# STATISTICKÉ VYHODNOCENÍ EPIDEMIOLOGICKÝCH STUDIÍ VYLOUČENÍ NÁHODY

## ASOCIACE:

**- statistická souvislost mezi dvěma proměnnými**

**- v epidemiologii mezi**
**exposicí (potenciálním rizikovým faktorem)**
**a**
**nemocí - výskytem** *(incidencí, prevalencí…)*
**- závažností průběhu**

# Jak se vyjadřuje riziko

Roční mortalita spojená s některými činnostmi v Holandsku

| | | |
|---|---|---|
| Včelí žihadlo | $2 \times 10^{-7}$ | 1 : 5 million |
| Zasažení bleskem | $5 \times 10^{-7}$ | 1 : 2 million |
| Létání | $1{,}23 \times 10^{-6}$ | 1 : 814 000 |
| Chůze | $1{,}85 \times 10^{-5}$ | 1 : 54 000 |
| Jízda na kole | $3{,}85 \times 10^{-5}$ | 1 : 26 000 |
| Řízení vozidla | $1{,}75 \times 10^{-4}$ | 1 : 5 700 |
| Jízda na motocyklu | $2 \times 10^{-4}$ | 1 : 1 000 |
| Kouření cigaret (20/den) | $5 \times 10^{-3}$ | 1 : 200 |

# TYPY VELIČIN VE STATISTICE

1. **Spojité:** Kontinuum v určitém rozsahu

2. **Diskrétní:** jen určité hodnoty

**a) Kategoriální = kvalitativní**

**dichotomické** (muž/žena; přežil/nepřežil; exponován/neexponován..)

**nominální** (krevní skupiny A/B/AB/0); stav svobodný/ženatý/rozvedený/vdovec…

**ordinální** – zlepšení/ beze změny/ zhoršení

**b) Numerické:** počty jevů či dějů (červů, klíšťat, dětí, ataků malárie…

# STATISTICKÉ VYHODNOCENÍ SPOJITÝCH VELIČIN

- 2 skupiny (populace A a B)
- Porovnáváme jejich parametry

## Porovnání průměrů:
## t-test

**Nulová hypotéza: oba průměry jsou stejné**

**P: Pravděpodobnost, že pokud nulová hypotéza platí, bude výsledek stejný nebo ještě extrémnější**

**Hladina významnosti: Pravděpodobnost chyby – zamítnutá nulová hypotéza ve skutečnosti platí**

**0,01; 0,05; 0,1**

**$P \leq 0,05$: Náhoda nepravděpodobná –zamítnutí nulové hypotézy**

**P > 0,05: Náhodu nelze vyloučit – nulovou hypotézu nezamítáme**

# t-test

- **Jednovýběrový:** hodnoty sledované populace porovnáváme s normou, tabulkovými hodnotami atp.
- **Dvojvýběrový:** Porovnáváme hodnoty u dvou populací
- **Jednostranný:** Předpokládáme, že hodnoty v populaci A budou vyšší než v populaci B (P je poloviční než u dvoustranného t-testu)
- **Dvoustranný:** Nevíme, zda hodnoty budou vyšší v populaci A nebo B
- **Párový:** Jediná skupina, měřena dvakrát po sobě. Hypotéza: Průměrný rozdíl při prvním a druhém měření bude 0.

# Dvouvýběrový dvoustranný T-test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$\bar{X}_1$ $\bar{X}_2$ : **průměry skupin 1 a 2**

$n_1$, $n_2$: **počty ve skupinách 1, 2**

**s : kombinace rozptylů obou skupin**

**t: tabulková hodnota pro určení P**

**s nárůstem t klesá hodnota P**

**T-test lze použít u výsledků s normálním rozdělením.**

**Jinak: transformovat výsledky logaritmováním, odmocněním atp.**

# STATISTICKÉ VYHODNOCENÍ DISKRÉTNÍCH VELIČIN

**Kohortové studie, studie případů a kontrol, průřezové studie: osoby řazeny do diskrétních kategorií:**

- **exponován/neexponován rizikovému faktoru**
- **onemocněl/neonemocněl**

| | Riziko + | Riziko - |
|---|---|---|
| **Nemoc +** | a | b |
| **Nemoc -** | c | d |

# VYHODNOCENÍ KOHORTOVÝCH STUDIÍ
# RELATIVNÍ RIZIKO

## - z kumulativní incidence, prevalence

**Riziko:** Pravděpodobnost vzniku nemoci v jedné skupině (u exponovaných, u neexponovaných)

- **Relativní riziko – relative risk (RR):** Kolikrát je větší riziko vzniku nemoci v exponované skupině než v neexponované

$$RR = \frac{a\,(b+d)}{b\,(a+c)}$$

# RELATIVNÍ RIZIKO

## - z kumulativní incidence, prevalence

- **RR > 1:** Faktor zvyšuje pravděpodobnost vzniku nemoci

  **RR = 1:** Faktor nemá vliv na vznik nemoci

  **RR < 1**: Faktor snižuje pravděpodobnost vzniku nemoci
- Byla – li vyšetřována úplně celá populace – RR se vztahuje k populaci
- Byl – li vyšetřen jen vzorek populace (správný výběr!), musíme výsledek vzorku k celé populaci vztáhnout.
- **95% interval spolehlivosti (confidence interval):** interval, ve kterém se s 95% pravděpodobností nachází skutečné RR celé populace
- **Leží – li v intervalu spolehlivosti i RR = 1, nelze vyloučit stejné riziko v exponované a neexponované skupině a hypotéza o statistické odlišnosti se zamítá**

