МАТФ
Универзитет у Београду
Математички факултет

Биостатистика и анализа података

Марија Цупарић

Једанаести час

Категорички предиктори

- ⋆ променљиве које не узимају вредности које се природно изражавају бројем (или узимају мали број различитих вредности)
- ⋆ могу се представити коришћењем индикатора (dammy variables)
- ⋆ Z обележје од k категорија $\Longrightarrow$ представља се помоћу $k-1$ индикатора $(X_1, ..., X_{k-1})$

Z	$(X_1, ..., X_{k-1})$
1	(1,0,...,0)
2	(0,1,...,0)
$\vdots$	$\vdots$
$k-1$	(0,0,...,1)
k	(0,0,...,0)

- ⋆ тиме се постиже интепретабилност коефицијената уз коришћење индикатора
- ⋆ ако Z има две категорије и $Y = \alpha + \beta X_1 + \varepsilon$, $X_1 = 1$ је прва категорија а $X_1 = 0$ је друга категорија, тада β представља разлику вредности зависне променљиве између прве и друге категорије

Пример. Посматрамо податке у бази Salaries из пакета car и посматрамо утицај пола и дужине радног стажа на плату.

```
library(car)
head(Salaries)
```

```
       rank discipline yrs.since.phd yrs.service  sex salary
1      Prof          B            19          18 Male 139750
2      Prof          B            20          16 Male 173200
3  AsstProf          B             4           3 Male  79750
4      Prof          B            45          39 Male 115000
5      Prof          B            40          41 Male 141500
6 AssocProf          B             6           6 Male  97000
```

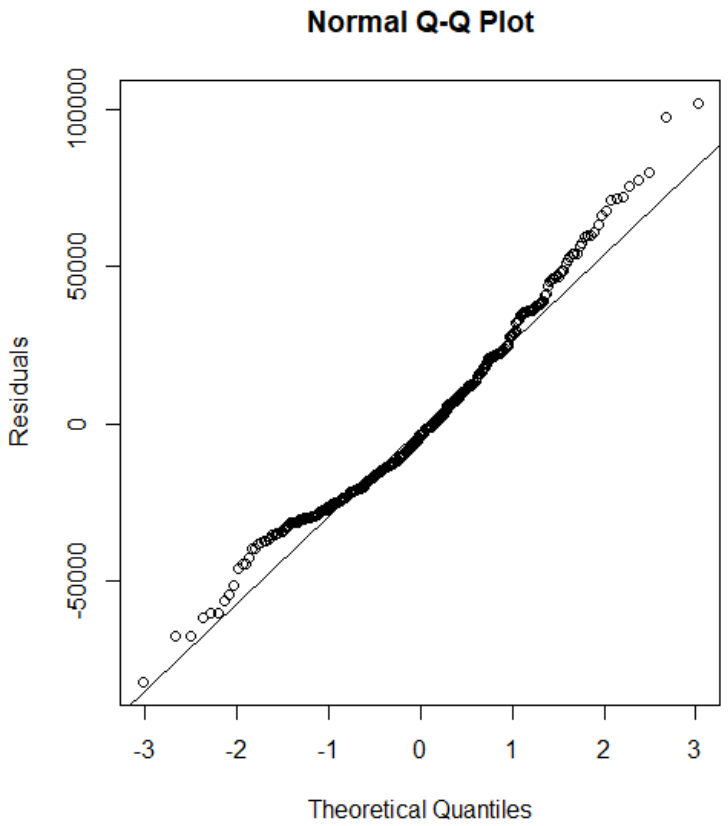
```
model<-lm(salary ~ sex + yrs.service, data = Salaries)
summary(model)
```

