ANALIZA OPISNIH SPREMENLJIVK

Posvetimo se najprej najbolj preprosti opisni spremenljivki, takšni, ki ima samo dve vrednosti (spol, izid zdravljenja, ki je lahko uspešen ali neuspešen, $\ldots$). Takšnim spremenljivkam rečemo, da so dihotomne.

Recimo, da opazujemo n enot, na katerih opazujemo neko dihotomno spremenljivko, vrednosti, ki ju lahko zavzame, pa označimo kar z A in $\bar{A}$. Privzemimo še, da so izidi neodvisni in da je verjetnost izida A za vsako enoto enaka $\pi = P(A)$.

Če je zdaj X slučajna spremenljivka, katere vrednost je število izidov A pri n enotah, ima X ob zgornjih predpostavkah **binomsko porazdelitev** s parametroma n in π in pišemo $X \sim b(n,\pi)$.

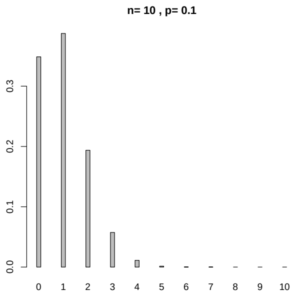
Verjetnost, da X zavzame vrednost k na vzorcu velikosti n, je

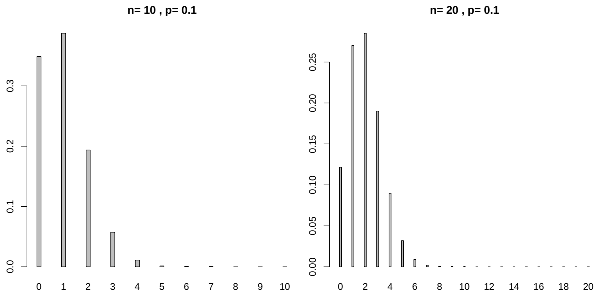
$$P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

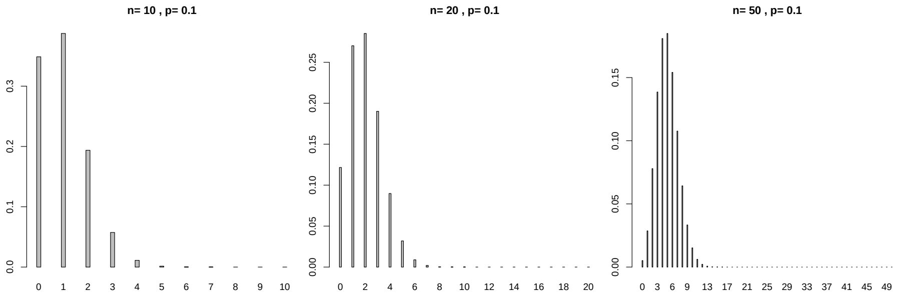
za $k = 0, 1, 2, \ldots, n$. Če je X porazdeljena po binomski porazdelitvi, potem velja

