UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO
FAKULTETA ZA MATEMATIKO IN FIZIKO

Amela Rakanović

Statistični gradniki za orodje Orange

DIPLOMSKO DELO
UNIVERZITETNI ŠTUDIJSKI PROGRAM PRVE STOPNJE
RAČUNALNIŠTVO IN MATEMATIKA

Mentor: izr. prof. dr. Janez Demšar

Ljubljana 2013

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Tržaška 25
1001 Ljubljana, Slovenija
telefon: 01 476 84 11
01 476 83 87
faks: 01 426 46 47
01 476 87 11
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 00019/2013
Datum: 05.04.2013

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko ter Fakulteta za matematiko in fiziko izdaja naslednjo nalogo:

Kandidat: **AMELA RAKANOVIĆ**

Naslov: **STATISTIČNI GRADNIKI ZA ORODJE ORANGE**
STATISTICAL WIDGETS FOR ORANGE PLATFORM

Vrsta naloge: Diplomsko delo univerzitetnega študija prve stopnje

Tematika naloge:

Orange je orodje za odkrivanje znanj iz podatkov. Njegov grafični vmesnik temelji na vizualnem programiranju. Uporabnik postavlja na platno grafične gradnike (angl. widgets), ki opravljajo posamezne operacije, in jih povezuje v sheme. Gradniki so razdeljeni v več skupin, kot so "Data" (branje podatkov, predprocesiranje, vzorčenje...), "Visualize" (različne vizualizacije podatkov), "Classify" (gradnja napovednih modelov) in podobno.

V trenutni različici Orange nima gradnikov za statistično analizo v smislu osnovnih statističnih testov, kot so t-test, ANOVA, analiza korelacij in podobni. V okviru diplomske naloge si zamislite gradnike, ki bodo vključevali tako osnovne teste kot njihovo primerno grafično predstavitev. Tako zamišljene gradnike implementirajte in testirajte.

Mentor:
izr. prof. dr.Janez Demšar

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

Dekan Fakultete za računalništvo in informatiko:
prof. dr. Nikolaj Zimic

Dekan Fakultete za matematiko in fiziko:
akad. prof. dr. Franc Forstnerič

UNIVERZA V LJUBLJANI • Fakulteta za matematiko in fiziko • 1

IZJAVA O AVTORSTVU

diplomskega dela

Spodaj podpisana Amela Rakanović,

z vpisno številko 63080458,

sem avtorica diplomskega dela z naslovom:

STATISTIČNI GRADNIKI ZA ORODJE ORANGE

STATISTICAL WIDGETS FOR ORANGE PLATFORM

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelala samostojno pod mentorstvom izr. prof. dr. Janeza Demšarja,
- so elektronska oblika diplomskega dela, naslov (slo., ang.), povzetek (slo., ang.) ter ključne besede (slo., ang.) identični s tiskano obliko diplomskega dela,
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 30. 8. 2013 Podpis avtorja:

Zahvala

Zahvaljujem se:

- mentorju za lep odnos, zaupanje in hitre popravke,
- družini, ki mi je omogočila super pogoje za študij,
- Simonu za lektoriranje,

ter vsem ostalim, ki so me spremljali in spodbujali pri nastanku tega dela. Svet je preveč kompleksen, da bi lahko v kratki zahvali (oz. sploh) razpentljala mrežo vzrokov, ki so pripeljali do tega, da danes tu sedim in se trudim oblikovati zahvalo. Dragi bralec, če se ti zdi, da si tudi ti prispeval k nastanku tega dela, si prisrčno vabljen, da čutiš mojo hvaležnost.

Kazalo

Povzetek

Orange je odprtokoden programski paket z interaktivnim konzolnim in grafičnim uporabniškim vmesnikom namenjen podatkovnem rudarjenju. Vsebuje mnogo naprednih metod strojnega učenja in grafičnih gradnikov, a pogreša osnovne statistične gradnike. Cilj diplomskega dela je bil zapolniti to vrzel, zato se je razvilo štiri uporabniku prijazne statistične gradnike za Orange. Vsak vsebuje interaktivni uporabniški vmesnik, potrebne statistične teste in vizualizacije podatkov. Z dodanimi gradniki je mogoče intuitivno in enostavno opravljati statistične analize, jih povezovati z ostalimi gradniki in shranjevati grafikone. Razviti so bili v programskem jeziku *Python 2* s pomočjo knjižnic *SciPy* in *NumPy*.

Ključne besede:

statistika, testi značilnosti, gradniki, uporabniški vmesnik, vizualizacije podatkov, grafikoni

Abstract

Orange is an open source data mining tool with command-line and graphical user interface. It offers plenty of machine learning algorithms and visualization, but almost none of the established statistical data analysis methods. The goal of this work was to fill that gap and develop four graphical widgets for Orange. Each widget contains interactive user interface, required statistical tests and visualizations of data. With added widgets it is easier and more intuitive to run statistical analysis, store the majority of charts and connect with other widgets. The widgets were developed in the programming language *Python 2* with the help of the libraries *SciPy* and *NumPy*.

Key words:

statistics, hypothesis tests, widgets, user interface, data visualization, plotting

Poglavje 1

Uvod

Računalniška obdelava podatkov v modernem svetu igra vedno večjo vlogo. Od skromnih začetkov v sredini prejšnjega stoletja je pod nadzor računalniskih sistemov prehajalo vedno več podatkov in poleg shranjevanja podatkov so računalniki časoma prevzeli tudi njihovo obdelavo in analizo.

Osnovne metode statistike so bile že od nekdaj jedro analize merskih podatkov v znanosti, financah in panogah uporabne matematike. S povečevanjem količine podatkov, predvsem po razcvetu svetovnega spleta, pa se je pojavila potreba po novih algoritmih, ki omogočajo obdelavo nestrukturiranih podatkov ter izluščenje dejstev iz sicer nepreglednih in neurejenih naborov informacij. Razvoj je pripeljal do metod umetne inteligence, strojnega učenja in rudarjenja podatkov, ki dandanes stojijo za velikimi internetnimi korporacijami in pripomorejo k učinkovitemu zbiranju podatkov in oglaševanju. Poleg tega tovrstne metode uporabljajo znanstveniki pri raziskovalnem delu. Razvoj algoritmov in prilagoditev na reševanje konkretnega problema na najrazličnejših področjih pa še vedno zahteva človeško posredovanje ter hiter in pregleden vmesnik za nadzor toka podatkov. To programsko nišo zasedajo splošni programi, kot na primer SPSS, R, Matlab, Root, Mathematica in Sage, ki ponujajo tudi možnosti interaktivne manipluacije podatkov. Za izbrane probleme obstajajo različni programski paketi, ki so prilagojeni za kar se da enostavno uporabo na danem področju. Razvoj enega izmed tovrstnih paketov – Orange – bo osnova za delo, predstavljeno v tem diplomskem delu.

Orange je uporabniku prijazen odprtokoden programski paket z interaktivnim konzolnim in grafičnim uporabniškim vmesnikom, že poln mnogih naprednih metod in gradnikov (ang. *widgets*) za obdelavo podatkov. Dobljene podatke lahko s statistiko organiziramo in povzamemo informacije ter jih poskušamo posplošiti preko dejansko zajetih podatkov. To v Orange težje

naredimo, dokler mu manjkajo statistični gradniki.

Za cilj diplomske naloge smo si zadali dopolniti Orange z naborom osnovnih statističnih grafičnih gradnikov: Primerjava povprečij oz. median (ang. *Comparison of means/medians*) (4.1), Neodvisnost diskretnih spremenljivk (ang. *Independence for discrete attributes*) (4.2), Korelacija zveznih spremenljivk (ang. *Correlation between continuous variables*) (4.3), Primerjava porazdelitev (ang. *Comparison of distributions*) (4.4). Z dodanimi gradniki, ki so napisani v duhu Oranga, lahko intuitivno in enostavno opravljamo osnovne statistične analize, shranjujemo večino vizualiziranih grafikonov ter jih povezujemo z ostalimi gradniki.

Poglavje 2

Orange

Orange, Data Mining Fruitful & Fun [1, 8] je modularen odprtokoden programski paket za podatkovno rudarjenje in strojno učenje. Njegova odprtokodna skupnost se še vedno povečuje, Orange pa se stalno dopolnjuje in posodablja. Razvili so ga v Laboratoriju za bioinformatiko na Fakulteti za računalništvo in informatiko, UL.

Orange ima skriptni uporabniški vmesnik v Pythonu in vizualni uporabniški vmesnik Orange Canvas. Prvi omogoča direktni dostop za hitro programiranje novih algorimov in razvijanje postopkov za kompleksno analizo podatkov, a se bomo v diplomskem delu osredotočili na drugega. Orange Canvas je zasnovan na gradnikih, razvrščenih po zavihkih glede na kategorijo, ki ji pripadajo. Taka zasnova omogoča preprosto vizualno programiranje.

Orange je opremljen s številnimi funkcijami za podatkovne analitike (na primer za obliko podatkov, vizualizacijo, klasifikacijo, itd.). Poleg osnovnih orodij so na voljo dodatki za bioinformatiko in tekstovno rudarjenje. Razvili so ga v programskih jezikih C++ in Python in je na voljo na operacijskih sistemih Linux, Microsoft Windows in OS X. Aktivno ga uporabljajo pri raziskovanju, napovedovanju vremena, v medicini in genetiki.

Vizualno programiranje (Slika 2.1) je realizirano na beli podlagi, kjer gradnike lahko skoraj poljubno povezujemo med seboj. Orange si nastavitve zapomni in predlaga najbolj uporabljene kombinacije ter sam izbere komunikacijske kanale med gradniki. Da si obliko podatkov in rezultatov lažje predstavljamo, lahko vizualizacijo z grafikoni po želji izberemo med diagrami razpršenosti, dendogrami, drevesi, mrežami, toplotnimi kartami in drugimi načini prikaza. Dejanja se neopazno izvajajo skozi shemo podatkovne analize. Z združevanjem različnih gradnikov lahko oblikujemo željeno delovno okolje, in če spreminjamo izbrane nastavitve nekega gradnika, se to odraža na celotni verigi.

Tako je analiza podatkov interaktivna. Kljub temu, da imamo na izbiro več kot 100 gradnikov, se izbira še povečuje, ker je Orange **razširljiv**: po potrebi si lahko razvijemo svoje gradnike.