# RELATIVNÍ RIZIKO-95% interval spolehlivosti MEZE

s 95% pravděpodobností

$$RR = e^{\pm 1{,}96\sqrt{\text{rozptyl}\,(\ln RR)}}$$

spodní mez $\frac{RR}{F}$

a,b,c,d musí být alespoň 10

$$F = e^{1{,}96\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}}$$

horní mez: RR.F

Hodnota 1,96

F: faktor chyby
e: základ přirozených logaritmů
Hodnota 1,96 vychází ze standardního normálního rozdělení a odpovídá hladině spolehlivosti 95 %; pro 99% je to 2,576

Interval spolehlivosti rozšiřuje
rozptyl - (i v důsledku **nízkého N** !)
relativní riziko

# RELATIVNÍ RIZIKO V OSOBODNECH I
# - z incidence

**$PY_e$ : Celková doba sledování exponovaných**
(v osobodnech)

**$PY_0$ : Celková doba sledování neexponovaných**
(v osobodnech)

**a, b: Počet nových případů nemoci**

| | Riziko + | Riziko - |
|---|---|---|
| Nemoc + | a | b |
| Celkem | $PY_e$ | $PY_0$ |

# RELATIVNÍ RIZIKO V OSOBODNECH II
## - z incidence

- **Relativní riziko – relative risk (RR):** kolikrát je riziko, že se osoba nakazí za jeden den sledování větší ve skupině exponovaných než ve skupině kontrolní
- Testování statistické významnosti:

  **interval spolehlivosti – confidence interval**

# Riziko (incidence) v osobodnech

## Incidence toxoplasmózy u HIV$^+$ pacientů

MACHALA, L., MALÝ, M., BERAN, O., JILICH, D., KODYM, P. Incidence and clinical and immunological characteristics of primary Toxoplasma gondii infection in HIV-infected patients. *International Journal of Infectious Diseases*. 2013, **17**(10), e892-e896. ISSN 1201-9712.

- **Seroconversion** indicating a recent T. gondii infection was observed in **14 patients.**
- The **total person-time of follow-up** of HIV-infected patients at risk of Toxoplasma seroconversion was **3046.3 years**. (infekce 1x za 217,6 let)
- The **incidence rate** of primary toxoplasmosis in the cohort was therefore 14:3046.3 = **0.0046**

**Toxoplasmosis in pregnant women (Palička et al., 1996):**

**0.0029 cases**

**of infection per one pregnancy (ca.3/4 of the year)**

# ATRIBUTIVNÍ RIZIKO (AR)
# ATTRIBUTABLE RISK

$$AR = R_e - R_0 = RR - 1$$

$R_e$ :Riziko u exponovaných

$R_0$ :Riziko u neexponovaných

O kolik je riziko větší u exponovaných než u neexponovaných; o kolik by se snížilo riziko, kdyby se odstranil sledovaný faktor **(absolutní efekt exposice)**

## ATRIBUTIVNÍ RIZIKO V PROCENTECH (AR%)

$$AR\% = \frac{AR}{R_e} \times 100$$

Podíl atributivního rizika – zdravotnický význam daného faktoru – o kolik % by kleslo riziko vzniku nemoci, kdyby se odstranil **(etiologická frakce)**

# ATRIBUTIVNÍ RIZIKO POPULAČNÍ (PAR)
# vychází z incidence

$$PAR = \frac{I_p - I_0}{I_p}$$

**$I_p$ : Incidence nemoci v populaci**

**$I_0$ : Incidence nemoci u neexp.**

Jakou část incidence nemoci lze připsat působení rizikového faktoru
Závisí na incidenci nemoci.

I malé atributivní riziko, působící nemoc s vysokou incidencí: velký dopad.

Lze vyjádřit i v % (PAR %)

# ODDS, ODDS RATIO

Odds = nerovnost, rozdíl, nestejnost

**Studie případů a kontrol**

**Odds: „šance na onemocnění“ – podíl exponovaných a neexponovaných v každé skupině**

| | **Riziko +** | **Riziko -** |
|---|---|---|
| **Nemoc +** | a | b |
| **Nemoc -** | c | d |

# ODDS RATIO

$$OR = \frac{a.d}{b.c}$$

**Odds ratio (OR):** Kolikrát je větší zastoupení exponovaných mezi případy než mezi kontrolami

# ODDS RATIO

- **HODNOCENÍ:**

**OR > 1:** Faktor zvyšuje pravděpodobnost vzniku nemoci

**OR = 1:** Faktor nemá vliv na vznik nemoci

**OR < 1:** Faktor snižuje pravděpodobnost vzniku nemoci

- Byla – li vyšetřována úplně celá populace – RR se vztahuje k populaci
- Byl – li vyšetřen jen vzorek populace (správný výběr!), musíme výsledek vzorku k celé populaci vztáhnout.
- **95% interval spolehlivosti (confidence interval):** interval, ve kterém se s 95% pravděpodobností nachází skutečné OR celé populace
- **Leží – li v intervalu spolehlivosti i OR = 1, nelze vyloučit stejné riziko v exponované a neexponované skupině a hypotéza o statistické odlišnosti se zamítá**

