritev.
Populacijsko povprečje i-tega športnika je sedaj slučajna spremenljivka, saj je na i-tem mestu v vzorcu vsakič drug posameznik.

.

Slučajno spremenljivko Y zapišimo enako kot pri končnem številu stratumov,

$$Y = \mu + X + \epsilon,$$

posamične vrednosti na vzorcu označimo kot

$$Y_{ij} = \mu + X_i + \epsilon_{ij},$$

namesto pogostosti posameznih vrednosti, nas bo sedaj zanimala varianca spremenljivke X, označimo jo z σ_b^2.
Zaradi preprostosti izpeljav predpostavimo, da je varianca pri vseh enotah i enaka: $\text{var}(\epsilon_{ij}) = \sigma_w^2$. Ker smo vzorčili neodvisno, velja: $\text{cov}(\epsilon_{ij}, \epsilon_{ik}) = 0$, $\text{cov}(\epsilon_{ij}, X_i) = 0$, $\text{cov}(\epsilon_{ij}, \epsilon_{lk}) = 0$

Bistvena razlika od prejšnjega primera je, da sedaj indeks i ne pomeni vedno skupine z istim povprečjem, temveč je na i-tem mestu vsakič druga skupina. Poglejmo, kako to spremeni izpeljave, denimo, da v vzorec izberemo $i = 1, \ldots, k$ skupin:

Nepristranska cenilka Zapišimo cenilko povprečja kot

$$\bar{Y}_s = c \sum_{i=1}^{k} \bar{Y}_{i\cdot}$$

radi bi določili vrednost c, da bo cenilka nepristranska.

Premislimo nepristranskost najprej intuitivno: v vsaki skupini smo vzeli naključen vzorec, njegovo povprečje je nepristranska cenilka povprečja tiste skupine. Po drugi strani je naključen tudi vzorec skupin, zato je tudi pričakovana vrednost povprečja spet nepristranska.

Velja

$$\bar{Y}_{i\cdot} = \mu + X_i + \frac{1}{n_i} \sum_{j=1}^{n_i} \epsilon_{ij},$$

zato $E(\bar{Y}_{i\cdot}) = \mu$ in

$$E(\bar{Y}_s) = cE(\sum_{i=1}^{k} \bar{Y}_{i\cdot}) = \mu ck$$

Velja torej $c = \frac{1}{k}$.

Poglejmo še malo bolj splošno.

Zapišimo cenilko povprečja kot

$$\bar{Y}_s = \sum_{i=1}^{k} c_i \bar{Y}_{i\cdot}$$

radi bi določili vrednosti c_i, da bo cenilka nepristranska.

Velja

$$\bar{Y}_{i\cdot} = \mu + X_i + \frac{1}{n_i} \sum_{j=1}^{n_i} \epsilon_{ij},$$

zato $E(\bar{Y}_{i\cdot}) = \mu$ in

$$E(\bar{Y}_s) = E(\sum_{i=1}^{k} c_i \bar{Y}_{i\cdot}) = \mu \sum_{i=1}^{k} c_i$$

Zamislimo si torej lahko več nepristranskih cenilk, bistveno je, da je vsota vrednosti c_i enaka 1. Smiselno bo torej izbrati vrednosti c_i tako, da bo standardna napaka minimalna, če so vzorci iz skupin različnih velikosti, variance v skupinah pa enake, bomo npr. vzeli uteži proporcionalne velikosti vzorcev. Za nadaljnje izpeljevanje zaradi preprostosti vzemimo, da so vrednosti c_i enake ne glede na indeks, torej $c_i = \frac{1}{k}$.

Imamo torej nepristransko cenilko, zanima nas njena varianca. Vemo, da imajo meritve iz posamezne skupine neko skupno lastnost in so tako med seboj odvisne. Oglejmo si kakšna je ta odvisnost.

Oglejmo si, kakšna je povezanost med meritvami:

Primer

Vemo, da posameznikove vrednosti nihajo okrog njegovega povprečja. Denimo, da želimo oceniti povprečno vrednost hemoglobina nekega športnika. Vzorčimo n neodvisnih vrednosti tega športnika. S pomočjo teh vrednosti ocenjujemo $\mu_i = \mu + x_i$, ker so neodvisne, je kovarianca poljubnega para vrednosti enaka 0.

Denimo, da posameznikove vrednosti uporabimo za ocenjevanje skupnega povprečja, μ. Poleg tega posameznika jih zberemo še nekaj, tako da imamo skupno k posameznikov, pri vsakem n meritev. Meritve pri istem posamezniku niso več neodvisne, saj imajo neko skupno lastnost (vrednost X_i). To se lahko prepričamo tudi takole: denimo, da vemo, da je populacijsko povprečje hemoglobina enako 148:

- Vzemimo dva posameznika. Meritev prvega je 132, kakšna je naša napoved za meritev drugega? Naša najboljša napoved je še vedno 148, saj nam vrednost prvega ne pove ničesar o vrednosti drugega, meritvi sta neodvisni; $\text{cov}(Y_{ij}, Y_{lm}) = 0$.

- Vzemimo dve meritvi istega posameznika. Prva meritev je enaka 132, kakšna je naša napoved za drugo meritev? Naša najboljša napoved je sedaj 132 in ne več 148. Meritvi sta torej odvisni; $\text{cov}(Y_{ij}, Y_{il}) \neq 0$.

.

Najprej izračunajmo kovarianco med dvema meritvama $i$tega posameznika v našem vzorcu (upoštevamo, da je kovarianca s konstanto enaka 0):

$$\begin{aligned}\operatorname{cov}(Y_{ij}, Y_{il}) &= \operatorname{cov}[\mu + X_i + \epsilon_{ij}, \mu + X_i + \epsilon_{il}] \\ &= \operatorname{cov}[\mu + X_i + \epsilon_{ij}, \mu + X_i + \epsilon_{il}] \\ &= \operatorname{cov}[X_i, X_i] + \operatorname{cov}[X_i, \epsilon_{il}] + \operatorname{cov}[\epsilon_{ij}, X_i] + \operatorname{cov}[\epsilon_{ij}, \epsilon_{il}] \\ &= \sigma_b^2\end{aligned}$$

Kovarianca ni enaka 0, je pozitivna, kar je v skladu z našimi pričakovanji (če je prva vrednost visoka sklepamo, da gre za posameznika z visokim povprečjem in podobno visoko napovemo tudi drugo vrednost)

Pri kovarianci interpretiramo le predznak (oz. ali je enaka 0), da bi jo lažje interpretirali zato izračunajmo korelacijo:

$$\operatorname{cor}(Y_{ij}, Y_{il}) = \frac{\operatorname{cor}(Y_{ij}, Y_{il})}{\sqrt{\operatorname{var}(Y_{ij})}\sqrt{\operatorname{var}(Y_{il})}}$$

Za izračun korelacije torej potrebujemo varianco meritev:

$$\begin{aligned}\operatorname{var}(Y_{ij}) &= \operatorname{var}\left[\mu + X_i + \epsilon_{ij}\right] = \operatorname{cov}\left[\mu + X_i + \epsilon_{ij}, \mu + X_i + \epsilon_{ij}\right] \\ &= \operatorname{cov}\left[\mu + X_i + \epsilon_{ij}, \mu + X_i + \epsilon_{ij}\right] \\ &= \operatorname{cov}\left[X_i, X_i\right] + \operatorname{cov}\left[X_i, \epsilon_{ij}\right] + \operatorname{cov}\left[\epsilon_{ij}, X_i\right] + \operatorname{cov}\left[\epsilon_{ij}, \epsilon_{ij}\right] \\ &= \sigma_b^2 + \sigma_w^2\end{aligned}$$

In zato

$$\begin{aligned}\operatorname{cor}(Y_{ij}, Y_{il}) &= \frac{\operatorname{cor}(Y_{ij}, Y_{il})}{\sqrt{\operatorname{var}(Y_{ij})}\sqrt{\operatorname{var}(Y_{il})}} \\ &= \frac{\sigma_b^2}{\sqrt{\sigma_b^2 + \sigma_w^2}\sqrt{\sigma_b^2 + \sigma_w^2}} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}\end{aligned}$$

Poizkusimo interpretirati gornjo formulo:

- Naj bodo povprečja vseh posameznikov enaka. Pote je $\sigma_b = 0$ in zato korelacija enaka 0. Če so povprečja vseh posameznikov enaka, nam informacija o prvi meritvi nekega posameznika ne pove ničesar o drugi

- Naj bodo vse meritve istega posameznika enake. Potem je $\sigma_w = 0$ in zato korelacija enaka 1. Če poznamo eno meritev posameznika, poznamo tudi vse ostale, saj so enake.

- Bolj ko so si posamezniki različni (v primerjavi z variabilnostjo znotraj posameznika), večja je korelacija med meritvami istega posameznika.

Napaka cenilke: Medtem ko pri nepristranskosti cenilke ni bilo bistvenih razlik glede na dosedanje primere, bo dejstvo, da sedaj vsakič izberemo druge skupine, igralo zelo pomembno vlogo.

Pri izračunu variance povprečja v vsaki skupini dobimo (naj bo število meritev pri vsaki enoti enako n):

$$\begin{aligned}
\mathrm{var}(\bar{Y}_{i\cdot}) &= \mathrm{var}\left[\mu + X_i + \frac{1}{n}\sum_{j=1}^{n}(\epsilon_{ij})\right] \\
&= \mathrm{var}(X_i) + \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{var}(\epsilon_{ij}) \\
&= \sigma_b^2 + \frac{\sigma_w^2}{n}
\end{aligned}$$

Upoštevamo, da so povprečja za različne i med seboj neodvisna

$$\begin{aligned}
\mathrm{var}(\bar{Y}_s) &= \mathrm{var}\left(\frac{1}{k}\sum_{i=1}^{k}\bar{Y}_{i\cdot}\right) \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\left[\sigma_b^2 + \frac{\sigma_w^2}{n}\right] \\
&= \frac{\sigma_b^2}{k} + \frac{1}{k^2}\sum_{i=1}^{k}\left[\frac{\sigma_w^2}{n}\right] \\
&= \frac{\sigma_b^2}{k} + \frac{\sigma_w^2}{nk}
\end{aligned}$$

Interpretirajmo formulo:

- Varianca cenilke je odvisna od variance med posamezniki in variance znotraj posameznikov.

- Glavni del standardne napake bo odvisen od variance med posamezniki, ki je deljena le s k.

- Število posameznikov ter število meritev pri vsakem posamezniku določimo glede na to, kaj vse nas zanima (samo skupno povprečje, ali še kaj drugega), glede na pričakovano razmerje med variancama ter glede na ceno/težavnost pridobivanja podatkov (običajno je težje zbrati več posameznikov, lažje pa več enot znotraj posameznika).

- Če sta varianci približno enake velikosti, bomo najmanjšo standardno napako dobili, kadar bomo iz vsake skupine vzeli le enega posameznika ($n = 1$). Če so enote med seboj neodvisne (vsaka ima drugačno vrednost X, indeksa j ne potrebujemo), velja

  $$\text{var}(Y_i) = \text{var}(X_i) + \text{var}(\epsilon_i) = \sigma_b^2 + \sigma_w^2$$

  in zato

  $$\text{var}(\bar{Y}) = \frac{\sigma_b^2 + \sigma_w^2}{n}$$

- Vidimo, da je pri tej vrsti vzorčenja bistveno upoštevati odvisnost med podatki (od istega posameznika). Če je ne upoštevamo, uporabimo napačno formulo za standardno napako - mislimo, da ocenjujemo bolj natančno kot je res (interval zaupanja ne bo pokril μ s pravo verjetnostjo). V skrajnem primeru bi lahko vse podatke vzeli od istega posameznika in ocenjevali, da je standardna napaka bistveno manjša kot je v resnici (ne bi se sploh zavedali, da obstaja variabilnost med posamezniki). Naša ocena bi bila v tem primeru pravilna le ,če razlik med posamezniki zares ni (in je $\sigma_b = 0$).

Ocena varianc: Za konec si poglejmo še, kako standardno napako ocenimo iz vzorca. V ta namen moramo seveda imeti ocene varianc med in znotraj posameznikov. Če so variance znotraj posameznikov različne, jih lahko ocenimo za vsakega posameznika posebej tako kot pri slučajnem vzorcu. Vzemimo, da so pri vseh posameznikih enake - združili bi radi informacijo vseh posameznikov v eno oceno.

Izpeljemo oceno variance σ_w^2 znotraj ene skupine na podlagi vzorca velikosti n (enako kot v primeru končnega števila stratumov):

$$\begin{aligned}
E[\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i.})^2] &= E[\sum_{j=1}^{n}Y_{ij}^2 - n\bar{Y}_{i.}^2] \\
&= \sum_{j=1}^{n}E[Y_{ij}^2] - nE[\bar{Y}_{i.}^2] \\
&= \sum_{j=1}^{n}[\mathrm{var}(Y_{ij}) + E(Y_{ij})^2] - n[\mathrm{var}(\bar{Y}_{i.}) + E(\bar{Y}_{i.})^2] \\
&= \sum_{j=1}^{n}[\mathrm{var}(Y_{ij}) + E(Y_{ij})^2] - n[\mathrm{var}(\bar{Y}_{i.}) + E(\bar{Y}_{i.})^2] \\
&= \sum_{j=1}^{n}[\sigma_b^2 + \sigma_w^2 + \mu^2] - n[\sigma_b^2 + \frac{\sigma_w^2}{n} + \mu^2] \\
&= n[\sigma_b^2 + \sigma_w^2] - n[\sigma_b^2 + \frac{\sigma_w^2}{n}] \\
&= n\sigma_w^2[1 - \frac{1}{n}] = \sigma_w^2(n-1)
\end{aligned}$$

Zapišimo cenilko kot

$$\hat{\sigma}_w^2 = c\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i.})^2$$

Velja

$$\begin{aligned}
E(\hat{\sigma}_w^2) &= c\sum_{i=1}^{k}E\sum_{j=1}^{n}[(Y_{ij}-\bar{Y}_{i.})^2] \\
&= c\sum_{i=1}^{k}\sigma_w^2(n-1) \\
&= ck\sigma_w^2(n-1)
\end{aligned}$$

Konstanta je torej enaka $c = \frac{1}{k(n-1)}$.

Na enak način izpeljimo še oceno za varianco med skupinami, sedaj vlogo X_i igra $\bar{Y}_{i.}$:

Izrek: X_i neodv, $E(X_i) = \mu_i$, $E(\bar{X}) = \mu$, $\text{var}(X_i) = \sigma^2$
Mi: $\bar{Y}_{i.}$ neodv, $E(\bar{Y}_{i.}) = \mu$, $E(\bar{Y}_{..}) = \mu$, $\text{var}(\bar{Y}_{i.}) = \sigma_b^2 + \frac{\sigma_w^2}{n}$

Dobimo

$$E\sum_{i=1}^{k}[(\bar{Y}_{i.} - \bar{Y}_{..})^2] = 0 + \sigma_b^2(k-1) + \frac{\sigma_w^2}{n}(k-1)$$

Zapišimo cenilko kot

$$\hat{\sigma}_b^2 = c\sum_{i=1}^{k}(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

Velja

$$\begin{aligned} E(\hat{\sigma}_b^2) &= ck[\sigma_b^2\frac{k-1}{k} + \frac{\sigma_w^2}{n}\frac{k-1}{k}] \\ &= c[\sigma_b^2(k-1) + \frac{\sigma_w^2(k-1)}{n}] \end{aligned}$$

Vidimo, da pričakovana vrednost tega izraza ni odvisna le od σ_b, zato cenilko zapišemo takole:

$$\hat{\sigma}_b^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\bar{Y}_{i.} - \hat{Y}_{..})^2 - \frac{\hat{\sigma}_w^2}{n}$$

Za konec zapišimo še nepristransko cenilko za SE^2:

$$\begin{aligned} \widehat{SE}^2 &= \frac{\hat{\sigma}_b^2}{k} + \frac{\hat{\sigma}_w^2}{nk} \\ &= \frac{\hat{\sigma}_w^2}{nk} + \frac{1}{k}\frac{1}{k-1}\sum_{i=1}^{k}[(\bar{Y}_{i.} - \hat{Y}_{..})^2] - \frac{\hat{\sigma}_w^2}{nk}] \\ &= \frac{1}{k(k-1)}\sum_{i=1}^{k}(\bar{Y}_{i.} - \hat{Y}_{..})^2 \end{aligned}$$

Povzetek

- Stratificirano vzorčenje: standardno napako (in s tem intervale zaupanja) lahko zmanjšamo, če imamo dodatne informacije - vemo, da so v skupinah različne variance, vemo, da so skupine različno močno zastopane v populaciji. Izkaže se, da si največji 'prihranek' zagotovimo, če poznamo velikosti skupin, nekoliko manj pa lahko oceno izboljšamo še, če poznamo variance (ponavadi tega tako ali tako ne poznamo).

- Vzorčenje po skupinah: bistveno je predvsem, da razumemo, kako so podatki odvisni med seboj, sicer bo ocena napake napačna. Vzorčenje po skupinah ima tu večjo std. napako kot če iz vsake skupine vzamemo po enega. Ker nimamo dodatne informacije, ne moremo zmanjšati SE, pomembno pa je, da jo navkljub odvisnosti v podatkih izračunamo pravilno.

Poglavje 2

Ocenjevanje parametrov, Rice 8

V poglavju o vzorčenju smo se ukvarjali z ocenjevanjem populacijskega povprečja in variance. Naravni cenilki za ti dve količini sta vzorčno povprečje in vzorčna varianca. Populacijsko povprečje in varianca določata normalno porazdelitev, kadar populacija ni normalno porazdeljena, nas bodo lahko zanimali tudi drugi parametri. Oglejmo si nekaj primerov različnih porazdelitev in njihovih parametrov, ki nas lahko zanimajo.

2.1 Primeri uporabe različnih porazdelitev

Poissonova porazdelitev:

- Štetje dogodkov v času ali prostoru
- Pogostost dogodkov je konstantna v času ali prostoru (ni najprej več okvar, kasneje manj, potem spet več)
- Dogodki v posameznih intervalih so neodvisni
- ni večkratnih dogodkov
- Primeri: prihajanje strank/telefonskih klicev/avtomobilov na semafor/zahtevkov v zavarovalnici ...

Če gornji pogoji držijo za naš primer, lahko predpostavimo, da je slučajna spremenljivka porazdeljena s Poissonovo porazdelitvijo. Vendar pa samo

oblika porazdelitve ni dovolj, da bi lahko računali poljubne verjetnosti, moramo poznati tudi parametre porazdelitve. Spomnimo se:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},\ k = 0,1,2,\dots$$

Parameter λ v našem primeru želimo oceniti iz podatkov. Torej, opravimo nekaj meritev in zberemo podatke. V tem poglavju bomo vedno predpostavljali, da smo zbrali naključen vzorec, torej da so $X_i,\ i = 1,\dots,n$ i.i.d spremenljivke.
Kako pa λ ocenimo s pomočjo podatkov? Vemo, da velja

$$E(X) = \lambda.$$

Cenilka za λ je torej lahko vzorčno povprečje, $\hat{\lambda} = 1/n \sum X_i$. Seveda vemo, da bomo na vsakem vzorcu dobili drugačno povprečje, zato je naša cenilka $\hat{\lambda}$ slučajna spremenljivka. Zanimala nas bo kvaliteta te cenilke, morda bi radi zapisali interval zaupanja. V ta namen potrebujemo vzorčno porazdelitev naše cenilke. Porazdelitev vzorčnega povprečja že poznamo, lahko ga aproksimiramo z normalno porazdelitvijo.
Postopali bi lahko tudi drugače, vemo namreč, da je $\mathrm{var}(X) = \lambda$ - bi morda λ ocenili z varianco? Bi bila taka ocena boljša (v kakšnem smislu)? Vsekakor bi bila malce drugačna, saj povprečje in varianca v splošnem ne bosta enaki.

Kaj lahko rečemo o lastnostih predlaganih cenilk za Poissonovo porazdelitev? Je taka cenilka nepristranska, torej ali velja $E(\hat{\lambda}) = \lambda$? Ali lahko to trdimo vsaj v neskončnosti (torej da konvergira k pravi vrednosti, doslednost)? Je ocena tudi učinkovita, torej kakšna je njena varianca (nepristranska ocena z zelo veliko standardno napako ne bo preveč koristna)?

Kaj pa če sta povprečje in varianca zelo različna? To je znak, da z našimi predpostavkami nekaj ni v redu in da je Poissonova porazdelitev zelo slab model za naše podatke. Tudi v splošnem bo torej vedno nujno vprašanje prileganja k neki porazdelitvi. S tem ko smo ocenili parameter, namreč upamo, da smo našli tisto izmed Poissonovih porazdelitev, ki se na nek način najbolje prilega našim podatkom, a kaj če Poissonova porazdelitev sploh ni smiselen model za naše podatke - o tem iz same ocene ne izvemo nič. O tem bomo govorili kasneje - v razdelku o prileganju.

Poglejmo si še nekaj primerov:

- Zanima nas čas med dvema dogodkoma - opišemo npr. z eksponentno porazdelitvijo

- Neka merjena spremenljivka je prispevek mnogih majhnih naključnih vplivov - predpostavimo normalno porazdelitev
- Na mnogih področjih so že pokazali, da se določene pojave da sorazmerno dobro predstaviti z neko porazdelitvijo, glej tudi Rice ...
- Zelo obširno poglavje so tudi modeli. Osnoven primer je npr. linearna regresija, predpostavimo, da je spremenljivka normalna, povprečje odvisno od neke kovariate X, $Y \sim N(\beta_0 + \beta X, \sigma^2)$. Ne bo nas zanimalo torej le povprečje, temveč tudi nek parameter β.

Večkrat bo iskanje smiselne cenilke dokaj preprosta naloga. Normalno porazdelitev npr. določata dva parametra, za oba smo v prejšnjem poglavju že našli nepristranski cenilki. Seveda pa bo še več kompleksnih primerov, ko intuicija pri iskanju cenilke ne bo zadostovala.

Povzemimo torej, kaj bi radi:

- Imeli nek generičen pristop k ocenjevanju parametrov
- Vedeli nekaj o lastnostih, torej ali je ocena nepristranska oziroma vsaj dosledna
- Presojali kvaliteto cenilke s pomočjo variance (oziroma srednja kvadratna napaka)
- Vedeli kaj o porazdelitvi cenilke
- Preverili tudi kvaliteto prileganja

Ponovimo formalno izrazoslovje, ki smo ga že uporabljali v prejšnjem razdelku, in dodajmo nekaj novih izrazov:
Omejili se bomo na naslednjo situacijo: predpostavljamo, da so opazovane vrednosti kot slučajne spremenljivke $X_1, \ldots, X_n$ neodvisne in enako porazdeljene, porazdelitev je lahko odvisna od nekih parametrov.

- Opazovane vrednosti: vrednosti, ki jih dejansko vidimo na nekem vzorcu. Označimo z $x_1, \ldots, x_n$
- Slučajne spremenljivke: $X_1, \ldots, X_n$, abstraktni opis 'nastajanja' opazovanih vrednosti

- Naboru vseh možnih vrednosti parametrov pravimo prostor parametrov, označimo s Θ.
  Primer: $X_i \sim \Gamma(a, \lambda),\ \Theta = \{(a, \lambda); a > 0, \lambda > 0\}$

- Ocena parametra (estimate) bo neka funkcija opazovanih vrednosti $\hat{\theta}(x_1, \ldots, x_n)$

- Funkciji $\hat{\theta}$ oziroma slučajni spremenljivki $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ pravimo cenilka (estimator)

- Vzorčna porazdelitev je porazdelitev slučajne spremenljivke $\hat{\theta}$

- Nepristranska cenilka je taka, pri kateri je pričakovana vrednost vzorčne porazdelitve enaka parametru, torej $E(\hat{\theta}) = \theta_0$, kjer θ_0 označuje pravo vrednost, $\hat{\theta}$ pa ocenjeno na vzorcu.