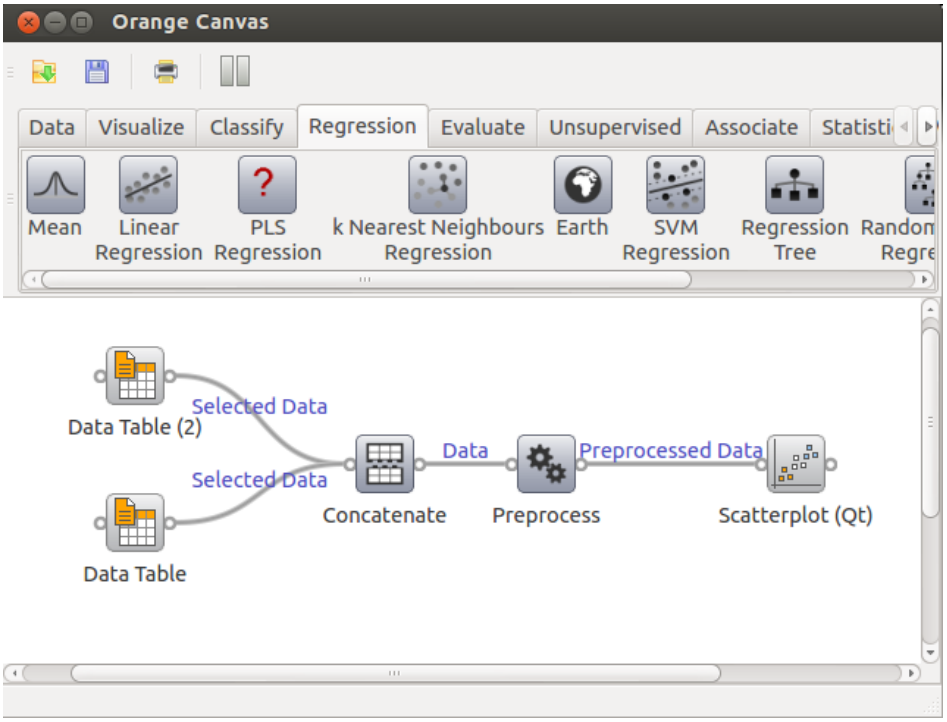
Pri vsaki analizi je v Orange potrebno najprej naložiti vhodno datoteko. To storimo tako, da na belo površino potegnemo gradnik *File*, ga dvokliknemo in izberemo željeno datoteko, ki mora biti formata *"C4.5"*, *"Assistant"*, *"Retis"* ali ‘‘*tab-delimited*" (Orangeov format). Tukaj je primer *"tab-delimited"* formata:

age	**prescription**	**astigmatic**	**tear_rate**	**lenses**
discrete	**discrete**	**discrete**	**discrete**	**discrete**
				class
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none

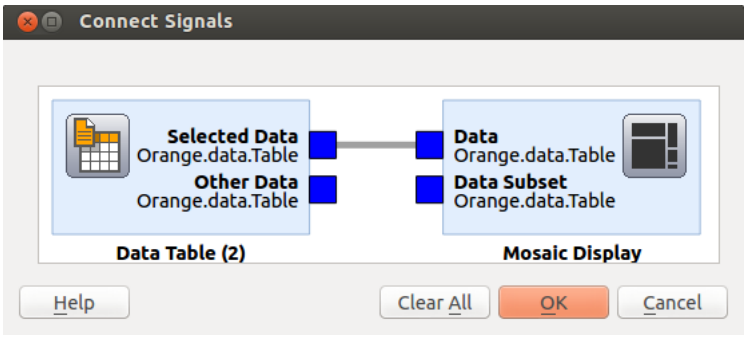
Tabela 2.1: *Primer "tab-delimited" formata. Prva vrstica pove ime atributa, druga pa njegov tip. V tretji vrstici lahko določimo kateri atribut je razred.*

Tipični gradniki (Slika 2.3) so sestavljeni iz levega **kontrolnega območja** (ang. *control area*), kjer se nahajajo radio gumbi (ang. *radio buttons*), potrditvena polja (ang. *check box*), spustna polja (ang. *combo box*), seznamska polja (ang. *list box*), vrtilno polje (ang. *spin box*) in ostale kontrolne komponente. Po njih lahko uporabnik klika in s tem izbira nastavitve vizualizacije na desnem **glavnem območju** (ang. *main area*) gradnika.

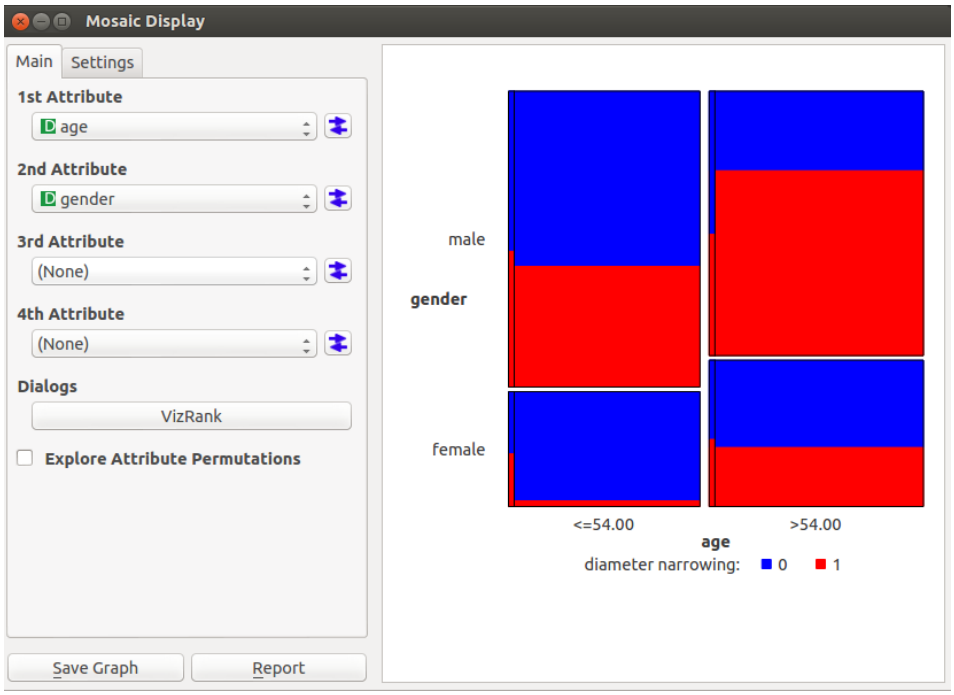
Gradnike med seboj povezujemo tako, da potegnemo vijugo med gradniki. Orange že sam določi glavni izhodni signal, a nekateri gradniki lahko sprejemajo in oddajajo več različnih signalov. Če bi radi spremenili izbiro, dvokliknemo na povezavo in izberemo željeni signal (Slika 2.2).

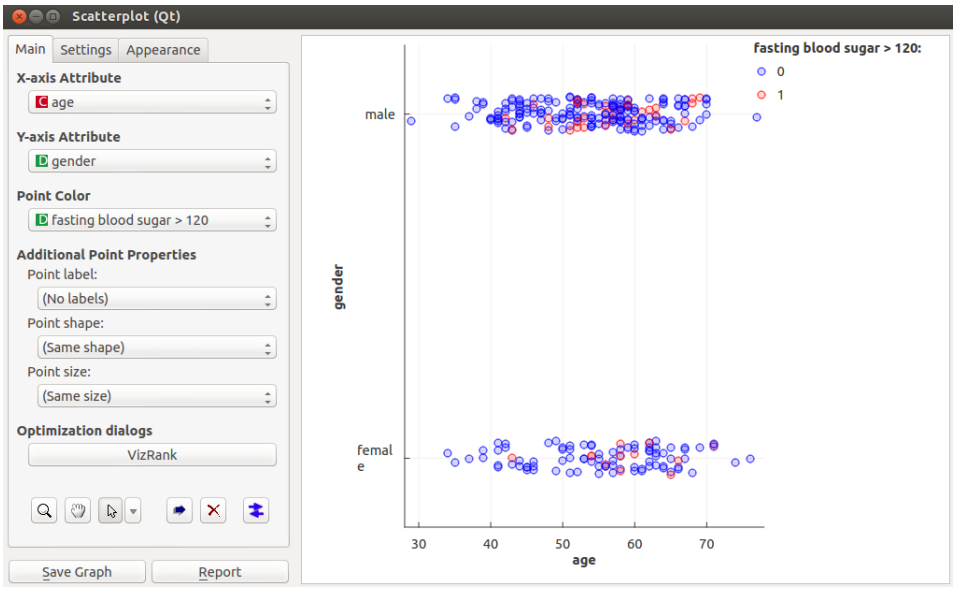


Slika 2.1: *Osnovno okno programa Orange, kjer gradnike najdemo zgoraj v shrambi in jih na beli površini poljubno povezujemo med seboj.*



Slika 2.2: *Prikaz povezovanja signalov v Orange. Povezana sta gradnika Data Table in Mosaic Display. Iz Data Table gre Selected Data v Data Mosaic Display.*





Slika 2.3: *Primera tipičnih gradnikov, ki imata na levi strani kontrolno območje in na desni vizualni prikaz rezultatov.*

2.1 Programiranje gradnikov

Za programiranje gradnikov je potrebno najprej vzpostaviti delovno okolje in se držati določenih pravil pri pisanju. Te posebnosti bomo predstavili v tem poglavju.

Uporaba repozitorija

Za programiranje gradnikov smo najprej vzpostavili okolje. Koda programa Orange se nahaja na spletnem repozitoriju Bitbucket, ki je tipa Mercurial. Večinoma je napisan v programskem jeziku Python in vključuje prevedeno *diff* implementacijo, napisano v C. Mercurial ima mnogo nastavitev (Slika 2.4) in podporo za Windows, Unix bazirane sisteme in Linux. Program je v osnovi ukazen, vendar obstajajo različice z grafičnim uporabniškim vmesnikom. Z repozitorijem komuniciramo preko programa *hg*, ki je poimenovan po kemijskem simbolu za živo srebro (ang. *mercury*).

Pred prevajanjem smo ustvarili virtualno okolje, da se preostali del sistema ne bi okrnil s programi, ki jih Orange potrebuje za svoje delovanje. Nato smo na repozitoriju naredili svojo vejo in si prenesli Orange na delovni računalnik ter prevedli v razvojno različico:

```
hg clone https://bitbucket.org/biolab/orange
python2 setup.py develop
```

Za delo z repozitorijem smo največ uporabljali ukaze *push* (potisne spremembe iz lokalnega repozitroija v glavni repozitorij) in *commit* (spravi spremembe iz delovnega direktorija v lokalni repozitorij).

Pisanje gradnikov

Programiranje gradnikov je dokaj enostavno z okoljem, ki poskrbi za zadeve, povezane z grafičnim uporabniškim vmesnikom. Najprej ustvarimo projekt, oz. datoteko (`NoviGradnik.py`), v kateri bomo pisali gradnik:

```
/orange/Orange/OrangeWidgets/
      setup.py
      kategorija/
                  __init__.py
                  NoviGradnik.py
```

Orange Canvas išče gradnike v mapi `/orange/Orange/OrangeWidgets`. V OrangeWidgets se nahajajo podmape, ki se imenujejo enako kot kategorije gradnikov. V njih je koda mnogih gradnikov. Vsak je predstavljen s svojo datoteko, ki se začne z vrsticami:

```
"""
<name>Primerjava porazdelitev</name>
<description>Gradnik, ki primerja porazdelitve.</description>
<icon>icons/ikona.svg</icon>
<priority>200</priority>
"""
```

Informacije o gradniku se nahajajo v glavi datoteke v obliki komentarja med trojnimi narekovaji. Ime, ki ga določimo med značkama *name*, postane ime gradnika v Orange Canvasu. Opis gradnika je med značkama *description* in se prikaže, ko z miško zapeljemo preko gradnikove ikone. Pot do ikone se poda med značkami *icon*. Prioriteta pove, kateri po vrsti se gradnik prikaže v njegovi kategoriji. Poleg tega je za delo potrebno vključiti knjižnice Orange, OWWidget in OWGUI. Potrebno je paziti, da se ime razreda ujema z imenom datoteke.

