Intervaly spolehlivosti

= intervalové odhady neznámého parametru (odhad pro π, μ, σ^2,…),

odvozují se z příslušné CLV

spolehlivost = $1-\alpha$

= pravděpodobnost, že neznámá hodnota parametru je intervalem pokryta;

nejčastěji volba $1-\alpha = 0{,}95$ (95% I.S.)

Oboustranné intervaly spolehlivosti

Pro střední hodnotu μ při známém σ:

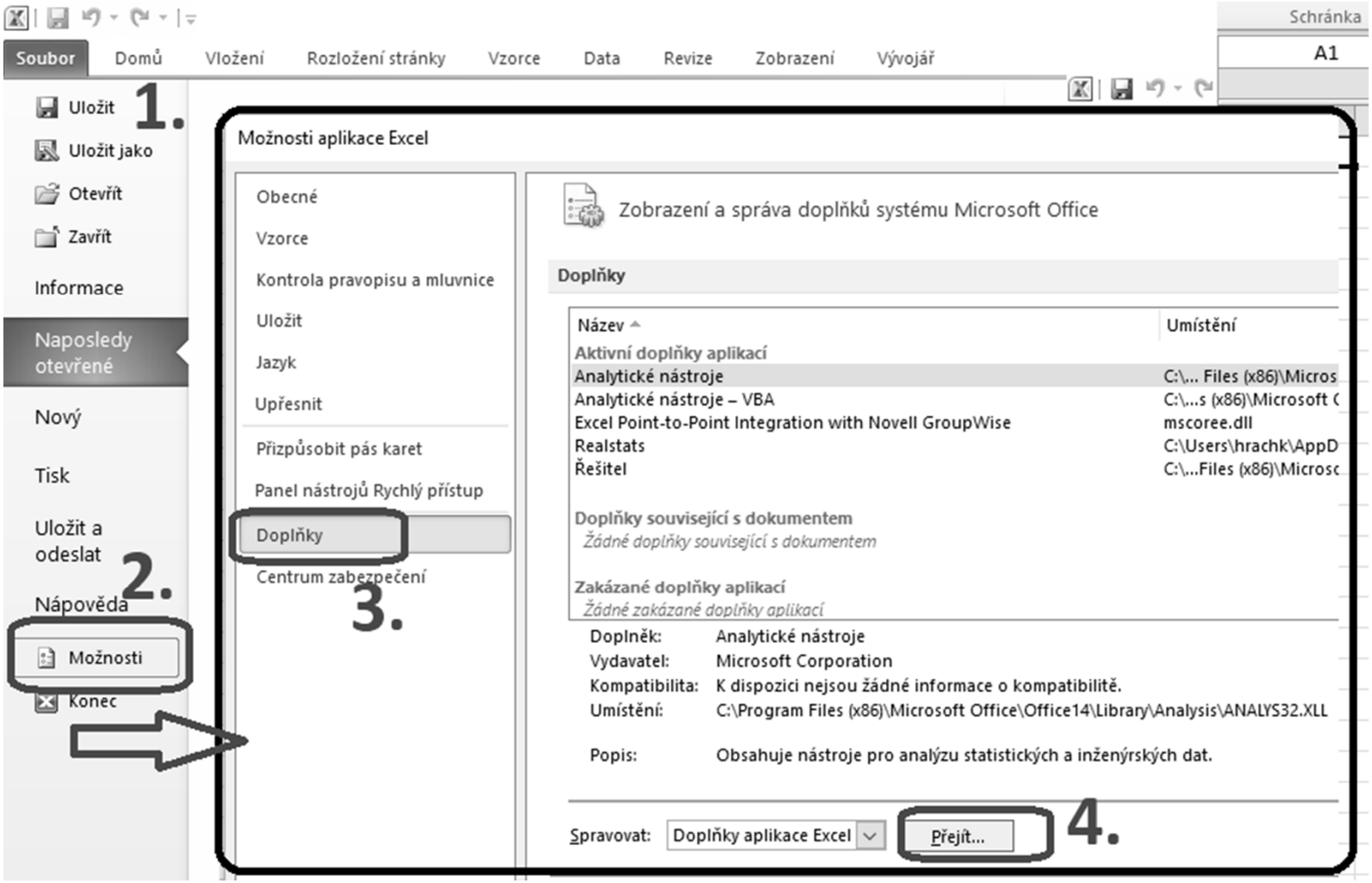
$$\bar{x} \pm u_{1-\alpha/2}\, \sigma/\sqrt{n}$$

Pro střední hodnotu μ při neznámém σ:

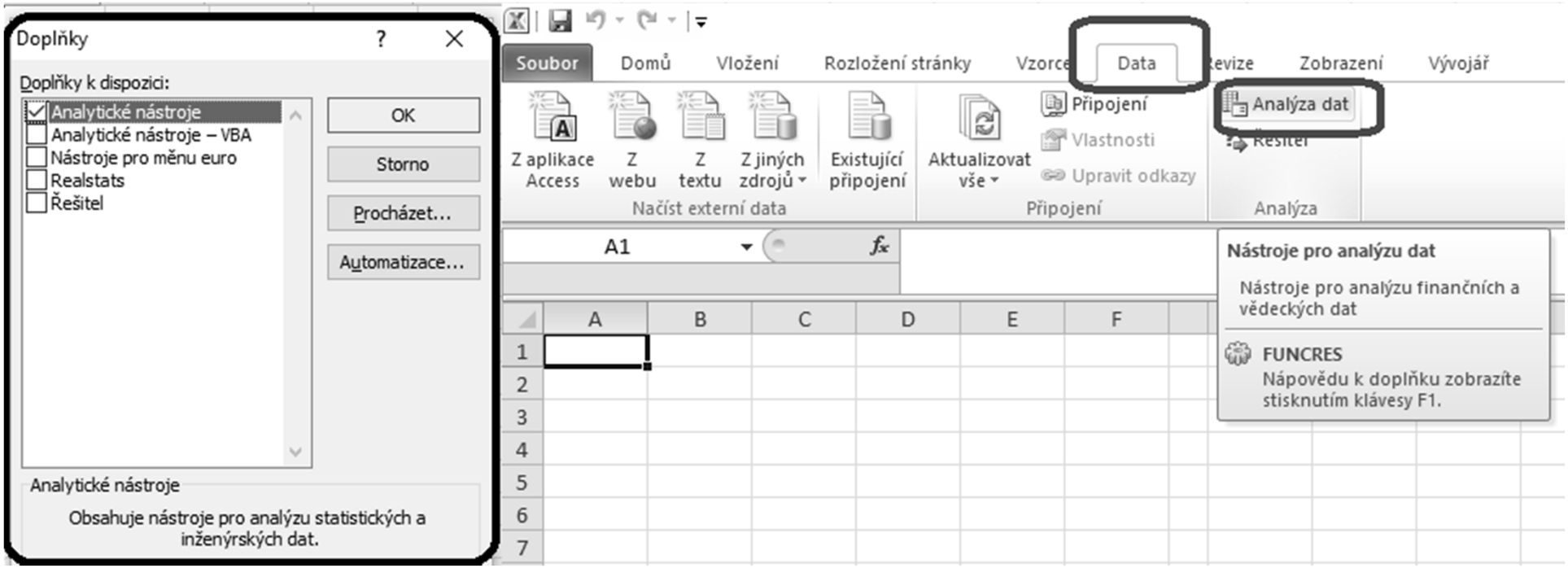
$$\bar{x} \pm t_{1-\alpha/2\ (n-1)}\, s/\sqrt{n}$$

kde n-1= počet stupňů volnosti (DF)