- Standardna napaka: $\sqrt{\text{var}(\hat{\theta})}$

- Kvaliteto cenilke bomo presojali tudi s pomočjo srednje kvadratne napake (mean square error): $E[(\theta_0 - \hat{\theta})^2]$. Uporabimo, da velja $\text{var}(X) = E(X^2) - E(X)^2$:

$$\begin{aligned} E[(\theta_0 - \hat{\theta})^2] &= \text{var}[(\theta_0 - \hat{\theta})] + (E[(\theta_0 - \hat{\theta})])^2 \\ &= \text{var}[\hat{\theta}] + [\theta_0 - E(\hat{\theta})]^2 \end{aligned}$$

  Srednja kvadratna napaka nam pove kako daleč lahko pričakujemo, da bo ocenjena vrednost od prave (povprečna razdalja ocenjene vrednosti od prave). Srednjo kvadratno napako lahko zapišemo kot vsoto dveh izrazov - variance cenilke (standardna napaka na kvadrat) in kvadrata pristranosti (bias). Če je cenilka nepristranska, je srednja kvadratna enaka varianci cenilke.

Ponovimo še nekoliko splošnješo definicijo intervala zaupanja
Naj vektor X predstavlja vzorčne vrednosti $X_1, \ldots X_n$. $(1 - \alpha)\%$ interval zaupanja za populacijski parameter θ je interval $[L(X), U(X)]$, za katerega velja

$$P(L(X) \leq \theta \leq U(X)] = 1 - \alpha$$

In takoj dodajmo še opombo - parameter θ je seveda fiksen, verjetnostna trditev se tako nanaša na zgornjo in spodnjo mejo, ki sta slučajni spremenljivki.

Primer: Ravnotežje Hardy-Weinberg

Vsak gen ima dva alela, možne so 3 kombinacije: AA, Aa, aa. Kadar so v ravnovesju (vzorčimo naključno iz neke populacije), so verjetnosti kombinacij enake: θ^2, $2\theta(1-\theta)$, $(1-\theta)^2$. Imamo podatke za nek vzorec žensk iz slovenske populacije ($n = 1000$):

Tabela 2.1

	AA	Aa	aa
Število	83	428	489

Kako bi ocenili parameter θ? Najlaže ga seveda ocenimo kot (n_1 je število AA)

$$\hat{\theta} = \sqrt{\frac{n_1}{n}} = \sqrt{\frac{83}{1000}} = 0{,}288$$

Kako natančna je ta ocena? Je ocena nepristranska? Kakšna je standardna napaka te cenilke? Intuicija nam pravi, da ta ocena ne bo optimalna, saj nismo izrabili vse informacije, ki jo imamo v podatkih - uporabili smo le število pojavitev variante AA.

.

2.2 Metoda momentov

Metoda momentov je preprosta metoda, ki pa žal pogosto da cenilke, ki niso optimalne, zato te metode v praksi ne uporabljamo pogosto. Uporabna je predvsem kot neka začetna možnost, kadar so druge metode prezahtevne. Mi si jo bomo pogledali predvsem zato, da bomo bolje razumeli kaj želimo in zakaj je metoda največjega verjetja tako uporabna.

K-ti moment neke porazdelitve (če obstaja, torej če $E(|X^k|) < \infty$) je definiran kot

$$\mu_k = E(X^k)$$

Naj bodo $X_1, \ldots, X_n$ i.i.d spremenljivke, potem je k-ti moment na vzorcu definiran kot

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

K-ti moment na vzorcu je cenilka k-tega momenta v populaciji. Ideja metode momentov je naslednja:
parametre populacije izrazimo z momenti najmanjšega možnega reda, da dobimo ustrezne cenilke za populacijske parametre potem momente nadomestimo z vzorčnimi ocenami.

Primer:

Poglejmo, kaj dobimo v primeru normalne porazdelitve. Radi bi ocenili dva parametra, μ in σ.

$$\mu = E(X) = \mu_1,\ \sigma^2 = E(X^2) - E(X)^2 = \mu_2 - \mu_1^2$$

Cenilki bosta torej

$$\hat{\mu} = \frac{1}{n}\sum X_i,\ \widehat{\sigma}^2 = \frac{1}{n}\sum X_i^2 - (\frac{1}{n}\sum X_i)^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$$

Kot vidimo, je cenilka za populacijsko povprečje kar vzorčno povprečje, za populacijsko varianco pa kar varianca vzorca. Vemo že, da je slednja ocena pristranska. Je pa cenilka dosledna (glej prejšnji razdelek).

.

Primer:

Vzemimo porazdelitev gama, $X \sim \Gamma(a, \lambda)$. Gostota te spremenljivke je $f_X(x) = \frac{\lambda^a e^{-\lambda x} x^{a-1}}{\Gamma(a)}$, iz predmeta Verjetnost vemo, da je $E(X) = \frac{a}{\lambda}$ in $\mathrm{var}(X) = \frac{a}{\lambda^2}$.

Ker velja $\text{var}(X) = \frac{E(X)}{\lambda}$, dobimo

$$\begin{aligned} \lambda &= \frac{E(X)}{\text{var}(X)} \\ \lambda &= \frac{\mu_1}{\mu_2 - \mu_1^2} \\ a &= E(X)\lambda = \frac{\mu_1^2}{\mu_2 - \mu_1^2} \end{aligned}$$

Dobimo torej naslednji cenilki:

$$\begin{aligned} \hat{\lambda} &= \frac{\frac{1}{n}\sum X_i}{\frac{1}{n}\sum (X_i - \bar{X})^2} \\ \hat{a} &= \frac{(\frac{1}{n}\sum X_i)^2}{\frac{1}{n}\sum (X_i - \bar{X})^2} \end{aligned}$$

.

Problemi metode: povedali nismo ničesar o lastnostih cenilke. Kot vidimo na zgornjem protiprimeru (normalna pozradelitev), cenilka ni nepristranska. Prav tako nimamo nobene generične informacije o porazdelitvi cenilke - vse bomo morali izpeljati za vsako cenilko posebej.

Primer: Ocenjevanje deleža

Pogosto nas zanima delež v populaciji. Na primer - zanima nas delež tujcev v Sloveniji, delež kreditojemalcev, ki odplačujejo redno, ipd.
Torej, imamo Bernoullijevo spremenljivko, X_i. Povprečje enot je rav no delež, torej zanima nas prvi moment, $\hat{\pi} = 1/n \sum X_i$.
Kako je porazdeljena ta ocena? Vsota i.i.d. Bernoullijevih spremenljivk je porazdeljena binomsko. Za lažje računanje lahko uporabimo normalno aproksimacijo, ki nam jo da centralni limitni izrek.

Vsota i.i.d Bernoullijevih spremenljivk prešteje, kolik enotam se je zgodil nek dogodek. Torej, X =število enot z dogodkom. Delež potem izračunamo kot X/N. Če poznamo porazdelitev X, preprosto izpeljemo tudi vse potrebno za delež:

Binomska porazdelitev:

$$\begin{aligned} E(X) &= E(\sum Y_i) = \sum E(Y_i) = N\pi \\ \text{var}(X) &= \text{var}(\sum Y_i) = \sum \text{var}(Y_i) = N\pi(1-\pi) \end{aligned}$$

Kaj pa delež? Očitno dobimo $E(D) = \pi$ in $\text{var}(D) = N\pi(1-\pi)/N^2 = \pi(1-\pi)/N$.

Kdaj torej pride v poštev Binomska porazdelitev? Zadostiti mora naslednjim kriterijem:

- Imejmo N enot, pri vsaki se dogodek lahko zgodi ali ne - Bernoullijeva porazdelitev
- Naj bodo enote med seboj neodvisne (kaj se zgodi pri enem človeku ne vpliva na to, kaj se zgodi pri drugem - protiprimer, dvojčki, nalezljive bolezni, prepisovanje ...)
- Naj bodo vse enote porazdeljene z enakim π (če ne o študentih ne vem nič dodatnega, imajo vsi enako verjetnost, da opravijo izpit), označimo vrednosti z Y_i.

.

V tem primeru smo torej imeli tudi porazdelitev. Izračunamo lahko interval zaupanja ...
Povzemimo: imamo splošno metodo, ki nam pomaga najti cenilko. Če je cenilka zvezna funkcija momentov, je dosledna (to nam da zakon velikih števil, etc), vendar ne vemo nič o tem, kako optimalna je. Je mogoče najti cenilko z manjšo varianco? Ničesar ne vemo o njeni porazdelitvi.

2.3 Metoda največjega verjetja

Metoda največjega verjetja je najpogosteje uporabljana metoda za iskanje točkovnih cenilk parametrov, pojavlja se na vseh področjih statistike. Je intuitivno smiselna in kot bomo videli ima nekaj zelo uporabnih lastnosti.

2.3.1 Ideja

Primer:
Recimo, da imamo nek vzorec velikosti n iz Bernoullijeve porazdelitve.

$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$$

Bernoullijeva porazdelitev je določena z enim parametrom, imenujmo ga p. Kako bi ga ocenili s pomočjo vrednosti vzorca?

Vemo že, da je smiselna ocena vzorčni delež enic (vzorčno povprečje). Na tem preprostem primeru katerega rešitev že razumemo, si poglejmo, kako poteka ocenjevanje po metodi največjega verjetja:

Recimo, da imamo vzorec velikosti 5, vrednosti so npr.: 1,0,1,1,1. Čeprav je smiselna ocena za p enaka 0,8, je seveda povsem možno, da podatki izhajajo iz Bernoullijeve porazdelitve z neko drugo vrednostjo p-ja.

Ali je možno, da je $p = 0{,}2$? Seveda je. Ne moremo govoriti o verjetnosti, da je $p = 0{,}2$ (p je vrednost parametra, ne pa slučajna spremenljivka. p je fiksen, le mi ga ne poznamo).

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1, p = 0{,}2) = 0{,}2^4 0{,}8^1 = 0{,}00128$$

Kaj pa za $p = 0{,}75$?

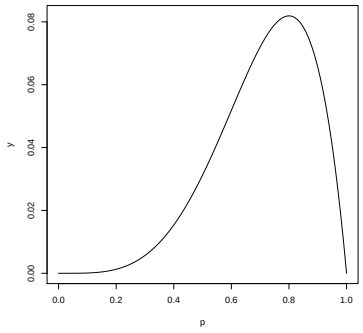
$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1, p = 0{,}75) = 0{,}75^4 0{,}25^1 = 0{,}079$$

Verjetnost dogodka izračunamo za poljuben p izračunamo kot

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, p) = p^k (1-p)^{n-k},$$

kjer je k število enk. Za naš primer podatkov lahko narišemo krivuljo, ki poveže verjetnosti opaženega dogodka za vse možne vrednosti p (prostor parametrov p so vse realne vrednosti med 0 in 1), slika 2.1.

Vrh funkcije lahko poiščemo z odvajanjem - odvajamo funkcijo $p^k(1-p)^{n-k}$ po p in izenačimo z 0 (lokalni maksimum). Poglejmo si za naš primer:



Slika 2.1: *Verjetnost opaženega dogodka glede na p.*

$$\begin{aligned} \left[p^4(1-p)^1\right]' &= [p^4 - p^5]' = 4p^3 - 5p^4 \\ 4\hat{p}^3 - 5\hat{p}^4 &= 0 \\ 4 - 5\hat{p} &= 0 \\ \hat{p} &= \frac{4}{5} \end{aligned}$$

Prišli smo do enake rešitve kot smo jo že uganili na začetku. Naša ocena po metodi največjega verjetja je v tem primeru enaka deležu, v splošnem pa lahko rečemo, da je to tista vrednost parametra, za katero je verjetnost našega opaženega dogodka največja možna.

Povzemimo še enkrat, zanima nas parameter p, tako da bo verjetnost

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1) = \prod P(X_i = x_i)$$

največja možna.

.

Primer:

Recimo, da imamo vrednost iz Gama porazdelitve. Nekaj gostot iz družine gama je prikazano na sliki 1.1. Denimo, da bi imeli samo eno vrednost. Stavili bi, da prihaja iz tiste gama porazdelitve, ki ima v tisti točki največjo gostoto. Kaj pa če imamo več točk - izberemo tisto gama porazdelitev, pri kateri je verjetnost, da dobimo take točke največja. Ker je gama zvezna

funkcija, nas bo namesto produkta verjetnosti zanimal produkt gostot. Ta produkt imenujemo verjetje.

.

2.3.2 Definicija

Kako bi to zapisali bolj formalno? Govorimo o verjetju (likelihood). Imejmo vzorec velikosti n, posamezne vrednosti vzorca označimo z x_i. Gostoto porazdelitve označimo z f, pri tem naj θ označuje vse možne parametre te porazdelitve. Funkcija verjetja je definirana kot (v diskretnem primeru vlogo f prevzamejo verjetnosti)

$$L(x, \theta) = \prod_{i=1}^{n} f(X_i = x_i, \theta)$$

Zanima nas, za katero vrednost (vrednosti) θ, bo funkcija L pri danih podatkih največja možna. Da bi poiskali maksimum, bomo odvajali.

Ker imamo vedno opravka s produktom in je produkt precej zoprno odvajati, bomo pogosto namesto funkcije L obravnavali raje funkcijo $logL$. Ker je preslikava z logaritmom monotona, bosta obe funkciji imeli maksimum v isti točki, $logL$ pa bo bolj pripravna za odvajanje, saj logaritem produkte spremeni v vsote, spomnimo se pravila za računanje z logaritmi:

$$\log(a \cdot b) = \log a + \log b$$

(Mimogrede: logaritem je definiran samo za pozitivne vrednosti, vendar to v našem primeru ni problem, saj so tudi gostote vedno pozitivne.) Uporabljali bomo oznako $l(x, \theta) = \log L(x, \theta)$, zapišemo torej lahko:

$$l(x, \theta) = \log(\prod_{i=1}^{n} f(x_i, \theta)) = \sum_{i=1}^{n} \log f(x_i, \theta).$$

Ker bomo logaritme odvajali, se spomnimo še pravil za odvajanje logaritma:

$$(\log x)' = \frac{1}{x}; \quad [\log(f(x)]' = \frac{f'(x)}{f(x)}$$

Poleg tega, da so vsote lažje za odvajanje, bomo vsoto uporabili tudi za dokazovanje teoretičnih lastnosti - kadar imamo opravka z vsotami, lahko uporabimo centralni limitni izrek.

Primer:

Vrnimo se k primeru Bernoullijeve porazdelitve.

$$\begin{aligned} L(x,p) &= \prod P(X_i = x_i, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ l(x,p) &= \log L(x,p) = \sum_{i=1}^{n} x_i \log(p) + (n - \sum_{i=1}^{n} x_i)\log(1-p) \end{aligned}$$

Sedaj izračunajmo odvod logaritma verjetja:

$$\begin{aligned} l(x,p)' &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} \\ &= \frac{\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i - p(n - \sum_{i=1}^{n} x_i)}{p(1-p)} \\ &= \frac{\sum_{i=1}^{n} x_i - pn}{p(1-p)} \end{aligned}$$

Izenačimo z 0, da dobimo maksimum (robne vrednosti p, torej 0 in 1 nas ne zanimajo):

$$\begin{aligned} \frac{\sum_{i=1}^{n} x_i - \hat{p}n}{\hat{p}(1-\hat{p})} &= 0 \\ \hat{p} &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

.

2.3.3 Lastnosti, izpeljave

Podobno kot pri metodi momentov imamo torej tudi tu splošen pristop k iskanju cenilke. Kaj lahko rečemo o lastnostih cenilk po metodi največjega

verjetja?

Izkaže se, da lahko izpeljemo kar nekaj lepih lastnosti, vendar pa je teorija tu nekoliko bolj poglobljena. Mi bomo dokaz lastnosti le nakazali, nekaj več (a ne vsi) korakov dokaza je v Rice-u, razdelek 8.5.2. Pri dokazu bomo naredili več predpostavk, predvsem bomo zahtevali, da ima funkcija gostote f lepe lastnosti (menjali bomo vrstni red odvajanja in integriranja, ipd). Dokaz bomo pokazali za primer, ko ocenjujemo en sam parameter (θ je enorazsežen). Zaradi lepih lastnosti in preprostosti uporabe se metoda največjega verjetja pojavlja na vseh področjih statistike, pokazali bomo torej tistih nekaj korakov, ki so ključnega pomena pri razumevanju teh lastnosti.
Dokazovali bomo asimptotske lastnosti, torej kaj velja na vzorcu, ko gre $n \to \infty$. Osrednjega pomena pri metodi največjega verjetja bo funkcija $l'(\theta)$, tej funkciji pravimo zbir (score function).
Zanima nas, kaj lahko rečemo o oceni, ki jo poda cenilka $\widehat{\theta}$ pridobljena po metodi največjega verjetja glede na pravo vrednost θ_0, torej kaj lahko rečemo o velikosti razlike $\widehat{\theta} - \theta_0$ oz. njeni porazdelitvi. To razliko lahko izrazimo iz Taylorjeve vrste (razvijemo vrsto za funkcijo $l'(\widehat{\theta})$ okrog θ_0, pri tem uporabimo, da velja $l'(\widehat{\theta}) = 0$, saj smo na ta način dobili $\widehat{\theta}$:

$$\begin{aligned} l'(\widehat{\theta}) &= l'(\theta_0) + (\widehat{\theta} - \theta_0) l''(\theta_0) + \text{ostanek} \\ 0 &\approx l'(\theta_0) + (\widehat{\theta} - \theta_0) l''(\theta_0) \\ \widehat{\theta} - \theta_0 &\approx -\frac{l'(\theta_0)}{l''(\theta_0)} \qquad (2.1) \end{aligned}$$

(2.2)

V zadnjem izrazu je $\widehat{\theta}$ seveda slučajna spremenljivka (medtem ko je θ_0 fiksna vrednost). Torej moramo tudi na desni strani izraza imeti slučajne spremenljivke - tako l' kot tudi l'' sta odvisni od X in s tem slučajni.
Ključnega vlogo pri lastnostih cenilke bosta torej igrala funkcija zbira in drugi odvod logaritma verjetja. Opazimo, da imamo tako v števcu kot tudi v imenovalcu vsoto, zapišimo izraza kot

$$\begin{aligned} l'(\theta_0) &= \sum_{i=1}^{n} [\log f(X_i, \theta_0)]' = \sum_{i=1}^{n} Y_i \\ l''(\theta_0) &= \sum_{i=1}^{n} [\log f(X_i, \theta_0)]'' = \sum_{i=1}^{n} Z_i \end{aligned}$$

V številcu imamo vsoto slučajnih spremenljivk, uporabimo lahko centralni limitni izrek, ki pravi, da vsota konvergira proti normalni porazdelitvi. Ko gre n proti neskončno, torej približno velja $\sum_{i=1}^{n} Y_i \sim N(nE(Y_i), n\text{var}(Y_i))$, oziroma

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \sim N(\sqrt{n}E(Y_i), \text{var}(Y_i)).$$

Poznamo torej asimptotsko porazdelitev v številcu, za celoten dokaz je torej potrebnih nekaj korakov:

- Pokazati, da je imenovalec v primerjavi s številcem praktično konstanten in ga zato lahko obravnavamo kot konstanto.
- Če to velja, je porazdelitev kvocienta enaka porazdelitvi števca, torej normalna. Potrebujemo le še njeno pričakovano vrednost
- in varianco.

Pokažimo najprej, da je slučajna spremenljivka v imenovalcu skoraj konstanta glede na spremenljivko v številcu:
Zapišimo izraz takole

$$\sqrt{n}(\widehat{\theta} - \theta_0) \approx -\frac{\frac{1}{\sqrt{n}} l'(\theta_0)}{\frac{1}{n} l''(\theta_0)} = -\frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i}{\frac{1}{n} \sum_{i=1}^{n} Z_i} \tag{2.3}$$

Nadaljujmo s pričakovano vrednostjo v številcu:

$$\begin{aligned} E(Y_i) = E([\log f(X, \theta_0)]') &= E\left(\frac{f'(X, \theta_0)}{f(X, \theta_0)}\right) \\ &= \int \left(\frac{f'(x, \theta_0)}{f(x, \theta_0)}\right) f(x, \theta_0) dx \\ &= \int f'(x, \theta_0) dx \\ &= \left[\int f(x, \theta_0) dx\right]' = 1' = 0 \end{aligned}$$

Tu smo naredili pogumen preskok (zamenjali vrstni red odvajanja in integriranja), za dokaz torej predpostavljamo, da je funkcija f dovolj lepa, da je to možno.

Za izraz v števcu smo že pokazali, da konvergira k normalni spremenljivki z varianco varY_i, ki je neodvisna od n, medtem ko je izraz v imenovalcu vzorčno povprečje, za katerega vemo, da varianca pada z n. Ko gre n proti neskončno, lahko torej izraz v imenovalcu obravnavamo kot konstanto.

Ker izraz v imenovalcu (2.3) obravnavamo kot konstanto, imamo še vedno normalno porazdelitev, prav tako imenovalec ne vpliva na pričakovano vrednost, ki je še vedno enaka 0. Da bomo v celoti poznali porazdelitev, moramo torej le še izračunati varianco:

$$\sqrt{n}(\widehat{\theta} - \theta_0) \sim N(0, \frac{\mathrm{var}(Y_i)}{E(Z_i)^2})$$

Označimo

$$I(\theta) = \frac{E(Z_i)^2}{\mathrm{var}(Y_i)},$$

tej funkciji pravimo Fisherjeva informacija.

Izkaže se, da velja

$$\mathrm{var}(Y_i) = E(Y_i^2) = E\left[([\log f(X, \theta_0)]')^2\right] = -E\left([\log f(X, \theta_0)]''\right) = -E(Z_i),$$

zato je varianca spremenljivke Y_i (ker je pričakovana vrednost 0, je enaka $E(Y_i^2)$) enaka negativni pričakovani vrednosti drugega odvoda, torej $-E(Z_i)$. Fisherjeva informacija je enaka

$$I(\theta) = -E(Z_i)$$

Varianco bomo torej izrazili s pomočjo izraza

$$I(\theta_0) = -E\left([\log f(X, \theta_0)]''\right)$$

ali

$$I(\theta_0) = E\left[([\log f(X, \theta_0)]')^2\right]$$

Povzemimo torej lastnosti ocene največjega verjetja in na koncu dodajmo še eno, zelo pomembno, a tokrat brez dokaza. Cenilka po metodi največjega verjetja je

- asimptotsko nepristranska (dosledna)
- se normalno porazdeljuje okrog prave vrednosti
- njena varianca je enaka $\frac{1}{nI(\theta_0)}$
- je asimptotsko učinkovita - ima najmanjšo varianco od vseh alternativnih cenilk (izrek Cramer-Rao poda spodnjo mejo za varianco nepristranskih cenilk, ki je enaka varianci pri metodi največjega verjetja)

Seveda pa kot vsaka metoda tudi ta ni brez težav:

- Vse omenjene lastnosti veljajo le za velike n, na majhnih vzorcih so lahko odstopanja precejšnja
- Funkcija ima lahko več ekstremov, lahko so lokalni maksimumi
- Lahko imamo težave z numerično nestabilnostjo

Še opomba: $I(\theta_0)$ seveda v praksi ne poznamo, zato bomo seveda uporabljali $I(\widehat{\theta})$, pokazati se namreč da, da je tudi porazdelitev $\sqrt{nI(\widehat{\theta})}(\widehat{\theta} - \theta_0)$ asimptotsko normalna za zvezne funkcije I.

$$P\left(z_{\alpha/2} \leq \sqrt{nI(\widehat{\theta})}(\widehat{\theta} - \theta_0) \leq z_{1-\alpha/2}\right) \approx 1 - \alpha,$$

zato lahko aproksimativni $100(1-\alpha)$ % interval zaupanja zapišemo kot

$$\widehat{\theta} \pm z_{1-\alpha/2} \frac{1}{\sqrt{nI(\widehat{\theta})}}$$

Primer:

Vrnimo se k Bernoullijevi porazdelitvi. Cenilka po metodi največjega verjetja je enaka $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Zapišimo oceno standardne napake

$$\begin{aligned} f(X,p) &= p^X(1-p)^{1-X} \\ I(p) &= -E\left[(\log f(X,p))''\right] \end{aligned}$$

Izračunati moramo torej drugi odvod. Prvi odvod:

$$(X \log p + (1 - X) \log(1 - p))' = \frac{X}{p} - \frac{1 - X}{1 - p}$$

Drugi odvod

$$\frac{X - p}{p(1 - p)}' = -\frac{X}{p^2} - \frac{1 - X}{(1 - p)^2}$$

Za izračun Fisherjeve informacije, potrebujemo še pričakovano vrednost. Slučajna spremenljivka tu je X, k sreči je naš izraz linearen v X, tako da lahko kot vedno uporabimo pravila za računanje pričakovane vrednosti ter pri tem upoštevamo, da je $E(X) = p$.

$$\begin{aligned} I(p) &= -E\left[-\frac{X}{p^2} - \frac{1 - X}{(1 - p)^2}\right] \\ &= \frac{E(X)}{p^2} + \frac{1 - E(X)}{(1 - p)^2} = \frac{p}{p^2} + \frac{1 - p)}{(1 - p)^2} \\ &= \frac{1}{p} + \frac{1}{(1 - p)} = \frac{1}{p(1 - p)} \end{aligned}$$

Kakšna je torej standardna napaka naše ocene?

$$\begin{aligned} \text{var}(p) &= \frac{1}{nI(p)} = \frac{p(1 - p)}{n} \\ SE &= \sqrt{\frac{p(1 - p)}{n}} \end{aligned}$$

Ta cenilka je enaka tisti, ki smo jo za ta primer že izpeljali. Kadar imamo na voljo eksaktno izpeljane porazdelitev in cenilke seveda ne potrebujemo asimptotskih metod kot je metoda največjega verjetja. Ta primer smo tako pokazali predvsem zato, da na preprostem primeru razumemo 'recept' metode največjega verjetja, pa tudi za to, ker se v le malo spremenjeni obliki verjetje pojavlja v modelih, npr. v modelu logistične regresije, kjer metoda največjega verjetja pride še kako prav.

