



Analýza dát

Prednáška 7

Testovanie hypotéz pre
kategoriálne dáta

χ^2 testy

χ^2 testy

χ^2 test dobrej zhody

- Pre jednu kategoriálnu premennú
- Príklad 1

χ^2 test nezávislosti

- Pre dve kategoriálne premenné
- Príklad 2

Príklad 1

V malom americkom okrsku bolo náhodne vybraných 275 súdnych porotcov a skúmalo sa, či ich príslušnosť k rase zodpovedá tomu, ako sú v danom štáte rozdelení voliči podľa rasy.

Všetci voliči majú podľa rasy nasledovné percentuálne zastúpenie:

Belosi	Černosi	Hispanci	Iní	Spolu
72 %	7%	12%	9%	100%

Príklad 1 - pokračovanie

275 súdnych porotcov bolo vzhľadom na rasu rozdelených takto:

Belosi	Černosi	Hispanci	Iní	Spolu
205	26	25	19	275

Môžeme na základe týchto údajov tvrdiť, že na súde v malom americkom okrsku dochádza k rasovej diskriminácii?

Príklad 1- pokračovanie

Ak by neexistovala na súde rasová diskriminácia,

- koľko porotcov-belochoch by sme očakávali medzi 275 porotcami?
- koľko porotcov-černochoch?
- koľko porotcov-hispáncov?
- koľko porotcov-iných?

Belosi	Černosi	Hispánci	Iní	Spolu
72 %	7%	12%	9%	100%

- Belosi:
- Černosi:
- Hispánci:
- Iní:

Príklad 1 - pokračovanie

	Belosi	Černosi	Hispanci	Iní	Spolu
Skutočnosť	205	26	25	19	275
Očakávané	198	19,25	33	24,75	275

Sú rozdiely medzi skutočnosťou a očakávaniami dostatočne veľké, aby sme mohli povedať, že existuje na súde rasová diskriminácia?

Nová testovacia štatistika

Vyskúšajme zmerať rozdiely medzi skutočnosťou a očakávaniami takto:

$$\frac{\textit{skutočnosť} - \textit{očakávanie}}{\sqrt{\textit{očakávanie}}}$$

jednotlivo pre všetky kategórie.

Nová testovacia štatistika - príklad

	Belosi	Černosi	Hispanci	Iní	Spolu
Skutočnosť	205	26	25	19	275
Očakávané	198	19,25	33	24,75	275

- Belosi:
- Černosi:
- Hispanci:
- Iní:
- A teraz všetko sčítame. Ale...

Vymyslíme niečo iné

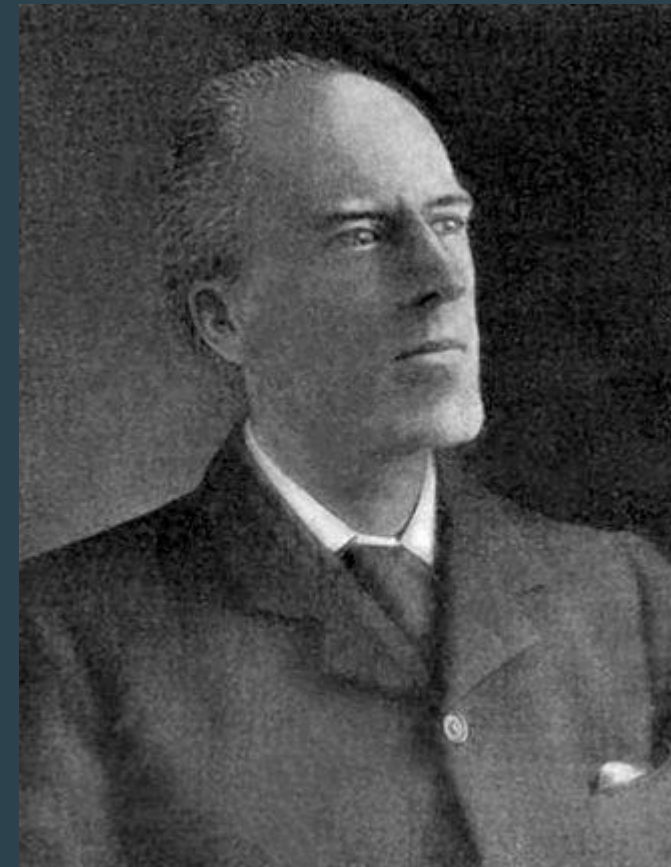
Každý zlomok pred sčítaním umocníme na druhú a až následne všetko spočítame:

Vymysleli sme χ^2
(chí kvadrát)

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - m_i)^2}{m_i}$$

kde:

- k je počet kategórií
- n_i znamená skutočný počet
- m_i znamená očakávaný počet



Karl Pearson (* 27. marec 1857, Londýn,
† 27. apríl 1936, Coldharbour)
bol anglický matematik a filozof,

Čo sme objavili a ako to funguje?

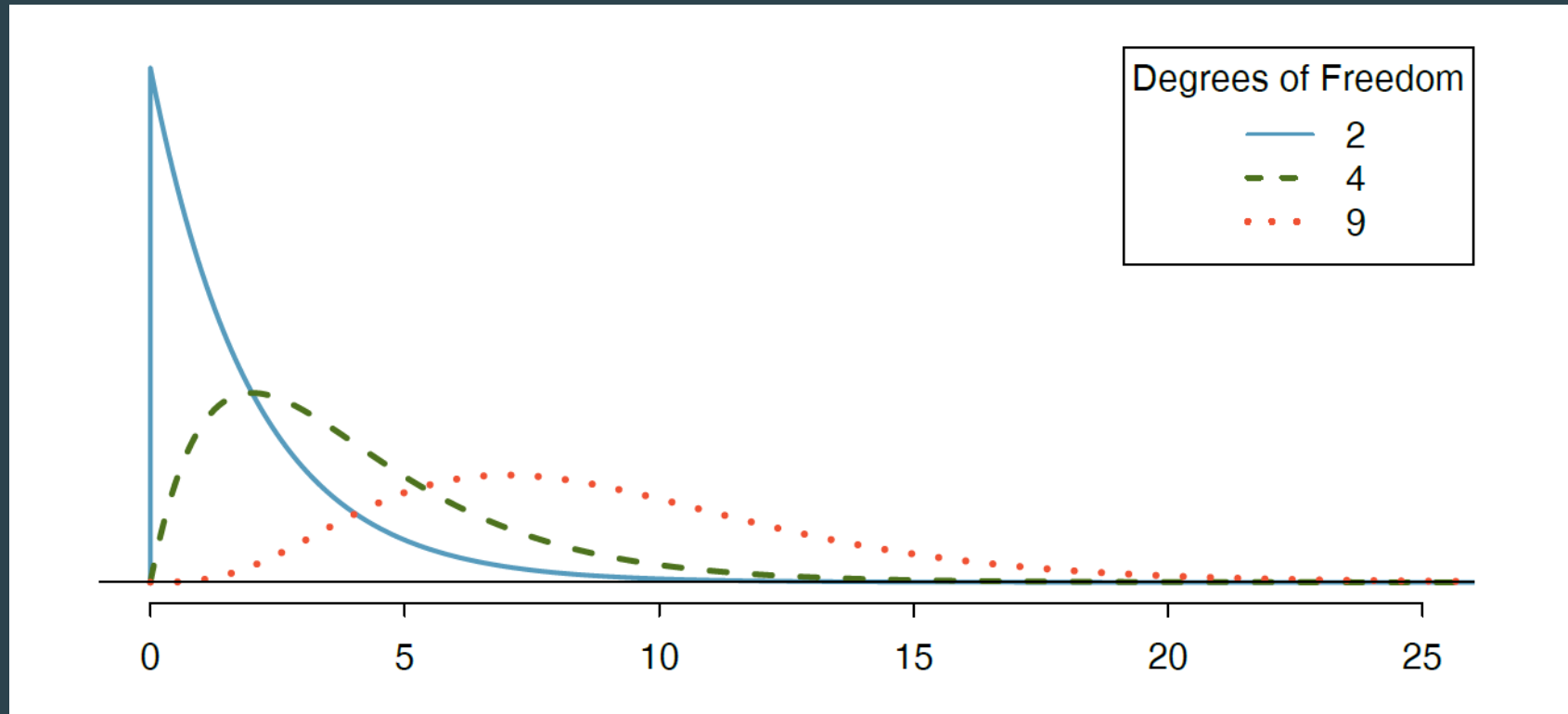
Vymysleli sme číslo χ^2 , ktoré meria ako veľmi sa líšia namerané početnosti od očakávaných.