# Vztah OR a RR

- **V dobře provedené studii případů a kontrol je OR dobrou aproximací RR**
- **OR je odolnější vůči výběrovému bias**
- **OR lze použít i při studiu vzácných nemocí (a, b je skoro nula, lze zanedbat):**

$$RR = \frac{a\,(b+d)}{b\,(a+c)} = \frac{a.d}{b.c} = OR \text{ (přibližně)}$$

* U běžných nemocí: 0 < RR < OR

# X-kvadrát

Jedinci klasifikovaní dle dvou **kategoriálních znaků.**

Každý znak má 2 nebo více kategorií.

Test zjišťuje, jaká je pravděpodobnost, že celkový počet osob bude rozdělen právě tímto způsobem.

| | Riziko + | Riziko – | Riziko +/- |
|---|---|---|---|
| Nemoc + | **a** | **b** | |
| Nemoc - | **c** | **d** | |

Počet stupňů volnosti: (počet sloupců-1 x počet řádků – 1)

# Fisherův exaktní test

Jedinci klasifikovaní dle dvou kategoriálních znaků.

Každý znak má 2 nebo více kategorií.

Test zjišťuje, jaká je pravděpodobnost, že celkový počet osob bude rozdělen právě tímto způsobem.

**Používá se hlavně při nízkých počtech jedinců ve skupinách.**

**Jednostranný, dvojstranný**

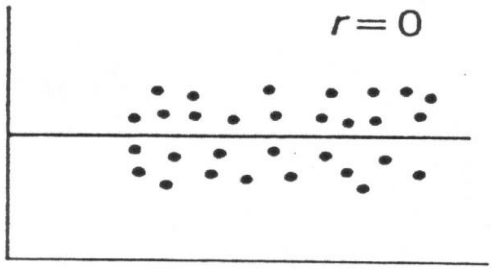
| | Riziko + | Riziko – | Riziko +/- |
|---|---|---|---|
| Nemoc + | **a** | **b** | |
| Nemoc - | **c** | **d** | |

Počet stupňů volnosti: (počet sloupců-1 x počet řádků – 1)

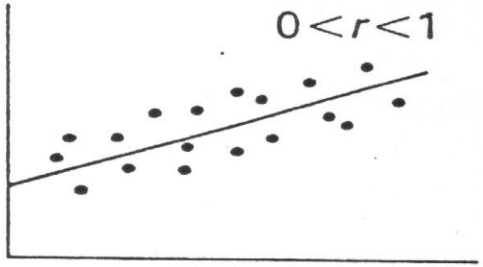
# ANALÝZA ZÁVISLOSTI JEDNÉ VELIČINY NA DRUHÉ

- **LINEÁRNÍ REGRESE**
- Pro spojité veličiny
- Proložení optimální přímky se určí metodou nejmenších čtverců

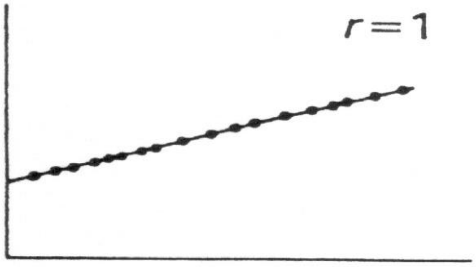
$$y = a + rx$$



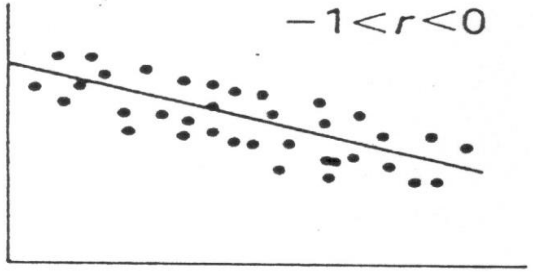

(a) No correlation



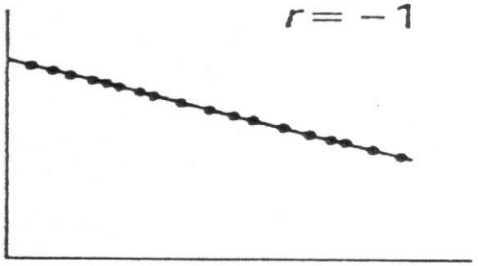

(b) Imperfect positive correlation




(c) Perfect positive correlation




(d) Imperfect negative correlation




(e) Perfect negative correlation

**Figure 9.2** Scatter diagrams illustrating different values of the correlation coefficient. Also shown are the regression lines.