Kako bomo v praksi izračunali interval zaupanja? V SE bomo seveda morali vstaviti ocene vrednosti p. Nato bomo uporabili normalno porazdelitev in torej prišteli ter odšteli ustrezen faktor $\widehat{SE}$. Ta interval zaupanja ni eksakten, izpeljali smo ga pri $n \to \infty$, na majhnih vzorcih bi morali preveriti, kako natančen dejansko je (kakšno je pokritje- v kakšnem deležu pokriva pravo vrednost). .

Primer: Ravnotežje Hardy-Weinberg

Izpeljimo cenilko po metodi največjega verjetja še za ta primer.
Naj bo $I_i(x)$ indikatorska funkcija za vsako izmed celic.

Verjetje zapišemo kot

$$L(x, \theta) = \prod_{i=1}^{n} (\theta^2)^{I_1(x_i)} (2\theta(1-\theta))^{I_2(x_i)} ((1-\theta)^2)^{I_3(x_i)}$$

Logaritem verjetja je torej enak

$$\begin{aligned} l(x, \mu) &= 2n_1 \log \theta + n_2 \log(2\theta(1-\theta)) + 2n_3 \log(1-\theta) \\ &= 2n_1 \log \theta + n_2 \log(2) + n_2 \log(\theta) + n_2 \log(1-\theta) + 2n_3 \log(1-\theta) \end{aligned}$$

Odvajamo

$$l'(x, \mu) = \frac{2n_1}{\theta} + \frac{n_2}{\theta} - \frac{n_2}{1-\theta} - \frac{2n_3}{1-\theta}$$

Izenačimo z 0 in dobimo

$$2n_1(1-\widehat{\theta}) + n_2(1-\widehat{\theta}) - n_2\widehat{\theta} - 2n_3\widehat{\theta}$$

Cenilka po metodi največjega verjetja je torej

$$\widehat{\theta} = \frac{2n_1 + n_2}{2n}$$

Kaj pa njena varianca?

$$(logf(X,\mu))' = \frac{2I_1}{\theta} + \frac{I_2}{\theta} - \frac{I_2}{1-\theta} - \frac{2I_3}{1-\theta}$$

$$(logf(x,\mu))'' = -\frac{2I_1}{\theta^2} + \frac{I_2}{\theta^2} - \frac{I_2}{(1-\theta)^2} - \frac{2I_3}{(1-\theta)^2}$$

Spomnimo, pričakovane vrednosti so enake

$$E(I_1) = \theta^2 \;\; E(I_2) = 2\theta(1-\theta) \;\; E(I_3) = (1-\theta)^2$$

Fisherjeva informacija je torej enaka

$$\begin{aligned} I(\theta) &= -E(l''(x,\mu)) = \frac{2\theta^2}{\theta^2} + \frac{2\theta(1-\theta)}{\theta^2} + \frac{2\theta(1-\theta)}{(1-\theta)^2} + \frac{2(1-\theta)^2)}{(1-\theta)^2} \\ &= 4 + \frac{2(1-\theta)}{\theta} + \frac{2\theta}{(1-\theta)} \\ &= 4 + \frac{2(1-\theta)^2}{\theta(1-\theta)} + \frac{2\theta^2}{(1-\theta)\theta} \\ &= 4 + \frac{2(1-2\theta+2\theta^2)}{\theta(1-\theta)} = \frac{4\theta - 4\theta^2 + 2 - 4\theta + 4\theta^2}{\theta(1-\theta)} \\ &= \frac{2}{\theta(1-\theta)} \end{aligned}$$

Torej,

$$\mathrm{var}(\widehat{\theta}) = \frac{\theta(1-\theta)}{2n}$$

.

Kako izračunamo varianco, kadar ocenjujemo več parametrov naenkrat? Velja soroden izrek, odvode nadomestimo s parcialnimi odvodi, Fisherjeva informacija postane matrika, člen a_{ij} v tej matriki je enak

$$a_{ij} = -E\left[\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log[f(X,\theta_0)]\right)\right],$$

namesto deljenja uporabimo inverz:

$$\mathrm{var}(\widehat{\theta}) = I^{-1}(\theta_0)/n$$

Primer:

Zanima nas, kako je prihodek podjetja v neki panogi odvisen od števila zaposlenih. Predpostavimo, da je prihodek podjetja normalno porazdeljen s povprečjem $\beta_0 + \beta_1 X$, kjer je X logaritem števila zaposlenih. Denimo, da imamo podatke o številu zaposlenih in prihodku za vzorec podjetij, radi bi ocenili parametra β_0 in β_1.
Zapišimo gostoto porazdelitve prihodka podjetja, če vemo, da je varianca enaka σ^2.
Predpostavljamo, da je $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$, torej

$$f(Y, X|\beta_0, \beta_1, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y-\beta_0-\beta_1 X)^2}{2\sigma^2}}$$

Zapišimo funkcijo verjetja. Kaj je funkcija, ki jo moramo maksimizirati? Dani so podatki (x_i, y_i), $i = 1, \ldots, n$.

$$\begin{aligned} L(y, x, \beta_0, \beta_1, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \end{aligned}$$

Logaritem te funkcije je

$$\log L(y, x, \beta_0, \beta_1, \sigma) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

Ker nas zanimata le parametra β_0 in β_1, je prvi del funkcije konstanta, maksimizirati je potrebno le izraz

$$-\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Izraz $y_i - (\beta_0 + \beta_1 x_i)$ predstavlja razdaljo med točkama $T(x_i, y_i)$ in $T(x_i, \beta_0 + \beta_1 x_i)$, to vrednost imenujemo ostanek (razdalja točke od premice). Oceni za parametra β_0 in β_1 določata premico, ki se najbolje prilega podatkom v smislu, da je vsota kvadriranih ostankov točk od premic najmanjša možna. To

oceno zato imenujemo ocena po metodi najmanjših kvadratov. Izračunajmo oceni β_0 in β_1 po metodi največjega verjetja
Najprej za β_0:

$$\begin{aligned}
&\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2\\
&= \quad -2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)
\end{aligned}$$

Če zgornji izraz izenačimo z 0, dobimo (izraz je enak nič za posebni vrednosti β_0 in β_1, ki ju označimo s strešico)

$$\begin{aligned}
-2\left(\sum_{i=1}^{n}y_i-n\widehat{\beta}_0-\widehat{\beta}_1\sum_{i=1}^{n}x_i\right) &= 0\\
\widehat{\beta}_0 &= \frac{1}{n}\left(\sum_{i=1}^{n}y_i-\widehat{\beta}_1\sum_{i=1}^{n}x_i\right)
\end{aligned}$$

Sedaj odvajamo še po β_1:

$$\begin{aligned}
&\frac{\partial}{\partial\beta_1}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2\\
&= \quad -2\sum_{i=1}^{n}x_i\left(y_i-\beta_0-\beta_1x_i\right)\\
&= \quad -2\left(\sum_{i=1}^{n}x_iy_i-\beta_0\sum_{i=1}^{n}x_i-\beta_1\sum_{i=1}^{n}x_i^2\right)
\end{aligned}$$

Če zgornji izraz izenačimo z 0, dobimo

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}x_iy_i-\widehat{\beta}_0\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2}$$

Združimo obe izpeljavi in (po malce premetavanja členov) dobimo

$$\widehat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Sedaj izračunajmo še standardno napako za obe oceni.
Za Fisherjevo matriko informacije moramo izračunati druge odvode. Logaritem funkcije verjetja je enak

$$\log f(Y, X|\beta_0, \beta_1, \sigma) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(Y - \beta_0 - \beta_1 X)^2}{2\sigma^2}$$

Prva odvoda sta enaka

$$\begin{aligned}
\frac{\partial}{\partial\beta_0}\log f(Y, X|\beta_0, \beta_1, \sigma) &= \frac{1}{\sigma^2}(Y - \beta_0 - \beta_1 X) \\
\frac{\partial}{\partial\beta_1}\log f(Y, X|\beta_0, \beta_1, \sigma) &= \frac{X}{\sigma^2}(Y - \beta_0 - \beta_1 X)
\end{aligned}$$

Drugi odvodi so potem

$$\begin{aligned}
\frac{\partial^2}{\partial\beta_0^2}\log f(Y, X|\beta_0, \beta_1, \sigma) &= -\frac{1}{\sigma^2} \\
\frac{\partial^2}{\partial\beta_1^2}\log f(Y, X|\beta_0, \beta_1, \sigma) &= -\frac{X^2}{\sigma^2} \\
\frac{\partial^2}{\partial\beta_1\beta_0}\log f(Y, X|\beta_0, \beta_1, \sigma) &= -\frac{X}{\sigma^2}
\end{aligned}$$

Členi Fisherjeve matrike informacije so negativne pričakovane vrednosti drugih odvodov. Ker pričakovane vrednosti X oziroma X^2 ne poznamo, ju ocenimo iz podatkov:

$$I(\beta_0, \beta_1) = \frac{1}{\sigma^2}\begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Inverz te matrike je potem

$$I^{-1}(\beta_0,\beta_1) = \frac{\sigma^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

in zato

$$\begin{aligned} \mathrm{var}(\widehat{\beta}_0) &= \frac{I_{11}^{-1}}{n} = \frac{1}{n}\frac{\sigma^2 \frac{1}{n}\sum_{i=1}^{n} x_i^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \\ &= \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \end{aligned}$$

ter

$$\begin{aligned} \mathrm{var}(\widehat{\beta}_1) &= \frac{I_{22}^{-1}}{n} = \frac{1}{n}\frac{\sigma^2}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \\ &= \frac{n\sigma^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \end{aligned}$$

.

Primer:

Radi bi ocenili vrednost parametra v eksponentni porazdelitvi:

$$f(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$$

Vemo, $E(X) = \theta$.

Uporabimo metodo največjega verjetja:

$$L(x,\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}}$$

Logaritem verjetja je torej enak

$$\begin{aligned} l(x,\mu) &= -n\log\theta - \sum_{i=1}^{n}\frac{x_i}{\theta} \\ &= -n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n}x_i \end{aligned}$$

Odvajamo

$$l'(x,\mu) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n}x_i}{\theta^2}$$

Izenačimo z 0 in dobimo

$$\begin{aligned} -n\widehat{\theta} + \sum_{i=1}^{n}x_i &= 0 \\ \widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}x_i \end{aligned}$$

Še varianca:

$$\begin{aligned} (logf(X,\theta))' &= -\frac{1}{\theta} + \frac{X}{\theta^2} \\ (logf(X,\theta))'' &= \frac{1}{\theta^2} - \frac{2X}{\theta^3} \\ -E(logf(X|\theta))'' &= -\frac{1}{\theta^2} + \frac{2E(X)}{\theta^3} = \frac{1}{\theta^2} \\ \mathrm{var}(\widehat{\theta}) &= \frac{\theta^2}{n} \end{aligned}$$

Kaj pa če bi gostoto parametrizirali drugače:

$$f(x|\eta) = \eta e^{-x\eta}$$

Kaj bomo dobili kot cenilko? Verjetno 1/povprečjem? Kaj lahko rečemo o lastnostih te cenilke? Zopet asimptotsko normalna?

Poglejmo, kaj dobimo:

$$L(x,\eta) = \prod_{i=1}^{n} \eta e^{-\eta x_i}$$

Logaritem verjetja je torej enak

$$\begin{aligned} l(x,\eta) &= n\log\eta - \sum_{i=1}^{n} \eta x_i \\ &= n\log\eta - \eta \sum_{i=1}^{n} x_i \end{aligned}$$

Odvajamo

$$l'(x,\mu) = \frac{n}{\eta} - \sum_{i=1}^{n} x_i$$

Izenačimo z 0 in dobimo

$$\begin{aligned} \frac{n}{\widehat{\eta}} - \sum_{i=1}^{n} x_i &= 0 \\ \widehat{\eta} = \frac{n}{\sum_{i=1}^{n} x_i} & \end{aligned}$$

Še varianca:

$$\begin{aligned} (logf(X,\eta))' &= \frac{1}{\eta} + X \\ (logf(X,\eta))'' &= -\frac{1}{\eta^2} \\ -E(logf(X|\eta))'' &= \frac{1}{\eta^2} \\ \mathrm{var}(\widehat{\eta}) &= \frac{\eta^2}{n} \end{aligned}$$

Bi si lahko prihranili računanje? Imamo $\tau(\theta) = \frac{1}{\theta} = \eta$.
Pokazali smo $\widehat{\theta} = \frac{1}{n}\sum x_i$, $\widehat{\eta} = \tau(\widehat{\theta}) = \frac{1}{\frac{1}{n}\sum x_i}$
Kaj vemo o lastnostih $\widehat{\tau(\theta)}$?

.

2.3.4 Metoda delta

Pogosto nas bo zanimala neka funkcija slučajne spremenljivke oz. njena asimptotska porazdelitev.

Primer za motivacijo: obeti

Naj bodo $X_1, X_2, \ldots, X_n$ n.e.p. s.s. porazdeljene z Bernoullijevo porazdelitvijo. Izkaže se, da delež (p) ni nujno najprimernejša mera, v biostatistiki nas bo resda pogosto zanimalo razmerje tveganj (npr. kakšno tveganje za nastanek raka dojke ima ženska z neko mutacijo gena), a bomo s pomočjo podatkov lahko ocenili le razmerje obetov. Obeti so definirani kot $\frac{p}{1-p}$. Če ima nekdo torej verjetnost 2/3, da ozdravi, rečemo, da so njegovi obeti za ozdravitev 2:1. Kako bi ocenili obete? Če je $\widehat{p} = \bar{X}$ cenilka za p, bo smiselna cenilka za obete $\frac{\widehat{p}}{1-\widehat{p}}$. Kakšne so lastnosti te cenilke? Obeti so zvezna funkcija p, torej imamo doslednost, kaj pa varianca, asimptotska porazdelitev?

.

Izrek - metoda delta:
Naj bo $\widehat{\theta}_n$ zaporedje cenilk za θ, za katerega velja

$$\sqrt{n}[\widehat{\theta}_n - \theta] \to N(0, \sigma^2)$$

v porazdelitvi. Potem za poljubno funkcijo g in dano vrednost θ (predpostavimo, da $g'(\theta)$ obstaja in da ni enako 0) velja

$$\sqrt{n}[g(\widehat{\theta}_n) - g(\theta)] \to N(0, [g'(\theta)]^2\sigma^2)$$

v porazdelitvi.

Dokaz

Razvijemo v Taylorjevo vrsto:

$$g(\widehat{\theta}_n) = g(\theta) + g'(\theta)(\widehat{\theta}_n - \theta) + \text{ostanek}$$

Ker $\widehat{\theta}_n$ konvergira θ v verjetnosti, gre ostanek proti 0 v verjetnosti. Na desni strani izraza

$$g(\widehat{\theta}_n) - g(\theta) = g'(\theta)(\widehat{\theta}_n - \theta) + \text{ostanek}$$

imamo torej dva člena - prvi konvergira proti normalni porazdelitvi v porazdelitvi, drugi proti 0 v verjetnosti. Uporabimo izrek Slutskyja in dokazali smo, da levi del izraza konvergira v verjetnosti proti normalni porazdelitvi. Oglejmo si še parametra te porazdelitve:

$$\begin{aligned} E[g(\widehat{\theta}_n)] &\approx g(\theta) + g'(\theta)E(\widehat{\theta}_n - \theta) \\ E[g(\widehat{\theta}_n)] &\approx g(\theta) \end{aligned}$$

Varianca pa je enaka

$$\begin{aligned} \text{var}[g(\widehat{\theta}_n)] &\approx \text{var}(g(\theta)) + (g'(\theta))^2\text{var}(\widehat{\theta}_n - \theta) \\ \text{var}[g(\widehat{\theta}_n)] &\approx (g'(\theta))^2\text{var}(\widehat{\theta}_n) \end{aligned}$$

Uporaba metode delta na primeru obetov

Naša funkcija g je enaka

$$g(x) = \frac{x}{1-x}$$

Odvod je enak

$$g'(x) = \frac{(1-x) - x(-1)}{(1-x)^2} = \frac{(1)}{(1-x)^2}$$

Torej, vemo da bo, ko gre $n \to \infty$ veljalo

$$E\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = E\left(g(\widehat{p})\right) = g(p) = \frac{p}{1-p}$$

Varianca naše cenilke pa bo enaka

$$\text{var}\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = (g'(p))^2\text{var}(\widehat{p}) = \left(\frac{1}{(1-p)^2}\right)^2 \frac{p(1-p)}{n} = \frac{p}{(1-p)^3 n}$$

.

Primer:

Vrnimo se k primeru eksponentne porazdelitve. Pokazali smo

$$\text{var}(\widehat{\theta}) \;=\; \frac{\theta^2}{n}$$

Uporabimo metodo delta za izračun variance

$$\begin{aligned}\text{var}(\tau(\widehat{\theta})) &= (\tau(\theta)')^2\text{var}(\widehat{\theta}) \\ &= (-\frac{1}{\theta^2})^2\frac{\theta^2}{n} \\ &= \frac{\theta^2}{n\theta^4} = \frac{1}{n\theta^2} = \frac{\lambda^2}{n}\end{aligned}$$

.

Primer:

Hardy Weinberg. Za ocenjevanje vrednosti θ smo izpeljali dve cenilki: intuitivno $\widehat{\theta} = \sqrt{\frac{N_1}{n}}$ in cenilko po metodi največjega verjetja (ki je enaka cenilki po metodi momentov): $\widehat{\theta} = \sqrt{\frac{2n_1+n_2}{n}}$. Za slednjo imamo tudi že izpeljano asimptotsko varianco ($var(\widehat{\theta}) = \frac{\theta(1-\theta)}{2n}$).
Da bi ju lahko primerjali, moramo izpeljati še asimptotsko varianco za prvo cenilko.
Uporabili bomo metodo delta.

Označimo $\tau = \theta^2$, pokažimo, da cenilka $\hat{\tau} = \frac{N_1}{n}$ izpolnjuje pogoje izreka. Cenilka je asimptotsko normalno porazdeljena (uporabimo CLI), njena pričakovana vrednost je enaka τ, N_1 je porazdeljen po binomski porazdelitvi in zato $\text{var}(N_1) = \tau(1-\tau)$.
Uporabimo metodo delta, izvemo, da velja (v porazdelitvi)

$$\sqrt{n}(\sqrt{\frac{N_1}{n}} - \theta) \to N(0, g'(\tau)^2 \text{var}(\widehat{\tau}))$$

Torej, cenilka je asimptotsko normalna in nepristranska. Izpeljimo še varianco:

$$\text{var}(\sqrt{\frac{N_1}{n}}) = (\frac{1}{2\sqrt{\tau}})^2 \frac{\tau(1-\tau)}{n} = \frac{1-\tau}{4n} = \frac{1-\theta^2}{4n}$$

Očitno je varianca intuitivne cenilke večja od variance po metodi MNV za vsak θ (vsi rezultati so le asimptotski).

$$\text{var}(\widehat{\theta}) = \frac{\theta^2}{n}$$

Uporabimo metodo delta za izračun variance

$$\begin{aligned} \text{var}(\tau(\widehat{\theta})) &= (\tau(\theta)')^2 \text{var}(\widehat{\theta}) \\ &= (-\frac{1}{\theta^2})^2 \frac{\theta^2}{n} \\ &= \frac{\theta^2}{n\theta^4} = \frac{1}{n\theta^2} = \frac{\lambda^2}{n} \end{aligned}$$

Poglavje 3

Preizkušanje domnev (Rice, 9)

Statistično sklepanje opisano v prejšnjih poglavjih je bilo osredotočeno na ocenjevanje vrednosti v populaciji s pomočjo vzorca. V tem razdelku bomo s pomočjo vzorca preverjali resničnost trditev o populaciji. Trditve o vrednosti nekega parametra v populaciji bomo imenovali **domneve** (hipoteze). Pri tem je bistveno, da je domneva vedno vezana na **populacijo**, preverjati pa jo želimo s pomočjo **vzorca**.
Nekaj primerov domnev:

- Kovanec je pošten, verjetnost grba ali cifre je 0,5.
  Domneva trdi, da je v populaciji parameter enak $p = 0{,}5$, vzorec nam predstavlja npr. 10 metov kovanca, na vzorcu lahko npr. ocenimo delež in s pomočjo njega sklepamo o populacijskem deležu.

- Ni razlik med povprečnim IQ-jem moških in žensk
  Predpostavimo, da se tako moški kot ženske porazdeljujejo normalno z enako varianco, μ naj označuje razliko med povprečjema, zanima nas vrednost μ.

- Število prometnih nesreč v času od sprejetja zakona linearno pada
  Za število prometnih nesreč v času predpostavimo neko porazdelitev oz. model - preverjamo tisti parameter, ki opisuje spreminjanje v času.

Vrednost, ki jo bomo izračunali na vzorcu, in s pomočjo katere se želimo odločati, imenujemo *testna statistika*. Testna statistika je seveda slučajna spremenljivka, označimo jo npr. s $T(X_1, \cdots, X_n)$. Da bomo lahko vedeli s kakšno gotovostjo se odločamo, bomo morali poznati njeno porazdelitev.

Začnimo s primerom za motivacijo:

Primer:
Računalnik skuša razlikovati med dvema viroma signalov. Prvi vir oddaja signale, katerih jakost je normalno porazdeljena z $N(0,3^2)$, drugi vir ima enako porazdelitev, a višjo povprečno jakost - $N(2,3^2)$. Računalnik prejme 9 signalov in se mora odločiti iz katerega vira so prišli. Posamezni signali iz istega vira so med seboj neodvisni.

Računalnik se torej odloča med domnevama

H_1: *Signal prihaja iz vira 1*

in

H_2: *Signal prihaja iz vira 2*

Ker se domnevi razlikujeta le glede povprečja, bo pri odločanju med njima vsekakor najbolj smiselno analizirati povprečje. Naša testna statistika naj bo torej vzorčno povprečje $T = \bar{X}$. Za vrednosti testne statistike nad 1 se nam zdi bolj verjetna domneva H_2, sicer domneva H_1. Recimo, da za mejo našega odločanja vzamemo ravno vrednost 1.

Denimo, da signal prihaja iz vira 1. Kakšna je verjetnost, da se bomo na podlagi vzorca odločili za H_1?
Signal prihaja iz vira 1, torej je porazdeljen normalno s povprečjem 0. Porazdelitev naše testne statistike je $T \sim N(0, \sigma^2/n) = N(0,1)$. Verjetnost, da bo povprečje desetih vrednosti manjše od 1, je

$$P(T < 1|H_1 \text{ je res}) = \mathrm{P}_1(\mathrm{T} < 1) = 0{,}84.$$

Kaj pa, če signal prihaja iz vira 2? Kakšna je verjetnost, da se bomo kljub temu na podlagi vzorca odločili za H_1?
Signal prihaja iz vira 2, torej je porazdeljen normalno s povprečjem 2, testna statistika je porazdeljena kot $T \sim N(2, \sigma^2/n) = N(2,1)$. Verjetnost, da bo povprečje desetih vrednosti manjše od 1 je (označimo standardizirano normalno spremenljivko z Z)

$$P(T < 1|H_2 \text{ je res}) = \mathrm{P}_2(\mathrm{T} < 1) = \mathrm{P}_2(\mathrm{T} - 2 < 1 - 2) = \mathrm{P}(\mathrm{Z} < -1) = 0{,}16$$

Je kritična vrednost 1 res tista, ki nam zagotavlja najboljše odločitve? Mi ta kritična vrednost omogoča, da se največkrat prav odločim?

