Univerza v Ljubljani
Fakulteta za farmacijo

Statistična analiza opisnih spremenljivk

Farmacevtska informatika
2011/2012, 1. letnik EMŠF

Doc. dr. Igor Locatelli, *mag. farm.*

Ljubljana, 20. 4. 2012

Osnovni pojmi – opisne (atributivne) spremenljivke

- Razdelitev glede na število kategorij (skupin)
  - Dihotomne ali binarne spremenljivke, zajemajo samo dve vrednosti oz. kategoriji;
    npr. spol (M ali Ž), preživetje (živ ali mrtev).
  - Politomne spremenljivke; imajo več kategorij;
    npr. genotip *CYP2C9,* barva las, opisna ocena.
- Razdelitev glede na urejenost v zaporedje
  - Nominalne spremenljivke, niso urejene po logičnem zaporedju;
    npr. krvna skupina (A, B, AB, 0).
  - Ordinalne spremenljivke, so urejene v zaporedje;
    npr. stopnja bolečine (brez, blaga, zmerna, huda, zelo huda).
  - Starost?

Urejanje opisnih spremenljivk

- Združevanje enot v skupine – kategorije:
  - Določitev števila enot v posamezni kategoriji (frekvenca)
- Spremenljivke z maloštevilnimi vrednostmi:
  enostavna razmejitev v kategorije
  Npr: spol, zakonski stan, krvne skupine (A, B, AB, 0),
  genotip *CYP2C9*.
- Spremenljivke z veliko vrednostmi in nejasnimi mejami
  Npr. barva las, barva oči
- Klasifikacije
  - Mednarodna klasifikacija bolezni (MKB),
  - anatomsko-terapevtska-kemična klasifikacija zdravil (ATC)

Statistična analiza opisnih spremenljivk

- Z-test za deleže (aproksimacija binomske porazdelitve)
- Chi-kvadrat test oz. χ^2 test
  - Kontingenčne tabele
  - Fisherjev natančni test
  - Razmerje obetov
- Dva odvisna vzorca: McNemarjev test
- Chi-kvadrat test za testiranje normalnosti porazdelitve oz. kakovosti prileganja

Z-test za deleže (Two-sample test for binominal proportions)

- Proučevanje povezave med pojavnostjo raka dojk in starostjo mater pri prvem porodu.
- Rak dojk (prisoten ali ne)
- Starost mater (≤29 ali ≥30)
- 683 od 3220 mater z rakom dojk in 1498 od 10245 mater v kontrolni skupini je bilo od prvem porodu starih =>30.

Verjetnosti..p

- p_1 = verjetnost za starost nad 30 let v skupini mater z rakom dojk
  - $p_1 = 683/3220 = 0{,}212$
- p_2 = verjetnost za starost nad 30 let v kontrolni skupini
  - $p_2 = 1498/10245 = 0{,}146$
- H_0: $p_1 = p_2 = p$
- H_A: $p_1 \neq p_2$
- Skupna verjetnost (p)

$$\mathrm{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{683+1498}{3220+10245} = 0{,}162$$

Testna statistika – z

- Če je $n_1p(1\text{-}p)\geq 5$ in $n_2p(1\text{-}p)\geq 5$, potem lahko uporabimo normalno aprosimacijo (z-test) binomski porazdelitvi
  - $n_1p(1\text{-}p) = 3220\cdot 0{,}162\cdot 0{,}838 = 437 \geq 5$
- $$\mathrm{z} = \frac{|p1 - p2| - \left(\frac{1}{2n_1}+\frac{1}{2n_2}\right)}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$
- z = 8,8 ali p<0,001 oz. $p \sim 1x10^{-18}$
- H_0 zavržemo in sprejmemo H_A

Chi-kvadrat test - kontingenčna tabela

Primer predstavljen v obliki kontingenčne tabele 2x2 z opazovanimi frekvencami.

Stanje	**Starost pri prvem porodu**		*Skupaj*
	≤29	**≥30**	
Rak dojk	2537	683	*3220*
Kontrola	8747	1498	*10245*
Skupaj	*11284*	*2181*	*13465*

Robne vrednosti stolpcev in vrstic.

