

# Základy biostatistiky

(MD710P09)

ak. rok 2008/2009

Karel Zvára

karel.zvara@mff.cuni.cz

<http://www.karlin.mff.cuni.cz/~zvara>

katedra pravděpodobnosti a matematické statistiky MFF UK

(naposledy upraveno 24. března 2009)



# kritické hodnoty normálního a Studentova $t$ -rozdělení

[Student distribution]

- ▶ **normální rozdělení**  $N(0, 1)$  [qnorm(1-alpha)]

$$Z \sim N(0, 1) : P(Z > z(\alpha)) = \alpha$$

ze symetrie platí  $P(|Z| > z(\alpha/2)) = \alpha$

- ▶ **Studentovo  $t$ -rozdělení**  $t_k$   
(podobné normálnímu, protože místo  $\sigma$  používá jeho odhad, má větší rozptyl)

$$T \sim t_k : P(|T| > t_k(\alpha)) = \alpha$$

- ▶ jsou to spíše kritické hodnoty  $|T|$  [qt(1-alpha/2,k)]

# kritické hodnoty normálního a Studentova $t$ -rozdělení

[Student distribution]

- ▶ **normální rozdělení**  $N(0, 1)$  [qnorm(1-alpha)]

$$Z \sim N(0, 1) : P(Z > z(\alpha)) = \alpha$$

ze symetrie platí  $P(|Z| > z(\alpha/2)) = \alpha$

- ▶ **Studentovo  $t$ -rozdělení**  $t_k$   
(podobné normálnímu, protože místo  $\sigma$  používá jeho odhad, má větší rozptyl)

$$T \sim t_k : P(|T| > t_k(\alpha)) = \alpha$$

- ▶ jsou to spíše kritické hodnoty  $|T|$  [qt(1-alpha/2,k)]

# kritické hodnoty normálního a Studentova $t$ -rozdělení

[Student distribution]

- ▶ **normální rozdělení**  $N(0, 1)$  [qnorm(1-alpha)]

$$Z \sim N(0, 1) : P(Z > z(\alpha)) = \alpha$$

ze symetrie platí  $P(|Z| > z(\alpha/2)) = \alpha$

- ▶ **Studentovo  $t$ -rozdělení**  $t_k$   
(podobné normálnímu, protože místo  $\sigma$  používá jeho odhad, má větší rozptyl)

$$T \sim t_k : P(|T| > t_k(\alpha)) = \alpha$$

- ▶ jsou to spíše kritické hodnoty  $|T|$  [qt(1-alpha/2,k)]

## některé kritické hodnoty

$\alpha$	0,50	0,25	0,10	<b>0,05</b>	0,01
$z(\alpha/2)$	0,674	1,150	1,645	<b>1,960</b>	2,576
$t_{100}(\alpha)$	0,677	1,157	1,660	<b>1,984</b>	2,626
$t_{20}(\alpha)$	0,687	1,185	1,725	<b>2,086</b>	2,845
$t_5(\alpha)$	0,727	1,301	2,015	<b>2,571</b>	4,032

- ▶  $T \sim t_k$  má jediný parametr  $k$  (počet stupňů volnosti)
- ▶ s rostoucím  $k$  se chování blíží normálnímu rozdělení
- ▶ pro  $Z \sim N(0, 1)$  je 95 % hodnot v intervalu  $(-1,960; 1,960)$
- ▶ pro  $T \sim t_5$  je 95 % hodnot v intervalu  $(-2,571; 2,571)$
- ▶ pro  $T \sim t_{20}$  je 95 % hodnot v intervalu  $(-2,086; 2,086)$
- ▶ pro  $T \sim t_{100}$  je 95 % hodnot v intervalu  $(-1,984; 1,984)$

## některé kritické hodnoty

$\alpha$	0,50	0,25	0,10	<b>0,05</b>	0,01
$z(\alpha/2)$	0,674	1,150	1,645	<b>1,960</b>	2,576
$t_{100}(\alpha)$	0,677	1,157	1,660	<b>1,984</b>	2,626
$t_{20}(\alpha)$	0,687	1,185	1,725	<b>2,086</b>	2,845
$t_5(\alpha)$	0,727	1,301	2,015	<b>2,571</b>	4,032

- ▶  $T \sim t_k$  má jediný parametr  $k$  (počet stupňů volnosti)
- ▶ s rostoucím  $k$  se chování blíží normálnímu rozdělení
- ▶ pro  $Z \sim N(0, 1)$  je 95 % hodnot v intervalu  $(-1,960; 1,960)$
- ▶ pro  $T \sim t_5$  je 95 % hodnot v intervalu  $(-2,571; 2,571)$
- ▶ pro  $T \sim t_{20}$  je 95 % hodnot v intervalu  $(-2,086; 2,086)$
- ▶ pro  $T \sim t_{100}$  je 95 % hodnot v intervalu  $(-1,984; 1,984)$

## některé kritické hodnoty

$\alpha$	0,50	0,25	0,10	<b>0,05</b>	0,01
$z(\alpha/2)$	0,674	1,150	1,645	<b>1,960</b>	2,576
$t_{100}(\alpha)$	0,677	1,157	1,660	<b>1,984</b>	2,626
$t_{20}(\alpha)$	0,687	1,185	1,725	<b>2,086</b>	2,845
$t_5(\alpha)$	0,727	1,301	2,015	<b>2,571</b>	4,032

- ▶  $T \sim t_k$  má jediný parametr  $k$  (počet stupňů volnosti)
- ▶ s rostoucím  $k$  se chování blíží normálnímu rozdělení
- ▶ pro  $Z \sim N(0, 1)$  je 95 % hodnot v intervalu  $(-1,960; 1,960)$
- ▶ pro  $T \sim t_5$  je 95 % hodnot v intervalu  $(-2,571; 2,571)$
- ▶ pro  $T \sim t_{20}$  je 95 % hodnot v intervalu  $(-2,086; 2,086)$
- ▶ pro  $T \sim t_{100}$  je 95 % hodnot v intervalu  $(-1,984; 1,984)$

## některé kritické hodnoty

$\alpha$	0,50	0,25	0,10	<b>0,05</b>	0,01
$z(\alpha/2)$	0,674	1,150	1,645	<b>1,960</b>	2,576
$t_{100}(\alpha)$	0,677	1,157	1,660	<b>1,984</b>	2,626
$t_{20}(\alpha)$	0,687	1,185	1,725	<b>2,086</b>	2,845
$t_5(\alpha)$	0,727	1,301	2,015	<b>2,571</b>	4,032

- ▶  $T \sim t_k$  má jediný parametr  $k$  (počet stupňů volnosti)
- ▶ s rostoucím  $k$  se chování blíží normálnímu rozdělení
- ▶ pro  $Z \sim N(0, 1)$  je 95 % hodnot v intervalu  $(-1,960; 1,960)$
- ▶ pro  $T \sim t_5$  je 95 % hodnot v intervalu  $(-2,571; 2,571)$
- ▶ pro  $T \sim t_{20}$  je 95 % hodnot v intervalu  $(-2,086; 2,086)$
- ▶ pro  $T \sim t_{100}$  je 95 % hodnot v intervalu  $(-1,984; 1,984)$



## některé kritické hodnoty

$\alpha$	0,50	0,25	0,10	<b>0,05</b>	0,01
$z(\alpha/2)$	0,674	1,150	1,645	<b>1,960</b>	2,576
$t_{100}(\alpha)$	0,677	1,157	1,660	<b>1,984</b>	2,626
$t_{20}(\alpha)$	0,687	1,185	1,725	<b>2,086</b>	2,845
$t_5(\alpha)$	0,727	1,301	2,015	<b>2,571</b>	4,032

- ▶  $T \sim t_k$  má jediný parametr  $k$  (počet stupňů volnosti)
- ▶ s rostoucím  $k$  se chování blíží normálnímu rozdělení
- ▶ pro  $Z \sim N(0, 1)$  je 95 % hodnot v intervalu  $(-1,960; 1,960)$
- ▶ pro  $T \sim t_5$  je 95 % hodnot v intervalu  $(-2,571; 2,571)$
- ▶ pro  $T \sim t_{20}$  je 95 % hodnot v intervalu  $(-2,086; 2,086)$
- ▶ pro  $T \sim t_{100}$  je 95 % hodnot v intervalu  $(-1,984; 1,984)$

## některé kritické hodnoty

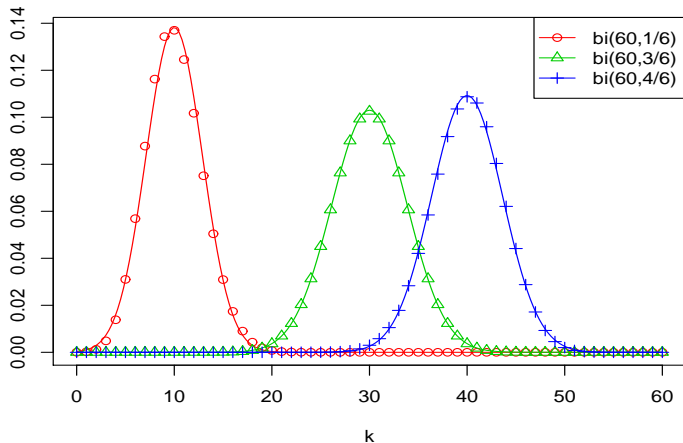
$\alpha$	0,50	0,25	0,10	<b>0,05</b>	0,01
$z(\alpha/2)$	0,674	1,150	1,645	<b>1,960</b>	2,576
$t_{100}(\alpha)$	0,677	1,157	1,660	<b>1,984</b>	2,626
$t_{20}(\alpha)$	0,687	1,185	1,725	<b>2,086</b>	2,845
$t_5(\alpha)$	0,727	1,301	2,015	<b>2,571</b>	4,032

