ANALÝZA DAT V R

9. VÝPOČET VELIKOSTI SOUBORU

Mgr. Markéta Pavlíková
Katedra pravděpodobnosti a matematické statistiky MFF UK

www.biostatisticka.cz

DATA, VÝZKUM, ANALÝZY

- ve výzkumu se střídají fáze prozkoumávací (exploratory) a fáze potvrzovací (confirmatory)
- tradičně se prezentují jako v opozici, ale v zásadě se spíše doplňují

EXPLORATORNÍ ANALÝZA

- exploratorní analýza
  - hledáme možné vazby, vzorce, tendence v datech
  - korelace, asociace slouží jako vodítko k vytváření hypotéz, případně k odhadu velikosti parametrů
  - „statistická významnost“ je spíše orientační hranicí
  - „už by to nemusela být náhoda“
  - „detektivní práce“: sbírání stop a důkazů

KONFORMATORNÍ ANALÝZA

- konfirmatorní analýza
  - potvrzujeme hypotézy stanovené předem
    - na základě našeho předchozího výzkumu
    - na základě výzkumu někoho jiného
    - na základě našich empirických zkušeností
  - jednoznačně postavená hypotéza, jednoznačně daný test
  - předem stanovená hranice chyby I. druhu, ideálně síla testu a tím velikost souboru
  - „soudní proces“, ve kterém jsou důkazy podrobeny testu

EXPLORE & CONFIRM

- na jedné náhodně vybrané části dat můžeme hypotézu generovat a druhou si nechat na potvrzování
- na jedněch datech lze potvrzovat starší hypotézu a současně prozkoumávat a generovat hypotézy nové
- oboje může využívat v principu stejné techniky
  - exploratorní: popisné statistiky, grafy, základní testy, konstrukce modelů, ...
  - konfirmatorní: typicky dvouvýběrový test, obecně konkrétní test nebo konkrétní model
  - (klasické frekventistické pojetí; bayesovská statistika pracuje trochu jinak)

KONFIRMACE - TESTOVÁNÍ HYPOTÉZY

1) jasně zvolená hypotéza a alternativa
2) zvolení minimální klinicky důležité odchylky / velikost efektu
3) vybraný konkrétní test
4) zvolená přijatelná hladina statistické významnosti
5) zvolená síla testu
6) vypočtená velikost souboru (prakticky navýšená)
7) nasbírání dat a provedení testu

1. HYPOTÉZA A ALTERNATIVA + 2. EFEKT

- stanovení hypotézy závisí na úkolu, který řešíme
- příklad (jednoduchý):
- *X* je počet úspěchů ve N pokusech, p je pravděpodobnost úspěchu
- *X*~Bi(p,N)
- H_0: p = 0.3 vs. H_1: p < 0.3
- ve skutečnosti pro výpočet síly nestačí H_1 v této podobě, ale v podobě nejmenšího rozdílu od H_0, který je nějak zajímavý
- H_0: p = 0.3 vs. H_1: p = 0.2

VÝZNAMNOST A SÍLA TESTU

	Ve skutečnosti platí H_0	Ve skutečnosti platí H_1
Tvrdíme, že platí H_0	$1-\alpha$ (správné rozhodnutí)	β (chyba II. druhu)
Zamítáme H0 (tvrdíme, že platí H_1)	α (chyba I. druhu) (hladina stat. významnosti)	$1-\beta$ (síla testu) (správné rozhodnutí)

- α volíme předem (jak moc nám vadí, že zamítneme H0, když platí)
- β je funkcí stanovené alternativy a velikosti souboru