Aká je minimálna hodnota? Kedy nastane?

Existuje maximálna hodnota?

Čo sa deje, ak sa hodnota χ^2 zväčšuje?

Graf χ^2 rozdelenia



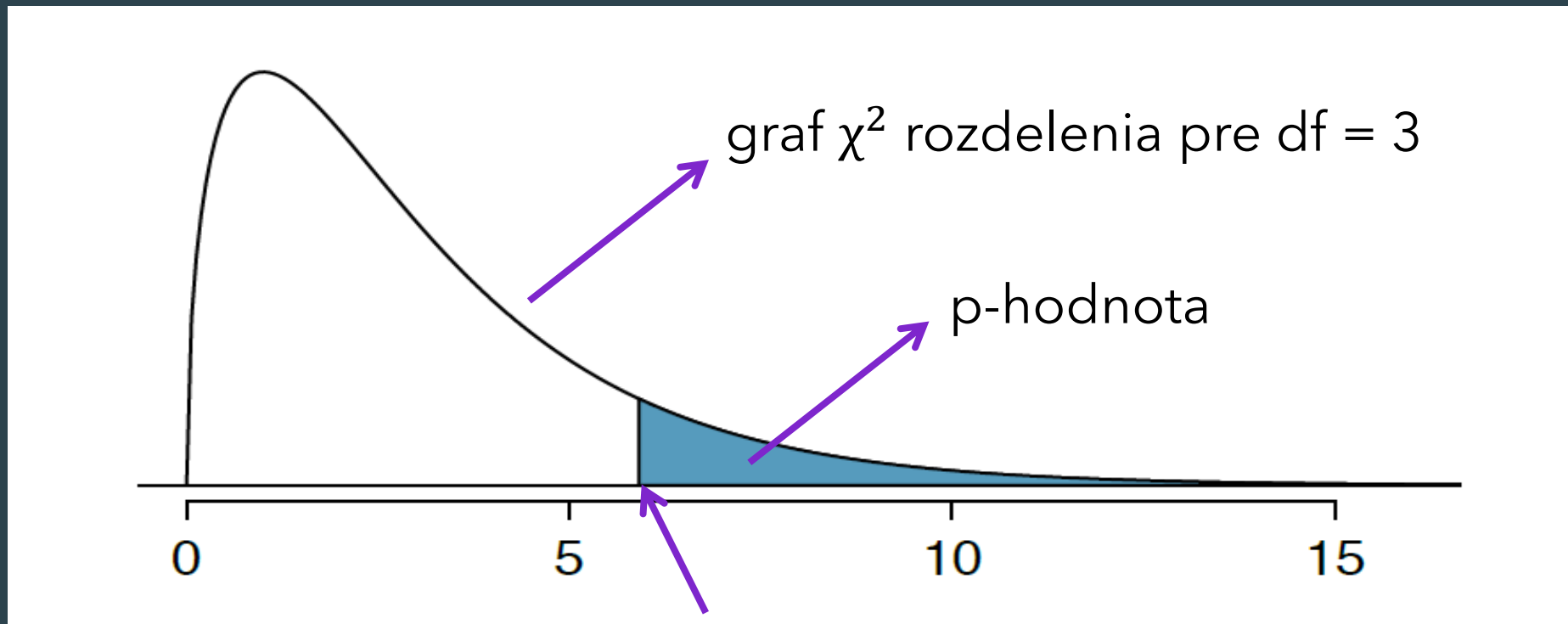
Degrees of Freedom = stupne voľnosti = $k - 1$

Príklad 1 - pokračovanie

Vráťme sa ku porotcom...

- Koľko kategórií sme mali? $k = 4$
- Koľko stupňov voľnosti máme? $DF = k - 1 = 3$
- Vypočítali sme, že rozdiely oproti očakávaniam sú $\chi^2 = 5,89$
- Je to veľa alebo málo?
- Ako vyzerá graf χ^2 rozdelenia?

