Matematika III – 11. týden
Kovariance, momentová funce a centrální limitní věta, zpět ke statistice

Jan Slovák

Masarykova univerzita
Fakulta informatiky

5. 12. – 9. 12. 2016

Obsah přednášky

Kde je dobré číst?

- Karel Zvára, Josef Štěpán, Pravděpodobnost a matematická pravděpodobnost statistika, Matfyzpress, 2006, 230pp.
- J. Slovák, M. Panák, M. Bulant, Matematika drsně a svižně, Muni Press, Brno 2013, v+773 s., elektronická edice `www.math.muni.cz/Matematika_drsne_svizne`
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, Teorie pravděpodobnosti a matematická statistika (sbírka příkladů), Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, Základní statistické metody, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.
- Riley, K.F., Hobson, M.P., Bence, S.J. Mathematical Methods for Physics and Engineering, second edition, Cambridge University Press, Cambridge 2004, ISBN 0 521 89067 5, xxiii + 1232 pp.

Kovariance veličin

Jsou-li X a Y dvě náhodné veličiny, pro které existují jejich konečné royptyly, pak definijeme jejich **kovarianci** vztahem

$$\operatorname{cov}(X, Y) = \mathrm{E}(X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y).$$

Evidentně je $\operatorname{cov}(X, X) = \operatorname{var} X$ a $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$.

Theorem

Nechť existují konečné rozptyly veličin X a Y. Pak

- $\operatorname{cov}(X, Y) = \mathrm{E}(XY) - (\mathrm{E}\,X)(\mathrm{E}\,Y)$
- *pro jakékoliv skaláry a, b, c, d platí* $\operatorname{cov}(a + bX, c + dY) = bd \operatorname{cov}(X, Y)$
- $\operatorname{var}(X + Y) = \operatorname{var} X + \operatorname{var} Y + 2 \operatorname{cov}(X, Y)$.

Od kovariance snadno odvodíme tzv. **korelační koeficient** dvou náhodných veličin X a Y. Definujeme jej jako kovarianci příslušných normovaných veličin:

$$\rho_{X,Y} = \operatorname{cov}\left(\frac{X - \mathrm{E}\,X}{\sqrt{\operatorname{var} X}}, \frac{Y - \mathrm{E}\,Y}{\sqrt{\operatorname{var} Y}}\right) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X \operatorname{var} Y}}.$$

Theorem

- $\rho_{a+bX,c+dY} = \operatorname{sign}(bd)\rho_{X,Y}$, *pro* $bd \neq 0$
- $\rho_{X,X} = 1$
- $\rho_{X,Y} = 0$, *pokud jsou veličiny* X *a* Y *nezávislé.*
- *pokud je* $\rho_{X,Y}$ *definován, pak je roven jedné právě, když existují konstanty* a, b, c *tak, že* $P(aX + bY = c) = 1$.

Varianční matice

Uvažme náhodný vektor $W = (X_1, \ldots, X_n)$ takový, že pro všechny jeho komponenty existuje rozptyl. Pak **varianční matice** $\operatorname{var} W$ je dána

$$\operatorname{var} W = \begin{pmatrix} \operatorname{var} X_1 & \operatorname{cov}(X_1, X_2) & \ldots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_2, X_1) & \operatorname{var} X_2 & \ldots & \operatorname{cov}(X_2, X_n) \\ & & \ldots & \\ \operatorname{cov}(X_n, X_1) & \operatorname{cov}(X_n, X_2) & \ldots & \operatorname{var} X_n \end{pmatrix}.$$

Theorem

Pro náhodný vektor X, skaláry a, matice skalárů B platí

$$\operatorname{var}(a + BX) = B \operatorname{var} X B^T.$$

Momenty

Podobně jako rozptyl můžeme uvažovat výrazy vyšších řádů:

$$\mu'_k = \mathrm{E}\, X^k, \quad \mu_k = \mathrm{E}(X - \mathrm{E}\, X)^k.$$

Nazýváme je k-tý moment a k-tý centrální moment náhodné veličiny X. Momenty lze všechny dostat jako koeficienty v mocninné řadě následujícím způsobem.
Pro volný reálný parametr t definujeme **momentovou vytvořující funkci** pro náhodnou veličinu X vztahem

$$M_X(t) = \mathrm{E}\, e^{tX}.$$

Tato funkce (za docela rozumných předpokladůná sledující věty) zcela určuje náhodné veličiny a má řadu užitečných vlastností (tj. *stejná momentová funkce na nějakém netriviálním intervalu* $\Longrightarrow$ *stejná distribuční funkce*).