# ANALÝZA VÍCE RIZIKOVÝCH FAKTORŮ SOUČASNĚ

## MNOHOROZMĚRNÁ REGRESNÍ ANALÝZA

- rozšíření modelu lineární regrese do vícerozměrného proístoru. Sleduje se závislost na více faktorech **– covariate.**

**VÍCENÁSOBNÁ LINEÁRNÍ REGRESE:** rozšíření modelu lineární regrese do vícerozměrného prostoru – model

„rozšířená rovnice přímky“:

$$Y = a + r_1x_1 + r_2x_2 \ldots + r_nx_n$$

$r_1$; $r_2 \ldots r_n$ korelční koeficienty jednotlivých rizikových faktorů

velikost r je úměrná síle faktorů

- Sledování , které z více faktorů mají vliv a které ne
- Odstranění confounding
- Využívá se v balíčcích statistických software

# Nepřátelé epidemiologa II

# Bias I

***Bias = svah, sklon, pud***

**Systémová chyba v provedení studie.**

## Výběrové bias: nesoulad kontroly a pokusné skupiny

Pokusná a kontrolní skupina se mohou lišit buď jen v přítomnosti/nepřítomnosti nemoci

nebo

jen v přítomnosti/nepřítomnosti rizikového faktoru,

**jinak úplně stejné!**

*Nemocniční pacienti: vstupuje do hry*

*důvod, proč byli hospitalizováni*

# Nepřátelé epidemiologa II

## Bias III

# ODSTRANĚNÍ BIAS

**Výběrové:**

**RANDOMIZACE: „Náhodné“ rozdělení do pokusných a kontrolních skupin tak, aby obě měly stejné složení (pohlaví, věk, etnikum….) – pomocí počítače, náhodných čísel – ne dle výběru konkrétní osoby**

**Informační:**

**DVOJITÉ ZASLEPENÍ (double blind experiment):**

**Experimentátor, který přichází s pokusnými osobami do styku, je nerozděluje do skupin a neví, kdo je ve skupině pokusné a kdo v kontrolní. To nevědí ani pokusné osoby.**

**TROJITÉ ZASLEPENÍ (triple blind experiment): Kdo je pokusná osoba a kdo patří do kontrolní skupiny neví ani pracovník, který data vyhodnocuje.**

**Kontrolní místo léku dostávají neúčinné a neškodné PLACEBO**

# MATCHING

Hlavně studie případů a kontrol (selekční bias).

- Ke každému jednotlivému případu přiřazena kontrola, která se liší jen tím, že nemá nemoc

**Ověření: Mc Namarovým chí-kvadrát testem**: jestli počet párů, které se neshodují v nějakém dalším parametru, neznamená statisticky významný rozdíl mezi skupinami případů a kontrol.

# Nepřátelé epidemiologa III

# Confounding

## Třetí faktor = confounder

**jsou zde dvě různé asociace:**

**confounder-nemoc**

**confounder-zkoumaný faktor**

Pokusná a kontrolní skupina se smějí lišit opravdu jen ve sledovaném faktoru !

# ODSTRANĚNÍ CONFOUNDING

# STRATIFIKACE

- Kategoriální veličiny, nebo se kategorie musí vytvořit. Je podezření, že by některá veličina mohla být confounderem:
- Vzorek se rozdělí podle této veličiny do vrstev
- V každé vrstvě se spočítá OR, RR. Z toho lze sledovat případný vliv confounderu: mění se RR

$$RR = \frac{a\,(b+d)}{b\,(a+c)}$$

**Lze spočítat adjustované RR/OR dle Mantel-Hanszela**

- $$RR_{MH} = \frac{\frac{\sum ad}{a+b+c+d}}{\frac{\sum bc}{a+b+c+d}}$$

* Rozpor mezi adjustovaným a skutečným OR/RR: confounding!

# STANDARDIZACE

- Porovnává-li se např. incidence, prevalence, mortalita, RR/OR atd. nemocí v různých populacích (nebo i v téže populaci po čase), musí se brát do úvahy jejich věkové složení
- Proto: stratifikace podle věkových kategorií, spočítat pro každou zvlášť
- Přepočítat na standardní (stejné) věkové složení, potom teprve porovnávat

*Table 4-4.* Crude and age-specific mortality rates from cancer in the United States, 1980

| Age in years | Number of cancer deaths | Population as of July 1, 1980 | Mortality rate per 100,000 |
|---|---|---|---|
| Under 5 | 686 | 16,348,000 | 4.2 |
| 5–9 | 777 | 16,700,000 | 4.7 |
| 10–14 | 720 | 18,242,000 | 3.9 |
| 15–19 | 1145 | 21,168,000 | 5.4 |
| 20–24 | 1538 | 21,319,000 | 7.2 |
| 25–29 | 2041 | 19,521,000 | 10.5 |
| 30–34 | 3040 | 17,561,000 | 17.3 |
| 35–39 | 4684 | 13,965,000 | 33.5 |
| 40–44 | 7786 | 11,669,000 | 66.7 |
| 45–49 | 14,230 | 11,090,000 | 128.3 |
| 50–54 | 26,800 | 11,710,000 | 228.9 |
| 55–59 | 41,600 | 11,615,000 | 358.2 |
| 60–64 | 53,045 | 10,088,000 | 525.8 |
| 65–74 | 127,430 | 15,581,000 | 817.9 |
| 75 + | 130,959 | 9,969,000 | 1313.7 |
| Total | 416,481 | 226,546,000 | 183.8 |

