LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

HIPOTĒŽU PĀRBAUDE AR NEPARAMETRISKAJIEM TESTIEM

DIPLOMDARBS

Autors: **Juris Rieksts-Riekstiņš**

Stud. apl. jr06016

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2011

Anotācija

Lai pārbaudītu statistiskas hipotēzes, ir izstrādāti daudzi statistiskie testi. Piemēram, hipotēžu pārbaudei par izlašu vidējo vērtību vienādību izplatītākie ir t-tests un vienfaktora dispersiju analīze (ANOVA), taču, tā kā tie ir parametriski testi, tad ir jāizpildās konkrētiem pieņēmumiem, lai šos testus varētu pielietot, turklāt diskrētā veidā uzdotiem datiem šie testi nemaz neder. Šajā darbā aplūkoti vairāki neparametriski testi, kas ir labas alternatīvas parametriskajiem testiem, kā arī paskaidrota atšķirība starp parametriskajiem un neparametriskajiem testiem. Katrā nodaļā aplūkotas atšķirīgas situācijas pēc izlašu skaita un to savstarpējās atkarības, savukārt apakšnodaļās aplūkoti piemērotākie testi atbilstošajam datu formātam. Darba noslēgumā iekļauti piemēri dažiem no aplūkotajiem testiem.

Atslēgas vārdi: hipotēžu pārbaude, rangi, diskrēti dati, testu jauda

Abstract

There have been developed several statistical tests for testing statistical hypothesis. For example, for testing hypothesis about the equality of sample means, the most common tests are the t-test and one-way analysis of variance (ANOVA), but, as they are parametric tests, there are certain assumptions about the data that have to be met before the test can be properly used, plus for discrete data, the parametric tests are invalid. In this paper several nonparametric tests are discussed, and the difference between parametric and nonparametric tests is reviewed. Each chapter is devoted for different situation based on the number of samples and their dependence, whereas each sub-chapter describes the most appropriate test for the type of data. At the end a few examples are included for some of the tests.

Keywords: hypothesis testing, ranks, discrete data, power of the tests

Saturs

Apzīmējumi

N izlases apjoms,

$N(\mu, \sigma^2)$ normāli sadalīts gadījuma lielums ar vidējo vērtību μ un dispersiju σ^2,

χ_k^2 χ^2 sadalīts gadījuma lielums ar k brīvības pakāpēm,

$Dexp(\mu, b)$ dubulteksponenciālais sadalījums ar parametriem μ un b,

$ln(a, b)$ lognormālais sadalījums ar parametriem a un b,

$binomial(n, p)$ binomiālais sadalījums ar apjomu n un varbūtību p,

$K(x_0, \gamma)$ Koši sadalījums ar parametriem x_0 un γ,

F izlases teorētiskā sadalījuma funkcija,

F_n izlases empīriskā sadalījuma funkcija,

P varbūtība,

$\stackrel{d}{\to}$ konverģence pēc sadalījuma,

$\stackrel{as}{\to}$ gandrīz droša konverģence,

$\underset{H_0}{\to}$ konverģence,kad H_0 spēkā,

I_α indikatorfunkcija, $I = 1$, kad ir spēkā α, un $I = 0$ pretējā gadījumā,

Ievads

Viena no būtiskākajām un plašāk pazīstamajām statistikas pamatproblēmām ir hipotēžu pārbaude. Praksē ļoti bieži ir nepieciešams savstarpēji salīdzināt divu vai vairāku izlašu rādītājus, lai noteiktu, vai starp šiem rādītājiem pastāv būtiskas atšķirības. Lai šādas problēmas varētu risināt, laika gaitā ir izstrādāti vairāki statistiski testi, kuru uzdevums ir pārbaudīt izvirzītās nulles hipotēzes. Plašāk pazīstamākās metodes ir t-tests divu izlašu vidējo vērtību salīdzināšanai un tā paplašinājums triju vai vairāk izlašu salīdzināšanai - vienfaktora dispersiju analīze (ANOVA), kas ir parametriskas datu analīzes metodes.

Parametriskie statistiskie testi ir ļoti efektīvs salīdzināšanas veids, taču tos ne vienmēr var pielietot. No teorijas viedokļa ir konkrēti pieņēmumi, kam jāizpildās, lai varētu korekti pielietot parametriskās metodes. Ir izpētīts, ka daudzos gadījumos novirzes no pieņēmumiem praktiski neietekmē rezultātus (skatīt [1]), taču ir arī situācijas, kad pastāv iespēja iegūt ļoti kļūdainus rezultātus. Dažiem datu sadalījuma veidiem parametriskās metodes vienkārši nestrādā, citiem tās uzrāda samērā vājus rezultātus, bet vēl citiem parametriskie testi darbojas labāk par neparametriskajiem par spīti to neatbilstībai. Šie teorētiskie uzskati ir pārbaudīti ar datu simulācijām, kuru apraksts ir atrodams pie darba praktiskās daļas.

Šī diplomdarba mērķis ir aplūkot dažus no pazīstamākajiem neparametriskajiem testiem tādā veidā, kas ļautu darba lasītājam ātri saprast, kāds būtu piemērotākais tests konkrētā situācijā, kā arī palīdzētu izvērtēt testa lietderību un efektivitāti. Lai materiāls būtu uzskatāmāks, ir pievienoti piemēri ar reāliem datiem, kā arī veiktas datu simulācijas testu jaudas un kļūdu pētīšanai. Apskatītie testi darbā sakārtoti pēc struktūras, kas attēlota 1. tabulā.

1. tabula Testu iedalījums

Datu veids	**Viena izlase**	**Divas atkarīgas izlases**	**Divas neatkarīgas izlases**	**Trīs atkarīgas izlases**	**Trīs neatkarīgas izlases**
Nomināli	Binomiālais tests, hī - kvadrāta tests	Maknemāra izmaiņu tests	Fišera tests, hī - kvadrāta tests	Kohrana Q tests	Hī - kvadrāta tests
Ordināli vai intervāla	Kolmogorova - Smirnova tests	Zīmju tests, Vilkoksona rangu zīmju tests	Mana - Vitneja U tests, Kolmogorova - Smirnova tests	Frīdmana divvirzienu rangu dispersijas analīze	Kruskala - Vallisa rangu dispersijas analīze, Džonkhīra terpsta tests

Darbs sastāv no divām ievadnodaļām, kurās sniegts ieskats hipotēžu pārbaudes un statistisko testu būtībā, paskaidrotas atšķirības starp parametriskajiem un neparametriskajiem testiem, kā arī sniegts detalizēts ieskats testu efektivitātes novērtēšanā. Tālāk ir piecas nodaļas ar vairākām apakšnodaļām, kas iedalītas pēc aplūkojamās problemātikas būtības un uzdoto datu veidiem. Katrai situācijai ir aprakstīts piemērotākais neparametriskais tests, kā arī pieminēta tā parametriskā alternatīva, ja tāda eksistē. Praktiskajā daļā ir aplūkoti piemēri dažiem populārākajiem testiem un rezultāti salīdzināti ar parametriskajiem testiem. Veiktas datu simulācijas ar programmu **R**, lai novērtētu testu jaudas un kļūdas dažādiem datu sadalījumiem. Pielikumā atrodams simulācijām izmantoto programmu kods.

1. Hipotēžu pārbaude, statistiskie testi un to kļūdas

Lai izdarītu secinājumus par kādu datu kopu, ir nepieciešams izvirzīt konkrētu un pamatotu hipotēzi. Atkarībā no tā, vai izvirzīto hipotēzi var vai nevar noraidīt, tiek pieņemti lēmumi par šo datu kopu. Taču, lai nonāktu līdz hipotēzes pārbaudei, ir jāsaprot, kāda metode būtu vispareizākā lēmuma pieņemšanai. Pastāv neskaitāmi statistiskie testi, ar kuru palīdzību tiek pārbaudītas hipotēzes, taču tie ir ļoti atšķirīgi, un nav tāda viena universālā testa, kas derētu jebkuriem datiem un jebkurai nulles hipotēzei. Lai izvēlētais tests būtu pareizs, ir jāņem vērā vairāki faktori. Ir svarīgi gan tas, kādā formātā ir aplūkojamie dati (nomināla, ordināla vai intervāla skala), gan salīdzināmo izlašu daudzums, apjoms un veids, gan arī tas, vai izpildās attiecīgā testa pieņēmumi. Taču, lai cik arī atbilstošu metodi mēs neizvēlētos, hipotēžu pārbaudē nekas nav zināms pilnīgi droši, tāpēc ir jāizvēlas būtiskuma līmenis, pie kura strādāt. Piemēram, ja tiek izvēlēts būtiskuma līmenis $\alpha = 0.05$, tad ar 95% ticamību var apgalvot, ka iegūtie rezultāti ir pareizi.

1.1. Nulles hipotēze

Hipotēze ir zinātnisks pieņēmums, kas ir loģisks un ticams, bet prasa pārbaudi un pierādījumu. Statistiskā hipotēze ir pieņēmums vienas vai vairāku statistisko kopu īpašībām, kuru pārbauda ar datu palīdzību, un iegūtajiem rezultātiem ir varbūtības raksturs [2]. Pirmais solis lēmuma pieņemšanas procesā ir nulles hipotēzes (H_0) izvirzīšana. Nulles hipotēze parasti tiek veidota tā, lai to varētu noraidīt un būtu spēkā alternatīvā hipotēze (H_1). Piemēram:

$$H_0 : \mu_1 = \mu_2,$$
$$H_1 : \mu_1 \neq \mu_2.$$

Šādi noformulēta hipotēze ļautu noteikt, vai starp divu izlašu vidējām vērtībām pastāv būtiskas atšķirības. Nulles hipotēze saka, ka būtiskas atšķirības nepastāv, savukārt alternatīvā hipotēze saka tieši pretējo, tas ir, ka būtiskas atšķirības pastāv. Ja testa rezultātā iegūtā p-vērtība ir lielāka par noteikto būtiskuma līmeni, tad saka, ka nulles hipotēzi nevar noraidīt pie noteiktā būtiskuma līmeņa, taču, ja tā ir mazāka par noteikto būtiskuma līmeni, tad saka, ka pie noteiktā būtiskuma līmeņa nulles hipotēze ir jānoraida, līdz ar to tiek pieņemta alternatīvā hipotēze. Tas, ka alternatīvā hipotēze tiek pieņemta, neļauj apgalvot, ka tā ir patiesa. Tā vienkārši tiek pieņemta kā alternatīva hipotēzei, kuru noraida. Jāpiebilst, ka šādi veidotu hipotēzi sauc par divpusēju hipotēzi. Ja alternatīvā hipotēze ir uzdota formā $H_1 : \mu_1 < \mu_2$ vai $H_1 : \mu_1 > \mu_2$, tad to sauc par vienpusēju hipotēzi.

1.2. Statistiskie testi

Statistikā ir izstrādāti ļoti daudzi testi, no kuriem ir iespējams atrast derīgu testu hipotēžu pārbaudei praktiski jebkuram pastāvošam pētījumu dizainam. Lai varētu droši apgalvot, kurš tests ir piemērotākais konkrētā situācijā, ir rūpīgi jāizanalizē dati. Tā kā šī darba pamattēma ir neparametriskie statistiskie testi, tad 2. nodaļa ir veltīta parametrisko un neparametrisko testu aprakstam un savstarpējam salīdzinājumam. Izvēle starp šiem testiem balstās tieši uz datiem, tāpēc ir svarīgi precīzi izpētīt visu mūsu rīcībā esošu informāciju. Izvēloties nepareizo testu, var gadīties, ka iegūtie rezultāti ir korekti (varbūt pat labāki nekā izvēloties pareizo testu), taču to nevar droši apgalvot. Šādas situācijas vēlāk aplūkošu ar datu simulācijām.

1.3. Kļūdas un testa jauda

Kā jau iepriekš minēju, lēmumam par hipotēzes pieņemšanu piemīt varbūtības raksturs. Tas nozīmē, ka pastāv arī varbūtība pieņemt kļūdainu lēmumu. Pat ja mēs esam par 95% droši par iegūto rezultātu, pastāv iespēja kļūdīties, turklāt pastāv divas dažādas kļūdas, kas definētas avotā [2].

Definīcija 1. Par pirmā veida kļūdu sauc varbūtību $P(\text{noraidīt } H_0|H_0) = \alpha$.

Definīcija 2. Par otrā veida kļūdu sauc varbūtību $P(\text{nenoraidīt } H_0|H_1) = \beta$.

Veicot testu, iegūtajam rezultātam ir 4 iespējamie varianti, kas apkopoti 2. tabulā.

2. tabula Statistiskā testa iespējamie iznākumi

Pieņemtais lēmums	H_0 patiesa (H_1 nepatiesa)	H_0 nepatiesa (H_1 patiesa)
Noraidīt H_0 (pieņemt H_1)	pirmā veida kļūda	pareizs lēmums
Pieņemt H_0 (noraidīt H_1)	pareizs lēmums	otrā veida kļūda

Iepriekš minētais būtiskuma līmenis faktiski ir maksimāli pieļaujamā pirmā veida kļūdas eksistences varbūtība. Jo mazāku izvēlēsimies būtiskuma līmeni α, jo mazāka iespēja nepamatoti noraidīt patiesu H_0. Otrā veida kļūdas varbūtība β ir apgriezti proporcionāla α, tas ir, ja pie nemainīga izlases apjoma N tiks samazināta α, tad β palielināsies. Lai samazinātu abu kļūdu pieļaušanas varbūtību, ir jāpalielina izlases apjoms N. No otrā veida kļūdas ir atkarīga arī testa jauda (skatīt avotā [3]).

Definīcija 3. Par testa jaudu sauc varbūtību $P(\text{noraidīt } H_0|H_1 = 1 - \beta$.

3. tabulā apkopotas varbūtības, kas ir zīmīgas statistiskajiem testiem.

3. tabula Ar statistisko testu asociētās varbūtības

Pieņemtais lēmums	H_0 patiesa $\big(H_1$ nepatiesa$\big)$	H_0 nepatiesa $\big(H_1$ patiesa$\big)$
Noraidīt H_0 $\big($pieņemt $H_1\big)$	būtiskuma līmenis $= \alpha$	testa jauda $1 - \beta$
Pieņemt H_0 $\big($noraidīt $H_1\big)$	$P(\text{pieņemt } H_0\|H_0) = 1 - \alpha$	$P(\text{pieņemt } H_0\|H_1) = \beta$

Izmantojot jaudas analīzi, ir iespējams aprēķināt minimāli nepieciešamo izlases apjomu, pie kura varētu iegūt noteikta lieluma statistisko efektu, kā arī var noteikt minimāli iegūstamo statistisko efektu pie dota izlases apjoma. Kā minēts avotā [4], jaudas jēdziens ļoti bieži tiek lietots arī, lai savstarpēji salīdzinātu dažādas statistiskās procedūras, piemēram, parametriskos un neparametriskos testus.