Chi-kvadrat test kontingenčna tabela: deleži

stanje * starost Crosstabulation				
		starost		Total
		≤29	**≥30**	
rak dojk	Count	2537	683	3220
	% within stanje	78,8%	21,2%	100,0%
	% within starost	22,5%	31,3%	23,9%
	% of Total	18,8%	5,1%	23,9%
kontrola	Count	8747	1498	10245
	% within stanje	85,4%	14,6%	100,0%
	% within starost	77,5%	68,7%	76,1%
	% of Total	65,0%	11,1%	76,1%
Total	*Count*	*11284*	*2181*	13465
	% within stanje	*83,8%*	*16,2%*	100,0%
	% within starost	*100,0%*	*100,0%*	100,0%
	% of Total	*83,8%*	*16,2%*	100,0%

Chi-kvadrat test - kontingenčna tabela: pričakovane frekvence

Stanje	**Starost pri prvem porodu**		*Skupaj*
	≤29	**≥30**	
Rak dojk	2537 n_1-x_1	683 x_1	*3220* n_1
Kontrola	8747 n_2-x_2	1498 x_2	*10245* n_2
Skupaj	*11284*	*2181* x_1+x_2	*13465* n_1+n_2

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

p_1 = 683/3220 = 0,212

p_2 = 1498/10245 = 0,146

$$E_{x1} = n_1 p = \frac{n_1(x_1 + x_2)}{n_1 + n_2} = 3220\text{x}2181/13465 = 521{,}6$$

$$E_{x2} = n_2 p = \frac{n_2(x_1 + x_2)}{n_1 + n_2} = 10245\text{x}2181/13465 = 1659{,}4$$

Chi-kvadrat test - kontingenčna tabela: pričakovane frekvence

Pričakovane frekvence -> Expected count

stanje * starost Crosstabulation				
		starost		Total
		≤29	**≥30**	
rak dojk	Count	2537	683	3220
	Expected Count	**2698,4**	**521,6**	
kontrola	Count	8747	1498	10245
	Expected Count	**8585,6**	**1659,4**	
	Total	11284	2181	13465

Chi-kvadrat test - ničelna in alternativna hipoteza

- Ničelna in alternativna hipoteza:
  - H_0: spremenljivki sta neodvisni ali $p_1 = p_2$;
    pričakovane frekvence (f_p) so enake opazovanim (f_o)
  - H_A spremenljivki sta odvisni ali $p_1 \neq p_2$;
    pričakovane frekvence (f_p) niso enake opazovanim (f_o)
- Izračunamo ekperimentalni chi-kvadrat oz. Pearsonov chi-kvadrat:

$$\chi^2_{\mathrm{exp}} = \sum_{i=1}^{k} \frac{(|f_o - f_p|)^2}{f_p}$$

2x2 kontingenčna tabela

Yatesova korektura ali
Yates‘s continuity correction

- k….število celic
  4 za 2x2, 6 za 2x3, itd.

Chi-kvadrat test - Yatesova korektura

- Namen je popraviti preveliko oceno, ki jo v primeru 2x2 kontingenčne tabele naredimo s Pearsonovim chi-kvadratom.
- Protiargument:
  - Popravek je prevelik in zato morda ni smiseln.

$$\chi^2_{\exp,corr} = \sum_{i=1}^{k} \frac{(\left| f_o - f_p \right| - 0{,}5)^2}{f_p}$$

Chi-kvadrat test - izračun s korekturo

$$X^2 = \frac{(|683 - 521.6| - .5)^2}{521.6} + \frac{(|2537 - 2698.4| - .5)^2}{2698.4}$$

$$+ \frac{(|1498 - 1659.4| - .5)^2}{1659.4} + \frac{(|8747 - 8585.6| - .5)^2}{8585.6}$$

$$= \frac{160.9^2}{521.6} + \frac{160.9^2}{2698.4} + \frac{160.9^2}{1659.4} + \frac{160.9^2}{8585.6}$$

$$= 49.661 + 9.599 + 15.608 + 3.017 = 77.89 \sim \chi_1^2 \text{ under } H_0$$

Chi-kvadrat test - tabelarični chi-kvadrat

χ^2 CRITICAL VALUES

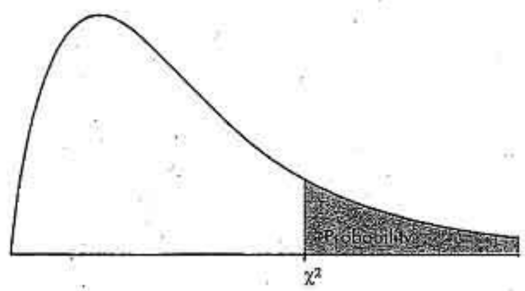


TABLE C: χ^2 CRITICAL VALUES

df	Tail probability p .25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	[illegible]	[illegible]

$$\chi^2_{tab(df=1;\alpha=0,05)} = 3,84$$

Stopinje prostosti:

df= (s-1)(v-1)