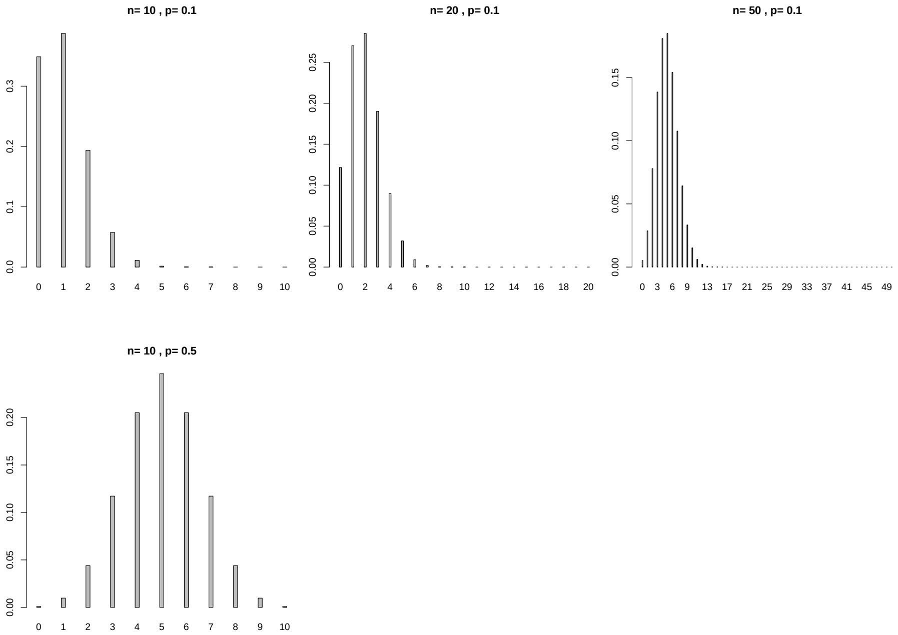
$$E(X) = n \cdot \pi$$

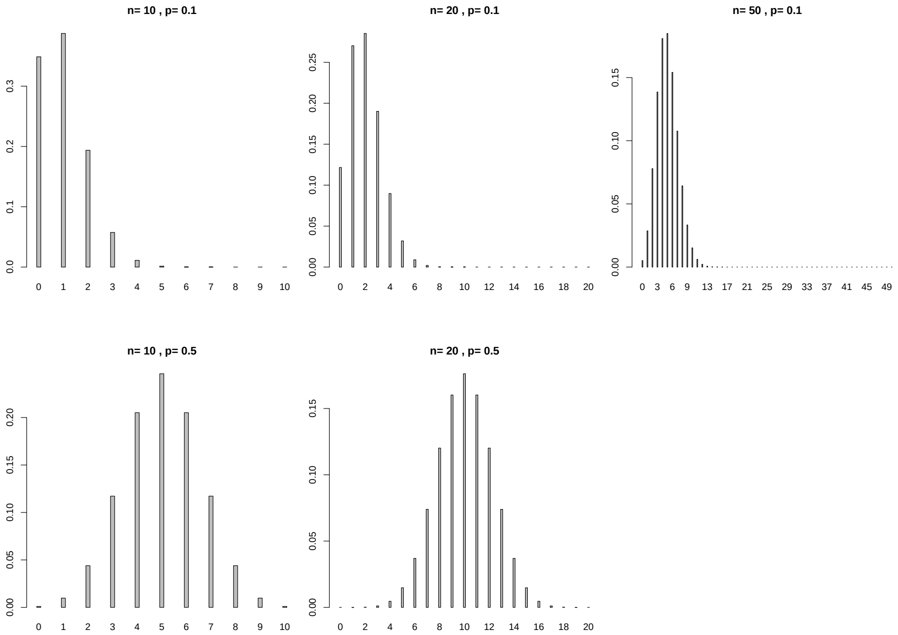
$$Var(X) = n \cdot \pi \cdot (1 - \pi)$$

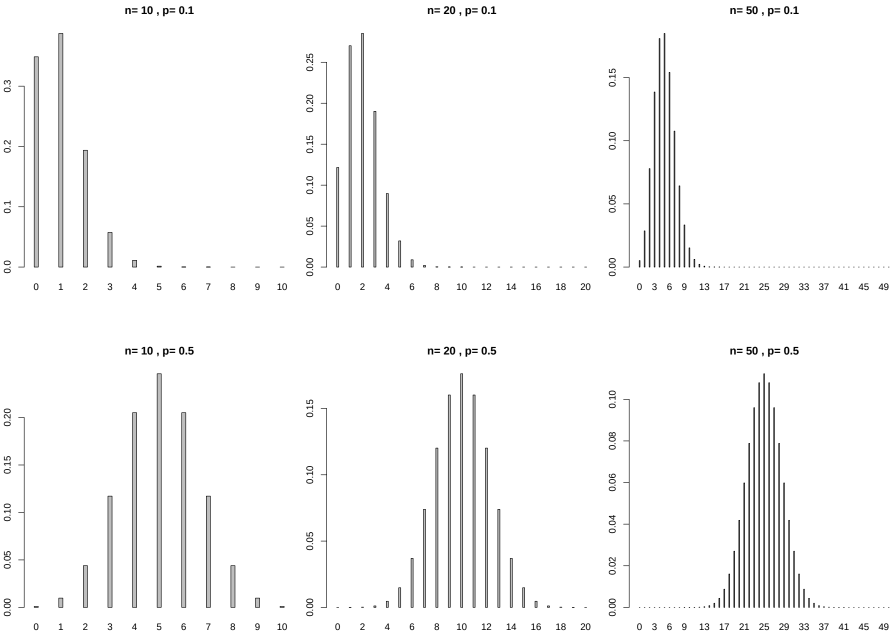

n= 10 , p= 0.1
0.0
0.1
0.2
0.3
0 1 2 3 4 5 6 7 8 9 10

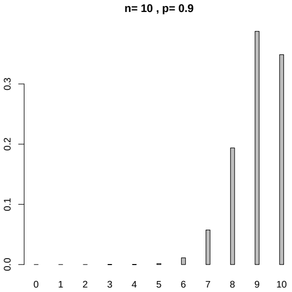

n= 10 , p= 0.1
n= 20 , p= 0.1
0.0
0.1
0.2
0.3
0.00
0.05
0.10
0.15
0.20
0.25
0 1 2 3 4 5 6 7 8 9 10
0 2 4 6 8 10 12 14 16 18 20