1.4. Testu efektivitāte

Iepriekš tika pieminēta testu efektivitāte, taču netika paskaidrots, kas īsti tā ir, un kā to aprēķināt. Pieņemsim, ka ir doti divi testi ar statistikām $T_{1,n}$ un $T_{2,n}$, kas pārbauda nulles hipotēzi $H_0 : \theta \in \Theta_0$, un statistikas veidotas tā, ka to lielas vērtības liecina par H_0 noraidīšanu, ts ir, $T_n > c_n \Rightarrow H_0^-$. α, β apzīmē pirmā veida kļūdas pieļaušanas varbūtību un testa jaudu, un ar θ apzīmēsim kādu noteiktu alternatīvu. Apzīmēsim ar $n(\alpha, \beta, \theta, T)$ tādu mazāko izlases apjomu, kam spēkā

$$P_\theta(T_n \geq c_n) \geq \beta \tag{1.1}$$

un

$$P_{H_0}(T_n \geq c_n) \geq \alpha. \tag{1.2}$$

Abus testus $T_{1,n}$ un $T_{2,n}$ var salīdzināt ar attiecību

$$\frac{n(\alpha, \beta, \theta, T_1)}{n(\alpha, \beta, \theta, T_2)}, \tag{1.3}$$

un tests $T_{1,n}$ ir atzīstams par labāku, ja šī attiecība ≤ 1.

Diemžēl robežlieluma $n(\alpha, \beta, \theta, T)$ aprēķināšana ir ļoti sarežģīta un reizēm pat neiespējama ļoti vienkāršos gadījumos. Turklāt robežlielumu attiecība var būt atkarīga no dažādām α, β vai θ izvēlētajām vērtībām. Par laimi, ja $\alpha \to 0$ $\beta \to 1$, vai arī $\theta \to \theta_0$, tad attiecība parasti vai nu konverģē un lielumu, kas atkarīgs tikai n θ, vai arī ir vienkārši konstante.

Uz šādiem pieņēmumiem savus efektivitātes pētīšanas modeļus ir izstrādājuši vairāki autori, no kuriem vienkāršākā un populārākā ir Pitmana metode.

Definīcija 4. Doti $X_1, ..., X_n$ neatkarīgi un vienādi sadalīti (turpmāk tekstā tādus apzīmēsim ar *iid*) novērojumi no sadalījuma $P_\theta, \theta \in \Theta$. Pieņemsim, ka mēs gribam pārbaudīt $H_0 : \theta \in \Theta_0$ pret $H_1 : \theta \in \Theta - \Theta_0$. Apzīmēsim ar $T_n = T_n(X_1, ..., X_n)$ tādu statistiku secību, pie kurām jānoraida H_0, kad T_n ir liels. Precīzāk, nofiksējam $0 < \alpha < 1, 0 < \beta < 1, \theta \in \Theta - \Theta_0$. Definēsim $c_n = c_n(\theta, \beta)$ kā $P_\theta(T_n > c_n) \leq \beta \leq P_\theta(T_n \geq c_n)$. Testa apjoms definējams kā $\alpha_n(\theta, \beta) = \sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T_n \geq c_n)$ un $N_T(\alpha, \beta, \theta) = \inf\, [n : \alpha_m(\theta, \beta) \leq \alpha \ \forall\ m \geq n]$

Tādējādi $N_T(\alpha, \beta, \theta)$ ir mazākais izlases apjoms, pie kura tests, kas balstīts uz secību T_n, ir ar jaudu β pie noteiktas alternatīvas θ un apjoma $\leq \alpha$, taču to ir ļoti sarežģīti vai pat neiespējami aprēķināt dotiem α, β un θ. Lai aprēķinātu $N_T(\alpha, \beta, \theta)$, ir jāzina precīzs T_n sadalījums jebkuram fiksētam θ visiem dotiem n.

Definīcija 5. (Pitmana efektivitāte) Dotām statistikām T_{1n} un T_{2n} ar

$$e_{T_2, T_1}(\alpha, \beta, \theta) = \frac{N_{T_1}(\alpha, \beta, \theta)}{N_{T_2}(\alpha, \beta, \theta)}$$

Pitmana efektivitāte ir

$$e_P(\alpha, \beta, \theta_0) = \lim_{\theta \to \theta_0} e_{T_2, T_1}(\alpha, \beta, \theta), \tag{1.4}$$

kur $\theta_0 \in \delta\Theta_0$, pieņemot, ka šāda robeža eksistē.

Pitmana efektivitāte (turpmāk tekstā - efektivitāte) tipiski nav atkarīga ne no α, ne β, līdz ar to tās aprēķināšana parasti ir visvienkāršākā. Aprēķini būs vienkārši, ja tiks apmierināti visi nepieciešamie nosacījumi, taču ir būtiski, ka šo efektivitāti var aprēķināt arī kā asimptotisku, kad statistikas T_{1n} un T_{2n} ir uzdotas kā novērtējumi. Tālāk aplūkosim divas teorēmas no grāmatas [1], kas palīdz veikt efektivitātes aprēķinus.

Teorēma 1. *Ja dotas statistikas T_{1n} un T_{2n} un abas apmierina sekojošus nosacījumus:*
(1) dotas funkcijas $\mu_n(\theta), \sigma_n^2(\theta)$ un eksistē $\delta > 0$, ka

$$\sup_{|\theta - \theta_0| \leq \delta} \sup_z \left| P_\theta \left(\frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)} \leq z \right) - \Phi(z) \right| \underset{n \to \infty}{\to} 0,$$

(2) $\mu_n' > 0$,

(3) $\frac{\sqrt{n}\sigma_n(\theta_0)}{\mu_n'(\theta_0)} = O(1)$,

(4) ja $|\theta_n - \theta_0| = O(\frac{1}{\sqrt{n}})$*, tad* $\frac{\mu_n'(\theta_n)}{\mu_n'(\theta_0)} \to 1$,

(5) ja $|\theta_n - \theta_0| = O(\frac{1}{\sqrt{n}})$*, tad* $\frac{\sigma_n(\theta_n)}{\sigma_n(\theta_0)} \to 1$,

tad efektivitāti var aprēķināt pēc formulas

$$e_P(T_2, T_1) = \left(\frac{\lim_{n\to\infty} \frac{\sqrt{n}\sigma_{1n}(\theta_0)}{\mu_{1n}'(\theta_0)}}{\lim_{n\to\infty} \frac{\sqrt{n}\sigma_{2n}(\theta_0)}{\mu_{2n}'(\theta_0)}} \right)^2. \tag{1.5}$$

Teorēma 2. *Pieņemsim, ka* $\theta_0 \in \delta\Theta_0$, $-\infty < h < \infty$ *un* $\theta_n = \theta_0 + \frac{h}{\sqrt{n}}$*. Ja dotas statistikas* T_{1n} *un* T_{2n} *un abas apmierina sekojošus nosacījumus:*

(1) eksistē funkcijas $\mu(\theta), \sigma(\theta)$ *tādas, ka visiem* h

$$\frac{\sqrt{n}(T_n - \mu(\theta_n))}{\sigma(\theta_n)} \xrightarrow[P_{\theta_n}]{d} N(0,1), \tag{1.6}$$

(2) $\mu'(\theta_0) > 0$,

(3) $\sigma(\theta_0) > 0$ *un* $\sigma(\theta)$ *ir nepārtraukti* θ_0

tad efektivitāti var aprēķināt pēc formulas

$$e_P(T_2, T_1) = \frac{\sigma_1^2(\theta_0)}{\sigma_2^2(\theta_0)} \left[\frac{\mu_2'(\theta_0)}{\mu_1'(\theta_0)} \right]^2. \tag{1.7}$$

Redzams, ka testa statistiskais efekts jeb efektivitāte ir atkarīga no izlases apjoma, datu sadalījuma, kā arī pieņēmumiem, kam jāizpildās. Jo mazāk pieņēmumu par datiem izsaka tests, jo mazāka būs tā efektivitāte, kas ir spēkā pie jebkāda izlases apjoma. Turklāt, jo lielāks izlases apjoms, jo lielāka būs testa efektivitāte. Taču var gadīties situācija, kad testa A jauda ir lielāka par testa B jaudu pie izlases apjoma $N = 30$, savukārt testa B jauda ir lielāka pie izlases apjoma $N = 30$ nekā testa A jauda pie apjoma $N = 20$. Tas nozīmē, ka, palielinot izlases apjomu, ar vājāku testu ir iespējams iegūt tikpat efektīvus rezultātus kā ar jaudīgāku testu pie mazāka izlases apjoma. Tas sniedz iespēju vai nu vienkāršot aprēķinus, vai arī atvieglot datu iegūšanas procesu, jo ir nepieciešams mazāks datu apjoms. Izvēle par labu kādam testam ir atkarīga no konkrētās situācijas, un šajā darbā tālāk tiek pētīts, kad un kādi testi būtu vispiemērotākie.

Lai gūtu labāku priekšstatu par efektivitātes salīdzinājumiem starp dažādiem testiem, daudzos avotos (piemēram, [1]) ir iespējams atrast tabulas ar aprēķinātām efektivitātēm dažādiem izlašu apjomiem pie konkrētām α un β vērtībām.

2. Parametrisko un neparametrisko testu salīdzinājums

Lai labāk saprastu atšķirību starp parametrisko un neparametrisko testu būtību, ir jāsaprot veidi, kādos var tikt uzdoti dati. Visvājākais mērījumu līmenis ir tad, kad objekti, personas vai īpašības tiek vienkārši klasificēti, izmantojot skaitļus vai citus simbolus.

Definīcija 6. Ja skaitļi (vai citi simboli) tiek izmantoti, lai noteiktu, pie kuras grupas pieder konkrētais objekts, tad saka, ka šie skaitļi (simboli) veido nominālo jeb kategorisko datu skalu.

Definīcija 7. Ja skaitļi parāda attiecīgas objektu pozīcijas, bet neparāda atšķirības pakāpi starp tiem, tad saka, ka šie skaitļi veido ordinālo skalu.

Definīcija 8. Ja skaitļi parāda attiecīgas objektu pozīcijas un ļauj arī novērtēt atšķirības starp objektiem, tad saka, ka šie skaitļi veido intervāla skalu.

Ja dati ir uzdoti ar nominālo vai ordinālo skalu, tad ir jāizmanto neparametriskie testi, savukārt, ja tie ir uzdoti ar intervāla skalu, tad ir jāpārbauda, vai izpildās parametrisko testu pieņēmumi. Ja tie izpildās, tad vēlams izmantot parametriskos testus, bet ja nē, tad neparametriskos.

2.1. Parametriskie testi

Galvenā parametrisko testu pazīme ir tāda, ka tie balstās un konkrētiem pieņēmumiem par datiem un to sadalījumu. Precīzāk, par ģenerālkopas sadalījumu, no kuras ir ņemta izlase. Tā kā parasti mums nav iespējas precīzi aprēķināt ģenerālkopas parametrus un noteikt tās sadalījumu, tad šie lielumi tiek novērtēti ar datiem no pieejamās izlases. Būtiskākais pieņēmums, kam jāizpildās, ir tāds, ka dati (vai to atlikumi) ir normāli sadalīti. Pastāv uzskats, ka pie lieliem izlases apjomiem datiem vajadzētu aproksimēt normālo sadalījumu, taču tas ir neviennozīmīgs uzskats, turklāt ļoti bieži ir jāstrādā ar neliela apjoma izlasēm, tāpēc šo pieņēmumu ir nepieciešams pārbaudīt. Veicot praktiskos pētījumus, ir novērots, ka nelielas novirzes no normalitātes var neietekmēt rezultātu, taču paļaušanās uz šādu novērojumu paliek katra paša ziņā. Ja tiek salīdzinātas vairākas izlases, tad ir jāpārbauda pieņēmums par šo izlašu dispersijas homogenitāti. Tas nozīmē, ka šo izlašu dispersijām vajadzētu būt vienādām. Saprotams, ka, lai izpildītos šie pieņēmumi, datiem ir jābūt no intervāla skalas, kas krietni ierobežo parametrisko testu pielietojuma iespējas. Toties, ja šie pieņēmumi ir spēkā, tad parametriskie testi viennozīmīgi ir jaudīgākie un efektīvākie, līdz ar to vislabākie.

2.2. Neparametriskie testi

Neparametriskie testi ir pazīstami arī kā "testi bez pieņēmumiem" un "testi bez sadalījuma". Tas tā ir tādēļ, ka šiem testiem ir mazāk pieņēmumu, kam jāizpildās, turklāt tiem nav būtisks datu sadalījums. Galvenā šo testu darbības būtība ir datu pārvēršana rangos. Par datu rangu sauc skaitlisku vērtību no 1 līdz N, kas tiek piešķirta katram novērojumam no izlases apjomā N, sakārtojot visus novērojumus augošā secībā un piešķirot vērtību 1 mazākajam novērojumam, 2 nākamajam mazākajam, un tā tālāk.

Tā kā analīze tiek veikta nevis ar pašiem datiem, bet gan to rangiem, tad var teikt, ka netiek pētītas reālās atšķirības starp datiem, bet gan tikai to savstarpējā attiecība (lielāks, mazāks, vienāds). No šī fakta arī izriet galvenais neparametrisko testu pretarguments, kas saka, ka tiek pazaudēta ļoti liela daļa informācijas, līdz ar to testi ir izšķērdīgi un netiek balstīti uz precīzo situāciju. Kā pretargumentu var minēt faktu, ka dzīvē ļoti bieži nākas strādāt ar datiem, kas jau ir praktiski rangu formā, līdz ar to var gadīties, ka nemaz nav citas alternatīvas kā lietot neparametriskos testus. Turklāt ir zināms, ka testu jaudu ir iespējams palielināt, ja palielina izlases apjomu, tātad ir iespējams iegūt vienlīdz labus rezultātus vienkārši ņemot vairāk datu. Iepriekš jau ir minēts, ka, ja dati ir uzdoti ar nominālo vai ordinālo skalu, tad parametriskos testus nemaz nevar pielietot. Ļoti bieži gadās strādāt arī ar pavisam neliela apjoma izlasēm, kurām ir sarežģīti precīzi noteikt teorētisko sadalījumu, līdz ar to neparametriskie testi ir pareizā izvēle arī šajā situācijā. Kā vēl vienu plusu var minēt neparametrisko testu relatīvo vienkāršību, kas ļauj cilvēkiem ar vājākām zināšanām labāk izprast testu būtību. Mūsdienās gan reti kurš aprēķinus neveic ar datorprogrammu palīdzību, līdz ar to aprēķinu vienkāršība nav visai aktuāla nepieciešamība, taču izpratne par to, ko tad īsti un kā tests salīdzina, ir vairāk nekā vēlama, lai iegūtos rezultātus varētu racionāli attēlot un argumentēt.

2.3. Piemērotākās metodes izvēle

No iepriekšējos paragrāfos lasītā varētu padomāt, ka neparametriskie testi ir labāki gandrīz jebkurā situācijā, taču tā gluži nav. Parametriskie testi viennozīmīgi ir labāki, vienkārši praksē ļoti bieži nav korekti tos pielietot. Lai iegūtu pārliecību, ka parametriskais tests ir pielietots korekti, nepieciešams pārbaudīt divus tā galvenos pieņēmumus - normalitāti un dispersijas homogenitāti. Šo pieņēmumu pārbaudei ir izveidoti speciāli statistiskie testi, kuru veikšana ir iespējama visās populārākajās statistiskajās datorprogrammās.