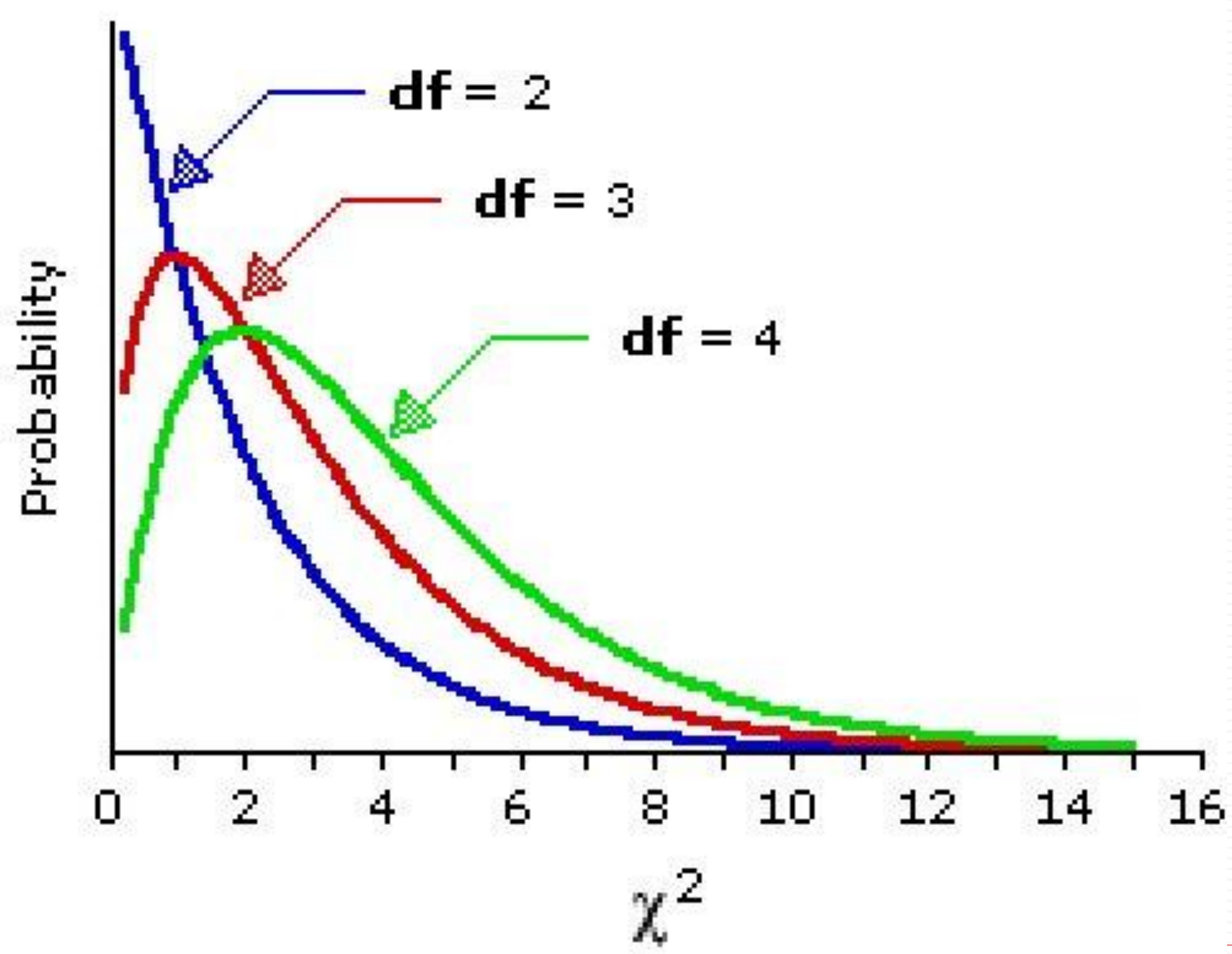
s: število stolpcev
v: število vrstic

Enostranski ali dvostranski?