- ▶  $T \sim t_k$  má jediný parametr  $k$  (počet stupňů volnosti)
- ▶ s rostoucím  $k$  se chování blíží normálnímu rozdělení
- ▶ pro  $Z \sim N(0, 1)$  je 95 % hodnot v intervalu  $(-1,960; 1,960)$
- ▶ pro  $T \sim t_5$  je 95 % hodnot v intervalu  $(-2,571; 2,571)$
- ▶ pro  $T \sim t_{20}$  je 95 % hodnot v intervalu  $(-2,086; 2,086)$
- ▶ pro  $T \sim t_{100}$  je 95 % hodnot v intervalu  $(-1,984; 1,984)$

# aproximace binomického rozdělení normálním

se stejnou střední hodnotou a stejným rozptylem

rozdělení  $bi(n, \pi)$  lze aproximovat pomocí  $N(n\pi, n\pi(1 - \pi))$



## další rozdělení související s normálním

[F-distribution, chi-square distribution]

- ▶  $V$  má rozdělení (musí být  $P(V > 0) = 1$  !!)

**logaritmicko-normální**, platí-li  $\ln V \sim N(\mu, \sigma^2)$

- ▶ Fisherovo  $F$ -rozdělení  $F_{k,m}$  [qf(1-alpha,k,m)]

$$F \sim F_{k,m} : P(F > F_{k,m}(\alpha)) = \alpha$$

- ▶ rozdělení chí-kvadrát  $\chi_k^2$  [qchisq(1-alpha,k)]

$$X^2 \sim \chi_k^2 : P(X^2 > \chi_k^2(\alpha)) = \alpha$$

- ▶ speciálně platí:

- ▶  $\chi_1^2(0,05) = 3,841 = 1,960^2$

- ▶  $\chi_1^2(\alpha) = z(\alpha/2)^2$

- ▶  $F_{1,m}(\alpha) = (t_m(\alpha))^2$

## další rozdělení související s normálním

[F-distribution, chi-square distribution]

- ▶  $V$  má rozdělení (musí být  $P(V > 0) = 1$  !!)

**logaritmicko-normální**, platí-li  $\ln V \sim N(\mu, \sigma^2)$

- ▶ **Fisherovo  $F$ -rozdělení**  $F_{k,m}$  [qf(1-alpha,k,m)]

$$F \sim F_{k,m} : P(F > F_{k,m}(\alpha)) = \alpha$$

- ▶ **rozdělení chí-kvadrát**  $\chi_k^2$  [qchisq(1-alpha,k)]

$$X^2 \sim \chi_k^2 : P(X^2 > \chi_k^2(\alpha)) = \alpha$$

- ▶ speciálně platí:

- ▶  $\chi_1^2(0,05) = 3,841 = 1,960^2$
- ▶  $\chi_1^2(\alpha) = z(\alpha/2)^2$
- ▶  $F_{1,m}(\alpha) = (t_m(\alpha))^2$

## další rozdělení související s normálním

[F-distribution, chi-square distribution]

- ▶  $V$  má rozdělení (musí být  $P(V > 0) = 1$  !!)

**logaritmicko-normální**, platí-li  $\ln V \sim N(\mu, \sigma^2)$

- ▶ **Fisherovo  $F$ -rozdělení**  $F_{k,m}$  [qf(1-alpha,k,m)]

$$F \sim F_{k,m} : P(F > F_{k,m}(\alpha)) = \alpha$$

- ▶ **rozdělení chí-kvadrát**  $\chi_k^2$  [qchisq(1-alpha,k)]

$$X^2 \sim \chi_k^2 : P(X^2 > \chi_k^2(\alpha)) = \alpha$$

- ▶ speciálně platí:

- ▶  $\chi_1^2(0,05) = 3,841 = 1,960^2$
- ▶  $\chi_1^2(\alpha) = z(\alpha/2)^2$
- ▶  $F_{1,m}(\alpha) = (t_m(\alpha))^2$

## další rozdělení související s normálním

[F-distribution, chi-square distribution]

- ▶  $V$  má rozdělení (musí být  $P(V > 0) = 1$  !!)

**logaritmicko-normální**, platí-li  $\ln V \sim N(\mu, \sigma^2)$

- ▶ **Fisherovo  $F$ -rozdělení**  $F_{k,m}$  [qf(1-alpha,k,m)]

$$F \sim F_{k,m} : P(F > F_{k,m}(\alpha)) = \alpha$$

- ▶ **rozdělení chí-kvadrát**  $\chi_k^2$  [qchisq(1-alpha,k)]

$$X^2 \sim \chi_k^2 : P(X^2 > \chi_k^2(\alpha)) = \alpha$$

- ▶ speciálně platí:

- ▶  $\chi_1^2(0,05) = 3,841 = 1,960^2$
- ▶  $\chi_1^2(\alpha) = z(\alpha/2)^2$
- ▶  $F_{1,m}(\alpha) = (t_m(\alpha))^2$

## další rozdělení související s normálním

[F-distribution, chi-square distribution]

- ▶  $V$  má rozdělení (musí být  $P(V > 0) = 1$  !!)

**logaritmicko-normální**, platí-li  $\ln V \sim N(\mu, \sigma^2)$

- ▶ **Fisherovo  $F$ -rozdělení**  $F_{k,m}$  [qf(1-alpha,k,m)]

$$F \sim F_{k,m} : P(F > F_{k,m}(\alpha)) = \alpha$$

- ▶ **rozdělení chí-kvadrát**  $\chi_k^2$  [qchisq(1-alpha,k)]

$$X^2 \sim \chi_k^2 : P(X^2 > \chi_k^2(\alpha)) = \alpha$$

- ▶ speciálně platí:

- ▶  $\chi_1^2(0,05) = 3,841 = 1,960^2$
- ▶  $\chi_1^2(\alpha) = z(\alpha/2)^2$
- ▶  $F_{1,m}(\alpha) = (t_m(\alpha))^2$



## další rozdělení související s normálním

[F-distribution, chi-square distribution]

- ▶  $V$  má rozdělení (musí být  $P(V > 0) = 1$  !!)

**logaritmicko-normální**, platí-li  $\ln V \sim N(\mu, \sigma^2)$

- ▶ **Fisherovo  $F$ -rozdělení**  $F_{k,m}$  [qf(1-alpha,k,m)]

$$F \sim F_{k,m} : P(F > F_{k,m}(\alpha)) = \alpha$$

- ▶ **rozdělení chí-kvadrát**  $\chi_k^2$  [qchisq(1-alpha,k)]

$$X^2 \sim \chi_k^2 : P(X^2 > \chi_k^2(\alpha)) = \alpha$$

- ▶ speciálně platí:

- ▶  $\chi_1^2(0,05) = 3,841 = 1,960^2$
- ▶  $\chi_1^2(\alpha) = z(\alpha/2)^2$
- ▶  $F_{1,m}(\alpha) = (t_m(\alpha))^2$

## další rozdělení související s normálním

[F-distribution, chi-square distribution]

- ▶  $V$  má rozdělení (musí být  $P(V > 0) = 1$  !!)

**logaritmicko-normální**, platí-li  $\ln V \sim N(\mu, \sigma^2)$

- ▶ **Fisherovo  $F$ -rozdělení**  $F_{k,m}$  [qf(1-alpha,k,m)]

$$F \sim F_{k,m} : P(F > F_{k,m}(\alpha)) = \alpha$$

- ▶ **rozdělení chí-kvadrát**  $\chi_k^2$  [qchisq(1-alpha,k)]

$$X^2 \sim \chi_k^2 : P(X^2 > \chi_k^2(\alpha)) = \alpha$$

- ▶ speciálně platí:

- ▶  $\chi_1^2(0,05) = 3,841 = 1,960^2$
- ▶  $\chi_1^2(\alpha) = z(\alpha/2)^2$
- ▶  $F_{1,m}(\alpha) = (t_m(\alpha))^2$

# populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- ▶ **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku)  $\Rightarrow$  rozdělení náhodné veličiny
- ▶ **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- ▶ **reprezentativní výběr** obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- ▶ **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- ▶ **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- ▶ **statistika** funkce náhodného výběru (pozorování)
- ▶ **odhad** statistika použitá k odhadu parametru

# populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- ▶ **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku)  $\Rightarrow$  rozdělení náhodné veličiny
- ▶ **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- ▶ **reprezentativní výběr** obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- ▶ **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- ▶ **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- ▶ **statistika** funkce náhodného výběru (pozorování)
- ▶ **odhad** statistika použitá k odhadu parametru

# populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- ▶ **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku)  $\Rightarrow$  rozdělení náhodné veličiny
- ▶ **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- ▶ **reprezentativní výběr** obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- ▶ **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- ▶ **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- ▶ **statistika** funkce náhodného výběru (pozorování)
- ▶ **odhad** statistika použitá k odhadu parametru

# populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- ▶ **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku)  $\Rightarrow$  rozdělení náhodné veličiny
- ▶ **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- ▶ **reprezentativní výběr** obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- ▶ **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- ▶ **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- ▶ **statistika** funkce náhodného výběru (pozorování)
- ▶ **odhad** statistika použitá k odhadu parametru

# populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- ▶ **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku)  $\Rightarrow$  rozdělení náhodné veličiny
- ▶ **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- ▶ **reprezentativní výběr** obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- ▶ **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- ▶ **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- ▶ **statistika** funkce náhodného výběru (pozorování)
- ▶ **odhad** statistika použitá k odhadu parametru

# populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- ▶ **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku)  $\Rightarrow$  rozdělení náhodné veličiny
- ▶ **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- ▶ **reprezentativní výběr** obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- ▶ **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- ▶ **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- ▶ **statistika** funkce náhodného výběru (pozorování)
- ▶ **odhad** statistika použitá k odhadu parametru



## populace a výběr

[population, (random) sample, representative, parameter, statistics, estimator]

- ▶ **populace (základní soubor)** soubor jednotek, o jejichž hromadných vlastnostech chceme vypovídat (všechny možné výsledky pokusu, všichni hoši zvoleného věku, všichni čolci v rybníčku)  $\Rightarrow$  rozdělení náhodné veličiny
- ▶ **výběr** náhodně vybraná vyšetřovaná část populace (vzorek)
- ▶ **reprezentativní výběr** obráží poměry v populaci (nutná vlastnost výběru, aby mohl vypovídat o populaci)
- ▶ **náhodný výběr** nezávislé náhodné veličiny se stejným rozdělením (model pro měření na výběru)
- ▶ **parametr** neznámé číslo popisující nějakou **vlastnost** populace, charakteristika rozdělení náhodné veličiny
- ▶ **statistika** funkce náhodného výběru (pozorování)
- ▶ **odhad** statistika použitá k odhadu parametru

# průměr z náhodného výběru

- ▶  $X_1, \dots, X_n$  **nezávislé**, stejné rozdělení  
 $\mu_{X_i} = E X_i = \mu$  (stejná střední hodnota)  
 $\sigma_{X_i}^2 = \text{var } X_i = \sigma^2$  (stejný rozptyl)

- ▶  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

- ▶  $\mu_{\bar{X}} = E \bar{X} = \mu$

- ▶ výběrový průměr  $\bar{X}$  je opět náhodná veličina
- ▶ je **nestranným** odhadem [unbiased estimator] parametru  $\mu$
- ▶ **nestranným** odhadem populačního průměru (střední hodnoty)
- ▶ když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

**náhodný výběr**  
populační průměr  
populační rozptyl

výběrový průměr

## průměr z náhodného výběru

- ▶  $X_1, \dots, X_n$  **nezávislé**, stejné rozdělení  
 $\mu_{X_i} = E X_i = \mu$  (stejná střední hodnota)  
 $\sigma_{X_i}^2 = \text{var } X_i = \sigma^2$  (stejný rozptyl)

- ▶  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

- ▶  $\mu_{\bar{X}} = E \bar{X} = \mu$

**náhodný výběr**  
populační průměr  
populační rozptyl

výběrový průměr

- ▶ výběrový průměr  $\bar{X}$  je opět náhodná veličina
- ▶ je **nestranným odhadem** [unbiased estimator] parametru  $\mu$
- ▶ **nestranným odhadem** populačního průměru (střední hodnoty)
- ▶ když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

## průměr z náhodného výběru

- ▶  $X_1, \dots, X_n$  **nezávislé**, stejné rozdělení  
 $\mu_{X_i} = E X_i = \mu$  (stejná střední hodnota)  
 $\sigma_{X_i}^2 = \text{var } X_i = \sigma^2$  (stejný rozptyl)

- ▶  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

- ▶  $\mu_{\bar{X}} = E \bar{X} = \mu$

- ▶ výběrový průměr  $\bar{X}$  je opět náhodná veličina
- ▶ je **nestranným** odhadem [unbiased estimator] parametru  $\mu$
- ▶ nestranným odhadem populačního průměru (střední hodnoty)
- ▶ když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

**náhodný výběr**  
populační průměr  
populační rozptyl

výběrový průměr

# průměr z náhodného výběru

- ▶  $X_1, \dots, X_n$  **nezávislé**, stejné rozdělení  
 $\mu_{X_i} = E X_i = \mu$  (stejná střední hodnota)  
 $\sigma_{X_i}^2 = \text{var } X_i = \sigma^2$  (stejný rozptyl)

**náhodný výběr**  
populační průměr  
populační rozptyl

- ▶  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

výběrový průměr

- ▶  $\mu_{\bar{X}} = E \bar{X} = \mu$

- ▶ výběrový průměr  $\bar{X}$  je opět náhodná veličina
- ▶ je **nestranným** odhadem [unbiased estimator] parametru  $\mu$
- ▶ nestranným odhadem populačního průměru (střední hodnoty)
- ▶ když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

# průměr z náhodného výběru

- ▶  $X_1, \dots, X_n$  **nezávislé**, stejné rozdělení  
 $\mu_{X_i} = E X_i = \mu$  (stejná střední hodnota)  
 $\sigma_{X_i}^2 = \text{var } X_i = \sigma^2$  (stejný rozptyl)

**náhodný výběr**  
populační průměr  
populační rozptyl

- ▶  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

výběrový průměr

- ▶  $\mu_{\bar{X}} = E \bar{X} = \mu$

- ▶ výběrový průměr  $\bar{X}$  je opět náhodná veličina
- ▶ je **nestranným** odhadem [**unbiased estimator**] parametru  $\mu$
- ▶ nestranným odhadem populačního průměru (střední hodnoty)
- ▶ když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

# průměr z náhodného výběru

- ▶  $X_1, \dots, X_n$  **nezávislé**, stejné rozdělení  
 $\mu_{X_i} = E X_i = \mu$  (stejná střední hodnota)  
 $\sigma_{X_i}^2 = \text{var } X_i = \sigma^2$  (stejný rozptyl)

**náhodný výběr**  
populační průměr  
populační rozptyl

- ▶  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

výběrový průměr

- ▶  $\mu_{\bar{X}} = E \bar{X} = \mu$

- ▶ výběrový průměr  $\bar{X}$  je opět náhodná veličina
- ▶ je **nestranným** odhadem [unbiased estimator] parametru  $\mu$
- ▶ nestranným odhadem populačního průměru (střední hodnoty)
- ▶ když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru

## průměr z náhodného výběru

- ▶  $X_1, \dots, X_n$  **nezávislé**, stejné rozdělení  
 $\mu_{X_i} = E X_i = \mu$  (stejná střední hodnota)  
 $\sigma_{X_i}^2 = \text{var } X_i = \sigma^2$  (stejný rozptyl)

**náhodný výběr**  
populační průměr  
populační rozptyl

- ▶  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

výběrový průměr

- ▶  $\mu_{\bar{X}} = E \bar{X} = \mu$

- ▶ výběrový průměr  $\bar{X}$  je opět náhodná veličina
- ▶ je **nestranným** odhadem [unbiased estimator] parametru  $\mu$
- ▶ nestranným odhadem populačního průměru (střední hodnoty)
- ▶ když pořizujeme výběry opakovaně, průměry kolísají kolem skutečné hodnoty populačního průměru



## rozptyl průměru z náhodného výběru

$$\sigma_{\bar{X}}^2 = \text{var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{var} X_i = \frac{\sigma^2}{n} = \left( \frac{\sigma}{\sqrt{n}} \right)^2 = (\text{S.E.}(\bar{X}))^2$$

- ▶  $\text{S.E.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$  – **střední chyba průměru**

[standard error of mean]

- ▶ variabilita průměrů (měřená rozptylem) z výběrů rozsahu  $n$  je  $n$ -krát menší, než variabilita jednotlivých pozorování  $\sigma^2$
- ▶ střední chyba průměru je  $\sqrt{n}$ -krát menší než  $\sigma$
- ▶ čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- ▶ speciálně pro normální rozdělení  $X_i \sim N(\mu, \sigma^2)$  nezávislé:

$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$$

(všimněte si závislosti na  $n$ )

## rozptyl průměru z náhodného výběru

$$\sigma_{\bar{X}}^2 = \text{var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{var} X_i = \frac{\sigma^2}{n} = \left( \frac{\sigma}{\sqrt{n}} \right)^2 = (\text{S.E.}(\bar{X}))^2$$

- ▶  $\text{S.E.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$  – **střední chyba průměru** [standard error of mean]
- ▶ variabilita průměrů (měřená rozptylem) z výběrů rozsahu  $n$  je  $n$ -krát menší, než variabilita jednotlivých pozorování  $\sigma^2$
- ▶ střední chyba průměru je  $\sqrt{n}$ -krát menší než  $\sigma$
- ▶ čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- ▶ speciálně pro normální rozdělení  $X_i \sim N(\mu, \sigma^2)$  nezávislé:

$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$$

(všimněte si závislosti na  $n$ )

## rozptyl průměru z náhodného výběru

$$\sigma_{\bar{X}}^2 = \text{var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{var} X_i = \frac{\sigma^2}{n} = \left( \frac{\sigma}{\sqrt{n}} \right)^2 = (\text{S.E.}(\bar{X}))^2$$