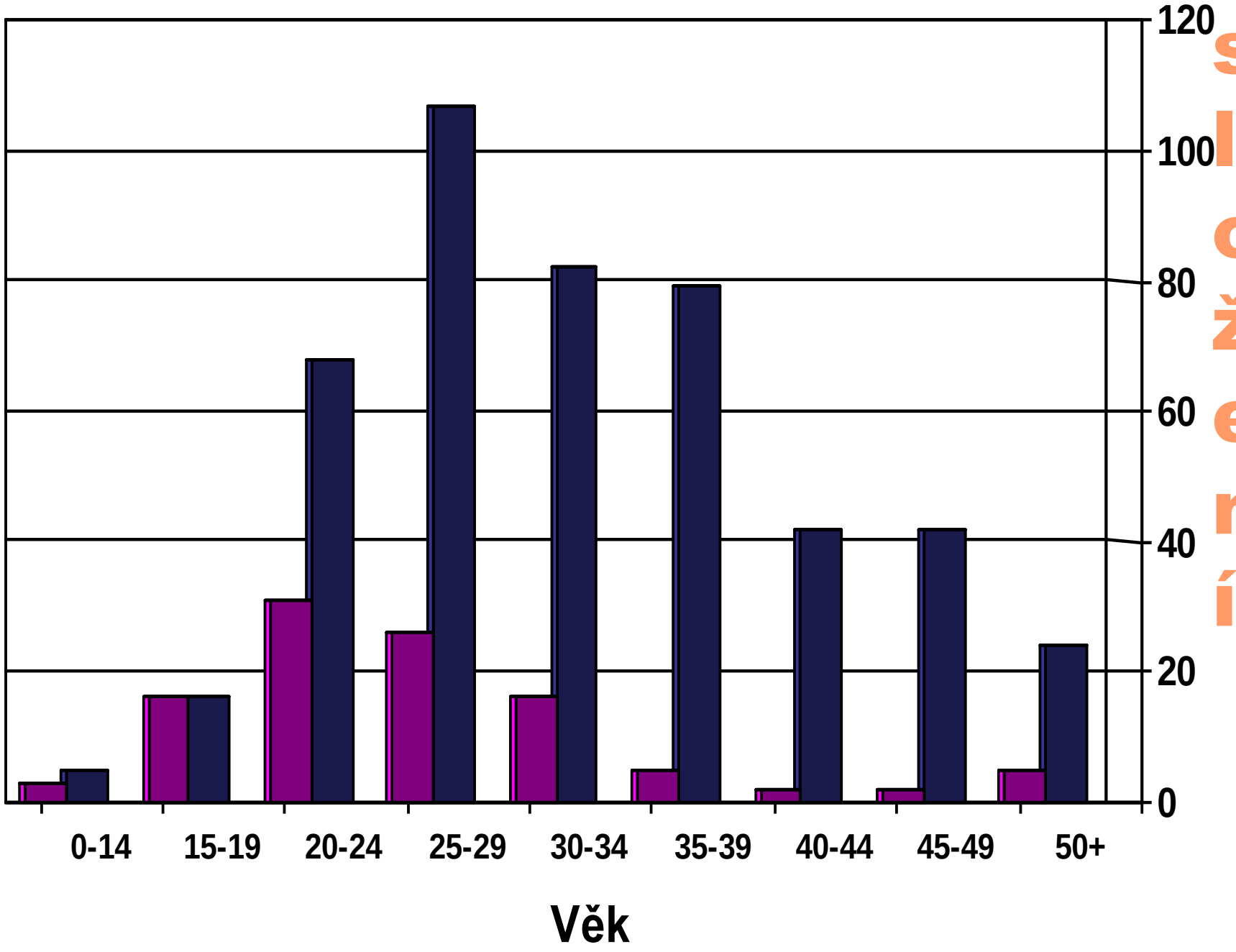
Source: Data from U.S. Bureau of the Census, *Statistical Abstract of the United States: 1984* (104th ed.). Washington, DC: 1983; and U.S. D.H.H.S., *Vital Statistics of the U.S., 1980*. Vol. II, Mortality, Part B. D.H.H.S. Publication No. (P.H.S.) 85–1102. Washington, DC: National Center for Health Statistics, 1985.