Glej na zgornji meji

df	.995	…	.95		.05	.025	.01
1	...	…	0.004	⋯	**3.841**	5.024	6.635
2	0.010	⋯	0.103	⋯	5.991	7.378	9.210

Hi kvadrat porazdelitev



Stopinje prostosti:

df= (s-1)(v-1)

s: število stolpcev
v: število vrstic

Chi-kvadrat test - sklep

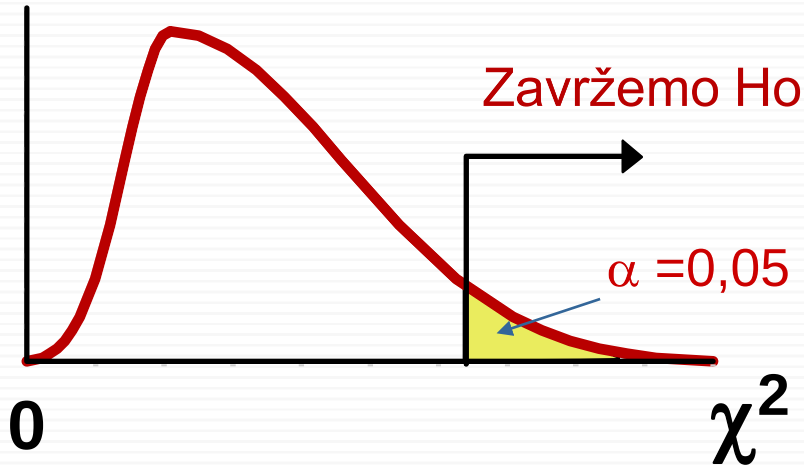
H_0: med spremenljivkama ni povezave

H_A: med spremenljivkama je povezava

$$\chi^2_{\exp} = \sum_{i=1}^{k} \frac{(\left| f_o - f_p \right| - 0,5)^2}{f_p}$$

$$\chi^2_{\exp} = 77,89$$

$$\chi^2_{tab(df=1;\alpha=0,05)} = 3,84$$



$$\chi^2_{\exp} > \chi^2_{tab}$$

p < α; α=0,05

H_0 zavržemo –> H_A sprejmemo

pojavnost raka dojk je statistično značilno povezana s starostjo matere pri prvem porodu

Chi-kvadrat test - SPPS output

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	78,370[a]	1	8,54E-19		
Continuity Correction[b]	77,885	1	1,09E-18		
Likelihood Ratio	74,604	1	5,75E-18		
Fisher's Exact Test				2,89E-18	5,48E-19
N of Valid Cases	13465				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 521,56.

b. Computed only for a 2x2 table.

Chi-kvadrat test - razmerje verjetij; likelihood ratio

Tudi razmerje verjetij se porazdeljuje po Chi-kvadrat statistiki, zato bodo tabelarične vrednosti enake. Eksperimentalne pa:

$$L\chi^2 = 2\left[\sum_{i=1}^{k} f_o \ln \frac{f_o}{f_p}\right]$$

Tak način je primeren za majhne vzorce.

Chi-kvadrat test

- Uporaben pri testiranju povezanosti dveh kategoričnih spremenljivk (npr. podatki podanih v kontingenčni tabeli: RxC)
- Ni uporaben pri ponavljajočih se načrtih: statistična enota (npr. pacient) se lahko pojavi samo enkrat.
- Pričakovane frekvence ne smejo biti nižje kot 5 (rešitev združevanje kategorij ali uporaba Fisherjevega natančnega testa)
  - Če je kontingenčna tabela večja, potem je ta omejitev omejena na 80% celic, pri ostalih mora biti pričakovana frekvenca vsaj nad 1.

Kontingenčna tabela 2x2

- Ali kava iz avtomata ohranja študente budne?
- Opazovane frekvence:

Placebo

	Kava brez kofeina	**Kava s kofeinom**	
Budni	2	23	25
Nebudni	5	30	35
	7	53	60

Kontingenčna tabela 2x2

- Pričakovane frekvence, če ni učinka:

	Placebo **Kava brez kofeina**	Placebo **Kava s kofeinom**	
Budni	?	?	25
Nebudni	?	?	35
	7	53	60

Kontingenčna tabela 2x2

- Pričakovane frekvence, če ni povezave (učinka):

	Placebo **Kava brez kofeina**	Placebo **Kava s kofeinom**	
Budni	$\frac{25 \times 7}{60} = 2,92$	$\frac{25 \times 53}{60} = 22,08$	25
Nebudni	$\frac{35 \times 7}{60} = 4,08$	$\frac{35 \times 53}{60} = 30,92$	35
	7	53	60

Pričakovane frekvence z vrednostjo manjšo od 5!