- ▶  $\text{S.E.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$  – **střední chyba průměru** [standard error of mean]
- ▶ variabilita průměrů (měřená rozptylem) z výběrů rozsahu  $n$  je  $n$ -krát menší, než variabilita jednotlivých pozorování  $\sigma^2$
- ▶ střední chyba průměru je  $\sqrt{n}$ -krát menší než  $\sigma$
- ▶ čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- ▶ speciálně pro normální rozdělení  $X_i \sim N(\mu, \sigma^2)$  nezávislé:

$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$$

(všimněte si závislosti na  $n$ )

## rozptyl průměru z náhodného výběru

$$\sigma_{\bar{X}}^2 = \text{var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{var} X_i = \frac{\sigma^2}{n} = \left( \frac{\sigma}{\sqrt{n}} \right)^2 = (\text{S.E.}(\bar{X}))^2$$

- ▶  $\text{S.E.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$  – **střední chyba průměru** [standard error of mean]
- ▶ variabilita průměrů (měřená rozptylem) z výběrů rozsahu  $n$  je  $n$ -krát menší, než variabilita jednotlivých pozorování  $\sigma^2$
- ▶ střední chyba průměru je  $\sqrt{n}$ -krát menší než  $\sigma$
- ▶ čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- ▶ speciálně pro normální rozdělení  $X_i \sim N(\mu, \sigma^2)$  nezávislé:

$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$$

(všimněte si závislosti na  $n$ )

## rozptyl průměru z náhodného výběru

$$\sigma_{\bar{X}}^2 = \text{var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{var} X_i = \frac{\sigma^2}{n} = \left( \frac{\sigma}{\sqrt{n}} \right)^2 = (\text{S.E.}(\bar{X}))^2$$

- ▶  $\text{S.E.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$  – **střední chyba průměru** [standard error of mean]
- ▶ variabilita průměrů (měřená rozptylem) z výběrů rozsahu  $n$  je  $n$ -krát menší, než variabilita jednotlivých pozorování  $\sigma^2$
- ▶ střední chyba průměru je  $\sqrt{n}$ -krát menší než  $\sigma$
- ▶ čím jsou rozsahy výběru větší, tím méně výběrové průměry kolísají (kolem populačního průměru)
- ▶ speciálně pro normální rozdělení  $X_i \sim N(\mu, \sigma^2)$  nezávislé:

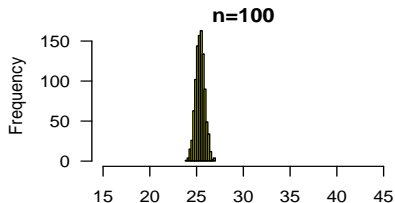
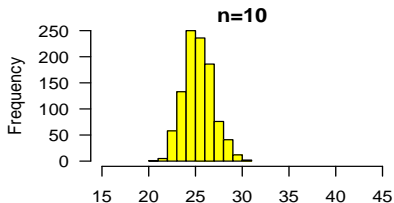
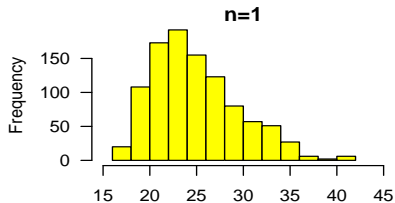
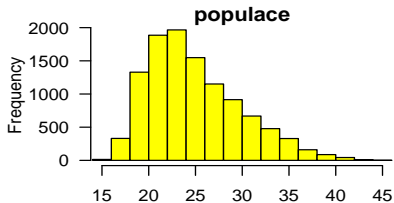
$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$$

(všimněte si závislosti na  $n$ )

## příklad: věk matek (umělá situace)

populace - 10 916 matek, opakované výběry rozsahu  $n = 1, 10, 100$

je patrná variabilita klesající s rostoucím  $n$



## příklad: věk matek – shrnutí

- ▶ velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- ▶ na rozdíl od běžné praxe **známe** populační průměr  $\mu$
- ▶ náhodně vybráno 1000 matek (vlastně průměry výběrů rozsahu  $n = 1$ ), nakreslen histogram
- ▶ 1000 krát náhodně vybráno vždy  $n = 10$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ 1000 krát náhodně vybráno vždy  $n = 100$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- ▶ skutečné rozptyly (odhady z 1000 realizací): 23,5; 2,20; 0,21

## příklad: věk matek – shrnutí

- ▶ velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- ▶ na rozdíl od běžné praxe **známe** populační průměr  $\mu$
- ▶ náhodně vybráno 1000 matek (vlastně průměry výběrů rozsahu  $n = 1$ ), nakreslen histogram
- ▶ 1000 krát náhodně vybráno vždy  $n = 10$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ 1000 krát náhodně vybráno vždy  $n = 100$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- ▶ skutečné rozptyly (odhady z 1000 realizací): 23,5; 2,20; 0,21



## příklad: věk matek – shrnutí

- ▶ velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- ▶ na rozdíl od běžné praxe **známe** populační průměr  $\mu$
- ▶ náhodně vybráno 1000 matek (vlastně průměry výběrů rozsahu  $n = 1$ ), nakreslen histogram
- ▶ 1000 krát náhodně vybráno vždy  $n = 10$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ 1000 krát náhodně vybráno vždy  $n = 100$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- ▶ skutečné rozptyly (odhady z 1000 realizací): 23,5; 2,20; 0,21

## příklad: věk matek – shrnutí

- ▶ velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- ▶ na rozdíl od běžné praxe **známe** populační průměr  $\mu$
- ▶ náhodně vybráno 1000 matek (vlastně průměry výběrů rozsahu  $n = 1$ ), nakreslen histogram
- ▶ 1000 krát náhodně vybráno vždy  $n = 10$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ 1000 krát náhodně vybráno vždy  $n = 100$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- ▶ skutečné rozptyly (odhady z 1000 realizací): 23,5; 2,20; 0,21

## příklad: věk matek – shrnutí

- ▶ velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- ▶ na rozdíl od běžné praxe **známe** populační průměr  $\mu$
- ▶ náhodně vybráno 1000 matek (vlastně průměry výběrů rozsahu  $n = 1$ ), nakreslen histogram
- ▶ 1000 krát náhodně vybráno vždy  $n = 10$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ 1000 krát náhodně vybráno vždy  $n = 100$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- ▶ skutečné rozptyly (odhady z 1000 realizací): 23,5; 2,20; 0,21

## příklad: věk matek – shrnutí

- ▶ velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- ▶ na rozdíl od běžné praxe **známe** populační průměr  $\mu$
- ▶ náhodně vybráno 1000 matek (vlastně průměry výběrů rozsahu  $n = 1$ ), nakreslen histogram
- ▶ 1000 krát náhodně vybráno vždy  $n = 10$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ 1000 krát náhodně vybráno vždy  $n = 100$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- ▶ skutečné rozptyly (odhady z 1000 realizací): 23,5; 2,20; 0,21

## příklad: věk matek – shrnutí

- ▶ velká populace dětí (a tedy jejich matek, téměř 11 tisíc)
- ▶ na rozdíl od běžné praxe **známe** populační průměr  $\mu$
- ▶ náhodně vybráno 1000 matek (vlastně průměry výběrů rozsahu  $n = 1$ ), nakreslen histogram
- ▶ 1000 krát náhodně vybráno vždy  $n = 10$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ 1000 krát náhodně vybráno vždy  $n = 100$  matek, vždy spočítán průměr, nakreslen histogram průměrů
- ▶ podle teorie by každý další rozptyl ze 100 průměrů měl být desetkrát menší
- ▶ skutečné rozptyly (odhady z 1000 realizací): 23,5; 2,20; 0,21

# centrální limitní věta (CLV, CLT)

[Central Limit Theorem]

- ▶ Necht'  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$  (nemusí pocházet z normálního rozdělení).  
Potom **pro velké**  $n$  má průměr  $\bar{X}$  přibližně rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , součet  $X_1 + \dots + X_n$  pak rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶ prakticky: **průměr** má pro dost velká  $n$  **normální rozdělení** s rozptylem  $n$ -krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení jednotlivých pozorování
- ▶ CLT je často důvodem předpokladu o normálním rozdělení, výsledná hodnota je ovlivněna součtem velkého počtu nahodilých malých vlivů
- ▶ příklad: průměrný věk matek z velkých výběrů má už (téměř) normální rozdělení

# centrální limitní věta (CLV, CLT)

[Central Limit Theorem]

- ▶ Necht'  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$  (nemusí pocházet z normálního rozdělení).  
Potom **pro velké**  $n$  má průměr  $\bar{X}$  přibližně rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , součet  $X_1 + \dots + X_n$  pak rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶ prakticky: **průměr** má pro dost velká  $n$  **normální rozdělení** s rozptylem  $n$ -krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení jednotlivých pozorování
- ▶ CLT je často důvodem předpokladu o normálním rozdělení, výsledná hodnota je ovlivněna součtem velkého počtu nahodilých malých vlivů
- ▶ příklad: průměrný věk matek z velkých výběrů má už (téměř) normální rozdělení

# centrální limitní věta (CLV, CLT)

[Central Limit Theorem]

- ▶ Necht'  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$  (nemusí pocházet z normálního rozdělení). Potom **pro velké**  $n$  má průměr  $\bar{X}$  přibližně rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , součet  $X_1 + \dots + X_n$  pak rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶ prakticky: **průměr** má pro dost velká  $n$  **normální rozdělení** s rozptylem  $n$ -krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení jednotlivých pozorování
- ▶ CLT je často důvodem předpokladu o normálním rozdělení, výsledná hodnota je ovlivněna součtem velkého počtu nahodilých malých vlivů
- ▶ příklad: průměrný věk matek z velkých výběrů má už (téměř) normální rozdělení