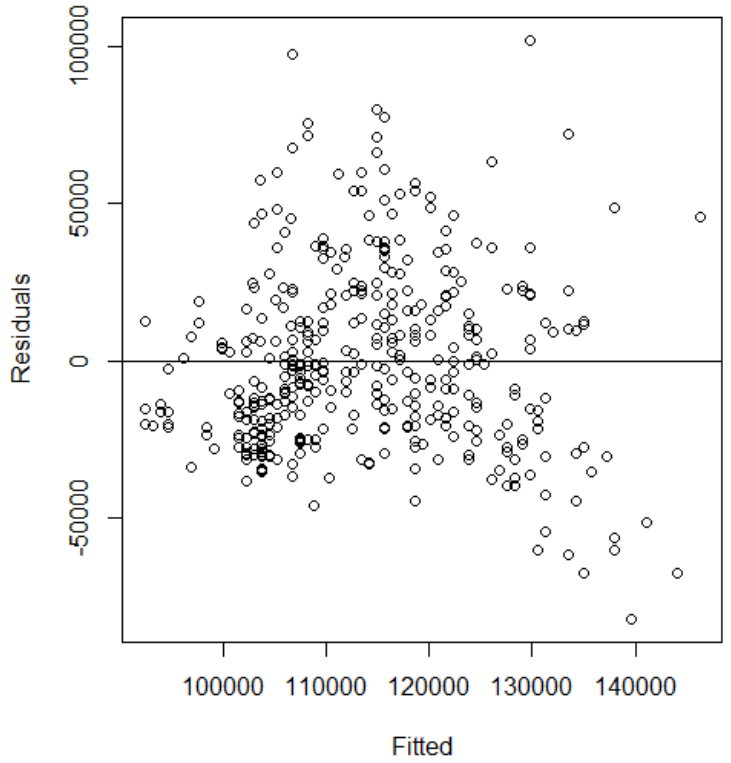
```
Call:
lm(formula = salary ~ sex + yrs.service, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-81757 -20614  -3376  16779 101707

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  92356.9     4740.2  19.484  < 2e-16 ***
sexMale       9071.8     4861.6   1.866   0.0628 .
yrs.service    747.6      111.4   6.711 6.74e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28490 on 394 degrees of freedom
Multiple R-squared:  0.1198,    Adjusted R-squared:  0.1154
F-statistic: 26.82 on 2 and 394 DF,  p-value: 1.201e-11
```





Може се рећи да су испуњене претпоставке модела.