Lai pārbaudītu datu atbilstību normālajam sadalījumam, jāpārbauda hipotēze

$$H_0 : X \sim N(\mu, \sigma^2)$$

pret alternatīvo hipotēzi

$$H_1 : X \nsim N(\mu, \sigma^2).$$

Visbiežāk izmantotie ir Kolmogorova - Smirnova tests ar Lilifora korekciju un Šapiro - Vilka tests. Abi šie testi salīdzina datus no izlases ar normāli sadalītu izlasi, kurai ir vienāda vidējā vērtība un dispersija, un pārbauda hipotēzi, ka izlase ņemta no normāli sadalītas populācijas. Kā minēts avotā [5], Šapiro - Vilka tests ir mazāk izplatīts, taču tas ir precīzāks un ar lielāku ticamību.

Definīcija 9. Doti $X_1, X_2, ...X_n$ *iid* gadījuma lielumi, kur X_i ir i-tais mazākais skaitlis izlasē, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Šapiro - Vilka testa statistiku var pierakstīt formā

$$W = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \tag{2.1}$$

kur a_i ir konstantes formā

$$(a_1, \ldots, a_n) = \frac{((m_1, \ldots, m_n)^T V^{-1})}{((m_1, \ldots, m_n)^T V^{-1} V^{-1} m)^{\frac{1}{2}}},$$

$m_1, ...m_n$ ir no normālā sadalījuma ņemtu *iid* gadījuma lielumu sagaidāmās vērtības sakārtotas augošā secībā, un V ir šo vērtību kovariāciju matrica. Testa statistikai ir spēkā īpašības

$$W \xrightarrow[H_0]{d} N(\mu, \sigma^2),$$
$$W \xrightarrow[H_1]{n \to \infty} \infty.$$

Abu testu lielākais mīnuss ir tāds, ka pie lieliem izlašu apjomiem ir ļoti liela iespējamība konstatēt minimālas novirzes no normalitātes, kas parādīsies kā būtiskas, tāpēc šādos gadījumos būtu vēlams normalitāti pārbaudīt arī grafiski, piemēram, zīmējot histogrammas vai kvantiļu-kvantiļu grafikus. Grafiskās metodes var sniegt uzskatāmu priekšstatu par datu sadalījumu, taču tā kā tas lielā mērā ir noderīgi liela apjoma izlasēm, bet neparametriskie testi ir plašāk pielietoti tieši mazām izlasēm, tad šīs metodes sīkāk netiks aplūkotas. Lai pārbaudītu izlašu dispersiju homogenitāti, jāpārbauda hipotēze

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2$$

pret alternatīvo hipotēzi

$$H_1 : \exists\ i, j \in (1, ..., n), i \neq j, \sigma_i^2 \neq \sigma_j^2.$$

Šim nolūkam visplašāk izmantotais ir Levena tests.

Definīcija 10. Aplūkosim k izlases ar apjomiem N_i un kopējo novērojumu skaitu N, i-tās grupas j-to novērojumu apzīmēsim ar $Y_i j$. Ja $Z_{ij} = \left|Y_{ij} - \bar{Y}_{i.}\right|$, $\bar{Y}_{i.}$ ir i-tās grupas vidējā vērtība, $Z_{..}$ ir visu Z_{ij} vidējais un $Z_{i.} = \frac{1}{N_i}\sum_{j=1}^{N_i} Z_{ij}$, tad Levena testa statistiku var aprēķināt pēc formulas

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - Z_{i.})^2}. \tag{2.2}$$

Statistikas būtiskuma pārbaudei izmanto īpašību (skatīt [6]), ka

$$W \xrightarrow[H_0]{d} F(\alpha, k-1, N-k),$$

$$W \underset{H_1}{\rightarrow} \infty.$$

3. Vienas izlases gadījums

Šajā nodaļā aplūkoti vairāki testi, ko var izmantot, lai pārbaudītu hipotēzes par kādu vienu izlasi. Ar šo testu palīdzību var noteikt, vai izlase ņemta no kādas konkrētas populācijas vai arī no konkrēta sadalījuma, respektīvi, ir iespējams pārbaudīt hipotēzes par atšķirību būtiskumu starp izlases un ģenerālkopas parametriem, par atšķirību būtiskumu starp novērotiem biežumiem un teorētiski sagaidāmiem biežumiem, kā arī hipotēzes par izlases datu atbilstību populācijai ar noteiktu sadalījumu vai formu.

3.1. Binomiālais tests

Doti $X_1, X_2, ..., X_n$ *iid* gadījuma lielumi kas pieņem vērtību 1 ar varbūtību p un vērtību 0 ar varbūtību q (Bernulli sadalījums). Tiek izvirzīta hipotēze

$$H_0 : p = p_0$$

un tās alternatīvā hipotēze ir

$$H_1 : p \neq p_0$$

divpusējā testa gadījumā, un

$$H_1 : p \geq p_0 \quad (H_1 : p \leq p_0)$$

vienpusējā testa gadījumā. Testa statistika aprēķināma (skatīt [7]) kā

$$Y = \sum_{i=1}^{n} X_i. \tag{3.1}$$

Varbūtība, ka tiks novērots konkrēts veiksmīgo realizāciju skaits k, ir aprēķināma pēc formulas

$$P[Y = k] = \frac{N!}{(k!(n-k)!)} p^k q^{n-k}. \tag{3.2}$$

Bieži vien ir nepieciešams aprēķināt varbūtību nevis precīzam veiksmīgo realizāciju skaitam, bet gan varbūtību, ka tie realizēsies vismaz noteiktu reižu skaitu. To var aprēķināt, summējot derīgo realizēšanās reižu varbūtības, līdz ar to vienādojums (3.2) var tikt pārrakstīts kā

$$P[Y \geq k] = P[Y = k] + P[Y = k+1] + ... + P[Y = n] = \sum_{i=1}^{n} \frac{n!}{i!(n-i)!} p^i q^{n-i}. \tag{3.3}$$

Jāpiebilst, ka aprēķini pēc šīm formulām kļūst arvien komplicētāki, kad n kļūst liels. Ir iespējams pierādīt, ka, palielinoties apjomam n, binomiālais sadalījums tiecas uz normālo sadalījumu, precīzāk,

$$Y \xrightarrow[H_0]{d} N(\mu, \sigma^2).$$

Šī tendence ir ļoti izteikta, kad p ir tuvs $\frac{1}{2}$, taču krietni lēnāka, kad p tuvojas 0 vai 1. Respektīvi, jo lielāka starpība starp p un q, jo lielākam jābūt N, lai aproksimācija būtu pietiekami tuva normālajam sadalījumam. Ir noteikts, ka pie $p \approx \frac{1}{2}$ pietiek, ja $n = 25$, savukārt uzmanīgiem jābūt, kad p vērtības ir tuvas 0 vai 1.

Teorēma 3. *Ja $p \to 0$ vai $p \to 1$, tad jāizpildās nevienādībai $npq > 9$, lai Y būtu aptuveni normāli sadalīts ar vidējo vērtību np un dispersiju npq, līdz ar to hipotēzes pārbaudei var izmantot statistiku*

$$z = \frac{Y - np}{\sqrt{npq}} \xrightarrow[H_0]{d} N(0,1). \tag{3.4}$$

$$z = \frac{Y - np}{\sqrt{npq}} \xrightarrow[H_1]{} \infty. \tag{3.5}$$

Tā kā binomiālais sadalījums atšķirībā no normālā nav nepārtraukts, ir vēlams pielietot nepārtrauktības korekciju. Vienādojums (3.4) jāpārveido formā

$$z = \frac{(Y \pm 0.5) - np}{\sqrt{npq}} \xrightarrow[H_0]{d} N(0,1), \tag{3.6}$$

kur $Y + 0.5$ tiek lietots, kad $Y < np$, bet $Y - 0.5$ lietots, kad $Y > np$, kas ļauj testa statistikas būtiskumu noteikt ar normālā sadalījuma palīdzību.

3.2. Hī-kvadrāta tests

Dota izlase ar *iid* diskrētiem gadījuma lielumiem $X_1, X_2, ..., X_N$, $X_i \sim F$. Nulles hipotēze izteikta formā

$$H_0 : F(x) = F_0(x) \ \forall x \in \mathbb{R}$$

un alternatīvā hipotēze ir

$$H_1 : \exists x \in \mathbb{R} : \ F(x) \neq F_0(x).$$

Definīcija 11. Testa statistiku aprēķina pēc formulas

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{k} \frac{n_i^2}{np_{i0}} - N, \tag{3.7}$$

kur $p_{i0} = P(X = X_{i0})$.

Teorēma 4. *Ja statistika aprēķināta pēc vienādojuma (3.7), tad tā konverģē uz hī - kvadrāta sadalījumu ar $k-1$ brīvības pakāpēm, kad H_0 ir spēkā.*

$$X^2 \xrightarrow[H_0]{d} \chi^2_{k-1}, \tag{3.8}$$

$$X^2 \underset{H1}{\rightarrow} \infty.$$

No vienādojuma (3.7) redzams, ka pie nelielām atšķirībām starp novērotajām un sagaidāmajām vērtībām arī testa statistika būs neliela, līdz ar to var secināt, ka, jo lielāka statistikas vērtība, jo mazāka iespējamība, ka novērojumi nāk no populācijas, kurai izpildās pie H_0 noteiktās sagaidāmās vērtības. Tā kā testa statistika ir asimptotiski sadalīta pēc hī - kvadrāta sadalījuma ar $k-1$ brīvības pakāpēm, tad tās būtiskumu var noteikt pēc šī sadalījuma tabulām.

Pielietojot hī - kvadrāta testu, ir jāpievērš uzmanība novērojumu apjomiem katrai no pieņemtajām diskrētajām vērtībām, jo testu nevar lietot, kad ir divas šādas vērtības un novērojumu skaits kādā no tām ir mazāks par 5, vai arī, kad iespējamo vērtību skaits ir lielāks par 2, bet 20% sagaidāmo biežumu ir mazāki par 5, vai arī kāds no sagaidāmajiem biežumiem mazāks par 1. Šādi apsvērumi jāņem vērā, jo testa statistika tikai aptuveni ir sadalīta pēc χ^2_{k-1}, nevis precīzi, līdz ar to sagaidāmo biežumu varbūtības var nebūt pietiekami tuvas, lai izdarītu secinājumus par novērotajiem biežumiem.

3.3. Kolmogorova - Smirnova tests

Nodaļā 2.3. ir pieminēts, ka ar Kolmogorova - Smirnova testa palīdzību var noteikt datu sadalījuma atbilstību normālajam sadalījumam, taču ar to ir iespējams novērtēt atbilstību arī starp datu empīrisko sadalījumu un kādu teorētisko sadalījumu, kam nav noteikti jābūt normālajam sadalījumam.

Definīcija 12. Doti $X_1, X_2, ..., X_N$ *iid* gadījuma lielumi. To empīriskā sadalījuma funkcija ir

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x},$$

kur $I_{X_i \le x}$ ir indikatora funkcija, kas pieņem vērtību 1, kad $X_i \le x$ un 0 pretējā gadījumā.

$$H_0 : F(x) = F_0(x) \ \forall x \in \mathbb{R}$$

$$H_1 : \exists x \in \mathbb{R} : \; F(x) \neq F_0(x)$$

Testa statistika apŗēkināma pēc formulas

$$D_n = \sup_x |F_n(x) - F(x)| \,. \tag{3.9}$$

Teorēma 5. *(Glivenko - Kantelli) Ir spēkā*

$$\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{as} 0.$$

Kā paskaidrots avotā [8], tiek noteikts punkts, pie kura starp teorētisko un empīrisko sadalījumu parādās maksimālās novirzes, un aplūkota varbūtība, ka šīs novirzes parādīšanās ir nejauša. Tā kā Kolmogorova - Smirnova tests pieņem, ka mainīgā sadalījums ir nepārtraukts, tad to var pielietot tikai datiem, kas uzdoti vismaz ar ordinālo skalu.

No testa apraksta var secināt, ka tas ir līdzīgs iepriekš aplūkotajam hī - kvadrāta testam. Lai arī atšķiras dati, kam var pielietot katru no testiem, ordinālās skalas datiem var izmantot abus testus. Ja tiek strādāts ar nelielām izlasēm, tad labāk būtu lietot Kolmogorova - Smirnova testu, jo tas ir precīzs, savukārt hī - kvadrāta tikai aptuveni precīzs. Ir grūti salīdzināt šo testu jaudas, jo katrs no tiem ir atkarīgs no citiem lielumiem, tāpēc izvēle par labu kādam no testiem, kad abi var tikt pielietoti, paliek katra paša ziņā.

4. Divu savstarpēji atkarīgu izlašu salīdzināšana

Izplatīta problēma statistikā ir kādas vienas izlases respondentu salīdzināšana divos dažādos apstākļos (piemēram, baktēriju daudzums organismā pirms un pēc antibiotiku lietošanas), vai arī divu izlašu respondentu salīdzināšana, kur katram respondentam no vienas izlases atbilst kāds cits respondents no otras izlases, par kuru var apgalvot, ka tie ir pakļauti vienādiem apstākļiem (piemēram, tiek salīdzināts tauku daudzums organismā divām cilvēku grupām, kur katrā grupā ir pa vienam cilvēkam no noteiktas svara kategorijas). Šādu un līdzīgu situāciju analīzei visizplatītākā metode ir t-tests, taču ne vienmēr ir iespējams iegūt datus, kam izpildās visi t-testa pieņēmumi, kā arī var gadīties, ka iegūtie dati ir tādā formā, ka t-tests to analīzei nemaz neder. Šajā nodaļā aplūkoti daži testi, kas ir labas alternatīvas parametriskajai metodei.

4.1. Maknemāra izmaiņu tests

Kā var noprast no testa nosaukuma, Maknemāra izmaiņu tests ir paredzēts izmaiņu būtiskuma noteikšanai kādai konkrētai izlasei pie diviem dažādiem nosacījumiem (piemēram, vai kādai cilvēku grupai dienas vidū ir jūtams nogurums, un vai tas samazinās pēc enerģijas dzēriena lietošanas). Tests ir piemērots darbam ar datiem, kas uzdoti nominālā vai ordinālā skalā, līdz ar to šim testam nav parametriskas alternatīvas.