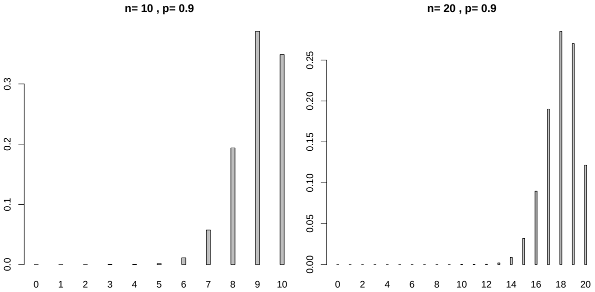

n= 10 , p= 0.1
n= 20 , p= 0.1
n= 50 , p= 0.1
0.0
0.1
0.2
0.3
0 1 2 3 4 5 6 7 8 9 10
0.00
0.05
0.10
0.15
0.20
0.25
0 2 4 6 8 10 12 14 16 18 20
0.00
0.05
0.10
0.15
0 3 6 9 13 17 21 25 29 33 37 41 45 49

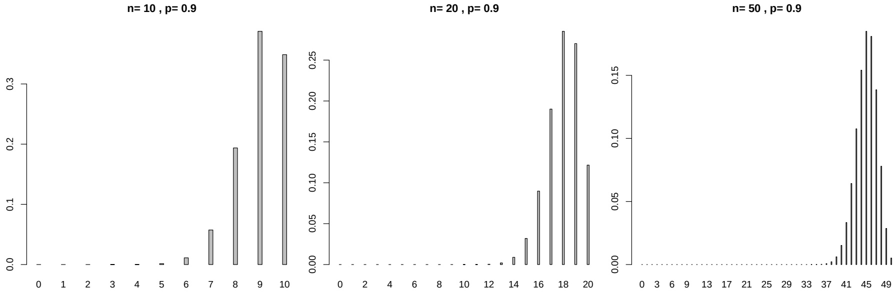

n= 10 , p= 0.1
0.0 0.1 0.2 0.3
0 1 2 3 4 5 6 7 8 9 10
n= 20 , p= 0.1
0.00 0.05 0.10 0.15 0.20 0.25
0 2 4 6 8 10 12 14 16 18 20
n= 50 , p= 0.1
0.00 0.05 0.10 0.15
0 3 6 9 13 17 21 25 29 33 37 41 45 49
n= 10 , p= 0.5
0.00 0.05 0.10 0.15 0.20
0 1 2 3 4 5 6 7 8 9 10


n= 10 , p= 0.1
n= 20 , p= 0.1
n= 50 , p= 0.1
n= 10 , p= 0.5
n= 20 , p= 0.5
0.0 0.1 0.2 0.3
0.00 0.05 0.10 0.15 0.20 0.25
0.00 0.05 0.10 0.15
0.00 0.05 0.10 0.15 0.20
0.00 0.05 0.10 0.15
0 1 2 3 4 5 6 7 8 9 10
0 2 4 6 8 10 12 14 16 18 20
0 3 6 9 13 17 21 25 29 33 37 41 45 49


n= 10 , p= 0.1
n= 20 , p= 0.1
n= 50 , p= 0.1
n= 10 , p= 0.5
n= 20 , p= 0.5
n= 50 , p= 0.5


n= 10 , p= 0.9
0.0
0.1
0.2
0.3
0
1
2
3
4
5
6
7
8
9
10


n= 10 , p= 0.9
0.0
0.1
0.2
0.3
0 1 2 3 4 5 6 7 8 9 10
n= 20 , p= 0.9
0.00
0.05
0.10
0.15
0.20
0.25
0 2 4 6 8 10 12 14 16 18 20