VSUVKA – aktivace nástroje ANALÝZA DAT v Excelu

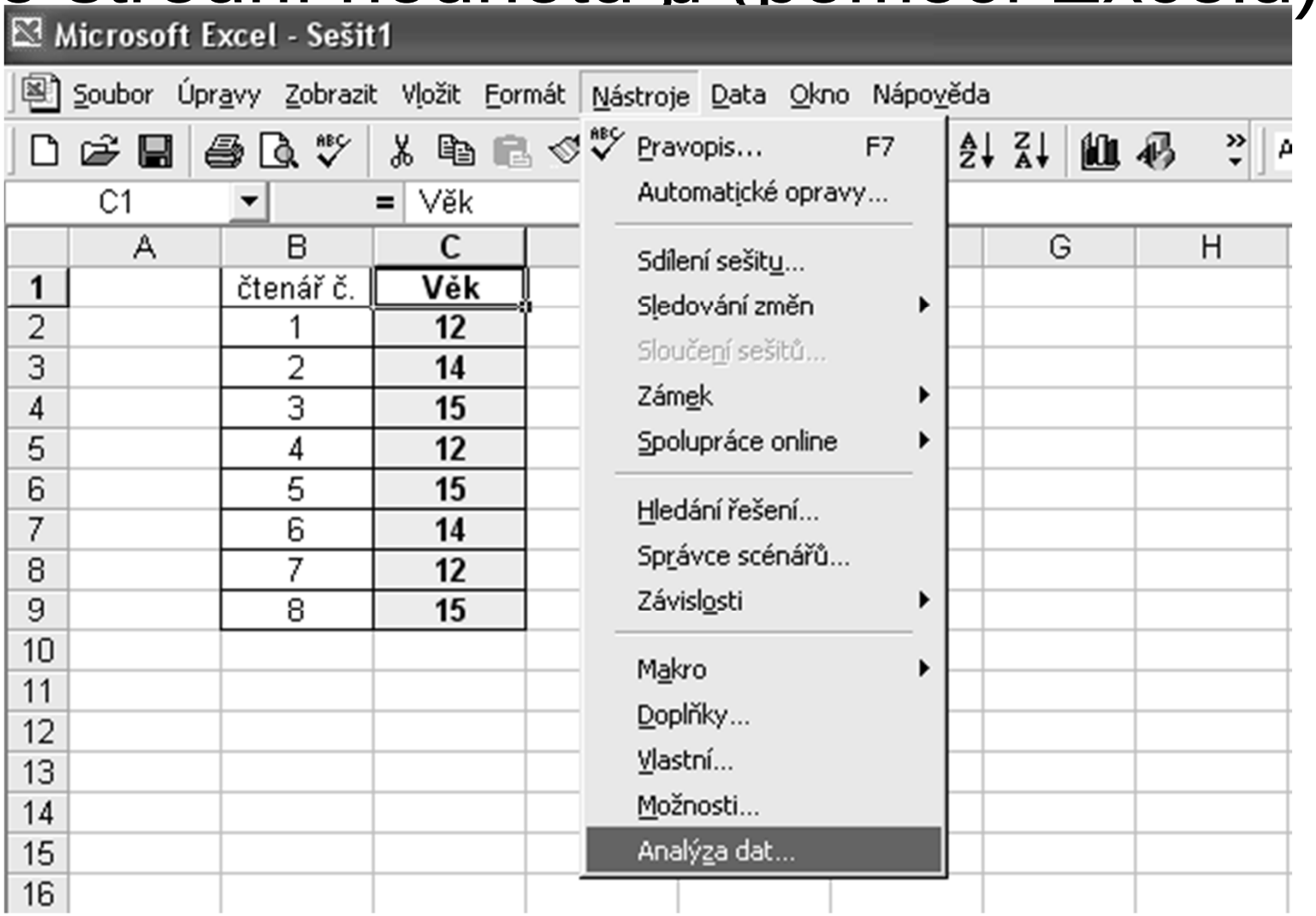


VSUVKA – aktivace nástroje ANALÝZA DAT v Excelu



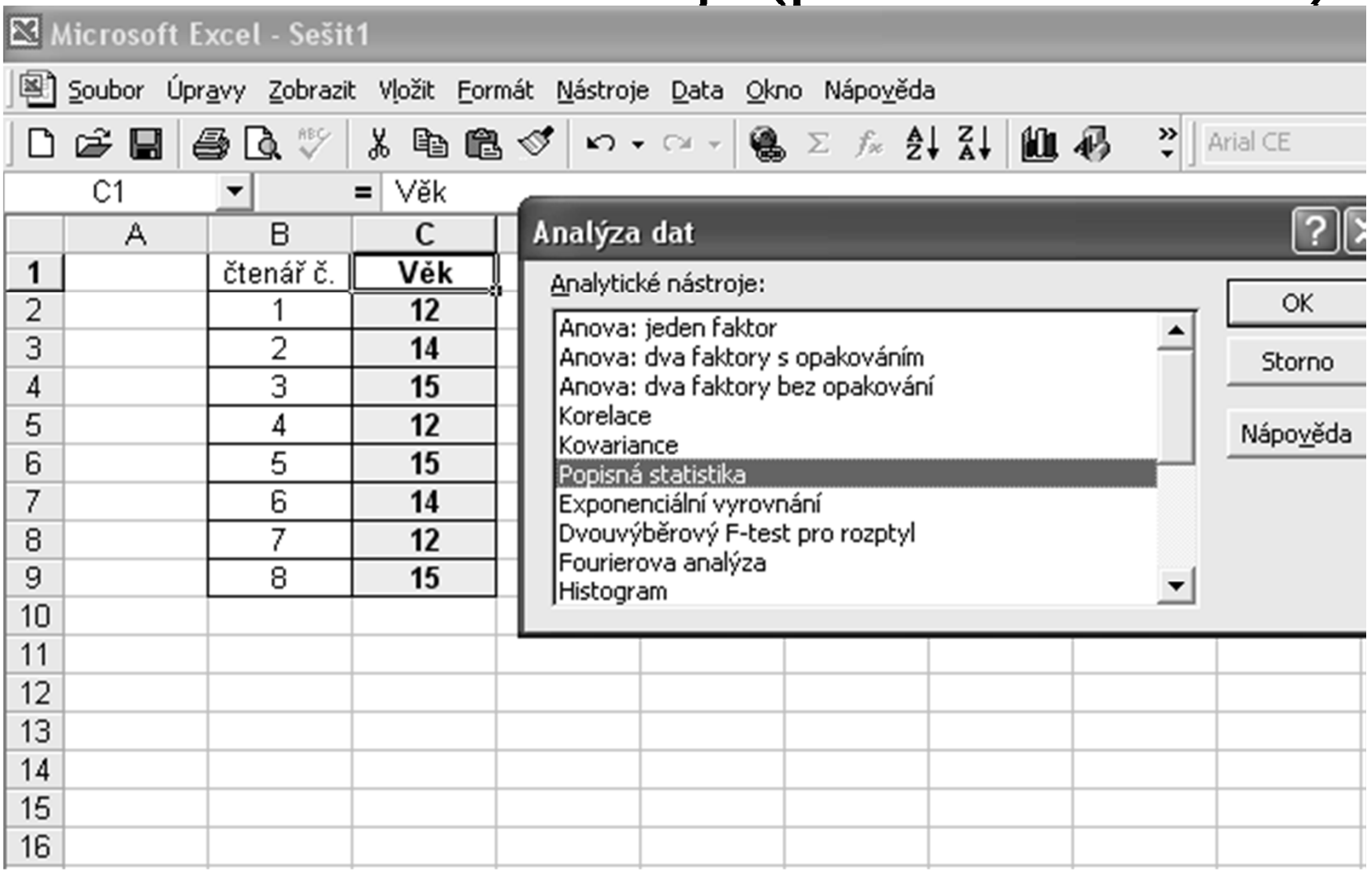
Oboustranné intervaly spolehlivosti

Pro střední hodnotu μ (pomocí Excelu):



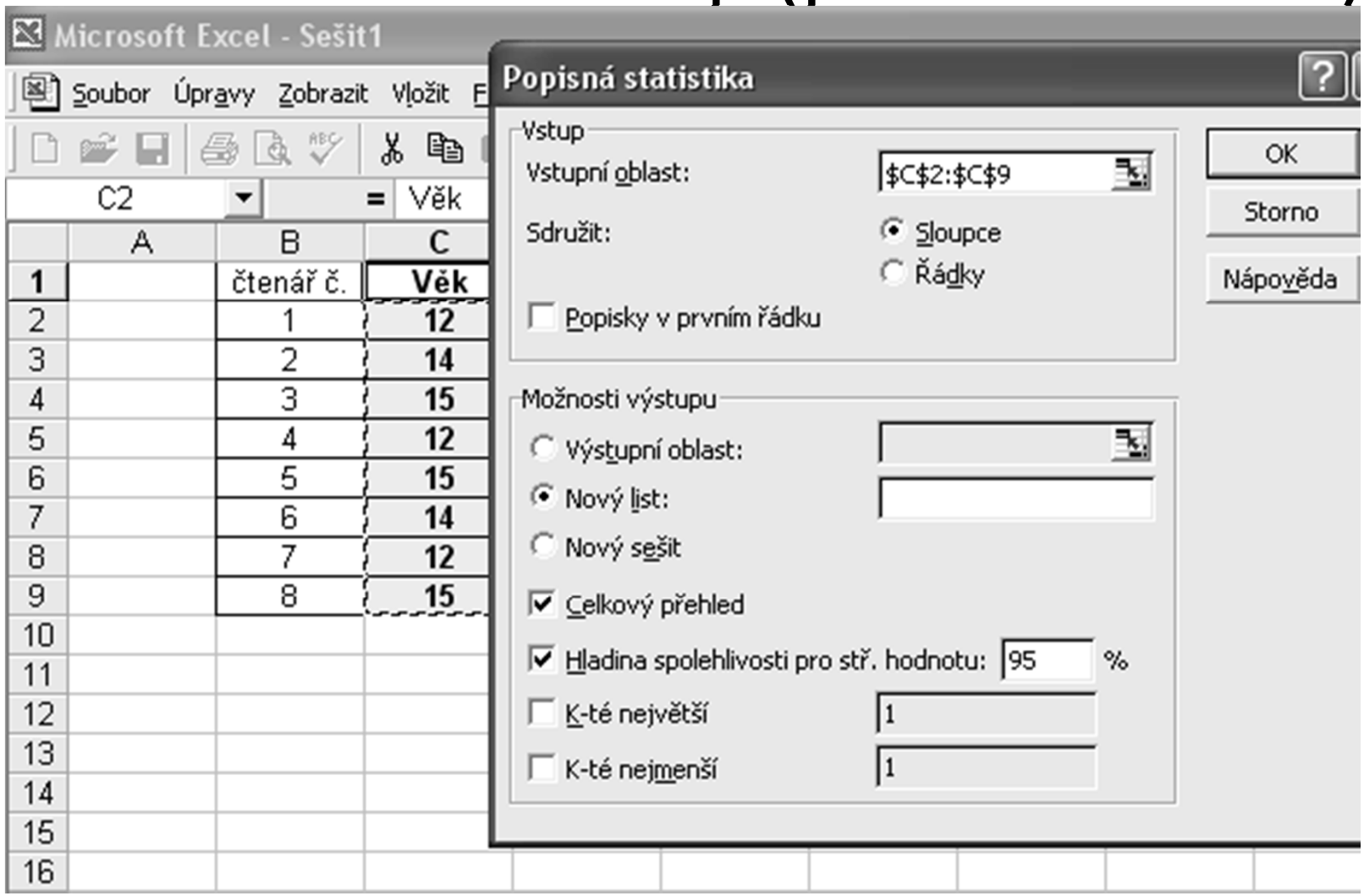
Oboustranné intervaly spolehlivosti

Pro střední hodnotu μ (pomocí Excelu):



Oboustranné intervaly spolehlivosti

Pro střední hodnotu μ (pomocí Excelu):



Oboustranné intervaly spolehlivosti

Pro střední hodnotu μ (pomocí Excelu):

	A	B
1	*Sloupec1*	
2		
3	Stř. hodnota	13,625
4	Chyba stř. hodnoty	0,498
5	Medián	14
6	Modus	12
7	Směr. odchylka	1,408
8	Rozptyl výběru	1,982
9	Špičatost	-2,135
10	Šikmost	-0,339
11	Rozdíl max-min	3
12	Minimum	12
13	Maximum	15
14	Součet	109
15	Počet	8
16	Hladina spolehlivosti (95,0%)	1,177

dolní mez:

13,625-1,177=

=12,448;

horní mez:

13,625+1,177=

=14,802

Oboustranné intervaly spolehlivosti

Pro střední hodnotu μ (odpověď):

S 95% spolehlivostí je střední věk čtenářů daného časopisu z rozmezí 12,448 až 14,802 roku.

Zpřesnění odhadu (tj. zúžení IS)?

a) zvýšit n (=změna dat);

b) snížit spolehlivost (data stejná);

c) snížit variabilitu (=změna populace).

Oboustranné intervaly spolehlivosti

Ilustrace vlivu zvýšení n (viz ZVČ):

Jednostranné intervaly spolehlivosti

⇔ hledáme jen jednu z obou mezí

Princip:

dle zadání úlohy hledáme jen dolní či jen horní mez podle „oboustranného“ vzorce s tou změnou, že výraz **1-α/2** ve vzorci nahradíme výrazem **1-α.**

Testování hypotéz

Otestujte / ověřte / prokažte…

že střední věk (tj. μ)

…činí 40 let (=40)

…je alespoň 40 let (>40)

Testování hypotéz

Otestujte / ověřte / prokažte…

zda hmotnost vejce

…závisí na jeho délce

…je různá dle typu snůšky

Testování hypotéz

Nulová hypotéza H_0:

pevně daná forma (nerozhoduje slovní formulace problému!); u parametrických testů obsahuje H_0 rovnost,

v jiných speciálních případech obsahuje H_0 např. tvrzení o nezávislosti

Alternativní hypotéza H_1:

doplněk k H_0

Testování hypotéz

Postup rozhodování při použití statistického SW (i např. Excel) – nelze „ručně“:

a) Z dat spočte počítač **p-hodnotu**
   (je vždy mezi 0-1)

b) Porovnáme p-hodnotu s předem zvolenou α :

c) Pokud je $p \leq \alpha$, zamítáme při daném α nulovou hypotézu ve prospěch hypotézy alternativní

d) Pokud naopak je $p > \alpha$, nelze při daném α zamítnout nulovou hypotézu ve prospěch hypotézy alternativní

Testování hypotéz

Jaký význam má ono volené α?

Možnosti při testování:	Doopravdy platí H_0	Doopravdy platí H_1
Dle dat vyberu H_0	OK	„chyba 2. druhu“
Dle dat zamítnu H_0	„chyba 1. druhu“	OK

α=P(chyby 1.druhu) … „hladina významnosti“

Jednovýběrový t-test

- Pro střední hodnotu μ :

a) $H_0: \mu=\mu_0$ $H_1: \mu\neq\mu_0$

(oboustranná alternativa)

b) $H_0: \mu=\mu_0$ $H_1: \mu>\mu_0$

c) $H_0: \mu=\mu_0$ $H_1: \mu<\mu_0$

(jednostranné alternativy)

Vždy μ_0 je konkrétní testovaná hodnota.

Jednovýběrový t-test

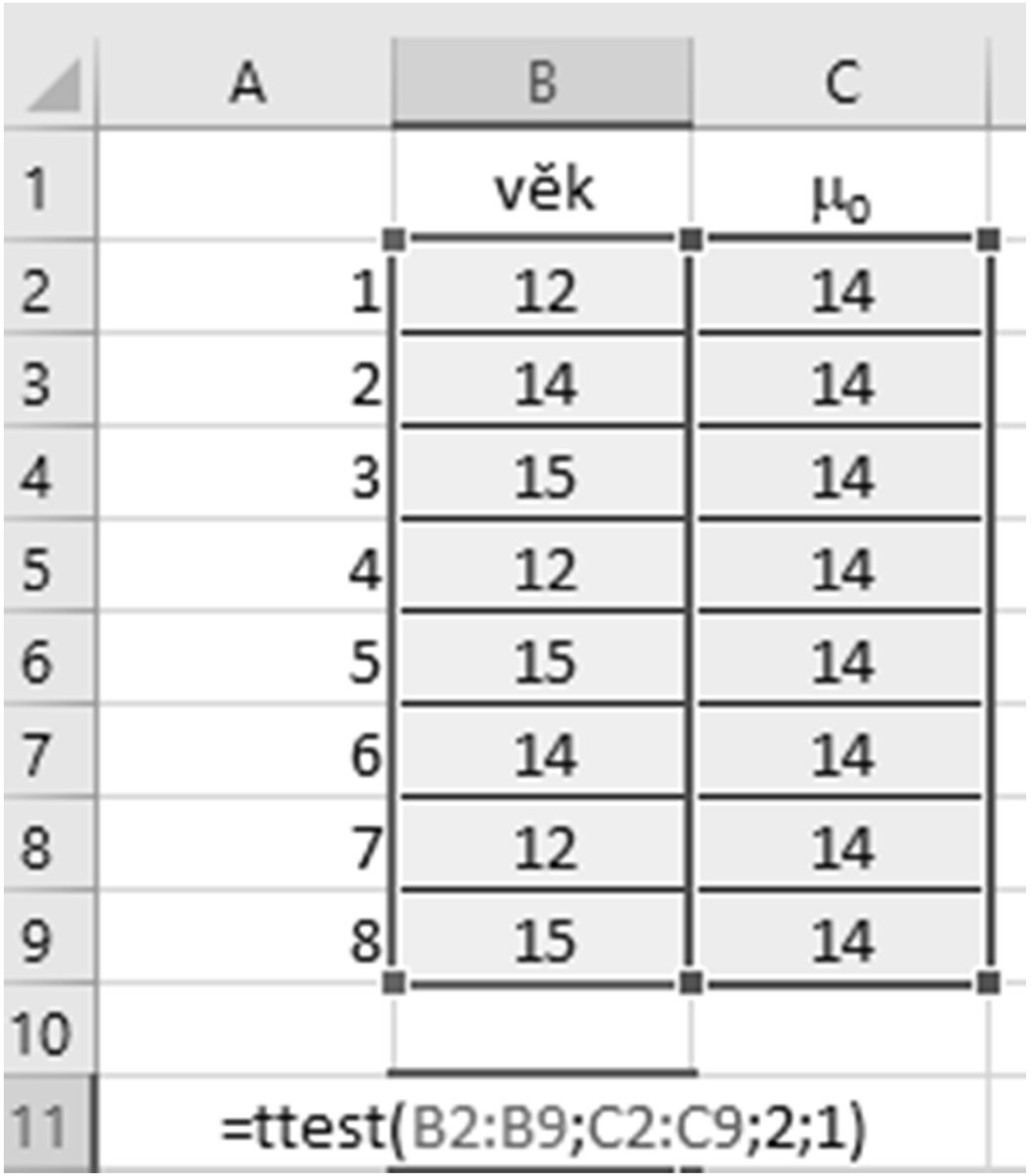
Příklad:

Dle věku osmi náhodně vybraných čtenářů dětského časopisu ověřte pravdivost tvrzení, že střední věk čtenářů tohoto časopisu je 14 let.