Iegūtos datus vislabāk var attēlot, kā tas parādīts 4. tabulā. A, B, C, D ir respondentu skaits, kuru atbildes bija tādas, kā atbilstošajā rindā un kolonnā rakstīts, piemēram, A ir to respondentu skaits, kas pirmajā mērījumā sniedza pozitīvu atbildi, bet otrajā - negatīvu. Līdz ar to kopējais respondentu skaits, kuru viedoklis mainījās starp mērījumiem, ir $A + D$. Nulles hipotēze izsaka pieņēmumu, ka sagaidāmais apjoms abās šūnās, kur ir respondenti ar izmaiņām atbildēs, būs vienāds, precīzāk $\frac{A+D}{2}$.

$$H_0 : A = D,$$

$$H_1 : A \neq D.$$

No vienādojuma (3.7) zināms, ka

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}.$$

Veicot Maknemāra testu, mūs interesē tikai gadījumi, kad konstatētas izmaiņas atbildēs. Tā kā A un D ir izmaiņu skaita gadījumi attiecīgajā virzienā, un $\frac{A+D}{2}$ ir izmaiņu skaita

4. tabula Izmaiņu būtiskuma pētīšanas palīgtabula

	Pēc: negatīvs	Pēc: pozitīvs
Pirms: pozitīvs	A	B
Pirms: negatīvs	C	D

sagaidāmā vērtība abos virzienos, tad vienādojumu (3.7) varam pārrakstīt formā

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \frac{(A - \frac{A+D}{2})^2}{\frac{A+D}{2}} + \frac{(D - \frac{A+D}{2})^2}{\frac{A+D}{2}} = \frac{(A-D)^2}{A+D} \tag{4.1}$$

Teorēma 6. *Ieviešot nepārtrauktības korekciju, vienādojumā (4.1) iegūtai statistikai ir spēkā*

$$X^2 = \frac{(|A-D|-1)^2}{(A+D)} \xrightarrow[H_0]{d} \chi_1^2. \tag{4.2}$$

Iepriekš jau minēts, ka statistiku X^2 var labi aproksimēt ar hī - kvadrāta sadalījumu tad, ja izlašu apjomi ir lieli, taču šai metodei ir alternatīva. Ja sagaidāmā vērtība $\frac{A+D}{2}$ ir mazāka par 5, tad ieviešam $N = A + D$ un ar x apzīmējam mazāko no lielumiem A un D, un analīzi veicam ar nodaļā 3.1. aplūkoto binomiālo testu.

Tā kā ar nomināliem datiem Maknemāra testam nav alternatīvas, tad šajā gadījumā tā efektivitāte nav būtiska, taču, ja ir iespējams pielietot parametrisko t-testu, tad Maknemāra testa efektivitāte ir apmēram 95%, kad $A + D = 6$, un šī efektivitāte asimptotiski samazinās līdz apmēram 63%, kad $A + D$ palielinās (skatīt [10]).

4.2. Zīmju tests

Ar zīmju testa palīdzību tiek pētītas situācijas, kad ir divas savstarpēji saistītas izlases un ir nepieciešams noteikt, vai to ietekmējošie apstākļi atšķiras. Vienīgais šī testa pieņēmums saka, ka mainīgais lielums ir ar nepārtrauktu sadalījumu, taču nav būtiska ne sadalījuma forma, ne arī tas, vai visi objekti nāk no vienas populācijas. Ļoti ērti šo testu pielietot ir situācijās, kad katrā novērojumu pārī ir iespējams noteikt vienīgi to, kurš no tiem ir lielākais. Abas izlases tiek salīdzinātas pa pāriem, un katram pārim tiek piešķirta zīme "+" vai "-" atkarībā no tā, vai pirmās izlases novērojums ir lielāks, vai mazāks par otrās izlases atbilstošo novērojumu.

Dotas divas izlases ar neatkarīgiem novērojumiem $X_1, ..., X_n$, $Y_1, ..., Y_n$.

Definīcija 13. Ja X_i ir i-tais novērojums no vienas izlases un Y_i ir tam atbilstošais novērojums no otras izlases, tad zīmju testa nulles hipotēze pierakstāma formā

$$H_0 : P[X_i > Y_i] = P[X_i < Y_i] = \frac{1}{2}. \tag{4.3}$$

Ieviesīsim apzīmējumu

$$Z_i = X_i - Y_i, i = 1, \ldots, n,$$

kur n ir katras izlases apjoms. Pozitīviem Z_i tiek piešķirtas "+" zīmes, bet negatīviem "-". Jāņem vērā, ka ir iespējama situācija, kad novērojumi ir vienādi. Tādā gadījumā vienkārši neņem vērā tos novērojumu pārus, kam nav atšķirības, bet strādā tikai ar atšķirīgajiem, iegūstot izlases apjomā $m \leq n$. Apzīmēsim ar W pāru skaitu, kad $Z_i > 0$.

Teorēma 7. *W ir sadalīts binomiāli, kad H_0 ir spēkā.*

$$W \xrightarrow[H_0]{d} binomial(m, \frac{1}{2}) \tag{4.4}$$

No teorēmas 7 var secināt, ka būtiskuma noteikšanai var izmantot 3.1. nodaļā aplūkoto binomiālo testu. Izlasēm, kas apjomā lielākas par 25, var pielietot arī normālā sadalījuma aproksimāciju ar nepārtrauktības korekciju.

4.3. Vilkoksona rangu zīmju tests

Iepriekšējā nodaļā aplūkotais zīmju tests izmanto tikai informāciju par atšķirībām datu pāros, taču, ja ir pieejama arī informācija par šo atšķirību lielumiem, tad var pielietot jaudīgāku testu - Vilkoksona rangu zīmju testu. Šī testa būtiskākā iezīme ir tāda, ka tiek piešķirts lielāks svars pārim, starp kura vērtībām ir lielāka atšķirība. Parametriskā alternatīva ir pāru t-tests, taču tā kā tas ir parametriskais tests, pastāv iespēja, ka to nevar pielietot pieņēmumu neizpildīšanās dēļ vai datu veida dēļ.

Definīcija 14. Pieņemsim, ka mums ir doti n datu pāri

$$(X, Y) = (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n).$$

Izslēdzot pārus, kam ir spēkā $X_i - Y_i = 0$, iegūstam pāru skaitu $m \leq n$. Starpības novērojumiem apzīmēsim ar $d_i = X_i - Y_i$, un piešķirsim rangus R_i visām starpībām d_i

pēc absolūtās vērtības augošā secībā, sākot no mazākās starpības. Ja vairākas vērtības ir vienādas, tad piešķiram katrai no tām vidējo rangu vērtību. Pozitīvās d_i vērtības aprakstīsim ar

$$\varphi_i = I(d_i > 0),$$

kur I ir indikatora funkcija: $\varphi_i = 1$, kad $d_i > 0$ un $\varphi_i = 0$ pretējā gadījumā. Līdz ar to statistikas vērtību atrodam kā

$$S = \min(W_+, W_-), \tag{4.5}$$

kur

$$W_+ = \sum_{i=1}^{m} \varphi_i R_i$$

un W_- atrod pēc analoģijas.

Teorēma 8. *Ja izlašu apjomi $m > 15$, tad*

$$W^+ \xrightarrow[H_0]{d} N(\mu_{W_+}, \sigma^2_{W_+}), \tag{4.6}$$

$$W^+ \underset{H_1}{\rightarrow} \infty, \tag{4.7}$$

kur $\mu_{W_+} = \frac{m(m+1)}{4}$ un $\sigma^2_{W_+} = \frac{m(m+1)(2m+1)}{24}$.

Efektivitāte

Materiālā [11] ir parādīts, ka Vilkoksona (patiesībā arī zīmju un Kruskala - Vallisa) testu efektivitāte normālajam sadalījumam nekad nebūs mazāka par 0.864.

Definīcija 15. Pieņemsim, ka $\beta_N(\theta)$ un $\beta^*_N(\theta)$ ir testu A un A^* , kas veikti vienai un tai pašai izlasei apjomā N, jaudas funkcijas pret to parametriskajām alternatīvām apzīmētām ar θ, un θ_0 ir hipotēzē noteiktā θ vērtība. Visi testi veikti pie būtiskuma līmeņa α, un $\alpha < \beta < 1$, kur β ir noteiktā jauda. Aplūkosim tādu alternatīvu secību, ka

$$\beta_N(\theta_N) \underset{N\to\infty}{\rightarrow} \beta, \tag{4.8}$$

un $N^* = h(N)$ tādi, ka

$$\beta^*_{N^*}(\theta_N) \underset{N\to\infty}{\rightarrow} \beta, \tag{4.9}$$

tad, ja eksistē robeža

$$e_{A^*,A} = \lim_{N\to\infty} \frac{N}{N^*}, \tag{4.10}$$

kas ir neatkarīga no α, β un izvēlētajām secībām θ un $h(N)$, tad $e_{A^*,A}$ sauc par testa A^* *relatīvo asimptotisko efektivitāti* attiecībā pret testu A.

Nodaļā 1.4. ir parādīts, kā pēc Pitmana metodes aprēķināt šo robežu (4.10). Pieņemsim, ka ir dotas izlases $X = X_1, ..., X_m$ un $Y = Y_1, ..., Y_n$ no nepārtrauktiem sadalījumiem F un G, un ir izvirzīta hipotēze $H_0 : F = G$. Patiesībā mūs interesē, vai starp sadalījumiem pastāv tikai nebūtiskas novirzes, tas ir, $G(u) = F(u - \theta)$ visiem u. Ja F ir normālais sadalījums, tad atbilstošais parametriskais tests ir Stjūdenta t-tests. Pitmana asimptotiskā efektivitāte Vikoksona testam attiecībā pret t-testu aprēķināma kā

$$e_{w,t} = 12\sigma^2 \left[\int f^2(x)dx \right]^2, \tag{4.11}$$

kur f ir sadalījuma F varbūtību blīvums, un σ^2 ir izlašu X un Y kopējā dispersija. Ir zināms, ka $e_{w,t} \approx 0.95$, kad f ir normāls blīvums, $e_{w,t} = 1$ vienmērīgā sadalījuma gadījumā un $e_{w,t} = 1.26$, kad $f(x) = \frac{x^2}{4} \exp^{-|x|}$. Visas šīs vērtības ir ļoti augstas, kas mudina aprēķināt zemāko iespējamo $e_{w,t}$ vērtību. Lai to noskaidrotu, pierādīsim sekojošo teorēmu.

Teorēma 9. *Pieņemsim, ka N^* apmierina nosacījumus (4.9), kur testi A un A^* ir t-tests un Vilkoksona tests nepārtraukta sadalījuma F noviržu pārbaudei. Tādā gadījumā (a) jebkuram sadalījumam F ir spēkā*

$$\lim_{N\to\infty} \inf \frac{N}{N^*} \geq \frac{108}{125} = 0.864. \tag{4.12}$$

Turklāt (b) šī apakšējā robeža ir iegūta sadalījumam ar blīvumu (4.20,4.21), kuram $e =$ 0.864.

Pierādījums. Ir zināms, ka, ja F ir nepārtraukts un

$$\lim_{\theta\to\infty} \int \frac{1}{\theta} \left[F(x+\theta) - F(x) \right] dF(x) = c, \tag{4.13}$$

tad pēc vienādojuma (4.10) aprēķināmā efektivitāte eksistē, un tā ir $12c^2\sigma^2$. Šis pierādījums arī parāda, ka vispārīgi

$$\lim_{N\to\infty} \inf \frac{N}{N^*} \geq 12\sigma^2 \left[\lim_{\theta\to 0} \inf \frac{1}{\theta} \int [F(x)+\theta) - F(x)]dF(x) \right]^2. \tag{4.14}$$

Pēc Fatou lemmas, nevienādības (4.14) labā puse ir lielāka vai vienāda ar

$$12\sigma^2 \left\{ \int \left[\lim_{\theta\to 0} \inf \frac{F(x+\theta) - F(x)}{\theta} \right] dF(x) \right\}^2. \tag{4.15}$$

No Dekompozīcijas Teorēmas seko, ka, ja F satur sigulāru komponenti,

$$\lim_{\theta \to 0} \frac{F(x+\theta) - F(x)}{\theta} = \infty \tag{4.16}$$

pozitīvo F kopai, līdz ar to (4.15) un e ir bezgalīgi. Tādēļ varam pieņemt, ka eksistē blīvums $F(x) = F'(x)$ visai F vērtību kopai, izņemot 0, kas nozīmē, ka izteiksme (4.15)vienkāršojas uz

$$12\sigma^2 \left[\int f^2(x)dx \right]^2 . \tag{4.17}$$

Ja $\sigma^2 = \infty$, tad no vienādojuma (4.13) seko, ka $e = \infty$, līdz ar to varam pieņemt, ka σ ir galīga. Tā kā (4.17) ir nemainīgs pie izvietojuma vai mēroga izmaiņām, varam ņemt $\sigma^2 = 1$, līdz ar ko (4.17) minimizēšana reducējas uz

$$\int f^2(x)dx \tag{4.18}$$

minimizēšanu ar nosacījumiem

$$\int xf(x)dx = 0; \int f(x)dx = \int x^2 f(x)dx = 1; \; f(x) \geq 0 \; \forall x. \tag{4.19}$$

Saskaņā ar nenoteikto reizinātāju metodi, ir pietiekami minimizēt

$$\int [f^2(x) + 2b(x^2 - a^2)f(x)]dx.$$

Nenegatīviem f to var panākt nosakot

$$f(x) = b(a^2 - x^2), \tag{4.20}$$

kad$x^2 \leq a^2$, un $f(x) = 0$ pretējā gadījumā. Konstantes a un b tiek noteiktas no nosacījumiem (4.19), lai

$$s = \sqrt{5}, \; b = \frac{3}{20}\sqrt{5}, \tag{4.21}$$

un ar šīm vērtībām izteiksme (4.17) ir vienāda ar 0.864, kas tādējādi ir izteiksmes (4.14) apakšējā robeža. Tā kā blīvumam (4.20) vienādojuma(4.13) robeža var tikt panest zem integrāļa zīmes, redzams, ka šajā gadījumā efektivitāte eksistē un ir vienāda ar apakšējo robežu, kas tādējādi nevarr tikt uzlabota. □

Augstāk aplūkotajā gadījumā parādīta efektivitāte sadalījumu atšķirību pētīšanā, taču dabā bieži ir sastopamas situācijas, kad patiesībā parādās ļoti lielas atšķirības starp pašiem datiem izlašu ietvaros, tas ir, daži novērojumi “izlec” no pārējā sadalījuma (tādi novērojumi tiek saukti par *outliers*). Ja tiek pētītas atšķirības starp izlašu vidējiem rādītājiem, tad šo "izlecēju" ietekme var būt gana būtiska, lai neuzrādītu korektus rezultātus par populāciju sadalījumiem. Ir izpētīts, ka šādos gadījumos Vilkoksona un citu rangu testu efektivitāte krietni pieaug, salīdzinājumā ar parametriskajiem testiem. [11]

5. Divu savstarpēji neatkarīgu izlašu salīdzināšana

Ļoti izplatīta problēma statistikā ir divu izlašu salīdzināšana. Iepriekšējā nodaļā aplūkojām gadījumu, kad izlases veidotas no savstarpēji saistītiem novērojumu pāriem, taču ne vienmēr abas izlases ir saistītas. Ir izstrādāti vairāki testi gadījumiem, kad abas aplūkojamās izlases ir pilnīgi neatkarīgas, piemēram, lai salīdzinātu divas populācijas, no katras tiek izvēlēta izlase no nejaušiem novērojumiem, un no šiem novērojumiem tiek veikts novērtējums par populāciju atšķirību. Šādos gadījumos arī izlašu apjomi var nebūt vienādi.

Arī šai situācijai atbilstošā parametriskā metode ir t-tests, kas intervāla skalas datiem noteikti arī ir vislabākā, ja vien izpildās pieņēmumi, taču nomināliem un ordināliem datiem parametrisku metožu nav, tāpēc tādos gadījumos noteikti jāizvēlas kāds no tālāk aprakstītajiem neparametriskajiem testiem vai to alternatīvas.