- ⋆ коефицијент уз предиктор yrs.service је значајан (на основу $t-$теста), док уз sex зависи са којим нивоом значајности гледамо
- ⋆ проверимо сада да ли је модел бољи ако се користи само један предиктор

```
model1=lm(salary~ yrs.service, data = Salaries)
anova(model1,model)
```

```
Analysis of Variance Table

Model 1: salary ~ yrs.service
Model 2: salary ~ sex + yrs.service
  Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
1    395 3.2259e+11
2    394 3.1977e+11  1 2825894127 3.4819 0.06279 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Дакле, поново је $p-$вредност теста на граници. Ако упоредимо коефоцијенте детерминације, види се да су једнаки 0.1198 и 0.1121 за моделе са два и једним предиктором редом. Према томе можемо закључити да пол не утиче значајно на плату.
Модел је: $\mu_{Y|yrs.service} = 99974.7 + 779.6 \cdot yrs.service$

Ако су категорије у самој бази наведене као стрингови, `R` ће их аутоматски посматрати као категорије. Ако су категорије наведене нумерички (нпр. 0 мушки пол и 1 женски пол) `R`-у се мора „рећи" да је у питању категоричка променљива уз помоћ функције `factor` примењене на саму променљиву.

У излазу функције `summary(model)` код индикаторских променљивих уз њихов назив се налази и вредност за коју је индикатор једнак 1. Вредност која фали представља референтну категорију.

Ако је x индикаторска променљива и $Y|x = \alpha + \beta x + \varepsilon$, тестирање $H_0 : \beta = 0$ је еквивалентно са тестирањем да је просечна вредност Y-а (које има нормалну расподелу) иста у обе категорије

Логистичка регресија

Да ли може да се моделира зависност променљиве Y од предиктора и када није нормална расподела у питању?
Може, али је потребно извршити неке модификације.

- ⋆ Претпоставимо да је Y индикатор (обележје које узима само две вредности, 0 и 1)

Пример

(1) Предвиђање да ли ће ћелија реаговати на терапију или не у зависности од нивоа експресије гена, концентрације лека, температуре и других фактора.

(2) Да ли ће доћи до фаталног исхода саобраћајне незгоде у зависности од типа возила, типа пута, сигнализације, временских услова,...

- ⋆ Нека је $p_i = P\{Y_i = 1\}$.

Како направити модел за p_i?

Пример. У истраживању једне популације биљака посматра се да ли биљка успешно развија цвет или не. Нека је Y висина биљке (зависна променљива), чија се средња вредност може моделовати линеарним моделом $Y|x_i = \alpha + \beta x_i + \varepsilon$, где је x_i количина доступне светлости. Међутим, уместо података $Y_1, \ldots, Y_n$ имамо само информацију да ли је биљка достигла минималну висину потребну за цветање (успешан развој) или не. Самим тим имамо узорак

$$Z_i = \begin{cases} 1, & Y \geq d \\ 0, & Y < d \end{cases}$$

Желимо да направимо модел за p_i где је

$$p_i = P\{Z_i = 1\} = P\{\alpha + \beta x_i + \varepsilon \geq d\} = P\{\varepsilon \leq d - \alpha - \beta x_i\}$$
$$= \Phi\left(\frac{d - \alpha - \beta x_i}{\sigma}\right),$$

где је $\sigma^2 = D(\varepsilon_i)$. Следи да је $\Phi^{-1}(p_i) = \frac{d-\alpha-\beta x_i}{\sigma} = A + B x_i$.

Једна могућност за моделовање p_i јесте да се трансформише њена вредност тако да трансформисана вредност буде у $\mathbb{R}$, а затим да се изврши моделовање линеарном регресијом.

У претходном примеру та трансформација је била Φ^{-1}.

Најчешћи типови такве регресије

пробит регресија - трансформација инверзом функције расподеле нормално расподељене случајне величине

логистичка регресија - трансформација инверзом функције расподеле логистички расподељене случајне величине; $F(x) = \frac{1}{1+e^{-x}}, x \in \mathbb{R}$ а њен инверз је $F^{-1}(x) = \log \frac{x}{1-x}, x \in (0,1)$

- ⋆ Модел логистичке регресије је: $\log \frac{p}{1-p} = \alpha + \beta x_i$.
- ⋆ Количник $\frac{p(x)}{1-p(x)}$ се назива квота

> Ако је квота 4, то значи да је, за фиксну вредност предиктора x, вероватноћа да је $Y = 1$ четири пута већа од вероватноће да је $Y = 0$.

- ⋆ Ако је $\beta < 0$, онда је $e^{\beta} < 1$, па долази до смањења квоте, док у супротном долази до њеног повећања
- ⋆ Када се параметри α и β оцене, добија се $\widehat{\log \frac{p}{1-p}} = A + Bx_i$
- ⋆ Трансформацијом претходног израза добија се да је оцена вероватноће да је $Y = 1$ када је вредност предиктора x једнака

$$\widehat{p}(x) = \frac{1}{1 + e^{-A-Bx}},$$

односно

$$\widehat{p}_i = \frac{1}{1 + e^{-A-Bx_i}}$$

Пример У јануару 1986. године дошло је до експлозије спејс-шатла Челинџер у коме је живот изгубило 7 чланова посаде. Као један од разлога катастрофе наводе се лоше перформансе спољних дихтунга на које је утицала изузетно ниска спољашња температура $31^\circ F$. Како би се испитала веза између попуштања дихтунга и спољне температуре лансирано је 6 летелица у различитим условима. Резултати се налазе у бази `challenger` (пакет `DATAstudio`). Променљива `origin` узима вредност 0 уколико није дошло до квара, а 1 ако је дошло до квара, а променљива `temp` означава температуру.

