

(12) Odhadovanie parametrov pravdepodobnostných rozdelení

1. Metóda momentov

S týmto sme sa už stretli, napríklad sme odhadovali parameter Poissonovho rozdelenia tak, že sme λ ako priemer s vysvetlením, že skutočná λ je strednou hodnotou, a preto ju odhadneme pomocou priemeru, ktorý je odhadom strednej hodnoty.

Takýto postup sa nazýva **metóda momentov**:

- Ak má rozdelenie jeden parameter, tak ho odhadneme tak, aby sa stredná hodnota získaného rozdelenia zhodovala s priemerom dát.
- Ak má rozdelenie dva parametre, tak ich odhadneme tak, aby sa
 - stredná hodnota získaného rozdelenia zhodovala s priemerom dát
 - stredná hodnota druhej mocniny zhodovala s priemerom druhých mocnín dát
- Iný prístup – chceme, aby sa
 - stredná hodnota získaného rozdelenia zhodovala s priemerom dát
 - disperzia získaného rozdelenia zhodovala s odhadom disperzie z dát
- Analogicky pri viacerých parametroch.

2. Metóda maximálnej vierohodnosti

<https://docs.scipy.org/doc/scipy-1.11.4/reference/generated/scipy.stats.fit.html>

- Doteraz sme predpokladali, že poznáme parameter rozdelenia a počítali sme pravdepodobnosti.
- Teraz je to naopak, máme dáta a chceme určiť parametre. Môžeme sa ale pýtať: Aká je pre danú hodnotu parametra pravdepodobnosť, že sa nadobudnú tie hodnoty, ktoré vidíme? Ako odhad zvolíme tú hodnotu parametra, ktorá uvedenú pravdepodobnosť maximalizuje.
- V prípade, že ide o spojité rozdelenie, uvažujeme funkčnú hodnotu združenej hustoty namiesto pravdepodobnosti.
- Terminológia:
 - Pravdepodobnosť pozorovaných dát, resp. hodnota združenej hustoty ako funkcia parametrov – **funkcia vierohodnosti, likelihood function**
 - Logaritmus tejto funkcie (analyticky sa jednoduchšie hľadá maximum, výhodnejší je aj numericky) – **logaritmická funkcia vierohodnosti, likelihood function**
 - Odhad parametrov získaný týmto spôsobom – **odhad metódou maximálnej vierohodnosti, maximum likelihood estimate (MLE)**
- Niekedy sa zhoduje s odhadom metódou momentov, ale nie vždy.

Príklad 1:

Pozorujeme hodnoty `data = [3, 10, 9, 8, 5, 7, 7]` a chceme ich modelovať pomocou Poissonovho rozdelenia. Parameter odhadneme metódou maximálnej virohodnosti.

Riešenie:

- Poissonovo rozdelenie:

$$P(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots$$

- Funkcia virohodnosti:

$$L(\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-n\lambda} \lambda^{\sum x_i} \prod \frac{1}{x_i!}$$

- Logaritmus funkcie virohodnosti:

$$\ell(\lambda) = \sum_{i=1}^n (x_i \log \lambda - \lambda - \log(x_i!))$$

```
import numpy as np
from scipy.stats import poisson

data = [3, 10, 9, 8, 5, 7, 7]

def poisson_likelihood(lam, data):
    likelihood = 1.0
    for x in data:
        likelihood *= poisson.pmf(x, lam)
    return likelihood

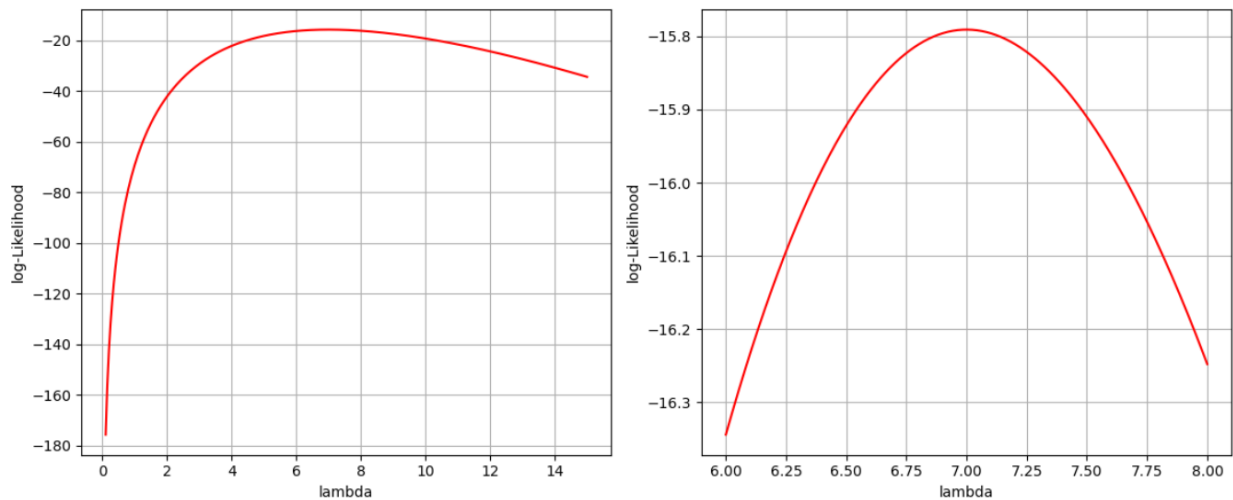
def poisson_loglikelihood(lam, data):
    log_likelihood = 0.0
    for x in data:
        log_likelihood += np.log(poisson.pmf(x, lam))
    return log_likelihood
```

Pre zadanú hodnotu parametra vieme vyčísliť funkčnú hodnotu.

```
lambda_val = 5
print(f"Likelihood: {poisson_likelihood(lambda_val, data):.10f}")
print(f"Log-Likelihood: {poisson_loglikelihood(lambda_val, data):.5f}")
```

```
Likelihood: 0.0000000115
Log-Likelihood: -18.27796
```

Chceme nájsť takú hodnotu parametra, pri ktorej je funkčná hodnota maximálna.



```
from scipy import stats
import numpy as np

data = [3, 10, 9, 8, 5, 7, 7]
dist = stats.poisson

# lambda: (0.00001; 1000), loc = 0
res = stats.fit(dist, data, bounds=[(1e-5, 1000), (0, 0)])
print(res)
```

```
params: FitParams(mu=np.float64(6.999999444060399), loc=np.float64(0.0))
success: True
message: 'Optimization terminated successfully.'
```

```
lam = res.params[0]
print(f"ML odhad parametra lambda: {lam:.5f}")
```

```
ML odhad parametra lambda: 7.00000
```

Lambdu môžeme ohraničiť na základe grafu:

```
res = stats.fit(dist, data, bounds=[(6, 8), (0, 0)]) # ohrnicenie pre lambda z grafu  
print(res)
```

```
params: FitParams(mu=np.float64(7.000001163976929), loc=np.float64(0.0))  
success: True  
message: 'Optimization terminated successfully.'
```

```
lam = res.params[0]  
print(f"ML odhad parametra lambda: {lam:.5f}")
```

```
ML odhad parametra lambda: 7.00000
```

Pre Poissonovo rozdelenie sa dá nájsť odhad analyticky:

$$\ell(\lambda) = \log L(\lambda) = \sum_{i=1}^n [x_i \log \lambda - \lambda - \log(x_i!)]$$

Hľadáme bod, kde je derivácia nulová:

$$\frac{d\ell}{d\lambda} = \sum_{i=1}^n \left(\frac{x_i}{\lambda} - 1 \right) = \frac{1}{\lambda} \sum_{i=1}^n x_i - n.$$

$$\frac{1}{\lambda} \sum_{i=1}^n x_i - n = 0 \Rightarrow \hat{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i$$

Strieškou sme označili výsledný odhad. Vidíme, že sa zhoduje s odhadom metódou momentov.