Věky popořadě (*viz interval spolehlivosti*):

12, 14, 15, 12, 15, 14, 12, 15.

$H_0: \mu=14 \qquad H_1: \mu\neq14$



	A	B	C
1		věk	μ_0
2	1	12	14
3	2	14	14
4	3	15	14
5	4	12	14
6	5	15	14
7	6	14	14
8	7	12	14
9	8	15	14
10			
11	=ttest(B2:B9;C2:C9;2;1)		

Výsledek =
p-hodnota

```
> vek=c(12,14,15,12,15,14,12,15)
> t.test(vek,mu=14)

        One Sample t-test

data:  vek
t = -0.75337, df = 7, p-value = 0.4758
alternative hypothesis: true mean is not equal to 14
95 percent confidence interval:
 12.44798 14.80202
sample estimates:
mean of x
   13.625
```

Jednovýběrový t-test

Příklad (*výsledek*):

p=0,48 (>0,05) $\Rightarrow$ nelze zamítnout H_0

Příklad (*odpověď*):

Na 5% hladině významnosti
nelze na základě dat zamítnout tvrzení,
že střední věk čtenářů činí 14 let.

Příklad (*k zamyšlení*):

Je nějaký vztah mezi tímto výsledkem a intervalem spolehlivosti [12,448; 14,802]?

Párový t-test

Sledujeme spojitou číselnou veličinu (např.):

- **Hmotnost zvířete před uložením do zimního spánku a po probuzení;**
- **Změření výšky stromu dvěma metodami (standardní a novou, experimentální); …**

Chceme prokázat:

- **Způsobuje zimní spánek významné snížení hmotnosti?**
- **Jsou výškové údaje zjištěné novou, experimentální metodou, srovnatelné s údaji zjištěnými osvědčenou standardní metodou?**

Párový t-test

Data jsou ve tvaru <u>párů</u>, tj. uspořádaných dvojic (též tzv. <u>závislé výběry</u> – dependent samples):

$$(y_1;z_1), \ldots, (y_n; z_n),$$ **kde např.**

- y_i **= hmotnost před zimním spánkem (i-té zvíře),**
  z_i **= hmotnost po probuzení (i-té zvíře)**

  i=1,…,n n=počet zvířat (2n=počet údajů)

- y_i **= výška i-tého stromu zjištěná standardně,**
  z_i **= výška i-tého stromu zjištěná novou metodou**

  i=1,…,n n=počet stromů (2n=počet údajů)

Párový t-test

Řešení:

- **Představíme-li si rozdíly** (*není je ale ani nutno skutečně počítat*)

$$x_i = y_i - z_i \qquad (i=1,\ldots,n)$$

- **vznikla by „rozdílová“ veličina X, jejíž střední hodnotu označíme** μ

Párový t-test

Řešení:

- **pro „rozdílovou“ veličinu X provedeme parametrický test s nulovou hypotézou** $H_0: \mu=0$ **(mezi hodnotami v párech <u>není</u> významný rozdíl)**

  a s alternativou $H_1: \mu\neq0$ **(mezi hodnotami v párech <u>je</u> významný rozdíl),**

  příp. (často) s alternativou jednostrannou

  (viz příklad se zimním spánkem)

Párový t-test

Příklad:

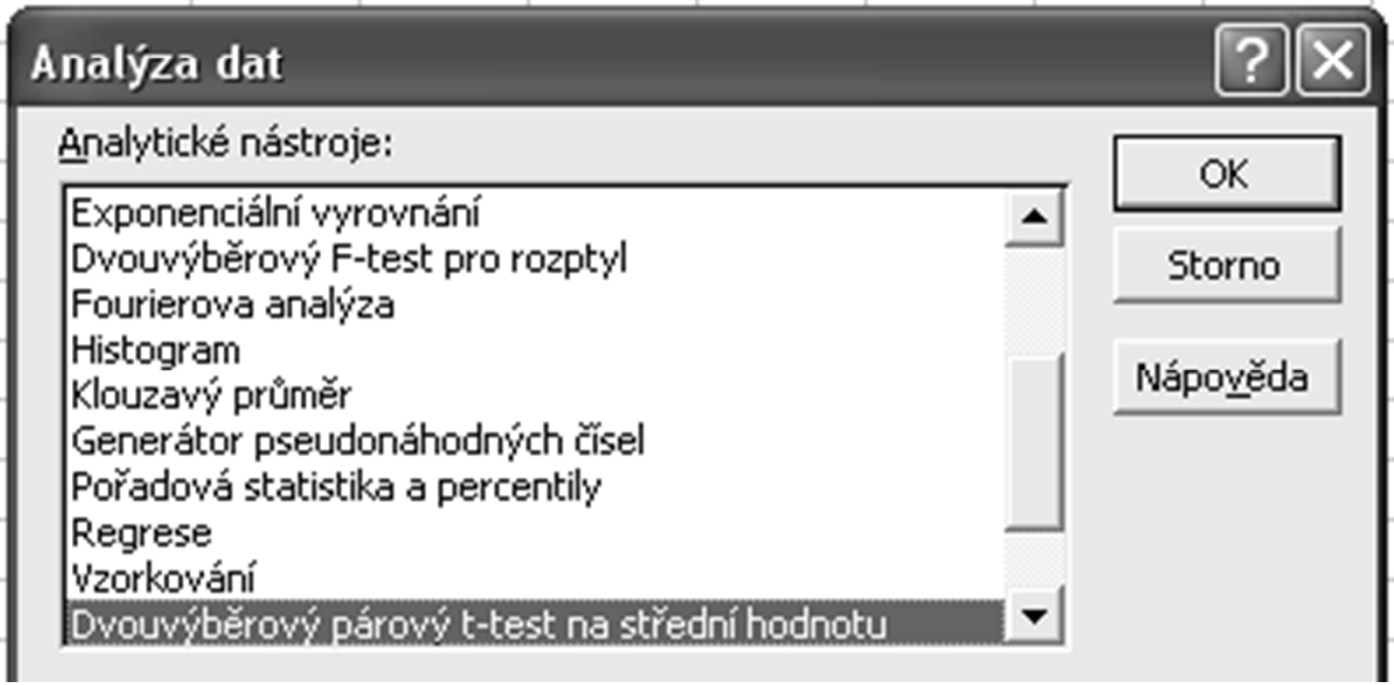
Posuďte na základě uvedených dat, zda obě metody určují v průměru výšku objektů srovnatelně. Osm objektů měřeno, každý oběma metodami:

Výška standardní metodou (m)	**1,2**	**2,4**	**1,6**	**1,8**	**3,2**	**2,7**	**2,0**	**1,9**
Výška experimen-tálně ověřovanou novou metodou (m)	**1,3**	**2,4**	**1,8**	**1,7**	**3,3**	**3,1**	**1,8**	**2,2**

Párový t-test

Počítačové řešení (Excelem):

	A	B	C	D	E	F	G	H	I
1	standard	1,2	2,4	1,6	1,8	3,2	2,7	2,0	1,9
2	experim.	1,3	2,4	1,8	1,7	3,3	3,1	1,8	2,2
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									



Párový t-test

	A	B	C	D	E	F	G	H	I
1	standard	1,2	2,4	1,6	1,8	3,2	2,7	2,0	1,9
2	experim.	1,3	2,4	1,8	1,7	3,3	3,1	1,8	2,2
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									

Párový t-test

	A	B	C	D	E	F
1	**Dvouvýběrový párový t-test na střední hodnotu**					
2						
3		*Soubor 1*	*Soubor 2*			
4	Stř. hodnota	2,1	2,2			
5	Rozptyl	0,4086	0,4914			
6	Pozorování	8	8			
7	Pears. korelace	0,9596				
8	Hyp. rozdíl stř. hodnot	0				
9	Rozdíl	7				
10	t stat	-1,4142				
11	P(T<=t) (1)	0,1001				
12	t krit (1)	1,8946				
13	**P(T<=t) (2)**	**0,2002**				
14	t krit (2)	2,3646				

p-hodnota

$p > \alpha$ (0,2002>0,05) $\Rightarrow$ nelze zamítnout H_0

Párový t-test

Excel „rychle“: =TTEST(B1:I1;B2:I2;2;1)

R:

```
> stand=c(1.2,2.4,1.6,1.8,3.2,2.7,2.0,1.9)
> experim=c(1.3,2.4,1.8,1.7,3.3,3.1,1.8,2.2)
> t.test(stand,experim,paired=T)

        Paired t-test

data:  stand and experim
t = -1.4142, df = 7, p-value = 0.2002
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.26720418  0.06720418
sample estimates:
mean of the differences
                  -0.1
```

Pro zkoumanou veličinu by ale měla platit NORMALITA; ne-li, co dělat?

Párový Wilcoxonův test

neparametrická verze párového testu;

netestujeme chování parametru μ

testujeme shodu (H_0), resp. rozdílnost (H_1) polohy obou závislých výběrů:

```
> stand=c(1.2,2.4,1.6,1.8,3.2,2.7,2.0,1.9)
> experim=c(1.3,2.4,1.8,1.7,3.3,3.1,1.8,2.2)
> wilcox.test(stand,experim,paired=T)

        Wilcoxon signed rank test with continuity correction

data:  stand and experim
V = 7, p-value = 0.2702
alternative hypothesis: true location shift is not equal to 0
```

Rozhodnutí: $p=0{,}2702 > 0{,}05$ →

Nelze zamítnout H_0. *aneb*:

Data neprokázala významnou odlišnost mezi oběma metodami. *aneb*:

Nová metoda měří srovnatelně s tou standardní.

```
> stand=c(1.2,2.4,1.6,1.8,3.2,2.7,2.0,1.9)
> experim=c(1.3,2.4,1.8,1.7,3.3,3.1,1.8,2.2)
> wilcox.test(stand,experim,paired=T)

        Wilcoxon signed rank test with continuity correction

data:  stand and experim
V = 7, p-value = 0.2702
alternative hypothesis: true location shift is not equal to 0
```

Dvouvýběrové testy

Sledujeme (spojité) číselné veličiny, např.:

- **výšku ve skupině mužů a ve skupině žen;**
- **hmotnost vajec ptačích druhů A a B;**

Chceme prokázat:

- **Je/není výška mužů a žen srovnatelná?**
- **Je/není mezi oběma ptačími druhy výrazný rozdíl ve hmotnosti vajec?**

Dvouvýběrové parametrické testy

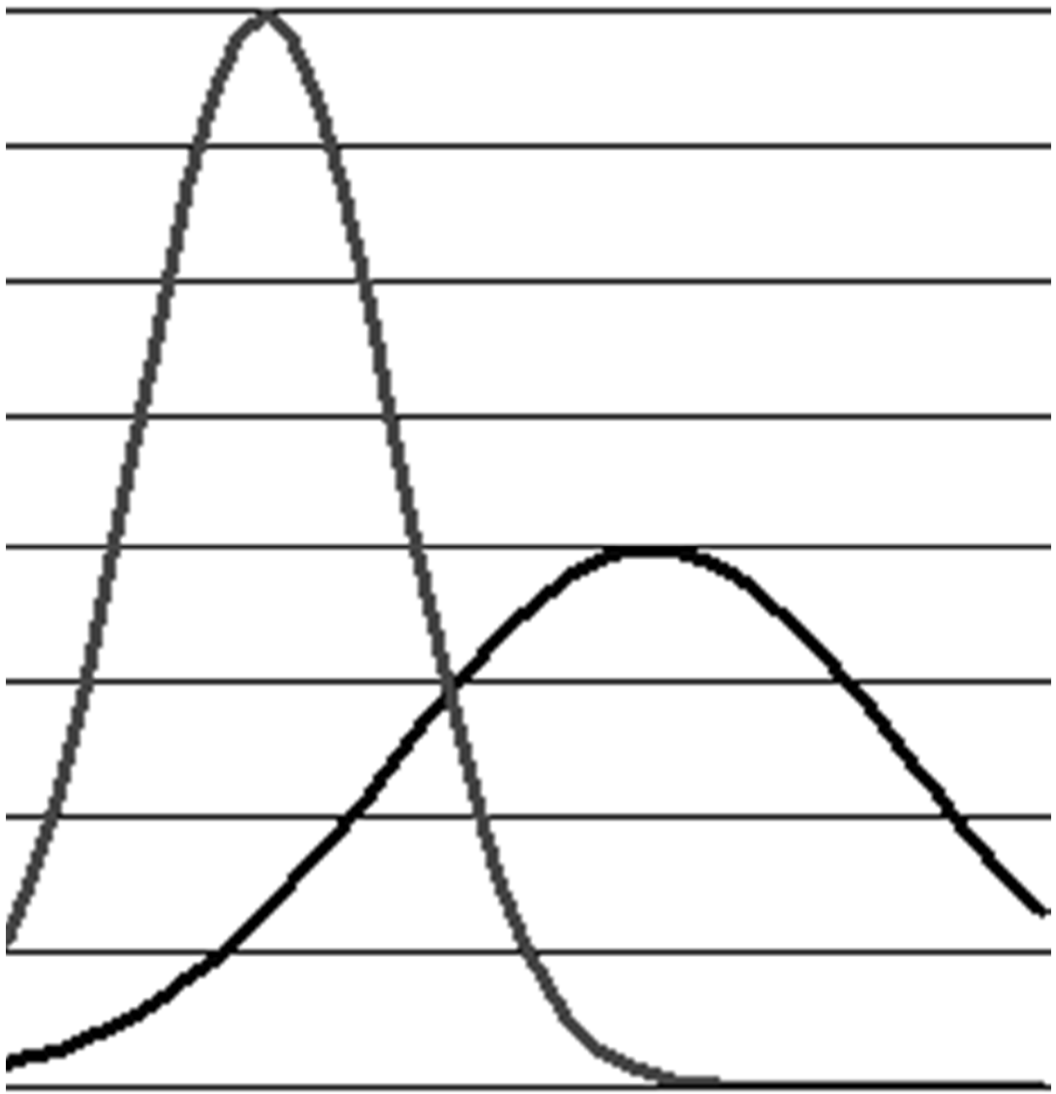
Předpoklady:

- **obě skupiny jsou nezávislé** *(např. u mužů a žen nejde o manželské páry)*
- **sledovaná veličina se v obou srovnávaných skupinách chová jako veličina normálně rozdělená**

Dvouvýběrové parametrické testy

Možnosti:

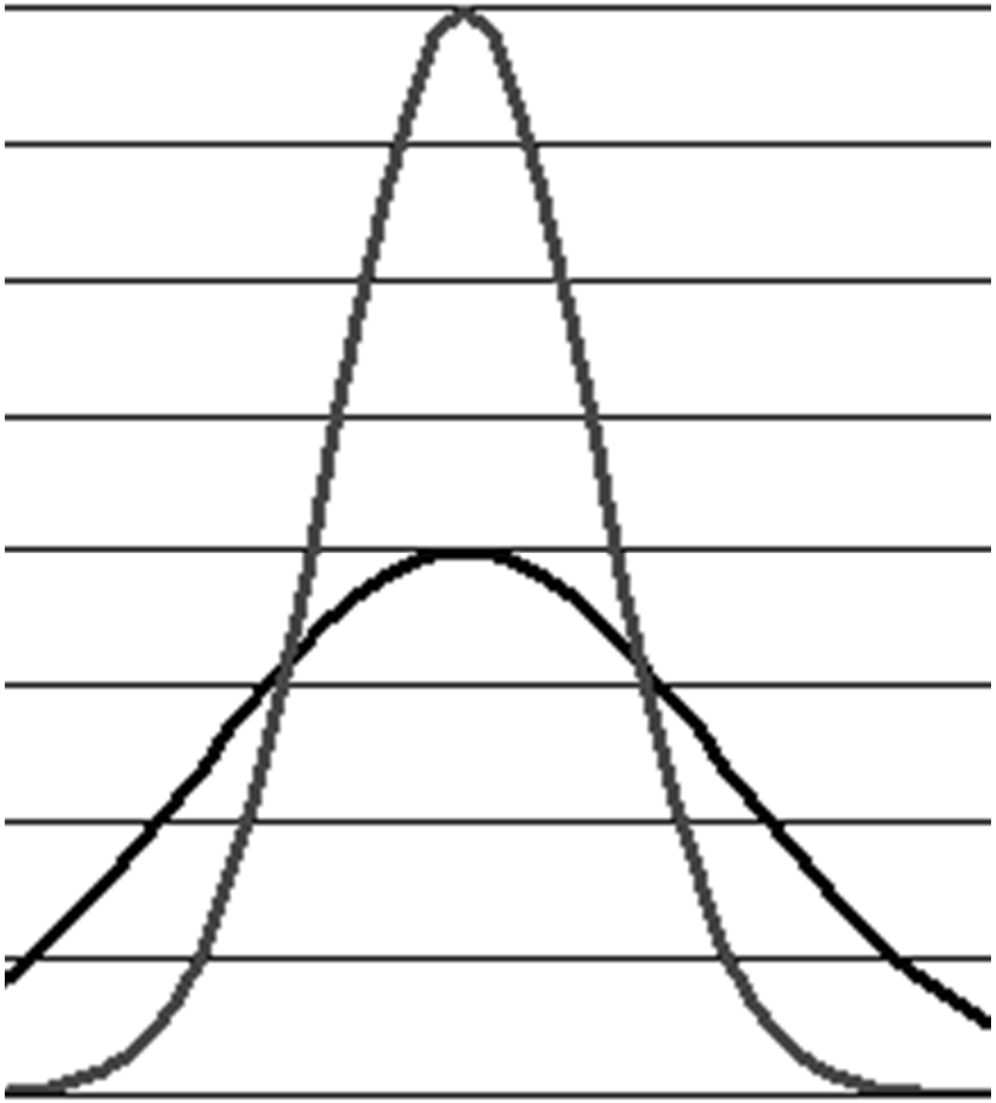
a) v obou skupinách odlišné
μ **i** σ^2



Dvouvýběrové parametrické testy

Možnosti:

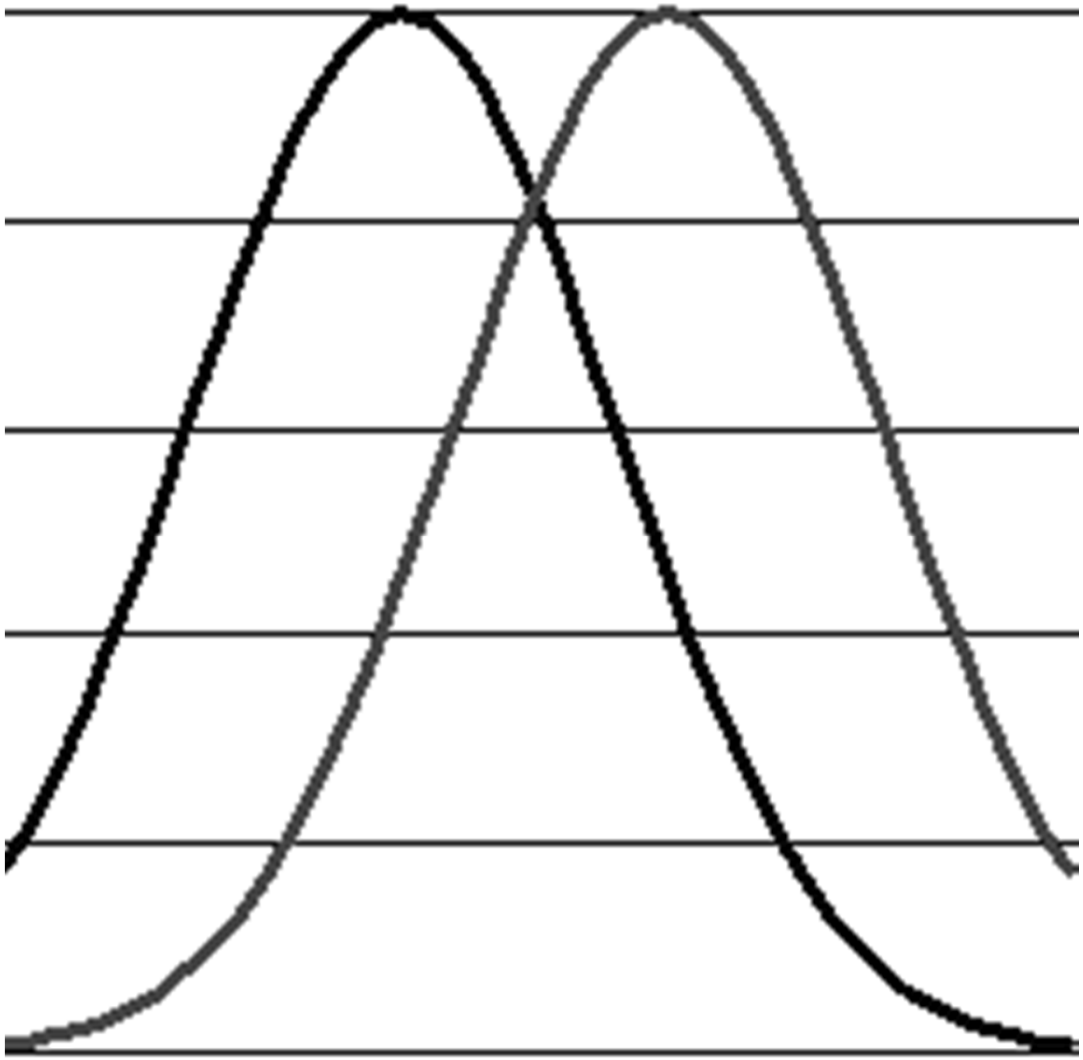
b) v obou skupinách shodné μ, σ^2 odlišné



Dvouvýběrové parametrické testy

Možnosti:

c) v obou skupinách odlišné μ, σ^2 shodné

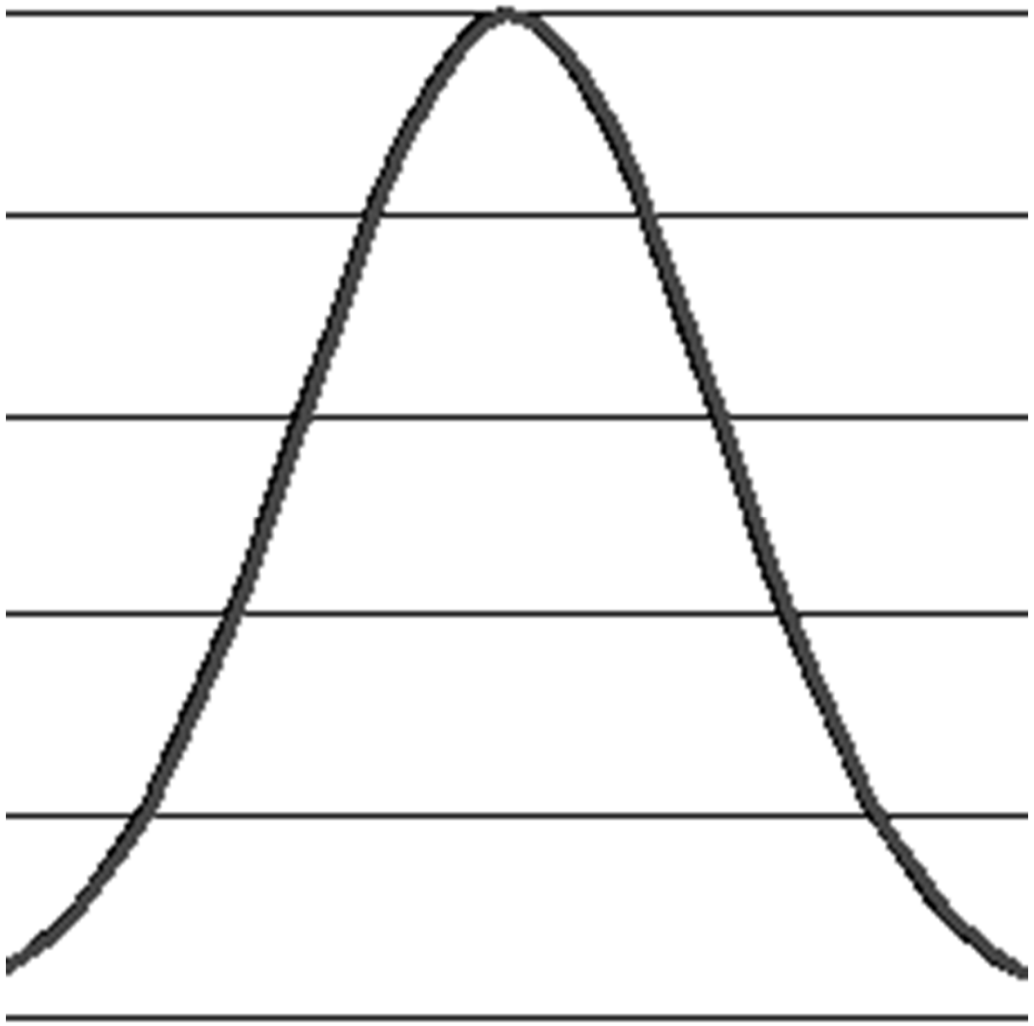


Dvouvýběrové parametrické testy

Možnosti:

d) v obou skupinách shodné

μ **i** σ^2



Dvouvýběrové parametrické testy

Testujeme tedy (popořadě):

1. **Je v obou skupinách srovnatelná či naopak výrazně odlišná variabilita?**

ANEB:

$H_0: \sigma_a^2 = \sigma_b^2$ versus $H_1: \sigma_a^2 \neq \sigma_b^2$

Jde o tzv. F-test homogenity rozptylů.