```
library(DATAstudio)
data(challenger)
model <- glm(oring ~ temperature, data = challenger, family =
    binomial)
summary(model)
```

```
Call:
glm(formula = challenger$oring ~ challenger$temperature, family = "binomial")

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           15.0429     7.3786   2.039   0.0415 *
challenger$temperature  -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```

p-вредност теста је 0.032 што упућује на значајан утицај температуре ваздуха на успех полетања. Коефицијент уз `temperature` је негативан што упућује на то да се смањењем температуре повећава вероватноћа неуспеха.

Добили смо модел: $\widehat{p}(x) = \frac{1}{1+e^{-(15.04-0.23x)}}$.

За $x = 31$ добија се $\widehat{p}(x) = 0.99963$.

```
predict(model, data.frame(temperature=31), type='response')
```

Можемо за конкретну вредност предиктора добити и 95% интервал поверења за $p(x)$.

$\eta(x) = \log \frac{p(x)}{1-p(x)}$ и $\eta(x) = \alpha + \beta x$

```
p <- predict(model, newdata = data.frame(temperature = 55), type =
    "link", se.fit = TRUE)
 eta_hat <- p$fit
 se_eta <- p$se.fit
CI_eta <- c(eta_hat - 1.96*se_eta, eta_hat + 1.96*se_eta)
 plogis(CI_eta)
```

Добијен је интервал за $p(55)$: (0.3350113, 0.9946935).

Коваријација

Интересује нас да ли постоји линеарна веза између променљивих X и Y, односно да ли постоје α и β такви да је $Y = \alpha + \beta X$

У регресионој анализи x није била случајна величина, а сада јесте.

Дефиниција: Нека су X и Y случајне величине са средњим вредностима μ_X и μ_Y. Коваријација између X и Y је

$$cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)).$$

Коваријација описује на који начин се X и Y истовремено одступају од својих средњих вредности

- ⋆ Ако су велике вредности X кад су велике вредности Y, $cov(X, Y) > 0$
- ⋆ Ако су велике вредности X кад су мале вредности Y, $cov(X, Y) < 0$
- ⋆ Ако су велике вредности X подједнако повезане и с великим и с малим вредностима Y, $cov(X, Y) = 0$

Коваријацију оцењујемо узорачком коваријацијом: $\widehat{cov(X, Y)} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{n-1}$

Пример. Проучава се веза између потрошње клијената банке у првом месецу коришћења платне картице (X) и укупна годишња потрошња (Y). Добијени су подаци $(10, 120), (5, 75), (25, 250), (100, 800), (30, 360)$. Приметимо да је $\bar{x} = 34, \bar{y} = 321, \sum xy = 98625$ Тада

$$\widehat{cov(X, Y)} = \frac{1}{n-1}(\sum xy - n\overline{xy}) = 11013.75.$$

Пошто је коваријација позитивна закључујемо да већа потрошња на годишњем нивоу је повезана са већом потрошњом у првом месецу.

R
```
x <- c(10, 5, 25, 100, 30)
y <- c(120, 75, 250, 800, 360)
cov(x,y)
```

Коваријација не мери јачину везе и не знамо како да тумачимо њене резултате. Шта је велико а шта мало? Зависи од проучавања.

Коефицијент корелације

Дефиниција: Нека су X и Y случајне величине са средњим вредностима μ_X и μ_Y и дисперзијама σ_X^2 и σ_Y^2. Коефицијент корелације између X и Y је

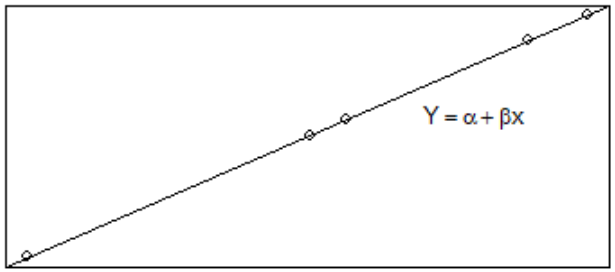
$$\rho = \frac{cov(X, Y)}{\sigma_X \sigma_Y}.$$

Теорема: Нека су α и $\beta \neq 0$ реални бројеви. Линеарна веза X и Y постоји, тј. $Y = \alpha + \beta X$ ако и само ако је $\rho = 1$ или $\rho = -1$.