# centrální limitní věta (CLV, CLT)

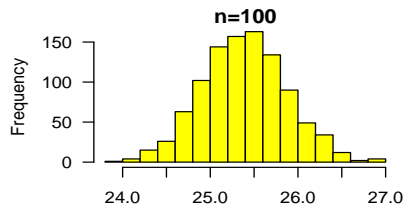
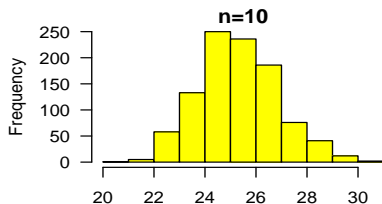
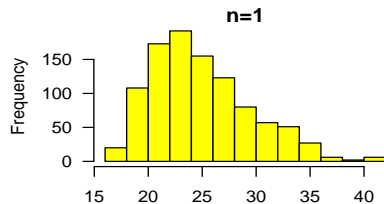
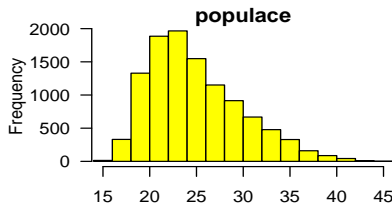
[Central Limit Theorem]

- ▶ Necht'  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$  (nemusí pocházet z normálního rozdělení). Potom **pro velké**  $n$  má průměr  $\bar{X}$  přibližně rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , součet  $X_1 + \dots + X_n$  pak rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶ prakticky: **průměr** má pro dost velká  $n$  **normální rozdělení** s rozptylem  $n$ -krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení jednotlivých pozorování
- ▶ CLT je často důvodem předpokladu o normálním rozdělení, výsledná hodnota je ovlivněna součtem velikého počtu nahodilých malých vlivů
- ▶ příklad: průměrný věk matek z velkých výběrů má už (téměř) normální rozdělení

# příklad: věk matek (nestejná měřítka!)

populace - 10 916 matek, opakované výběry rozsahu  $n = 1, 10, 100$

je patrné, že s rostoucím  $n$  se histogram blíží histogramu norm. rozdělení



## příklad: věk matek

průměrný věk matek v opakovaných výběrech

rozsah výběru $n$	průměr průměrů	směr. odch. průměrů	šikmost průměrů	špičatost průměrů
1	24,74	4,848	0,682	-0,040
10	25,14	1,482	0,743	-0,199
100	25,40	0,455	0,087	-0,076
populace	$\mu = 25,41$	$\sigma = 4,932$	$\gamma_1 = 0,771$	$\gamma_2 = 0,189$

# interval spolehlivosti pro $\mu$ (výběr z $N(\mu, \sigma^2)$ )

[confidence interval]

- ▶ víme, že  $\bar{X} \sim N(\mu, \sigma^2/n)$ , tedy  $Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$

$$P(|Z| < 1,96) = P\left(\frac{|\bar{X} - \mu|}{\sigma} \sqrt{n} < 1,96\right) = 0,95$$

- ▶ což je totéž, jako ( $\mu$  se od  $\bar{X}$  liší nejvýše ...)

$$P\left(|\bar{X} - \mu| < 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

- ▶ tedy (všimněte si zkracování intervalu s rostoucím  $n$ )

$$P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

- ▶ dostali jsme 95% **interval spolehlivosti pro parametr  $\mu$**

## interval spolehlivosti pro $\mu$ (výběr z $N(\mu, \sigma^2)$ )

[confidence interval]

- ▶ víme, že  $\bar{X} \sim N(\mu, \sigma^2/n)$ , tedy  $Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$

$$P(|Z| < 1,96) = P\left(\frac{|\bar{X} - \mu|}{\sigma} \sqrt{n} < 1,96\right) = 0,95$$

- ▶ což je totéž, jako ( $\mu$  se od  $\bar{X}$  liší nejvýše ...)

$$P\left(|\bar{X} - \mu| < 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

- ▶ tedy (všimněte si zkracování intervalu s rostoucím  $n$ )

$$P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

- ▶ dostali jsme 95% interval spolehlivosti pro parametr  $\mu$

## interval spolehlivosti pro $\mu$ (výběr z $N(\mu, \sigma^2)$ )

[confidence interval]

- ▶ víme, že  $\bar{X} \sim N(\mu, \sigma^2/n)$ , tedy  $Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$

$$P(|Z| < 1,96) = P\left(\frac{|\bar{X} - \mu|}{\sigma} \sqrt{n} < 1,96\right) = 0,95$$

- ▶ což je totéž, jako ( $\mu$  se od  $\bar{X}$  liší nejvýše ...)

$$P\left(|\bar{X} - \mu| < 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

- ▶ tedy (všimněte si zkracování intervalu s rostoucím  $n$ )

$$P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

- ▶ dostali jsme 95% interval spolehlivosti pro parametr  $\mu$

# interval spolehlivosti pro $\mu$ (výběr z $N(\mu, \sigma^2)$ )

[confidence interval]

- ▶ víme, že  $\bar{X} \sim N(\mu, \sigma^2/n)$ , tedy  $Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$

$$P(|Z| < 1,96) = P\left(\frac{|\bar{X} - \mu|}{\sigma} \sqrt{n} < 1,96\right) = 0,95$$

- ▶ což je totéž, jako ( $\mu$  se od  $\bar{X}$  liší nejvýše ...)

$$P\left(|\bar{X} - \mu| < 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

- ▶ tedy (všimněte si zkracování intervalu s rostoucím  $n$ )

$$P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

- ▶ dostali jsme 95% **interval spolehlivosti pro parametr  $\mu$**

## interpretace intervalu spolehlivosti

- ▶ je to **intervalový** odhad hodnoty  $\mu$
- ▶  $\bar{X}$  je **bodový** odhad
- ▶ **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé  $\mu$  (odhadovaný parametr)**
- ▶ kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu  $\mu$ , ve zbylých asi 5 % zůstane skutečné  $\mu$  mimo interval spolehlivosti
- ▶ pro obecné  $\alpha$  (spolehlivost  $1 - \alpha$ ):

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right) = 1 - \alpha$$

- ▶ **POZOR** na nesprávné interpretace, vypovídá o neznámé **konstantě  $\mu$** , nikoliv o **náhodných veličinách  $X$  nebo  $\bar{X}$**



## interpretace intervalu spolehlivosti

- ▶ je to **intervalový** odhad hodnoty  $\mu$
- ▶  $\bar{X}$  je **bodový** odhad
- ▶ **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé  $\mu$  (odhadovaný parametr)**
- ▶ kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu  $\mu$ , ve zbylých asi 5 % zůstane skutečné  $\mu$  mimo interval spolehlivosti
- ▶ pro obecné  $\alpha$  (spolehlivost  $1 - \alpha$ ):

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right) = 1 - \alpha$$

- ▶ **POZOR** na nesprávné interpretace, vypovídá o neznámé **konstantě  $\mu$** , nikoliv o **náhodných veličinách  $X$  nebo  $\bar{X}$**

## interpretace intervalu spolehlivosti

- ▶ je to **intervalový** odhad hodnoty  $\mu$
- ▶  $\bar{X}$  je **bodový** odhad
- ▶ **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé**  $\mu$  (**odhadovaný parametr**)
- ▶ kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu  $\mu$ , ve zbylých asi 5 % zůstane skutečné  $\mu$  mimo interval spolehlivosti
- ▶ pro obecné  $\alpha$  (spolehlivost  $1 - \alpha$ ):

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right) = 1 - \alpha$$

- ▶ POZOR na nesprávné interpretace, vypovídá o neznámé **konstantě**  $\mu$ , nikoliv o **náhodných veličinách**  $X$  nebo  $\bar{X}$

## interpretace intervalu spolehlivosti

- ▶ je to **intervalový** odhad hodnoty  $\mu$
- ▶  $\bar{X}$  je **bodový** odhad
- ▶ **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé**  $\mu$  (**odhadovaný parametr**)
- ▶ kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu  $\mu$ , ve zbylých asi 5 % zůstane skutečné  $\mu$  mimo interval spolehlivosti
- ▶ pro obecné  $\alpha$  (spolehlivost  $1 - \alpha$ ):

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2)\right) = 1 - \alpha$$

- ▶ POZOR na nesprávné interpretace, vypovídá o neznámé **konstantě**  $\mu$ , nikoliv o **náhodných veličinách**  $X$  nebo  $\bar{X}$

## interpretace intervalu spolehlivosti

- ▶ je to **intervalový** odhad hodnoty  $\mu$
- ▶  $\bar{X}$  je **bodový** odhad
- ▶ **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé**  $\mu$  (**odhadovaný parametr**)
- ▶ kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu  $\mu$ , ve zbylých asi 5 % zůstane skutečné  $\mu$  mimo interval spolehlivosti
- ▶ pro obecné  $\alpha$  (spolehlivost  $1 - \alpha$ ):

$$P \left( \bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) \right) = 1 - \alpha$$

- ▶ POZOR na nesprávné interpretace, vypovídá o neznámé konstantě  $\mu$ , nikoliv o náhodných veličinách  $X$  nebo  $\bar{X}$