Formulirajmo to malo bolj natančno. Lahko se vprašamo npr. "Kakšna je verjetnost, da sprejmemo pravo domnevo?"
Izkaže se, da za odgovor na to vprašanje nimamo dovolj podatkov. Izračunati bi morali

$$P(\text{sprejmemo pravo domnevo}) =$$
$$= P(\text{sprejmemo } H_1 | H_1 \text{ je res}) P(H_1 \text{ je res}) + P(\text{sprejmemo } H_2 | H_2 \text{ je res}) P(H_2 \text{ je res})$$

Ker ne poznamo verjetnosti posameznih domnev, tega izraza ne moremo izračunati. Teh verjetnosti tudi ne bomo mogli oceniti iz podatkov. Bayesovska statistika jih predpostavlja (prior), mi se bomo osredotočili na verjetnost pravilne oziroma napačne odločitve. .

Primer:
Izpeljimo primer še nekoliko bolj formalno: naj bosta porazdelitvi $N(0, \sigma^2)$ in $N(a, \sigma^2)$. Pri meji $\frac{a}{2}$ dobimo

$$\begin{aligned} P(T \le \frac{a}{2} | H_1 \text{ je res}) &= P_1\left(\frac{T}{\sigma/\sqrt{n}} < \frac{a}{2\sigma/\sqrt{n}}\right) = P\left(Z < \frac{a}{2\sigma/\sqrt{n}}\right) = \Phi(\frac{a}{2\sigma/\sqrt{n}}) \\ P(T < \frac{a}{2} | H_2 \text{ je res}) &= P_2(T < \frac{a}{2}) = P_2(\frac{T-a}{\sigma/\sqrt{n}} < \frac{\frac{a}{2}-a}{\sigma/\sqrt{n}}) \\ &= P\left(Z < -\frac{a}{2\sigma/\sqrt{n}}\right) = \Phi(-\frac{a}{2\sigma/\sqrt{n}}) \end{aligned}$$

kjer je Φ kumulativna porazdelitvena funkcija normalne porazdelitve. .

3.1 Osnovni pojmi pri statističnem preizkušanju domnev

Kako torej določimo kritično (mejno) vrednost, na podlagi katere se bomo odločali?

Jasno je: **domneve statistično ne moremo dokazati**. Ne moremo npr. dokazati, da je nek parameter v populaciji natanko enak 2 ali pa da podatki izhajajo iz normalne porazdelitve. Ker naključno vzorčenje povzroči, da na vsakem vzorcu dobimo druge podatke, vzorčne ocene ne bodo

praktično nikoli enake populacijskim. Nemogoče je torej z vzorcem pokazati, da je npr. povprečje v populaciji enako 2, če je na vsakem vzorcu različno od te vrednosti.

Primer:

Kot primer si zamislimo ničelno domnevo: 45 % nesreč povzročijo vinjeni vozniki.
Denimo, da imamo vzorec 10 nesreč. Na tem vzorcu ocenimo delež. Jasno je, da delež na vzorcu ne bo natanko enak deležu v populaciji, jasno je tudi, da na podlagi vzorca ne moremo ničesar z gotovostjo trditi o populaciji. Četudi med 10 nesrečami niti ene ni povzročil vinjen voznik, ne moremo trditi, da tudi v populaciji ni nobene take. Vseeno pa nekaj vemo - če zares imamo rezprezentativen vzorec 10 nesreč in nismo na tem vzorcu imeli niti enega vinjenega voznika, je to pri predpostavki, da jih je v populaciji 45 % precej nenavadno. Še več - izračunamo lahko verjetnost tega dogodka:

$$P_0(\sum_{i=1}^{10} X_i = 0) = 0{,}55^{10} = 0{,}0025.$$

Vidimo, da več kot nenavadno - tak rezultat bomo uporabili, da bomo s precejšnjo gotovostjo (s tveganjem manjšim kot 3 promile) zaključili, da je v populaciji manj kot 45 % vinjenih povzročiteljev nesreč.

.

3.1.1 Neyman-Pearsonova paradigma

Ker torej na podlagi vzorca ne moremo govoriti o 'enakosti' v populaciji, dokazovanje obrnemo in domnevo zastavimo tako, da jo bomo poizkusili ovreči, torej z veliko verjetnostjo trditi, da nekaj ni res. Osnovno postavitev problema preizkušanja domnev v frekventistični statistiki imenujemo *Neyman-Pearsonova paradigma*. Temelji na dveh nasprotnih domnevah - *ničelni* (H_0) in *alternativni* (H_A). Velja lahko le ena izmed njiju. Ničelna domneva pri tem ponavadi predstavlja neko dosedanje znanje oziroma prepričanje in pri njej ostajamo, razen če nismo močno prepričani v nasprotno. Vpeljimo terminologijo

- Če ničelno domnevo zavrnemo, čeprav je res, to imenujemo *napaka I.*

vrste. Njeno verjetnost

$$P(\text{zavrnemo } \mathrm{H}_0|\mathrm{H}_0 \text{ je res})$$

označimo z α. Ponavadi bo to majhna vrednost, saj želimo zavrniti H_0, le, če smo v to precej prepričani, torej te napake ne želimo narediti, najpogosteje je $\alpha = 0{,}05$. Pogosto bomo srečali tudi poimenovanje *velikost testa* ali *stopnja značilnosti* (razliko bomo natančneje definirali kasneje), določili jo bomo pred preizkusom in bo podlaga za določanje meje zavrnitve. Čeprav bi vrednost α lahko določali poljubno, je v navadi, da izberemo 0,05, 0,01 ali 0,001.

- Verjetnost, da ne uspemo zavrniti ničelne domneve, čeprav bi jo morali, imenujemo *napaka II. vrste*:

  $$P(\text{ne zavrnemo } \mathrm{H}_0|\mathrm{H}_0 \text{ ni res})$$

  Označimo jo z β. Vrednost $1 - \beta$ imenujemo moč testa.

- Glede na vrednost α določimo *območje zavrnitve*, ki označuje vrednosti testne statistike, pri katerih ničelno domnevo zavrnemo, in *območje sprejema*, ki mu je komplementarno. Točko (oz. več točk), ki razmejuje območji, imenujemo kritična vrednost oziroma meja zavrnitve.

Primer dveh signalov:

Vrnimo se k primeru dveh signalov, naša testna statistika $T = \bar{X}$ je pod ničelno domnevo porazdeljena standardno normalno. Odločimo se za vrednost $\alpha = 0{,}05$, potem je meja zavrnitve enaka 1,64, verjetnost, da to vrednost presežemo po naključju (torej čeprav je ničelna domneva res), je namreč enaka

$$\alpha = P_0(T > 1{,}64) = P(Z > 1{,}64) = 0{,}05,$$

kjer z Z označujemo standardno normalno spremenljivko, s P_0 pa smo označili verjetnost dogodka, če ničelna domneva drži, torej $P_0(A) = P(A|H_0 \text{ velja})$. Interval $(1{,}64, \infty)$ torej predstavlja območje zavrnitve, interval $(-\infty, 1{,}64]$ pa območje sprejema. Izračunajmo še moč testa, torej verjetnost, da je testna statistika v območju zavrnitve, kadar drži alternativna domneva in je T porazdeljena kot $N(2,1)$:

$$1 - \beta = P_A(T > 1{,}64) = P_A(T - 2 > 1{,}64 - 2) = P(Z > -0{,}36) = 0{,}64$$

Moč testa je torej enaka 0,64

.

Ko želimo presojati moč testa, smo vezani na neko alternativno domnevo in porazdelitev, ki jo ta določa. Izračunati bomo namreč morali verjetnost, da uspemo zavrniti ničelno domnevo, kadar drži alternativna. Da bi razumeli, od česa je moč testa odvisna, se vrnimo k primeru dveh virov signalov in ga zapišimo splošneje

Primer - dva vira signalov:
Naj bo pod ničelno domnevo $X \sim N(0, \sigma^2)$, pod alternativno pa $X \sim N(a, \sigma^2)$. Testna statistika $T = \frac{\bar{X}}{\sigma/\sqrt{n}}$ je pod ničelno domnevo porazdeljena kot $T \sim N(0,1)$. Izračunajmo mejo zavrnitve za nek α:

$$P_0(T > c) = \alpha \;\Rightarrow\; P\left(Z > z_{1-\alpha}\right) = \alpha,$$

kjer je Z standardizirana normalna spremenljivka. Velja torej

$$c = z_{1-\alpha}$$

Izračunajmo torej verjetnost zavrnitve, kadar drži alternativna domneva:

$$\begin{aligned} P_A(T > c) &= P_A(T > z_{1-\alpha}) = P_A\left(\frac{\bar{X}}{\sigma/\sqrt{n}} > z_{1-\alpha}\right) \\ &= P_A(\frac{\bar{X} - a}{\sigma/\sqrt{n}} > z_{1-\alpha} - \frac{a\sqrt{n}}{\sigma}) = P(Z > z_{1-\alpha} - \frac{a\sqrt{n}}{\sigma}) \end{aligned}$$

Manjši kot je izraz na desni strani, večja bo moč testa.
Interpretirajmo rezultat: Moč je odvisna od štirih količin:

- vrednosti α: manjši α, manjša moč
- vrednosti parametra pod alternativno domnevo: večji a, večja moč
  Bolj kot sta povprečji narazen, lažje to opazimo s pomočjo vzorca.
- velikosti vzorca: večji n, večja moč
  Večji kot imamo vzorec, bolj zanesljivi so naši odgovori. Velikost vzorca ne vpliva na napako prve vrste (ta je postavljena na α), je pa ključnega pomena pri postavitvi meje zavrnitve in s tem moči testa.

- vrednosti σ: večja varianca, manjša moč
  Če so podatki v populaciji precej razpršeni, bomo težje opazili razliko med povprečjema, saj ne bo nič čudnega, če bodo posamezne vrednosti precej daleč od povprečja pod ničelno domnevo.

Katere izmed gornjih količin zares lahko spreminjamo? Vrednost α je ponavadi fiksna, večina znanstvenih člankov ne dovoljuje, da bi bila večja od 0,05. Za katero vrednost (izmed vseh vrednosti pod alternativno domnevo) računamo moč - za tisto, ki je najmanjša med strokovno pomembnimi. Radi bi razliko vsaj takšno ali še večjo, če bo manjša, nam tudi statistična značilnost ne bo dosti pomagala. Na to vrednost torej prav tako ne moremo vplivati. Jasno je, da lahko vplivamo na velikost vzorca, kaj pa na varianco? Lahko in sicer tako, da zožimo svoje cilje ter omejimo populacijo, ki nas zanima tako, da jo naredimo bolj homogeno. Npr., namesto da gledamo bolnike vseh starosti in večih različnih diagnoz, se omejimo na starost 30-50 let. Seveda pa moramo vedeti, da to pomeni, da bomo tudi rezultati na koncu morali interpretirati na tej podpopulaciji, vsako sklepanje izven nje, bo čisto ugibanje. .

Primer - sodišče:

Ravnotežje vrednosti α in β lahko razumemo z analogijo dokazovanja na sodišču. Ničelna domneva na sodišču je

Obtoženi ni kriv

V sistemu zahodnega sveta ničelno domnevo zavrnemo, le če smo v dokaze zelo zelo prepričani - 'onkraj razumnega dvoma'. Ta 'dvom' v statistiki opišemo z α, dvom sme biti zelo majhen, saj ne želimo obsojati nedolžnih. Na sodišču je α zelo zelo majhen, pet odstotkov po krivem obsojenih bi bilo ogromno.
Seveda pa nam je jasno, da z manjšanjem α povečujemo napako II. vrste - veča se verjetnost, da bomo dejansko krivega izpustili.
Hkrati vemo tudi, da obstajajo sodni sistemi, kjer je α nastavljena bistveno višje, tam so mnenja, da cena obsodbe nedolžnih ni tako velika, da ne bi opravičila dejstva, da uspejo obtožiti veliko več krivih (večja moč). .

Na mnogih področjih raziskovanja je uveljavljena vrednost $\alpha = 0{,}05$. Ta vrednost ohranja neko ravnotežje med tem, da imamo večjo moč $(1 - \beta)$ za odkrivanje novega in hkrati ne naredimo preveč napačnih zaključkov (napaka I. reda).

Povejmo še eno definicijo:

Domneva je **enostavna**, če natanko določa porazdelitev (torej, natanko določa vrednosti parametrov neke porazdelitve). V nasprotnem primeru govorimo o sestavljeni domnevi.

Domnevi, opisani v našem primeru, sta bili primera enostavnih domnev. Če enostavna domneva drži, je porazdelitev testne statistike natanko določena. Nasprotno sestavljena domneva lahko zajema več vrednosti parametrov in tako porazdelitve ne poznamo (in ne moremo računati verjetnosti) tudi če vemo, da domneva drži. Pogosto bo nek parameter lahko zavzel vrsto različnih vrednosti (npr. parameter μ v normalni porazdelitvi vse realne vrednosti, parameter σ pa vse pozitivne vrednosti) in bo ničelna domneva enostavna (npr. $\mu = 0$), alternativna pa sestavljena (npr. $\mu \neq 0$), lahko pa bo seveda sestavljena tudi že ničelna domneva (npr. $\mu \leq 0$).

Pisati bomo torej morali nekoliko splošnejše: naj Θ označuje prostor vseh možnih vrednosti parametra (ali večih parametrov) θ, ki so lahko generirali naše podatke. Ničelno in alternativno domnevo zapišemo kot

$$H_0 : \theta \in \Theta_0,\ H_A : \theta \in \Theta / \Theta_0.$$

Naj bo T zvezna testna statistika, katere velike vrednosti kažejo v prid alternativne domneve, c naj bo meja zavrnitve pri nekem α, velja torej

$$\alpha = \sup_{\theta \in \Theta_0} P(\text{zavrnemo } \mathrm{H}_0) = \sup_{\theta \in \Theta_0} \mathrm{P}(\mathrm{T} \geq \mathrm{c})$$

Razložimo gornjo definicijo še z besedami:
Če je ničelna domneva H_0 enostavna, potem množica Θ_0 obsega le en element, imenujmo ga θ_0 - če predpostavimo, da ničelna domneva drži, s tem poznamo tudi njeno porazdelitev. Gornji izraz se poenostavi v

$$\alpha = \sup_{\theta \in \Theta_0} P(\text{zavrnemo } \mathrm{H}_0) = \mathrm{P}_{\theta_0}(\text{zavrnemo } \mathrm{H}_0)$$

Če pa je domneva sestavljena, je v množici Θ_0 več vrednosti. Za vsako lahko izračunamo verjetnost $P_\theta(\text{zavrnemo } \mathrm{H}_0)$, največja izmed teh verjetnosti je enaka α.

Opomba: pri zvezno porazdeljenih testnih statistikah lahko za poljubno vrednost α najdemo ustrezno mejo c, pri diskretnih pa to ni nujno mogoče. Pri željeni stopnji značilnosti 0,05 bomo izbrali tako mejno vrednost c, da velikost testa ne bo presegala te stopnje, bo pa morda ustrezno manjša. Dejanska velikost testa bo pri diskretnih spremenljivkah tako manjša ali enaka željeni velikosti testa (ki jo bomo imenovali stopnja značilnosti).

Primer:

V našem primeru: $\Theta = \{0{,}2\}$, $\Theta_0 = \{0\}$, torej $\Theta/\Theta_0 = \{2\}$.

.

Primer:

Zanima nas povprečna vrednost IQ pri ljubljanskih gimnazijcih. Pri predpostavki, da je normalno porazdeljen s standardnim odklonom $\sigma = 15$, nas zanima, ali je nadpovprečen (torej nad 100), v vzorec zberemo $n = 25$ dijakov. Ničelna domneva naj bo $H_0 : \mu \leq 100$, alternativna pa $H_A : \mu > 100$. Velja torej: $\Theta = (-\infty, \infty)$, $\Theta_0 = (-\infty, 100]$, torej $\Theta/\Theta_0 = (100, \infty)$.

Poiščimo mejo zavrnitve za $\alpha = 0{,}05$. Naša testna statistika je enaka $\bar{X}$ in je pod ničelno domnevo porazdeljena kot $\bar{X} \sim N(\theta, \frac{15^2}{n})$ za nek $\theta \in \Theta_0$. Zavrnili jo bomo za vrednosti $\bar{X}$, ki bodo dovolj nad 100. Manjša kot je vrednost θ (v našem primeru μ), bolj levo ima povprečje porazdelitev naše testne statistike in zato je tem manjša verjetnost, da je nad neko vrednostjo c. Največjo verjetnost torej dobimo pri $\theta = 100$. Konstanto torej določimo iz izraza

$$\alpha = \sup_{\theta \in \Theta_0} P(\text{zavrnemo } \mathrm{H}_0) = \sup_{\theta \in \Theta_0} \mathrm{P}(\bar{\mathrm{X}} \geq \mathrm{c}) = \mathrm{P}_{\theta=100}(\bar{\mathrm{X}} \geq \mathrm{c})$$

Izračunati moramo torej ustrezen c za $\bar{X}$, ki je porazdeljena kot $N(100, \frac{15^2}{n})$. Vidimo, da smo supremum preprosto zamenjali z izračunom verjetnosti pri mejni vrednosti iz množice parametrov. Izračunajmo mejo zavrnitve:

$$\begin{aligned} \alpha &= P_{\theta=100}(\bar{X} \geq c) \\ &= P_{\theta=100}(\frac{\bar{X} - 100}{3} \geq \frac{c - 100}{3}) = P(Z \geq \frac{c - 100}{3}) \end{aligned}$$

za $\alpha = 0{,}05$:

$$\begin{aligned} \frac{c-100}{3} &= 1{,}64 \\ c &= 104{,}92 \end{aligned}$$

.

3.1.2 Vrednost p

Zgoraj opisana Neyman-Pearsonova paradigma pogosto deluje nekoliko robotsko, saj sta možna le dva odgovora - domnevo zavrnemo ali obdržimo. Na ta način smo celotno informacijo o podatkih zreducirali v preprosto DA/NE trditev, kar je zagotovo prevelika izguba informacije. Zato poleg (oziroma namesto) te osnovne odločitve navadno poročamo tudi *vrednost p*.
Definirajmo torej vrednost p:
Vzemimo primer enosmerne alternativne domneve (zavračamo za velike vrednosti testne statistike). Naj bo na podatkih vrednost testne statistike T enaka t. Vrednost p je količina, za katero velja

$$p = \sup_{\theta \in \Theta_0} P(T \geq t).$$

Torej - to je verjetnost, da na podatkih dobimo tako ali še bolj ekstremno vrednost, če ničelna domneva drži.

Seveda je jasno, da za $t \geq c$ velja

$$p = \sup_{\theta \in \Theta_0} P(T \geq t) \leq \sup_{\theta \in \Theta_0} P(T \geq c) = \alpha,$$

za vrednosti večje od kritične, je vrednost p torej manjša od 0,05. Kadar je $p < \alpha$ in smo torej v območju zavrnitve, pravimo, da je rezultat statistično značilen.

Primer dveh signalov:
Denimo, da na podatkih izračunamo $T = 1{,}8$. Potem je pod ničelno domnevo verjetnost (v našem primeru nas zanima parameter μ):

$$p = \sup_{\theta \in \Theta_0} P(T \geq t) = P_{\mu=0}(T \geq 1{,}8) = P(Z \geq 1{,}8) = 0{,}036$$

V tem primeru bi torej ničelno domnevo pri $\alpha = 0{,}05$ zavrnili, p pa nam predstavlja verjetnost, da bi na podatkih dobili tako ali še bolj ekstremno

vrednost testne statistike, čeprav ničelna domneva drži. .

Primer IQ testa:
Denimo, da na podatkih izračunamo $T = 108$. Potem je pod ničelno domnevo verjetnost (v našem primeru nas zanima parameter μ):

$$\begin{aligned} p &= \sup_{\theta \in \Theta_0} P(\bar{X} \geq t) = P_{\mu=100}(\bar{X} \geq 108) \\ &= P(\frac{\bar{X} - 100}{15/5} \geq \frac{108 - 100}{3}) = P(Z > 2{,}7) = 0{,}0038 \end{aligned}$$

V tem primeru bi prav tako ničelno domnevo pri $\alpha = 0{,}05$ zavrnili, p pa nam predstavlja verjetnost, da bi na podatkih dobili tako ali še bolj ekstremno vrednost testne statistike, čeprav ničelna domneva drži. .

Mimogrede: Vrednost p je v statistiko vpeljal sloviti angleški statistik R.A. Fisher, njegova paradigma preverjanja domnev zahteva le ničelno domnevo, za izračun vrednosti p določitev alternativne domneve ni potrebna. Manjša kot je vrednost p, močnejši je naš dokaz proti ničelni domnevi. Alternativnih domnev je lahko več, zanimale nas bodo, ko bomo govorili o moči testa.

Vrednost p nam omogoča, da se vsaj nekoliko izognemo ostremu odločanju glede zavrnitve ničelne domneve. Jasno je, da je nesmiselno trditi, da smo pri rezultatu $p = 0{,}049$ kaj bolj prepričani o ničelni domnevi kot pri rezultatu $p = 0{,}051$. Tidva rezultata morata vsekakor biti interpretirana na smiselno enak način.

3.1.3 Postopek preizkušanja domnev

Postopek preizkušanja domnev je vedno enak, zapišimo ga v korakih:

1. Zapišemo ničelno domnevo. Določimo velikost napake prve stopnje α. Postavimo ostale predpostavke.

2. Definiramo smiselno testno statistiko. Poiščemo njeno porazdelitev pod ničelno domnevo in za dano vrednost α poiščemo mejo zavrnitve ter s tem določimo območje zavrnitve.

3. Izračunamo vrednost testne statistike na podatkih. Glede na njeno vrednost zavrnemo oz. obdržimo ničelno domnevo, hkrati poročamo tudi vrednost p.

Nekaj opomb:

- Druga točka je tista, kjer nastopimo **statistiki**, iskanje smiselne testne statistike in njene porazdelitve je zagotovo najtežji korak (kar pa seveda ne pomeni, da ni veliko napak tudi pri ostalih dveh). Pri tem si pogosto pomagamo z asimptotskimi približki porazdelitve - uporabimo centralni limitni izrek.

- **Podatke** pogledamo šele v tretji točki. Če želimo, da bo velikost testa zares takšna kot smo jo določili, podatki ne smejo vplivati na prvo točko, saj sicer vrednost α ni taka kot želimo. To vodi v problem večkratnega preizkušanja domnev, ki ga tu le omenimo, se pa bomo k njemu zaradi pogostosti teh vrst napak še vrnili - temeljiteje ga bomo obdelali v razdelku 3.7

- Bistvena postavka so tudi **predpostavke**. S statističnim testom preverjamo vrednost nekega parametra v populaciji, pri tem pa praktično vedno moramo nekaj predpostaviti o porazdelitvi. Resničnost te predpostavke še kako vpliva na rezultat testa, zato je nevarno interpretirati rezultate ne da bi vedeli karkoli o predpostavkah in njihovi resničnosti. Če predpostavke ne držijo, potem so vsi nadaljnji izračuni vprašljivi. Pri primeru z inteligenčnim količnikom smo tako predpostavili normalno porazdelitev IQ med ljubljanskimi dijaki, prav tako smo predpostavili znano varianco. Če predpostavki nista resnični, verjetnosti seveda ne moremo računati na tak način.