Gradniki medsebojno komunicirajo (Slika 2.2) zato vsakemu gradniku določimo vhodne in izhodne signale. V primeru spodaj gradnik sprejme in vrne podatke tipa `ExampleTable`.

```
self.inputs = [("Test Data Set", ExampleTable, self.data)]
self.outputs = [("Results", ExampleTable, self.ldata)]
```

Kadar želimo poslati obdelane podatke iz gradnika, to storimo s funkcijo `self.send()`, v kateri naprej z nizom poimenujemo izhod, nato dodamo še objekt, ki ga pošiljamo:

```
def commit(self):
    self.send("Sampled Data", self.sample)
```

Ker je vnovično zagananje celotnega Orange Canvasa zavoljo testiranja sprememb v delovanju gradnika zamudno, želimo omogočiti poganjanje samo izbranega gradnika. Ponavadi na dnu datoteke gradnika napišemo:

```
if __name__=="__main__":
    appl = QApplication(sys.argv)
    ow = OWDataSamplerA()
    ow.show()
    dataset = Orange.data.Table('iris.tab')
    ow.data(dataset)
    appl.exec_()
```

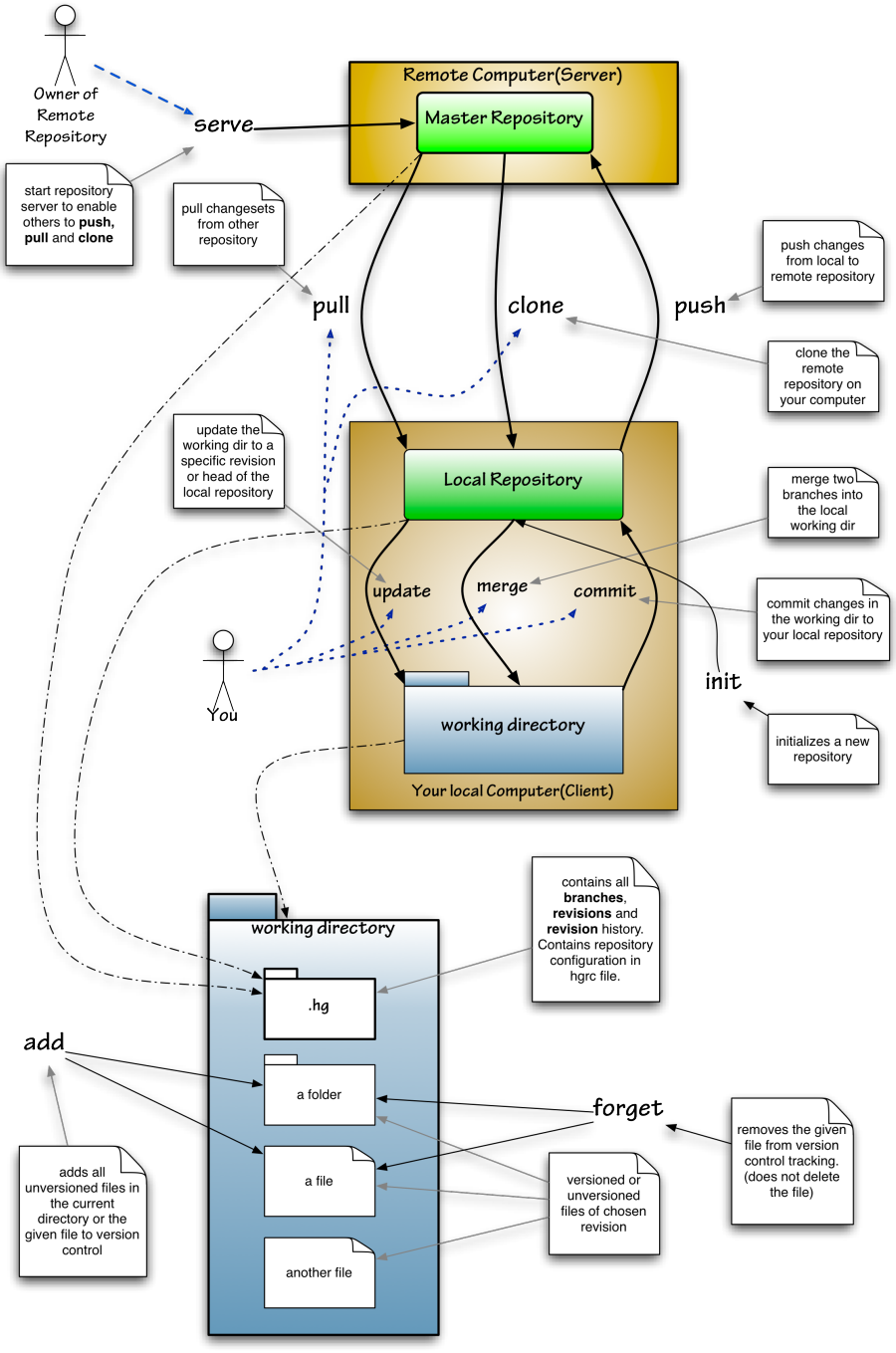
Pomembna je tudi metoda `data()`, ki jo napišemo v gradniku. Z njo definiramo, kako bomo obdelali vhodne podatke. Omenili smo jo tudi pri definiciji `self.inputs`.

Nastavitve gradnika podamo v `settings_list`, takoj na začetku razreda. Nato jih uporabljamo za stanje kontrolnega dela gradnika. Če pokličemo `loadSettings()`, prepišemo prirejene nastavitve, v primeru, da so se v prejšnji uporabi gradnika spremenile.

```
class OWDataSamplerB(OWWidget):
    settingsList = ['proportion', 'commitOnChange']
```

To so osnovne nastavitve, ki jih je potrebno pred delom upoštevati. Za postavljanje objektov na gradnik ponavadi definiramo tip objekta, njegov položaj in območje. Tovrstnih objekov je dosti: gumbi, radio gumbi, oznake,

potrditvena polja, itd. Točna navodila za njihovo uporabo najdemo v uradni dokumentaciji Orangea. Za izrisovanje grafov smo uporabili knjižico *orangeqt*. Nekateri objekti niso opisani v dokumentaciji, zato je bil potreben sprehod skozi njihovo programsko kodo. Ko smo imeli zgoraj opisane parametre nastavljene, smo začeli programirati.



Slika 2.4: *Nekatere pomembne Mercurialove operacije in njihove relacije. Vir [9].*

Poglavje 3

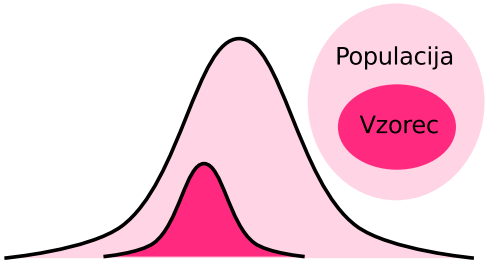
Statistika

3.1 Osnovni pojmi

V tem poglavju bomo predstavili nekaj osnovnih pojmov [6, 2, 4], ki se bodo v naslednjih poglavjih večkrat pojavljali.

Populacija je končna ali neskončna množica elementov, ki jim lahko opazujemo ali merimo neko merljivo količino. Če pogledamo matematično, je populacija neprazna množica G, na kateri je definirana merljiva funkcija X. Vsaka meritev da vrednost $X(e)$ te funkcije pri izbranem elementu $e \in G$. Predpostavimo, da elemente, na katerih merimo X, izbiramo slučajno. Tedaj imamo *slučajno spremenljivko* X na verjetnostnem prostoru G.

Vzorec je majhen del populacije, iz katerega poskušamo veljavno sklepati, kakšne so razmere na celotni populaciji. Da je vzorec reprezentativen, mora biti izbran nepristransko in mora biti dovolj velik. Matematično gledano je slučajni vzorec slučajni vektor $(X_1, X_2, ..., X_n)$. *Enostaven slučajni vzorec* je tisti, pri katerem predpostavimo, da imajo vse spremenljivke $X_1, X_2, ..., X_n$ enako porazdelitev, in sicer isto kot spremenljivka X, ter da so spremenljivke X_i med seboj neodvisne.



Slika 3.1: *Populacija in vzorec*

Statistična hipoteza je domneva o porazdelitvi slučajne spremenljivke X na populaciji. Ko začnemo s testiranjem hipoteze, najprej določimo *ničelno hipotezo* H_0, ki je enostavna domneva (natanko določa tip in točno vrednost parametra), ponavadi v nasprotju s tem, kar želimo dokazati. Poleg nje postavimo tudi *alternativno hipotezo* H_1, ki je vsaka dopustna hipoteza v nasprotju s H_0. *Parameter* je numerična značilnost, ki jo ima statistična spremenljivka na populaciji. Če poznamo tip porazdelitve in raziskujemo domnevo na nekem parametru, je hipoteza *parametrična*, sicer je *neparametrična*. Hipoteza je enostavna, če natančno določa tip in točno vrednost parametra, drugače je sestavljena.

Primer 3.1.1. Označimo povprečje z μ, ter disperzijo s σ. Hipoteza $H : \mu = 0$ je enostavna, ko poznamo porazdelitev X (normalna $N(\mu, \sigma)$) in njeno σ. Kadar σ ne poznamo, ali postavimo hipotezo $H : \mu > 0$, je hipoteza sestavljena.

Parametri

Parameter je številska značilnost statistične spremenljivke na populaciji. Najpomembnejši populacijski parametri so mere centralne tendence (povprečje, mediana, modus), mere razpršenosti, korelacijski koeficienti, itd. V diplomski nalogi bomo omenjali predvsem naslednje parametre:

Populacijsko povprečje $\mu = E(X)$ je pričakovana vrednost spremenljivke X na populaciji. *Vzorčno povprečje* $\overline{X}$ je cenilka za populacijsko povprečje in ga izračunamo po obrazcu:

$$\overline{X} = \frac{x_1 + x_2, ... + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.1}$$

kjer so $x_1, x_2, ..., x_n$ vzorčni podatki. Vzorčno povprečje ni robustna mera, ker imajo na povprečje velik vpliv ekstremne vrednosti.

Varianca ali *disperzija* je mera, ki opisuje statistično razpršenost. Povprečni odklon od povprečja ni uporaben, ker se pozitivni in negativni odkloni izničijo. Smiselna vrednost je varianca, definirana z $\sigma^2 = E((X - E(X))^2)$. Varianca je torej povprečen kvadriran odklon od populacijskega povprečja. *Vzorčna varianca* je cenilka za populacijsko varianco in jo izračunamo po obrazcu:

$$S = \frac{1}{n-1} \sum_{i} (x_i - \overline{x})^2 \tag{3.2}$$

.

Mediana je srednja vrednost zaporedja števil. Zaporedje razdeli na dve enaki polovici po številu elementov. Je bolj robustna od vzorčnega povprečja, saj podatki, ki ekstremno odstopajo, manj vplivajo na njeno vrednost. Pri zaporedju $(x_1, x_2, \ldots, x_n)$ z n elementi, ki je razvrščeno po velikosti, se izračuna po predpisu:

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & n = \text{liho} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & n = \text{sodo} \end{cases} \tag{3.3}$$

Kvartili so vrednosti, ki delijo urejeno zaporedje na štiri enake dele.

- Prvi kvartil (Q_1) je spodnji kvartil. Je vrednost, od katere je 25% vrednosti slučajne spremenljivke manjših in 75% večjih.
- Drugi kvartil (Q_2) je mediana.
- Tretji kvartil (Q_3) imenujemo zgornji kvartil. Je vrednost, od katere je 75% vrednosti slučajne spremenljivke manjših in 25% večjih.

Standardni odklon je mera, s katero merimo razpršenost vrednosti iz populacije. Označen je s σ, je koren variance in njegova vzorčna cenilka je:

$$\sigma = \sqrt{S} \tag{3.4}$$

Nizek standardni odklon pomeni, da so vrednosti precej koncentrirane okrog povprečja, medtem ko visok σ opisuje razpršenost statističnih enot preko velikega intervala vrednosti. Za razliko od variance je izražena v enakih enotah kot spremenljivka X.