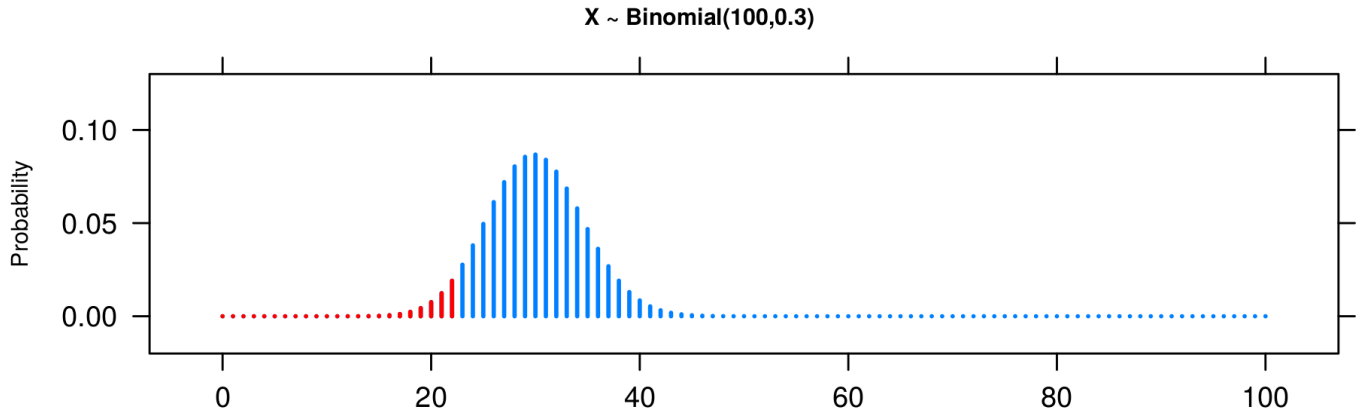
3. VHODNÝ TEST

- vhodný test pro danou situaci je binomický test / test proporcí
- v R `binom.test, prop.test`

$$P(X \leq x \mid p = p_0) = \sum_{k=0}^{x} \binom{N}{k} (p_0)^k (1 - p_0)^{N-k}$$

4. HLADINA VÝZNAMNOSTI A KRITICKÁ HODNOTA

- stanovíme hladinu významnosti $\alpha = 0.05$
- jaká je kritická hodnota testu?

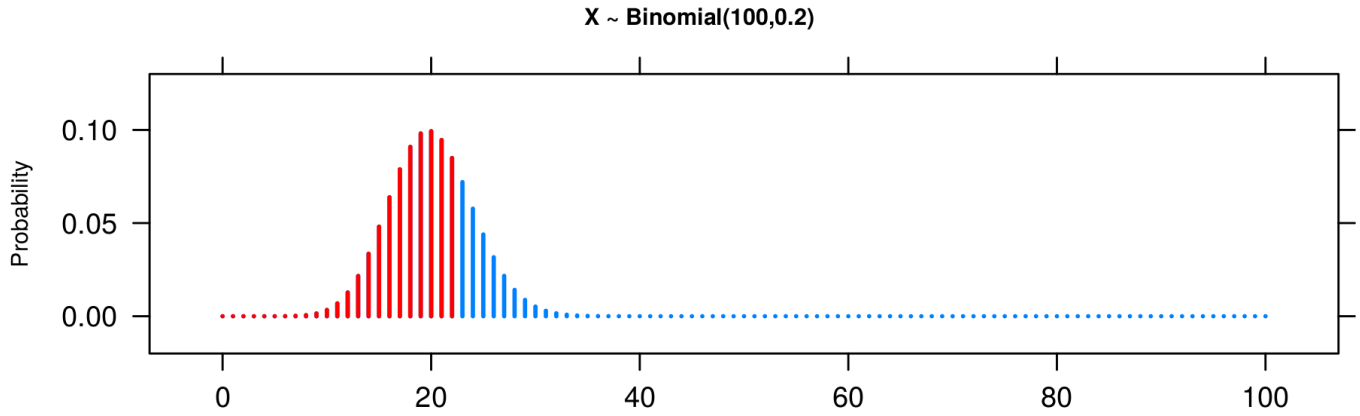



```
> k = qbinom(0.05, 100, 0.3)
[1] 23
> pbinom(k, 100, 0.3)
[1] 0.07553077
> pbinom(k - 1, 100, 0.3)
[1] 0.04786574
```

$$\begin{aligned} \alpha &= \mathrm{P}(X \leq 22 \,|\, p = 0.3) \\ &= \sum_{k=0}^{22} \binom{100}{k} (0.3)^k (0.7)^{100-k} \\ &\doteq 0.0479 \end{aligned}$$