Poročanje vrednosti p je ustaljen način intepretacije rezultatov v statistiki, ki je tudi zelo uporaben, saj lahko vrednosti poljubne testne statistike priredimo vrednost p - ne glede na to, kakšno testno statistiko smo uporabili, z vrednostjo p vedno povzamemo naše mnenje o ničelni domnevi. To pa seveda vodi k pretiranim poenostavitvam, na primer:

- Vrednost p je pogosto **interpretirana** kot verjetnost, da je ničelna domneva res. Ta interpretacija je seveda napačna - ničelna domneva podaja vrednost parametra v populaciji, ki je konstanta in ne slučajna spremenljivka. Vrednost p tako podaja našo verjetnost napake I. stopnje, če sledimo opisanemu postopku preverjanja domnev.

- **Statistično značilen** rezultat ni vedno tudi **strokovno pomemben**. Pove le, da ničelna domneva v populaciji ne drži, sama statistična značilnost pa ne pove nič o tem, kako daleč proč od ničelne domneve smo. Pri velikih vzorcih bodo rezultati zaradi manjše standardne napake bolj značilni, ne da bi bila dejanska odstopanja kaj večja.
  Primer: denimo, da primerjamo po velikosti populacijo Ljubljančanov in Mariborčanov. Statistično značilen rezultat pomeni, da z veliko verjetnostjo verjamemo, da sta populaciji različno veliki. Vzemimo skrajni primer in v vzorec zajemimo prav vse meščane. Denimo, da dobimo razliko 1 mm. V tem primeru izračun vrednosti p ni potreben, saj nimamo negotovosti - imamo namreč celo populacijo. Vrednost p (tveganje, da smo se zmotili, ker imamo le vzorec in ne cele populacije) je enaka 0. Razlika je torej statistično značilna, a prav nič strokovno pomembna. Pri velikih vzorcih bo rezultat praktično vedno statistično značilen, ne da bi to sploh bilo pomembno.

3.1.4 Statistična značilnost in interval zaupanja

V naši teoriji smo oznako α do sedaj uporabljali za označevanje stopnje značilnosti ter za $(1-\alpha)$ odstotne intervale zaupanja. Uporaba iste oznake ni naključje - pri tem namreč govorimo o isti količini. Poglejmo si najprej primer

Primer:

Vrednosti hemoglobina pri vzdržljivostnih športnikih naj bi bile porazdeljene kot $N(148; 12{,}9^2) = N(\mu_0, \sigma^2)$. Kolesarji, pri katerih te vrednosti uporabljajo kot izhodiščne vrednosti pri anti-dopinških testih, so mnenja, da so specifike njihovega športa drugačne in zato povprečna vrednost hemoglobina drugačna. Izvedejo raziskavo in zberejo vzorec n=10 vrednosti profesionalnih kolesarjev:
144 154 141 173 156 141 158 162 159 148
Želimo zavrniti ničelno domnevo, da imajo kolesarji povprečje enako 148 in zapisati interval zaupanja za dejansko vrednost njihovega povprečja.
Lotimo se statističnega sklepanja v opisanih korakih:

1. Ničelna domneva: H_0 : povprečje je pri kolesarjih enaka kot pri ostalih vzdržljivostnih športnikih, $\mu = \mu_0 = 148$. Privzamemo $\alpha = 0{,}05$. Zanima nas dvostranska alternativna domneva, torej $\mu \neq \mu_0$. Predpostavimo, da je porazdelitev vrednosti tudi pri kolesarjih normalna in

da je vrednost σ enaka kot pri ostalih vzdržljivostnih športnikih.

2. Ker nas zanima povprečje, bo smiselna testna statistika seveda vzorčno povprečje $\bar{X}$. Vemo, da je vzorčno povprečje pod ničelno domnevo porazdeljeno kot $N(\mu_0, \sigma^2/n)$, če vzorčno povprečje standardiziramo dobimo testno statistiko ($SE = \sigma/\sqrt{n}$)

$$T = \frac{\bar{X} - \mu_0}{SE} = \frac{\bar{X} - 148}{12{,}9/\sqrt{10}} = \frac{\bar{X} - 148}{4{,}1}$$

Slučajna spremenljivka T (= naša testna statistika) je pod ničelno domnevo porazdeljena standardno normalno. Ker nas zanima dvostranska alternativna domneva, je meja zavrnitve $z_{1-\alpha/2} = 1{,}96$. Velja namreč

$$P(Z \geq z_{1-\alpha/2}) = \alpha/2$$

3. Sedaj izračunajmo vrednost testne statistike na naših podatkih. Dobimo $\bar{X} = 153{,}6$, vrednost testne statistike je torej $T = 1{,}37$. Vrednost je manjša od $z_{1-\alpha/2}$, ničelne domneve torej ne moremo zavreči. Izračunamo še vrednost p (zanimajo nas odstopanja v obe smeri):

$$\begin{aligned} p &= \sup_{\theta \in \Theta_0} P(|T| \geq 1{,}37) = P_{\mu=148}(|T| \geq 1{,}37) \\ &= P(|Z| \geq 1{,}37) = 2 \cdot 0{,}085 = 0{,}17 \end{aligned}$$

Izračunajmo še 95% interval zaupanja. Vemo, $\bar{X} \sim N(\mu_0, \sigma^2/n) = N(\mu_0, SE)$, $(1-\alpha)\%$ interval zaupanja zapišemo kot $[\bar{X} - z_{1-\alpha/2}SE, \bar{X} + z_{1-\alpha/2}SE]$, v našem primeru dobimo $153{,}6 \pm 1{,}96 \cdot 4{,}1 = [145{,}6, 161{,}6]$. S 95% zaupanjem lahko trdimo, da je dejanska vrednost ena izmed vrednosti na tem intervalu (nismo pa v nobeno izmed teh vrednosti bolj ali manj prepričani). Opazimo, da je na tem intervalu tudi vrednost 148.

Tako za preizkušanje domneve kot tudi za izračun intervala zaupanja smo uporabili dejstvo, da je $\bar{X} \sim N(\mu_0, \sigma^2/n)$, pri čemer smo vrednost μ_0 enkrat preizkušali, drugič ocenjevali. Jasno je, da bosta zaradi uporabe iste teorije oba principa interpretacije rezultatov tesno povezana. Poglejmo še enkrat, kako:

Ničelne domneve ne zavrnemo za $Z \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]$. Ker vemo, da velja

$$Z = \frac{\bar{X} - \mu_0}{SE} \Rightarrow Z \cdot SE = \bar{X} - \mu_0$$

lahko zgornji izraz prepišemo v

$$Z \cdot SE \in [-z_{1-\alpha/2} \cdot SE, z_{1-\alpha/2} \cdot SE]$$

Vidimo torej, da ničelne domneve ne zavrnemo za

$$\bar{X} - \mu_0 \in [-z_{1-\alpha/2} \cdot SE, z_{1-\alpha/2} \cdot SE]$$

torej takrat, kadar je razdalja vzorčnega povprečja od tistega pod ničelno domnevo manj kot $z_{1-\alpha/2} \cdot SE$:

$$|\bar{X} - \mu_0| \leq z_{1-\alpha/2} \cdot SE$$

Sedaj se spomnimo še, kako izračunamo interval zaupanja: izhajamo iz dejstva, da imamo verjetnost $1 - \alpha$, da vzorčno povprečje ne bo predaleč proč od prave vrednosti, torej $P(|\bar{X} - \mu| \leq z_{1-\alpha/2}SE) \geq 1 - \alpha$. Interval zaupanja tvorijo ravno tiste vrednosti μ od katerih vzorčno povprečje ni oddaljeno za več kot $z_{1-\alpha/2}SE$. Če bi torej testirali ničelno domnevo za katerikoli μ z intervala zaupanja, ničelne domneve ne bi zavrnili. .

Trditev se da pokazati tudi mnogo splošneje, a je mi ne bomo dokazovali. Povejmo le:

Ničelno domnevo $H_0 : \theta = \theta_0$ pri stopnji značilnosti α zavrnemo natanko takrat, ko θ_0 ni v $(1 - \alpha)\%$ intervalu zaupanja.

Premislimo še tale stavek: $(1 - \alpha)\%$ interval zaupanja za neko vrednost cenilke X je množica tistih parametrov θ, za katere je vrednost cenilke X v območju sprejema.

Interval zaupanja in vrednost p torej hodita z roko v roki, glede na to, kako želimo podatke interpretirati lahko poročamo eno izmed njiju ali obe hkrati.

Primer:

Poglejmo si še primer testne statistike, ki je binomsko porazdeljena. Zanima nas, ali obstaja povezanost med spolom in telefoniranjem med vožnjo.
Kako bi preizkusili to domnevo - kakšne podatke imamo na voljo: na voljo imamo podatke policije, pri vsakem ustavljenem vozniku, ki je telefoniral, imamo podatek o spolu. Premislimo, ali je tak vzorec naključen:

- Pomembno je, zakaj je bil posameznik ustavljen:
  - ker je policist opazil, da oseba telefonira
  - policija ustavlja avtomobile, v katerih vozniki po njihovem mnenju bolj verjetno telefonirajo (večji avtomobili, službeni avtomobili)
  - policija ustavlja zaradi drugih prekrškov in hkrati kaznuje tudi telefoniranje.

  Težava z drugo in tretjo točko je, da bo morda zaradi tega ustavljenih več predstavnikov enega spola (več mosških) in se bo s tem povečala tudi verjetnost, da oseba tistega spola telefonira.

- Pa tudi, če velja prva točka. Ne vemo, ali je v populaciji enako število voznikov obeh spolov. Ne moremo torej govoriti o tem, kdo več telefonira, ampak o tem, ali je med tistimi, ki telefonirajo, razlika v spolu. Če je precej manj voznic žensk, bo namreč tudi delež tistih, ki jih ujamemo pri telefoniranju manjši. Če nas zanima, kateri spol je bolj nagnjen k telefoniranju, potrebujemo drugačne podatke: podatke za n ustavljenih vozil - vsakič bi zbrali podatek o spolu in o telefoniranju.

Kaj lahko torej naredimo z našimi podatki: če bi radi zmanjšali delež telefoniranja in pri tem ciljali na enega izmed spolov, je naša raziskava pravilno zastavljena. Če pa bi radi razumeli, kaj vse je povezano s telefoniranjem, pa nam ti podatki ne dajo pravih odgovorov.

1. Ničelna domneva: $\pi = 0{,}5$. $\alpha = 0{,}05$. Predpostavke: neodvisnost avtomobilov ni sporna, o naključnosti smo že govorili.

2. Testna statistika bo preprosto število žensk, ki telefonirajo. To število bo porazdeljeno kot $Bin(n, \pi)$. Za $\alpha = 0{,}05$ in $n = 50$ lahko pripravimo območje zavrnitve:

$$\begin{aligned} P_{\pi=0,5}(T \leq 17) &= 0{,}016 \\ P_{\pi=0,5}(T \leq 18) &= 0{,}032 \end{aligned}$$

   Območje zavrnitve je torej $(0{,}17] \cup [33{,}50]$, območje sprejema $[18{,}32]$.

3. Na naših podatkih je bilo med 50 posamezniki 31 žensk, vrednost p:

$$p = P_{\pi=0,5}(X \geq 31) + P_{\pi=0,5}(X \leq 19) = 0{,}118$$

Ničelne domneve torej v tem primeru ne moremo zavrniti. Zapišimo še $(1-\alpha)\%$ interval zaupanja. Interval zaupanja zajema vse tiste vrednosti π, pri katerih ničelne domneve za naše podatke ne bi zavrnili.
Kaj so torej vrednosti π, pri katerih je vrednost 31 v intervalu zaupanja? Očitno je taka vrednosti $\pi = 0{,}5$. Poiščimo spodnjo mejo intervala zaupanja (gledamo le enostransko verjetnost):

$$\begin{aligned} P_{\pi=0,4}(T \geq 31) &= 0{,}001 \\ P_{\pi=0,47}(T \geq 31) &= 0{,}024 \\ P_{\pi=0,48}(T \geq 31) &= 0{,}033 \end{aligned}$$

Ničelne domneve $H_0 : \pi = 0{,}48$ torej na naših podatkih ne bi zavrnili, domnevo $H_0 : \pi = 0{,}47$ pa bi. Če zaokrožujemo na dve decimalki, je spodnja meja intervala zaupanja torej 0,48. Na enak način najdemo, da je zgornja meja

$$\begin{aligned} P_{\pi=0,75}(T \leq 31) &= 0{,}029 \\ P_{\pi=0,76}(T \leq 31) &= 0{,}019 \end{aligned}$$

$(1-\alpha)\%$ interval zaupanja je torej enak $[0{,}48, 0{,}75]$.

.

Statistična značilnost, torej dejstvo, da je vrednost p manjša od α in da domnevo zavrnemo, sta osnovni cilj mnogih raziskav. Pogosto bo raziskovalec navdušen nad statistično značilnim rezultatom in razočaran v primeru statistično neznačilnega, v mnogih znanstvenih revijah so namreč naklonjeni le objavljanju statistično značilnih rezultatov. Kot pa smo že omenili, statistično značilen rezultat še ne pomeni strokovno pomembnega in obratno. Zato je pogosto koristno, če interpretaciji rezultatov dodamo še interval zaupanja.

Primer:

Športnikove vrednosti hemoglobina igrajo ključno vlogo v boju proti dopingu, vrednosti, ki pretirano odstopajo, namreč kažejo na manipulacije s krvjo. Hemoglobin se meri na 1 g/l natančno, razlike za manj kot 1 so nepomembne. V našem primeru je bil 95% interval zaupanja enak $[145{,}6, 161{,}6]$. Rezultat ni statistično značilen, s podatki torej nismo uspeli dokazati, da so kolesarji različni od ostalih vzdržljivostnih športnikov. Vendar pa tudi nismo pokazali,

da so enaki, nasprotno, prav mogoče je, da imajo v resnici povprečje enako 155 ali 160, kar pa bi bila ogromna in strokovno zelo pomembna razlika. Za kakršenkoli odgovor na to vprašanje bi torej morali zbrati večji vzorec, s pomočjo katerega bi potem dobili ožji interval zaupanja, ki bi lahko vseboval vrednost 148 ali pa tudi ne.
Naš rezultat torej ni statistično značilen, so pa v intervalu tudi strokovno pomembne vrednosti, zato zaključkov ne moremo delati, edino priporočilo je, da se zbere večji vzorec.
Pogljemo si še ostale možnosti rezultatov oziroma interpretacije. Denimo, da bi v raziskavi dobili naslednje rezultate

1. 95% interval zaupanja je [150,1, 156,1]:
   interval zaupanja je precej širok, a ne zajema vrednosti 148 (vrednosti parametra pod ničelno domnevo). Rezultat je statistično značilen, pa tudi strokovno pomemben, saj lahko s 95% zaupanjem trdimo, da je hemoglobin za vsaj dva g/l večji od drugih športnikov.

2. 95% interval zaupanja je [147,3, 148,7]:
   Interval zaupanja zajema 148, rezultat torej ni statistično značilen. Interval zaupanja je tudi precej ozek, imamo 95 % zaupanje, da razlika ni večja od 0,7. V tem primeru smo statistično dokazali, da med kolesarji in ostalimi športniki ni strokovno pomembnih razlik - nabiranje večjega vzorca ni smiselno.

3. 95% interval zaupanja je [148,1, 148,8]:
   Interval zaupanja ne zajema vrednosti 148, je torej statistično značilen, kar pa še ne pomeni, da je strokovno pomemben - nasprotno, imamo 95 % zaupanje, da razlika ni večja od 0,8.

4. 95% interval zaupanja je [148,1, 160,6]:
   Rezultat je statistično značilen, vendar vidimo, da je vrednost 148 praktično na meji. To pa pomeni, da morda ni nobene strokovno pomembne razlike, morda pa tudi je. Vrednost p je zelo blizu 0,05, omenili pa smo že, da da ni posebej smiselno razlikovati med vrednostmi okoli 0,05. Zato nad rezultatom navkljub statistični značilnosti ne moremo biti pretirano navdušeni.

.

Majhna vrednost p tako ne kaže nujno na to, da smo odkrili nekaj pomemb-

nega. Z večanjem vzorca se manjša standardna napaka in s tem širina intervala zaupanja, pri zelo velikih vzorcih je zato lahko zelo majhna. Če v populaciji obstaja majhna razlika, lahko vedno najdemo dovolj velik vzorec, da bo ta razlika na vzorcu statistično značilna. Na zelo velikih vzorcih (sploh npr. kadar preučujemo celotno populacijo) zato vrednosti p postanejo nebistvene za odločanje.
Nasprotno so intervali zaupanja na majhnih vzorcih lahko zelo široki. V takem primeru ničelne domneve ne bomo mogli zavrniti, ne glede na to, da je alternativna domneva morda res (imamo zelo majhno moč). Velike vrednosti p torej nikakor ne smemo interpretirati kot dokaz resničnosti ničelne domneve. Kdaj torej vseeno lahko delamo zaključke kot je npr. 'ni razlik med volilnimi preferencami glede na spol'? Le takrat, kadar je interval zaupanja dovolj ozek, da ne vsebuje strokovno pomembnih razlik.

3.2 Test t, Rice 11

Test t je eden najpogosteje uporabljanih testov v statistiki, zato si ga bomo v tem razdelku podrobneje ogledali. Obstaja več variant tega testa, glede na to kakšne podatke imamo (ena skupina, primerjava dveh skupin, enaki oz. različni varianci). Vsem je skupno to, da je testna statistika porazdeljena po porazdelitvi t (Studentovi porazdelitvi).

3.2.1 Test z za en vzorec (one-sample z-test)

Preden si ogledamo test t, začnimo s testom z, pri katerem je lažje razumeti izračun moči, prav tako bomo lažje razumeli razlike med testom t in z.

Denimo, da imamo vzorec iz populacije, ki je normalno porazdeljena, poznamo njeno varianco, ki je enaka σ^2. Naša ničelna domneva je $H_0 : \mu = \mu_0$, torej da je povprečje enako neki vrednosti. Ker preverjamo povprečje, je smiselna testna statistika $\bar{X}$. Ker poznamo njena parametra, jo lahko standardiziramo:

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Primer:

Spomnimo se primera z vaj, zanima nas ali tudi slovenska populacija mladih v povprečju začne kaditi pri 13 letih, zbrali smo vzorec velikosti 25, recimo, da vemo, da je v Evropi std. odklon enak 2,5, predpostavimo, da je enak tudi v Sloveniji:

Naša ničelna domneva je $H_0 : \mu = 13$, predpostavljamo normalno porazdelitev populacije, postavimo si $\alpha = 0{,}05$.

Testna statistika bo porazdeljena std. normalno, zapišemo lahko mejo zavrnitve pri naši vrednosti α, ki je (za dvosmerni test) enaka $z_{1-\alpha/2} = 1{,}96$.

Oglejmo si podatke, denimo, da je $\bar{X} = 14{,}2$, ocena standardnega odklona pa $\widehat{\sigma} = 2{,}5$. Potem je vrednost testne statistike enaka

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{14{,}2 - 13}{2{,}5/5} = \frac{1{,}2}{0{,}5} = 2{,}4$$

Vidimo, da lahko ničelno domnevo zavrnemo pri stopnji značilnosti $\alpha = 0{,}05$. Izračunajmo še vrednost p:

$$p = \sup_{\theta \in \Theta_0} P(|T| \geq t) = P_{\mu=13}(|T| \geq 2{,}4) = P(|Z| \geq 2{,}4) = P(Z \geq 2{,}4) + P(Z \leq -2{,}4) = 0{,}016.$$

Vrednosti p je za ta primer torej enaka 0,016. Poglejmo si še interval zaupanja za našo oceno:

$$\bar{X} \pm z_{1-\alpha/2} SE = 14{,}2 + 1{,}96 \cdot 0{,}5 = [13{,}22; 15{,}18]$$

Ker smo ničelno domnevo zavrnili, vrednosti 13 (ničelna domneva) ni v intervalu zaupanja. V tem primeru je očitno, da vrednosti 13 ni v intervalu zaupanja natanko takrat kadar ničelno domnevo zavrnemo (obakrat uporabljamo isto formulo, le obrnjeno).

Kako pa bi izračunali moč za ta primer? Recimo, da nas zanima moč, da najdemo razliko, če je ta dejansko enaka 1 ($\mu = 14$). Testna statistika T sedaj ni porazdeljena kot t:

$$\begin{aligned}
P_{\mu=14}(|T| \geq 1{,}96) &= P_{\mu=14}(T \geq 1{,}96) + P_{\mu=14}(T \leq -1{,}96) \\
&= P_{\mu=14}\left(\frac{\bar{X} - 13}{\sigma/\sqrt{n}} \geq 1{,}96\right) + P_{\mu=14}\left(\frac{\bar{X} - 13}{\sigma/\sqrt{n}} \leq -1{,}96\right) \\
&= P_{\mu=14}\left(\frac{\bar{X} - 14}{\sigma/\sqrt{n}} \geq 1{,}96 - \frac{1}{\sigma/\sqrt{n}}\right) + \\
&\quad P_{\mu=14}\left(\frac{\bar{X} - 14}{\sigma/\sqrt{n}} \leq -1{,}96 - \frac{1}{\sigma/\sqrt{n}}\right) \\
&= P_{\mu=14}\left(\frac{\bar{X} - 14}{0{,}5} - \frac{1}{0{,}5} \geq 1{,}96 - \frac{1}{0{,}5}\right) \\
&\quad + P_{\mu=14}\left(\frac{\bar{X} - 13}{0{,}5} - -\frac{1}{0{,}5} \leq -1{,}96 - \frac{1}{0{,}5}\right) \\
&= P(Z \geq 1{,}96 - 2) + P(Z \leq -1{,}96 - 2) \\
&= P(Z \geq -0{,}04) + P(Z \leq -3{,}96) = 0{,}516
\end{aligned}$$

3.2.2 Test t za en vzorec (one-sample t-test)

Denimo, da imamo vzorec iz populacije, ki je normalno porazdeljena. Naša ničelna domneva je $H_0 : \mu = \mu_0$, torej da je povprečje enako neki vrednosti. Ker preverjamo povprečje, je smiselna testna statistika $\bar{X}$. Vemo, da je ta statistika normalno porazdeljena, vendar pa ne poznamo njene variance in zato ne poznamo porazdelitve. Testno statistiko zato standardiziramo z ocenjeno standardno napako, dobimo

$$T = \frac{\bar{X} - \mu_0}{\widehat{\sigma}/\sqrt{n}}$$

Vemo že, da je ta testna statistika porazdeljena po porazdelitvi t_{n-1} (glej razdelek 1.3.2). Podajmo primer:

Primer:

Denimo, da imamo enak problem kot zgoraj, ne upamo si nič predpostaviti o varianci. Zanima nas ali tudi slovenska populacija mladih v povprečju začne kaditi pri 13 letih, zbrali smo vzorec velikosti 25:

Naša ničelna domneva je $H_0 : \mu = 13$, predpostavljamo normalno porazdelitev populacije, postavimo si $\alpha = 0{,}05$.

Testna statistika bo porazdeljena kot t_{24}, zapišemo lahko mejo zavrnitve pri naši vrednosti α, ki je (za dvosmerni test) enaka $t_{24,\alpha/2} = 2{,}06$.

Oglejmo si podatke, denimo, da je $\bar{X} = 14{,}2$, ocena standardnega odklona pa $\widehat{\sigma} = 2{,}5$. Potem je vrednost testne statistike enaka

$$T = \frac{\bar{X} - \mu_0}{\widehat{\sigma}/\sqrt{n}} = \frac{14{,}2 - 13}{2{,}5/5} = \frac{1{,}2}{0{,}5} = 2{,}4$$

Vidimo, da lahko ničelno domnevo zavrnemo pri stopnji značilnosti $\alpha = 0{,}05$. Izračunajmo še vrednost p:

$$p = \sup_{\theta \in \Theta_0} P(|T| \geq t) = P_{\mu=13}(|T| \geq 2{,}4) = 0{,}0245.$$

Vrednosti p je za ta primer torej enaka 0,0245. Poglejmo si še interval zaupanja za našo oceno:

$$\bar{X} \pm t_{24}\widehat{SE} = 14{,}2 + 2{,}06 \cdot 0{,}5 = [13{,}17; 15{,}23]$$

Ker smo ničelno domnevo zavrnili, vrednosti 13 (ničelna domneva) ni v intervalu zaupanja. V tem primeru je očitno, da vrednosti 13 ni v intervalu zaupanja natanko takrat kadar ničelno domnevo zavrnemo (obakrat uporabljamo isto formulo, le obrnjeno).