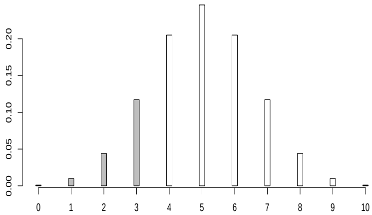
n= 10 , p= 0.9
0.0
0.1
0.2
0.3
0 1 2 3 4 5 6 7 8 9 10
n= 20 , p= 0.9
0.00
0.05
0.10
0.15
0.20
0.25
0 2 4 6 8 10 12 14 16 18 20
n= 50 , p= 0.9
0.00
0.05
0.10
0.15
0 3 6 9 13 17 21 25 29 33 37 41 45 49

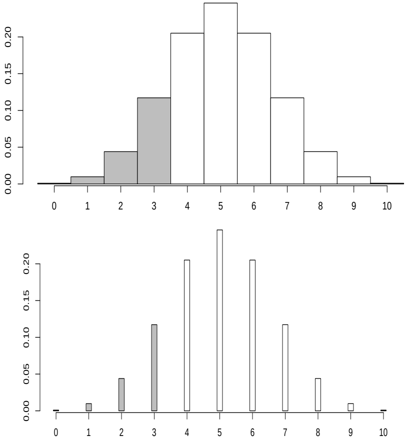
Primer: Naj bo X binomsko porazdeljena s parametroma $n = 10$ in $\pi = 0{,}5$, torej $X \sim b(10, 0{,}5)$. Vprašanje: Koliko je $P(X \leq 3)$?

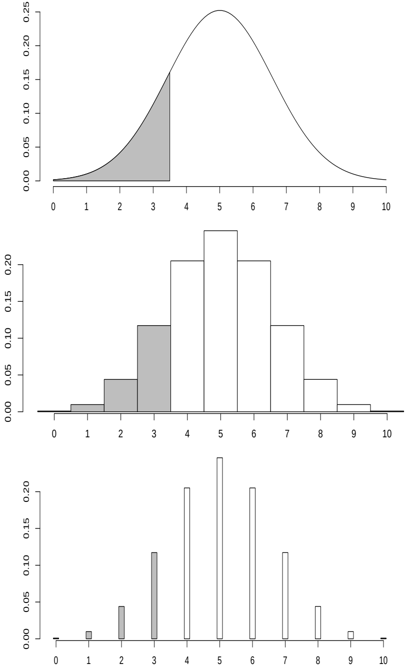
Točen odgovor seveda dobimo takole:

$$P(X \leq 3) = \sum_{k=0}^{3} P(X = k) = 0{,}1719$$

vendar bo tak postopek zelo težaven pri velikih n. Danes imamo sicer ustrezne računalniške programe, še vedno pa pride prav **normalna aproksimacija**.


0.00
0.05
0.10
0.15
0.20
0 1 2 3 4 5 6 7 8 9 10


0.00
0.05
0.10
0.15
0.20
0
1
2
3
4
5
6
7
8
9
10
0.00
0.05
0.10
0.15
0.20
0
1
2
3
4
5
6
7
8
9
10


0.00 0.05 0.10 0.15 0.20 0.25
0 1 2 3 4 5 6 7 8 9 10
0.00 0.05 0.10 0.15 0.20
0 1 2 3 4 5 6 7 8 9 10
0.00 0.05 0.10 0.15 0.20
0 1 2 3 4 5 6 7 8 9 10

Vsota višin (= verjetnosti) prvih štirih stolpcev v stolpičnem diagramu je enaka vsoti ploščin prvih štirih pravokotnikov v histogramu. Ta ploščina pa je približno enaka ploščini pod normalno krivuljo za *x* **manjše od 3,5**.

Približni odgovor torej dobimo takole:

$$E(X) = n \cdot \pi = 5 \quad \text{in} \quad Var(X) = n \cdot \pi \cdot (1 - \pi) = 2{,}5$$

Vzemimo spremenljivko $Y \sim \mathcal{N}(5, 2{,}5)$ in naj bo $Z \sim \mathcal{N}(0{,}1)$. Potem je

$$P(X \leq 3) \approx P(Y \leq 3{,}5) = P\left(Z \leq \frac{3{,}5 - 5}{\sqrt{2{,}5}}\right) = 0{,}1714$$

Normalna aproksimacija ni vedno enako dobra, a dobro deluje, če se le obe vrednosti spremenljivke pojavljata s frekvenco večjo od 5.

Blood Types on a Tropical Island

When we have collected data on only one qualitative variable with several categories, we often are interested in comparing the proportions in these categories with some standard. This is often known as a goodness-of-fit test.

Consider the following example:

We have explored a new tropical island and are interested in the ABO blood types of the natives (more specifically, their ABO phenotypes). It might be interesting to see if their blood-type distribution is similar to the distribution seen in the United States.