Korelacijski koeficient je številska mera, ki opisuje linearno povezanost slučajnih spremenljivk X in Y. Definiran je kot:

$$\rho = \frac{K(X,Y)}{\sigma_X \sigma_Y} \tag{3.5}$$

kjer je $K(X,Y) = E(XY) - E(X)E(Y)$ *kovarianca* slučajnih spremenljivk. Kovarianca je število, ki meri, koliko sta dve naključni spremenljivki povezani. Varianca je poseben primer kovariance. σ_X in σ_Y sta standardna odklona spremenljivk X in Y. Korelacijski koeficient lahko zavzema vrednosti iz intervala $[-1, 1]$. $\rho = \pm 1$ pomeni popolno linearno povezanost spremenljivk X in Y. Če je $\rho = 0$ sta, spremenljivki X in Y nekorelirani.

Primer 3.1.2. Imamo množico vrednosti: 26, 45, 32, 54, 23, 42. Želimo izračunati vzorčno povprečje, varianco, standardni odklon in kvartile.

- vzorčno povprečje: $\overline{X} = (26 + 45 + 32 + 54 + 23 + 42)/6 = 37$
- varianca:
  $\sigma^2 = \frac{1}{5}((26-37)^2 + (45-37)^2 + (32-37)^2 + (54-37)^2 + (23-37)^2 + (42-37)^2) = 144$
- standardni odklon: $\sigma = \sqrt{144} = 12$
- kvartili: najprej uredimo zaporedje 23, 26, 32, 42, 45, 54
  - $Q_1 = 27.5$
  - $Q_2 = (42 + 32) \cdot 0.5 = 37$
  - $Q_3 = 44.25$

Neodvisnost in nekoreliranost

Slučajne spremenljivke $X_1, X_2, \ldots, X_n$ so med seboj *neodvisne*, če pri poljubnih vrednostih $x_1, x_2, \ldots, x_n \in \mathbb{R}$ velja

$$F(x_1, x_2, \ldots, x_n) = F_1(x_1)F_2(x_2)\cdots F_n(x_n). \tag{3.6}$$

Tukaj je F porazdelitvena funkcija slučajnega vektorja $(X_1, X_2, \ldots, X_n)$, $F_1, F_2, \ldots F_n$ pa so porazdelitvene funkcije posameznih komponent tega vektorja. Če zgornja enačba ne velja, pravimo, da so spremenljivke *odvisne*. Slučajni spremenljivki X in Y sta *nekorelirani*, ko velja

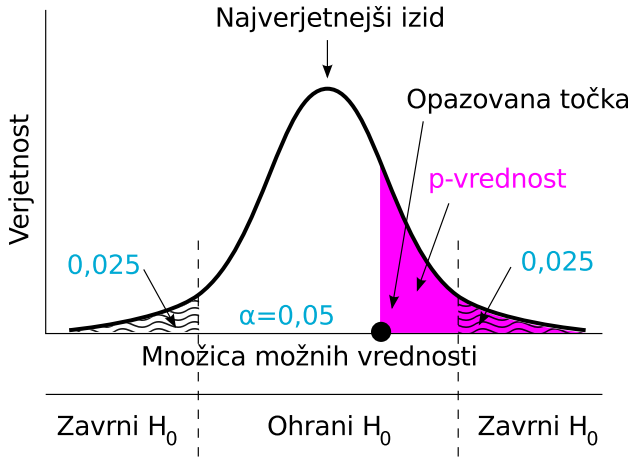
$$E(XY) = E(X)E(Y) \qquad \text{oziroma} \qquad K(X, Y) = 0. \tag{3.7}$$

V obratnem primeru pravimo, da sta *korelirani*. Iz korelacijskega koeficienta sledi, da sta neodvisni slučajni spremeljivki vedno nekorelirani. Nasprotno ne velja.

3.2 Testiranje statističnih hipotez

V statistiki pogosto želimo preverjati predpostavke o vrednostih določenih parametrov populacije na osnovi podatkov, dobljenih na vzorcu.

Test značilnosti [6, 7, 5] je postopek, po katerem na podlagi vzorca ničelno hipotezo H_0 zavrnemo ali o njej ne odločimo. Testiramo H_0 proti H_1 glede na stopnjo značilnosti testa α. Poznamo parametrične in neparametrične teste. Test je *neparametričen*, ko je porazdelitev neznana in je hipoteza o vrsti porazdelitve. Pri *parametričnem* testu je znan tip porazdelitve, hipoteza se pa sklicuje na parameter porazdelitve. Glede na to, ali je kritično območje sestavljeno iz dveh polodprtih intervalov ali enega, ločimo *dvostranske* in *enostranske* parametrične teste.



Slika 3.2: *Primer za dvostrani statistični test. Valovite površine zavzemajo 5% površine pod krivuljo. Obarvana ploščina je enaka p-vrednosti.*

Čeprav je ničelna hipoteza H_0 jedro statističnega testa, je ponavadi druga skrb raziskovalca. Razloga za to sta:

- Raziskovalna hipoteza H_1 je ponavadi preveč kompleksna, da bi se jo testiralo direktno.
- Ker je raziskovalna hipoteza enaka alternativni hipotezi, bo zavrnitev ničelne hipoteze močno podprla raziskovalno hipotezo. Odločitev, da ničelno hipotezo zavržemo, je močnejša od odločitve, da jo obdržimo.

Primer 3.2.1. Gospa, ki je okušala čaj, je znan primer testiranja statistične hipoteze. Fisherjeva kolegica Dr. Muriel Bristol je trdila, da lahko ugotovi, ali je bil v skodelico najprej dodan čaj ali mleko. Fisher ji ni verjel na besedo, zato ji je dal v naključnem zaporedju poskusiti 8 različno polnjenih skodelic. Njena naloga je bila, da uspešno izbere 4 skodelice določenega postopka. Ničelna hipoteza je bila, da gospa nima take sposobnosti. Test je enostavno štel število uspešno izbranih 4 skodelic. Kritična meja je bil edini primer 4 uspehov od 4 možnosti uspeha. Na osnovi kriterija ($< 5\%; 1$ od 70 permutacij $\approx 1.4\%$) je Fisher določil, da alternativna hipoteza ni bila potrebna. Gospa je identificirala vse skodelice pravilno, kar se šteje za statistično signifikanten rezultat. [7]

3.2.1 Postopek testiranja značilnosti

1. Imamo hipotezo, za katero ne vemo, če je resnična.

2. Najprej določimo relevantno ničelno in alternativno hipotezo. Ničelna hipoteza mora biti izbrana tako, da lahko po testu zaključimo ali je alternativna hipoteza sprejeta ali ostane neodločeno. Pri postavljanju hipotez moramo upoštevati dvostranskost ali enostranskost testa.

3. Glede na velikost vzorca in obliko problema se odločimo, kateri test značilnosti T je primeren.

4. Izberemo stopnjo značilnosti α, pod katero bo ničelna hipoteza zavrnjena ali neodločena. (Pogosti izbiri sta 5% in 1%).

5. Glede na parameter α in porazdelitev testne statistike T določimo kritično območje testa ω_0. Pri tem velja:

$$P(T \in \omega_0 \mid H_0) \leq \alpha$$

6. S testom značilnosti izračunamo eksperimentalno vrednost t_e. Lahko izračunamo tudi p-vrednost.

7. Glede na t_e ali p-vrednost se odločimo ali ničelno hipotezo zavrnemo ali ne. Če $t_e \in \omega_0$, ničelno hipotezo H_0 zavrnemo, sicer jo obdržimo. Podobno, če je p vrednost manjša od α, ničelno hipotezo H_0 zavrnemo.

3.2.2 Sprejemanje odločitev

Čeprav sledimo testu, obstaja verjetnost, da ne sprejmemo pravilne odločitve. Poleg pravilnih odločitev poznamo napake tipa I in tipa II:

Odločitev	Pravilna H_0	Napačna H_0
Obdrži H_0	Pravilna odločitev	Napaka tipa II
Zavrni H_0	Napaka tipa I	Pravilna odločitev

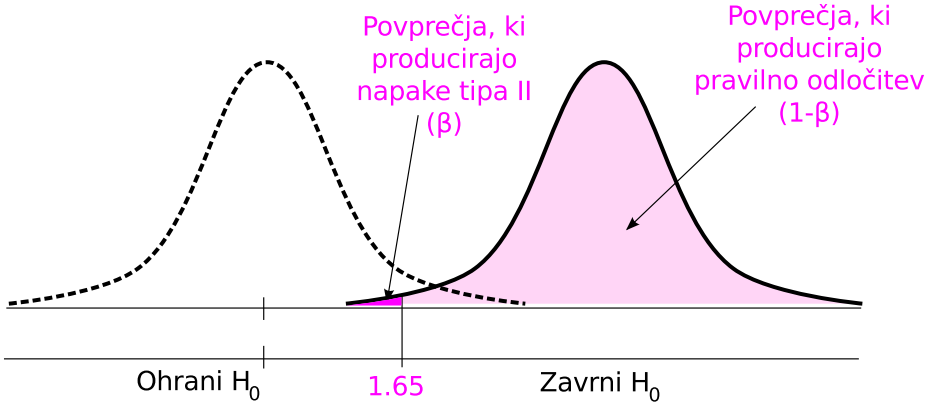
Pri dobrem testu morata biti verjetnosti za obe napaki čim manjši. Verjetnost napake tipa II se ponavadi poveča, kadar želimo zmanjšati napako tipa I in obratno.

Napaka tipa I

Napaka tipa I se zgodi, če H_0 zavrnemo, kljub temu da je pravilna. To se zgodi, ko imamo smolo in je povprečje naših izbranih podatkov ravno v območju za zavrnitev H_0, ki je označeno z ω_0. Verjetnost tega tipa napake je enaka stopnji značilnosti α.

Napaka tipa II

Zgodi se, da H_0 ne zavržemo, čeprav je napačna. Temu primeru pravimo napaka tipa II, katere verjetnost označimo z β. Pomembno vlogo igra oddaljenost prave porazdelitve od predvidene. Če sta porazdelitvi bolj oddaljeni, je β majhen in je verjetnost za pravilno odločitev velika. Velja tudi obratno: če sta porazdelitvi zelo blizu, je verjetnost za pravilno odločitev manjša.



Slika 3.3: *Hipotetična in prava distribucija ter področje napake tipa II (β)*

3.2.3 Studentov t-test dveh neodvisnih vzorcev

Poznamo več različnih Studentovih t-testov. Tukaj bomo opisali različice testa, uporabljene v gradniku.

Studentov t-test dveh neodvisnih vzorcev je parametričen test značilnosti, kjer *testna statistika* T sledi *Studentovi t porazdelitvi*. Najpogosteje se uporablja za primerjavo povprečij dveh neodvisnih vzorcev populacij. V gradniku smo uporabili dvostransko različico (ničelna hipoteza $H_0 : \mu_1 = \mu_2$). Za t-test zahtevamo normalno porazdelitev vzorčnih povprečij.

Enake velikosti vzorcev

Če sta velikosti vzorcev enaki in lahko predpostavimo tudi, da sta varianci vzorcev enaki, lahko uporabimo t-test.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}, \tag{3.8}$$

kjer je

$$S_{X_1X_2} = \sqrt{\frac{1}{2}(S^2_{X_1} + S^2_{X_2})}. \tag{3.9}$$

S_{X_1} in S_{X_2} sta standardna odklona slučajnih spremenljivk X_1 in X_2. Število prostostnih stopenj za ta test je $2n - 2$, kjer je n število udeležencev v vsakem vzorcu.

Različni velikosti vzorcev

Če želimo uporabiti t-test za različne velikosti vzorcev, ni treba privzeti, da sta varianci vzorcev enaki. Vrednost t statistike izračunamo po obrazcu:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{3.10}$$

kjer je

$$S_{X_1X_2} = \sqrt{\frac{(n_1 - 1)S^2_{X_1} + (n_2 - 1)S^2_{X_2}}{n_1 + n_2 - 2}}. \tag{3.11}$$

S_{X_1} in S_{X_2} sta standardna odklona slučajnih spremenljivk X_1 in X_2. Število prostostnih stopenj za ta test je $n_1 + n_2 - 2$, kjer je n_1 število udeležencev prvem vzorcu in n_2 v drugem vzorcu.