Kako pa bi izračunali moč za ta primer? Recimo, da nas zanima moč, da najdemo razliko, če je ta dejansko enaka 1 ($\mu = 14$). Predpostavimo, da je $\sigma = 2{,}5$. Testna statistika T sedaj ni porazdeljena kot t:

$$\begin{aligned} P_{\mu=14}(|T| \geq 2{,}06) &= P_{\mu=14}(T \geq 2{,}06) + P_{\mu=14}(T \leq -2{,}06) \\ &= P_{\mu=14}(\frac{\bar{X} - 13}{\widehat{\sigma}/\sqrt{n}} \geq 2{,}06) + P_{\mu=14}(\frac{\bar{X} - 13}{\widehat{\sigma}/\sqrt{n}} \leq -2{,}06) \\ &= P_{\mu=14}(\frac{\bar{X} - 14 + 1}{\widehat{\sigma}/\sqrt{n}} \geq 2{,}06) + P_{\mu=14}(\frac{\bar{X} - 14 + 1}{\widehat{\sigma}/\sqrt{n}} \leq -2{,}06) \end{aligned}$$

Od tu naprej zdaj ne moremo nadaljevati na enak način kot pri testu z, saj bi na levi res dobili spremenljivko porazdeljeno kot t, a bi nam ostal še člen s slučajno porazdeljenim $\widehat{\sigma}$. Uporabiti moramo ne-centralne porazdelitve t. Velja, da je spremenljivka porazdeljena kot ne-centralni t s parametrom necentralnosti μ, če jo lahko zapišemo kot

$$\frac{Z + \mu}{\sqrt{U/(n-1)}}$$

kjer sta spremenljivki Z in U tako kot pri porazdelitvi t porazdeljeni normalno oz. hi-kvadrat in med seboj neodvisni (v programu R: pt, argument ncp).

Nadaljujemo:

$$\begin{aligned} P_{\mu=14}(|T| \geq 2{,}06) &= P_{\mu=14}(\frac{\frac{\bar{X}-14+1}{\sigma/\sqrt{n}}}{\widehat{\sigma}/\sigma} \geq 2{,}06) + ... \\ &= P_{\mu=14}(\frac{Z + \frac{1}{\sigma/\sqrt{n}}}{\widehat{\sigma}/\sigma} \geq 2{,}06) + ... \\ &= P(T_{nc} \geq 2{,}06) + ... = 0{,}486 \end{aligned}$$

Zgoraj je spremenljivka T_{nc} porazdeljena kot necentralni t s parametrom $1/0{,}5 = 2$. Rezultat desnega dela je zanemarljiv (velikostni razred 10^{-5}).

Dobili smo nekoliko manjšo moč kot pri z, kar je pričakovano - če vemo manj (naredimo manj predpostavk), imamo manjšo moč. .

3.2.3 Test t za dva neodvisna vzorca

Denimo sedaj, da imamo dve skupini (dva vzorca), obe iz populacij, ki sta normalno porazdeljeni. Radi bi primerjali povprečje dveh skupin. Naša ničelna domneva je $H_0 : \mu_1 = \mu_2$, torej da sta povprečji skupin enaki. Predpostavljamo, da je varianca v obeh populacijah enaka. Ker preverjamo povprečji, je smiselna testna statistika $\bar{X}_1 - \bar{X}_2$. Vemo, da je ta statistika normalno porazdeljena, izpeljimo še njeno varianco (velikost vzorca iz prve skupine naj bo n_1, iz druge pa n_2:

$$\mathrm{var}(\bar{X}_1 - \bar{X}_2) = \mathrm{var}(\bar{X}_1) + \mathrm{var}(\bar{X}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Ker variance ne poznamo, jo moramo oceniti iz podatkov. Z izpeljavo nepristranske ocene smo se že ukvarjali na strani 1.4. Vemo, da jo v vsaki skupini nepristransko ocenimo kot

$$\widehat{\sigma}_w^2 = \frac{1}{n_i - 1} \sum_{i=1}^{n_1} (X_{ij} - \bar{X}_{i.})^2,$$

vsaka utežena vsota teh cenilk bo zato nepristranska. Utežimo tako, da vsi ostanki enako prispevajo k cenilki - vzemimo vsoto

$$\sum_{i=1}^{n_1} (X_{1j} - \bar{X}_1)^2 + \sum_{i=2}^{n_1} (X_{2j} - \bar{X}_2)^2 = (n_1 - 1)\hat{\sigma_1}^2 + (n_2 - 1)\hat{\sigma_2}^2$$

S katero konstanto moramo deliti to vsoto, da dobimo nepristransko cenilko?

$$\sigma^2 = E\left(c(n_1 - 1)\hat{\sigma_1}^2 + (n_2 - 1)\hat{\sigma_2}^2 \right) = c((n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2)$$

Očitno je torej $c = 1/(n_1 + n_2 - 2)$.

Skupno varianco (pooled variance) izračunamo kot

$$\begin{aligned} s_p &= \frac{n_1 - 1}{n - 2} \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (X_{1j} - \bar{X}_{1.})^2 + \frac{n_2 - 1}{n - 2} \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_{2.})^2 \\ &= \frac{1}{n - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 \end{aligned}$$

Kmalu bomo videli, zakaj je uteževanje z $\frac{n_i-1}{n-2}$ primernejše od uteževanja z $\frac{n_i}{n}$.

Cenilka $s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$ nepristransko ocenjuje varianco razlike povprečij. Tako kot v primeru za en vzorec torej definiramo cenilko

$$T=\frac{\bar{X}_1-\bar{X}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Pokažimo, da je ta cenilka zares porazdeljena s porazdelitvijo t. Spomnimo se, kaj potrebujemo v ta namen (1.3.2):

Definicija porazdelitve t: naj bo $Z\sim N(0,1)$ in $U\sim\chi^2_n$ ter Z in U neodvisni. Potem je kvocient $Z/\sqrt{U/n}$ porazdeljen po porazdelitvi t_n

Testno statistiko izrazimo kot

$$\begin{aligned}\frac{\bar{X}_1-\bar{X}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} &= \frac{\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\frac{s_p}{\sigma}}\\ &= \frac{\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{s_p^2}{\sigma^2}}}\\ &= \frac{\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{s_p^2(n-2)}{\sigma^2}/(n-2)}}\end{aligned}$$

Izraz v števcu je standardno normalno porazdeljena spremenljivka, poglejmo si izraz v imenovalcu:

$$\frac{s_p^2(n-2)}{\sigma^2}=\frac{1}{\sigma^2}\sum_{j=1}^{n_1}(X_{1j}-\bar{X}_{1.})^2+\frac{1}{\sigma^2}\sum_{j=1}^{n_2}(X_{2j}-\bar{X}_{2.})^2$$

V razdelku 1.3.2 smo pokazali, da sta oba gornja sumanda porazdeljena kot χ^2, z n_1-1 oziroma n_2-1 stopinjama prostosti. Ker je vzorčenje v skupinah med seboj neodvisno, je vsota porazdeljena kot $\chi^2_{n_1+n_2-2}$.
Kot zadnje potrebujemo le še neodvisnost spremenljivk v števcu in imenovalcu. V razdelku 1.3.2 smo pokazali, da sta vzorčno povprečje in odmiki od njega med seboj neodvisni v vsaki skupini posebej, ker obravnavamo dva

neodvisna vzorca (vzorčenje v skupinah je med seboj neodvisno), takoj sledi tudi zahtevana neodvisnost. S tem smo pokazali, da je testna statistika porazdeljena kot t_{n-2}.

Kaj pa če varianci nista enaki? Potem velja

$$\mathrm{var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

in je potrebno vsako varianco oceniti iz svojega vzorca. Vendar pa porazdelitev testne statistike

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

ni več t. Navkljub temu se da pokazati, da lahko tudi to testno statistiko dobro aproksimiramo s porazdelitvijo t, vendar pa se pri tem stopinje prostosti izračunajo po posebni formuli.

Opombi:

- Potrebno je razumeti, da vzorca nista nujno enako velika - velikost testa bo navkljub temu pravilna, nam pa enaka velikost pri dani skupni velikosti vzorca da največjo moč.
- Kaj pričakujemo, če bomo uporabili napačen test: če imamo v resnici enaki varianci in uporabimo test, ki to predpostavko uporabi, pričakujemo, da bo zaradi dodatne informacije SE manjša in zato moč testa večja. Če uporabimo predpostavko o enakih variancah, četudi ne drži, lahko dobimo napačno velikost testa pod ničelno domnevo. Tak je tudi splošen princip: dodatna informacija, ki jo prinašajo predpostavke nam zmanjša standardno napako cenilke. Zato bomo imeli večjo moč za preizkušanje domnev, vendar je v primeru, da je predpostavka napačna, velikost pod ničelno domnevo lahko napačna.

3.2.4 Test t za dva odvisna vzorca

Denimo sedaj, da so enote prvega in drugega vzorca odvisne in sicer na tak način, da vsaki enoti iz prvega vzorca ustreza ena enota iz drugega. Primer so na primer ponovljene meritve na istih enotah (morda z nekim vmesnim

dogodkom). Podatke lahko zapišemo kot pare (X_i, Y_i) za $i = 1, \ldots n$, kjer je n število enot. Naša ničelna domneva je zopet, da sta povprečji enaki, zapišimo jo kot $H_0 : \mu_X = \mu_Y$.
Bistvena razlika od situacije z neodvisnima vzorcema je, da so enote odvisne, označimo $\sigma_{XY} = \mathrm{cov}(X, Y)$, oziroma $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$.
Testna statistika bo zopet

$$T = \frac{\bar{X} - \bar{Y}}{\widehat{SE}}$$

bistvena razlika od situacije z neodvisnimi vzorci pa je pri izračunu standardne napake. Varianco razlike namreč sedaj zapišemo kot

$$\begin{aligned}
\mathrm{var}(\bar{X} - \bar{Y}) &= \mathrm{var}(\bar{X}) + \mathrm{var}(\bar{Y}) - 2\mathrm{cov}(\bar{X}, \bar{Y}) \\
&= \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n} - 2\frac{1}{n^2}\mathrm{cov}(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} Y_j) \\
&= \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n} - 2\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{cov}(X_i, \sum Y_j) \\
&= \frac{1}{n}\left[\sigma_X^2 + \sigma_Y^2 - 2\frac{1}{n}\sum_{i=1}^{n}\mathrm{cov}(X_i, Y_i)\right] \\
&= \frac{1}{n}\left[\sigma_X^2 + \sigma_Y^2 - 2\mathrm{cov}(X_i, Y_i)\right]
\end{aligned}$$

Vidimo torej, da je ob pozitivni povezanosti med enotami standardna napaka manjša. Kadar sta varianci σ_X^2 in σ_Y^2 enaki, dobimo

$$\mathrm{var}(\bar{X}_1 - \bar{Y}_2) = \frac{1}{n}\left[\sigma^2 + \sigma^2 - 2\rho\sigma^2\right] = \frac{2\sigma^2}{n}(1 - \rho)$$

Jasno je torej, da bolj kot je korelacija pozitivna, bolj se nam splača načrtovati raziskavo tako, da so enote v parih.

Zgoraj smo napisali testno statistiko, kako v tem primeru ocenimo standardno napako iz podatkov? Najprej opazimo, da velja

$$\bar{X} - \bar{Y} = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i = \overline{X - Y}$$

Definiramo torej novo slučajno spremenljivko $D_i = X_i - Y_i$. Velja $E(D) = \mu_X - \mu_Y = \mu_D$, torej preizkušamo ničelno domnevo $H_0 : \mu_D = 0$. Pri predpostavki, da sta X in Y normalno porazdeljeni spremenljivki, ki sta robni

spremenljivki neke multivariatne normalne, je tudi D normalno porazdeljena.

Vidimo, da lahko problem odvisnih vzorcev prevedemo na test t za en vzorec. Standardno napako ene slučajne spremenljivke seveda že znamo oceniti, prav tako vemo, da je testna statistika porazdeljena s porazdelitvijo t. Ko preverjamo predpostavke, nas zanima le predpostavka normalnosti D, spremenljivki X in Y nas vsaka zase ne zanimata.

3.3 Razmerje verjetij

Kadar se odločamo med dvema enostavnima domnevama, lahko smiselno testno statistiko zapišemo z razmerjem verjetij.

Primer:

Vrnimo se k primeru dveh virov signalov. Imamo naslednji domnevi

H_0: Signal prihaja iz vira z gostoto $f_0(x)$ in *H_A: Signal prihaja iz vira z gostoto $f_A(x)$*

Denimo, da je signal pod ničelno domnevo porazdeljen kot $N(0, \sigma^2)$, pod alternativno pa $N(a, \sigma^2)$.

Oglejmo si razmerje gostot - bolj ko bo razmerje različno od 1, bolj bomo prepričani v eno izmed domnev:

$$\begin{aligned}
T &= \prod_{i=1}^{n} \frac{f_A(X_i)}{f_0(X_i)} \\
&= \prod_{i=1}^{n} \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(X_i-a)^2}{2\sigma^2}\}}{\frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(X_i)^2}{2\sigma^2}\}} \\
&= \prod_{i=1}^{n} \frac{\exp\{-\frac{(X_i-a)^2}{2\sigma^2}\}}{\exp\{-\frac{(X_i)^2}{2\sigma^2}\}} \\
&= \exp\{-\sum_{i=1}^{n} \frac{(X_i-a)^2 - (X_i)^2}{2\sigma^2}\} \\
&= \exp\{-\sum_{i=1}^{n} \frac{-2ax_i + a^2}{2\sigma^2}\} \\
&= \exp\{\sum_{i=1}^{n} \frac{2ax_i - a^2}{2\sigma^2}\} \\
&= \exp\{\frac{na}{\sigma^2}\left(\bar{X} - \frac{a}{2}\right)\}
\end{aligned}$$

Da bi postavili mejo zavrnitve oziroma znali izračunati vrednost p, moramo poznati porazdelitev testne statistike pod ničelno domnevo. Ker se zdi zgornja testna statistika nekoliko zapletena, jo seveda lahko transformiramo - vsaka bijektivna preslikava bo imela identične lastnosti. Definirajmo torej

$$Y = \frac{na}{\sigma^2}\left(\bar{X} - \frac{a}{2}\right)$$

Večja kot bo vrednost Y, bolj bo to dokaz proti ničelni domnevi. Da bi vedeli, katere vrednosti so 'velike', moramo poznati porazdelitev slučajne spremenljivke Y. Pod ničelno domnevo so vrednosti X_i standardno normalno porazdeljene. Ker so a in σ^2 konstante (znane vrednosti), je Y linearna kombinacija neodvisnih normalnih spremenljivk in zato normalno porazdeljena. Sedaj bi lahko nadaljevali na dva načina - lahko poiščemo povprečje in standardni odklon spremenljivke Y, lahko pa testno statistiko še poenostavimo:

$$Z = \frac{\bar{X}}{\sigma/\sqrt{n}}$$

Testna statistika Z je standardizirana normalna spremenljivka. Vidimo tudi, da smo dobili enako testno statistiko kot smo jo 'uganili' že ob prejšnjem obisku tega primera.

.

Za testno statistiko, ki je enaka razmerju verjetij (oziroma je bijektivna preslikava tega) velja izrek:

Neyman-Pearsonova lema Naj bosta H_0 in H_A enostavni domnevi. Testna statistika T naj bo enaka razmerju verjetij, njeno mejo zavrnitve označimo s c, velikost testa pa z α, zavračamo za velike vrednosti T. Potem ima vsaka druga testna statistika velikosti manj ali enako α manjšo moč kot T.

3.3.1 Posplošeni test razmerja verjetij

Podobno kot pri iskanju cenilk za nek populacijski parameter (metoda največjega verjetja) tudi pri preverjanju domnev obstaja generična metoda, ki poda testno statistiko in njeno porazdelitev. Imenuje se posplošeni test razmerja verjetij.
Omenili smo že, da je razmerje verjetij smiselna testna statistika, ki ima v primeru enostavnih domnev tudi lepo lastnost. Če sta obe domnevi enostavni, poznamo obe porazdelitvi - v števcu je verjetje (produkt gostot) pod alternativno domnevo, v imenovalcu pa verjetje pod ničelno domnevo.
Če katera izmed domnev ni enostavna, verjetja ne moremo zapisati, saj gostote ne poznamo - možnih je več vrednosti verjetja, pač za vse parametre, ki so možni pod tisto domnevo. Ideja posplošenega testa razmerja verjetij

je, da zapišemo verjetje v tisti vrednosti parametra (ali parametrov), za katero je to verjetje največje. V imenovalcu zapišemo maksimum verjetja pod ničelno domnevo, v števcu pa največje verjetje v celem prostoru vrednosti parametrov, s formulo:

$$\Lambda = \frac{\sup_{\theta \in \Theta} L(x, \theta)}{\sup_{\theta \in \Theta_0} L(x, \theta)}$$

Vrednost Λ bo vedno večja od 1, ničelno domnevo bomo zavrnili za velike vrednosti. Testno statistiko Λ imenujemo Wilksov Λ.
Da bi lahko poiskali ustrezno mejo zavrnitve, moramo poznati porazdelitev Λ. Izkaže se, da vrednost $2 \log \Lambda$ pod ničelno domnevo konvergira k porazdelitvi χ^2_k, kjer je k razlika v številu parametrov, ki jih ocenjujemo pod ničelno in pod alternativno domnevo, torej dimenzija Θ minus dimenzija Θ_0.

Primer:

Poglejmo si, kako zapišemo posplošeni test razmerja verjetij za primer IQ-ja: Ničelna domneva $H_0 : \mu = 100$, alternativna domneva $H_A : \mu \neq 100$. Ničelna domneva je torej enostavna, gostoto poznamo in ker je parameter samo eden, je supremum v imenovalcu odveč - imenovalec je enak tistemu iz enostavnega testa razmerja verjetij: v imenovalcu torej imamo

$$\sup_{\theta \in \Theta_0} L(x, \theta) = L(x, \theta_0) = \prod_{i=1}^{n} f_0(x_i) = \prod_{i=1}^{n} f(x_i, \mu = 100)$$

Kaj pa v števcu? Ker je alternativna domneva sestavljena, bi lahko verjetje zapisali v katerikoli vrednosti μ. V kateri vrednosti μ je verjetje na naših podatkih največje - v tisti vrednosti, ki jo s podatkov ocenimo z metodo največjega verjetja. Cenilka po metodi največjega verjetja za μ je ravno vzorčno povprečje, števec je torej enak

$$\sup_{\theta \in \Theta} L(x, \theta) = L(x, \bar{x})$$

Vrednost Λ je torej enaka

$$\begin{aligned}\Lambda &= \frac{\sup\limits_{\theta\in\Theta} L(x,\theta)}{\sup\limits_{\theta\in\Theta_0} L(x,\theta)} = \frac{L(x,\bar{x})}{L(x,\theta_0)} \\ &= \frac{\prod\limits_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x_i-\bar{x})^2}{2\sigma^2}\}\right)}{\prod\limits_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x_i-100)^2}{2\sigma^2}\}\right)}\end{aligned}$$

.

Primer:

Poglejmo si, kako uporabimo posplošeni test razmerja verjetij v linearni regresiji:
Predpostavljamo, da podatki izhajajo iz modela

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon$$

kjer je $\epsilon \sim N(0,\sigma^2)$, porazdelitev$Y$ pri neki vrednosti x je $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$.
Naša ničelna domneva je $H_0 : \beta_1 = 0$, alternativna pa $\beta_1 \neq 0$. Alternativna domneva je sestavljena, enako velja za ničelno, saj β_0 ne poznamo in gostota torej ni natanko določena.
Velja torej:
Pod ničelno domnevo je $Y \sim N(\beta_0, \sigma^2)$, pod alternativno domnevo $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$.
Pod alternativno domnevo torej za zapis gostote in s tem verjetja potrebujemo oceni β_0 in β_1. Dobimo ju po metodi največjega verjetja (pokazali smo že, da sta to enaki oceni kot tisti, ki ju dobimo po metodi najmanjših kvadratov). Označimo ju z $\widehat{\beta}_0^A$ in $\widehat{\beta}_1^A$. Pod ničelno domnevo za zapis verjetja potrebujemo oceno β_0. Ta parameter tu enostavno predstavlja populacijsko povprečje, vemo že, da je cenilka po metodi največjega verjetja zanj ravno vzorčno povprečje. Zapišimo torej Wilksov Λ:

$$\begin{aligned}\Lambda &= \frac{\sup\limits_{\theta\in\Theta} L(y,\theta)}{\sup\limits_{\theta\in\Theta_0} L(y,\theta)} = \frac{L(y,\widehat{\beta}_0^A,\widehat{\beta}_1^A)}{L(y,\widehat{\beta}_0^0)} \\ &= \frac{\prod\limits_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(y_i-(\widehat{\beta}_0^A+\widehat{\beta}_1^A y_i))^2}{2\sigma^2}\}\right)}{\prod\limits_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(y_i-\bar{y})^2}{2\sigma^2}\}\right)}\end{aligned}$$

Ker smo v števcu ocenili 2 parametra, v imenovalcu pa enega, je razlika v številu ocenjevanih parametrov 1 in zato $2\log\Lambda \sim \chi_1^2$. .

Opomba: Pogosto bo ulomek pri definiciji Λ obrnjen (npr. v Rice-u), potem je $-2\log\Lambda$ porazdeljena s χ^2. Ker je vrednost ulomka omejena z 1, v praksi tako ali tako nikoli ne bo dvoma, saj bodo za logaritem testne statistike možne le pozitivne oz. le negativne vrednosti.

Oris dokaza za en ocenjevani parameter in enostavno ničelno domnevo:

- Pod ničelno domnevo je ocena parametra v celotnem prostoru $\hat{\theta}$ blizu pravi vrednosti (θ^0). Logaritem funkcije verjetja v pravi vrednosti parametra razvijemo v Taylorjevo vrsto

$$\sum_{i=1}^{n} logf(\theta^0) = \sum_{i=1}^{n} logf(\hat{\theta}) + (\theta^0 - \hat{\theta})\sum_{i=1}^{n} logf'(\hat{\theta}) + \frac{(\theta^0-\hat{\theta})^2}{2}\sum_{i=1}^{n} logf''(\hat{\theta}) + ost.$$

- Prvi odvod logaritma verjetja je funkcija zbira, ta je enak 0 za ocenjeno vrednost parametra.

- $\frac{1}{n}\sum\limits_{i=1}^{n} logf''(\hat{\theta})$ konvergira proti $E(logf''(\hat{\theta}))$

- Iz teorije največjega verjetja vemo, da varianco cenilke lahko aproksimiramo z $\mathrm{var}(\hat{\theta}) = -(nE(logf''(\hat{\theta})))^{-1}$

- Cenilka $\widehat{\theta}$ je normalno porazdeljena (cenilka po metodi največjega verjetja), pod ničelno domnevo je blizu θ^0, pričakovana vrednost razlike je enaka 0.

- Kvadratni člen smo tako standardizirali.

- Polovico odpravimo z dvakratnikom logaritma, kvadrat normalne porazdelitve je porazdeljen kot χ^2.

- Najtežji del dokaza je pokazati, da so nadaljnji členi razvoja zanemarljivi in pa razširitev na sestavljeno ničelno domnevo.