Theorem

Nechť X je náhodná veličina pro kterou na intervalu $(-a, a)$ existuje její analytická momentová vytvořující funkce. Pak na tomto intervalu je $M_X(t)$ dána absolutně konvergující řadou

$$M_t(X) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathrm{E}\, X^k.$$

Theorem

Pro součet náhodných veličin platí:

$$M_{X+Y}(t) = M_X(t) M_Y(t).$$

Momentová vytvořující funkce pro $X \sim \mathrm{Bi}(0, 1)$

Často je jednodušší počítat momenty z jejich vytvořující funkce než přímo.
Pro alternativní rozdělení náhodné veličiny $Y \sim A(p)$ spočteme snadno

$$M_Y(t) = \mathrm{E}\, e^{tY} = e^0(1-p) + e^t p = p(e^t - 1) + 1.$$

Protože je binomické rozdělení $X \sim \mathrm{Bi}(n, p)$ dáno jako součet n alternativních rozdělení $Y_i \sim \mathrm{A}(p)$, je zjevně v tomto případě

$$M(t) = M_X(t) = (p(e^t - 1) + 1)^n.$$

Obecně platí $\mu'_k = \frac{d^r}{dt^r} M_X(t)|_{t=0}$. Je tedy např. první moment binomického rozdělení skutečně np (první derivace $M(t)$ v nule), což je střední hodnota. Druhý moment je $np(1-p)$, čímž jsme ověřili výsledek pro rozptyl.

Momentová vytvořující funkce pro $Z \sim \mathrm{N}(0,1)$

$$
\begin{aligned}
M_Z(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2-2tx+t^2-t^2}{2}\right) dx \\
&= \exp(t^2/2) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2}\right) dx \\
&= \exp(t^2/2).
\end{aligned}
$$

(V předposledním řádku je integrálem dána pravděpodobnost jakékoliv hodnoty pro normální rozdělení, proto je to jednička.)
Derivováním: $(M_Z)'(0) = 0$ a $(M_Z)''(0) = (te^{t^2/2})'(0) = 1$. Je tedy skutečně

$$\mathrm{E}\, Z = 0, \quad \mathrm{var}\, Z = 1.$$

Uvažme nezávislé náhodné veličiny Y_1, Y_2, ..., které mají všechny stejné rozdělení se střední hodnotou 0 a rozptylem 1.
Předpokládejme, že třetí absolutní moment $E|Y_i|^3$ je konečný.
Pro náhodnou veličinu $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$ spočtěme momentovou funkci (koeficient $n^{-1/2}$ je volen tak, aby rozptyl S_n byl stále 1)

$$M_{S_n} = \prod_{i=1}^{n} \mathrm{E}\, e^{(t/\sqrt{n})Y_i} = (M_Y(t/\sqrt{n}))^n,$$

kde M_Y je společná momentová funkce všech veličin Y_i. Nyní

$$M_Y(t/\sqrt{n}) = 1 + 0\frac{t}{\sqrt{n}} + 1\frac{t^2}{2n} + \mathrm{o}(t^2/n)$$

a v limitě proto dostáváme

$$\lim_{n\to\infty} M_{S_n}(t) = \lim_{n\to\infty} \left(1 + \frac{t^2}{2n} + \mathrm{o}(1/n)\right)^n = e^{t^2/2}.$$

To je právě momentová funkce pro rozdělení $\mathrm{N}(0,1)$**!**.

Tím jsme skoro dokázali:

Theorem (Centrální limitní věta)

Nechť $Y_1, Y_2, \ldots$ jsou nezávislé náhodné veličiny se společnou střední hodnotou $\mathrm{E}\, Y_i = \mu$, *rozptylem* $\mathrm{var}\, Y_i = \sigma^2 > 0$ *a konečným třetím absolutním momentem* $E|Y_i|^3$. *Pro distribuční funkce náhodných veličin*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\sigma}(Y_i - \mu)$$

platí

$$\lim_{n\to\infty} P(S_n < x) = \Phi(x),$$

kde $\Phi(x)$ *je distribuční funkce normálního rozdělení* $\mathrm{N}(0,1)$.