5. SÍLA TESTU

- jaká je síla tohoto testu při p = 0.02
- stanovujeme konkrétní hodnotu (klinicky významný rozdíl)



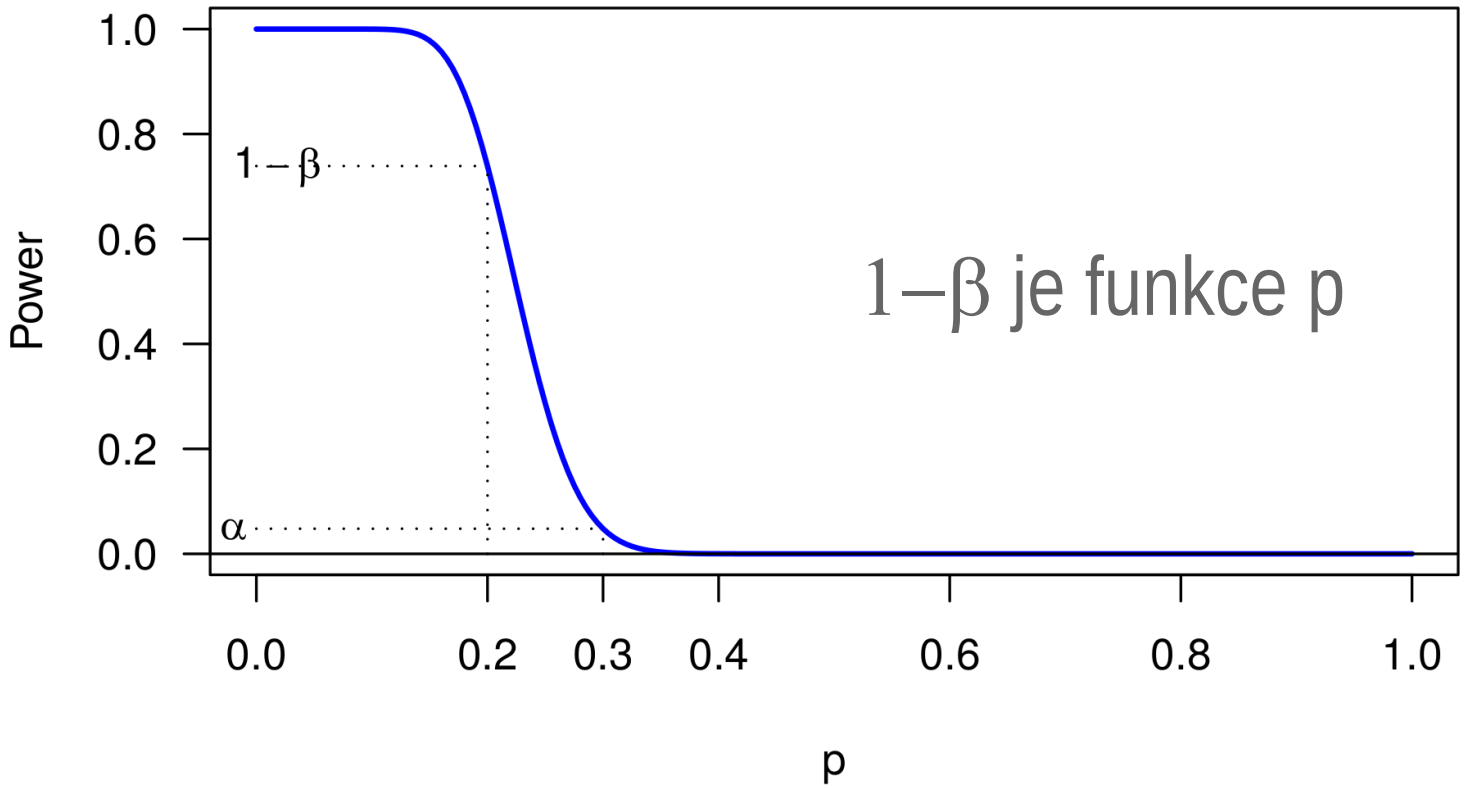

```
> pbinom(k, 100, 0.2)
[1] 0.7389328
```