Kateri test uporabiti?

Fisherjev natančni test verjetnosti

- $p_1 = 2/7 = 0{,}29$ verjetnost, da ostanemo budni, če spijemo placebo
- $p_2 = 23/53 = 0{,}43$ verjetnost da ostanemo budni, če spijemo kavo
- H_0: $p_1 = p_2 = p$
- H_A: $p_1 \neq p_2$

	Kava brez kofeina	Kava s kofeinom	Skupaj
Budni	a	b	a+b
Nebudni	c	d	c+d
Skupaj	a+c	b+d	n

- $p_{(a,b,c,d)} = \dfrac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$
- Postopek: spreminjamo frekvenco v polju a od 0 od vrednosti a, pri čemer fiksiramo robne vrednosti

Fisherjev natančni test verjetnosti

- Tabelo preuredimo tako da v celico (1,1) imamo najmanjšo frekvenco.

Enumeration of all possible tables with fixed margins and their associated probabilities, based on the hypergeometric distribution for Example 10.19

0	25
7	28

.017

1	24
6	29

.105

2	23
5	30

.252

3	22
4	31

.312

4	21
3	32

.214

5	20
2	33

.082

6	19
1	34

.016

7	18
0	35

.001

The question now is: What should be done with these probabilities to evaluate the significance of the results? The answer depends on whether a one-sided or a two-sided alternative is being used. In general, the following method can be used.

Fisherjev natančni test verjetnosti

- Tabelo preuredimo tako da v celico (1,1) imamo najmanjšo frekvenco.

Fisher's Exact Test: General Procedure and Computation of *p*-Value

To test the hypothesis H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$, where the expected value of at least one cell is <5 when the data are analyzed in the form of a 2×2 contingency table, use the following procedure:

(1) Enumerate all possible tables with the same row and column margins as the observed table, as shown in Equation 10.10.

(2) Compute the exact probability of each table enumerated in step 1, using either the computer or the formula in Equation 10.7.

(3) Suppose the observed table is the *a* table and the last table enumerated is the *k* table.

(a) To test the hypothesis H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$, the *p*-value = $2 \times \min[Pr(0) + Pr(1) + \ldots + Pr(a), Pr(a) + Pr(a+1) + \ldots + Pr(k), .5]$.

(b) To test the hypothesis H_0: $p_1 = p_2$ vs. H_1: $p_1 < p_2$, the *p*-value = $Pr(0) + Pr(1) + \ldots + Pr(a)$.

Fisherjev natančni test verjetnosti končni izračun

- Vsota verjetnosti na levi strani od dejanskih frekvenc (p(2)):
  - p(0)+p(1)+p(2)=0,017+0,105+0,252=0,375
- Vsota verjetnosti na desni strani od dejanskih frekvenc (p(2)):
  - p(2)+p(3)+p(4)+p(5)+p(6)+p(7)= =0,252+0,312+0,214+0,082+0,016+0,001=0,878
- p-vrednost testa = 2 x min(0,375 in 0,878) = 2 x 0,375 = 0,750
- Ho ne moremo zavreči.

Vpliv števila oseb na rezultat

- Ali kava iz avtomata ohranja študente budne?
- Opazovane frekvence:

Placebo

	Kava brez kofeina	**Kava s kofeinom**	
Budni	20	230	250
Nebudni	50	300	350
	70	530	600

Primer 10x večjega vzorca

- p_1 = 20/70=0,29 verjetnost, da ostanemo budni, če spijemo placebo.
- p_2 = 230/530=0,43 verjetnost da ostanemo budni, če spijemo kavo.

$$H_0: p_1 = p_2 = p$$

$$H_A: p_1 \neq p_2$$

Opazovane frekvence:

	Kava brez kofeina	**Kava s kofeinom**	
Budni	20	230	250
Nebudni	50	300	350
	70	530	600

Primer 10x večjega vzorca

- H_0: spremenljivki sta neodvisni ali $p_1 = p_2$; pričakovane frekvence so enake opazovanim
- H_A spremenljivki sta odvisni ali $p_1 \neq p_2$; pričakovane frekvence niso enake opazovanim.