Introduction to Chi-Square

Goodness-of-Fit Test

Page 1 of 3

Inference for Qual. Data

Inference Based on Counts

Topics Testing Options

An Electronic Companion to Biostatistics TM

Calculation of Expected Values for Blood-Type Example

	Type O	Type A	Type B	Type AB	Total
Observed counts	125	195	120	60	500
Expected counts (based on U.S. population)	$\frac{4}{10} \times 500 = 200$	$\frac{4}{10} \times 500 = 200$	$\frac{1}{10} \times 500 = 50$	$\frac{1}{10} \times 500 = 50$	500

In our tropical-island example, suppose we find, in a sample of 500 natives, the observed counts shown in the first row of the table on the left.

We know from blood-bank data that the ratio of these blood types in the United States is approximately

$$\pi_O : \pi_A : \pi_B : \pi_{AB} = 4 : 4 : 1 : 1.$$

Do we have evidence that the ratio of blood types on the island is different from that seen in the United States?

We need to compute the expected counts, as shown in the second row of the table.

Introduction to Chi-Square

Goodness-of-Fit Test

Page 2 of 3

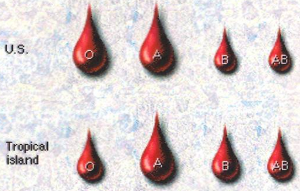
Inference for Qual. Data | Inference Based on Counts | Topics | Testing | Options

An Electronic Companion to Biostatistics TM

Differences in the Distribution of Blood Types Between Tropical Islanders and the U.S. Population

$$\chi^2 = \frac{(125-200)^2}{200} + \frac{(195-200)^2}{200} + \frac{(120-50)^2}{50} + \frac{(60-50)^2}{50}$$

$$\chi^2 = 129.4$$



The island has approximately the same proportion of types A and AB as the U.S., but it has less type O and more type B.

Introduction to Chi-Square

Goodness-of-Fit Test

To complete the blood-type example, we compute χ^2 (chi-square) as $\chi^2 = 129.4$.

To assess the size of the χ^2 statistic, we refer to tables of the distribution of this quantity, called the χ^2 distribution. This, like the t distribution, is indexed by the degrees of freedom. In this case, degrees of freedom $= k - 1 = 3$. From the table, we find that $\chi^2 = 129.4$ is overwhelmingly significant. This is not surprising, considering differences in the two sets of proportions shown in the graph.

To use the chi-square test properly, all expected counts should be > 1 and no more than 20% of expected counts should be < 5. To obtain such expected counts, it may be necessary to pool some categories.

PRIMERJAVA DVEH NEODVISNIH DELEŽEV

Primer: Primerjava dveh načinov zdravljenja

V randomiziranem kliničnem poskusu hočemo primerjati novo zdravljenje A s starim B. Dobili smo naslednje rezultate

Zdravljenje	Število pacientov	Umrlih
A	257	41
B	244	64
Skupaj	501	105

Vprašanje: Ali je tveganje v dveh skupinah različno? Oziroma: Ali drži ničelna hipoteza

$$H_0 : \pi_1 = \pi_2 = \pi$$

Tudi ta problem lahko rešujemo z normalno aproksimacijo k binomski porazdelitvi. Vendar si bomo raje ogledali drugačen pristop, ki ima to prednost, da je zlahka razširljiv na več skupin. Zapišimo podatke v prejšnji tabeli drugače

Ugotovljene frekvence	Izid		
	Smrt	Preživetje	Skupaj
Zdravljenje A	41	216	257
B	64	180	244
Skupaj	105	396	501

Takšni tabeli rečemo **kontingenčna tabela**. V danem primeru imamo $2 \cdot 2 = 4$ kombinacij kategorij dveh spremenljivk, ki določata vrstice in stolpce tabele. V tabeli so **ugotovljene frekvence** pojavljanja teh kombinacij, torej frekvence izidov po skupinah. V zadnji vrstici in zadnjem stolpcu so vsote stolpcev in vrstic, rečemo jim robne vsote.

Če velja ničelna hipoteza $\pi_1 = \pi_2 = \pi$, je število smrti v skupini A porazdeljeno kot $b(257{,}\pi)$. Torej pričakujemo $257 \cdot \pi$ smrti v skupini A. Ker π ne poznamo, ga ocenimo iz podatkov. Ker je med 501 ljudmi umrlo 105, je ta ocena $\hat{\pi} = 105/501$ in pričakovano število smrti potem $257 \cdot \frac{105}{501} = 53{,}862$.