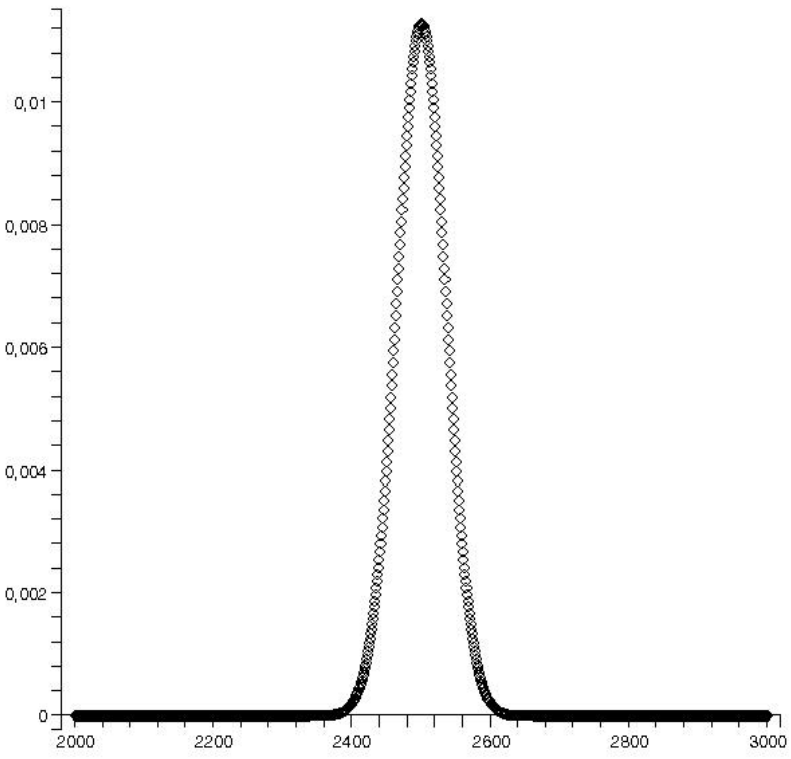
Všimněme si: součty $X_n = \sum_{i=1}^{n} Y_i$ mají střední hodnotu $n\mu$ a rozptyl $n\sigma^2$. Veličiny S_n jsou tedy právě normované veličiny X_n.

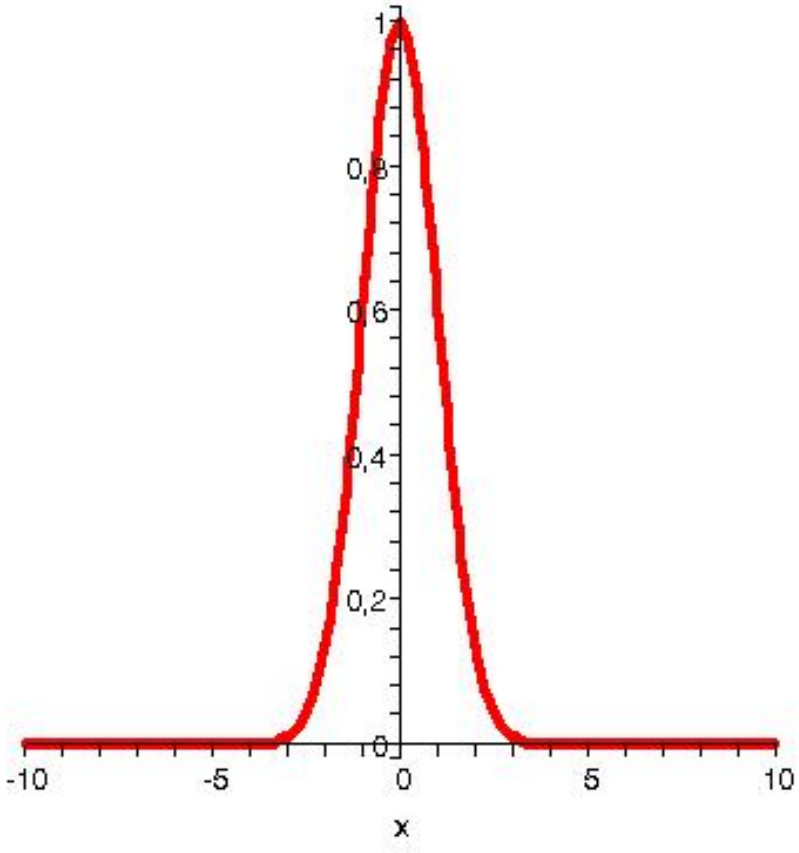
Pokud jsou $Y_i \sim \mathrm{A}(p)$ nezávislé, pak $\mathrm{E}(Y_i)^3 = p < \infty$ a všechny podmínky centrální limitní věty jsou splněny, $\mu = p$, $\sigma^2 = p(1-p)$. Součtové veličiny $X_n = \sum_{i=1}^{n} Y_i$ pak představují právě binomická rozdělení $\mathrm{Bi}(n, p)$ a příslušné normované veličiny jsou