## interpretace intervalu spolehlivosti

- ▶ je to **intervalový** odhad hodnoty  $\mu$
- ▶  $\bar{X}$  je **bodový** odhad
- ▶ **základní vlastnost**: 95% interval spolehlivosti **překryje** s pravděpodobností 95 % **neznámé**  $\mu$  (**odhadovaný parametr**)
- ▶ kdybychom postup prováděli opakovaně, pak asi v 95 % případů interval překryje skutečnou hodnotu  $\mu$ , ve zbylých asi 5 % zůstane skutečné  $\mu$  mimo interval spolehlivosti
- ▶ pro obecné  $\alpha$  (spolehlivost  $1 - \alpha$ ):

$$P \left( \bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z(\alpha/2) \right) = 1 - \alpha$$

- ▶ **POZOR** na nesprávné interpretace, vypovídá o neznámé **konstantě**  $\mu$ , nikoliv o **náhodných veličinách**  $X$  nebo  $\bar{X}$

## interval spolehlivosti při neznámém $\sigma$

- ▶ pro  $X_i$  s normálním rozdělením je třeba použít kritické hodnoty Studentova  $t$ -rozdělení (pozor na **jinak značené** kritické hodnoty Studentova  $t$ -rozdělení)

$$P\left(\bar{X} - \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S_x}{\sqrt{n}} t_{n-1}(\alpha)\right) = 1 - \alpha$$

- ▶ jako odhad  $\sigma^2$  se použije výběrový rozptyl

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

- ▶ při velkých  $n$  ( $n \geq 50$ ) stačí použít  $z(\alpha/2)$  místo  $t_{n-1}(\alpha)$
- ▶ interval spolehlivosti se počítá i při odhadu jiných parametrů
- ▶ je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

## interval spolehlivosti při neznámém $\sigma$

- ▶ pro  $X_i$  s normálním rozdělením je třeba použít kritické hodnoty Studentova  $t$ -rozdělení (pozor na **jinak značené** kritické hodnoty Studentova  $t$ -rozdělení)

$$P\left(\bar{X} - \frac{S_x}{\sqrt{n}}t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S_x}{\sqrt{n}}t_{n-1}(\alpha)\right) = 1 - \alpha$$

- ▶ jako odhad  $\sigma^2$  se použije výběrový rozptyl

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

- ▶ při velkých  $n$  ( $n \geq 50$ ) stačí použít  $z(\alpha/2)$  místo  $t_{n-1}(\alpha)$
- ▶ interval spolehlivosti se počítá i při odhadu jiných parametrů
- ▶ je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

## interval spolehlivosti při neznámém $\sigma$

- ▶ pro  $X_i$  s normálním rozdělením je třeba použít kritické hodnoty Studentova  $t$ -rozdělení (pozor na **jinak značené** kritické hodnoty Studentova  $t$ -rozdělení)

$$P\left(\bar{X} - \frac{S_x}{\sqrt{n}}t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S_x}{\sqrt{n}}t_{n-1}(\alpha)\right) = 1 - \alpha$$

- ▶ jako odhad  $\sigma^2$  se použije výběrový rozptyl

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

- ▶ při velkých  $n$  ( $n \geq 50$ ) stačí použít  $z(\alpha/2)$  místo  $t_{n-1}(\alpha)$
- ▶ interval spolehlivosti se počítá i při odhadu jiných parametrů
- ▶ je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**



## interval spolehlivosti při neznámém $\sigma$

- ▶ pro  $X_i$  s normálním rozdělením je třeba použít kritické hodnoty Studentova  $t$ -rozdělení (pozor na **jinak značené** kritické hodnoty Studentova  $t$ -rozdělení)

$$P\left(\bar{X} - \frac{S_x}{\sqrt{n}}t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S_x}{\sqrt{n}}t_{n-1}(\alpha)\right) = 1 - \alpha$$

- ▶ jako odhad  $\sigma^2$  se použije výběrový rozptyl

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

- ▶ při velkých  $n$  ( $n \geq 50$ ) stačí použít  $z(\alpha/2)$  místo  $t_{n-1}(\alpha)$
- ▶ interval spolehlivosti se počítá i při odhadu jiných parametrů
- ▶ je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

## interval spolehlivosti při neznámém $\sigma$

- ▶ pro  $X_i$  s normálním rozdělením je třeba použít kritické hodnoty Studentova  $t$ -rozdělení (pozor na **jinak značené** kritické hodnoty Studentova  $t$ -rozdělení)

$$P\left(\bar{X} - \frac{S_x}{\sqrt{n}}t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S_x}{\sqrt{n}}t_{n-1}(\alpha)\right) = 1 - \alpha$$

- ▶ jako odhad  $\sigma^2$  se použije výběrový rozptyl

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

- ▶ při velkých  $n$  ( $n \geq 50$ ) stačí použít  $z(\alpha/2)$  místo  $t_{n-1}(\alpha)$
- ▶ interval spolehlivosti se počítá i při odhadu jiných parametrů
- ▶ je to interval, který s požadovanou pravděpodobností překryje odhadovaný parametr – **intervalový odhad**

## příklad: věk matek I

normální rozdělení dáno CLT a velkým  $n$

- ▶ 95% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left( 25,7 - 1,98 \cdot \frac{4,1}{\sqrt{99}}; 25,7 + 1,98 \cdot \frac{4,1}{\sqrt{99}} \right) = (24,9; 26,5)$$

`[confint(lm(vek.m~1,data=Kojeni))]`

- ▶ 99% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek (bude užší nebo širší?)
- ▶ větší jistota způsobí delší interval spolehlivosti (méně vypovídající tvrzení)

$$\left( 25,7 - 2,63 \cdot \frac{4,1}{\sqrt{99}}; 25,7 + 2,63 \cdot \frac{4,1}{\sqrt{99}} \right) = (24,6; 26,8)$$

`[confint(lm(vek.m~1,data=Kojeni),level=0.99)]`

## příklad: věk matek I

normální rozdělení dáno CLT a velkým  $n$

- ▶ 95% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left( 25,7 - 1,98 \cdot \frac{4,1}{\sqrt{99}}; 25,7 + 1,98 \cdot \frac{4,1}{\sqrt{99}} \right) = (24,9; 26,5)$$

`[confint(lm(vek.m~1,data=Kojeni))]`

- ▶ 99% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek (bude užší nebo širší?)
- ▶ větší jistota způsobí delší interval spolehlivosti (méně vypovídající tvrzení)

$$\left( 25,7 - 2,63 \cdot \frac{4,1}{\sqrt{99}}; 25,7 + 2,63 \cdot \frac{4,1}{\sqrt{99}} \right) = (24,6; 26,8)$$

`[confint(lm(vek.m~1,data=Kojeni),level=0.99)]`

## příklad: věk matek I

normální rozdělení dáno CLT a velkým  $n$

- ▶ 95% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left( 25,7 - 1,98 \cdot \frac{4,1}{\sqrt{99}}; 25,7 + 1,98 \cdot \frac{4,1}{\sqrt{99}} \right) = (24,9; 26,5)$$

`[confint(lm(vek.m~1,data=Kojeni))]`

- ▶ 99% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek (bude užší nebo širší?)
- ▶ větší jistota způsobí delší interval spolehlivosti (méně vypovídající tvrzení)

$$\left( 25,7 - 2,63 \cdot \frac{4,1}{\sqrt{99}}; 25,7 + 2,63 \cdot \frac{4,1}{\sqrt{99}} \right) = (24,6; 26,8)$$

`[confint(lm(vek.m~1,data=Kojeni),level=0.99)]`

## příklad: věk matek II

normální rozdělení dáno CLT a velkým  $n$

- ▶ 90% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left( 25,7 - 1,66 \cdot \frac{4,1}{\sqrt{99}}; 25,7 + 1,66 \cdot \frac{4,1}{\sqrt{99}} \right) = (25,0; 26,4)$$

[`confint(lm(vek.m~1,data=Kojeni),level=0.9)`]

- ▶ příklady **nesprávné interpretace** 90% intervalu spolehlivosti:
  - ▶ 90 % žen má věk v intervalu (25,0; 26,4)  
např. mezi našimi 99 matkami je jen 12 ve věku 25 a 10 ve věku 26 roků, navíc, s rostoucím  $n$  se interval zužuje
  - ▶ výběrový průměr věku matek je s pravděpodobností 90 % v intervalu (25,0; 26,4)  
výběrový průměr je uprostřed (tedy uvnitř) intervalu vždy

## příklad: věk matek II

normální rozdělení dáno CLT a velkým  $n$

- ▶ 90% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left( 25,7 - 1,66 \cdot \frac{4,1}{\sqrt{99}}; 25,7 + 1,66 \cdot \frac{4,1}{\sqrt{99}} \right) = (25,0; 26,4)$$

[`confint(lm(vek.m~1,data=Kojeni),level=0.9)`]

- ▶ příklady **nesprávné interpretace** 90% intervalu spolehlivosti:
  - ▶ 90 % žen má věk v intervalu (25,0; 26,4)  
např. mezi našimi 99 matkami je jen 12 ve věku 25 a 10 ve věku 26 roků, navíc, s rostoucím  $n$  se interval zužuje
  - ▶ výběrový průměr věku matek je s pravděpodobností 90 % v intervalu (25,0; 26,4)  
výběrový průměr je uprostřed (tedy uvnitř) intervalu **vždy**

## příklad: věk matek II

normální rozdělení dáno CLT a velkým  $n$

- ▶ 90% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left( 25,7 - 1,66 \cdot \frac{4,1}{\sqrt{99}}; 25,7 + 1,66 \cdot \frac{4,1}{\sqrt{99}} \right) = (25,0; 26,4)$$