Dvouvýběrové parametrické testy

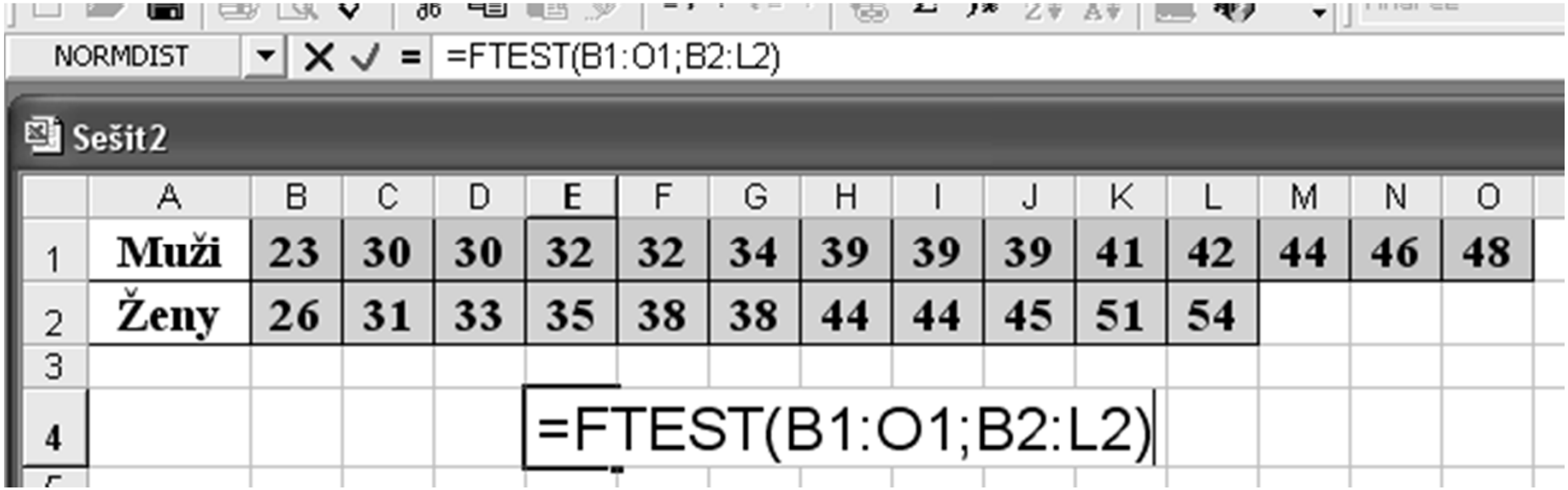
Příklad: Zjistěte, zda ve sledované populaci <u>závisí</u> chování veličiny věk na pohlaví.

Jde o příklad testu (ne)závislosti spojité veličiny na veličině alternativní.

Jinak řečeno, porovnáváme chování veličiny věk u mužů a u žen… Je to tedy opravdu 2-výběrový test.

Nejprve zjistíme, zda je v obou skupinách srovnatelná variabilita:

Dvouvýběrové parametrické testy



Výsledkem je p-hodnota (zde =0,522); protože p>0,05, nelze zamítnout H_0. Znamená to, že variabilita obou skupin je srovnatelná.

Dvouvýběrové parametrické testy

Následně testujeme:

2. Jsou v obou skupinách srovnatelné či naopak výrazně odlišné střední hodnoty?

ANEB:

$$H_0: \mu_a = \mu_b \qquad \text{versus} \qquad H_1: \mu_a \neq \mu_b$$

Jde o tzv. 2-výběrový t-test. Na F-test navazuje proto, že existuje ve dvou variantách: při různých a při shodných rozptylech.

Dvouvýběrové parametrické testy

Příklad - pokračování: Už víme, že veličina věk má ve sledované populaci srovnatelný rozptyl u mužů a u žen. Teď ověříme, zda je i střední hodnota (střední věk) u obou pohlaví srovnatelná, nebo zda se výrazně liší u mužů a u žen.

Dvouvýběrové parametrické testy

Dvouvýběrový t-test s rovností rozptylů		
	Soubor 1	*Soubor 2*
Stř. hodnota	37,071	39,909
Rozptyl	50,533	73,291
Pozorování	14	11
Společný rozptyl	60,428	
Hyp. rozdíl stř. hodr	0	
Rozdíl	23	
t stat	-0,906	
P(T<=t) (1)	0,187	
t krit (1)	1,714	
P(T<=t) (2)	**0,374**	
t krit (2)	2,069	

p-hodnota = 0,374

p>0,05⇒nelze zamítnout H_0

Nejen rozptyl věku, ale také střední věk mužů a žen je srovnatelný.

Jinak řečeno (celkově) – ve sledované populaci nebyla zjištěna závislost věku na pohlaví (situaci odpovídají Gaussovy křivky dle d)).

Dvouvýběrové parametrické testy

Excel „rychleji“ (příkaz určující jen p-hodnotu):

=TTEST(data1;data2;1;2) … jednostranně, shodné rozptyly

=TTEST(data1;data2;2;2) … oboustranně, shodné rozptyly

=TTEST(data1;data2;1;3) … jednostranně, různé rozptyly

=TTEST(data1;data2;2;3) … oboustranně, různé rozptyly

R:

(verze obecná, při neshodě rozptylů)

```
> muzi=c(23,30,30,32,32,34,39,39,39,41,42,44,46,48)
> zeny=c(26,31,33,35,38,38,44,44,45,51,54)
> t.test(muzi,zeny)

        Welch Two Sample t-test

data:  muzi and zeny
t = -0.88537, df = 19.392, p-value = 0.3868
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.536735  3.861411
sample estimates:
mean of x mean of y
 37.07143  39.90909
```

Mann-Whitney test

=neparametrická verze 2-výběrového testu

netestujeme chování parametru μ

testujeme shodu (H_0), resp. rozdílnost (H_1) polohy obou nezávislých výběrů:

```
> muzi=c(23,30,30,32,32,34,39,39,39,41,42,44,46,48)
> zeny=c(26,31,33,35,38,38,44,44,45,51,54)
> wilcox.test(muzi,zeny,paired=F)

        Wilcoxon rank sum test with continuity correction

data:  muzi and zeny
W = 64, p-value = 0.4929
alternative hypothesis: true location shift is not equal to 0
```

ANOVA

Sledujeme (např.):

- **Hmotnost jedinců tří psích plemen;**
- **Délku klasů u čtyř odrůd pšenice;**

Chceme prokázat (typické praktické otázky):

- **je u všech tří plemen hmotnost jedinců srovnatelná, nebo se významně liší?**
- **závisí délka klasu na odrůdě?**

ANOVA

Předpoklady:

- **všechny skupiny jsou nezávislé**
- **sledovaná veličina** *(hmotnost, délka,…atp.)* **se ve všech srovnávaných skupinách chová jako veličina normálně rozdělená, a to se stejnou variabilitou** *(tzv. podmínka homogenity rozptylů)*

ANOVA

Komentáře:

- **jde tedy o zobecnění 2-výběrových testů na případ dvou či více porovnávaných skupin**
- **zajímá nás vlastně, zda chování sledované normálně rozdělené veličiny *Y*** *(plat respondenta, resp. délka výlisku,…)* **závisí na příslušnosti do té či oné kategorie**

ANEB zda *Y* závisí na tzv. kategoriálním faktoru (na vzdělání, na typu stroje,…), odtud označení „jednofaktorová ANOVA“

ANOVA

Značení:

r …počet rozlišovaných kategorií u daného faktoru ($r \geq 2$)

μ_i …střední hodnota *Y* v i-té kategorii (i=1…r)

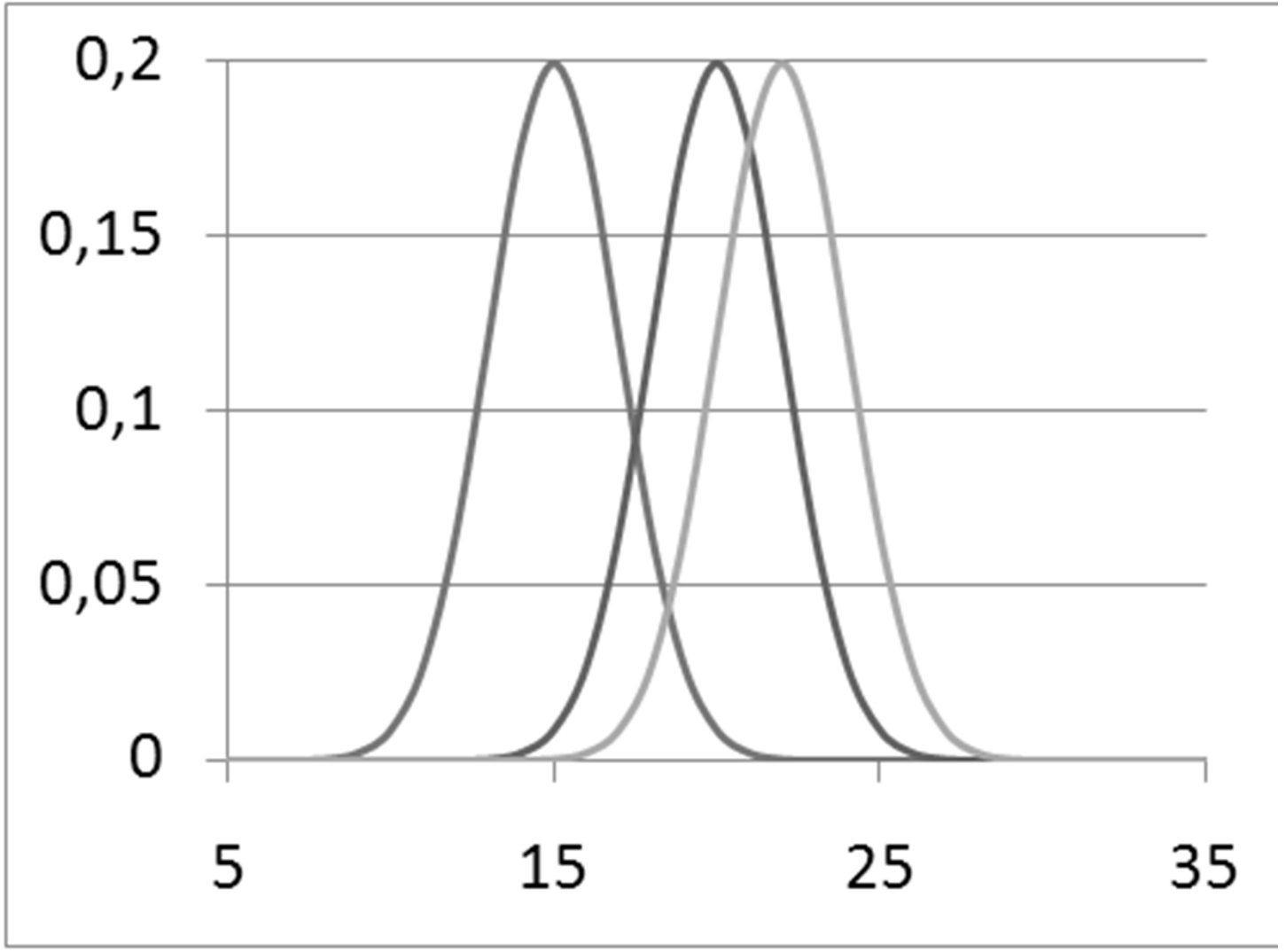
Testujeme:

H_0: $\mu_1 = \ldots = \mu_r$ ANEB nezávislost na faktoru

H_1: non **H_0** **ANEB závislost na faktoru**

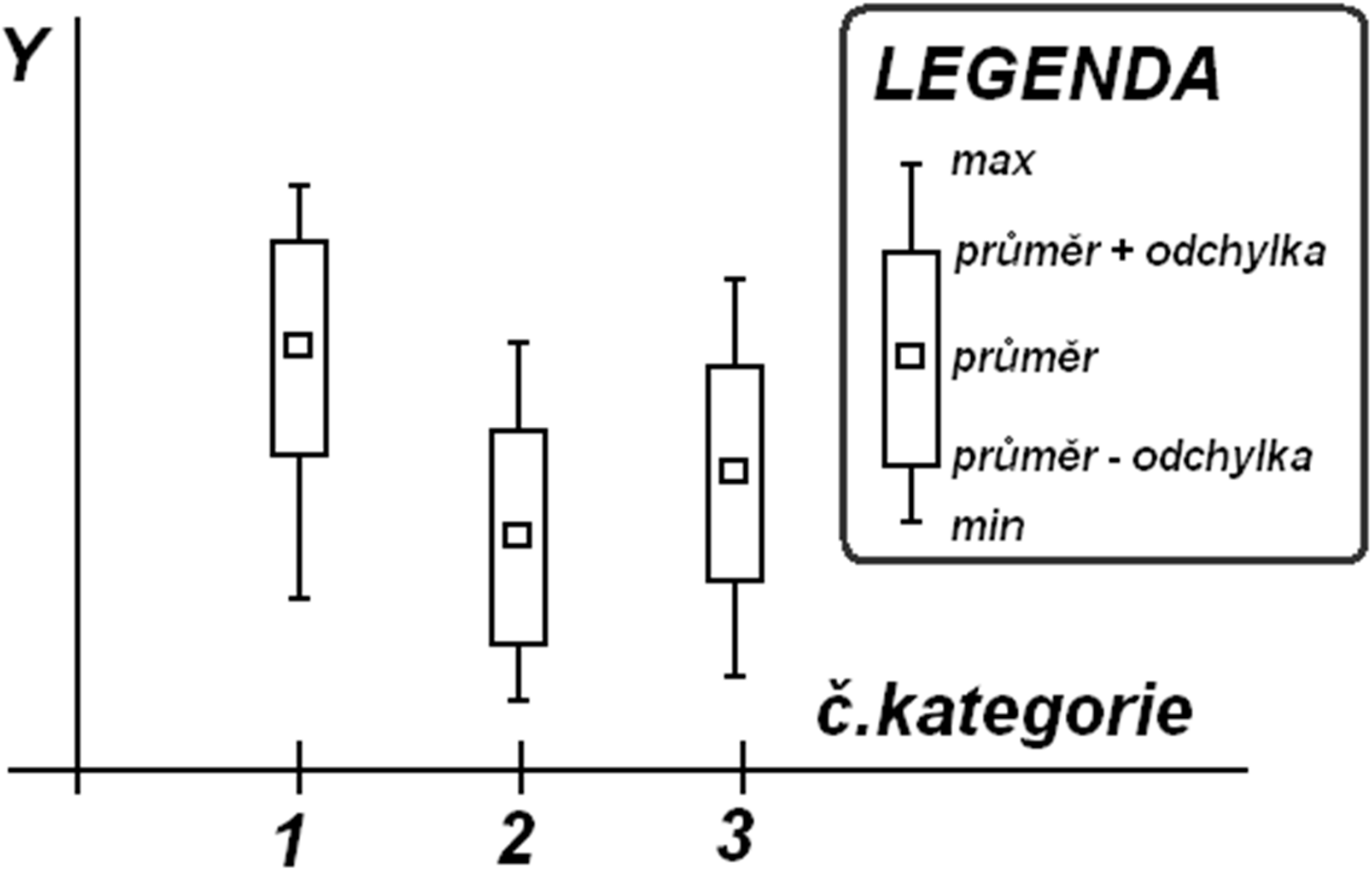
ANOVA

Gaussovy křivky (3-kategoriální faktor):