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{Y_i - p}{\sqrt{p(1-p)}} \right) = \frac{X_n - np}{\sqrt{np(1-p)}}.$$

Podle centrální limitní věty má tato veličina pro velká n rozdělení velmi podobné rozdělení $\mathrm{N}(0, 1)$.
Jinými slovy, rozdělení $\mathrm{Bi}(n, p)$ je velice blízké rozdělení $\mathrm{N}(np, np(1-p))$ pro velká n. To je obsahem tzv. Laplaceovy–Moivreovy věty. To jsme už viděli minule na obrázcích:

Pro hodnoty Bi(5000, 0.5) je výsledek vidět na obrázku níže. Druhá křivka na obrázku je grafem funkce $f(x) = e^{-x^2/2}$.





Aproximace binomického rozdělení normálním se často považuje v praxi za dostatečnou, jestliže $np(1-p) > 9$

Při praktických průzkumech zpravidla věříme „zákonu velkých čísel“. Potřebujeme přitom rozhodnout, jak velký vzorek už postačuje.
Typickým příkladem je např. tato úloha: Chceme zjistit poměr p osob s danou krevní skupinou A v populaci. U kolika osob je třeba krevní skupinu skutečně zjistit, abychom měli 90% pravděpodobnost, že naše zjištění se nebude lišit o více než 5%.
Propočítáním zjistíme, že (nezávisle na p) vždy stačí odhadnout $p = X/n$, kde X je náhodná veličina udávající počet osob majících požadovanou skupinu, pro vzorek 270 lidí.

Rozdělení χ^2

Ve statistice budeme pracovat s charakteristikami náhodných vektorů, které budou obdobné výběrovému průměru a rozptylu, ale také s relativními poměry takových charakteristik atd. Podíváme se teď na několik takových případů.
Uvažme $Z \sim \mathrm{N}(0,1)$ a spočtěme hustotu $f_Y(x)$ pro $Y = Z^2$.
Evidentě je $f_Y(x) = 0$ pro $x \leq 0$, pro kladná x

$$F_Y(x) = P(Y < x) = P(-\sqrt{x} < Z < \sqrt{x})$$
$$= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}\, dz = \int_0^x \frac{1}{\sqrt{2\pi}} t^{-1/2}\, \mathrm{e}^{-t/2}\, dt.$$

Hustotu dostaneme derivací

$$f_Y(x) = \frac{d}{dx} F_Y(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2}\, \mathrm{e}^{-x/2}\,.$$

Tomuto rozdělení se říká χ^2 **s jedním stupněm volnosti**, píšeme $Y \sim \chi^2$.

Gama rozdělení $Y \sim \Gamma(a, b)$

Výběrový rozptyl bude odpovídat součtům takovýchto nezávislých veličin.
Uvažme hustotu (trochu obecnějšího tvaru než u χ^2)

$$f_X(x) = cx^{a-1}\,\mathrm{e}^{-bx}$$

pro $x > 0$, zatímco $f_X(x) = 0$ pro nekladná x (χ^2 odpovídá volbě $a = b = 1/2$).
Je třeba volit $c = \frac{b^a}{\Gamma(a)}$ a jde o rozdělení $\Gamma(a, b)$.

k-tý moment takové veličiny X je

$$\begin{aligned}\mathrm{E}\, X^k &= \int_0^\infty x^k \frac{b^a}{\Gamma(a)} x^{a-1}\, \mathrm{e}^{-bx}\, dx \\ &= \frac{\Gamma(a+k)}{\Gamma(a) b^k} \int_0^\infty \frac{b^{a+k}}{\Gamma(a+k)} x^{a-1+k}\, \mathrm{e}^{-bx}\, dx \\ &= \frac{\Gamma(a+k)}{\Gamma(a) b^k}\end{aligned}$$