[confint(lm(vek.m~1,data=Kojeni),level=0.9)]

- ▶ příklady **nesprávné interpretace** 90% intervalu spolehlivosti:
  - ▶ 90 % žen má věk v intervalu (25,0; 26,4)  
např. mezi našimi 99 matkami je jen 12 ve věku 25 a 10 ve věku 26 roků, navíc, s rostoucím  $n$  se interval zužuje
  - ▶ *výběrový průměr věku matek je s pravděpodobností 90 % v intervalu (25,0; 26,4)*  
výběrový průměr je uprostřed (tedy uvnitř) intervalu **vždy**



## příklad: věk matek II

normální rozdělení dáno CLT a velkým  $n$

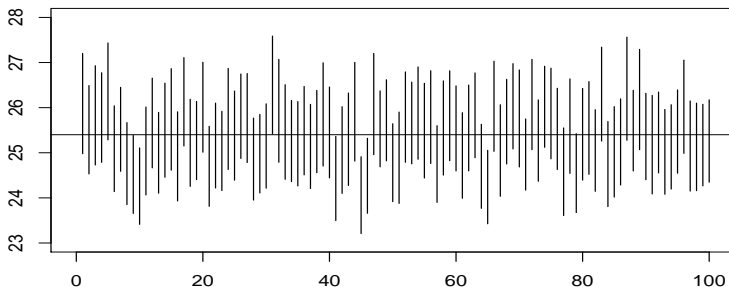
- ▶ 90% interval spolehlivosti pro populační průměr věku *všech* matek na základě výběru 99 matek

$$\left( 25,7 - 1,66 \cdot \frac{4,1}{\sqrt{99}}; 25,7 + 1,66 \cdot \frac{4,1}{\sqrt{99}} \right) = (25,0; 26,4)$$

[`confint(lm(vek.m~1,data=Kojeni),level=0.9)`]

- ▶ příklady **nesprávné interpretace** 90% intervalu spolehlivosti:
  - ▶ 90 % žen má věk v intervalu (25,0; 26,4)  
např. mezi našimi 99 matkami je jen 12 ve věku 25 a 10 ve věku 26 roků, navíc, s rostoucím  $n$  se interval zužuje
  - ▶ výběrový průměr věku matek je s pravděpodobností 90 % v intervalu (25,0; 26,4)  
výběrový průměr je uprostřed (tedy uvnitř) intervalu **vždy**

## simulované výběry pro $n = 100$ (věk matek)



znázorněno celkem 100 95% intervalů spolehlivosti pro  $\mu$   
ve skutečnosti mimořádně víme, že  $\mu = 25,4$   
v 7 případech je  $\mu$  nepřekryto

## centrální limitní věta pro četnosti

- ▶ (CLT obecně:) Nechtě  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$ . Potom pro velké  $n$  má průměr z nich přibl. rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , jejich součet přibl. rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶  $Y \sim \text{bi}(n, \pi)$ :  $Y$  je absolutní četnost výskytu jevu s pstí  $\pi$  v  $n$  nezáv. pokusech
- ▶  $Y = \sum_{i=1}^n X_i$  je součet nezávislých náhodných veličin  $X_i$  s alternativním rozdělením,  $X_i \sim \text{alt}(\pi)$ ,  $\text{var } X_i = \pi(1 - \pi)$
- ▶ podle CLT proto přibližně  $Y \sim N(n\pi, n\pi(1 - \pi))$
- ▶ relativní četnost  $Y/n = \bar{X}$  je průměr veličin s alternativním rozdělením, označme  $\hat{\pi} = Y/n$
- ▶ podle CLT je přibližně  $\hat{\pi} \sim N(\pi, \pi(1 - \pi)/n)$
- ▶  $\hat{\pi}$  je **nestranný** odhad  $\pi$

## centrální limitní věta pro četnosti

- ▶ (CLT obecně:) Nechtě  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$ . Potom pro velké  $n$  má průměr z nich přibl. rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , jejich součet přibl. rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶  $Y \sim \text{bi}(n, \pi)$ :  $Y$  je absolutní četnost výskytu jevu s pstí  $\pi$  v  $n$  nezáv. pokusech
- ▶  $Y = \sum_{i=1}^n X_i$  je součet nezávislých náhodných veličin  $X_i$  s alternativním rozdělením,  $X_i \sim \text{alt}(\pi)$ ,  $\text{var } X_i = \pi(1 - \pi)$
- ▶ podle CLT proto přibližně  $Y \sim N(n\pi, n\pi(1 - \pi))$
- ▶ relativní četnost  $Y/n = \bar{X}$  je průměr veličin s alternativním rozdělením, označme  $\hat{\pi} = Y/n$
- ▶ podle CLT je přibližně  $\hat{\pi} \sim N(\pi, \pi(1 - \pi)/n)$
- ▶  $\hat{\pi}$  je **nestranný** odhad  $\pi$

## centrální limitní věta pro četnosti

- ▶ (CLT obecně:) Nechtě  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$ . Potom pro velké  $n$  má průměr z nich přibl. rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , jejich součet přibl. rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶  $Y \sim \text{bi}(n, \pi)$ :  $Y$  je absolutní četnost výskytu jevu s pstí  $\pi$  v  $n$  nezáv. pokusech
- ▶  $Y = \sum_{i=1}^n X_i$  je součet nezávislých náhodných veličin  $X_i$  s alternativním rozdělením,  $X_i \sim \text{alt}(\pi)$ ,  $\text{var } X_i = \pi(1 - \pi)$
- ▶ podle CLT proto přibližně  $Y \sim N(n\pi, n\pi(1 - \pi))$
- ▶ relativní četnost  $Y/n = \bar{X}$  je průměr veličin s alternativním rozdělením, označme  $\hat{\pi} = Y/n$
- ▶ podle CLT je přibližně  $\hat{\pi} \sim N(\pi, \pi(1 - \pi)/n)$
- ▶  $\hat{\pi}$  je **nestranný** odhad  $\pi$

## centrální limitní věta pro četnosti

- ▶ (CLT obecně:) Nechtě  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$ . Potom pro velké  $n$  má průměr z nich přibl. rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , jejich součet přibl. rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶  $Y \sim \text{bi}(n, \pi)$ :  $Y$  je absolutní četnost výskytu jevu s pstí  $\pi$  v  $n$  nezáv. pokusech
- ▶  $Y = \sum_{i=1}^n X_i$  je součet nezávislých náhodných veličin  $X_i$  s alternativním rozdělením,  $X_i \sim \text{alt}(\pi)$ ,  $\text{var } X_i = \pi(1 - \pi)$
- ▶ podle CLT proto přibližně  $Y \sim N(n\pi, n\pi(1 - \pi))$
- ▶ relativní četnost  $Y/n = \bar{X}$  je průměr veličin s alternativním rozdělením, označme  $\hat{\pi} = Y/n$
- ▶ podle CLT je přibližně  $\hat{\pi} \sim N(\pi, \pi(1 - \pi)/n)$
- ▶  $\hat{\pi}$  je **nestranný** odhad  $\pi$

## centrální limitní věta pro četnosti

- ▶ (CLT obecně:) Nechtě  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$ . Potom pro velké  $n$  má průměr z nich přibl. rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , jejich součet přibl. rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶  $Y \sim \text{bi}(n, \pi)$ :  $Y$  je absolutní četnost výskytu jevu s pstí  $\pi$  v  $n$  nezáv. pokusech
- ▶  $Y = \sum_{i=1}^n X_i$  je součet nezávislých náhodných veličin  $X_i$  s alternativním rozdělením,  $X_i \sim \text{alt}(\pi)$ ,  $\text{var } X_i = \pi(1 - \pi)$
- ▶ podle CLT proto přibližně  $Y \sim N(n\pi, n\pi(1 - \pi))$
- ▶ relativní četnost  $Y/n = \bar{X}$  je průměr veličin s alternativním rozdělením, označme  $\hat{\pi} = Y/n$
- ▶ podle CLT je přibližně  $\hat{\pi} \sim N(\pi, \pi(1 - \pi)/n)$
- ▶  $\hat{\pi}$  je **nestranný** odhad  $\pi$

## centrální limitní věta pro četnosti

- ▶ (CLT obecně:) Nechtě  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$ . Potom pro velké  $n$  má průměr z nich přibl. rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , jejich součet přibl. rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶  $Y \sim \text{bi}(n, \pi)$ :  $Y$  je absolutní četnost výskytu jevu s pstí  $\pi$  v  $n$  nezáv. pokusech
- ▶  $Y = \sum_{i=1}^n X_i$  je součet nezávislých náhodných veličin  $X_i$  s alternativním rozdělením,  $X_i \sim \text{alt}(\pi)$ ,  $\text{var } X_i = \pi(1 - \pi)$
- ▶ podle CLT proto přibližně  $Y \sim N(n\pi, n\pi(1 - \pi))$
- ▶ relativní četnost  $Y/n = \bar{X}$  je průměr veličin s alternativním rozdělením, označme  $\hat{\pi} = Y/n$
- ▶ podle CLT je přibližně  $\hat{\pi} \sim N(\pi, \pi(1 - \pi)/n)$
- ▶  $\hat{\pi}$  je **nestranný** odhad  $\pi$