ANOVA

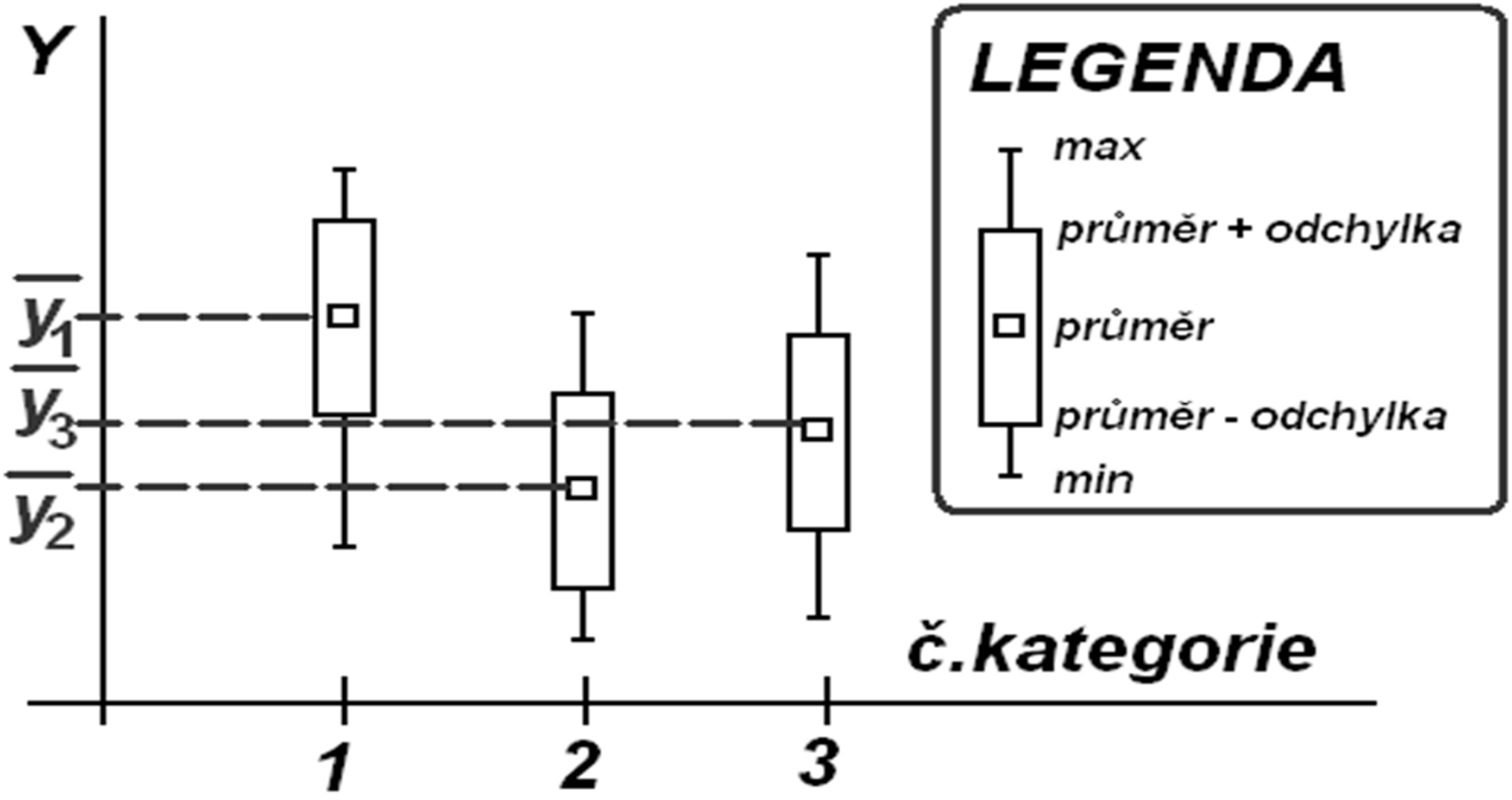
Data (3-kategoriální faktor):



ANOVA

Značení:

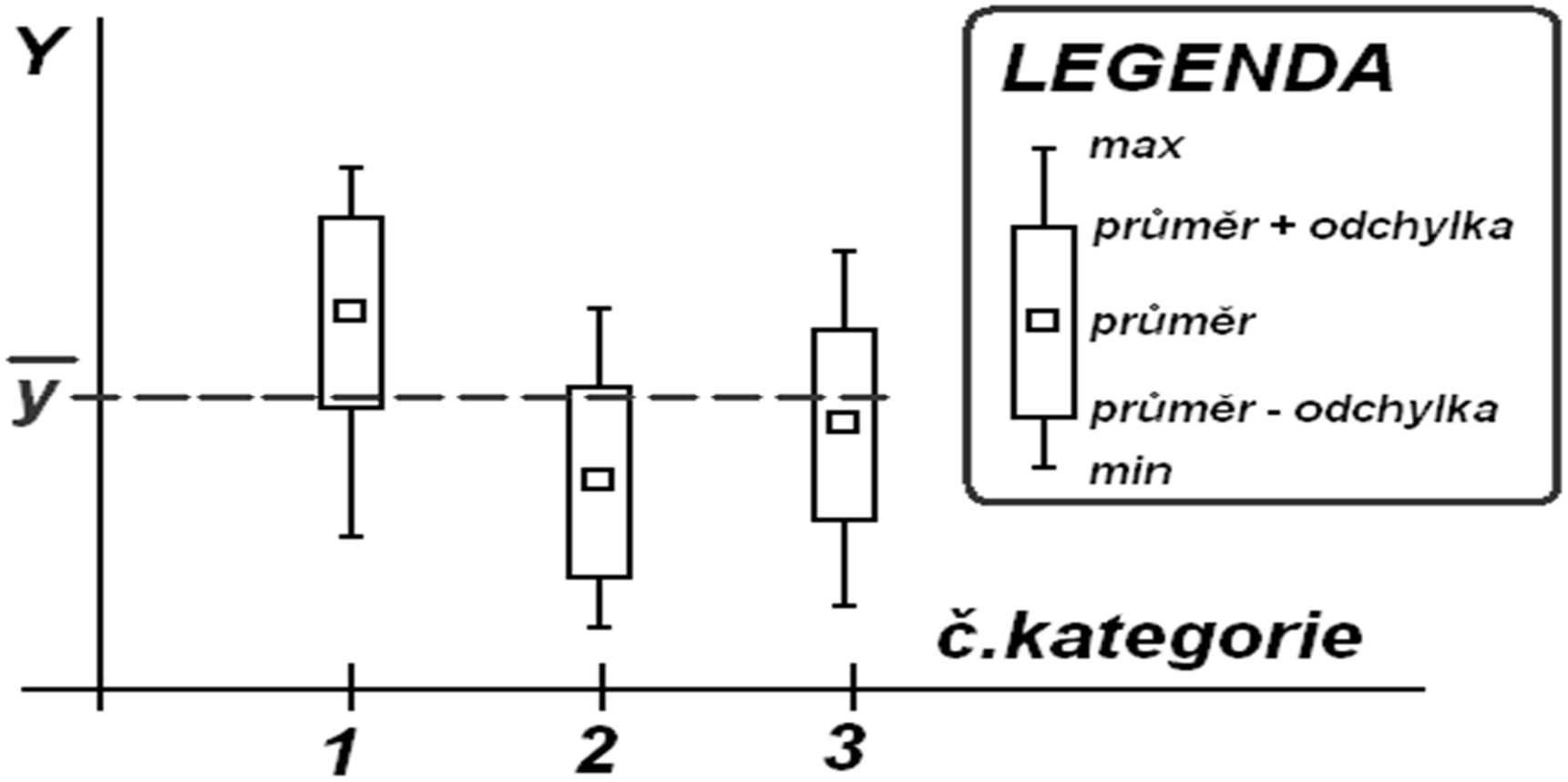
- **„podmíněné“ průměry (po kategoriích)**



ANOVA

Značení:

- **celkový průměr (všech skupin)**



ANOVA

Výpočty:

- **„součty čtverců“Q_{TOT} , Q_m , Q_v**
- **viz přehled vzorců:**

$H_0: \mu_1 = \ldots = \mu_r \qquad T = \dfrac{Q_m/(r-1)}{Q_v/(n-r)}$

$Q_{TOT} = s^2 \cdot (n-1) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = Q_v + Q_m$

$Q_m = \sum_{i=1}^{r} (\bar{y}_i - \bar{y})^2 \cdot n_i$

$Q_v = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$

$W = \langle F_{1-\alpha}(r-1, n-r); \infty)$

ANOVA

Výsledky – tabulka ANOVY:

hodnota	součet čtverců	stupně volnosti	podíl
mezi-skupinová	Q_m	r–1	$Q_m/(r–1)$
vnitro-skupinová	Q_v	n–r	$Q_v/(n–r)$
suma	Q_{TOT}	n–1	

ANOVA

Příklad: **Byly sledovány výnosy čtyř odrůd brambor (označme odrůdy A, B, C, D). Každá odrůda byla pěstována na sedmi srovnatelných polích. Zjistěte, zda je typ odrůdy faktorem, který ovlivňuje hektarový výnos brambor.**

Data – jednotlivé výnosy (v Excelu):

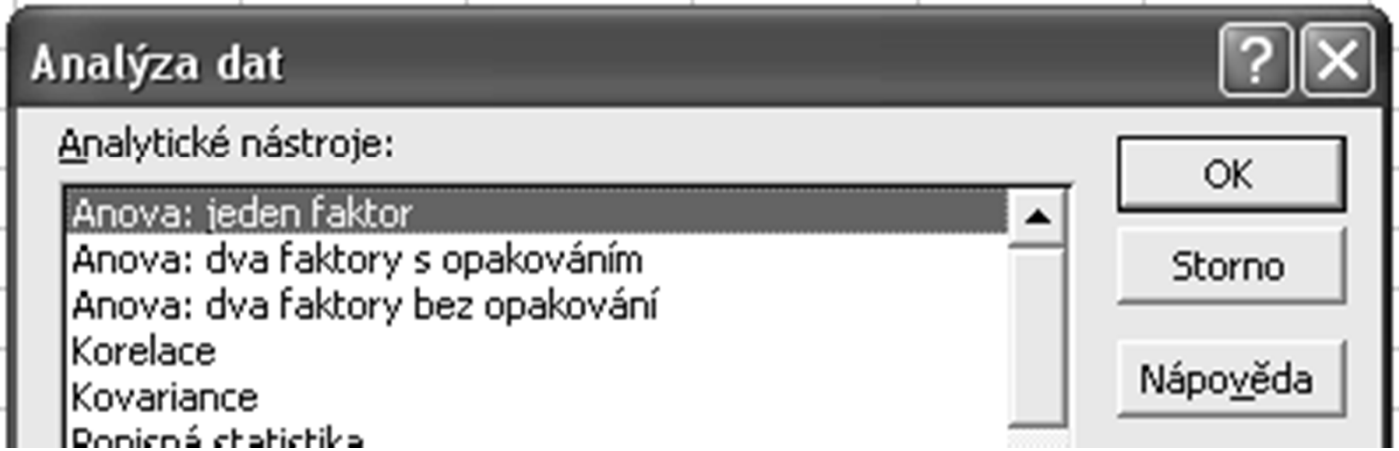
	A	B	C	D	E	F	G	H
1	**A**	**19,3**	**18,0**	**21,6**	**22,4**	**20,9**	**20,1**	**24,0**
2	**B**	**23,1**	**26,5**	**25,2**	**25,0**	**24,3**	**21,4**	**26,7**
3	**C**	**23,7**	**20,8**	**19,8**	**24,1**	**22,2**	**22,6**	**22,9**
4	**D**	**17,2**	**16,6**	**16,9**	**17,7**	**21,3**	**15,2**	**19,0**

H_0: nezávislost výnosů na odrůdě

H_1: závislost (aneb výnosy se významně liší)

ANOVA

	A	B	C	D	E	F	G	H
1	**A**	**19,3**	**18,0**	**21,6**	**22,4**	**20,9**	**20,1**	**24,0**
2	**B**	**23,1**	**26,5**	**25,2**	**25,0**	**24,3**	**21,4**	**26,7**
3	**C**	**23,7**	**20,8**	**19,8**	**24,1**	**22,2**	**22,6**	**22,9**
4	**D**	**17,2**	**16,6**	**16,9**	**17,7**	**21,3**	**15,2**	**19,0**
5								
6								
7								
8								
9								
10								
11								
12								