S splošnimi oznakami je naša tabela takšna

Splošne oznake	Izid: Smrt	Izid: Preživetje	Skupaj
Zdravljenje A	a	b	n_1
B	c	d	n_2
Skupaj	m_1	m_2	N

Pričakovano število smrti v skupini A je potem $= \frac{m_1 \cdot n_1}{N}$ in podobno za ostale pričakovane frekvence.

Pričakovane frekvence	Izid		
	Smrt	Preživetje	Skupaj
Zdravljenje A	53,862	203,138	257
B	51,138	192,862	244
Skupaj	105	396	501

Razlika frekvenc	Izid Smrt	Preživetje	Skupaj
Zdravljenje A	$-12{,}862$	$12{,}862$	0
B	$12{,}862$	$-12{,}862$	0
Skupaj	0	0	0

Kot vidimo, se razlike ugotovljenih in pričakovanih frekvenc sicer seštejejo v nič, vendar pa

- večje razlike govorijo proti ničelni hipotezi,
- je dana razlika bolj pomembna, če je pričakovana frekvenca majhna.

Potemtakem se zdi primerna mera ujemanja med ugotovljenimi (f_u) in pričakovanimi (f_p) frekvencami naslednji izraz

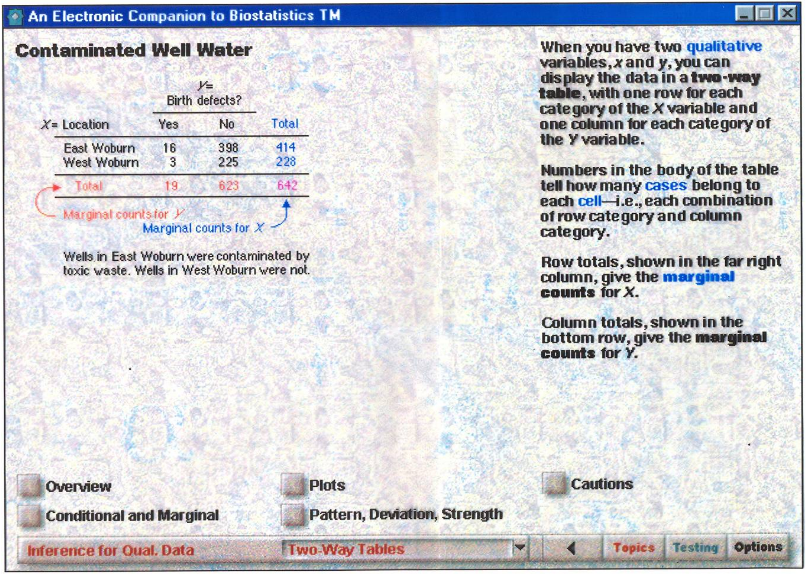
$$X^2 = \sum_{\text{vse celice}} \frac{(f_u - f_p)^2}{f_p} \tag{1}$$

Izkaže se, da je izraz (1) porazdeljen približno kot χ^2 **z eno stopinjo prostosti**. Približek je boljši pri večjih pričakovanih frekvencah, za zadovoljiv približek pa je potrebno, da so vse pričakovane frekvence večje od 5.

V našem primeru so vse pričakovane frekvence dovolj velike, torej po (1) izračunamo

$$\begin{aligned} X^2 &= \frac{(-12{,}862)^2}{53{,}862} + \frac{12{,}862^2}{203{,}138} + \frac{(-12{,}862)^2}{51{,}138} + \frac{12{,}862^2}{192{,}862} \\ &= 7{,}978 \end{aligned}$$

Verjetnost, da bi bila vrednost χ^2 pri eni stopinji prostosti večja ali enaka dobljeni vrednosti, je 0,0047. Ničelno hipotezo torej z lahkim srcem zavrnemo.



An Electronic Companion to Biostatistics TM

Contaminated Well Water

X= Location	Y= Birth defects? Yes	No	Total
East Woburn	16	398	414
West Woburn	3	225	228
Total	19	623	642

Marginal counts for Y

Marginal counts for X

Wells in East Woburn were contaminated by toxic waste. Wells in West Woburn were not.

When you have two qualitative variables, x and y, you can display the data in a **two-way table**, with one row for each category of the X variable and one column for each category of the Y variable.

Numbers in the body of the table tell how many cases belong to each cell—i.e., each combination of row category and column category.