Primer:

Poglejmo si, kaj da posplošeni test razmerja verjetij v preprostem primeru, ko sicer porazdelitev testne statistike že poznamo (nadaljevanje primera zgoraj). Naj bo ničelna domneva $H_0 : E(X) = \mu_0$, predpostavljamo, da je $X \sim N(0, \sigma^2)$. Alternativna domneva je $\mu \neq \mu_0$. Prostor vrednosti parametra μ je torej $\Theta = \{\mu; \mu \in (-\infty, \infty)\}$, pod ničelno domnevo pa $\Theta_0 = \{\mu_0\}$. Zapišimo Wilksov Λ, upoštevajmo, da je povprečje cenilka za μ po metodi največjega verjetja, in da nam pod ničelno domnevo ni potrebno oceniti

nobenega parametra:

$$\begin{aligned}
\Lambda &= \frac{\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x_i-\widehat{\mu})^2}{2\sigma^2}\}\right)}{\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x_i-\mu_0)^2}{2\sigma^2}\}\right)} \\
&= \frac{\prod_{i=1}^{n}\left(\exp\{-\frac{(x_i-\bar{x})^2}{2\sigma^2}\}\right)}{\prod_{i=1}^{n}\left(\exp\{-\frac{(x_i-\mu_0)^2}{2\sigma^2}\}\right)} \\
&= \frac{\exp\{-\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{2\sigma^2}\}}{\exp\{-\sum_{i=1}^{n}\frac{(x_i-\mu_0)^2}{2\sigma^2}\}} \\
&= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x_i-\bar{x})^2-(x_i-\mu_0)^2]\} \\
&= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x}-x_i+\mu_0)(x_i-\bar{x}+x_i-\mu_0)\} \\
&= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-\bar{x}+\mu_0)(2x_i-\bar{x}-\mu_0)\} \\
&= \exp\{-\frac{1}{2\sigma^2}(-\bar{x}+\mu_0)\sum_{i=1}^{n}(2x_i-\bar{x}-\mu_0)\} \\
&= \exp\{-\frac{1}{2\sigma^2}(-\bar{x}+\mu_0)\sum_{i=1}^{n}(\bar{x}-\mu_0)\} \\
&= \exp\{\frac{(\bar{x}-\mu_0)^2 n}{2\sigma^2}\}
\end{aligned}$$

Velja torej

$$2\log\Lambda = \frac{(\bar{x}-\mu_0)^2}{\sigma^2/n}$$

Teorija posplošenega testa razmerja verjetij pravi, da je $2\log\Lambda$ porazdeljen kot χ^2_1. Asimptotska porazdelitev je v tem primeru tudi enaka natančni

- vemo, da je

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1),$$

kvadrat te testne statistike je potem porazdeljen kot χ^2_1.

DN: Oglejte si še, kakšno testno statistiko dobimo, če vrednost σ ni znana.

.

Opomba: da bi lahko uporabili posplošeni test razmerja verjetij, mora biti k seveda enak vsaj 1. S tem testom torej ne moremo odločati med domnevama $\mu \leq \mu_0$ in $\mu > \mu_0$.

Oglejmo si še nekaj primerov uporabe posplošenega testa razmerja verjetij.

Primer:

Zanima nas, kako se moški in ženske v srednjih letih razlikujejo glede kadilskega statusa. Naj bodo možni štirje odgovori: 'nikoli nisem kadil', 'kadim in ne skušam prenehati', 'sem bivši kadilec', 'skušam prenehati, a do sedaj neuspešno'. Označimo slučajno spremenljivko, ki ponazarja odgovore z X, odgovore pa z 1,2,3 in 4. Zanima nas ali je porazdelitev spremenljivke X enaka pri moških kot pri ženskah, v vzorec naberimo n predstavnikov vsakega spola.
Oznake: število posameznih odgovorov žensk: z_1, z_2, z_3, z_4, število posameznih odgovorov moških m_1, m_2, m_3, m_4.
Naj bodo prave verjetnosti žensk $p_{z1}, p_{z2}, p_{z3}, p_{z4}$, prave verjetnosti moških pa $p_{m1}, p_{m2}, p_{m3}, p_{m4}$. Preveriti želimo ničelno domnevo

$$H_0 : p_{z1} = p_{m1}, p_{z2} = p_{m2}, p_{z3} = p_{m3}, p_{z4} = p_{m4}$$

Alternativna domneva je, da je vsaj ena od verjetnosti različna.
Zapišimo testno statistiko:
Najprej si oglejmo verjetje na celotnem prostoru parametrov. Vsi $p$ji so med seboj lahko različni, so pa omejeni z vsoto - vsota pri ženskah in pri moških mora biti enaka 1. Kako bi zapisali verjetnost posameznih opaženih dogodkov?

$$P(X_1 = 1, X_2 = 3, X_3 = 7, \ldots, Y_1 = 2, Y_2 = 5, \ldots) = p_{z1}^{z_1} \cdot p_{z2}^{z_2} \cdot \ldots p_{z4}^{z_4} \cdot p_{m1}^{m_2} \cdot \ldots \cdot p_{m4}^{m_4}$$

To je torej verjetje na celotnem prostoru parametrov Θ
Kaj pa pod ničelno domnevo. Prostor parametrov pod ničelno domnevo dovoljuje le štiri verjetnosti, označimo jih z p_1 do p_4, sešteti se morajo v 1. Verjetje je enako

$$P(X_1 = 1, X_2 = 3, X_3 = 7, \ldots, Y_1 = 2, Y_2 = 5, \ldots) = p_1^{z_1+m_1} \ldots p_4^{z_4+m_4}$$

Kako bi ocenili parametre pod ničelno domnevo oz. na celem prostoru?
Izkaže se, da so ocene na celotnem prostoru po metodi največjega verjetja enake kar deležem: $\hat{p}_{zk} = \frac{z_k}{n}$. Pod ničelno domnevo pa podobno: $\hat{p}_k = \frac{z_k+m_k}{2n}$.
Zapišimo Λ

$$\Lambda = \frac{\hat{p}_{z1}^{z_1} \hat{p}_{z2}^{z_2} \ldots \hat{p}_{m4}^{m_4}}{\hat{p}_1^{z_1+m_1} \ldots \hat{p}_4^{z_4+m_4}}$$

Koliko stopinj prostosti imamo? Zavedati se moramo, da v resnici ne ocenjujemo vseh štirih parametrov temveč le 3, saj se mora vse sešteti v 1. Torej jih je na celotnem prostoru 6, pod ničelno pa trije, dobimo torej $\chi^2_{(3)}$.

.

3.4 Presojanje lastnosti testov

Pri testu razmerja verjetij smo že omenili, kako bomo presojali lastnosti testov oz. kakšno testno statistiko bomo želeli poiskati.
Da bi bile alternativne možnosti za testno statistiko med seboj lažje primerljive, bomo pri vseh določili mejo zavrnitve tako, da bo enaka α (oziroma manjša enaka α v diskretnih primerih). Potem nas bo zanimalo, katera izmed testnih statistik ima pri določeni alternativni domnevi največjo moč. Kot vidimo, je razmerje verjetij dober kandidat za testno statistiko, a žal Neyman-Pearsonova lema velja le za enostavne domneve.
Medtem ko je ničelna domneva pogosto enostavna (vrednost parametra je 0), je alternativna skoraj vedno sestavljena. Testno statistiko, ki ima največjo moč za katerokoli enostavno alternativno domnevo, imenujemo 'enakomerno najmočnejša' (uniformly most powerful).

Za lažjo predstavo naštejmo nekaj primerov, nekateri sicer presegajo naše trenutno znanje, a kažejo na situacije, v katerih je možnih več alternativnih domnev.

- Preverjali smo, ali imajo kolesarji drugačno porazdelitev od drugih vzdržljivostnih športnikov. Primerjali smo le povprečji, predpostavili smo enaki (znani) varianci. Tudi, če predpostavke ne bi naredili, bi še vedno preverjali le enakost povprečij, s testom ne bi imeli moči za primer enakih povprečij a različnih varianc.

- Pri sestavljenih alternativnih domnevah bo testna statistika lahko imela precejšnjo moč za nek podprostor parametrov v alternativni domnevi, medtem ko za druge ne bo imela moči.
  Denimo, da nas zanima povezanost med dvema spremenljivkama. Linearno povezanost bomo lahko preverili s testom, ki temelji na Pearsonovem koeficientu. Ne bomo pa imeli nikakršne moči za odkrivanje povezanosti v obliki črke U.

- Neparametrični testi, npr. test Mann-Whitney, preverjajo ničelno domnevo, da sta porazdelitvi enaki. Seveda se porazdelitvi lahko razlikujeta na več načinov, alternativnih domnev je torej veliko. Test Mann-Whitney bo imel lahko precejšnjo moč za dve normalni porazdelitvi z različnim (premaknjenim) povprečjem, nima pa nikakršne moči za dve normalni porazdelitvi z enakim povprečjem, a različno varianco.

- V analizi preživetja nas zanima ničelna domneva, da je nek parameter v času konstanten (razmerje tveganj dveh posameznikov). Tudi tu je možnih veliko alternativnih domnev - lahko v resnici monotono narašča oziroa pada, lahko pa niha gor in dol. Testni statistiki, ki sta občutljivi za eno oziroma drugo alternivno domnevo se precej razlikujeta.

3.4.1 Vrednost p kot slučajna spremenljivka

Je vrednost p slučajna spremenljivka? Seveda - na vsakem vzorcu dobimo drugo vrednost. Vrednost p je slučajna spremenljivka, ki lahko zavzame vrednosti med 0 in 1. Če je slučajna spremenjlivka, potem ima tudi neko porazdelitev. Poizkusimo povedati kaj o tej porazdelitvi. Naj bo T naša testna statistika, zaradi preprostosti vzemimo, da sta ničelna in alternativna domneva enostavni, torej, da je porazdelitev testne statistike pri obeh domnevah natanko določena.
Denimo, da velja ničelna domneva. Kakšna je verjetnost, da bo $p \leq \alpha$? S kakšno verjetnost bomo pri neki vrednosti α zavrnili ničelno domnevo? Ta verjetnost je ravno α, saj smo ga tako definirali. Torej $P(p \leq \alpha) = \alpha$ in to za vsak α med 0 in 1. Vidimo, da je vrednost p pod ničelno domnevo porazdeljena enakomerno.

Izpeljimo porazdelitev vrednosti p pod ničelno domnevo še formalno (recimo, da je ničelna domneva enostavna, imamo zvezno testno statistiko, zavračamo za velike vrednosti): pod ničelno domnevo označimo z $F_0(t)$ kumulativno porazdelitveno funkcijo slučajne spremenljivke T ($F_T(t) = F_0(t)$), velja $p = P_0(T > t) = 1 - F_0(t)$. Vrednost p obravnavamo kot slučajno spremenljivko, zato pišimo $\mathcal{P} = 1 - F_0(T)$.

$$\begin{aligned} F_{\mathcal{P}}(x) &= P(\mathcal{P} \leq x) = 1 - P(\mathcal{P} > x) = 1 - P(1 - F_0(T) > x) \\ &= 1 - P(F_0(T) \leq 1 - x) = 1 - P(T \leq F_0^{-1}(1 - x)) \\ &= 1 - P(T \leq F_0^{-1}(1 - x)) = 1 - F_0(F_0^{-1}(1 - x)) = 1 - (1 - x) = x \end{aligned}$$

Dokaz lahko tudi skrajšamo (uporabimo probability integral transform): po definiciji je $\mathcal{P} = 1 - F_0(T)$, ker je pod ničelno domnevo $F_T = F_0$ in je $F_T(T)$ enakomerno porazdeljena spremenljivka, velja to tudi za p.

Kaj pa pod alternativno domnevo? Če ničelna domneva ne velja, seveda upamo, da bomo zavrnili večkrat kot z verjetnostjo α, sicer ima testna statistika bolj malo smisla. Torej, pod alternativno domnevo porazdelitev ni

enakomerna, temveč pričakujemo, da je vsaj nekoliko asimetrična v desno. Bolj kot je pomaknjena proti levi, večjo moč ima naš test.

Opomba: če je ničelna domneva sestavljena, p vrednosti niso enakomerno porazdeljene pod ničelno domnevo, da se pokazati, da je njihova porazdelitev stohastično večja od enakomerne.

3.4.2 Lastnosti testov na majhnih vzorcih

Centralni limitni izrek nam v mnogih situacijah omogoča, da se izognemo predpostavkam o porazdelitvi. Vendar pa vemo, da centralni limitni izrek poda porazdelitev le za $n \to \infty$, porazdelitev testne statistike na majhnih vzorcih lahko precej odstopa. Meja zavrnitve, pri kateri je napaka I. stopnje za $n \to \infty$ enaka α, pri majhnih vzorcih lahko da drugačne rezultate, ničelno domnevo lahko zavrnemo premalokrat ali prevečkrat. Pogosto približki niti ne bodo tako zelo slabi, vseeno pa je tudi zaradi tega razlikovanje med vrednostima $p = 0{,}051$ in $p = 0{,}049$ nesmiselno.
Recimo, da je populacija porazdeljena z eksponentno porazdelitvijo in da ničelna domneva govori o povprečju populacije. Smiselna testna statistika bo vzorčno povprečje, njeno porazdelitev lahko aproksimiramo z normalno. Vendar pa aproksimacija za majhne vzorce lahko ni optimalna, dejanske verjetnosti so nekoliko drugačne od tistih za normalno porazdelitev.

Eksaktne porazdelitve testne statistike marsikdaj ne poznamo, saj bi morali v ta namen poznati porazdelitev neke slučajne spremenljivke v populaciji. Pri izpeljavi porazdelitve testne statistike si tako pomagamo s centralnim limitnim izrekom, zato je porazdelitev testne statistike le asimptotska. Primer te uporabe je npr. posplošeni test razmerja verjetij (ki v dokazu uporabi CLT). V takem primeru se seveda najprej vprašamo, ali je velikost testa vsaj približno prava tudi na majhnih vzorcih, torej ali nam tudi na majhnih vzorcih test zavrača z verjetnostjo α.
To je tudi primer, ko večinoma uporabljamo izraz 'velikost testa': postavimo neko vrednost α in glede na to vrednost izračunamo mejo zavrnitve s pomočjo asimptotske porazdelitve testne statistike. Potem s simulacijami izračunamo delež zavrnitev pri tej meji - temu pravimo velikost testa. Seveda želimo, da bi bila velikost testa čim bližje α. Če velikost testa ni natanko enaka α, si želimo, da bi bila manjša od α, torej da se motimo kvečjemo manjkrat kot z verjetnostjo α, v tem primeru pravimo, da je test konzervativen. Če bi uporabnik testa namreč bil prepričan, da je napaka prve vrste enaka α, a je v

resnici večja, bi bilo to zavajujoče - poročali bi o mnogih statistično značilnih rezultatih, ki to niso. Bo pa pri konzervativnih testih seveda nekoliko manjša moč.
Premislimo še, kako se napačna velikost testa kaže na intervalih zaupanja. Spomnimo se: interval zaupanja je množica tistih vrednosti parametra, ki ga preizkušamo, pri katerih ničelne domneve ne bi zavrnili:

- Če test zavrača premalokrat: interval zaupanja bo širši, kot bi lahko bil. Čeprav bomo izračunali 95% interval zaupanja (s pomočjo asimptotske porazdelitve), bomo v resnici imeli nekoliko večje zaupanje, da smo pokrili pravi rezultat. S tem seveda ni nič narobe, le naša natančnost je nekoliko manjša (širina IZ)

- Če test zavrača prevečkrat: interval zaupanja bo ožji, kot bi moral biti. Verjetnost, da nam uspe pokriti pravo vrednost bo premajhna, rezultati bodo zavajujoči - izgledalo bo, da imamo o pravi vrednosti več informacije kot je res, večja bo verjetnost, da smo pravo vrednost zgrešili.

3.4.3 α in β pri diskretnih slučajnih spremenljivkah

Pri diskretnih slučajnih porazdelitvah je (vsaj pri majhnih vzorcih) pogosto nemogoče postaviti mejo zavrnitve tako, da bo velikost testa natanko α.

Primer:

Kot primer si poglejmo test, s katerim preverjamo ali je kovanec pošten, torej verjetnost obeh izidov enaka $p = 0{,}5$. Denimo, da kovanec vržemo 10x, naša velikost vzorca je torej $n = 10$, testna statistika pa $\bar{X}$. Želimo postaviti mejo zavrnitve, da bo velikost testa enaka α:
Izkaže se, da velja

$$P_0(\bar{X} \geq 9) = 0{,}01,\ P_0(\bar{X} \geq 8) = 0{,}055$$

V nobenem primeru velikost testa torej ne bo natanko enaka 0,05. Kot mejo zavrnitve moramo izbrati tisto vrednost, pri kateri je verjetnost napake I. stopnje še vedno pod 0,05. .

3.4.4 Post-hoc računanje moči

Za osvežitev teme o moči najprej premislimo naslednji primer ____**Primer:** Naj bo velikost testa natanko enaka α. Recimo, da imamo situacijo, v kateri je moč majhna. Kaj je majhna moč? Kakšna je najmanjša možna moč? Premislimo na primeru primerjave dveh povprečij. Alternativna domneva pravi, da je razlika med povprečjema različna od nič. Večja kot je razlika, večja bo moč. Ko se razlika manjša (gre proti 0), se moč manjša. Kakšna je moč, kadar je razlika enaka 0? Moč je verjetnost, da zavrnemo ničelno domnevo, v situaciji, ko je razlika enaka 0, je torej enaka α. Če je razlika malo različna od nič, je moč malo večja od α. Infimum je torej enak α.
Vsaka smiselna testna statistika velikosti α bo torej imela moč vsaj α. Vendar pa to ne velja za diskretno porazdeljene testne statistike, pri katerih je tudi verjetnost zavrnitve pod ničelno domnevo v resnici manjša od α. .____

Recenzenti v znanstvenih revijah pogosto dobro raziskavo opredelijo kot tako, ki ima pred začetkom izračunano moč, oz. ki je načrtovana tako, da je izračunana potrebna velikost vzorca za neko željeno moč (ponavadi 0,8). Marsikatera raziskava žal ni narejena na ta način, raziskovalci zberejo vzorec, ki jim je na voljo, in upajo na najboljše. Pogosto se zgodi, da recenzenti pri takih raziskavah zahtevajo naknadno računanje moči, s čemer želijo razumeti vrednost raziskave. Premislimo:

- Recimo najprej, da je bil rezultat neznačilen. Kot smo že omenili, je rezultat lahko neznačilen zaradi majhnosti vzorca ali zaradi dejstva, da v populaciji dejansko ni razlik. Pri interpretaciji je ključnega pomena interval zaupanja, ki nam jasno pokaže ali je smiselno zbrati večji vzorec (torej ali je bila moč premajhna)

- Kaj pa če je bil rezultat značilen. Vprašajmo se malce drugače: recimo, da imamo majhno moč, da bi odkrili neko razliko, a jo vendarle odkrijemo. Ali nam to nakazuje, da moramo naše zaključke interpretirati bolj pazljivo? Ne, če je rezultat značilen, potem ni važno, kakšna je bila moč. Moč je verjetnost, ki je relevantna pred začetkom raziskave. Ko dobimo rezultat, nas ta verjetnost ne zanima več, zanima nas rezultat. Seveda je pri interpretaciji zopet koristen interval zaupanja - če smo imeli majhno moč, bo verjetno precej širok in bo morda zajemal tudi strokovno nepomembne vrednosti.

Povzemimo: vso informacijo nam podajajo intervali zaupanja, dodatno naknadno računanje moči nam ne prinese ničesar novega in je irelevantno. Je pa res, da je v nekaterih primerih intervale zaupanja težje definirati (npr. neparametrični testi), zato je potrebno tam posebej premisliti o interpretaciji.

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of. (R.A.Fisher)

3.5 Test Mann-Whitney

Vsak test ima seveda svoje predpostavke, osnovna predpostavka pri testu t je normalna porazdelitev spremenljivke, ki nas zanima. Pri testu t nas zanima parameter porazdelitve, zato pravimo, da je ta test parametričen. V tem razdelku si bomo ogledali primer *neparametričnega* testa, ki ga lahko razumemo kot alternativo testu t.

Ideja testa Mann-Whitney je, da vrednosti opažene na vzorcu nadomestimo z rangi. Pri tem združimo obe skupini (oba vzorca) in rangiramo vse enote. Najmanjši enoti dodelimo rang 1, morebitnim vezem (ties - enake vrednosti) damo vmesne range. Nato si ogledamo range v posamezni skupini. V nadaljevanju bomo povsod predpostavili situacijo brez vezi (ki jo pričakujemo v primeru zveznih porazdelitev), čeprav bodo vsi zaključki delovali tudi v primeru zmernega deleža vezi.
Ničelna domneva pri testu Mann-Whitney je zelo splošna:

$$H_0 : \text{porazdelitvi, iz katerih izhajata vzorca, sta enaki.}$$

Lahko zapišemo tudi $H_0 : F = G$, kjer F in G označujeta porazdelitveni funkciji.

Če ničelna domneva drži, imajo vse enote enake verjetnosti za vsakega izmed rangov. Če imamo dva vzorca velikosti m in n, so torej možni vsi rangi od 1 do $m + n$.

Smiselna izbira za testno statistiko je vsota rangov v eni skupini (oz. povprečje). Označimo z $X_1, \ldots, X_n$ in $Y_1, \ldots, Y_m$ vrednosti na vzorcih, testna statistika naj bo

$$T_Y = \sum_{i=1}^{m} Y_i$$

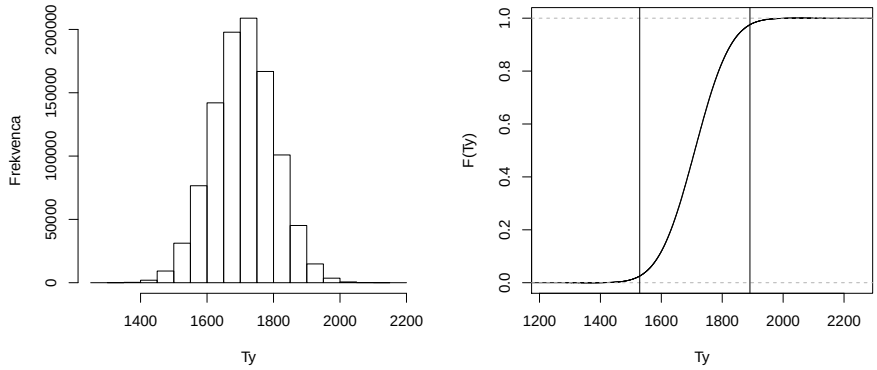
Pri tem je dovolj, da gledamo le vsoto v eni skupini, saj je druga vsota s tem enolično določena - z eno od vsot smo že zajeli celotno informacijo, ki jo dajejo podatki.
Porazdelitev vrednosti testne statistike lahko za neka dana m in n seveda simuliramo. Tu ni nobenih predpostavk o porazdelitvi - enostavno si ogledamo pogostost posameznih vrednosti T_Y, ki jih dobimo, ko dane range naključno delimo v dve skupini. Tak graf nato primerjamo s tistim, kar se nam je dejansko zgodilo.

Primer:

Denimo, da imamo dva vzorca, velikosti $n = 30$ in $m = 45$. Oglejmo si porazdelitev vrednosti T_Y pod ničelno domnevo, generirajmo veliko število vzorcev te velikosti:

```
    > nexact <- 100000
> ty <- rep(NA,nexact){
> for(dt in 1:nexact)
> yoni <- sample(1:(m+n),size=m,replace=F)
> ty[dt] <- sum(yoni)
> }
```

Histogram vrednosti T_Y (`ty`) in empirična porazdelitvena funkcija te spremenljivke sta podana na sliki 3.1. Označeni sta vrednosti, ki odrežeta spodnjih in zgornjih 2,5 % (v našem primeru 1529 in 1891). Vrednosti znotraj teh dveh mej predstavljajo območje sprejetja za naš primer. .



Slika 3.1: *Porazdelitev vrednosti T_Y. Levo histogram, desno empirična kumulativna porazdelitvena funkcija z vrisanimi mejami zavrnitve za $\alpha = 0{,}05$.*

Kako izračunamo vrednost p? Za konkretno vrednost $T_Y = t$ na naših podatkih si ogledamo delež $P(T_Y \leq t)$ oziroma $P(T_Y \geq t)$ in ga množimo z 2 (tistega od njiju, ki predstavlja verjetnost v repu, torej je manjši od 0,5).

Testno statistiko smo tu podali empirično (z zelo velikim številom simulacij), verjetnosti posameznih izidov bi seveda lahko izpeljali tudi teoretično, kar pa bi zahtevalo precej več dela. Dobra lastnost je dejstvo, da moramo za določeni velikosti vzorcev porazdelitev izpeljati le enkrat in jo nato lahko uporabljamo. Včasih so v ta namen obstajale tabele, danes je vsekakor veliko lažje in hitreje generirati 100000 ali še več primerov (za podani izračun mej na miljon primerih je R porabil 20s), s čimer teoretične verjetnosti aproksimiramo na več decimalk natančno.

Pri večjih velikostih vzorcih bo izračun seveda nekoliko zahtevnejši, zaradi večjega števila različnih možnosti bo tudi za kvaliteten empirični izračun potrebnih več simulacij. Ali lahko kaj rečemo o aproksimativni porazdelitvi pod ničelno domnevo? Upamo lahko, da je vsota približno normalno porazdeljena, navkljub dejstvu, da ni vsota neodvisnih slučajnih spremenljivk. Če to približno velja, za poznavanje porazdelitve potrebujemo še oba parametra - pričakovano vrednost in varianco.