*Table 4-8.* Calculation of age-adjusted cancer mortality rates in the U.S., using the 1940 U.S. population as the standard

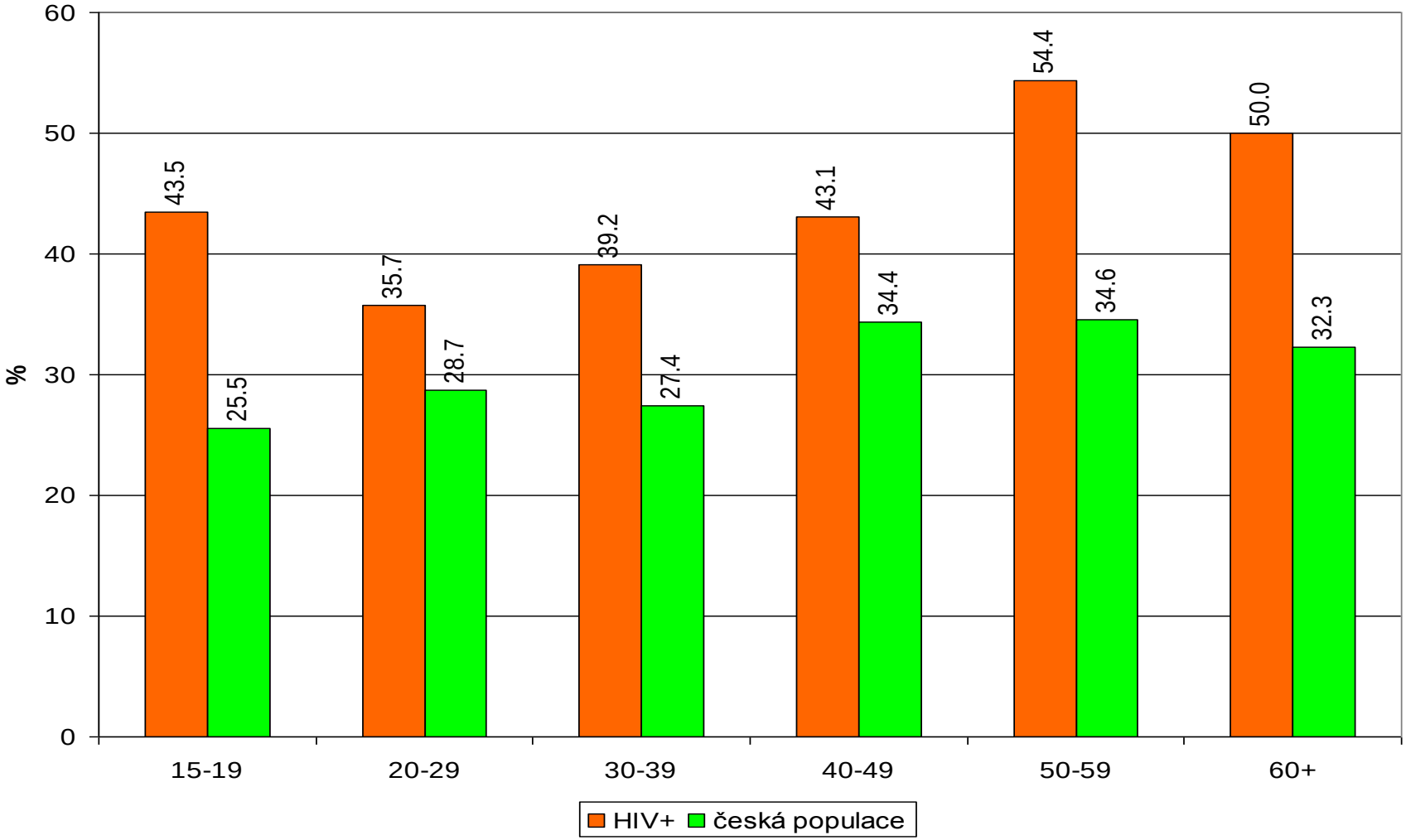
| Age in years | 1940 population (in thousands) | 1980 cancer mortality rates per 100,000 | Expected deaths in 1980 |
|---|---|---|---|
| < 5 | 10,541 | 4.2 | 442.7 |
| 5–9 | 10,685 | 4.7 | 502.2 |
| 10–14 | 11,746 | 3.9 | 458.1 |
| 15–19 | 12,334 | 5.4 | 666.0 |
| 20–24 | 11,588 | 7.2 | 834.3 |
| 25–29 | 11,097 | 10.5 | 1165.2 |
| 30–34 | 10,242 | 17.3 | 1771.9 |
| 35–39 | 9,545 | 33.5 | 3197.6 |
| 40–44 | 8,788 | 66.7 | 5861.6 |
| 45–49 | 8,255 | 128.3 | 10591.2 |
| 50–54 | 7,257 | 228.9 | 16611.3 |
| 55–59 | 5,844 | 358.2 | 20933.2 |
| 60–64 | 4,728 | 525.8 | 24859.8 |
| 65–74 | 6,377 | 817.9 | 52157.5 |
| 75 + | 2,643 | 1313.7 | 34721.1 |
| Total | 131,670 | | 174,773.7 |

$$\text{Age-adjusted cancer mortality rate} = \frac{174{,}773.7}{131{,}670{,}000} = 132.7/10^5$$