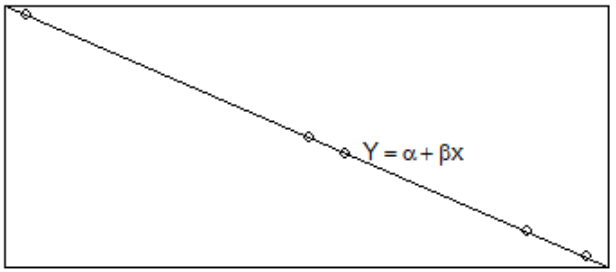
$\rho = 1$ постоји савршена линеарна веза с позитивном корелацијом

$\rho = -1$ постоји савршена линеарна веза с негативном корелацијом

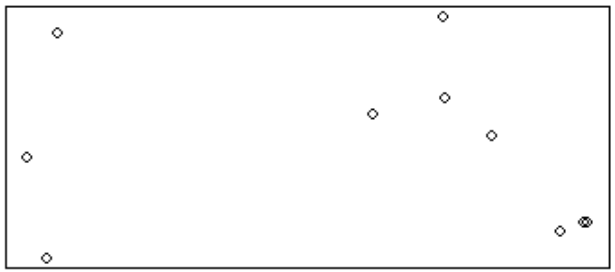
$\rho = 0$ случајне променљиве су некорелисане, па ако постоји веза међу њима, она никако није линеарна



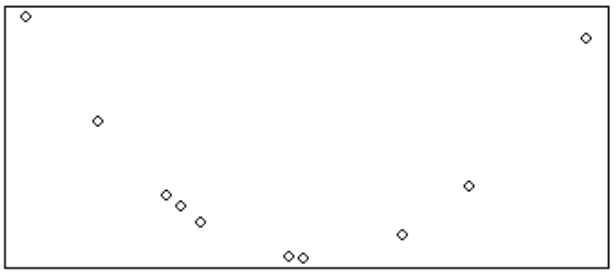
$\rho = 1$



$\rho = -1$



$\rho = 0$



$\rho = 0$

Дефиниција: Узорачки коефицијент корелације R дефинише се као

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)(\sum Y^2 - (\sum Y)^2)}}.$$

- ⋆ Вредности R блиске 1 (веће од 0.75) или -1 (мање од -0.75) сматрамо добром линеарном везом
- ⋆ Вредности R између 0.5 и 0.75, односно, -0.75 и -0.5, сматрамо осредњом линеарном везом
- ⋆ Остале вредности R сматрамо слабом линеарном везом

Пример (наставак). Проучава се веза између потрошње клијента банке у првом месецу коришћења платне картице (X) и укупна годишња потрошња (Y). Добијени су подаци $(10, 120), (5, 75), (25, 250), (100, 800), (30, 360)$.
Приметимо да је $\sum x = 170, \sum y = 1605, \sum xy = 98625, \sum x^2 = 11650, \sum y^2 = 852125$. Тада

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(\sum y^2 - (\sum y)^2)}} = 0.991.$$

Вредност r је близу 1 па постоји снажна позитивна линеарна веза X и Y. Снажна линеарна веза између потрошње у првом месецу и укупне годишње потрошње не значи да висока потрошња у првом месецу узрокује већу годишњу потрошњу; могуће је да постоји трећи заједнички узрочник, као што су приход клијента, начин живота или навике у коришћењу платне картице.