Príklad 2:

Odvodíme ML odhad parametra exponenciálneho rozdelenia.

Riešenie:

Funkcia vierohodnosti:

$$L(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

Jej logaritmus:

$$\ell(\lambda) = \log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n x_i$$

Hľadáme maximum:

$$\frac{d\ell}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i$$

$$\frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{X}}$$

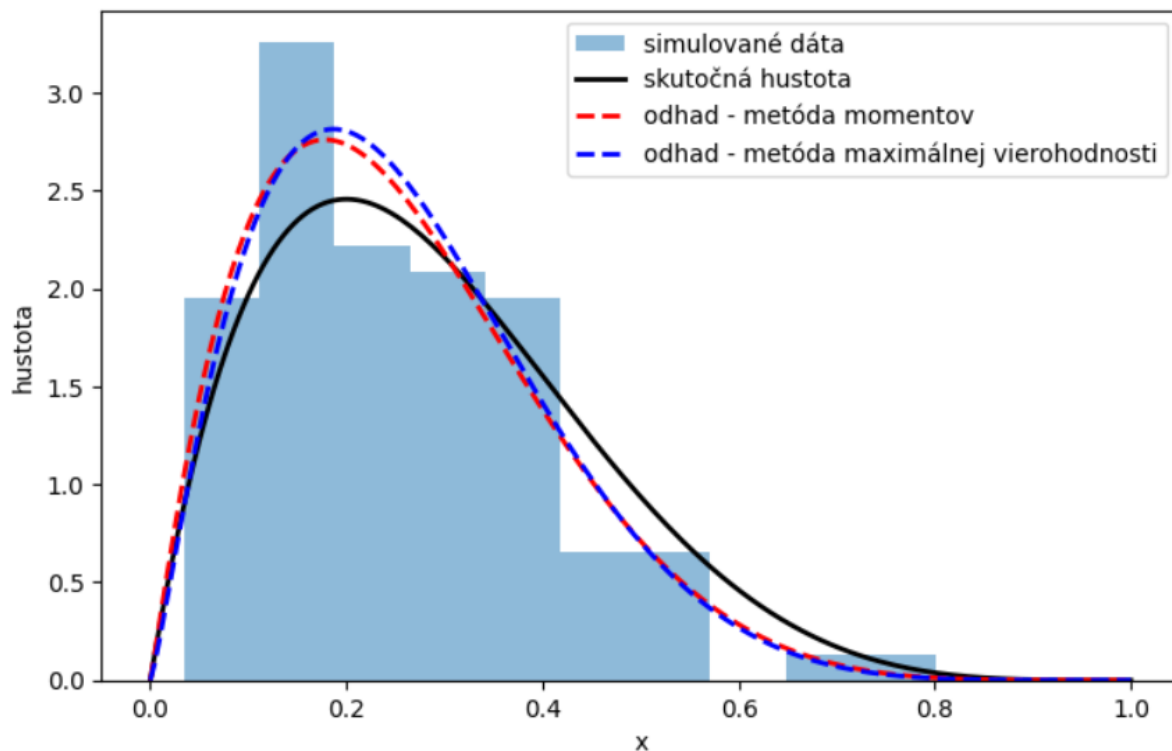
Parameter *lambda* sa odhadne ako prevrátená hodnota priemeru. To sa znovu zhoduje s metódou momentov.

Príklad 3:

Uvažujme tzv. **beta rozdelenie**, ktoré nadobúda hodnoty z intervalu (0, 1) a je charakterizované dvoma parametrami – <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.beta.html>
Vygenerujeme z neho dáta a späťne odhadneme parametre – uvedieme len výsledok ako ukážku toho, že metóda momentov a metóda maximálnej vierohodnosti nemusí dať rovnaký výsledok:

Metóda momentov: alpha=2.07588, beta=5.96535

Metóda maximálnej vierohodnosti: alpha=2.22291, beta=6.32296



3. Nevychýlenosť odhadov

Majme dáta, o ktorých vieme, že sú to nezávislé náhodné premenné, ktoré pochádzajú z rovnomerného rozdelenia na intervale $(0, b)$. Chceme odhadnúť parameter b .