# Věková struktura



# Séroprevalence toxoplasmózy u mužů (%) – porovnání s obecnou populací



# ANALÝZY PŘEŽITÍ

## Survivorship analyses

**Doba**

**od vstupu do studie** (infekce, terapeutický zásah…)

**do smrti nebo nástupu příznaků** či výrazného zhoršení stavu

# TABULKA PŘEŽITÍ

## Life table

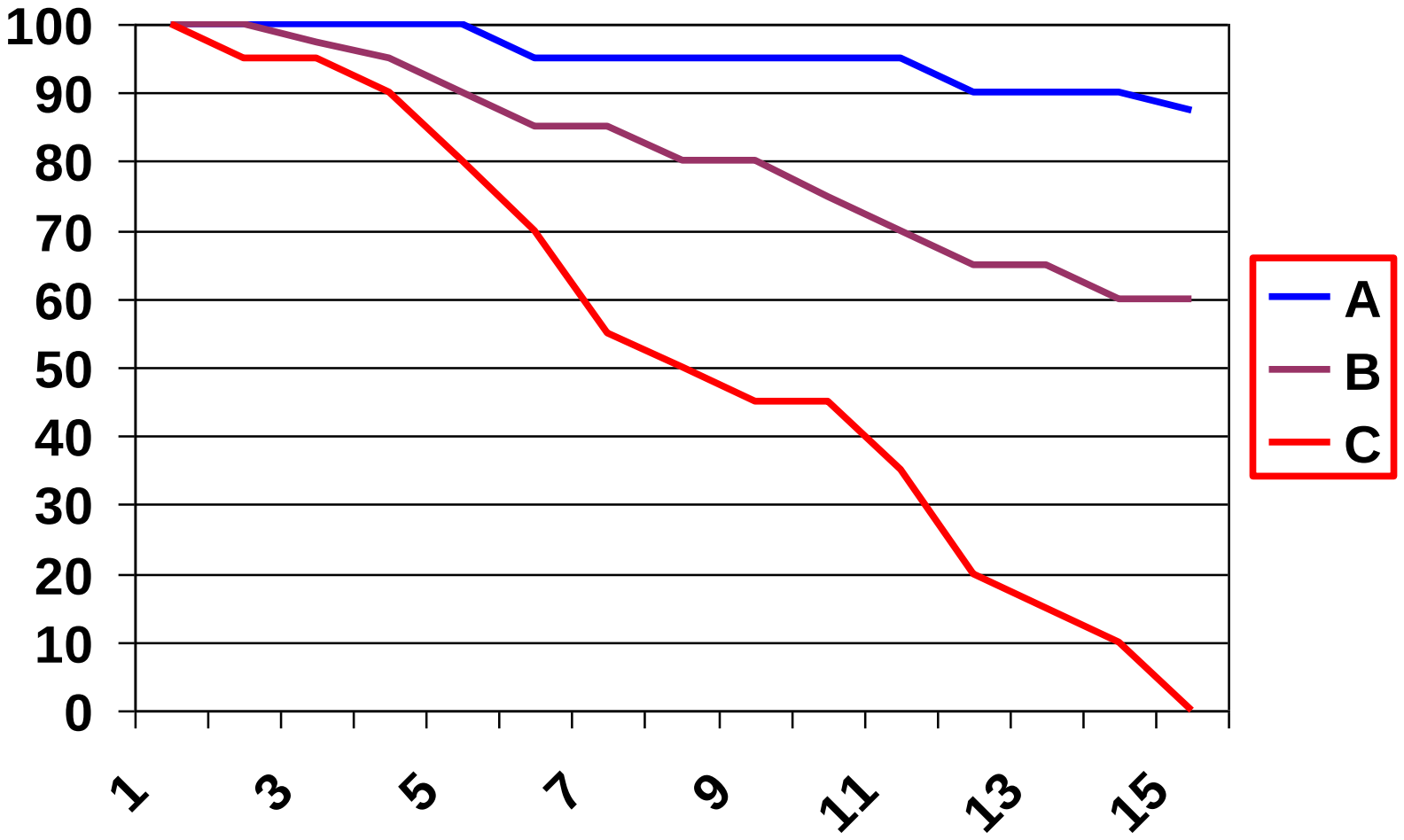
**Table A–15.** Example of a Life Table for 500 Persons Using a New Therapy for Parkinson's Disease

| YEAR ($x$) | NUMBER OF PERSONS WITH STABLE DISEASE STATUS AT START OF YEAR ($Q_x$) | NUMBER OF PERSONS WHOSE DISEASE STATUS DECLINED DURING YEAR ($_nd_x$) | PROBABILITY OF DISEASE STATUS DECLINING ($_nq_x$) | PROBABILITY OF DISEASE STATUS REMAINING STABLE ($_np_x$) |
|---|---|---|---|---|
| 1 | 500 | 100 | 0.22 | 0.78 |
| 2 | 400 | 100 | 0.29 | 0.71 |
| 3 | 300 | 150 | 0.67 | 0.33 |
| 4 | 150 | 90 | 0.86 | 0.14 |
| 5 | 60 | 38 | 0.93 | 0.07 |

Pravděpodobnost zhoršení za rok: počet zhoršených za rok / Počet lidí v riziku zhoršení v polovině roku