R
```
x <- c(10, 5, 25, 100, 30)
y <- c(120, 75, 250, 800, 360)
cor(x,y)
```

Веза регресије и корелације

- ⋆ Узорачки коефицијент корелације R у тесној је вези с нагибом регресионе праве B

$$R = \frac{B\sqrt{S_{xx}}}{\sqrt{S_{yy}}}$$

- ⋆ Знак коефицијента корелације одређује нам и знак нагиба регресије: када је $R > 0$, регресиона права расте како расте x, а када је $R < 0$ опада; када је $R = 0$, нагиб је такође једнак нули па регресиони модел није применљив
- ⋆ Приметимо још да је

$$R^2 = 1 - \frac{SSE}{S_{yy}}$$

- ⋆ Како је S_{yy} укупно одступање Y , а SSE одступање настало услед необјашене грешке, R^2 нам је проценат објашеног одступања регресионом линијом, односно коефицијент детерминације

Пример. Нека x представља број сати које је студент провео у учењу за тест из статистике, а Y број поена коју је добио на тесту. На узорку од 28 студената добијени су резултати:
(0.25,60), (0.75,60), (1.0,65), (1.5,70), (1.75,70), (2.5,75), (3.0,85), (0.5,62), (0.75,69), (1.0,70), (1.5,75), (2.0,73), (2.5,82), (3.0,82), (0.5,65), (0.75,71), (1.25,68), (1.5,65), (2.0,79), (2.75,85), (3.0,88), (0.5,80), (1.0,58), (1.25,71), (1.75,78), (2.0,60), (2.75,79), (3.5,90)

Како је $r^2 = 0.60$. То значи да је 60% одступања броја поена које студенти добију на тесту могуће објаснити бројем сати које су провели у учењу, док преосталих 40% одступања модел не може да објасни и сматрамо их случајном грешком.

Категорички подаци

⋆ Категоричке променљиве (дискретне номиналне или ординалне) су променљиве чији се исходи могу сврстати у ограничен број категорија.

Пример

(1) променљива: тип станишта – категорије: шума, ливада, мочвара, река

(2) променљива: здравствено стање – категорије: здрава јединка, заражена јединка

(3) променљива: развојни стадијум – категорије: ларва, јувенилна јединка, одрасла јединка

Ако проучавамо повезаност две категоричке променљиве X и Y које имају r и k категорија, за то не можемо користити коефицијент корелације.

Табеле контигенције

Претпоставимо да се разматрају (категоричке) случајне величине X и Y које имају r и k категорија редом. Тада цео узорак можемо поделити у $r \cdot k$ категорија и направити табелу контигенције.
Када обе променљиве имају само по две категорије, онда се користе табеле контигенције $2x2$. Сваки елемент узорка упада у тачно једну ћелију.

	Y	
X	категорија y_1	категорија y_2
категорија x_1	x_1 и y_1	x_1 и y_2
категорија x_2	x_2 и y_1	x_2 и y_2

Пример Спроведена је студија у једној популацији биљака како би се испитало да ли је начин узгоја повезан са успешним достизањем фазе цветања. Изабран је случајан узорак од 200 биљака које су узгајане стандарном или новом методом. За сваку биљку утврђено је да ли се успешно развила до фазе цветања. Од биљака узгајаних стандардном методом, 70 је успешно достигло фазу цветања док 30 није, док су код нове методе 82 биљке успешно развиле цвет а 18 није.

Случајне променљиве су: метод узгоја (стандардни и нови) и фаза цветања (успешно и неуспешно).

	метод узгоја		
фаза цветања	стандардни	нови	
успешна	$n_{11} = 70$	$n_{12} = 82$	$n_{1\bullet} = 152$
неуспешна	$n_{21} = 30$	$n_{22} = 18$	$n_{2\bullet} = 48$
	$n_{\bullet 1} = 100$	$n_{\bullet 2} = 100$	$n = 200$

Тестирање зависности између две категоричке променљиве

- ⋆ Тестирамо да не постоји зависност између променљивих против алтернативе да постоји
- ⋆ Разликујемо два случаја

тест независности: испитујемо да ли су неке две случајне променљиве независне - извлачимо узорак обима n и сваки елемент сврставамо у одговарајуће категорије, без претходног знања колико ће их у којој категорији бити;

тест хомогености: испитујемо да ли је код обе категорије случајне променљиве X подједнако заступаљена свака од категорија случајне променљве Y - извлачимо узорак обима $n_{1\bullet}$ и $n_{2\bullet}$ из сваке категорије за X (укупно n), а затим их сврставамо у категорије за Y.