Sprejemanje odločitev:

Ker smo uporabili dvostransko različico, je pri testni statistiki $t \geq t_{1-\alpha/2}(df)$ število prostostnih stopenj odvisno od izbire t-testa. Opisali smo metodi, pri katerih lahko prizamemo enaki varianci vzorcev. Pri različnih variancah vzorcev lahko uporabimo *Welchov t-test*.

3.2.4 Mann Whitneyev U-test za dva neodvisna vzorca

Mann Whitney U-test za dva neodvisna vzorca je neparametričen test, ki testira ničelno hipotezo, da sta dve porazdelitvi enaki. Vzorca sta neodvisna in lahko tudi različnih velikosti.

Hipoteze:

$$H_0 : \text{porazdelitev distribucije 1} = \text{porazdelitev distribucije 2} \tag{3.12}$$
$$H_1 : \text{porazdelitev distribucije 1} \neq \text{porazdelitev distribucije 2} \tag{3.13}$$

Test:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \tag{3.14}$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2, \tag{3.15}$$
$$U = \min(U_1, U_2), \tag{3.16}$$

kjer je R vsota rangov vzorca, ter n_i velikost i-tega vzorca.

Sprejemanje odločitev:

Ničelno hipotezo H_0 zavržemo, če je testna statistika U manjša ali enaka kritični vrednosti U, ki jo razberemo iz tabele. Vrnjena p-vrednost metode velja za enostransko hipotezo. Če želimo rezultat za dvostransko hipotezo, vrnjeno p-vrednost pomnožimo z 2.

3.2.5 ANOVA

ANOVA je test, ki se uporablja za analizo variance. Kadar opravljamo enak poizkus pri istih pogojih, opažamo v rezultatu poskusa *slučajna odstopanja*. Z analizo variance ugotavljamo ali so ta odstopanja naključna ali ne. Poznamo več vrst ANOVAe. V gradniku smo uporabili *enojno klasifikacijo*, ki testira hipotezo, da imata dve skupini ali več enako populacijsko povprečje. Skupine so lahko različnih velikosti.

Vzorec velikosti n slučajno razdelimo v r skupin. Z vsako skupino naredimo poskus pri drugi vrednosti edinega parametra, ki ga imamo. Rezultate v prvi skupini $x_{11}, x_{12}, ..., x_{1n}$ imenujemo za vrednosti prve spremenljivke X_1, rezultate v drugi skupini $x_{21}, x_{22}, ..., x_{2n}$ pa imejmo za vrednosti druge spremenljivke X_2 in tako dalje do r.

Test ANOVA ima določene predpostavke: Spremenljivke $X_1, X_2, ..., X_n$ so med seboj nedovisne in normalno porazdeljene z enako razpršenostjo σ^2, ki je ne poznamo. S testom ANOVA testiramo hipotezo $H_0 : \mu_0 = \mu_1 = ... = \mu_k$ proti alternativni hipotezi, da je vsaj eno grupno povprečje različno. Izračunajmo:

$$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \quad n = \sum_{i=1}^{k} n_i, \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^{k}\frac{n_i}{n}\overline{X}_i$$

$$S_B^2 = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2, \quad S_W^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2. \tag{3.17}$$

Kadar ničelna hipoteza H_0 velja, sta S_B^2 in S_W^2 neodvisni ter $S_B^2/\sigma^2 \sim \chi^2(k-1)$ in $S_B^2/\sigma^2 \sim \chi^2(k-1)$ in $S_W^2/\sigma^2 \sim \chi^2(n-k)$, zato velja:

$$F = \frac{S_B^2/(k-1)}{S_W^2/(n-k)} \sim F(k-1, n-1),$$

kjer je $F(k-1, n-1)$ Snedecorjeva porazdelitev. $(k-1, n-1)$ je število prostostnih stopenj pri F-testu.

Sprejemanje odločitev:

Ko izračunamo testno statistiko F, sprejmemo odločitev glede na različico testa:

- Dvostranska: H_0 zavrnemo, ko $F \leq F_{\alpha/2}(df_1, df_2)$ ali $F \geq F_{1-\alpha/2}(df_1, df_2)$,
- Enostranska v desno: H_0 zavrnemo ko $F \geq F_{1-\alpha}(df_1, df_2)$,
- Enostranska v levo: H_0 zavrnemo ko $F \leq F_{\alpha}(df_1, df_2)$,

kjer $F_p(df_1, df_2)$ pomeni kvantil Snedecorjeve porazdelitve z (df_1, df_2) prostostnimi stopnjami za verjetnost p.

3.2.6 Kruskal-Wallisov H test

Kruskal-Wallis H je neparametrična verzija ANOVA in se uporablja, kadar predpostavke za ANOVA niso izpolnjene. Testira se, ali vzorci izhajajo iz enake porazdelitve, torej hipotezo $H_0 : \mu_0 = \mu_1 = ... = \mu_k$ proti alternativni hipotezi, da je vsaj eno grupno povprečje različno. Uporablja se za primerjavo več kot dveh neodvisnih vzorcev. Test ne zazna, katera skupina se razlikuje od ostalih. Opisali bomo test z enojno klasifikacijo, ki ga uporabljamo v gradniku *Primerjava povprečij oz. median.*

Vzorec razdelimo na g skupin. Število vseh opazovanj označimo z N. Nato določimo range skupin.

$$H = \frac{12}{N(N+1)} \left[\sum_{i=1}^{g} \frac{R_i^2}{n_i} \right] - 3(N+1) \tag{3.18}$$

kjer je:

- n_i število opazovanj v skupini i
- r_{ij} rank vrednosti j v skupini i
- $\overline{R}_i = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$
- število prostostnih stopenj $= g - 1$

Primer 3.2.2. Imamo 2 skupini s podatki: $\{23, 45, 67, 20\}$ in $\{77, 55, 66, 44\}$. Ko podatke uredimo naraščajoče, dobimo: $\{20, 23, 44, 45, 55, 66, 67, 77\}$. Tako dobimo ranke skupin: $\{2, 4, 7, 1\}$ in $\{8, 5, 6, 3\}$.

Sprejemanje odločitev:

p-vrednost je približek verjetnostne porazdelitve $Pr(\chi^2_{g-1} \geq K)$. Če imamo tabelo verjetnostne porazdelitve χ^2, lahko najdemo kritično vrednost pod $g - 1$ prostostnimi stopnjami. Nato pogledamo izbrano α vrednost. Ničelno hipotezo zavrnemo, kadar velja $K \geq \chi^2_{\alpha:g-1}$. Če rezultati kažejo na signifikantnost, to pomeni, da obstaja razlika med vsaj dvema skupinama.
Opomba: Če so nekatere vrednosti n_i majhne (t.j. manjše od 5), se lahko verjetnostna porazdelitev K precej razlikuje od porazdelitve χ^2.

3.2.7 Pearsonov korelacijski koeficient

Pearsonov korelacijski koeficient r meri linearno korelacijo med dvema normalno porazdeljenima slučajnima spremenljivkama X in Y. Zavzema vrednosti med $+1$ (popolna pozitivna linearna korelacija) in -1 (popolna negativna linearna korelacija), kjer 0 implicira nič korelacije. Definiran je kot:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \tag{3.19}$$

3.2.8 Spearmanov korelacijski koeficient

Spearmanov korelacijski koeficient ρ neparametrično meri monotonost med dvema slučajnima spremenljivkama X in Y. Za razliko od Pearsonovega korelacijskega koeficienta, Spearmanov koeficient ne predvideva, da sta oba nabora podatkov normalno porazdeljena. Ravno tako zavzema vrednosti med med -1 (popolna negativna linearna korelacija) in 1 (popolna pozitivna linearna korelacija). Izračunamo ga enako kot Pearsonov korelacijski koeficient, le da namesto slučajnimih spremenljivk X in Y vstavimo njihove range x_i in y_i.

3.2.9 Pearsonov χ^2 test

Pearsonov χ^2-test [3] je statistični test, ki preveri, če je odstopanje med dvema naboroma podatkov naključno ali sistematično. Pokriva situacije, kjer sta več kot dve kategoriji možnih izidov; na primer pacienti, kategorizirani glede na to, kakšno je njihovo stanje po zdravljenju (poslabšano, nespremenjeno, izboljšano). Uporablja se za testiranje ustreznosti vzorca in za test neodvisnosti. Test predpostavlja, da je disperzija sorazmerna s frekvenco vzorca, $\sigma^2 = E_i$ (da je vzorčenje Poissonovo). X^2 statistiko izračunamo po obrazcu:

$$E_i = \frac{N}{n} \qquad X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{3.20}$$

kjer je N število opazovanj, n število predalčkov, O_i opazovana frekvenca in E_i pričakovana frekvenca. X^2 pomeni veliki grški χ, in ga moramo ločit od simbola za spremenljivko X.

Sprejemanje odločitev:

Primerjamo testno statistiko X^2 s kritično vrednostjo iz χ^2 porazdelitve z $d-1$ prostostnimi stopnjami.

Primer 3.2.3. Imamo tristrano kocko. Na njej so narisani slon, želva in krokodil. Vemo, da je kocka pravična in, da je za vsako žival tretjina možnosti, da pade. Kocko vržemo 300 krat in dobimo rezultate:

	sloni	želve	krokodili	vsi
Opazovane frekvence	89	120	91	300
Pričakovane frekvence	100	100	100	300

Zanima nas, ali so rezultati signifikantno različni od pričakovane frekvence. Sedaj dobljene podatke vstavimo v formulo:

$$\frac{(\text{opazovana vrednost} - \text{pričakovana vrednost})^2}{\text{pričakovana vrednost}}$$

Rezultati za živali so: sloni : $1,21$, želve : $4,0$, krokodili : $0,81$. Seštevek teh vrednosti je statistika $X^2 = 6,02$. Nato pogledamo v tabelo porazdelitve χ^2 z 2 prostostnima stopnjama in zaključimo, ali je naša meritev signifikantno različna.

3.2.10 Kolmogorov–Smirnov test (dvo-vzorčni)

Kolmogorov–Smirnov test je eden najbolj uporabnih neparametričnih testov dveh neodvisnih slučajnih spremenljivk. Občutljiv je na razlike v lokaciji in obliki empirične kumulativne porazdelitve funkcije dveh vzorcev. K-S test meri razdaljo med empirično porazdelitveno funkcijo dveh vzorcev. Testira ničelno hipotezo, da vzorca izhajata iz iste zvezne porazdelitve.

Empirična porazdelitvena funkcija je F_n na n neodvisnih in enako porazdeljenih opazovanj X_i definirana kot:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}, \tag{3.21}$$

kjer je $I_{X_i \leq x} = \begin{cases} 1 & X_i \leq x \\ 0 & \text{sicer} \end{cases}$.