5.1. Fišera tests

Fišera tests ir ļoti laba metode nelielu izlašu salīdzināšanai gadījumos, kad dati var pieņemt tikai divas dažādas vērtības.

Dotas izlases ar $X_1, ..., X_n$ un $Y_1, ..., Y_m$ *iid* gadījuma lielumiem. Apzīmēsim ar A un C vienas izlases novērojumu apjomus un ar B un D otras izlases novērojumu apjomus attiecīgi katrai no iespējamajām vērtībām. Tātad, novērojumu apjoms pirmajai vērtībai būs $A + B$, bet otrajai $C + D$, un kopējais novērojumu apjoms pa abām izlasēm $N = A + B + C + D = m + n$.

Teorēma 10. *Varbūtība, ka tiks novērots tieši šāds novērojumu sadalījums pēc vērtībām, ir uzdota ar hiperģeometrisko sadalījumu*

$$p = \frac{\frac{(A+C)!}{A!C!}\frac{(B+D)!}{B!D!}}{\frac{N!}{(A+B)!(C+D)!}} = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!} \tag{5.1}$$

Nulles hipotēze saka, ka abās izlasēs novērojumiem ir vienāda varbūtība pieņemt konkrētu vērtību. Alternatīvā hipotēze var būt gan vienpusēja, gan divpusēja.

$$H_0 : P(X_i = k) = P(Y_j = k),\ \forall i = 1, ..., n, j = 1, ..., m,$$

$$H_1 : \exists i = 1, ..., n, j = 1, ..., m : P(X_i = k) \neq P(Y_j = k)$$

divpusējai hipotēzei un

$$H_1 : \exists i = 1, ..., n, j = 1, ..., m : P(X_i = k) \geq P(Y_j = k)$$

vienpusējai.

Lai pārbaudītu divpusējo hipotēzi, mums atliek tikai salīdzināt iegūto varbūtību p ar interesējošo būtiskuma līmeni α.

Lai veiktu vienpusēju testu, mums ir jāaplūko visi iespējamie gadījumi, kad sadalījums ir tāds, kā noteikts alternatīvajā hipotēzē, vai arī vēl ekstrēmāks attiecīgajā virzienā. Attiecīgi iegūtās varbūtības ir jāsummē, un tad šī summārā varbūtība jāsalīdzina ar būtiskuma līmeni $\frac{\alpha}{2}$.

Pastāv uzskats, ka Fišera tests ir viens no labākajiem šāda veida problemātikas testiem, taču ir arī cilvēki, kas uzskata, ka var rasties neprecizitātes ar nulles hipotēzes noraidīšanas līmeni. Alternatīva šim testam ir Barnarda tests (*Barnard's test*), kas dažos avotos [12] tiek uzskatīts pat par jaudīgāku, taču Fišera tests ir izplatītāks.

5.2. Hī - kvadrāta tests

Līdzīgi kā Fišera tests, arī hī - kvadrāta tests salīdzina divu izlašu proporciju pa kategorijām atšķirību būtiskumu. Tests vislabāk piemērots diskrētiem datiem, jo tiek aplūkoti tikai biežumi, ar kādiem izlašu dati pieder konkrētajām kategorijām, bet netiek ņemta vērā precīza informācija par mērījumiem. Dati ir izteikti diskrētā veidā, un tiek analizēts, vai novērojuma vērtība liecina par tā piederību konkrētai izlasei, vai arī otrādi, ka novērojuma piederība konkrētai izlasei liecina par lielu iespējamību, ka tas pieņems konkrētu vērtību (sīkāku aprakstu skatīt [13]).

Definīcija 16. Pieņemsim, ka ir doti $X_1, ..., X_N$ *iid* novērojumi, kas iedalās c izlasēs un var pieņemt r dažādas vērtības. Apzīmēsim ar n_{ij} novērojumu skaitu ar i-to vērtību j-tajā izlasē un ar $N = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$ kopējo novērojumu skaitu. Līdz ar to testa statistika aprēķināma pēc formulas

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{E_{ij}} - N, \tag{5.2}$$

kur

$$E_{ij} = \frac{\sum_{j=1}^{c} n_{ij} \sum_{i=1}^{r} n_{ij}}{N} \tag{5.3}$$

ir sagaidāmais novērojumu apjoms ar i-to vērtību j-tajā izlasē.

Teorēma 11.

$$X^2 \xrightarrow[H_0]{as} \chi^2_{(r-1)(c-1)} \tag{5.4}$$

Balstoties uz teorēmu 11, var izdarīt secinājumus par nulles hipotēzi. Ir noskaidrots, ka hī - kvadrāta tests darbojas diezgan slikti pie nelieliem apjomiem ($n \leq 5$), jo tiek aproksimēts nepārtraukts sadalījums. Kad kopējais novērojumu skaits $N \leq 20$, labāk lietot Fišera testu, jo tas rēķina precīzas varbūtības, nevis novērtētas, savukārt pie lieliem apjomiem ($N > 40$) statistikai jāievieš nepārtrauktības korekcija (tikai tādā gadījumā, ja ir tikai divas iespējamās vērtības).

5.3. Mana - Vitneja U tests

Ja dati ir uzdoti vismaz ordinālās skalas veidā, tad tiem ir iespējams piešķirt rangus.

Definīcija 17. Ja dota skaitļu kop $X_1, X_2, ..., X_n$ apjomā n, tad i-tā novērojuma rangs ir

$$r_i = \#\{k : X_k \leq X_i\}. \tag{5.5}$$

Mana - Vitneja U tests veic tieši šo rangu analīzi. Doti $X_1, ..., X_N$ *iid* gadījuma lielumi, kas iedalās divās izlasēs $\boldsymbol{x_1}$ un $\boldsymbol{x_2}$. Tiek izvirzīta nulles hipotēze, ka šīm izlasēm ir vienādas mediānas

$$H_0 : \theta_1 = \theta_2$$

un alternatīvā hipotēze

$$H_1 : \theta_1 \neq \theta_2.$$

Tests ir ļoti labi pielietojams un ir lieliska alternatīva parametriskajam t-testam ar pieņēmumu par dispersiju nevienādību, taču ir nelielas variācijas paša testa ietvaros. Nelielām izlasēm iespējams veikt precīzus aprēķinus, savukārt lielām izlasēm jālieto normālā sadalījuma aproksimācija. Vēl ir jāpārbauda, vai nav novērojamas situācijas, kad diviem vai vairākiem novērojumiem ir vienādi rangi.

Definīcija 18. Ja dotas n_i izlases un

$$R_i = \sum_{j: X_j \in x_i} r_{ij}$$

ir šo izlašu rangu summas, tad testa statistikas aprēķina kā

$$U_i = R_i - \frac{n_i(n_i+1)}{2}, i = 1, 2. \tag{5.6}$$

Veicot nelielus algebriskus aprēķinus, var noskaidrot, ka

$$U_1 + U_2 = n_1 n_2.$$

Tā kā šādi aprēķini kļūtu arvien komplicētāki, kad $N \to \infty$, tad šādos gadījumos izmantosim normālā sadalījuma aproksimāciju.

Teorēma 12.

$$z = \frac{U \pm 0,5 - \mu_U}{\sigma_U} = \frac{U \pm 0,5 - \frac{n_1(N+1)}{2}}{\sqrt{\frac{n_1 n_2 (N+1)}{12}}} \xrightarrow[H_0^+]{d} N(0,1), \tag{5.7}$$

kur 0.5 *jāpieskaita, kad tiek meklētas varbūtības sadalījuma kreisajā pusē, bet jāatņem, kad sadalījuma labajā pusē.*

Situācijās, kad ir vairāki novērojumi ar vienādiem rangiem, tiek veikta dispersijas izlabošana

$$\sigma^2 = \frac{n_1 n_2}{N(N-1)} \left(\frac{N^3 - N}{12} - \sum_{j=1}^{g} \frac{t_j^3 - t_j}{12} \right). \tag{5.8}$$

Ievietojot dispersijas vērtību no vienādojuma (5.8) vienādojumā (5.7) redzams, ka nav nekādu izmaiņu, ja vienādu rangu nemaz nav, taču pretējā gadījumā rezultāts tiks uzlabots. Kā jau iepriekš minēts, Mana - Vitneja U tests ir līdzīgs parametriskajam t-testam. To var droši pielietot arī gadījumos, kad izpildās pieņēmums par normalitāti, jo U testa efektivitāte ir aptuveni 95% pat pie vidēji lielām izlasēm, salīdzinot ar t-testu. Turklāt ir novēroti arī gadījumi, kad U tests ir pat efektīvāks par parametrisko alternatīvu. Jāņem vērā arī tas, ka Mana - Vitneja testam ir ļoti neliela varbūtība uzrādīt zīmīgas atšķirības starp abām izlasēm dažu izteikti neiederīgu (*outliers*) datu dēļ, jo aprēķini tiek veikti, balstoties uz rangiem. Citiem vārdiem sakot, U tests ir robusts.

5.4. Kolmogorova - Smirnova tests

Nodaļā 3.3. Kolmogorova - Smirnova tests tika aplūkots vienas izlases gadījumā, ar kura palīdzību noteikt, vai dati nāk no normāli sadalītas populācijas, tas ir, tika salīdzinātas divas sadalījuma funkcijas - empīriskā un teorētiskā. Divu izlašu gadījumā tests pārbauda, vai abas izlases nāk no vienas un tās pašas populācijas (vai arī no populācijām ar vienādiem sadalījumiem), tātad tiek salīdzinātas abu izlašu empīriskās sadalījuma funkcijas.

$$H_0 : F_1(x) = F_2(x) \ \forall x \in \mathbb{R},$$

$$H_1 : \exists x \in \mathbb{R} : F_1(x) \neq F_2(x).$$

Ja starp abiem sadalījumiem tiek konstatētas gana lielas starpības, tad ir pamats noraidīt H_0.

Definīcija 19. Vienpusējā testa statistika aprēķināma kā

$$D_{m,n} = \sup_x [F_{2,m}(x) - F_{1,n}(x)], \tag{5.9}$$

kur $F_{1,n}(x)$ un $F_{2,m}(x)$ ir atbilstošo izlašu empīriskās sadalījuma funkcijas.

Definīcija 20. Tā kā divpusējais tests pēta atšķirības tikai starp abiem sadalījumiem, nenorādot virzienu, tad testa statistika izsakāma ar sadalījumu starpību absolūtajām vērtībām

$$D_{m,n} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| . \tag{5.10}$$

Salīdzinot Kolmogorova - Smirnova testu ar t-testu, ir noskaidrots, ka tā efektivitāte ir aptuveni 95% mazām izlasēm, taču tā nedaudz samazinās, pieaugot izlases apjomam. Tests visās situācijās ir jaudīgāks par tā alternatīvām - mediānas un hī - kvadrāta testiem, taču par Mana - Vitneja testu tas ir efektīvāks tikai pie ļoti mazām izlasēm.

6. Trīs vai vairāku savstarpēji atkarīgu izlašu salīdzi-nāšana

Līdz šim esam aplūkojuši gadījumus, kad tiek pētītas atšķirības vai nu starp vienu izlasi un kādu teorētisku sadalījumu, vai arī starp divām izlasēm. Teorētiski varētu domāt, ka vairāku izlašu gadījumā var vienkārši aplūkot visas izlases pa pāriem un veikt atbilstošos divu izlašu testus, taču tā nebūtu racionāla metode, jo, pirmkārt, būtu jāveic ļoti daudz šādu testu, un, otrkārt, veicot k testus ar būtiskuma līmeni α, pirmā veida kļūdas pieļaušanas iespēja tiek palielināta k reizes. Tāpēc ir izstrādāti statistiskie testi, kas paredzēti tieši šādām situācijām. Pazīstamākais no tādiem noteikti ir vienfaktora dispersiju analīze jeb ANOVA tabulas, taču, tā kā tas ir parametriskais tests, tad ir jāizpildās vairākiem nosacījumiem, lai to varētu pielietot. Šajā nodaļā aplūkosim divus neparametriskus testus, no kuriem viens paredzēts darbam ar nomināliem datiem, bet otrs - ar vismaz ordināliem.

6.1. Kohrana Q tests

Kohrana Q tests pēc būtības ir nodaļā 4.1. aplūkotā Maknemāra izmaiņu testa paplašinājums, kas paredzēts 3 vai vairāku savstarpēji atkarīgu izlašu salīdzināšanai, kuru novērojumi uzdoti ar diskrētiem datiem.

Pieņemsim, ka ir dotas k izlases apjomā N un dati ir sakārtoti tabulā ar N rindām un k kolonnām, kur katra kolonna reprezentē vienu no k mērījumiem, kas veikti *iid* novērojumiem $X_1, ..., X_N$. Lai pārbaudītu, vai kādas konkrētas vērtības g proporcija $G_j = \sum_{i=1}^{N} I_{X_i=g}$ ir vienāda visās izlasēs, izvirza nulles hipotēzi

$$H_0 : G_1 = G_2 = ... = G_k,$$

kuru pārbauda pret alternatīvo hipotēzi

$$H_1 : \exists\ i, j \in (1, k) : i \neq j, G_i \neq G_j.$$

Definīcija 21. Testa statistika ir aprēķināma pēc formulas

$$Q = \frac{k(k-1)\sum_{j=1}^{k}(G_j - \overline{G})^2}{k\sum_{i=1}^{N} L_i - \sum_{i=1}^{N} L_i^2}, \tag{6.1}$$

kur G_j ir interesējošās vērtības kopskaits j-tajā kolonnā (izlasē) un L_i interesējošās vērtības kopskaits i-tajā rindā.

Teorēma 13. *Ja $N \geq 4$ un $Nk \geq 24$, tad*

$$Q \xrightarrow[H_0]{d} \chi^2_{k-1}. \tag{6.2}$$

Kā jau iepriekš minēts, tad nomināliem datiem parametriskos testus nav korekti pielietot, līdz ar to Kohrana tests šādos gadījumos ir piemērotākais, taču ordināliem vai intervāla datiem var tikt pazaudēta daļa informācijas par datiem, līdz ar to iegūtie rezultāti var būt kļūdaini.

6.2. Frīdmana divvirzienu rangu dispersijas analīze

Ja dati ir uzdoti vismaz ordinālā veidā, tad labākā neparametriskā metode, lai pārbaudītu, vai izlases nāk no vienas un tās pašas populācijas, ir Frīdmana rangu dispersijas analīze. Gluži tāpat kā iepriekšējā nodaļā, ir dotas k izlases apjomā N (apjomi visām izlasēm ir vienādi, jo tās ir savstarpēji atkarīgas). Izkārtojot datus tabulā ar N rindām un k kolonnām, piešķiram visiem N savstarpēji atkarīgajiem novērojumiem (novērojumiem katras rindas ietvaros) rangus no 1 līdz k. Apzīmējot izlases mediānu ar θ, varam noformulēt nulles hipotēzi

$$H_0 : \theta_1 = \theta_2 = ... = \theta_k,$$

un alternatīvo hipotēzi

$$H_1 : \exists\ i, j \in (1, k) : i \neq j, \theta_i \neq \theta_j$$

tas ir, ka visas izlases ir no populācijām ar vienādām mediānām.