Тест независности

У узорку обима n елементи се класификују на основу категорија сваке од променљивих. Ако су случајне величине независне онда класификоване вредности једне променљиве не дају информације о класификацији друге променљиве.

хипотезе: $H_0 : X$ и Y су независне и $H_1 : X$ и Y нису независне

идеја теста: упоредити стварни број елемената узорка у свакој категорији с очекиваним бројем елемената када би X и Y биле независне

тест статистика: $X_0^2 = \sum_{\text{по свим пољима}} \frac{(\widehat{E}_{ij} - n_{ij})^2}{\widehat{E}_{ij}}$, где је $\widehat{E}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ оцењени број елемената у $i-$том реду и $j-$тој колони, има χ_1^2 расподелу

$p-$вредност теста: површину десно од X_0^2, вредности коју је статистика X_0^2 узела у узорку.

Овај тест се може примењивати када је n велико. Обично се узима да је n довољно велико ако да је свако $\widehat{E}_{ij} > 5$. У супротном су $p-$вредности непрецизне.

Пример (наставак)

	метод узгоја		
фаза цветања	стандардни	нови	
успешна	$n_{11} = 70$ (76)	$n_{12} = 82$ (76)	$n_{1\bullet} = 152$
неуспешна	$n_{21} = 30$ (24)	$n_{22} = 18$ (24)	$n_{2\bullet} = 48$
	$n_{\bullet 1} = 100$	$n_{\bullet 2} = 100$	$n = 200$

$$\widehat{E}_{11} = \frac{n_{1\bullet} n_{\bullet 1}}{n} = \frac{152 \cdot 100}{200} = 76, \widehat{E}_{12} = \frac{n_{1\bullet} n_{\bullet 2}}{n} = \frac{152 \cdot 100}{200} = 76$$

$$\widehat{E}_{21} = \frac{n_{2\bullet} n_{\bullet 1}}{n} = \frac{48 \cdot 100}{200} = 24, \widehat{E}_{22} = \frac{n_{2\bullet} n_{\bullet 2}}{n} = \frac{48 \cdot 100}{200} = 24$$

$$X_0^2 = \sum_{\text{по свим пољима}} \frac{(\widehat{E}_{ij} - n_{ij})^2}{\widehat{E}_{ij}}$$

$$= \frac{(76-70)^2}{76} + \frac{(76-82)^2}{76} + \frac{(24-30)^2}{24} + \frac{(24-18)^2}{24} = 3.95$$

Користећи χ_1^2 расподелу, добија се да је $p-$вредност теста је 0.047 (`1-pchisq(3.95,1)`). Можемо рећи да постоји веза између метода узгоја и успешности цветања. Очекивали смо да се новом методом развије 76 биљака до цветања, а развијено је 82.

Тест хомогености

Вредности на једној маргини су фиксиране од стране истраживача док је друга случајна, односно узорак делимо у две групе (према једној категоричкој променљивој) унапред одређене величине

хипотезе: нулта хипотеза је да исти проценат елемената има одређено својство (друга категоричка променљива) у обе групе, тј. $H_0 : p_{11} = p_{21}$, а алтернативна је да је тај проценат различит $H_1 : p_{11} \neq p_{21}$

тест статистика: $X_0^2 = \sum_{\text{по свим пољима}} \frac{(\widehat{E}_{ij} - n_{ij})^2}{\widehat{E}_{ij}}$, где је $\widehat{E}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ оцењени број елемената у $i-$том реду и $j-$тој колони, има χ_1^2 расподелу

$p-$вредност теста: површину десно од X_0^2, вредности коју је статистика X_0^2 узела у узорку.

Пример Проучава се ефикасност нове методе узгоја биљака. Узети су узорци од 100 биљака узгајаних стандардном методом и 100 биљака узгајаних новом методом. За сваку биљку забележено је да ли се успешно развила до фазе цветања. Резултати су дати у табели.