(protože integrál z hustoty rozdělení $\Gamma(a+k, b)$ v posledním upravovaném výrazu je nutně roven jedné)
Zejména tedy vidíme, že $\mathrm{E}\, X = \frac{\Gamma(a+1)}{b\Gamma(a)} = \frac{a}{b}$, zatímco

$$\mathrm{var}\, X = \frac{\Gamma(a+2)}{b^2\Gamma(a)} - \frac{a^2}{b^2} = \frac{(a+1)a - a^2}{b^2} = \frac{a}{b^2}.$$

Momentová vytvořující funkci pro všechny hodnoty $-b < t < b$ je

$$\begin{aligned} M_X(t) &= \int_0^\infty \mathrm{e}^{tx} \frac{b^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-bx}\, dx \\ &= \frac{b^a}{(b-t)^a} \int_0^\infty \frac{(b-t)^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-(b-t)x}\, dx \\ &= \frac{b^a}{(b-t)^a}. \end{aligned}$$

Pro součet nezávislých rozdělení $Y = X_1 + \cdots + X_n$ s rozděleními $X_i \sim \Gamma(a_i, b)$ tedy okamžitě dostáváme momentovou vytvořující funkci (pro hodnoty $|t| < b$)

$$M_Y(t) = \left(\frac{b}{b-t}\right)^{a_1+\cdots+a_n},$$

tj. $Y \sim \Gamma(a_1 + \cdots + a_n, b)$. (Velmi podstatný je přitom předpoklad, že všechna gamma rozdělení sdílí stejnou hodnotu b).

rozdělení χ^2

Jako okamžitý důsledek nyní dostáváme hustotu rozdělení veličiny $Y = Z_1^2 + \cdots + Z_n^2$, kde všechna $Z_i \sim \mathrm{N}(0,1)$. Jde totiž o gamma rozdělení $Y \sim \grave{}(n/2, 1/2)$ a má hustotu

$$f_Y(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1}\, \mathrm{e}^{-x/2}\,.$$

Tomuto speciálnímu případu gamma rozdělení říkáme rozdělení χ^2 s n **stupni volnosti**. Značíme jej zpravidla $Y \sim \chi^2_n$.

F-rozdělení

Při prorovnání výběrových rozptylů potkáme veličiny, které jsou dány podílem

$$U = \frac{X/k}{Y/m}$$

$X \sim \chi^2_k$ a $Y \sim \chi^2_m$.
Náhodná veličina $U = \frac{X/k}{Y/m}$ má hustotu $f_U(u)$

$$f_U(u) = \frac{\Gamma((k+m)/2)}{\Gamma(k/2)\Gamma(m/2)} \left(\frac{k}{m}\right)^{k/2} u^{k/2-1} \left(1 + \frac{k}{m}u\right)^{-(k+m)/2}.$$

Takovému rozdělení se říká **Fisherovo-Snedecorovo rozdělení s** k **a** m **stupni volnosti**, zkráceně také **F-rozdělení**.

t-rozdělení

Další potřebné rozdělení se objevuje při zkoumání podílu veličin $Z \sim \mathrm{N}(0,1)$ a $\sqrt{X/n}$, kde $X \sim \chi^2_n$ (tj. zajímá nás poměr Z a směrodatné odchylky nějakého výběru).
Dostaneme náhodnou veličinu

$$T = \frac{Z}{\sqrt{X/n}}$$

a hustotou $f_T(t)$

$$f_T(t) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}.$$

Tomuto rozdělení říkáme **Studentovo t-rozdělení s n stupni volnosti**.

matematická statistika

Zkoumáme statistiky u nějakého výběru z daného základního souboru (populace).
Matematická statistika se snaží postihnout, do jaké míry jsou zjištěné výsledky relevantní pro celou populaci, případně se ze zjištěných dat pokouší zjistit nebo upřesnit vhodný teoretický model pro chování celého souboru (a z něj pak třeba odhadovat pravděpodobnost nějakého budoucího jevu).

Dva základní přístupy:

- **frekvenční statistika** (nebo také klasická statistika)
- **bayesovská statistika**.

frekvenční přístup

- Vychází z matematické abstrakce, že skutečné pravděpodobnosti jsou dány četnostmi výskytů jevů v tak velkých vzorcích dat, že je můžeme dobře aproximovat nekonečnými modely a využít pro odhady spolehlivosti centrální limitní věty.
- Statistik zde na pravděpodobnost pohlíží jako na idealizaci relativní četnosti případů, v nichž se vyskytne určitý výsledek při opakovaných pokusech.
- Tato zdánlivá výhoda/rigoróznost se může ale rychle stát nevýhodou, jakmile se začneme zabývat spolehlivostí samotných dat a vhodností zvoleného experimentu.
- Stejně tak je obtížné frekvenční statistiku dobře použít pro odhad pravděpodobnosti výskytu jednorázového děje.