Teorēma 14. *Ja $k > 5$, tad testa statistika F_r pēc sadalījuma konverģē uz χ^2_{k-1}.*

$$F_r = \left[\frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2\right] - 3N(k+1) \xrightarrow[H_0]{d} \chi^2_{k-1} \tag{6.3}$$

Gadījumā, ja ir vairāki savstarpēji atkarīgie novērojumi, kam ir vienādi rangi, vienādojums (6.3) jāpārveido formā

$$F_r = \frac{12 \sum_{j=1}^{k} R_j^k - 3N^2k(k+1)^2}{Nk(k+1) + \frac{Nk - \sum_{i=1}^{N} \sum_{j=1}^{s_i} t_{i.j}^3}{k-1}}, \tag{6.4}$$

kur s_i ir kopu skaits ar vienādiem rangiem i-tajā grupā un $t_{i.j}$ ir i-tās grupas j-tās vienādo rangu kopas apjoms.

Pielietojot Frīdmana testu normāli sadalītiem datiem, tā efektivitāte pret F testu (ANOVA) ir 64%, kad $k = 2$, un tā pamazām pieaug līdz 91%, kad $k = 20$, līdz ar to ir skaidrs, ka, ja vien tas ir iespējams, labāk pielietot parametrisko testu.

7. Trīs vai vairāku savstarpēji neatkarīgu izlašu salīdzināšana

Veicot pētījumus, bieži vien ir nepieciešams noteikt, vai pastāv būtiskas atšķirības starp vairākām populācijām. Lai to noteiktu, no populācijām tiek ņemtas gadījuma izlases, kuru vērtības tiek salīdzinātas. Taču, tā kā gadījuma izlašu vērtības gandrīz vienmēr būs atšķirīgas kaut kādā mērā, tad uzdevums ir noteikt, vai šīs starpības norāda uz patiešām atšķirīgām vērtībām, vai arī tās ir tādas, kādas ir sagaidāmas, izvēloties gadījuma izlases.

Šajā nodaļā tiks aplūkoti 3 neparametriski statistiskie testi, ar kuru palīdzību var noteikt atšķirību būtiskumu starp populācijām. Parametriskā alternatīva šīm procedūrām ir vienfaktora dispersiju analīze (ANOVA) jeb F tests, kura pielietošanai ir jāizpildās vairākiem nosacījumiem. Ja dati uzdoti nominālā vai ordinālā skalā, vai arī tie nav normāli sadalīti un ar homogēnu dispersiju, tad jāizvēlas piemērotākais no šajā nodaļā aplūkotajiem statistiskajiem testiem.

7.1. Hī - kvadrāta tests

Iepriekšējās nodaļās aplūkojām hī - kvadrāta testus vienas un divu izlašu gadījumiem. Šajā nodaļā apskatītais tests ir nodaļā 5.2. jau pieminētā testa paplašinājums, kas piemērots 3 vai vairāku ar diskrētiem datiem uzdotu izlašu salīdzināšanai, turklāt tas ir gandrīz identisks divu izlašu gadījumam.

Doti *iid* gadījuma lielumi $X_1, X_2, ..., X_N$. Pieņemsim, ka tie iedalās k izlasēs un tie var pieņemt r dažādas vērtības. Apzīmējot izlases mediānu ar θ, varam noformulēt nulles un alternatīvo hipotēzi.

$$H_0 : \theta_1 = \theta_2 = ... = \theta_k,$$

$$H_1 : \exists\ i, j \in (1, k) : i \neq j, \theta_i \neq \theta_j.$$

Ja n_{ij} ir novērotais gadījumu skaits ar i-to vērtību j-tajā izlasē, bet $E_{ij} = np_{ij0}$ ir šī lieluma sagaidāmā vērtība, tad testa statistika pierakstāma formā

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{n_{ij}^2}{E_{ij}} - N. \tag{7.1}$$

Teorēma 15. *Ir spēkā*

$$X^2 \xrightarrow[H_0]{d} \chi^2_{(r-1)(k-1)}. \tag{7.2}$$

Lai iegūtie rezultāti būtu labi, nepieciešams, lai sagaidāmie biežumi nebūtu pārāk mazi. Ja brīvības pakāpju skaits ir lielāks par 1, tad ieteicams, ka ne vairāk kā 20% no visiem $E_{ij} < 5$, kā arī nevienam sagaidāmajam biežumam nevajadzētu būt mazākam par 1.

7.2. Kruskala - Vallisa rangu dispersijas analīze

Kruskala - Vallisa tests ir ļoti noderīga metode, ar kuras palīdzību noteikt, vai vairākas izlases nāk no vienas un tās pašas populācijas vai arī no identiskām populācijām ar vienādām mediānām, tas ir, lai pārbaudītu nulles hipotēzi

$$H_0 : \theta_1 = \theta_2 = ... = \theta_k.$$

pret alternatīvo

$$H_1 : \exists\ i, j \in (1, k) : i \neq j, \theta_i \neq \theta_j$$

Lai šo testu pielietotu, datiem jābūt uzdotiem vismaz ordinālā formā, kā arī visiem novērojumiem $X_1, ..., X_N$ jābūt *iid.* Metodes pamatā ir datu aizstāšana ar rangiem. Tiek ņemti pilnīgi visi novērojumi kopā, un katram tiek piešķirta atbilstoša ranga vērtība. Tālāk tiek aprēķinātas rangu summas katrai izlasei, un tad no tām tiek noteiktas izlašu vidējās rangu vērtības. Ja izlases ir no vienādām populācijām, tad arī to vidējām rangu vērtībām būtu jābūt vienādām, taču pretējā gadījumā tām būtu jāuzrāda būtiskas atšķirības.

Definīcija 22. Kruskala - Vallisa testa statistika aprēķināma pēc formulas

$$KW = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j (\overline{R}_j - \overline{R})^2, \tag{7.3}$$

kur n_j ir novērojumu skaits j-tajā izlasē, R_j ir rangu summa j-tajā izlasē un $\overline{R} = \frac{N+1}{2}$ ir vidējā ranga vērtība, ņemot visus datus kopā.

Gadījumā, ja vairākām vērtībām jāpiešķir vienādi rangi, statistikas vērtību iespējams uzlabot, izdalot vienādojumā (7.3) iegūto statistiku ar

$$1 - \frac{\sum_{i=1}^{g} (t_i^3 - t_i)}{N^3 - N}, \tag{7.4}$$

kur g ir dažādo grupu skaits ar vienādiem rangiem un t_i ir vienādo rangu skaits i-tajā grupā. Šī uzlabojuma mērķis ir palielināt statistikas KW vērtību, tādējādi padarot rezultātu būtiskāku. Ja ir neliels skaits grupu ar vienādiem rangiem, tad uzlabojums arī būs pavisam neliels, bet pie liela vienādu rangu skaita uzlabotā statistikas vērtība jūtami atšķirsies.

Teorēma 16. *Statistika KW konverģē uz hī - kvadrāta sadalījumu ar $k-1$ brīvības pakāpēm, kad H_0 ir spēkā*

$$KW \xrightarrow[H_0]{d} \chi^2_{k-1}. \tag{7.5}$$

Piezīme 17. *Grāmatā [14] ir apgalvots, ka aproksimācija ar χ^2_{k-1} ir laba, ja visi $n_j \geq 6$ un $k \geq 3$. Visos gadījumos pastāv cieša saikne starp šiem lielumiem un novērtējuma precizitāti.*

Salīdzinot ar parametrisko F testu, Kruskala - Vallisa testa efektivitāte asimptotiski tiecas uz $95,5\%$ gadījumos, kad ir korekti pielietot arī F testu, tāpēc to var uzskatīt par ļoti labu alternatīvu parametriskajai metodei. Turklāt Kruskala - Vallisa testa efektivitātei ir spēkā 4.3. nodaļā aprakstītās Vilkoksona testa efektivitātes īpašības,par kurām smalkāks izklāsts atrodams grāmatā [11].

7.3. Džonkhīra - Terpsta tests

Kruskala - Vallisa testā tiek pārbaudīta nulles hipotēze, ka k izlases ir ar vienādām mediānām, pret alternatīvo hipotēzi, ka viena vai vairākas izlašu mediānas atšķiras no citām. Ja ir nepieciešams pārbaudīt konkrētāku alternatīvo hipotēzi, tas ir, ka izlašu mediānas ieņem kādu noteiktu skaitlisku secību, nevis vienkārši atšķiras, tad ir jāpielieto Džonkhīra - Terpsta tests. Šādā gadījumā Kruskala - Vallisa testa alternatīvā hipotēze ir derīga, taču pārāk vispārēja.

Doti $X_1, ...X_N$ *iid* gadījuma lielumi, kas iedalīti k neatkarīgās izlasēs.

$$H_0 : \theta_1 = \theta_2 = ... = \theta_k$$

$$H_1 : \theta_1 < \theta_2 < ... < \theta_k$$

Definīcija 23. Par Mana - Vitneja skaitu sauc statistiku

$$U_{ij} = \sum_{h=1}^{n_i} \boxtimes(X_{hi}, j), \tag{7.6}$$

kur $\boxtimes(X_{hi}, j)$ ir gadījumu skaits, kad novērojums X_{hi} ir mazāks par novērojumu izlasē j, kad $i < j$.

Definīcija 24. Par Džonkhīra - Terpsta testa statistiku J sauc visu iespējamo Mana - Vitneja skaitu summu

$$J = \sum_{i<j}^{k} U_{ij} = \sum_{i=1}^{k} -1 \sum_{j=i+1}^{k} U_{ij}. \tag{7.7}$$

Teorēma 18. *Kad $N \to \infty$, tad ir spēkā*

$$J \xrightarrow[H_0]{d} N(\mu_J, \sigma_J^2), \tag{7.8}$$

kur

$$\mu_J = \frac{N^2 - \sum_{j=1}^{k} n_j^2}{4}$$

$$\sigma_J^2 = \frac{1}{72}\left[N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right].$$

Džonkhīra - Terpsta testa asimptotiskā efektivitāte ir 95.5%, salīdzinot ar atbilstošo parametrisko testu, taču tā ir aptuveni vienāda ar Kruskala - Vallisa testa efektivitāti.

8. Praktiskā daļa

8.1. Piemēri

Latvijas Valsts mežzinātnes institūts "Silava" nodarbojas ar dažādu apstākļu ietekmes uz koku parametriem pētīšanu. Šim nolūkam visā Latvijā ir ierīkoti eksperimentāli stādījumi, kuros regulāri tiek veikti dažādi mērījumi. Ievācamie dati var atšķirties atkarībā no aktuālā pētījuma, taču pilnvērtīgai analīzei tiek ievākta pēc iespējas visa pieejamā informācija. Kad dati ir ievākti, tie tiek apkopoti un tiek veiktas atbilstošas analīzes, kuru rezultāti tiek publicēti dažādos zinātniskos izdevumos vai projektu ietvaros.

Koku pazīmes var tikt novērtētas dažādos veidos. Piemēram, augstums un stumbru diametrs ir uzdots ar intervāla skalas datiem, zaru skaits vai skujbires novērtējums (skalā no 1 līdz 5) ir uzdots ar ordināliem datiem, bet kādas īpašības esamību vai neesamību var apzīmēt ar vērtībām 0 un 1.

Piemērs 1

Tiek aplūkota priežu plantācija, kurā aug vienā laikā stādīti koki. Viena no koku kvalitatīvajām pazīmēm ir augusta dzinumu rašanās, kas var būtiski ietekmēt stumbra kvalitāti. Ir dati par augusta dzinumu esamību 2009. un 2010. gadā, un ir nepieciešams noskaidrot, vai starp šiem gadiem atšķiras koku daudzums ar augusta dzinumiem. Tā kā dati var pieņemt tikai vērtības 0 vai 1 un abas izlases ir atkarīgas, tad pārbaudei var izmantot Maknemāra izmaiņu testu. Ar datorprogrammu SPSS 15.0 veiktie aprēķini parādīja, ka $p = 0.0012$, līdz ar to jānoraida hipotēze par augusta dzinumu daudzumu vienādību 2009. un 2010. gadā.

Piemērs 2

Aplūkosim to pašu plantāciju. Vairākus gadus pēc kārtas visiem kokiem ir nomērīts augstums. Nepieciešams noteikt, vai pieaugums, kas izveidojies 2009. gadā būtiski atšķiras no pieauguma 2010. gadā. Abi mērījumi ir savstarpēji atkarīgi un dati ir uzdoti nominālas skalas veidā, līdz ar to varam pielietot zīmju testu, Vilkoksona rangu zīmju testu, kā arī parametrisko t-testu. Ar datorprogrammu SPSS 15.0 veiktie aprēķini parādīja, ka $p = 0.0047$, līdz ar to visi trīs testi noraida nulles hipotēzi par abu izlašu vidējo vērtību vienādību.

Piemērs 3

Iepriekš noskaidrojām, ka augusta dzinumu daudzums pa gadiem atšķiras vienā plantācijā, taču vajadzētu noskaidrot, vai tas atšķiras arī vienā gadā divās ģeogrāfiski attālās plantācijās. Salīdzināsim datus no Ogres un Smiltenes plantācijām. Šajā gadījumā izlases ir neatkarīgas, līdz ar to jāpielieto vai nu Fišera, vai hī - kvadrāta tests, taču tā kā izlašu apjomi ir ļoti lieli, tad jālieto tieši hī - kvadrāta tests. Ar programmu **R** aprēķināts, ka testa statistikai $p = 0.00008$, līdz ar to jānoraida hipotēze, ka abās plantācijas augusta dzinumu daudzums ir vienāds.

Piemērs 4

Zaru skaits augstākajā mieturī arī ir būtiska koku kvalitatīvā pazīme, kas pieņem diskrētas vērtības. Lai pārbaudītu, vai abās plantācijas zaru skaiti ir vienādi, var pielietot vai nu Mana - Vitneja, vai Kolmogorova - Smirnova testu. Veicot abos šos testus ar

SPSS, tika iegūts $p = 0.0051$, līdz ar to nulles hipotēze jānoraida un nevar apgalvot, ka abās plantācijās zaru skaiti ir vienādi. Salīdzinājumam tika veikts arī t-tests, kas uzrādīja aptuveni tādu pašu rezultātu.

Piemērs 5

Ogres plantācijā stādījums ir ierīkots vienas un tās pašas koku ģimenes stādot 4 blakus esošos laukos. Lai noteiktu, vai arī tik nelielas ģeogrāfiskā novietojuma atšķirības var būtiski ietekmēt koku skujbiri, jāveic vai nu Kruskala - Vallisa tests, vai arī parametriskā vienfaktora dispersiju analīze. Tā kā skujbire ir novērtēta ballēs no 1 līdz 5, tad skaidrs, ka priekšroka tiek dota neparametriskajam testam. Ar Kruskala - Vallisa testu iegūts, ka $p = 0.054$, kas liedz noraidīt H_0 pie būtiskuma līmeņa $\alpha = 0.05$, savukārt ar ANOVA iegūts $p = 0.039$. Tātad rezultāti ir nepārliecinoši. Lai pārbaudītu, cik korekti ir pielietot ANOVA, veiksim datu normalitātes un dispersijas homogenitātes pārbaudi. Ar Levena testu iegūts, ka dispersijas nav homogēnas, savukārt ar Šapiro - Vilka un Kolmogorova - Smirnova testu ar Lilifora korekciju iegūts, ka dati neatbilst normālajam sadalījumam, līdz ar to varam secināt, ka Kruskala - Vallisa tests šajā gadījumā sniedz labākus rezultātus.