## centrální limitní věta pro četnosti

- ▶ (CLT obecně:) Nechtě  $X_1, X_2, \dots, X_n$  jsou nezávislé náhodné veličiny se stejným rozdělením, se střední hodnotou  $\mu$  a rozptylem  $\sigma^2 > 0$ . Potom pro velké  $n$  má průměr z nich přibl. rozdělení  $N\left(\mu, \frac{\sigma^2}{n}\right)$ , jejich součet přibl. rozdělení  $N(n\mu, n\sigma^2)$ .
- ▶  $Y \sim \text{bi}(n, \pi)$ :  $Y$  je absolutní četnost výskytu jevu s pstí  $\pi$  v  $n$  nezáv. pokusech
- ▶  $Y = \sum_{i=1}^n X_i$  je součet nezávislých náhodných veličin  $X_i$  s alternativním rozdělením,  $X_i \sim \text{alt}(\pi)$ ,  $\text{var } X_i = \pi(1 - \pi)$
- ▶ podle CLT proto přibližně  $Y \sim N(n\pi, n\pi(1 - \pi))$
- ▶ relativní četnost  $Y/n = \bar{X}$  je průměr veličin s alternativním rozdělením, označme  $\hat{\pi} = Y/n$
- ▶ podle CLT je přibližně  $\hat{\pi} \sim N(\pi, \pi(1 - \pi)/n)$
- ▶  $\hat{\pi}$  je **nestranný** odhad  $\pi$

## interval spolehlivosti pro pravděpodobnost $\pi$

- ▶ odmocnina z rozptylu odhadu  $\hat{\pi}$  je  $\sqrt{\frac{\pi(1-\pi)}{n}}$
- ▶ střední chyba relativní četnosti = směrodatná odchylka relativní četnosti
- ▶ pravděpodobnost  $\pi$  neznáme, odhadneme ji pomocí relativní četnosti  $\hat{\pi} = Y/n$
- ▶ odtud je  $100(1 - \alpha)\%$  přibližný interval spolehlivosti pro  $\pi$

$$\left( \hat{\pi} - z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}; \hat{\pi} + z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} \right)$$

- ▶ existují přesnější (pracnější) postupy  
[`prop.test(y,n,correct=FALSE)`]  
[`binom.test(y,n)`]

## interval spolehlivosti pro pravděpodobnost $\pi$

- ▶ odmocnina z rozptylu odhadu  $\hat{\pi}$  je  $\sqrt{\frac{\pi(1-\pi)}{n}}$
- ▶ střední chyba relativní četnosti = směrodatná odchylka relativní četnosti
- ▶ pravděpodobnost  $\pi$  neznáme, odhadneme ji pomocí relativní četnosti  $\hat{\pi} = Y/n$
- ▶ odtud je  $100(1 - \alpha)\%$  přibližný interval spolehlivosti pro  $\pi$

$$\left( \hat{\pi} - z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}; \hat{\pi} + z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} \right)$$

- ▶ existují přesnější (pracnější) postupy  
[`prop.test(y,n,correct=FALSE)`]  
[`binom.test(y,n)`]

## interval spolehlivosti pro pravděpodobnost $\pi$

- ▶ odmocnina z rozptylu odhadu  $\hat{\pi}$  je  $\sqrt{\frac{\pi(1-\pi)}{n}}$
- ▶ střední chyba relativní četnosti = směrodatná odchylka relativní četnosti
- ▶ pravděpodobnost  $\pi$  neznáme, odhadneme ji pomocí relativní četnosti  $\hat{\pi} = Y/n$
- ▶ odtud je  $100(1 - \alpha)\%$  přibližný interval spolehlivosti pro  $\pi$

$$\left( \hat{\pi} - z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}; \hat{\pi} + z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} \right)$$

- ▶ existují přesnější (pracnější) postupy  
[prop.test(y,n,correct=FALSE)]  
[binom.test(y,n)]

## interval spolehlivosti pro pravděpodobnost $\pi$

- ▶ odmocnina z rozptylu odhadu  $\hat{\pi}$  je  $\sqrt{\frac{\pi(1-\pi)}{n}}$
- ▶ střední chyba relativní četnosti = směrodatná odchylka relativní četnosti
- ▶ pravděpodobnost  $\pi$  neznáme, odhadneme ji pomocí relativní četnosti  $\hat{\pi} = Y/n$
- ▶ odtud je  $100(1 - \alpha)\%$  přibližný interval spolehlivosti pro  $\pi$

$$\left( \hat{\pi} - z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}; \hat{\pi} + z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} \right)$$

- ▶ existují přesnější (pracnější) postupy  
`[prop.test(y,n,correct=FALSE)]`  
`[binom.test(y,n)]`

## interval spolehlivosti pro pravděpodobnost $\pi$

- ▶ odmocnina z rozptylu odhadu  $\hat{\pi}$  je  $\sqrt{\frac{\pi(1-\pi)}{n}}$
- ▶ střední chyba relativní četnosti = směrodatná odchylka relativní četnosti
- ▶ pravděpodobnost  $\pi$  neznáme, odhadneme ji pomocí relativní četnosti  $\hat{\pi} = Y/n$
- ▶ odtud je  $100(1 - \alpha)\%$  přibližný interval spolehlivosti pro  $\pi$

$$\left( \hat{\pi} - z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}; \hat{\pi} + z(\alpha/2) \cdot \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} \right)$$

- ▶ existují přesnější (pracnější) postupy  
[prop.test(y,n,correct=FALSE)]  
[binom.test(y,n)]

## příklad: hody s hrací kostkou

- ▶ odhadujeme pravděpodobnost šestky,  $\alpha = 0,05$
- ▶ kostka A:  $n = 100, y = 17, \hat{\pi}_A = 0,17$

$$\left( 0,17 - 1,96 \cdot \sqrt{\frac{0,17 \cdot 0,83}{100}}; 0,17 + 1,96 \cdot \sqrt{\frac{0,17 \cdot 0,83}{100}} \right) = (0,10; 0,24)$$

- ▶ kostka B:  $n = 100, y = 41, \hat{\pi}_B = 0,41$

$$\left( 0,41 - 1,96 \cdot \sqrt{\frac{0,41 \cdot 0,59}{100}}; 0,41 + 1,96 \cdot \sqrt{\frac{0,41 \cdot 0,59}{100}} \right) = (0,31; 0,51)$$

- ▶ důležitý rozdíl: u kostky A patří  $1/6 = 0,167$  do 95% intervalu spolehlivosti; u kostky B nikoliv

## příklad: hody s hrací kostkou

- ▶ odhadujeme pravděpodobnost šestky,  $\alpha = 0,05$
- ▶ kostka A:  $n = 100, y = 17, \hat{\pi}_A = 0,17$

$$\left( 0,17 - 1,96 \cdot \sqrt{\frac{0,17 \cdot 0,83}{100}}; 0,17 + 1,96 \cdot \sqrt{\frac{0,17 \cdot 0,83}{100}} \right) = (0,10; 0,24)$$

- ▶ kostka B:  $n = 100, y = 41, \hat{\pi}_B = 0,41$

$$\left( 0,41 - 1,96 \cdot \sqrt{\frac{0,41 \cdot 0,59}{100}}; 0,41 + 1,96 \cdot \sqrt{\frac{0,41 \cdot 0,59}{100}} \right) = (0,31; 0,51)$$

- ▶ důležitý rozdíl: u kostky A patří  $1/6 = 0,167$  do 95% intervalu spolehlivosti; u kostky B nikoliv



## příklad: hody s hrací kostkou

- ▶ odhadujeme pravděpodobnost šestky,  $\alpha = 0,05$
- ▶ kostka A:  $n = 100, y = 17, \hat{\pi}_A = 0,17$

$$\left( 0,17 - 1,96 \cdot \sqrt{\frac{0,17 \cdot 0,83}{100}}; 0,17 + 1,96 \cdot \sqrt{\frac{0,17 \cdot 0,83}{100}} \right) = (0,10; 0,24)$$

- ▶ kostka B:  $n = 100, y = 41, \hat{\pi}_B = 0,41$

$$\left( 0,41 - 1,96 \cdot \sqrt{\frac{0,41 \cdot 0,59}{100}}; 0,41 + 1,96 \cdot \sqrt{\frac{0,41 \cdot 0,59}{100}} \right) = (0,31; 0,51)$$

- ▶ důležitý rozdíl: u kostky A patří  $1/6 = 0,167$  do 95% intervalu spolehlivosti; u kostky B nikoliv

## příklad: hody s hrací kostkou

- ▶ odhadujeme pravděpodobnost šestky,  $\alpha = 0,05$
- ▶ kostka A:  $n = 100, y = 17, \hat{\pi}_A = 0,17$

$$\left( 0,17 - 1,96 \cdot \sqrt{\frac{0,17 \cdot 0,83}{100}}; 0,17 + 1,96 \cdot \sqrt{\frac{0,17 \cdot 0,83}{100}} \right) = (0,10; 0,24)$$

- ▶ kostka B:  $n = 100, y = 41, \hat{\pi}_B = 0,41$

$$\left( 0,41 - 1,96 \cdot \sqrt{\frac{0,41 \cdot 0,59}{100}}; 0,41 + 1,96 \cdot \sqrt{\frac{0,41 \cdot 0,59}{100}} \right) = (0,31; 0,51)$$

- ▶ důležitý rozdíl: u kostky A patří  $1/6 = 0,167$  do 95% intervalu spolehlivosti; u kostky B nikoliv