Ak použijeme metódu momentov, dostaneme odhad, ktorého stredná hodnota je skutočná hodnota parametra:

X_i - nezávislé náhodné premenné s rovnomerným rozdelením na intervale $(0, b)$

$$E(X_i) = \frac{1}{2}b \rightarrow \text{odhadnime parameter } b \text{ ako } \hat{b} = 2\bar{X}$$

$$E(\hat{b}) = 2E(\bar{X}) = 2 \cdot \frac{1}{2}b = b$$

To je dobrá vlastnosť, lebo hovorí, že "v priemere" sa odhad rovná skutočnej hodnote.

Tento konkrétny odhad však má aj nevýhodu, že môže byť menší ako maximálna pozorovaná hodnota. Teda napríklad naše dáta obsahujú hodnotu 10, ale my odhadneme parameter b ako napríklad 8, teda tvrdíme, že dáta sú z rovnomerného rozdelenia na $(0, 8)$. Aby sme sa takémuto protirečeniu vyhli, nájdeme iné odhady.

Výjdeme z myšlienky odhadnúť b ako maximum z dát, ktoré "o niečo" zväčšíme. To "o niečo" bude tak, aby sa stredná hodnota odhadu rovnala skutočnej hodnote parametra.

$$\tilde{b} = c \cdot \max_{i=1, \dots, n} X_i, \quad c \text{ bude závisieť od počtu dát } n$$

$$\text{Požadujeme splnenie podmienky } E(\tilde{b}) = b$$

$$M = \max_{i=1, \dots, n} X_i$$

$$\begin{aligned} F_M(m) &= P(\max X_i \leq m) = P(X_1 \leq m, \dots, X_n \leq m) \stackrel{\text{nezávislosť}}{=} \\ &= P(X_1 \leq m) \cdot P(X_2 \leq m) \cdots P(X_n \leq m) = \left(\frac{m}{b}\right)^n = \\ &= m^n \cdot \frac{1}{b^n} \end{aligned}$$

$$f_M(m) = F'_M(m) = n \cdot m^{n-1} \cdot \frac{1}{b^n} \quad \text{pre } m \in (0, b), \quad \text{inak } f=0$$

$$E(M) = \int_{-\infty}^{\infty} x f_M(x) dx = \int_0^b x \cdot \frac{n \cdot x^{n-1}}{b^n} dx = \frac{n}{b^n} \int_0^b x^n dx = \frac{n}{b^n} \left[\frac{x^{n+1}}{n+1} \right]_0^b = \frac{n}{n+1} \frac{b^{n+1}}{b^n}$$

$$E(\tilde{b}) = E(cM) = cE(M) = c \cdot \frac{n}{n+1} b = \frac{n}{n+1} b$$

Ak ma' byť $E(\tilde{b}) = b$, musíme zobrať $c = \frac{n+1}{n}$

Odhad teda bude $\tilde{b} = \frac{n+1}{n} \cdot \max_{i=1, \dots, n} X_i$

Vyskúšajme rovnakým spôsobom využiť minimum:

$$m = \min_{i=1, \dots, n} X_i$$

$$\begin{aligned} F_m(x) &= P(\min X_i \leq x) = 1 - P(\min X_i > x) = \\ &= 1 - P(X_1 > x, \dots, X_n > x) \stackrel{\text{nez.}}{=} \\ &= 1 - P(X_1 > x) \dots P(X_n > x) = \\ &= 1 - \left(\frac{b-x}{b} \right)^n = 1 - \frac{1}{b^n} (b-x)^n \end{aligned}$$

$$f_m(x) = -\frac{n}{b^n} (b-x)^{n-1} \cdot (-1) = \frac{n}{b^n} (b-x)^{n-1}$$

$$\begin{aligned} E(m) &= \int_{-\infty}^{\infty} x f_m(x) dx = \int_0^b x \cdot \frac{n}{b^n} (b-x)^{n-1} dx = \frac{n}{b^n} \int_0^b x (b-x)^{n-1} dx = \\ &= \frac{n}{b^n} \cdot \frac{b^{n+1}}{n(n+1)} = \frac{1}{n+1} b \end{aligned}$$