8.2. Simulācijas

Lai pārbaudītu pirmā un otrā veida kļūdas pieļaušanas varbūtības, tika veiktas simulācijas programmā **R**. Divu neatkarīgu izlašu gadījumā tika salīdzināts t-tests un Mana - Vitneja U tests. Pārbaudei tika izvēlēti 4 dažādi datu sadalījumi - normālais, lognormālais, Košī un dubulteksponenciālais (saukts arī par Laplasa sadalījumu).

Divu izlašu gadījumā tika ģenerētas divas izlases ar vienādiem sadalījumiem. Pirmā veida kļūdas pārbaudei abām izlasēm tika izvēlēti vienādi parametri, piemēram, $N(0, 1)$, taču otrā veida kļūdas pārbaudei katrai izlasei bija atšķirīgi parametri. Parametru atšķirības attēlotas 5. tabulā.

Vairāku izlašu gadījumā tika ģenerētas 5 izlases ar vienādiem sadalījumiem, kur pirmā veida kļūdas pārbaudei visām izlasēm bija vienādi sadalījuma parametri, savukārt otrā veida kļūdas pārbaudei vienai izlasei parametri atšķīrās no pārējo izlašu parametriem tieši tāda pašā veidā, kā attēlots 5. tabulā.

Lai pētītu testu uzvedību gadījumos, kad daļa datu ir izteikti neiederīgi, salīdzinot ar kopējo datu sadalījumi, tika veikta tā sauktā datu piesārņošana. Šim nolūkam tika

5. tabula Izlašu sadalījumi un to parametri

viena izlase	otra izlase
$N(0,1)$	$N(1,1)$
$ln(0,1)$	$ln(1,1)$
$K(0,1)$	$K(1,1)$
$Dexp(0,10)$	$Dexp(1,10)$

izmantota programmā **R** iebūvētā funkcija "contaminate". Šīs procedūras būtība ir ar iepriekš noteiktu varbūtības līmeni aizstāt elementu gadījuma izlasē ar no šī elementa atkarīgu funkciju. Šajā darbā tika izmantots aizstāšanas varbūtības līmenis $\epsilon = 0.05$ un elements x tika aizstāts ar $f(x) = x * 10$. Tika aplūkotas gan situācijas, kad piesārņotu datu nav, gan situācijas, kad piesārņotie dati ir tikai vienā izlasē, gan arī situācijas, kad tādi ir visās izlasēs.

Pielikumā atrodamajās tabulās dati ir izteikti kā nepareizi noraidīto pareizo nulles hipotēžu apjoms pret veikto testu daudzumu pirmā veida kļūdai, savukārt otrā veida kļūdas vietā izteikta testu jauda, kas aprēķināta kā pareizi noraidīto nepareizu nulles hipotēžu apjoms pret veikto testu daudzumu. Piesārņotie dati apzīmēti kā *outliers*.

Pirmā veida kļūdas pārbaudei tika salīdzinātas izlases ar vienādām vidējām vērtībām un novērtēta varbūtība noraidīt pareizu H_0. Šim nolūkam tika aplūkotas izlases ar apjomiem $N = (5, 10, 25, 30, 50, 100)$. Katram gadījumam simulācijas atkārtotas 1000 reižu un tām iegūts nepareizi noraidīto hipotēžu skaits. Tabulās redzams, ka abu testu pirmā veida kļūdas varbūtības visiem sadalījumiem ir ļoti līdzīgas, lielākoties tās atbilst būtiskuma līmenim $\alpha = 0.05$. Košī sadalījumam t-tests uzrāda krietni labākus rezultātus, taču interesanta tendence parādās lognormālajam sadalījumam gadījumā, kad daļa datu ir izteikti neiederīgi (*outliers*). Šajā situācijā redzams, ka pirmā veida kļūdas varbūtība konverģē uz 1, kad izlašu apjoms $N \to \infty$, taču Mana - Vitneja testam šī konverģence ir krietni lēnāka, kas liecina par testa robustību. Vēl zīmīgi, ka normāli sadalītiem datiem neparametriskais tests uzrāda labākus rezultātus pie ļoti mazām izlasēm.

Otrā veida kļūdas pieļaušanas varbūtības β vietā aplūkota testu jauda $1-\beta$. Simulētas tādas pašas situācijas kā pirmā veida kļūdas gadījumā, taču šoreiz izlases ģenerētas ar atšķirīgiem sadalījumu parametriem. Visos gadījumos

$$1 - \beta \underset{N \to \infty}{\to} 1,$$

taču šī konverģence atšķiras gan starp testiem, gan starp sadalījumiem. Normālajam un lognormālajam sadalījumam mazliet labākus rezultātus uzrāda t-tests, taču visos pārējos gadījumos Mana - Vitneja tests ir jaudīgāks. Izteikti neiederīgu datu ieviešana t-teta rezultātus padara vēl sliktākus, turklāt Košī sadalījumam tā jauda ir izteikti slikta. Dubulteksponenciālajam sadalījumam ar izteikti neiederīgiem datiem abi testa uzrāda sliktus rezultātus, taču arī šeit neparametriskais tests ir labāks.

Lai salīdzinātu parametrisko vienfaktora dispersiju analīzi (ANOVA) ar neparametrisko Kruskala - Vallisa, veiktas līdzīgas simulācijas kā divu izlašu gadījumā. Ģenerētas 5 neatkarīgas izlases ar kopējo apjomu $N = (25, 50, 100, 250, 500, 1000, 2000, 5000)$. Izvēlēti 3 datu sadalījumi - normālais, Košī un dubulteksponenciālais. Aplūkotas situācijas, kad dati ir bez īpaši neiederīgiem, kā arī situācijas, kad īpaši neiederīgi dati ir, turklāt vienā variantā tie ir vienmērīgi sadalīti starp visām izlasēm, bet otrā tie ir tikai vienā izlasē.

Pielikumā atrodamajās tabulās redzams, ka kopumā abiem testiem pirmā veida kļūdas pieļaušanas varbūtība ir ļoti līdzvērtīga. Dažos gadījumos tā mazliet labāka vienam testam, dažos otram, taču kopumā mazliet labāka tendence ir t-testam. Kļūdas pieļaušanas varbūtība ļoti reti pārsniedz būtiskuma līmeni $\alpha = 0.05$, vienīgi gadījumā, kad vienā izlasē ir izteikti neiederīgi dati, ANOVA kļūda krietni pieaug pie lieliem izlašu apjomiem.

Testu jaudas normāli sadalītiem datiem ļoti strauji sasniedz 1, turklāt redzams, ka izpildās teorija, ka Kruskala - Vallisa testa jauda ir nedaudz vājāka, izņemot gadījumus, kad ir izteikti neiederīgi dati. Košī sadalījumam ANOVA jauda ir tik vāja, ka var apgalvot, ka tests nedarbojas, savukārt Kruskala - Vallisa testam jauda palielinās pie lielām izlasēm, turklāt tā konverģē uz 1. Dubulteksponenciālajam sadalījumam abi testi uzrāda ļoti sliktus rezultātus. Lai arī abiem testiem jauda

$$1 - \beta \underset{N \to \infty}{\to} 1,$$

tas notiek ļoti lēnām. Mazliet straujāka konverģence piemīt Kruskala - Vallisa testam.

Secinājumi

Darbā aplūkoti dažādi neparametriski testi, kas piemēroti atšķirīgu situāciju analīzēm. Tika aplūkoti vieni no populārākajiem un plašāk pielietotajiem neparametriskajiem testiem, kas jālieto gadījumos, kad datu struktūra ir nepiemērota vai arī neizpildās parametrisko testu nosacījumi. Lielākā daļa aplūkoto testu pieejami populārākajās statistiskajās datorprogrammās, piemēram, SPSS, **R**, SAS.

Darba mērķis bija analizēt piemērotākās metodes izvēli datu pētīšanai konkrētā situācijā, kā arī izprast atšķirību starp parametriskām un neparametriskām metodēm. Kā galvenie novērtējumi testu kvalitātei darbā apskatītas pirmā un otrā veida kļūdas, kā arī testu Pitmana asimptotiskā efektivitāte.

Praktiskajā daļā ir parādīti piemēri, kur ikdienā tiek pielietota hipotēžu pārbaude ar neparametriskajiem testiem. testi veikti programmās **R** un SPSS. Tieši saskare ar pielietojumu praksē mudināja mani veikt padziļinātu analīzi par neparametriskajiem testiem, lai ikdienā veiktās analīzes būtu korektas un pamatotas.

Ar datu simulācijām parādīts, ka patiešām neparametriskie testi testi var uzrādīt gandrīz vienlīdz labus vai atsevišķos gadījumos pat labākus rezultātus nekā parametriskie testi. Apstiprinājās teorētiskais pieņēmums, ka neparametriskie testi daudzos gadījumos ir labāki mazām izlasēm un datiem, kuros ir daži izteikti neiederīgi dati, kā arī tas, ka dažiem sadalījumiem parametriskās metodes strādā ļoti slikti.

Tā kā darbā aplūkotie testi pārsvarā ir domāti izlašu vidējo vērtību salīdzināšanai, tad būtu noderīgi padziļināti aplūkot arī ticamības intervālu konstruēšanu un statistiku būtiskuma novērtēšanu, piemēram, ar Monte Karlo simulācijām vai butstrapa metodēm. Darbā aprakstīta tikai Pitmana asimptotiskā efektivitāte, taču ir aprēķināmas arī cita veida efektivitātes.

Izmantotā literatūra un avoti

[1] A. DasGupta. *Asymptotic theory of statistics and probability*. Springer Verlag, 2008.

[2] I. Arhipova and S. Bāliņa. *Statistika ekonomikā: risinājumi ar SPSS un Microsoft Excel*. Datorzinību centrs, Rīga, 2003.

[3] M. Lynch and B. Walsh. *Genetics and analysis of quantitative traits*, volume 24. Sinauer Sunderland, MA, 1998.

[4] Nicholas J. Gotelli and Aaron M. Ellison. *A Primer of Ecological Statistics*. Sinauer Sunderland, MA, 2004.

[5] A.P. Field. *Discovering statistics using SPSS*. SAGE publications Ltd, 2009.

[6] M.B. Brown and A.B. Forsythe. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367, 1974.

[7] `http://udel.edu/~mcdonald/statexactbin.html`.

[8] G.W. Corder and D.I. Foreman. *Nonparametric statistics for non-statisticians: a step-by-step approach*. John Wiley & Sons Inc, 2009.

[9] `http://en.wikipedia.org/wiki/Wald-Wolfowitz_runs_test`.

[10] S. Siegel and N.J. Castellan Jr. *Nonparametric statistics for the behavioral sciences* McGraw-Hill Book Company, 1988.

[11] J.L. Hodges and E.L. Lehmann. The efficiency of some nonparametric competitors of the t-test. *The Annals of Mathematical Statistics*, 27(2):324–335, 1956.

[12] `http://en.wikipedia.org/wiki/Fisher\LV\textquoterights_exact_test`.

[13] C. Robson. *Real world research: A resource for social scientists and practitioner-researchers*. Wiley-Blackwell, 2002.

[14] E. Brunner and U. Munzel. *Nichtparametrische Datenanalysen.* Springer Verlag, 2002.

1. Pielikums

1.1. Simulāciju ceļā iegūtie rezultāti

6. tabula Divas neatkarīgas izlases, pirmā veida kļūda

N	t-tests	U tests	t-tests	U tests	t-tests	U tests	t-tests	U tests
bez *outliers*	$N(0,1)$		$ln(0,1)$		$K(0,1)$		$Dexp(0,10)$	
5	0.047	0.032	0.043	0.032	0.014	0.031	0.032	0.030
10	0.049	0.042	0.050	0.044	0.018	0.042	0.045	0.045
25	0.047	0.044	0.049	0.048	0.020	0.053	0.048	0.051
30	0.051	0.053	0.053	0.051	0.019	0.047	0.052	0.050
50	0.051	0.052	0.051	0.050	0.021	0.054	0.050	0.050
100	0.052	0.053	0.052	0.052	0.020	0.047	0.047	0.049
1000	0.051	0.052	0.048	0.048	0.023	0.051	0.053	0.054
abās *outliers*								
5	0.009	0.037	0.001	0.005	0.005	0.036	0.005	0.030
10	0.008	0.036	0.003	0.024	0.012	0.046	0.014	0.034
25	0.028	0.060	0.004	0.031	0.014	0.049	0.019	0.042
30	0.029	0.042	0.003	0.029	0.023	0.053	0.023	0.042
50	0.039	0.052	0.007	0.052	0.024	0.048	0.033	0.055
100	0.035	0.046	0.005	0.040	0.022	0.050	0.026	0.042
1000	0.041	0.038	0.014	0.036	0.022	0.053	0.055	0.045
vienā *outliers*								
5	0.010	0.023	0.008	0.045	0.005	0.030	0.009	0.024
10	0.024	0.046	0.012	0.039	0.012	0.045	0.017	0.032
25	0.023	0.051	0.033	0.056	0.017	0.040	0.017	0.052
30	0.026	0.054	0.050	0.060	0.023	0.046	0.036	0.052
50	0.039	0.057	0.068	0.054	0.014	0.052	0.027	0.037
100	0.049	0.042	0.13	0.079	0.016	0.055	0.032	0.053
1000	0.059	0.061	0.98	0.377	0.018	0.053	0.040	0.037

7. tabula Divas neatkarīgas izlases, testu jauda

N	t-tests	U tests	t-tests	U tests	t-tests	U tests	t-tests	U tests
bez *outliers*	$N(0,1)$		$ln(0,1)$		$K(0,1)$		$Dexp(0,10)$	
5	0.265	0.201	0.250	0.198	0.055	0.096	0.178	0.139
10	0.553	0.512	0.559	0.514	0.062	0.182	0.360	0.399
25	0.935	0.922	0.931	0.920	0.077	0.428	0.697	0.805
30	0.969	0.963	0.967	0.958	0.074	0.503	0.765	0.870
50	0.999	0.998	0.999	0.998	0.074	0.737	0.935	0.980
100	1	0.999	1	1	0.077	0.954	0.998	1
1000	1	1	1	1	0.082	1	1	1
abās *outliers*								
5	0.021	0.161	0.013	0.065	0.011	0.079	0.009	0.026
10	0.134	0.446	0.093	0.372	0.035	0.164	0.010	0.038
25	0.468	0.890	0.283	0.885	0.067	0.418	0.023	0.060
30	0.568	0.939	0.352	0.929	0.073	0.487	0.039	0.057
50	0.769	0.995	0.621	0.996	0.080	0.750	0.045	0.075
100	0.970	1	0.864	1	0.076	0.957	0.049	0.103
1000	1	1	1	1	0.089	1	0.178	0.496
vienā *outliers*								
5	0.058	0.243	0.047	0.292	0.021	0.111	0.012	0.031
10	0.221	0.541	0.193	0.576	0.051	0.199	0.032	0.052
25	0.727	0.941	0.629	0.960	0.099	0.482	0.018	0.056
30	0.800	0.966	0.705	0.973	0.057	0.523	0.037	0.057
50	0.973	1	0.874	1	0.088	0.742	0.045	0.084
100	0.999	1	0.981	1	0.095	0.959	0.058	0.092
1000	1	1	1	1	0.105	1	0.218	0.495