stanje * napitek Crosstabulation

			napitek		Total
			placebo	kava	
stanje	budni	Count	20	230	250
		Expected Count	29,2	220,8	
	zaspali	Count	50	300	350
		Expected Count	40,8	309,2	
Total		Count	70	530	600

Primer 10x večjega vzorca

- H_0: spremenljivki sta neodvisni ali $p_1 = p_2$; pričakovane frekvence so enake opazovanim
- H_A spremenljivki sta odvisni ali $p_1 \neq p_2$; pričakovane frekvence niso enake opazovanim.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5,591[a]	1	,018		
Continuity Correction[b]	4,998	1	,025		
Likelihood Ratio	5,811	1	,016		
Fisher's Exact Test				,020	,012

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 29,17.

b. Computed only for a 2x2 table

$p < \alpha$; $\alpha=0{,}05$

H_0 zavržemo -> H_A sprejmemo

Zaspanost je statistično značilno povezana s izbiro napitka.

McNemarjev test

- Testiranje povezanosti dveh opisnih spremenljivk, pri ponavljajočih se poskusih (repeated measure experiments)
  - Neko opisno spremenljivko spremljamo v dveh časovnih točkah.
  - Neko opisno spremenljivko merimo na dva načina (npr. primerjava dveh diagnostičnih testov).

McNemarjev test: procedura

Table 24.2 Observed frequencies of pairs in which the characteristic is present or absent.

Circumstance 2	Circumstance 1 Present	Absent	Total no. of pairs
Present	w	x	$w + x$
Absent	y	z	$y + z$
Total	$w + y$	$x + z$	$m = w + x + y + z$

1 Define the null and alternative hypotheses under study

H_0: the proportions with the characteristic are equal in the two groups in the population

H_1: these population proportions are not equal.

2 Collect relevant data from two samples

3 Calculate the value of the test statistic specific to H_0

$$\chi^2 = \frac{(|x - y| - 1)^2}{x + y}$$

which follows the Chi-squared distribution with 1 degree of freedom.

McNemarjev test: karies

1 H_0: the two methods of assessment identify the same percentage of teeth with cavities in the population

H_1: these percentages are not equal.

2 The frequencies for the matched pairs are displayed in the table:

	Radiographic diagnosis		
Diagnosis on section	Cavities absent	Cavities present	Total
Cavities absent	45	4	49
Cavities present	17	34	51
Total	62	38	100

3 Test statistic, $\chi^2 = \frac{(|17-4|-1)^2}{17+4} = 6.86$

McNemarjev test: karies

1 H_0: the two methods of assessment identify the same percentage of teeth with cavities in the population
H_1: these percentages are not equal.
2 The frequencies for the matched pairs are displayed in the table:

	Radiographic diagnosis		
Diagnosis on section	Cavities absent	Cavities present	Total
Cavities absent	45	4	49
Cavities present	17	34	51
Total	62	38	100

3 Test statistic, $\chi^2 = \frac{(|17-4|-1)^2}{17+4} = 6.86$

Chi-kvadrat (tab) = 3.84 in je manjši od eksperimentalnega
$p < \alpha$; α=0,05
H_0 zavržemo –> H_A sprejmemo
Metodi ne dajeta enakih rezultatov, z radiografsko metoda vedno ne odkrijemo kariesa, kadar je ta prisoten.

Prilagajanje normalne porazdelitve empirični porazdelitvi

- Normalno porazdelitev (z vsemi karakteristikami) prilagoditi tako, da sta povprečna vrednost in standardni odklon enaka tistima, ki ju ima empirična porazdelitev
- Še vedno veljajo omejitve Chi-kvadrat testa:
  - Pričakovane frekvence ne smejo biti nižje kot 5