**Pravděpodobnost, že se za 5 let nezhorší = 0,78 x 0,71 x 0,33 x 0,14 x 0,07 = 0,0016**

# Křivky přežití



# Ztracená léta života v důsledku nemoci - DALY

- - **Disability Adjusted Life Year** je měřítko či ukazatel míry zátěže způsobené nemocí v lidské populaci (WHO).
- 1 DALY = 1 ztracený rok zdravého života.
- DALY = počet let, o který nemoc zkrátí život + počet let neplnohodnotného života (poškození kvality života nemocí)
- Součet všech DALY celé populace vyjadřuje zátěž způsobenou nemocí: rozdíl mezi skutečným stavem a ideálním stavem – bez nemoci.
- Koeficient závažnosti nemoci při výpočtu DALY: v rozpětí od 0 (dokonalý zdravý život) po 1 (smrt)

# Metaanalýza

- Z dat starších studií vytvořit studie nové
- Prověřit správnost provedení starších studií
- Kde je to možné, odstranit bias, confounding
- Využít maximum dat

# INTERAKCE
## = modifikace účinku

Při určitém souběhu některých rizikových faktorů se se výsledné relativní riziko nesčítá, ale násobí.

Například nekojení (někdy i odstavení) v nízkém (do 2 měsíců) věku enormně zvyšuje riziko úmrtí na průjmová onemocnění (Brazílie:  OR=23,3!).

**Mechanismus** ne vždy objasněn.

**Zjištění:** regresní modely, stratifikovaná analýza.

# Kritéria kauzálního vztahu rizikového faktoru k nemoci (dle BRADFORDA HILLA)

1. Je vyloučena náhoda, bias a confounding
2. SÍLA ASOCIACE: – vysoké relativní riziko
3. PLAUSIBILITA: zjištěný vztah je v souladu s předpokládaným biologickým mechanismem, též ANALOGIE s jinými nemocemi.
4. KOHERENCE: ke stejnému závěru dospělo více nezávislých studií
5. ČASOVÝ SLED: příčina předchází následek
6. VZTAH DÁVKY A ÚČINKU: kvantitativní souvislost mezi dávkou (exposicí) a frekvencí (průběhem) nemoci

# Kauzální vztah faktoru k nemoci je doložen, když

1. Je vyloučena náhoda, bias a confounding
2. Asociace je velmi silná
3. Zjištěný vztah je v souladu s předpokládaným biologickým mechanismem
   PLAUSIBILITA =„dává to smysl“
4. Ke stejnému výsledku dospělo více nezávislých studií
5. Existuje kvantitativní souvislost mezi dávkou a frekvencí (průběhem) nemoci

# Nepřátelé epidemiologa I

## Náhoda

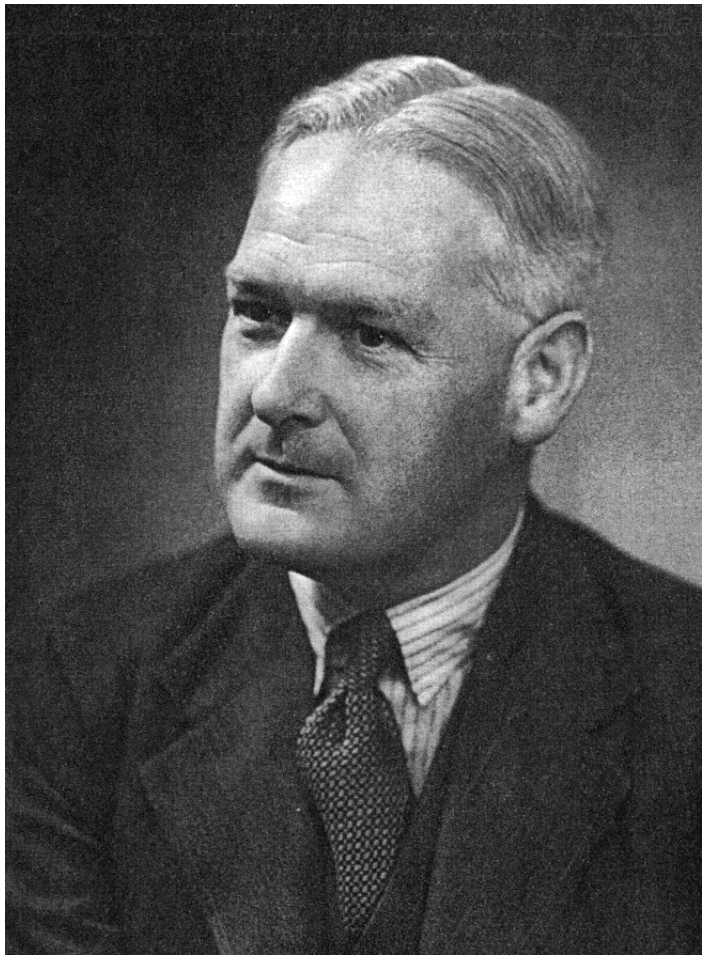
**P: pravděpodobnost, že asociace je nahodilá;**

**P < 5 %**

**Signifikance nevyjadřuje pravdu, ale pravděpodobnost**

*Signifikantní: výsledek může být nahodilý, ale je to málo pravděpodobné*

*(záleží na velikosti vzorku)*



**„The glitter of the t-table diverts attention of the inadequacy of the fare“**

**(Sir Austin Bradford Hill)**