8. tabula Piecas neatkarīgas izlases, pirmā veida kļūda

N	F	KW	F	KW	F	KW
bez *outliers*	$N(0,1)$		$K(0,1)$		$Dexp(0,10)$	
25	0.040	0.022	0.025	0.035	0.047	0.044
50	0.042	0.043	0.024	0.042	0.034	0.032
100	0.055	0.047	0.015	0.044	0.052	0.056
250	0.043	0.044	0.018	0.043	0.048	0.057
500	0.046	0.038	0.015	0.047	0.037	0.031
1000	0.048	0.051	0.026	0.063	0.044	0.045
2000	0.043	0.050	0.016	0.063	0.047	0.045
5000	0.048	0.050	0.013	0.040	0.038	0.035
visās *outliers*						
25	0.023	0.030	0.036	0.030	0.026	0.036
50	0.024	0.044	0.027	0.036	0.018	0.057
100	0.025	0.041	0.015	0.063	0.030	0.042
250	0.032	0.045	0.012	0.044	0.040	0.052
500	0.034	0.047	0.012	0.035	0.034	0.047
1000	0.056	0.044	0.014	0.044	0.047	0.045
2000	0.049	0.050	0.020	0.042	0.050	0.051
5000	0.041	0.044	0.015	0.040	0.044	0.050
vienā *outliers*						
25	0.027	0.036	0.020	0.038	0.028	0.033
50	0.035	0.045	0.022	0.053	0.040	0.049
100	0.038	0.036	0.018	0.054	0.038	0.048
250	0.051	0.047	0.022	0.060	0.049	0.048
500	0.073	0.063	0.015	0.046	0.065	0.053
1000	0.075	0.048	0.018	0.049	0.069	0.043
2000	0.065	0.034	0.019	0.051	0.078	0.062
5000	0.082	0.044	0.021	0.049	0.076	0.043

9. tabula Piecas neatkarīgas izlases, testu jauda

N	F	KW	F	KW	F	KW
bez *outliers*	$N(0,1)$		$K(0,1)$		$Dexp(0,10)$	
25	0.270	0.202	0.061	0.060	0.052	0.032
50	0.529	0.486	0.032	0.145	0.044	0.050
100	0.883	0.863	0.028	0.326	0.055	0.057
250	1	1	0.044	0.758	0.061	0.061
500	1	1	0.031	0.974	0.070	0.082
1000	1	1	0.029	1	0.098	0.115
2000	1	1	0.024	1	0.135	0.174
5000	1	1	0.027	1	0.316	0.464
visās *outliers*						
25	0.073	0.206	0.051	0.085	0.019	0.037
50	0.163	0.466	0.031	0.189	0.018	0.051
100	0.407	0.850	0.027	0.322	0.026	0.044
250	0.880	0.999	0.027	0.692	0.038	0.055
500	0.997	1	0.029	0.971	0.049	0.076
1000	1	1	0.026	1	0.063	0.096
2000	1	1	0.034	1	0.076	0.209
5000	1	1	0.028	1	0.133	0.460
vienā *outliers*						
25	0.139	0.240	0.051	0.088	0.030	0.026
50	0.413	0.508	0.041	0.192	0.036	0.041
100	0.819	0.867	0.047	0.374	0.042	0.065
250	0.977	1	0.041	0.784	0.069	0.069
500	1	1	0.046	0.978	0.107	0.078
1000	1	1	0.029	1	0.125	0.122
2000	1	1	0.038	1	0.164	0.181
5000	1	1	0.052	1	0.305	0.482

1.2. Izveidoto programmu kods

```
## 2 neatkarigas izlases##
## normalais sadalijums#
N<-100
norm_t=0
norm_u=0
norm_kop=0
n_norm1<-c(1:10000)
n_norm2<-c(1:10000)
for (i in 1:10000)
{
norm1<-rnorm(N, 0,1)
norm2<-rnorm(N, 1,1)
p[i]<-t.test(norm1,norm2)$p.value
w[i]<-wilcox.test(norm1,norm2)$p.value
if (p[i]<0.05) {norm_t=norm_t+1}
if (w[i]<0.05) {norm_u=norm_u+1}
if (p[i]<0.05 && w[i]<0.05) {norm_kop=norm_kop+1}
}
norm_t
norm_u
norm_kop
##lognormalais##
N<-100
lnorm_t=0
lnorm_u=0
lnorm_kop=0
n_lnorm1<-c(1:10000)
n_lnorm2<-c(1:10000)
for (i in 1:10000)
{
lnorm1<-rnorm(N, 0,1)
lnorm2<-rnorm(N, 1,1)
p[i]<-t.test(lnorm1,lnorm2)$p.value
w[i]<-wilcox.test(lnorm1,lnorm2)$p.value
if (p[i]<0.05) {lnorm_t=lnorm_t+1}
if (w[i]<0.05) {lnorm_u=lnorm_u+1}
if (p[i]<0.05 && w[i]<0.05) {lnorm_kop=lnorm_kop+1}
```

```
}
lnorm_t
lnorm_u
lnorm_kop
###
##koshii##
N<-1000
cauchy_t=0
cauchy_u=0
cauchy_kop=0
n_cauchy1<-c(1:10000)
n_cauchy2<-c(1:10000)
for (i in 1:10000)
{
cauchy1<-rcauchy(N, 0,1)
cauchy2<-rcauchy(N, 1,1)
p[i]<-t.test(cauchy1,cauchy2)$p.value
w[i]<-wilcox.test(cauchy1,cauchy2)$p.value
if (p[i]<0.05) {cauchy_t=cauchy_t+1}
if (w[i]<0.05) {cauchy_u=cauchy_u+1}
if (p[i]<0.05 && w[i]<0.05) {cauchy_kop=cauchy_kop+1}
}
cauchy_t
cauchy_u
cauchy_kop
###
##laplasa dub exp##
N<-25
lap_t=0
lap_u=0
lap_kop=0
n_lap1<-c(1:10000)
n_lap2<-c(1:10000)
for (i in 1:10000)
{
lap1<-urlaplace(N, 0,10)
lap2<-urlaplace(N, 0,10)
p[i]<-t.test(lap1,lap2)$p.value
w[i]<-wilcox.test(lap1,lap2)$p.value
```

```
if (p[i]<0.05) {lap_t=lap_t+1}
if (w[i]<0.05) {lap_u=lap_u+1}
if (p[i]<0.05 && w[i]<0.05) {lap_kop=lap_kop+1}
}
lap_t
lap_u
lap_kop
###
library(simFrame)
png(filename="contaminate_%03d_med.png", width=480, height=480)
showClass("VirtualContControl")
##divas izlses, visaas izlases ir izlecējdati##
## normalais#
N<-5
norm_t=0
norm_u=0
norm_kop=0
for (i in 1:1000)
{
foo<-generate(size=N, distribution=rnorm, dots=list(mean=1))
bar<-contaminate(foo, "DARContControl", target= "V1",
 epsilon=0.05, fun=function(x) x* 10)
boo<-generate(size=N, distribution=rnorm, dots=list(mean=0))
bor<-contaminate(boo, "DARContControl", target= "V1",
 epsilon=0.05, fun=function(x) x* 10)
p[i]<-t.test(bar$V1,bor$V1)$p.value
w[i]<-wilcox.test(bar$V1,bor$V1)$p.value
if (p[i]<0.05) {norm_t=norm_t+1}
if (w[i]<0.05) {norm_u=norm_u+1}
if (p[i]<0.05 && w[i]<0.05) {norm_kop=norm_kop+1}
}
norm_t
norm_u
##lognormalais##
N<-1000
lnorm_t=0
lnorm_u=0
lnorm_kop=0
for (i in 1:1000)
```

```
{
foo<-generate(size=N, distribution=rlnorm, dots=list(mean=1))
bar<-contaminate(foo, "DARContControl", target= "V1",
 epsilon=0.05, fun=function(x) x* 10)
boo<-generate(size=N, distribution=rlnorm, dots=list(mean=0))
bor<-contaminate(boo, "DARContControl", target= "V1",
 epsilon=0.05, fun=function(x) x* 10)
p[i]<-t.test(bar$V1,bor$V1)$p.value
w[i]<-wilcox.test(bar$V1,bor$V1)$p.value
if (p[i]<0.05) {lnorm_t=lnorm_t+1}
if (w[i]<0.05) {lnorm_u=lnorm_u+1}
if (p[i]<0.05 && w[i]<0.05) {lnorm_kop=lnorm_kop+1}
}
lnorm_t
lnorm_u
##koshii##
N<-1000
cauchy_t=0
cauchy_u=0
cauchy_kop=0
for (i in 1:1000)
{
foo<-generate(size=N, distribution=rcauchy, dots=list(location=1))
bar<-contaminate(foo, "DARContControl", target= "V1",
 epsilon=0.05, fun=function(x) x* 10)
boo<-generate(size=N, distribution=rcauchy, dots=list(location=0))
bor<-contaminate(boo, "DARContControl", target= "V1",
 epsilon=0.05, fun=function(x) x* 10)
p[i]<-t.test(bar$V1,bor$V1)$p.value
w[i]<-wilcox.test(bar$V1,bor$V1)$p.value
if (p[i]<0.05) {cauchy_t=cauchy_t+1}
if (w[i]<0.05) {cauchy_u=cauchy_u+1}
if (p[i]<0.05 && w[i]<0.05) {cauchy_kop=cauchy_kop+1}
}
cauchy_t
cauchy_u
##laplasa dub exp##
N<-5000
lap_t=0
```

```
lap_u=0
lap_kop=0
for (i in 1:1000)
{
foo<-generate(size=N, distribution=urlaplace, dots=list(location=1, scale=10))
bar<-contaminate(foo, "DARContControl", target= "V1",
 epsilon=0.05, fun=function(x) x* 10)
boo<-generate(size=N, distribution=urlaplace, dots=list(location=0, scale=10))
bor<-contaminate(boo, "DARContControl", target= "V1",
 epsilon=0.05, fun=function(x) x* 10)
p[i]<-t.test(bar$V1,bor$V1)$p.value
w[i]<-wilcox.test(bar$V1,bor$V1)$p.value
if (p[i]<0.05) {lap_t=lap_t+1}
if (w[i]<0.05) {lap_u=lap_u+1}
if (p[i]<0.05 && w[i]<0.05) {lap_kop=lap_kop+1}
}
lap_t
lap_u
lap_kop
###
## 5 neatkarigas izlases##
##normalais##
N<-200
k<-4
Nn<-round(N/k)
Nn+N
Lk<-LETTERS[1:k]
Lkk<-LETTERS[k+1]
norm_aov<-0
norm_kw<-0
for (i in 1:1000){
a<-rnorm(N,0,1)
b<-rnorm(Nn,1,1)
d<-data.frame(x=c(a,b), fac=c(sample(Lk, N,
 replace=TRUE),sample(Lkk, Nn, replace=TRUE)))
p[i]<-summary(aov(d$x~d$fac))[[1]][["Pr(>F)"]]
w[i]<-kruskal.test(d$x~d$fac)$p.value
if (p[i]<0.05) {norm_aov=norm_aov+1}
if (w[i]<0.05) {norm_kw=norm_kw+1}
```

```
}
norm_aov
norm_kw
##koshii####
N<-1600
k<-4
Nn<-round(N/k)
Nn+N
Lk<-LETTERS[1:k]
Lkk<-LETTERS[k+1]
kos_aov<-0
kos_kw<-0
for (i in 1:1000){
a<-rcauchy(N,0,1)
b<-rcauchy(Nn,1,1)
d<-data.frame(x=c(a,b), fac=c(sample(Lk, N,
 replace=TRUE),sample(Lkk, Nn, replace=TRUE)))
p[i]<-summary(aov(d$x~d$fac))[[1]][["Pr(>F)"]]
w[i]<-kruskal.test(d$x~d$fac)$p.value
if (p[i]<0.05) {kos_aov=kos_aov+1}
if (w[i]<0.05) {kos_kw=kos_kw+1}
}
kos_aov
kos_kw
##laplasa dub exp####
N<-4000
k<-4
Nn<-round(N/k)
Nn+N
Lk<-LETTERS[1:k]
Lkk<-LETTERS[k+1]
lap_aov<-0
lap_kw<-0
for (i in 1:1000){
a<-urlaplace(N,0,10)
b<-urlaplace(Nn,1,10)
d<-data.frame(x=c(a,b), fac=c(sample(Lk, N,
 replace=TRUE),sample(Lkk, Nn, replace=TRUE)))
p[i]<-summary(aov(d$x~d$fac))[[1]][["Pr(>F)"]]
```

```
w[i]<-kruskal.test(d$x~d$fac)$p.value
if (p[i]<0.05) {lap_aov=lap_aov+1}
if (w[i]<0.05) {lap_kw=lap_kw+1}
}
lap_aov
lap_kw
##5 neatkarigas izlases, dati ar izleceejiem##
##cont##
N<-4000
k<-4
Nn<-round(N/k)
Nn+N
Lk<-LETTERS[1:k]
Lkk<-LETTERS[k+1]
norm_aov<-0
norm_kw<-0
for (i in 1:1000){
foo<-generate(size=N, distribution=urlaplace, dots=list(location=0, scale=10))
a<-contaminate(foo, "DARContControl", target= "V1",
 epsilon=0.05, fun=function(x) x* 10)
boo<-generate(size=Nn, distribution=urlaplace, dots=list(location=1, scale=10))
b<-contaminate(boo, "DARContControl", target= "V1",
 epsilon=0.05, fun=function(x) x* 10)
d<-data.frame(x=c(foo$V1,b$V1), fac=c(sample(Lk, N,
 replace=TRUE),sample(Lkk, Nn, replace=TRUE)))
p[i]<-summary(aov(d$x~d$fac))[[1]][["Pr(>F)"]]
w[i]<-kruskal.test(d$x~d$fac)$p.value
if (p[i]<0.05) {norm_aov=norm_aov+1}
if (w[i]<0.05) {norm_kw=norm_kw+1}
}
norm_aov
norm_kw
```

Diplomdarbs "Hipotēžu pārbaude ar neparametriskajiem testiem" izstrādāts LU Fizikas un Matemātikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Juris Rieksts-Riekstiņš

____________________ ________________
(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.
Vadītājs: doc. Dr.math. Jānis Valeinis

____________________ ________________
(paraksts) (datums)

Recenzents: doc. Dr. Math. Nadežda Siņenko

Darbs iesniegts Matemātikas nodaļā ________________
(datums)

____
(darbu pieņēma)

Diplomdarbs aizstāvēts valsts pārbaudījuma komisijas sēdē

________________ prot. Nr. ________, vērtējums________
(datums)

Komisijas sekretāre: asoc. prof. Dr.math. Inese Bula ________________
(paraksts)