Izračunajmo najprej pričakovano vrednost in varianco slučajne spremenljivke R_Y (za R_{Yi} označimo rang enote i, ki ima vrednost Y_i). Vrednosti R_Y pod ničelno domnevo predstavljajo slučajni vzorec m vrednosti iz populacije z $m+n$ vrednostmi (brez ponavljanja). Velja

$$E(R_Y) = \frac{1}{m+n}\sum_{i=1}^{m+n} k = \frac{1}{m+n}\frac{(m+n)(m+n+1)}{2} = \frac{m+n+1}{2}$$

in

$$\mathrm{var}(R_Y) = \frac{1}{m+n}\sum_{i=1}^{m+n} k^2 - \frac{m+n+1}{2} = \frac{(m+n)^2-1}{12}$$

Vemo že, da je potem

$$E(\bar{R}_Y) = E(R_Y) = \frac{m+n+1}{2}$$

in zato

$$E(T_Y) = \frac{m(m+n+1)}{2}$$

varianco cenilke pa izračunamo s pomočjo dejstva

$$\mathrm{var}(\bar{R}_Y) = \frac{\sigma^2}{m}\frac{m+n-m}{m+n-1} = \frac{\sigma^2}{m}\frac{n}{m+n-1}$$

in zato

$$\begin{aligned} \text{var}(T_Y) &= m^2 \frac{(m+n)^2-1}{12m} \frac{n}{(m+n-1)} \\ &= \frac{m^2(m+n-1)(m+n+1)n}{12m(m+n-1)} \\ &= \frac{mn(m+n+1)}{12} \end{aligned}$$

Poznamo torej asimptotsko porazdelitev T_Y.

Primer:
Lahko jo primerjamo z našo empirično - asimptotsko povprečje je v našem primeru 1710, varianca 8550, dobimo meji

$$1710 \pm 1{,}96 * \sqrt{8550} = [1528{,}8; 1891{,}2]$$

Vidimo, da se porazdelitvi povsem ujemata (na desnem grafu slike 3.1 bi se krivulji povsem prekrivali, zato nista narisani obe). .

Testna statistika testa Mann-Whitney je tesno povezana s T_Y (pokazali bomo, da med njima obstaja bijektivna preslikava), ima pa tudi zanimivo interpretacijo, količina, ki nas v tej situaciji zanima, je namreč:

$$\pi = P(X < Y)$$

Če vrednosti X in Y izhajajo iz enake porazdelitve, je $\pi = 0{,}5$. Testna statistika U_Y je definirana kot

$$U_Y = \sum_{i=1}^{n} \sum_{j=1}^{m} I[X_i < Y_j]$$

Štejemo torej, kolikokrat je X_i manjši od Y_i, vrednost $U_Y/(mn)$ torej ravno ocenjuje π. Člene gornje vsote lahko premečemo in dobimo

$$U_Y = \sum_{i=1}^{n} \sum_{j=1}^{m} I[X_i < Y_{(j)}]$$

kjer $Y_{(i)}$ opisuje i-to vrednost y, ko so ti urejeni po velikosti. To pa pomeni, da je vrednost U_Y enaka številu vrednosti X, ki so manjše od najmanjšega

Y + število vrednosti X, ki so manjše od naslednjega Y, ipd.

To pa lahko z rangi zapišemo kot

$$U_Y = \sum_{j=1}^{m}(R_{Yj} - j) = T_Y - \sum_{j=1}^{m} j = T_Y - \frac{m(m+1)}{2}$$

Dobimo

$$\begin{aligned} E(U_Y) &= \frac{mn}{2} \\ \text{var}(U_Y) &= \frac{m(m+n+1)}{2} \end{aligned}$$

Testna statistika enaka je

$$T = \frac{U_Y - E(U_Y)}{\sqrt{\text{var}(U_Y)}}$$

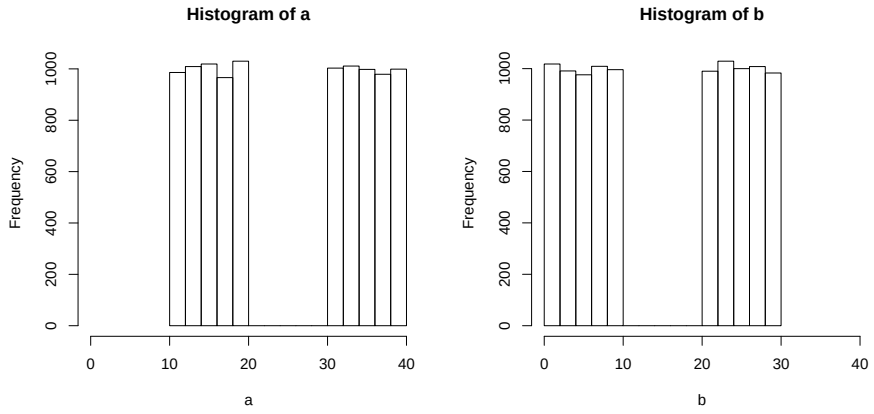
aproksimativna porazdelitev za T je standardna normalna.

Povzetek:
Test Mann-Whitney ne primerja povprečij ali kateregakoli drugega parametra, test je neparameteričen. Ne potrebuje nikakršnih predpostavk glede porazdelitev iz katere izhajajo vrednosti, vse kar potrebujemo za poznavanje porazdelitev testne statistike je velikost vzorcev. Na večjih vzorcih lahko porazdelitev testne statistike dobro aproksimiramo z normalno porazdelitvijo. Mann-Whitneyev test bi lahko interepretirali tudi kot test, ki namesto povprečij vrednosti primerja povprečje rangov. S tem je tudi očitna alternativa testu t, ki seveda potrebuje predpostavke o porazdelitvi populacije. Test Mann-Whitney se pogosto uporablja za primerjavo vzorcev iz bolj asimetričnih porazdelitev, s tem ko se namesto za konkretne vrednosti odločimo za range, se izognemo izstopajočim vrednostim. Test je primeren tudi za urejenostne opisne spremenljivke, kjer je računanje povprečij dvomljivo (razmiki med dvema kategorijama niso nujno enako veliki).
Navkljub dejstvu, da pri menjavi vrednosti za range izgubimo nekaj informacije, se izkaže, da moč testa Mann-Whitney niti pri normalnih porazdelitvah ni bistveno manjša.

Kaj lahko rečemo o pomanjkljivostih testa Mann-Whitney? Morda to, da je izračunati vrednost p preprosto, tvoriti intervale zaupanja pa precej bolj problematično. Intervali zaupanja, ki jih izpiše R (in večina drugih programov) so vezani na premik, alternativna domneva je: $H_0 : F(x) = G(x + \Delta)$. Seveda smo pri tem naredili močno predpostavko, ki ni nujno smiselna. Veliko bolj bi bil zanimiv interval zaupanja za delež vrednosti x, ki so manjše od y (ničelna domneva je 0,5), žal pa je tega precej težje izračunati. Medtem, ko je gornja testna statistika normalno porazdeljena pod ničelno domnevo, dobimo pod alternativno drugačno varianco, zato intervala zaupanja ne moremo enostavno obrniti (ne bo pravo pokritje). Računanje tega intervala zaupanja je nekoliko težje in žal ni del večine statističnih paketov. Morda pa kmalu bo ...

Pa še ena opomba:
Pogosto se test Mann-Whitney interpretira s pomočjo median. To ni res - to ni test enakosti median. Zamislimo si npr. primer:
Vzorec 1, vrednosti `a <- c(11:19,20,31:40)`
Vzorec 2, vrednosti `b <- c(1:10,21,22:31)`

Vidimo, da sta mediani enaki (21), razlika v rangih pa bo značilna, $p = 0{,}01$. Primer porazdelitev je na sliki 3.2.Skupini imata enako mediano (20), vendar sta zelo različni Mann-Whitneyev test bo imel veliko moč za opažanje teh razlik.



Slika 3.2: *Porazdelitev dveh skupin. Prva ima verjetnost* 0,501*, da je vrednost med 10 in 20, druga ima verjetnost* 0,501*, da je vrednost med 20 in 30.*

3.6 Hi-kvadrat test, goodness of fit

Primer:
Predsedniške volitve v ZDA leta 1992. Analiziramo s testom χ^2, dobimo vre-

Tabela 3.1

	Moški	Ženske	Skupaj
Clinton	467	472	939
Ostali	337	571	908
Skupaj	804	1043	1847

dnost testne statistike 29,9, vrednost $p < 0{,}001$. Uporabili smo porazdelitev χ^2 z eno stopinjo prostosti. Zakaj je to prava testna statistika, kakšen je dokaz v ozadju? .

Izkaže se, da dokaz porazdelitve pri testu χ^2 zahteva nekaj dela, pokazali bomo sorodno testno statistiko - ki nam jo da posplošeni test razmerja verjetij.

Ničelna domneva, ki jo preverjamo v gornjem primeru (2x2), je, da je delež Clintonovih volilcev neodvisen od spola. Zapisali jo bomo nekoliko bolj splošno (ter tako uvedli nekoliko bolj splošno testno statistiko) in sicer $H_0 : \pi = \pi(\theta)$, kjer je π vektor parametrov - členi tega vektorja π_i so verjetnosti v posameznih celicah naše tabele. Pod ničelno domnevo vse verjetnosti izrazimo s parametrom θ (oz. nekaj parametri θ).
V našem primeru (2x2), imamo štiri celice in zato štiri parametre, ničelna domneva jih izrazi vse z eno količino:

$$H_0 : \pi_1 = \pi_2 = \theta;\ \pi_3 = \pi_4 = (1 - \theta)$$

Na celotnem prostoru parametrov izrazimo vrednosti z dvema količinama (parametri so vezani po stolpcih):

$$H_A \cup H_0 : \pi_1 = \theta_1; \pi_2 = \theta_2\ \pi_3 = (1 - \theta_1);\ \pi_4 = (1 - \theta_2)$$

Vpeljimo nekaj oznak: naj bodo $O_1 = n_{11}$, $O_2 = n_{12}$, $O_3 = n_{21}$, $O_4 = n_{22}$ opazovane frekvence (zato oznaka O) vrednosti v vsaki celici. Vsoti frekvenc po stolpcih naj bosta označeni z $n_{.1}$ in $n_{.2}$, vsoti frekvenc po vrsticah pa $n_{1.}$ in $n_{2.}$.

Poglejmo si oceni parametrov po metodi največjega verjetja: verjetje zapišemo kot

$$L(O, \pi) = \prod_{i=1}^{k} \pi_i^{O_i}$$

- Ničelna domneva:

$$\begin{aligned} L(\theta) &= \theta^{O_1}\theta^{O_2}(1-\theta)^{O_3}(1-\theta)^{O_4} \\ l(\theta) &= (O_1 + O_2)\log\theta + (O_3 + O_4)\log(1-\theta) \\ l'(\theta) &= \frac{O_1 + O_2}{\theta} - \frac{O_3 + O_4}{1-\theta} \end{aligned}$$

  Izenačimo z 0 in dobimo

$$\begin{aligned} 0 &= (O_1 + O_2)(1-\widehat{\theta}) - (O_3 + O_4)(\widehat{\theta}) \\ 0 &= (O_1 + O_2) - \widehat{\theta}(O_1 + O_2 + O_3 + O_4) \\ \widehat{\theta} &= \frac{O_1 + O_2}{O_1 + O_2 + O_3 + O_4} = \frac{n_{1.}}{n} \end{aligned}$$

- Celoten prostor:

$$\begin{aligned} L(\theta) &= \theta_1^{O_1}\theta_2^{O_2}(1-\theta_1)^{O_3}(1-\theta_2)^{O_4} \\ l(\theta) &= O_1\log\theta_1 + O_2\log\theta_2 + O_3\log(1-\theta_1) + O_4\log(1-\theta_2) \\ \frac{\partial l(\theta)}{\partial\theta_1} &= \frac{O_1}{\theta_1} - \frac{O_3}{1-\theta_1} \end{aligned}$$

  Izenačimo z 0 in dobimo

$$\begin{aligned} 0 &= O_1(1-\widehat{\theta_1}) - O_3(\widehat{\theta_1}) \\ 0 &= O_1 - \widehat{\theta_1}(O_1 + O_3) \\ \widehat{\theta_1} &= \frac{O_1}{O_1 + O_3} = \frac{O_1}{n_{.1}} \end{aligned}$$

$$\Lambda = \frac{\prod_{i=1}^{k} \pi_i^1(\widehat{\theta}_1, \widehat{\theta}_2)^{O_i}}{\prod_{i=1}^{k} \pi_i^0(\widehat{\theta})^{O_i}} = \prod_{i=1}^{k} \left(\frac{\pi_{1i}(\widehat{\theta}_1, \widehat{\theta}_2)}{\pi_{0i}(\widehat{\theta})} \right)^{O_i}$$

Zapišimo gornjo testno statistiko še malce drugače. Označimo z E_i pričakovano frekvenco v vsaki celici, torej

$$E_i = \frac{n_{1.} n_{.1}}{n}$$

Dobimo:

$$\begin{aligned} 2 \log \Lambda &= 2 \sum_{i=1}^{k} O_i \log \frac{\frac{O_i}{n_{.i}}}{\frac{n_{i.}}{n}} \\ &= 2 \sum_{i=1}^{k} O_i \log \frac{O_i}{E_i} \end{aligned}$$

(Tu je $n_{3.} = n_{1.}$, torej robna vrednost po tistem stolpcu, ki vsebuje to celico)

Asimptotska porazdelitev te testne statistike je χ^2, v našem primeru z eno stopinjo prostosti (zgoraj ocenjujemo dva parametra, spodaj enega).

Po drugi strani je Pearsonova χ^2 testna statistika enaka:

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Tudi v tem primeru poznamo le asimptotsko porazdelitev, ki pa jo je nekoliko težje dokazati. Pokažimo raje, da sta testni statistiki asimptotsko ekvivalentni:

Funkcijo $f(x) = x \log \frac{x}{E}$ razvijemo v Taylorjevo vrsto okrog E (pod ničelno domnevo bo O_i zelo blizu E_i). Velja

$$\begin{aligned} f(E) &= E \log \frac{E}{E} = 0 \\ f'(x) &= \log \frac{x}{E} + x \frac{E}{x} \frac{1}{E} = 1 + \log \frac{x}{E} \;\rightarrow\; f'(E) = 1 \\ f''(x) &= \frac{E}{x/E} = \frac{1}{x} \;\rightarrow\; f''(E) = \frac{1}{E} \end{aligned}$$

$$f(O_i) = 0 + (O_i - E_i)E_i + \frac{1}{2}(O_i - E_i)^2 \frac{1}{E_i} + \dots$$

Dobimo torej

$$2\sum_{i=1}^{k} O_i \log \frac{O_i}{E_i} = 2\sum_{i=1}^{k} \left[(O_i - E_i) + \frac{1}{2}(O_i - E_i)^2 \frac{1}{E_i} + \dots \right]$$

Prvi izraz na desni je enak 0, saj je vsota verjetnosti po vseh celicah v obeh primerih enaka n, ostane nam torej drugi le drugi izraz

$$2 \log \Lambda \approx= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Testni statistiki sta torej asimptotsko ekvivalentni. Vrednost χ^2 nam dela težave, kadar je v kateri celici pričakovano število enako 0, medtem ko nam testna statistika gLRT dela težave tudi, kadar je opazovano število 0. To je tudi eden izmed razlogov, zakaj se večinoma uporablja prva.

Opomba: Izpeljali smo testno statistiko za splošen primer multinomne porazdelitve. Lahko jo uporabljamo tako za kontingenčne tabele kot tudi za tabele neke porazdelitve (ki je definirana z enim ali večimi parametri θ).

Primer:

Na strani statističnega urada lahko dobimo podatke o rojstnih dnevih prebivalcev Slovenije. Takoj lahko vidimo, da so nekateri dnevi bolj pogosti od drugih - najpogostejši je prvi januar, najmanj pogost pa 31.12 (29. februar bodisi izločimo, bodisi pomnožimo s 4). Zanima nas, ali je to lahko le posledica naključja. Postavimo ničelno domnevo:
H_0 : vsi rojstni dnevi so enako verjetni
Kako bi preverili to domnevo? Uporabimo test χ^2. Opazovane vrednosti so kar števila rojstev, pričakovane so skupno število/366. Test ima 365 stopinj prostosti.
Dobimo $p < 0{,}0001$, odstopanja so zelo močno statistično značilna. Izkaže se, da razlogi za ta odstopanja niso preveč poetični - posamezniki, pri katerih datum rojstva ni znan, imajo datum postavljen na 1.1., podobno je s prvimi dnevi v posameznih mesecih. .

Primer: Ravnotežje Hardy-Weinberg

Vrnimo se k primeru Hardy-Weinbergovega ravnotežja. Medtem, ko nas je do sedaj zanimala predvsem ocena θ, se tokrat vprašajmo ali je naš model (torej predpostavka, da so geni v Hardy-Weinbergovem ravnotežju) zares smiseln.

Zanima nas torej ničelna domneva $H_0 : \pi = (\theta^2, 2\theta(1-\theta), (1-\theta)^2)$, alternativna domneva je $H_A : \pi = (\theta_1, \theta_2, 1-\theta_1-\theta_2)$

Izpeljali smo cenilko

$$T = 2\log\Lambda = 2\sum_{i=1}^{k} O_i \log\frac{O_i}{E_i}$$

Vemo, da je pod ničelno domnevo porazdeljena kot χ^2_1 (razlika v številu parametrov, ki jih ocenjujem). V našem primeru so O_i vrednosti v posameznih celicah, E_i pa so pričakovane frekvence pod ničelno domnevo. Spomnimo se, da je cenilka največjega verjetja za θ enaka

$$\widehat{\theta} = \frac{2n_1 + n_2}{2n}$$

Vrednost E_1 je torej $E_1 = \left(\frac{2n_1+n_2}{2n}\right)^2 n$. Na naših podatkih dobimo pričakovane vrednosti ($\widehat{\theta} = 0{,}297$):

Tabela 3.2

	AA	Aa	aa
O_i	83	428	489
E_i	88,2	494,2	417,6

Vrednost testne statistike je enaka

$$T = 2\log\Lambda = 2\sum_{i=1}^{k} O_i \log\frac{O_i}{E_i} = 0{,}626$$

Vrednost $p = 0{,}438$.

Podobno je vrednost testne statistike hi-kvadrat enaka

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 0{,}622,\ p = 0{,}430$$

.

3.7 Problem večkratnega preizkušanja domnev

Fishing for hypotheses is like throwing a dart at a wall and then drawing a target around it. (A. Monette)

He uses statistics as a drunken man uses lamp-posts–for support rather than illumination. (A. Lang)

Primer:

Primer: Recimo, da nas zanima, katero nebesno znamenje ima največ sreče v ljubezni. Denimo, da vemo (da bo naš izračun lažji), da je v populaciji starosti 18-35 let s svojo srečo v ljubezni zadovoljnih 40 % posameznikov. Zberemo naključen vzorec po 30 posameznikov iz vsakega nebesnega znamenja ter si ogledamo rezultate. Vidimo, da je največ srečnih (17 od 30) med biki, zato postavimo ničelno domnevo
H_0: Največ 40% posameznikov rojenih v znamenju bika ima srečo v ljubezni in jo skušamo zavrniti. Testna statistika naj bo število srečnih v ljubezni, označimo jo z X. Če je X velik, bomo ničelno domnevo zavrnili, vemo že, da je X binomsko porazdeljena slučajna spremenljivka. Izračunajmo vrednost p

$$p = P_0(X \geq 17) = 0{,}048$$

Vrednost $p < \alpha$, zato veselo sklenemo, da so biki srečnejši v ljubezni. Rezultate objavimo v ugledni astrološki reviji. Meja zavrnitve je bila 17
Poglejmo, kaj je narobe z zgornjim primerom. Domnevo smo postavili po ogledu podatkov - če bi bi rezultat 17 dobili pri drugem nebesnem znamenju, bi ničelno domnevo postavili glede na tisto znamenje.
V našem izračunu torej nismo računali verjetnosti, da je vrednost testne statistike pri bikih večja oz. enaka 17, temveč verjetnost, da je največja od 12 vrednosti večja ali enaka 17.

$$P_0(\max_{1\leq i\leq 12}(X) \geq 17) = 1 - P_0(X_1 \leq 16, \ldots, X_{12} \leq 16) = 1 - P(X_i \leq 16)^{12} = 0{,}447$$

Naša napaka prve stopnje je torej enaka 0,45 namesto 0,05. Na tak način se seveda da dokazati marsikaj.

Poglejmo še, kako bi našo ničelno domnevo (H_0 : s srečo v ljubezni je ne glede na znamenje zadovoljnih 40% posameznikov) preizkusili pravilno? Uporabimo lahko test χ^2. Opazovane frekvence so številke v kontingenčni

tabeli:

Tabela 3.3

	oven	bik	dvojčka	rak	lev	devica	tehtn.	škorp.	strel.	koz.	vodn.	ribi
d	13	17	11	9	13	10	12	13	11	10	11	8
n	17	13	19	21	17	20	18	17	19	20	19	22

Pričakovane vrednosti so velikost vzorca krat 0,4 oz. 0,6. Kaj pa stopinje prostosti? Pod ničelno domnevo ne ocenimo nobenega parametra, na celem prostoru jih je 12. Torej 12 stopinj prostosti. Za naš primer dobimo $T = 8{,}9$, $p = 0{,}71$.

.

Verjetno največji vir napačnih rezultatov je problem večkratnega preizkušanja domnev. Pogosto raziskovalci preizkusijo več domnev, na koncu pa interpretirajo tisto, ki se izkaže za statistično značilno. Poglejmo si, kaj se dogaja z napako I. vrste v takem primeru: denimo, da na podlagi podatkov preizkusimo dve domnevi, ki sta med seboj nepovezani (recimo, da sta testni statistiki neodvisni slučajni spremenljivki), torej da na primer posebej preverjamo ali IQ večji od 100 pri ženskah in pri moških. Kakšna je verjetnost, da uspemo zavrniti vsaj eno, čeprav sta obe ničelni domnevi res?

$$\begin{aligned} P(T_1 \geq c_1 \text{aliT}_2 \geq \text{c}_2) &= 1 - P(T_1 < c_1, T_2 < c_2) \\ &= 1 - P(T_1 < c_1)P(T_2 < c_2) = 1 - 0{,}95^2 = 0{,}0975 \end{aligned}$$

Če bi imeli še več domnev, bi bila napaka prve stopnje še večja.
V angleščini za ta pojav uporabljajo izraz 'Inflated type I error'.

Če zelimo razumeti, kaj se dogaja z verjetnostjo napake I. vrste, moramo povsem definirati strategijo in jo preizkusiti s simulacijami.

Ključen je vrstni red: najprej domneve, potem testi. Eksploratorna raziskava vs. preizkušanje domnev : ne moremo hkrati iskati domneve in jo potrjevati.

Problem pristranskosti poročanja (publication bias):
Omenimo še pristranskost poročanja: nad majhnimi vrednostmi p niso 'navdušeni' le raziskovalci, temveč tudi uredniki znanstvenih revij. To pa pomeni, da ima članek več možnosti, da bo objavljen, če poroča o statistično značilnih rezultatih. Vendar pa to vodi k velikim težavam pri meta-analizi znanstvenih člankov, torej analizi, ki jo naredimo tako, da zberemo literaturo, ki vsebuje raziskave na neko temo. Več raziskovalcev je preverjalo neko ničelno domnevo, a objavljeni so le članki tistih raziskovalcev, ki so dobili statistično značilen rezultat. Jasno je, da bomo zaradi tega dobili vsaj nekoliko izkrivljen pogled na dejstva.

Pomembnost velikosti vzorca pri zaupanju v raziskavo:
Prijatelj pride z novico iz dnevnega časopisja:
'Raziskovalci so proučevali povezanost vere in depresije (so verni ljudje manj ali bolj depresivni).' Raziskava je pokazala, da vernimi in ateisti ni razlik v pogostosti depresije. Sedaj pa je bila opravljena nova, večja raziskava, v 20 državah sveta, ki kaže nasprotne rezultate - med vernimi je depresije manj.