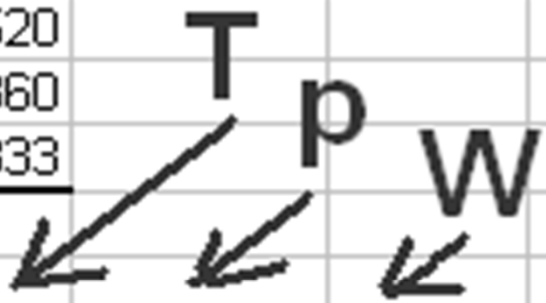


ANOVA


A
B
C
D
19,3
18,0
21,6
22,4
20,9
20,1
24,0
6,7
2,9
9,0
Anova: jeden faktor
Vstup
Vstupní oblast:
$A$1:$H$4
Sdružit:
Sloupce
Řádky
Popisky v prvním sloupci
Alfa:
0,05
Možnosti výstupu
Výstupní oblast:
Nový list:
Nový sešit
OK
Storno
Nápověda

ANOVA

	A	B	C	D	E	F	G
1	**Anova: jeden faktor**						
2							
3	Faktor						
4	*Výběr*	*Počet*	*Součet*	*Průměr*	*Rozptyl*		
5	A	7	146,3	20,9	3,993		
6	B	7	172,2	24,6	3,520		
7	C	7	156,1	22,3	2,360		
8	D	7	123,9	17,7	3,833		
9							
10	ANOVA						
11	*Zdroj variability*	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
12	Mezi výběry	174,912	3	58,304	**17,015**	**3,87E-06**	3,009
13	Všechny výběry	82,240	24	3,427			
14	Celkem	257,153	27				



$p = 3{,}87 \cdot 10^{-6} = 4 \cdot 10^{-6} = 0{,}000\ 004$

ANOVA

Výsledek:

p=0,000 004 < 0,05 $\Rightarrow$ zamítáme H_0

Data prokázala, že výnosy jednotlivých 4 odrůd se významně liší

ANEB

že odrůda je faktorem, na němž výnos významně závisí.

ANOVA

Poznámky:

a) ANOVA = zkratka z „analysis of variance“

b) DF = zkratka z „degrees of freedom“
(stupně volnosti)

c) pokud r=2, lze závislost na faktoru porovnat jak ANOVOU, tak 2-výběrovými testy *(ty mají oproti ANOVĚ tu výhodu, že existují i v jednostranných variantách a lze tudíž případně posoudit, která z obou porovnávaných kategorií je „lepší“)*

ANOVA

Poznámky:

d) Sice jsme prokázali významné rozdíly ve výnosech, můžeme dokonce porovnat zjištěné podmíněné průměry (viz Excel), ale NELZE hned tvrdit, že ANOVA prokázala, která odrůda je horší či lepší – zatím víme jen, že „jsou významné rozdílnosti“

ANOVA – post hoc testy

ANOVA detekuje rozdílnost $\Rightarrow$ zjišťujeme dvou-výběrovými porovnáváními jednotlivých kategorií, které dvě se navzájem významně odlišují … tzv. post-hoc testy

Bonferroniho korekce: $\alpha^* = \alpha/m$

(m … počet vzájemných porovnání)

Kruskal-Wallis test

= neparametrická obdoba ANOVA testu

testujeme shodu (H_0), resp. rozdílnost (H_1) polohy všech nezávislých výběrů

R: kruskal.test(list(data1,data2,…,datar))

nebo

kruskal.test(datakomplet~faktorkomplet)

Test normality

Jak posoudit, kdy postupovat ne/parametricky?

Shapiro-Wilkův test normality

H_0: shoda dat s normálním modelem

R: shapiro.test(data)

TEST χ^2 DOBRÉ SHODY

Sledujeme kategoriální veličinu X (např.):

- **pohlaví (zastoupení samců a samic);**
- **kvalita výrobku (I.jakost, II.jakost, zmetek);**

Chceme prokázat:

- **jsou obě pohlaví zastoupena rovnoměrně, tedy v poměru 1:1 (50:50 %)?**
- **jsou výrobky dle jakosti zastoupeny v poměru 3:1:1 (60:20:20 %)?**

TEST χ^2 DOBRÉ SHODY

Testovaná dvojice hypotéz
- obecně pro veličinu X s r-kategoriemi:

- **H_0: $P(x_1) = \pi_1$; $P(x_2) = \pi_2$;…; $P(x_r) = \pi_r$**
- **H_1: non H_0**

kde π_1,…,π_r jsou konkr. čísla: π_1+…+π_r = 1

TEST χ^2 DOBRÉ SHODY

Testovaná dvojice hypotéz konkr.:

- u příkladu samci vs samice:

- **H_0: $P(x_1) = 0,5$; $P(x_2) = 0,5$**

- u příkladu s jakostí:

- **H_0: $P(x_1) = 0,6$; $P(x_2) = 0,2$; $P(x_3) = 0,2$**

TEST χ^2 DOBRÉ SHODY

Z dat určíme absolutní, tzv.

pozorované četnosti $n_1; n_2; \ldots; n_r$

přičemž $n_1+\ldots+n_r = n$

Pro jednotlivé kategorie spočteme tzv.

očekávané četnosti $o_1; o_2; \ldots; o_r$

a to podle vzorce:

$$o_i = n \cdot \pi_i \qquad (i=1,\ldots,r)$$

TEST χ^2 DOBRÉ SHODY

Např. nechť při kontrole jakosti bylo

110 výrobků I. jakosti (n_1= 110),

56 výrobků II. jakosti (n_2= 56)

34 zmetků (n_3= 34), tj. n= 200;

při testu, zda π_1=0,6; π_2=0,2; π_3=0,2

dostaneme jaké očekávané četnosti?

$o_1 = n \cdot \pi_1 = 200 \cdot 0{,}6 = 120$

$o_2 = n \cdot \pi_2 = 200 \cdot 0{,}2 = 40$

$o_3 = n \cdot \pi_3 = 200 \cdot 0{,}2 = 40$

TEST χ^2 DOBRÉ SHODY

Podstatou testové statistiky je porovnání četností pozorovaných s očekávanými:

$$T = \sum (n_i - o_i)^2 / o_i \qquad (i=1\ldots r)$$

Př. (pokrač.):

$$T = (110-120)^2/120 +$$
$$+ (56-40)^2/40 +$$
$$+ (34-40)^2/40 =$$
$$= 0{,}83 + 6{,}40 + 0{,}90 = 8{,}13$$

TEST χ^2 DOBRÉ SHODY

Řešení pomocí Excelu:

	A	B	C	D
1		π_i	n_i	o_i
2		0,6	110	120
3		0,2	56	40
4		0,2	34	40
5	suma	1	200	200
6				
7	=chitest(C2:C4;D2:D4)			
8				

CHITEST(aktuální; očekávané)

p=0,017

…tj.?

TEST χ^2 DOBRÉ SHODY

Poznámka:

V principu lze takto testovat i normalitu;

Sledovanou číselnou veličinu „rozsekáme“ intervalovým rozdělením do kategorií a porovnáme pozorované četnosti v těchto kategoriích s očekávanými, které by odpovídaly pravděpodobnostem dle „gaussovského“ modelu

TEST χ^2 NEZÁVISLOSTI

Sledujeme dvojici kategoriálních veličin X,Y

- **např. u každého respondenta jeho pohlaví (M-Ž) a krevní skupinu (A-B-AB-0);**
- **nebo u každého výrobku jeho kvalitu (I.jakost, II.jakost, zmetek) a to, během jaké směny vznikl (dopolední – odpolední - noční směna);**

TEST χ^2 NEZÁVISLOSTI

Chceme prokázat:

- **závisí nebo nezávisí krevní skupina na pohlaví?**

(ve smyslu, zda jsou nebo nejsou mezi muži a ženami významné rozdíly v zastoupení jednotlivých krevních skupin)

TEST χ^2 NEZÁVISLOSTI

Resp. (příklad s výrobky) chceme prokázat:

- **<u>závisí nebo nezávisí</u> kvalita výrobku na tom, během jaké směny vznikl?**

(ve smyslu, zda jsou nebo nejsou mezi jednotlivými směnami významné <u>rozdíly</u> v zastoupení jednotlivých kvalitativních kategorií)

TEST χ^2 NEZÁVISLOSTI

Testovaná dvojice hypotéz:

H_0: nezávislost (mezi X a Y)

H_1: non H_0 (tj. závislost mezi X a Y)

TEST χ^2 NEZÁVISLOSTI

Data:

	A	B	C
1	výrobek č.	jakost	směna
2	1	I	d
3	2	I	n
4	3	z	d
5	4	II	o
6	5	I	o
7		...	
81	80	II	n

TEST χ^2 NEZÁVISLOSTI

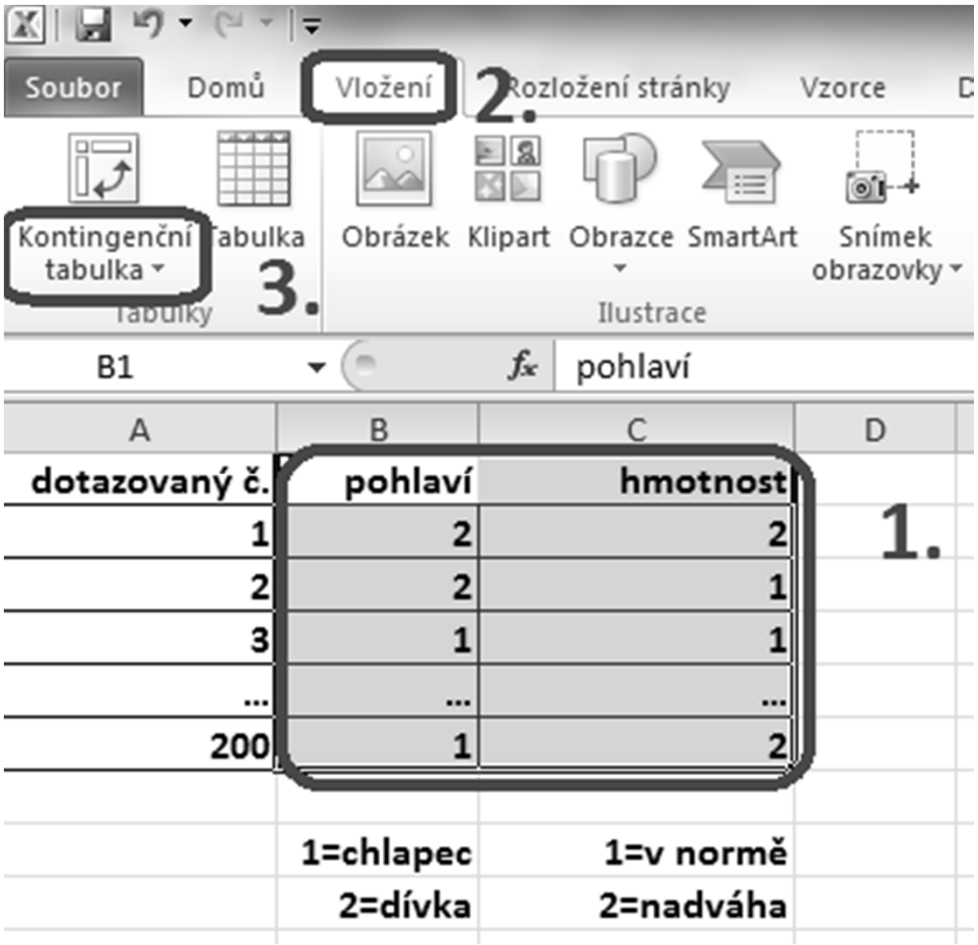
Data přehledně – kontingenční tabulka pozorovaných absolutních četností:

Pozorované četnosti	$Y = y_1$	$Y = y_2$	...	$Y = y_s$	suma
$X = x_1$	n_{11}	n_{12}	...	n_{1s}	$n_{1\bullet}$
$X = x_2$	n_{21}	n_{22}	...	n_{2s}	$n_{2\bullet}$
...	...	...	...	...	...
$X = x_r$	n_{r1}	n_{r2}	...	n_{rs}	$n_{r\bullet}$
suma	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet s}$	n