Testiranje normalnosti

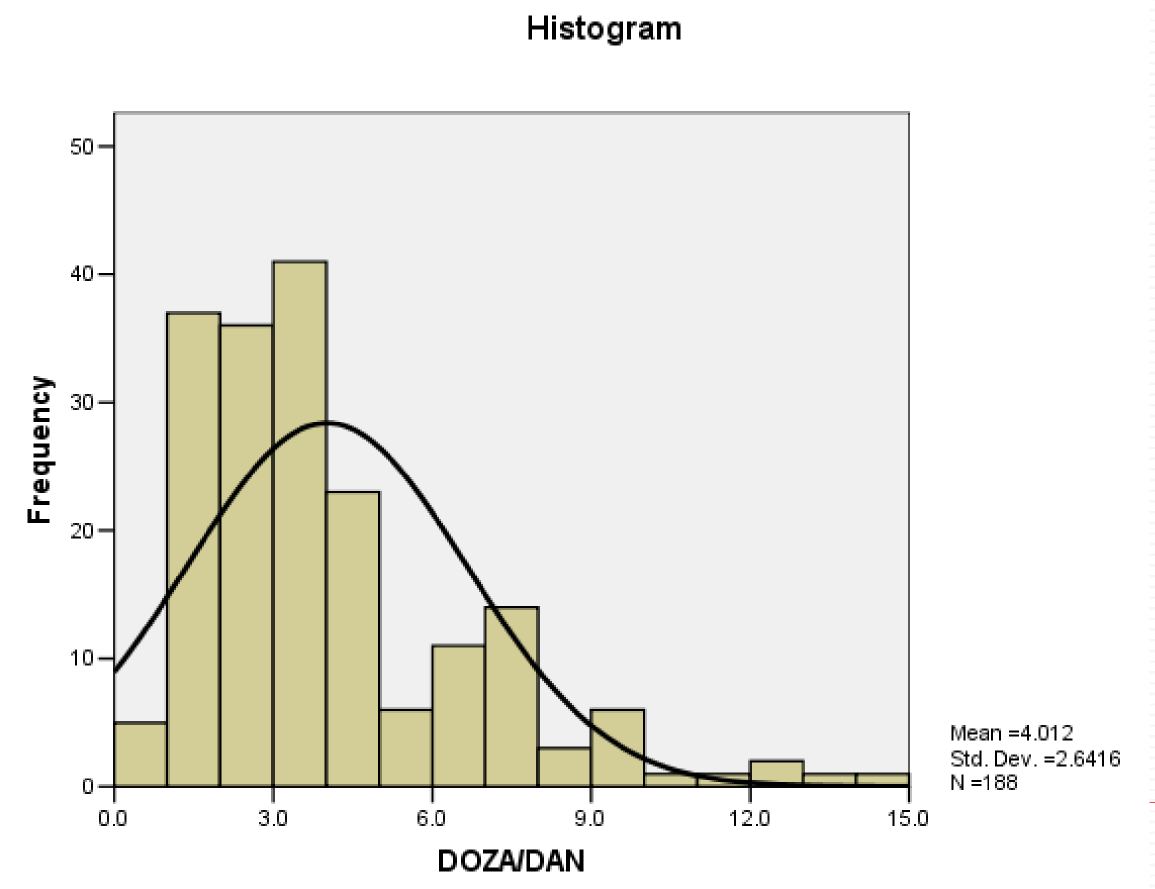
- Normalno porazdelitev prilagajamo frekvenčni
- Test temelji na Chi-kvadrat statistiki

$$\chi^2_{\exp} = \sum_{i=1}^{k} \frac{(f_o - f_p)^2}{f_p}$$

- χ^2_{tab} iz tabel, stopinje prostosti: m= k-r-1
  - k= št. razredov
  - r= št. preizkušanih parametrov porazdelitve (r=2, povprečna vrednost in standardni odklon)

Primer s podatki o odmerkih varfarina

H_o: frekvenčna porazdelitev se porazdeljuje normalno
H_A: frekvenčna porazdelitev se ne porazdeljuje normalno



spodnja meja	zgornja meja	opazovane frekv. (f_o)	z-vrednosti	kumulat. delež	delež	pričakov. frekv. (f_p)	$(f_o - f_p)^2$	$(f_o - f_p)^2/f_p$
0	1	5	-1,14	12,7%	12,7%	23,9	357,0	14,9
1	2	37	-0,76	22,3%	9,6%	18,1	358,9	19,9
2	3	36	-0,38	35,1%	12,8%	24,0	143,9	6,0
3	4	41	0,00	49,8%	14,7%	27,7	176,8	6,4
4	5	23	0,37	64,6%	14,8%	27,8	22,6	0,8
5	6	6	0,75	77,4%	12,8%	24,1	328,6	13,6
6	7	11	1,13	87,1%	9,7%	18,2	52,0	2,9
7	8	14	1,51	93,4%	6,3%	11,9	4,3	0,4
8	15	15	4,16	100,0%	6,6%	12,3	7,2	0,6