Máme k dispozici (velký) základní statistický soubor s N jednotkami, který nazýváme **populace**, a zároveň nějaký číselný znak pro každou z jednotek, tj. soubor hodnot $(x_1, \ldots, x_N)$. Z něj ovšem máme k dispozici pouze **výběrový soubor** s hodnotami $(X_1, \ldots, X_n)$.
Abychom se vyhnuli diskusi skutečné velikosti základního statistického souboru s N jednotkami, budeme předpokládat, že vybíráme položky výběrového souboru jednu po druhé a každou vybranou jednotku poté do populace vracíme. Zároveň předpokládáme, že každá položka má stejnou pravděpodbnost výběru $1/N$. Hovoříme pak o **náhodném výběru**.

Pracujeme tedy s vektorem $(X_1, \ldots, X_n)$ nezávislých náhodných veličin a všechny tyto veličiny mají stejné rozdělení pravděpodobnosti. Zejména tedy budou sdílet distribuční funkci $F_X(x)$ a momenty

$$\mathrm{E}\, X_i = \mu, \quad \mathrm{var}\, X_i = \sigma^2.$$

Dalším naším krokem musí být odvození charakteristik výběrového průměru $\bar{X}$ a **výběrového rozptylu**

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

přičemž následující věta dává hned zdůvodnění, proč volíme koeficient $\frac{1}{n-1}$ místo $\frac{1}{n}$.

Theorem

Pro výběrový průměr $\bar{X}$ spočítaný z náhodného výběru rozsahu n z rozdělení s konečnou střední hodnotou μ a konečným rozptylem σ^2 platí

$$\mathrm{E}\,\bar{X} = \mu, \quad \mathrm{var}\,\bar{X} = \frac{1}{n}\sigma^2.$$

Pro výběrový rozptyl S^2 platí

$$\mathrm{E}\,S^2 = \sigma^2.$$

Naším úkolem je odhadovat charakteristiky, jako jsou průměr μ hodnot znaku $\bar{x}$ nebo jejich rozptyl σ^2 pro celou populaci pomocí obdobných charakteristik pro náš daleko menší výběr, které budeme značit pomocí velkých písmen, např. $\bar{X}$, S^2.

Zde vstupuje do hry pravděpodobnost – budeme chtít znát pravděpodobnost přiblížení hodnot pro náš výběr těm pro celou populaci.
Říkáme, že $\bar{X}$ je nestranným odhadem střední hodnoty znaku pro populaci, zatímco výběrový rozptyl je nestranným odhadem rozptylu.
V případě, že bychom realizovali výběr z populace bez vracení, bude výběrový průměr stále nestranným odhadem střední hodnoty, výběrový rozptyl ale již ne (vyskočí tam faktor $(N-1)/N$).

V praktických úlohách je třeba znát nejen číselné charakteristiky výběrového průměru a rozptylu, ale jejich úplné rozdělení pravděpodobnosti. To můžeme samozřejmě odvodit, pouze známe-li konkrétní rozdělení pravděpodobnosti X_i. Jako užitečnou ilustraci se podíváme na náhodný výběr z normálního rozdělení.
Výběrový průměr bude mít normální rozdělení a protože již známe jeho střední hodnotu a rozptyl, bude $\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n}\sigma^2)$.
O něco složitější je to s odvozením rozdělení pravděpodobnosti výběrového rozptylu. Uvažme vektor Z normovaných normálních veličin

$$Z_i = \frac{X_i - \mu}{\sigma}.$$

Theorem

Je-li $(X_1, \ldots, X_n)$ *náhodný výběr z rozdělení* $\mathrm{N}(\mu, \sigma^2)$, *pak jsou* $\bar{X}$ *a* S^2 *nezávislé veličiny a platí*

$$\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n}\sigma^2), \quad \frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}\,.$$

Okamžitým důsledkem je, že normalizovaný výběrový průměr

$$T = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

má studentovo t-rozdělení pravděpodobnosti s $n - 1$ stupni volnosti.