Posledný integrál: `import sympy as sp`

```
x, b = sp.symbols('x b', real=True, positive=True)
n = sp.symbols('n', integer=True, positive=True)

integrand = x * (b - x)**(n-1)

integral = sp.integrate(integrand, (x, 0, b))
print("Integrál funkcie 'x * (b - x)**(n-1)' od 0 do b:")
print(integral)
```

Integrál funkcie 'x * (b - x)**(n-1)' od 0 do b:
 $b^{n+1} / (n+1)$

Takže aj odhad $b^* = (n+1) \min_{i=1..n} X_i$ bude spĺňať $E(b^*) = b$.

Všimnime si však, že aj v tomto prípade môže byť niektoré pozorovanie väčšie ako odhad hornej hranice.

Toto nenastane v predchádzajúcom prípade (maximum vynásobené konštantou) anie v tomto nasledujúcom:

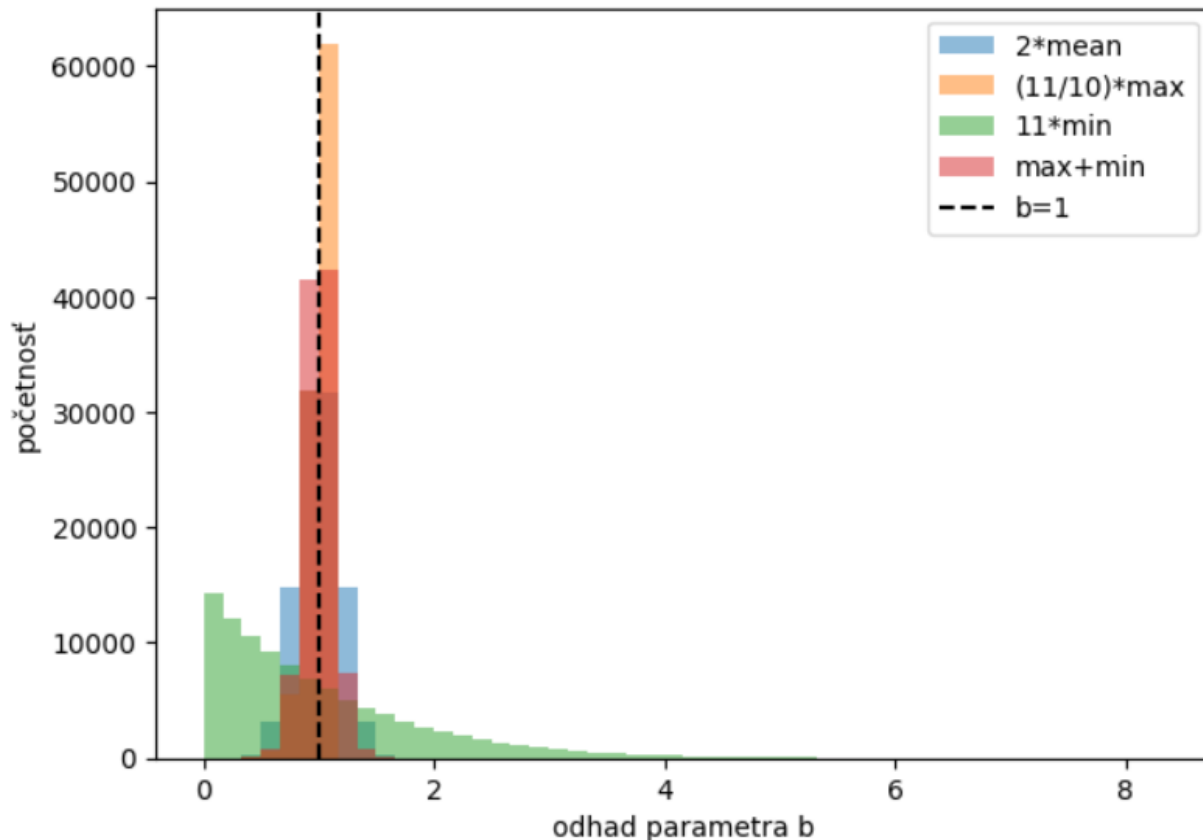
$$b^{**} = \max_{i=1..n} X_i + \min_{i=1..n} X_i$$