	развој до фазе цветања		
метод	да	не	
нови	82		100
старандардни		30	100

Тестирати одговарајућу хипотезу.
Пошто су унапред одређене вредности на једној маргини, у питању је тест хомогености.

хипотезе: $H_0 : p_{11} = p_{21}$ и $H_1 : p_{11} \neq p_{21}$

реализована вредност тест статистике:

	развој до фазе цветања		
метод	да	не	
нови	82 (76)	18 (24)	100
стандардни	70 (76)	30 (24)	100
	152	48	

$$X_0^2 = \sum_{\text{по свим пољима}} \frac{(\widehat{E}_{ij} - n_{ij})^2}{\widehat{E}_{ij}} = 3.95$$

$p-$**вредност теста:** Користећи χ_1^2 расподелу, добија се да је $p-$вредност теста је 0.047 (`1-pchisq(3.95,1)`). Закључујемо да проценат биљака које су развиле цвет није исти. Већи је проценат код новог метода.

Табеле контигенције $r \times k$

Претпоставимо да се разматрају (категоричке) случајне величине X и Y које имају r и k категорија редом. Тада цео узорак можемо поделити у $r \cdot k$ категорија и направити табелу контигенције.

	Y				
X	1	2	$\cdots$	k	
1	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1\bullet}$
2	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rk}	$n_{r\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet k}$	n

хипотезе: исте као и раније за тест независности и хомогености

тест статистика: $X_0^2 = \sum_{\text{по свим пољима}} \frac{(\widehat{E}_{ij} - n_{ij})^2}{\widehat{E}_{ij}}$, где је $\widehat{E}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ оцењени број елемената у $i-$том реду и $j-$тој колони, има $\chi^2_{(r-1)(k-1)}$ расподелу

$p-$вредност теста: површину десно од X_0^2, вредности коју је статистика X_0^2 узела у узорку.

Пример. Продавац жели да утврди да ли постоји разлика између продаје током дана у недељи који је изабран за снижење и величине продаје. Током 2 године, 40 распродаја је спроведено током уторка, среде и четвртка, а распродаје су подељене у слабе, средње и велике. Добијени су резултати:

	величина продаје			
дан	слаба	средња	велика	
уторак	12	20	8	40
среда	12	22	6	40
четвртак	0	18	22	40

Извршити тестирање.

	величина продаје			
дан	слаба	средња	велика	
уторак	12 (8)	20 (20)	8 (12)	40
среда	12 (8)	22 (20)	6 (12)	40
четвртак	0 (8)	18 (20)	22 (12)	40
	24	60	36	120

$$X_0^2 = \sum_{\text{по свим пољима}} \frac{(\widehat{E}_{ij} - n_{ij})^2}{\widehat{E}_{ij}} = 25.067$$

$p-$вредност теста: Користећи χ^2_4 расподелу, добија се да је $p-$вредност теста приближно 0 (`1-pchisq(25.067,4)`). Можемо рећи да постоји (статистички значајна) повезаност између дана у недељи и величине продаје. Код слабе продаје уочавамо да је уторком и средом вредност већи од очекиване, а за велике мања од очекиване, док је четвртком обрнуто. Тако да су велике распродаје чешће четвртком него осталим данима.

```
prodaja <- matrix( c(12, 20, 8, 12, 22, 6, 0, 18, 22), nrow = 3,
    byrow = TRUE)
rownames(prodaja) <- c("utorak", "sreda", "cetvrtak")
colnames(prodaja) <- c("slaba", "srednja", "velika")
prodaja
chisq.test(prodaja)
```

	slaba	srednja	velika
utorak	12	20	8
sreda	12	22	6
cetvrtak	0	18	22

```
        Pearson's Chi-squared test

data:  prodaja
X-squared = 25.067, df = 4, p-value = 4.878e-05
```

Тест је прецизан само за велике узорке, а n сматрамо довољно великим, ако ни у једном пољу очекивани број није мањи од 1 и у барем 80% поља није мањи од 5.