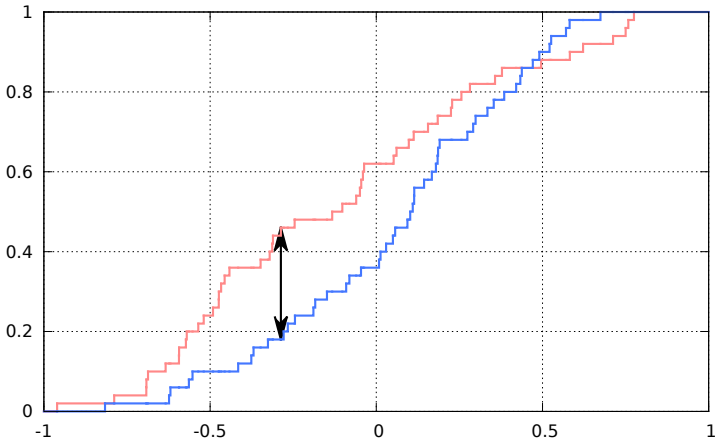
Statistiko Kolmogorov-Smirnova izračunamo po obrazcu:

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|. \tag{3.22}$$

Sprejemanje odločitev:

V praksi statistika potrebuje precejšnje število točk, da pravilno zavrže ničelno hipotezo. H_0 pri stopnji značilnosti α zavržemo, če velja $D_{n,n'} > c(\alpha)\sqrt{\frac{n+n'}{nn'}}$, kjer so vrednosti $c(\alpha)$ podane v tabeli:

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95



Slika 3.4: *Ilustracija dvo-vzorčnega K-S testa. Črna puščica predstavlja dvo-vzorčno K-S statistiko in meri največji razmik med kumulativnima porazdelitvama (rdeča in modra črta).*

3.2.11 Shapiro-Wilkov test normalnosti

Shapiro-Wilkov test testira ničelno hipotezo, da so vzorci $x_1, x_2, ..., x_n$ porazdeljeni po normalni porazdelitvi. Testno statistiko izračunamo po obrazcu:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{3.23}$$

kjer je x_i i-ta vrednost v naraščajočem vrstnem redu in a_i vrednosti so določene z obrazcem:

$$(a_1, \ldots, a_n) = \frac{m^\top V^{-1}}{||m^T V^{-1}||},$$

kjer $m = (m_1, \ldots, m_n)^\top$, $m_1, ..., m_n$ so pričakovane vrednosti urejene statistike, V je kovariančna matrika [1]. Prikažemo jo lahko s Q-Q plotom.

Sprejemanje odločitev:

Če je izračunana p-vrednost manjša od izbrane stopnje značilnosti α, potem zavrnemo ničelno hipotezo.

3.2.12 Univariantna linearna regresija

Linearna regresija je model funkcijske povezave med različnimi spremenljivkami. Predpostavlja, da je med spremenljivkama x in y linearna zveza ($y = \alpha + \beta x$). Naprej sestavimo *stohastični (verjetnostni) model*, ker praktična odstopanja od linearne zveze obravnavamo kot vrednosti slučajne spremenljivke U,

$$Y = \alpha + \beta X + U, \tag{3.24}$$

kjer so X, Y, U slučajne spremenljivke. Ker želimo vsaj v povprečju linearen model, predpostavimo $E(U) = 0$. Slučajni vzorec je realizacija slučajnega vektorja:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n). \tag{3.25}$$

Vpeljemo spremenljivke $U_i = Y_i - \alpha - \beta X_i$ za $i = 1, 2, \ldots, n$. Pri linearni regresiji predpostavljamo, da so spremenljivke U_i med seboj neodvisne in enako porazdeljene z matematičnim upanjem 0 in disperzijo σ^2, torej $E(U_i) = 0$, $E(U_i^2) = \sigma^2$ in $E(U_i U_j) = 0$ za $i \neq j$. Ponavadi privzamemo, da lahko točno kontroliramo vrednosti spremenljivk X_i tudi pri večkratnem ponavljanju meritev.

Denimo, da je približek za premico $y = \alpha + \beta x$ premica $y = a + bx$. Koeficienta a in b določimo po postopku *najmanjših kvadratov*. Torej iščemo minimum funkcije:

$$f(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{3.26}$$

pod pogoji

$$\frac{\partial f}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0,$$

$$\frac{\partial f}{\partial b} = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0.$$

Dobimo znane cenilke za α in β:

$$B = \frac{C_{xy}}{C_{xx}}, \qquad A = \bar{Y} - B\bar{X},$$

kjer je

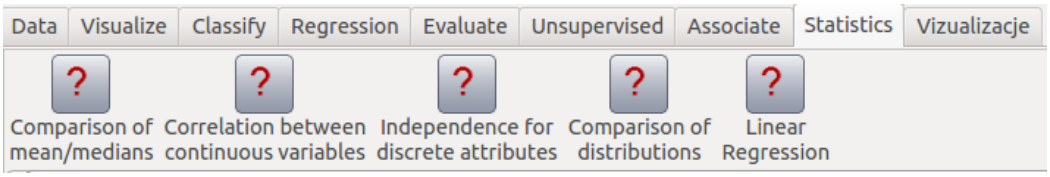
$$C_{xy} = \bar{xy} - \bar{x}\bar{y}$$

$$C_{xx} = \bar{xx} - \bar{x}\bar{x}.$$

Poglavje 4

Gradniki

V tem poglavju bomo opisali funkcije, uporabljene teste in vizualizacije gradnikov (Slika 4.1), ki smo jih sestavili v sklopu diplomske naloge. Vsi vsebujejo statistične teste opisane v poglavju 3. Nahajajo se na spletnem repozitoriju: `https://bitbucket.org/amii/orange/`.



Slika 4.1: *Gradniki v svoji kategoriji Statistics.*

4.1 Primerjava povprečij oz. median

Primerjava povprečij oz. median je gradnik, namenjen primerjanju povprečij, median in petštevilčnega povzetka.

Uporabnik najprej poveže gradnik z drugim gradnikom, ki ima za izhodni podatek tabelo tipa `ExampleTable`. Ta lahko vsebuje zvezne ali diskretne atribute, kateri se glede na svoj tip postavijo v pravo škatlo za izbiranje.

Uporabnik na levi strani uporabniškega vmesnika določi zvezni atribut in diskretni atribut. Nato gradnik grupira vrednosti zveznega atributa po vrednostih izbranega diskretnega atributa ter za vsako grupo izračuna petštevilčni povzetek (glej primer 5.1.1). Izvede tudi parametrični (t-test (3.2.3) ali ANOVA (3.2.5)) in neparametrični (Mann-Whitney (3.2.4) ali Kruskal Wallisov (3.2.6)

test. Poleg izračunanih vrednosti gradnik izriše škatlasti prikaz, ki si ga lahko uporabnik po želji shrani v obliki poročila.

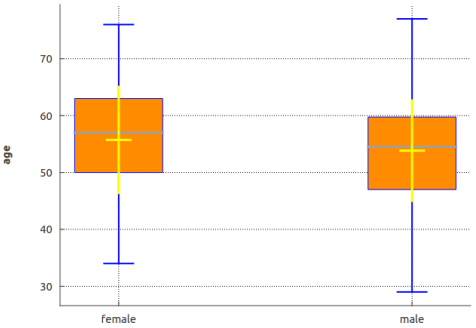
Primer 4.1.1. Denimo, da imamo v tabeli med drugim podatke o značilnostih oseb:

Višina	barva
164	modra
181	rjava
175	rjava
172	modra
...	...

Višina je zvezni atribut in barva je diskretni atribut. Zato gradnik glede na barvo grupira skupaj {164, 172, ...} ter {181, 175, ...} ter za vsako grupo izriše eno škatlo.

4.1.1 Vizualizacija

Petštevilčni povzetek je prikazan s škatličnim grafikonom (Slika 4.2). To je diagram, ki prikazuje numerične petštevilčne povzetke (kvartile in povprečje) za slučajne spremenljivke. Na vodoravni osi se nahajata dve ali več zveznih slučajnih spremenljivk, ki jih primerjamo. Škatla predstavlja območje, v katerem se nahaja 50% vseh podatkov, medtem ko navpične črte okoli njih prikazujejo območje, na katerem se na vsaki strani nahaja 25% podatkov. Iz asimetrije škatle glede na mediano lahko ocenimo poševnost porazdelitve. Na levi strani gradnika lahko izbiramo vrstni red škatel: glede na oznako, glede na povprečje in glede na mediano.

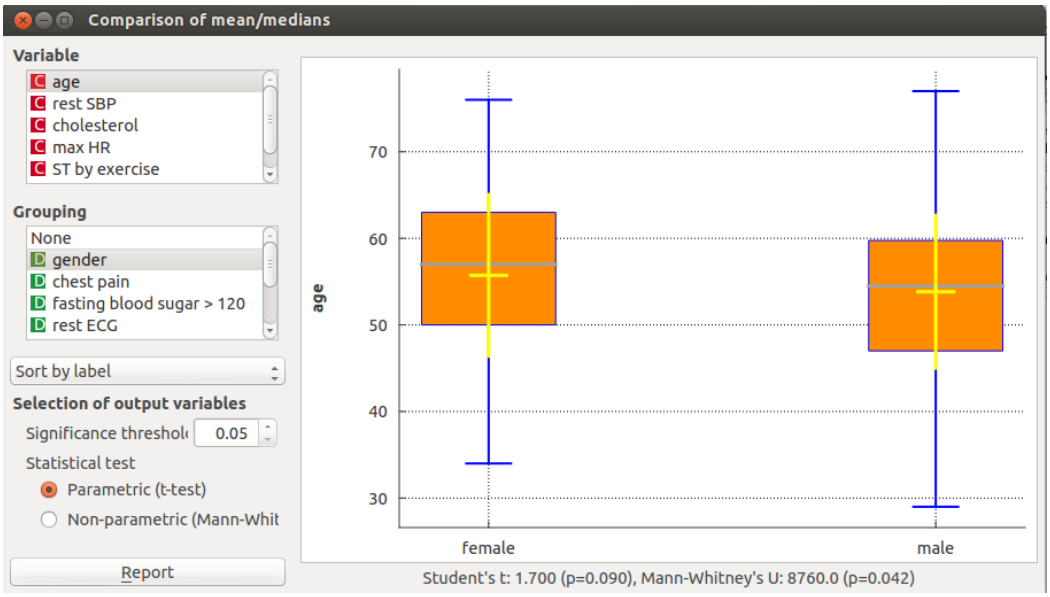


Slika 4.2: *Škatlični diagram, prikazuje dve škatli. Na vodoravni osi sta dve vrednosti diskretnih spremenljivk, na navpični osi pa dolžina življenja.*

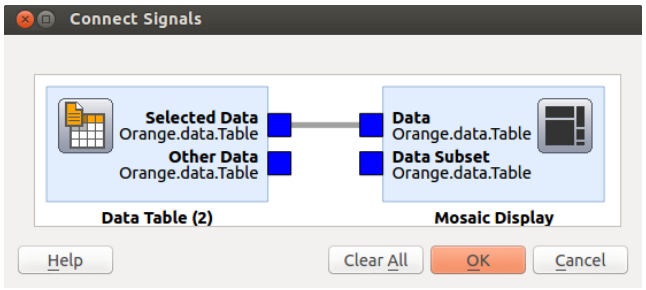
4.1.2 Izhodni podatki

Gradnik nam poleg prikaza podatkov ponudi tudi dve tabeli tipa `ExampleTable` na izhodnih signalih.

Izbiramo lahko med osnovno statistiko (petštevilčni povzetek) in podatki, ki jih gradnik izloči glede na nastavljeno stopnjo značilnosti. Izhodne podatke lahko povežemo z vsemi gradniki, ki za vhodni podatek sprejmejo tabele tipa `ExampleTable`. Če želimo zamenjati izbiro izhodnih podatkov, dvokliknemo na povezavo med gradnikoma in se nam spet pojavi okno za izbiro izhodnih podatkov (Slika 4.4).