Row totals, shown in the far right column, give the **marginal counts** for X.

Column totals, shown in the bottom row, give the **marginal counts** for Y.

Overview

Plots

Cautions

Conditional and Marginal

Pattern, Deviation, Strength

Inference for Qual. Data

Two-Way Tables

Topics

Testing

Options

An Electronic Companion to Biostatistics TM

Race and the Death Penalty

Counts

Race of: Accused	Race of: Victim	Accused convicted of first-degree murder? Yes — Death sentence? Yes	Accused convicted of first-degree murder? Yes — Death sentence? No	Accused convicted of first-degree murder? No	Total
Black	White	39	44	110	193
White	White	49	120	288	457
Black	Black	11	45	232	288
White	Black	0	3	17	20
Total		99	212	647	958

Percentages

Race of: Accused	Race of: Victim	Accused convicted of first-degree murder? Yes — Death sentence? Yes	Accused convicted of first-degree murder? Yes — Death sentence? No	Accused convicted of first-degree murder? No	Total
Black	White	20	23	57	100
White	White	11	26	63	100
Black	Black	4	16	80	100
White	Black	0	15	85	100
All combined		10	22	68	100

In 1987, the U.S. Supreme Court heard arguments that blacks convicted of murder are more likely than whites to be sentenced to death. The data here are for the state of Florida. Rows (predictor X) tell the race of the accused and of the victim. Columns (response Y) tell the outcome of the case.

Conditional vs. Marginal: The bottom row gives the marginal distribution for the outcome Y. The other four rows give the conditional distributions for Y, one distribution for each category of the X variable.

Counts vs. Percentages: Conditional distributions are easier to compare if we convert counts to row percentages. For example, the 110 in the upper right corner becomes $(110/193) \times 100\% = 57\%$.

Overview | Plots | Cautions

Conditional and Marginal | Pattern, Deviation, Strength

Inference for Qual. Data | Two-Way Tables | Topics | Testing | Options

Mere povezanosti

Ena možnost za opis razlik v deležih med dvema vzorcema je njuna **razlika**

$$RD = \pi_1 - \pi_2$$

Razliko tipično uporabimo pri testiranju ničelne hipoteze, redkeje pa kot opisno statistiko. Razlog je v tem, da so populacijske verjetnosti pojavljanja bolezni praviloma zelo majhne in zato razlike manj dramatične. Veliko bolj pogosto se uporablja **relativno tveganje**

$$RR = \frac{\pi_1}{\pi_2}$$

Primer: naj bo tveganje za pljučnega raka med nekadilci 0,001, med kadilci pa 0,009. Razlika tveganj je 0,008 (enako kot med 0,419 in 0,411), relativno tveganje pa 9!

Primer: Uganite od kod so ti podatki!

Primer: Uganite od kod so ti podatki!

Spol	Umrlo	Preživelo	Tveganje
moški	1364	367	$1364/1731 = 0{,}79$
ženske	126	344	$126/470 = 0{,}27$

Primer: Uganite od kod so ti podatki!

Spol	Umrlo	Preživelo	Tveganje
moški	1364	367	$1364/1731 = 0{,}79$
ženske	126	344	$126/470 = 0{,}27$

Starostna skupina	Umrlo	Preživelo	Tveganje
otroci	52	57	$52/109 = 0{,}48$
odrasli	1438	654	$1438/2092 = 0{,}69$

Primer: Uganite od kod so ti podatki!

Spol	Umrlo	Preživelo	Tveganje
moški	1364	367	$1364/1731 = 0{,}79$
ženske	126	344	$126/470 = 0{,}27$

Starostna skupina	Umrlo	Preživelo	Tveganje
otroci	52	57	$52/109 = 0{,}48$
odrasli	1438	654	$1438/2092 = 0{,}69$

Potovalni razred	Umrlo	Preživelo	Tveganje
posadka	673	212	$673/885 = 0{,}76$
1.	122	203	$122/325 = 0{,}38$
2.	167	118	$167/285 = 0{,}59$
3.	528	178	$528/706 = 0{,}75$

Obeti

Obeti so razmerje med verjetnostjo, da se nek dogodek zgodi in verjetnostjo, da se ne zgodi. Torej

$$\text{obeti} = \frac{\pi}{1 - \pi}$$

Primer: če je dogodek smrt in njegova verjetnost 0,75, so obeti enaki 3, ker je verjetnost smrti trikrat večja od verjetnosti preživetja.