povprečje	4,012
stdev	2,6416
n	188

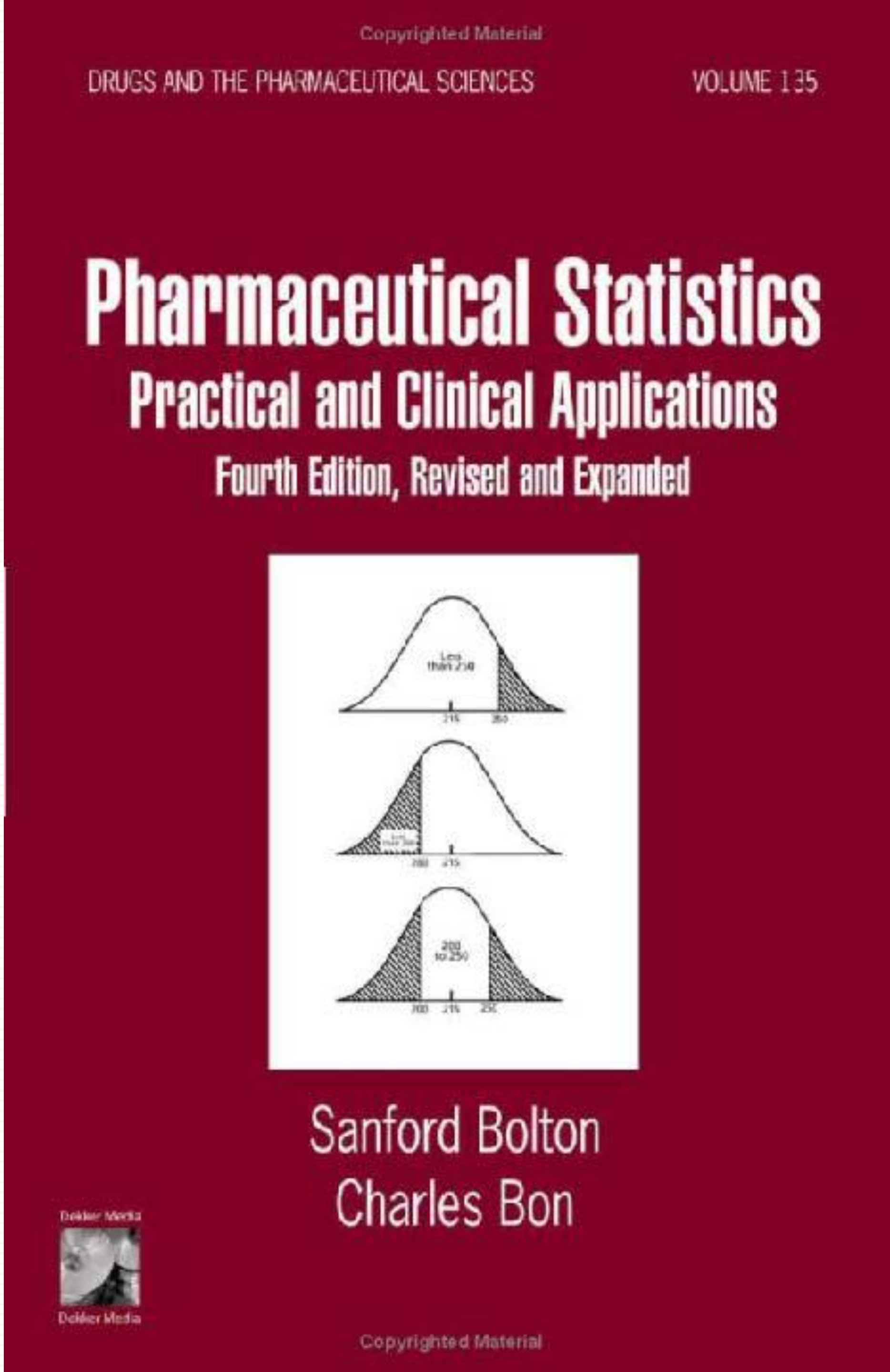
vsota 65,4

χ^2_{eks}

stopinje prostosti: m = k-r-1 = 9-2-1 = 6

$\chi^2_{tab(d.f. = 6;\ \alpha = 0,05)} = 12,6 < \chi^2_{eks};\ p < 0.05 \rightarrow$

H_0 zavržemo in sprejmemo H_A; frekvenčna porazdelitev se ne porazdeljuje normalno



Še vedno poglavje 15!