$$1-\beta(p) = \mathrm{P}(X \le 22 \mid p) = \sum_{k=0}^{22} \binom{100}{k} p^k (1-p)^{100-k}$$

1–β je funkce p

$$1-\beta(0.2) \doteq 0.7389$$

5. SÍLA TESTU



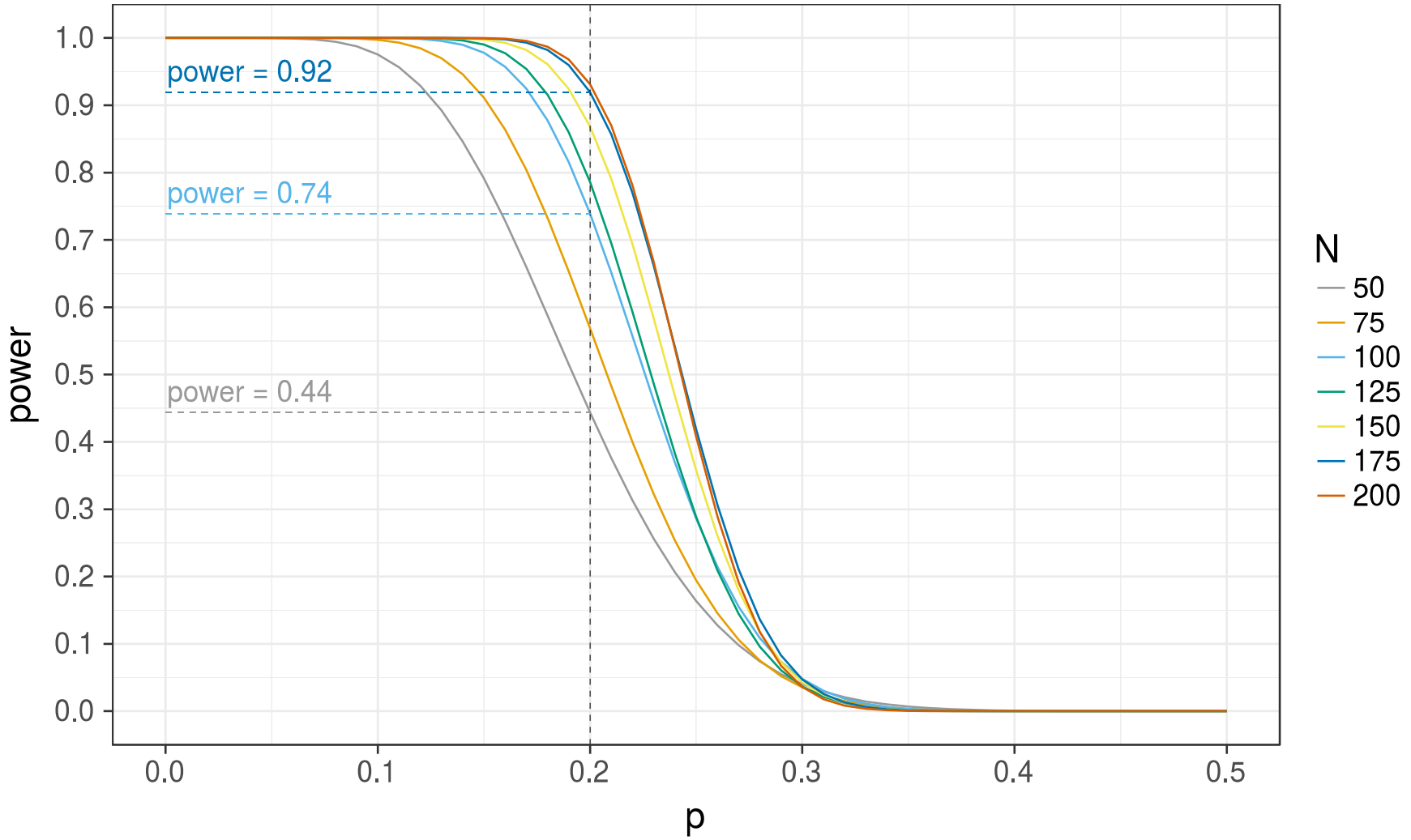
- pro spolehlivé přijetí H_1: p = 0.2 je síla 0.74 možná příliš nízká (vyšší pravděpodobnost nepřijetí, i když platí)
- jak zajistíme, aby byla síla testu 0.9?