Slika 4.3: *Gradnik: Primerjava povprečij oz. median. Na sliki sta izbrana atributa "age" in "gender". Iz slike lahko vidimo, da je pri ženskah povprečna življenjska doba višja kot pri moških.*

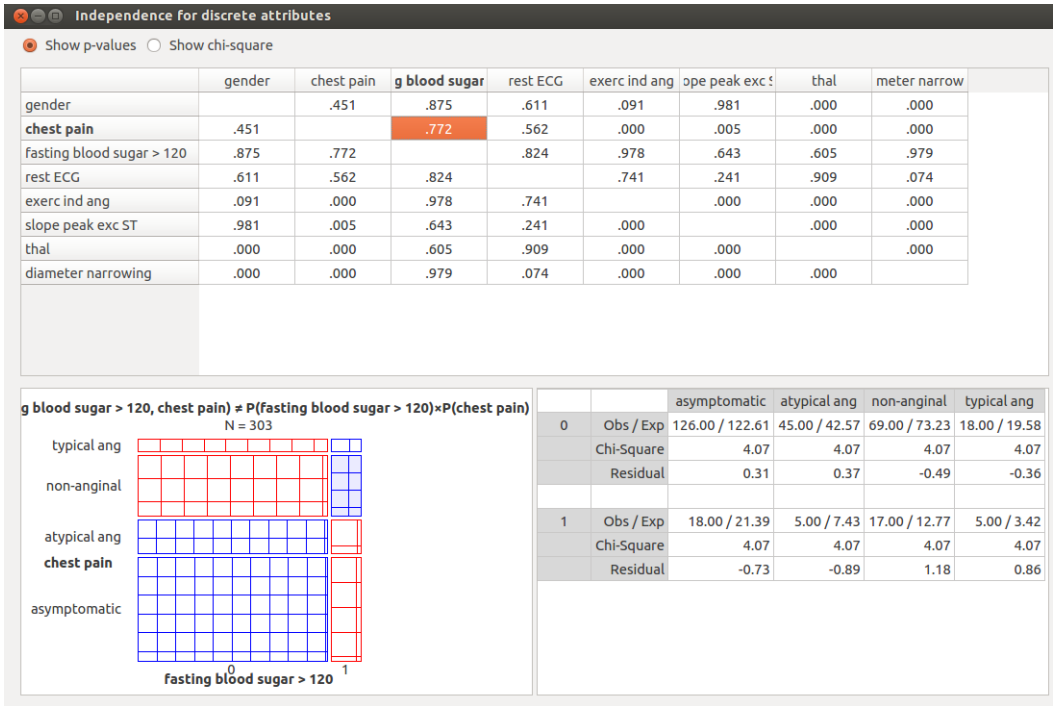


Slika 4.4: *Izbira izhodnih podatkov poteka tako, da z miško kliknemo na modri kvadrat levo ponujenih opcij in jih povežemo z željenim vhodom (desno ponujena izbira).*

4.2 Neodvisnost diskretnih spremenljivk

Neodvisnost diskretnih spremenljivk je gradnik, namenjen analizi odvisnosti diskretnih atributov.

Za vhodni podatek gradnik sprejme tabelo tipa `ExampleTable`, iz katere izbere samo diskretne atribute. Izvede se Pearsonov χ^2 test (3.2.9) in uporabnik lahko z radijski gumbi izbira med prikazom p-vrednosti ali statistiko X^2. Pod radijskimi gumbi se nahaja tabela z izbranimi vrednostmi diskretnih atributov. Spodnji del pokaže vizualizacijo za par, ki je izbran v tabeli. Levo se nahaja parketni diagram, desno je tabela s številčnimi vrednostmi diagrama med vsemi diskretnimi kombinacijami: izmerjena in pričakovana vrednost, vrednost χ^2 in Pearsonov ostanek.



	gender	chest pain	g blood sugar	rest ECG	exerc ind ang	ope peak exc	thal	meter narrow
gender		.451	.875	.611	.091	.981	.000	.000
chest pain	.451		.772	.562	.000	.005	.000	.000
fasting blood sugar > 120	.875	.772		.824	.978	.643	.605	.979
rest ECG	.611	.562	.824		.741	.241	.909	.074
exerc ind ang	.091	.000	.978	.741		.000	.000	.000
slope peak exc ST	.981	.005	.643	.241	.000		.000	.000
thal	.000	.000	.605	.909	.000	.000		.000
diameter narrowing	.000	.000	.979	.074	.000	.000	.000	

		asymptomatic	atypical ang	non-anginal	typical ang
0	Obs / Exp	126.00 / 122.61	45.00 / 42.57	69.00 / 73.23	18.00 / 19.58
	Chi-Square	4.07	4.07	4.07	4.07
	Residual	0.31	0.37	-0.49	-0.36
1	Obs / Exp	18.00 / 21.39	5.00 / 7.43	17.00 / 12.77	5.00 / 3.42
	Chi-Square	4.07	4.07	4.07	4.07
	Residual	-0.73	-0.89	1.18	0.86

Slika 4.5: *Gradnik: Neodvisnost diskretnih spremenljivk. V glavni tabeli je izbran prikaz p-vrednosti. Spodaj se nam prikazuje vizualizacija obarvanega polja v tabeli s parketnim diagramom in tabelo številčnih vrednosti diagrama.*

Vizualizacija

Na mrežnem diagramu so z rdečo predstavljena območja, kjer so meritve nižje od pričakovanih in z modro območja, kjer so meritve višje od pričakovanih. Ko z miško zapeljemo čez diagram se nam prikažejo vrednosti vizualicije v opisnem okencu.

4.3 Korelacija zveznih spremenljivk

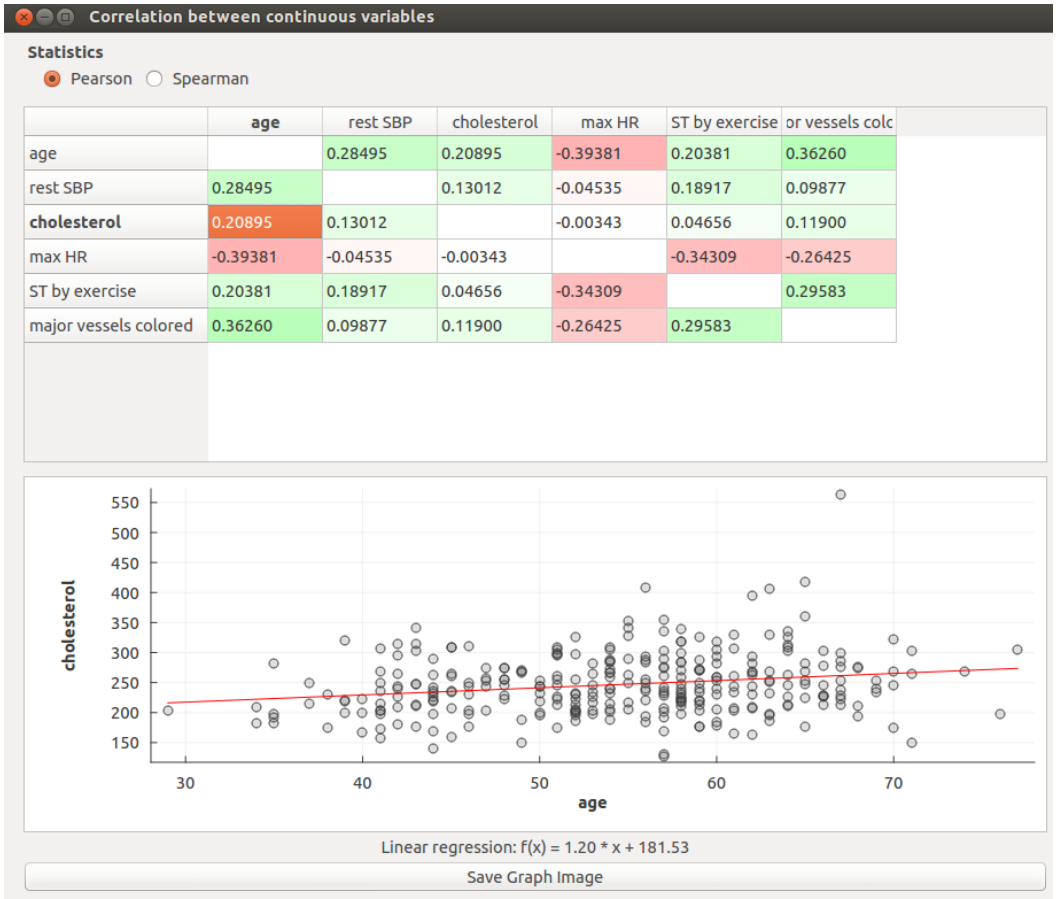
Korelacija zveznih spremenljivk je gradnik, namenjen analizi korelacije dveh zvezno porazdeljenih atributov na podlagi dveh vzorcev.

Za vhodni podatek sprejme tabelo tipa `ExampleTable`, v kateri obdrži zvezne atribute. Za vsak par izračuna korelacijski koeficient. Uporabnik ima na izbiro Pearsonov koeficient (3.2.7) in Spearmanov koeficient (3.2.8). Korelacijski koeficienti, ki ležijo med -1 (podatki ležijo na premici in padajo navzol), preko 0 (atributa sta nekorelirana) do 1 (podatki ležijo na premici in rastejo navzgor), so predstavljeni v tabeli. Celice v tabeli so za lažjo uporabo obarvane: zeleno za največje pozitivne korelacije, belo za korelacijski koeficient enak 0, in rdeče za negativne korelacije, pri čemer so za vmesne vrednosti barve zvezno interpolirane med skrajnimi vrednostmi.

Uporabnik ima s klikom na vnos v tabeli možnost izbire para atributov, za katera gradnik prikaže raztreseni graf, z izbranima atributoma na koordinatnih oseh. Graf je oplemeniten s premico linearne regresije, izračunano z uporabo funkcije `scipy.stats.linregress` iz SciPy paketa. Tako je korelacija tudi nazorno razvidna iz raztresenosti podatkov glede na prikazano premico.

Uporabniški vmesnik

Uporabnik najprej zagotovi vhodne atribute tako, da poveže gradnik "File" ali gradnik "DataTable" z gradnikom za korelacijo neodvisnih spremenljivk. Nato lahko s klikom na radijski gumb izbira med Pearsonovim ali Spearmanovim koeficientom. Pod izbiro se mu prikaže obarvana tabela, na kateri lahko s klikom izbira celice, ki pripadajo parom vhodnih atributov. Po izbiri celice se mu spodaj osveži graf, na katerem so prikazane vrednosti izbranih atributov in premica linearne regresije. Uporabnik lahko po želji doda sliko grafa na poročilo.