Primer: Titanic

Obeti

Obeti so razmerje med verjetnostjo, da se nek dogodek zgodi in verjetnostjo, da se ne zgodi. Torej

$$\text{obeti} = \frac{\pi}{1-\pi}$$

Primer: če je dogodek smrt in njegova verjetnost 0,75, so obeti enaki 3, ker je verjetnost smrti trikrat večja od verjetnosti preživetja.

Primer: Titanic

Spol	π	$1-\pi$	Obeti
moški	0,79	0,21	3,76
ženske	0,27	0,73	0,37

Razmerje obetov

V prejšnjem primeru smo izračunali obete posebej pri moških in posebej pri ženskah. Če bi bilo tveganje v obeh skupinah enako, bi bili enaki tudi obeti. Primerjava (kvocient) bi torej utegnila imeti smisel.

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}$$

Vendar, če že imamo relativno tveganje, zakaj bi človek računal še razmerje obetov?

Relativno tveganje in razmerje obetov

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\pi_1}{\pi_2} \cdot \frac{1-\pi_2}{1-\pi_1} = RR \cdot \frac{1-\pi_2}{1-\pi_1}$$

Zanimivo, vendar ali tudi uporabno?

Nadaljujmo s primerom Titanica in smrti po spolu. V tabeli vpeljimo splošne oznake

	Izid		
Spol	Smrt	Preživetje	Skupaj
Moški	$n_{11} = 1364$	$n_{12} = 367$	$n_{1+} = 1731$
Ženske	$n_{21} = 126$	$n_{22} = 344$	$n_{2+} = 470$
Skupaj	$n_{+1} = 1490$	$n_{+2} = 711$	$n = 2201$

S temi oznakami je razmerje obetov

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\frac{n_{11}/n_{1+}}{n_{12}/n_{1+}}}{\frac{n_{21}/n_{2+}}{n_{22}/n_{2+}}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

S temi oznakami je razmerje obetov

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\frac{n_{11}/n_{1+}}{n_{12}/n_{1+}}}{\frac{n_{21}/n_{2+}}{n_{22}/n_{2+}}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Razmerje obetov torej lahko izračunamo tako, da navzkrižno množimo frekvence v tabeli in zmnožka delimo. To pa ni le računska ugodnost. Postavimo si sedaj tole vprašanje: če nas zanimajo le mrtvi, kolikšna je verjetnost (tveganje), da so moški? Dobimo seveda n_{11}/n_{+1} in podobno med preživelimi n_{12}/n_{+2}. Ustrezni obeti so potem n_{11}/n_{21} pri mrtvih in n_{12}/n_{22} pri živih. Razmerje obetov (**za biti moški!**) je potem

S temi oznakami je razmerje obetov

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\frac{n_{11}/n_{1+}}{n_{12}/n_{1+}}}{\frac{n_{21}/n_{2+}}{n_{22}/n_{2+}}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Razmerje obetov torej lahko izračunamo tako, da navzkrižno množimo frekvence v tabeli in zmnožka delimo. To pa ni le računska ugodnost. Postavimo si sedaj tole vprašanje: če nas zanimajo le mrtvi, kolikšna je verjetnost (tveganje), da so moški? Dobimo seveda n_{11}/n_{+1} in podobno med preživelimi n_{12}/n_{+2}. Ustrezni obeti so potem n_{11}/n_{21} pri mrtvih in n_{12}/n_{22} pri živih. Razmerje obetov (**za biti moški!**) je potem

$$OR_m = \frac{\frac{n_{11}}{n_{21}}}{\frac{n_{12}}{n_{22}}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

S temi oznakami je razmerje obetov

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\frac{n_{11}/n_{1+}}{n_{12}/n_{1+}}}{\frac{n_{21}/n_{2+}}{n_{22}/n_{2+}}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Razmerje obetov torej lahko izračunamo tako, da navzkrižno množimo frekvence v tabeli in zmnožka delimo. To pa ni le računska ugodnost. Postavimo si sedaj tole vprašanje: če nas zanimajo le mrtvi, kolikšna je verjetnost (tveganje), da so moški? Dobimo seveda n_{11}/n_{+1} in podobno med preživelimi n_{12}/n_{+2}. Ustrezni obeti so potem n_{11}/n_{21} pri mrtvih in n_{12}/n_{22} pri živih. Razmerje obetov (**za biti moški!**) je potem

$$OR_m = \frac{\frac{n_{11}}{n_{21}}}{\frac{n_{12}}{n_{22}}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

!!!