Príklad 1 - pokračovanie



$$\chi^2 = 5,89$$

Ako vypočítame p-hodnotu pre χ^2 ?

- pomocou tabuliek – nepresné, neodporúčam;
- pomocou softvéru (napr. Excel):

= CHISQ.DIST.RT(x;stupeň voľnosti)

= CHISQ.DIST.RT(5,89;3)= 0,117

χ^2 test dobrej zhody

1. Formulovanie nulovej a alternatívnej hypotézy.
2. Stanovenie hladiny významnosti.
3. Kontrola podmienok: očakávaných početností musí byť v každej kategórii aspoň 5 ($m_i \geq 5$)
4. Výpočet χ^2 a p-hodnoty:

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - m_i)^2}{m_i};$$

$$\text{p-hodnota} = \text{CHISQ.DIST.RT}(\chi^2; k-1)$$

χ^2 test dobrej zhody

5. Rozhodnutie:

- ak je p-hodnota menšia ako hladina významnosti, tak zamietame nulovú hypotézu v prospech alternatívnej hypotézy
- ak je p-hodnota väčšia prípadne rovná hladine významnosti, nulovú hypotézu nemožno zamietnuť

Príklad 1 – testovanie

Príklad 2

Majiteľov zábavného centra zaujímalo, či obľúbenosť hororových filmov nejakým spôsobom súvisí s pohlavím, teda, či napr. mužskí návštevníci viac obľubujú horory alebo je to naopak. Náhodným výberom vybrali 100 zo svojich zákazníkov a spýtali sa ich, či obľubujú hororové filmy alebo nie.

Výsledok tohto prieskumu je v nasledujúcej kontingenčnej tabuľke:

Poznámka na okraj

Kontingenčná tabuľka je tabuľka, v ktorej sú zaznamenané hodnoty absolútnych početností dvoch kategoriálnych premenných (v ang. two-way table). Počet riadkov v kontingenčnej tabuľke budeme označovať r a počet stĺpcov budeme označovať s .

V našom príklade 2 ide o premenné:

- Pohlavie
- Oblúbenosť hororov

Príklad 2

Skutočnosť	Oblúbujú horory	Neoblúbujú horory	súčet
Muži	32	13	45
Ženy	28	27	55
súčet	60	40	100

Možno na základe týchto údajov tvrdiť, že mužskí zákazníci zábavného centra oblúbujú horory viac ako ženské návštevníčky?

χ^2 test o nezávislosti

1. Formulovanie nulovej a alternatívnej hypotézy:

H_0 : premenné pohlavie a obľúbenosť hororov sú nezávislé (horory sú rovnako obľúbené alebo neobľúbené u oboch pohlaví)

H_1 : premenné pohlavie a obľúbenosť hororov sú závislé (muži a ženy majú rozdielny pohľad na obľúbenosť hororov)

Príklad 2

Predpokladajme, že platí nulová hypotéza, teda že premenné pohlavie a obľúbenosť hororov sú nezávislé.

Za tohto predpokladu:

- koľko očakávame mužov, ktorí obľubujú horory?
- Koľko očakávame žien, ktoré obľubujú horory?
- Koľko očakávame mužov, ktorí neobľubujú horory?
- Koľko očakávame žien, ktoré neobľubujú horory?

Očakávanie	Oblúbujú horory	Neoblúbujú horory	súčet
Muži			45
Ženy			55
súčet	60	40	100

Muži oblúbujúci horory:

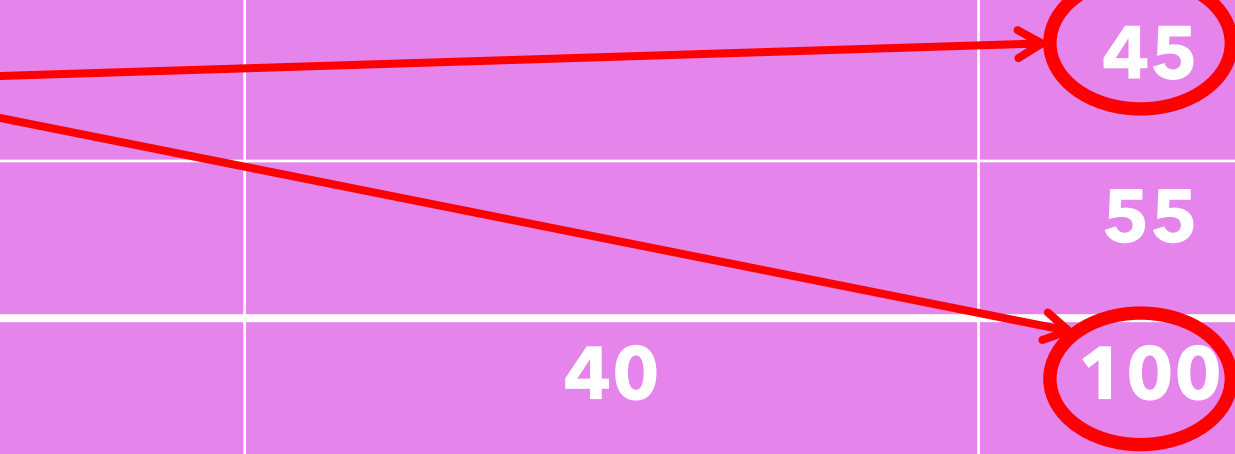
Muži neoblúbujúci horory:

Ženy oblúbujúce horory:

Ženy neoblúbujúce horory:

Príklad 2 - Očakávané početnosti

	Oblúbujú horory	Neoblúbujú horory	súčet
Muži	$\frac{60 * 45}{100}$		45
Ženy			55
súčet	60	40	100



Porovnanie skutočnosti a očakávaní

Skutočnosť	Oblúbujú horory	Neoblúbujú horory	súčet
Muži	32	13	45
Ženy	28	27	55
súčet	60	40	100

Očakávanie	Oblúbujú horory	Neoblúbujú horory	súčet
Muži	27	18	45
Ženy	33	22	55
súčet	60	40	100

Príklad 2 –pokračujeme v testovaní

2. Stanovenie hladiny významnosti: $\alpha = 0,05$.

3. Kontrola podmienok: očakávaných početností musí byť v každej kategórii aspoň 5.

4. Výpočet χ^2 a p-hodnoty:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

p-hodnota= CHISQ.DIST.RT (x;stupeň voľnosti)

Stupne voľnosti = $(r - 1) \cdot (s - 1)$, r = počet riadkov v kontingenčnej tabuľke a s = počet stĺpcov v kontingenčnej tabuľke

Príklad 2 - výpočet

Skutočnosť	Oblubujú horory	Neoblubujú horory	súčet
Muži	32	13	45
Ženy	28	27	55
súčet	60	40	100

Očakávanie	Oblubujú horory	Neoblubujú horory	súčet
Muži	27	18	45
Ženy	33	22	55
súčet	60	40	100

Príklad 2-rozhodnutie

5. Rozhodnutie:

keďže p-hodnota (0,04) je menšia ako hladina významnosti (0,05), nulovú hypotézu (že premenné sú nezávislé) **môžeme zamietnuť** a teda, premenné sú závislé: muži a ženy sa líšia v názore na obľúbenosť hororových filmov a **muži v zábavnom centre majú viac v obľube hororové filmy ako ženy.**

2. kreatívne zadanie

Zrealizujte „pokus“, na ktorom si vyskúšate použitie χ^2 **testu dobrej zhody alebo** χ^2 **testu nezávislosti** . Svoj postup zdokumentujte (video, fotky...) a samozrejme vykonajte príslušný výpočet (test) s bodmi 1-5. Pozor na jedinú podmienku použitia testu (aspoň 5 očakávaných početností v každej kategórii/v každom políčku kontingenčnej tabuľky).

BACHA: treba si vybrať 1 kategoriálnu premennú v prípade χ^2 testu dobrej zhody alebo 2 kategoriálne premenné v prípade χ^2 testu nezávislosti!!!

2. kreatívne zadanie

Možné námety:

- Je v balíčku gumených macíkov/lentiliiek/M&M's (?) rovnomerný počet každej farby?