6. VELIKOST SOUBORU

$$1 - \beta(p_A) = 1 - P(X \leq x \mid p = p_A) = \sum_{k=0}^{x} \binom{N}{k} (p_A)^k (1 - p_A)^{N-k}$$

- síla je funkce *N*, p a kritické hodnoty k (a tím tedy α)
- α, β, p a N jsou tedy svázané
- stanovíme α, p a N, můžeme vypočítat β
- stanovíme α, p a β, můžeme vypočítat N

6. VELIKOST SOUBORU



6. VELIKOST SOUBORU

- hledáme N empiricky
- $\alpha = 0.05$, $\beta = 0.90$
- $p_0 = 0.3$, $p_A = 0.2$

n	k	alpha	power
100	22	0.04786574	0.7389328
125	28	0.03682297	0.7856383
150	35	0.04286089	0.8683183
175	**42**	**0.04733449**	**0.9193571**
200	48	0.03594782	0.9309691

- přesný výpočet: `library(stat), library(pwr)`

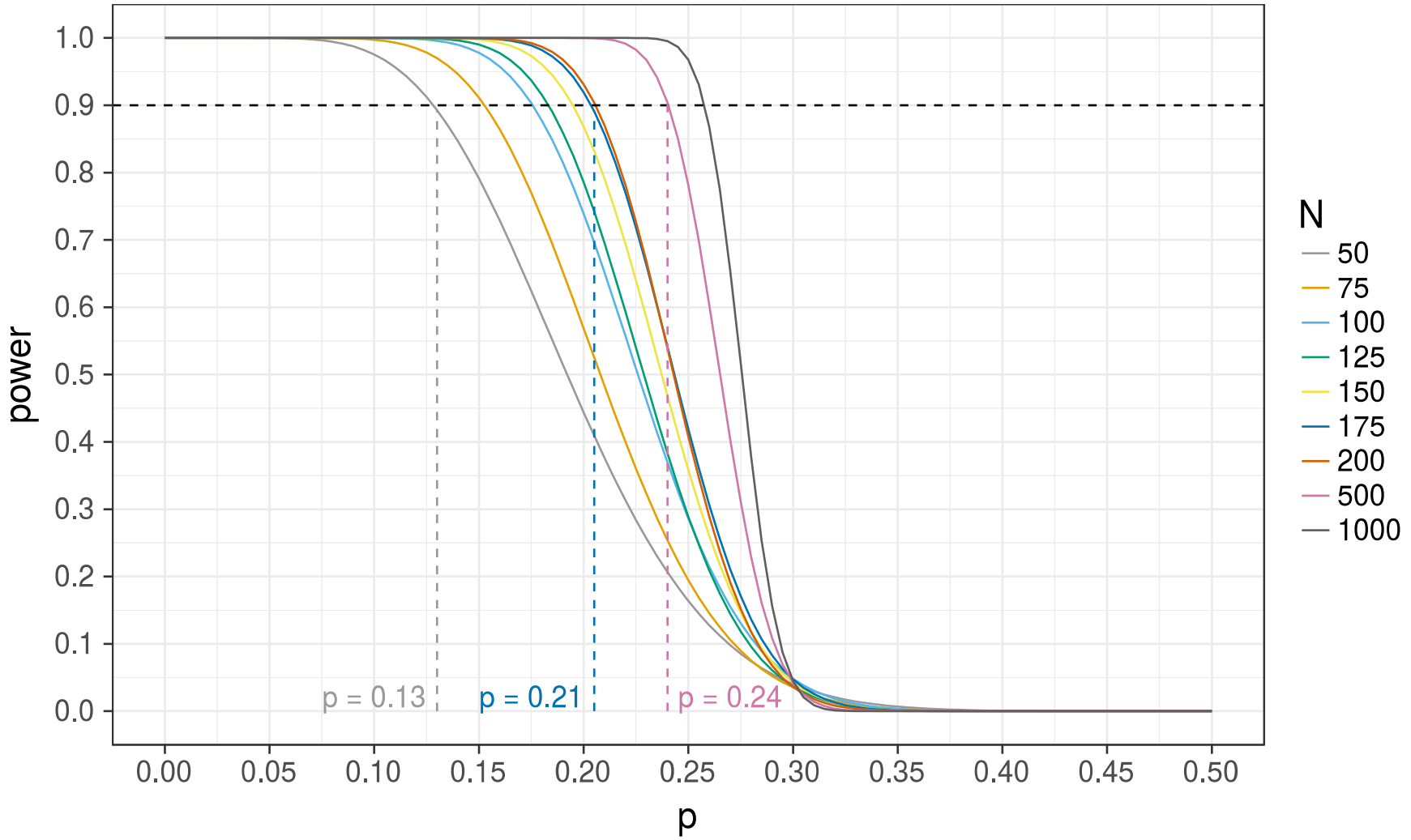
```
pwr.p.test(ES.h(0.2,0.3),sig.level=0.05,power = 0.9,alternative="less")


proportion power calculation for binomial distribution (arcsine transformation)

              h = -0.232
              n = 159.1
      sig.level = 0.05
          power = 0.9
    alternative = less
```