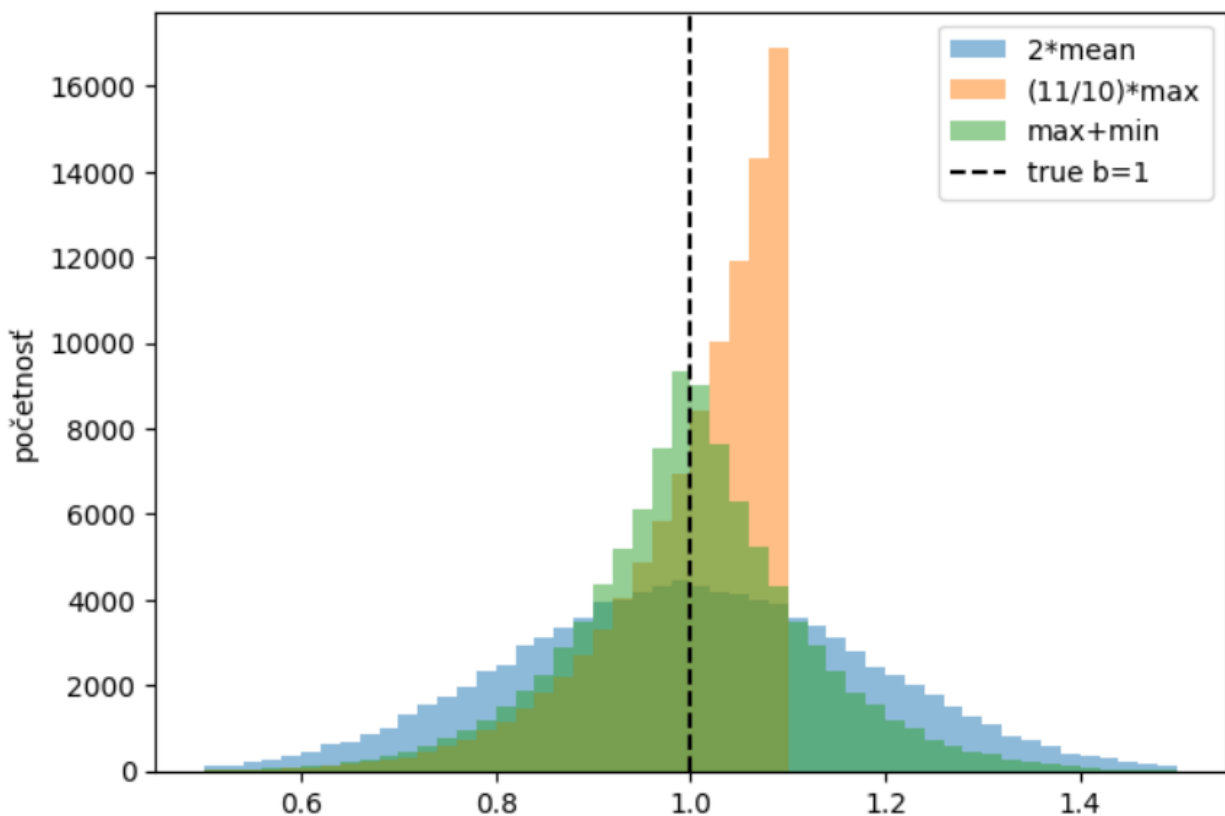
$$E(b^{**}) = E(\max X_i) + E(\min X_i) = \frac{n}{n+1} b + \frac{1}{n+1} b = \left(\frac{n}{n+1} + \frac{1}{n+1} \right) b = b$$

Odhad, ktorého stredná hodnota sa rovná skutočnej hodnote parametra, sa nazýva nevychýlený odhad.