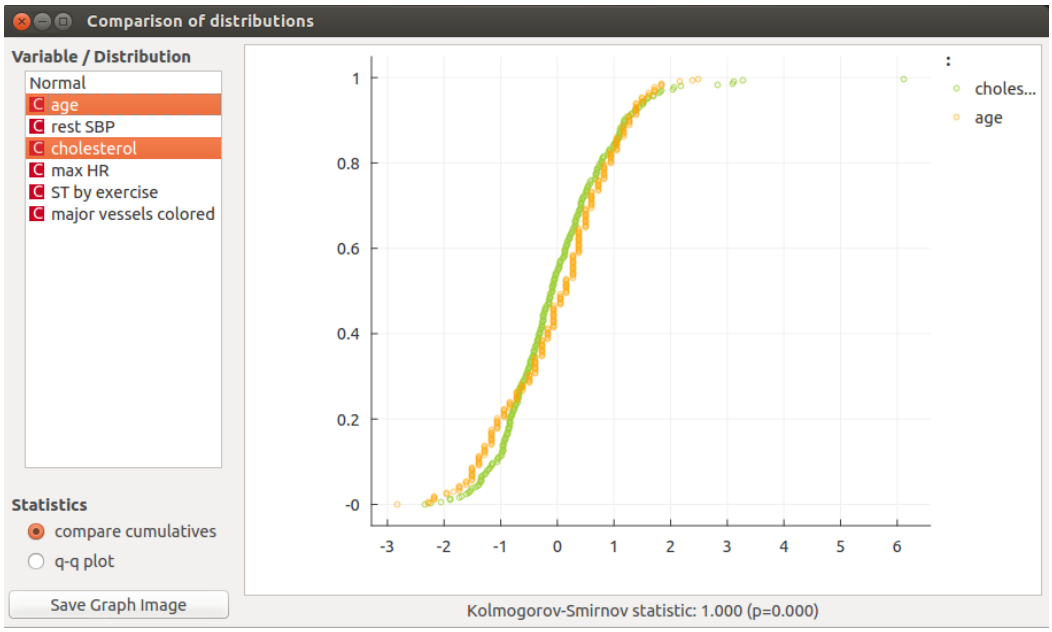


Slika 4.6: *Gradnik: Korelacija zveznih spremenljivk, Na primeru smo povezali gradnik "File" z "Correlation between continuous variables" in s tem zagotovili vhodne podatke* heart_disease.tab. *Izbrano imamo celico atributov* age *in* cholesterol, *ki prikazuje Pearsonov koeficient. Pod tabelo je izrisan raztreseni graf ter premica linearne regresije, iz katerih lahko sklepamo na trend naraščanja holesterola v krvi z leti.*

4.4 Primerjava porazdelitev

Primerjava porazdelitev (Slika 4.7) je gradnik, namenjen primerjavi porazdelitev dveh zvezno porazdeljenih atributov.

Za vhodni podatek sprejme tabelo tipa `ExampleTable`, v kateri obdrži zvezne atribute in jih vstavi v škatlo za izbiranje. Uporabnik izbere dva atributa in željeno vizualizacijo podatkov (na voljo sta Primerjava kumulativ in Q-Q grafikon). Po želji lahko primerjamo porazdelitev tudi z normalno porazdelitvijo. Gradnik primerja porazdelitve tudi numerično s testom Shapiro-Wilk (3.2.11) za testiranje z normalno porazdelitvijo in test Kolmogorov-Smirnov (3.2.10) za ostale kombinacije porazdelitev. Po želji lahko shranimo sliko grafikona v obliki *pdf* poročila.

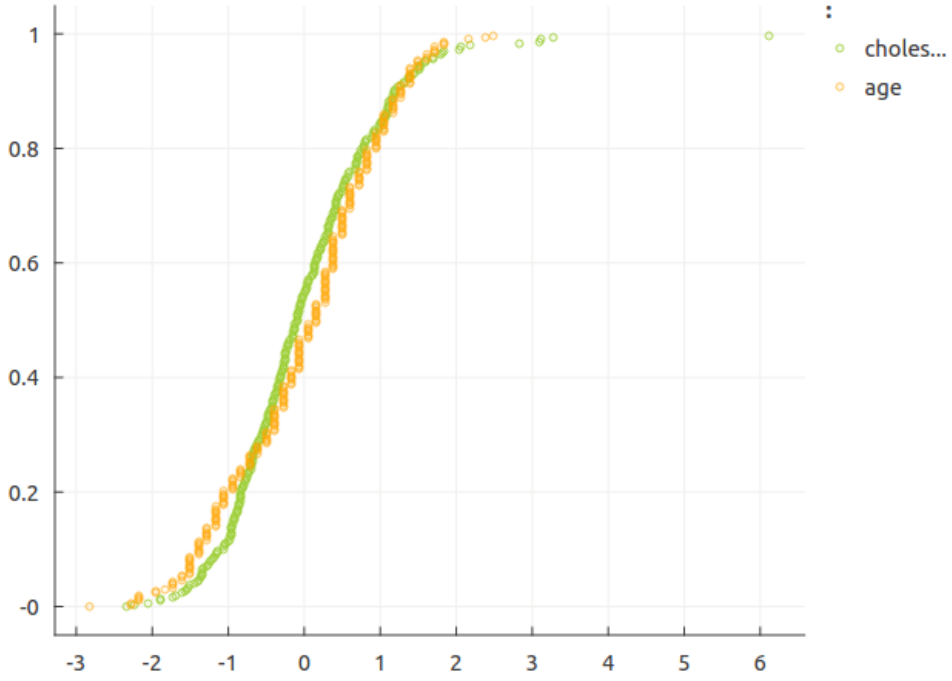


Slika 4.7: *Gradnik: Primerjava porazdelitev. Levi imamo kontrolno območje na katerem imamo izbrana dva atributa ter vizualizacijo s kumulativi. Desno je izrisan prikaz glede na izbrane nastavitve.*

4.4.1 Vizualizacije

Primerjava kumulativ

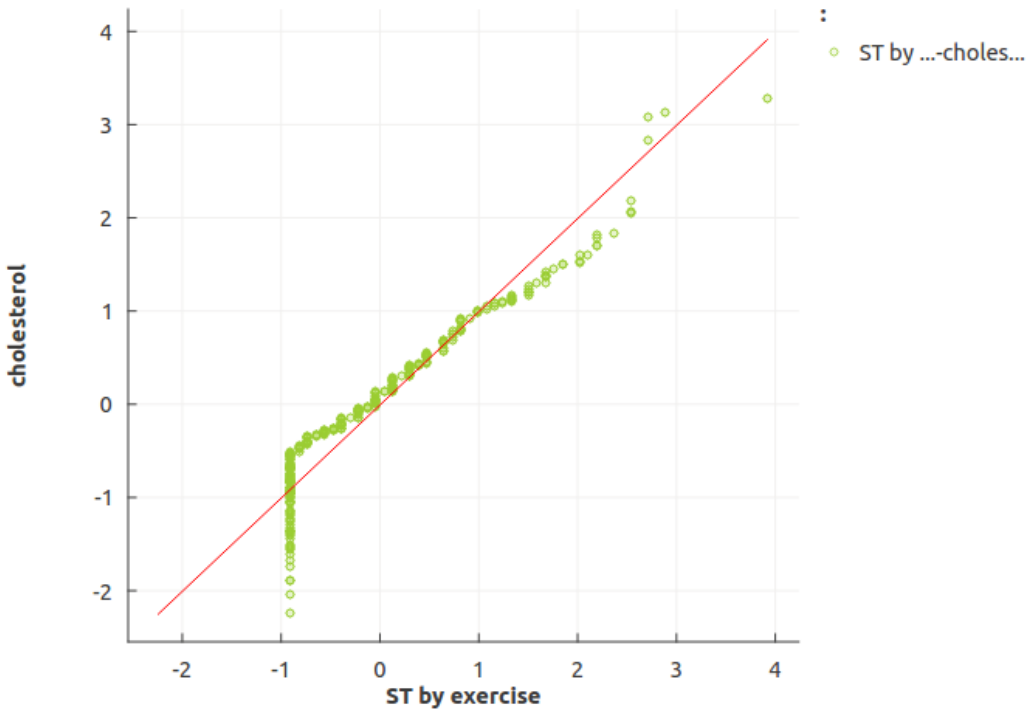
Pred primerjavo kumulativ (tekočih seštevkov) gradnik za oba atributa uvede novo spremenljivko, tako da imata obe porazdelitvi povprečje pri nič in enotsko disperzijo (Slika 4.8). S tem dosežemo, da lahko primerjamo obliko porazdelitev tudi za atribute, ki sicer ne opisujejo istih merskih količin. Na grafikonu sta narisani porazdelitvi vsaka s svojo barvo.



Slika 4.8: *Grafikon primerjave kumulativ. Obe porazdelitvi sta prilagojeni, da imata povprečje pri nič in imata enako širino, zato primerjava poudari razlike v obliki porazdelitve. Iz grafa lahko sklepamo, da obe spremenljivki sledita isti porazdelitvi.*

Q-Q grafikon

Q-Q grafikon (Slika 4.9) se uporablja za primerjavo porazdelitev in je grafikon kvantilov dveh distribucij. Kot pri primerjavi kumulativ, tudi tukaj pred primerjavo gradnik uvede novo spremenljivko. Iz vzorca točk na diagramu lahko sklepamo, koliko sta si dve porazdelitvi podobni. Q-Q graf je vedno naraščajoč. Če sta dve porazdelitvi enaki, točke sledijo premici $x = y$.



Slika 4.9: *Q-Q grafikon. Porazdelitvi sta nanešeni vsaka na svojo os, primerjava z diagonalo pa pokaže odstopanje med porazdelitvama. Tudi tukaj sta obe porazdelitvi pred izrisom centrirani in nastavljeni na enotsko širino.*

Poglavje 5

Zaključek

V diplomskem delu je bilo treba razviti štiri statistične gradnike za Orange, kar smo uspešno naredili. S tem smo Orange oplemenitili z naborom gradnikov, potrebnih za osnovno statistično analizo podatkov. Gradniki so prijazni za uporabo, se povezujejo z ostalimi gradniki preko signalov in omogočajo vizualno programiranje. Vsak vsebuje interaktivni uporabiški vmesnik na kontrolnem območju in glavno območje, na katerem se prikazuje vizualizacija. Vizualizacije se lahko s klikom vključi v poročilo.

Sedaj lahko tudi v Orange statistično organiziramo in povzamemo informacije. Z gradnikom *Primerjava povprečij oz. median* primerjamo povprečja in testiramo ali so si med seboj signifikantno različna. Z gradnikom *Neodvisnost diskretnih spremenljivk* analiziramo korelacijo med diskretnimi atributi in z *Korelacija zveznih spremenljivk* preučujemo povezanost zveznih atributov. Poleg tega lahko z gradnikom *Primerjava porazdelitev* primerjamo porazdelitve različnih podatkov. Gradniki vsebujejo mnogo testov za testiranje statističnih hipotez. Eden od prvih in največjih izzivov je bilo usposobiti delovno okolje in se ga naučiti uporabljati. Poleg tega je bil izziv z gradniki zagotoviti dobro uporabniško izkušnjo.

Nadaljnje delo

Glede na to, da je med izdelavo diplomskega dela izšel novi Orange, napisan v Python 3, s popolnoma spremenjenim Orange Canvasom, bo treba gradnike posodobiti. Poleg tega bi se podatkom lahko dodalo in nato tudi upoštevalo napake meritve. Lahko bi se razvilo še več statističnih gradnikov in s tem izpopolnilo nabor orodij za bolj dovršeno in prilagodljivo analizo podatkov.

Literatura

[1] J. Demšar, B. Zupan, G. Leban, T. Curk, "Orange: From Experimental Machine Learning to Interactive Data Mining", v zborniku *Knowledge Discovery in Databases*, 2004, str. 537–539.

[2] M. Hladnik, "Verjetnost in statistika (Zapiski predavanj)", Založba FE in FRI, 2002.

[3] (2013) R. Lowry "Chi-Square Procedures for the Analysis of Categorical Frequency Data", pogl. 8. Dostopno na:
`http://vassarstats.net/textbook/ch8pt1.html`

[4] J. Mauko, "Testiranje statističnih hipotez", Diplomsko delo, Fakulteta za naravoslovje in matematiko, 2010.

[5] (2013) M. Raič, "Vaje iz verjetnosti in statistike", Dostopno na:
`http://valjhun.fmf.uni-lj.si/~raicm/`

[6] J. A. Rice, "Mathematical Statistics and Data Analysis", Wadsworth Publishing Company, 1995.

[7] R. S. Witte, J. S. Witte, "Statistics", John Wiley & Sons, 2004.

[8] (2013) Orange - Data mining Fruitful & Fun. Dostopno na:
`http://orange.biolab.si/`

[9] (2013) Wikipedia. Dostopno na:
`http://wikipedia.org/`