rozdíly: R má přesnější aproximaci α

6. VELIKOST SOUBORU



s 50 prokážeme spolehlivě na 90% rozdíl 0.17, se 175 rozdíl 0.09, s 500 rozdíl 0.06

KONFIRMACE - TESTOVÁNÍ HYPOTÉZY

- čím menší rozdíl chceme prokázat, tím strměji roste počet potřebných pozorování
- naopak, máme-li moc pozorování, prokážeme i rozdíl, který není klinicky zajímavý
- praktické hledisko: k vypočtenému počtu pozorování přidat očekávaný úbytek (zpravidla cca 10%)
- výpočet provádíme pro tzv. hlavní cíl (main endpoint)
- hlavní cíl stanovujeme tak, aby byl adekvátní počet rozumně dosažitelný
  - novorozenecká úmrtnost 1‰ vs. 2‰: 30576 v každé větvi
  - kompozitní výsledek úmrtnost + vážné poškození zdraví 1% vs. 2%: 3017 v každé větvi
- výsledná studie může být jak observační tak experimentální, u obojího lze provádět konfirmaci hypotéz na správně velkém souboru

KONFIRMACE - TESTOVÁNÍ HYPOTÉZY

- základní knihovna {stat} a větší library(pwr) má nástroje pro různé typy designu
  - porovnání dvou poměrů power.p.test, pwr.2p.test
  - porovnání dvou poměrů, skupiny nevyvážené pwr.2p2n.test
  - pro t.test, ANOVA, Fisherův test, chi-kvadrát test, test korelací, ..
- co s neparametrickými testy?
  - neznáme rozdělení, takže výpočet nejde
  - můžeme si pomoci simulací
  - rule-of-thumb: přidej 15% oproti parametrické variantě (Lehmann)
- co s regresí?
  - https://stats.stackexchange.com/questions/10079/rules-of-thumb-for-minimum-sample-size-for-multiple-regression

DĚKUJI ZA POZORNOST

www.biostatisticka.cz