Všetky uvedené odhady sú teda nevychýlené.

Nie sú však rovnocenné, spravme simuláciu pre odhadovanie z 10 dát:





odhad 2*mean: priemer = 0.9996, disperzia = 0.0333
 odhad (11/10)*max: priemer = 1.0001, disperzia = 0.0084
 odhad 11*min: priemer = 1.0044, disperzia = 0.8414
 odhad max+min: priemer = 1.0005, disperzia = 0.0153

4. Priemer a výberová disperzia

Teraz vieme vysvetliť, prečo je pri výpočte výberovej disperzie v menovateli **n-1**, nie **n**. Dôvod je ten, že takto je to nevychýlený odhad disperzie:

- Pre normálne rozdelenie:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2, \text{ označme túto nah. prem. } Y, \text{ t.j. } Y = \frac{(n-1)S^2}{\sigma^2}$$

$$E(S^2) = E\left(\frac{\sigma^2}{(n-1)} Y\right) = \frac{\sigma^2}{n-1} \underbrace{E(Y)}_{=n-1} = \frac{\sigma^2}{n-1} \cdot (n-1) = \sigma^2$$

Využili sme tu, že stredná hodnota chí-kvadrátu je počet jeho stupňov voľnosti.

- Vo všeobecnosti:

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n ((X_i - \mu) - (\bar{X} - \mu))^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

$$E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = \sum_{i=1}^n E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2]$$

V sume sa sčítajú disperzie náhodnej premennej X , stredná hodnota v druhom člene je vlastne disperzia priemeru. Teda:

$$E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

Takže aj tu dostaneme, že výberová disperzia s menovateľom **n-1** je nevychýlený odhad disperzie.

Odhad s menovateľom **n-1** to nespĺňa, ale “nie je veľmi zlý” – v limite pre n idúce do nekonečna sa jeho stredná hodnota tiež rovná skutočnej hodnote. Takýto odhad sa nazýva **asymptoticky nevychýlený**.

$$\begin{aligned} E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) &= E\left(\frac{n-1}{n} \cdot \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right) = \\ &= \frac{n-1}{n} \underbrace{E\left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right)}_{\sigma^2} = \frac{n-1}{n} \sigma^2 = \\ &= \left(1 - \frac{1}{n}\right) \sigma^2 \rightarrow \sigma^2 \text{ pre } n \rightarrow \infty \end{aligned}$$

Platí tiež, že **priemer je nevychýlený odhad strednej hodnoty**.

Aj pre priemer, aj pre výberovú disperziu platí, že ich disperzia konverguje k nule, keď počet dát **n** ide do nekonečna:

- Priemer ako odhad strednej hodnoty:

$$E(X_i) = \mu, \quad D(X_i) = \sigma^2$$

$$\Rightarrow E(\bar{X}) = \mu$$

$$D(\bar{X}) = \frac{\sigma^2}{n} \rightarrow 0 \text{ pre } n \rightarrow \infty$$

- Výberová disperzia (menovateľ **n-1**) ako odhad disperzie, spravíme dôkaz len pre normálne rozdelenie (môžeme využiť vlastnosti chí-kvadrát rozdelenia, čo výpočet zjednoduší):

$$D(S^2) = D\left(\frac{\sigma^2}{n-1} Y\right) = \left(\frac{\sigma^2}{n-1}\right)^2 \underbrace{D(Y)}_{2(n-1)} = \frac{\sigma^4}{(n-1)^2} \cdot 2(n-1) = \frac{2\sigma^4}{n-1}$$

$$\Rightarrow D(S^2) \rightarrow 0 \text{ pre } n \rightarrow \infty$$

Využili sme tu, že disperzia chí-kvadrátu je dvojnásobok počtu stupňov voľnosti.