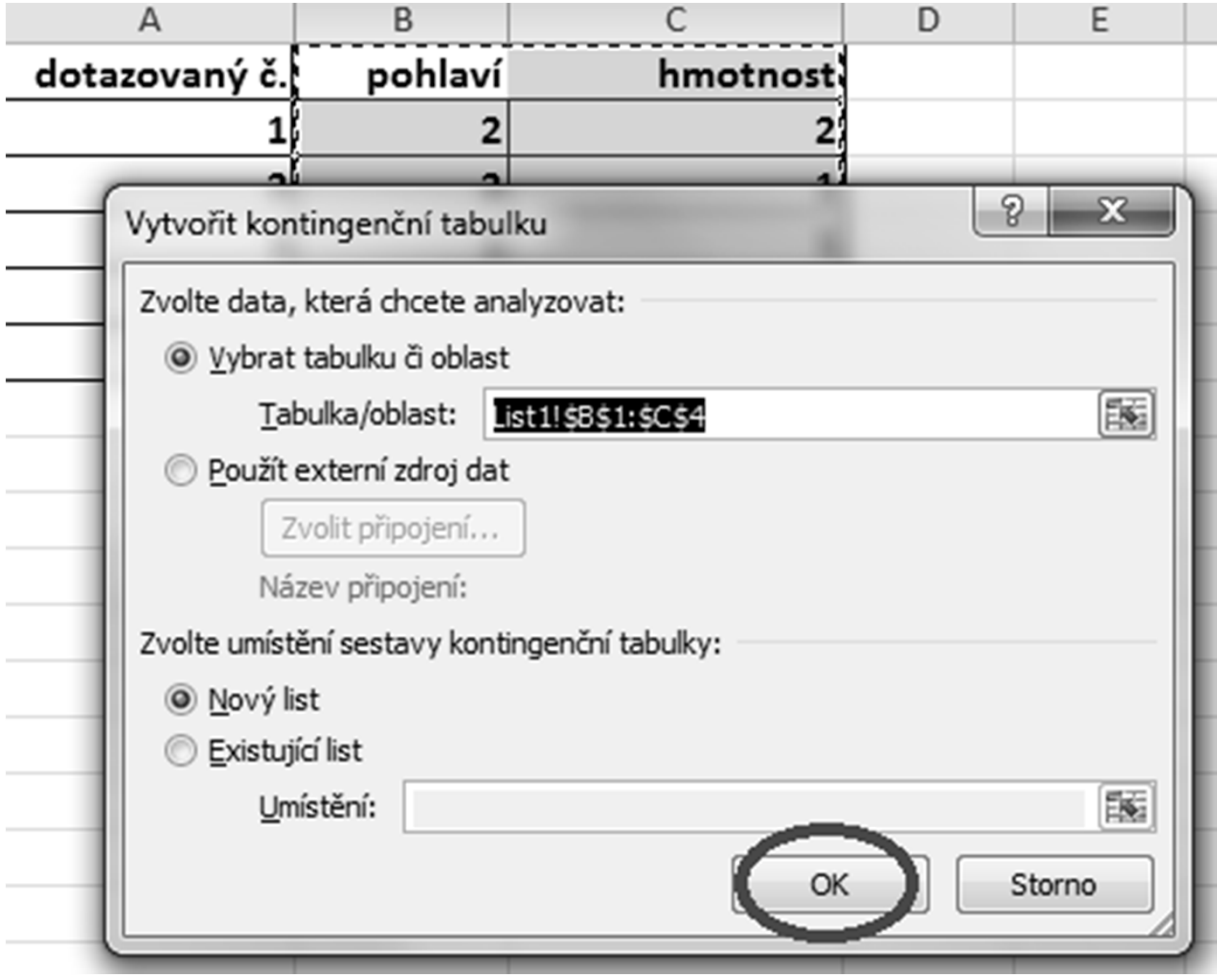
r = počet „řádkových“ kategorií

s = počet „sloupcových“ kategorií

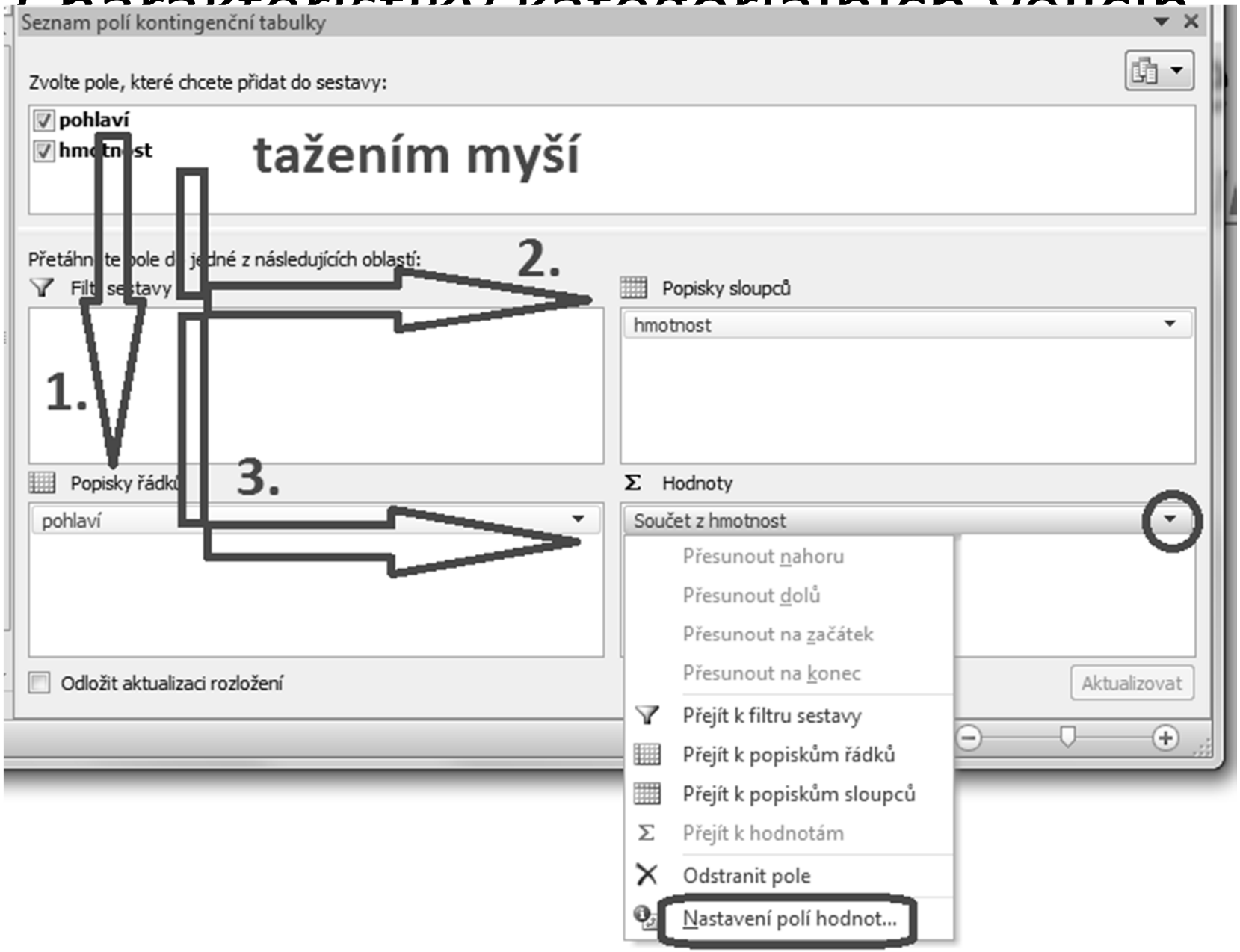
VSUVKA: KONTINGENČNÍ TABULKA (vytvoření)



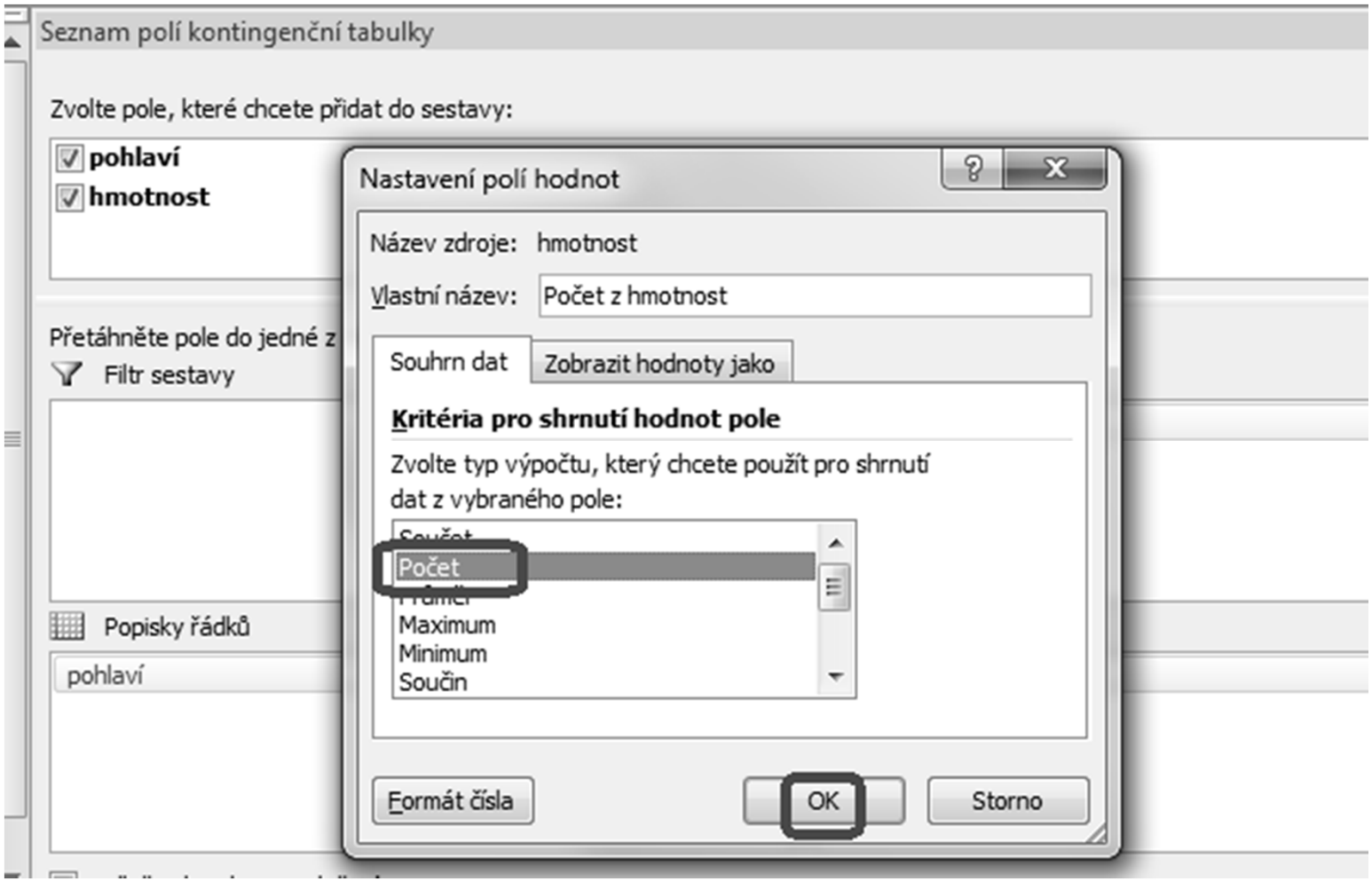
VSUVKA: KONTINGENČNÍ TABULKA (vytvoření)



Charakteristiky kategoriálních veličin



VSUVKA: KONTINGENČNÍ TABULKA (vytvoření)



TEST χ^2 NEZÁVISLOSTI

Kontingenční tabulka - příklad:

	dopol	odpol	noc	suma
I. jakost	12	15	11	38
II. jakost	7	9	11	27
zmetky	4	5	6	15
suma	23	29	28	80

např. $n_{12} = 15$ $n_{21} = 7$

$n_{1\cdot} = 38$ $n_{\cdot 1} = 23$

TEST χ^2 NEZÁVISLOSTI

Připravíme tabulku <u>očekávaných</u> četností:

Očekávané četnosti	$Y = y_1$	$Y = y_2$	...	$Y = y_s$	suma
$X = x_1$	o_{11}	o_{12}	...	o_{1s}	$n_{1\bullet}$
$X = x_2$	o_{21}	o_{22}	...	o_{2s}	$n_{2\bullet}$
...	...	...	...	...	...
$X = x_r$	o_{r1}	o_{r2}	...	o_{rs}	$n_{r\bullet}$
suma	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet s}$	n

$o_{ij} = n_{i\bullet} \cdot n_{\bullet j} / n$

např. $o_{12} = n_{1\bullet} \cdot n_{\bullet 2} / n$

TEST χ^2 NEZÁVISLOSTI

Očekávané četnosti – příklad (pokrač.):

	dopol	odpol	noc	suma
I. jakost	10,9	13,8	13,3	38
II. jakost	7,8	9,8	9,5	27
zmetky	4,3	5,4	5,3	15
suma	23	29	28	80

např. $o_{12} = n_{1.} \cdot n_{.2} / n = 38 \cdot 29 / 80 = 13{,}8$

! součty stejné jako původně (až na zaokr.)!

TEST χ^2 NEZÁVISLOSTI

Řešení pomocí Excelu:

	A	B	C	D	E
1		dopol	odpol	noc	suma
2	I. jakost	12	15	11	38
3	II. jakost	7	9	11	27
4	zmetky	4	5	6	15
5	suma	23	29	28	80
6					
7		dopol	odpol	noc	suma
8	I. jakost	10,9250	13,7750	13,3000	38
9	II. jakost	7,7625	9,7875	9,4500	27
10	zmetky	4,3125	5,4375	5,2500	15
11	suma	23	29	28	80
12					
13		=chitest(B2:D4;B8:D10)			

p=0,883…tj.?

TEST χ^2 NEZÁVISLOSTI

Podmínka použití:

- ***pozor – u obou typů testu (dobré shody i nezávislosti) musí být všechny kategorie dostatečně zastoupeny, aneb všechny očekávané četnosti mají být aspoň 5;***

- ***není-li splněno, doporučuje se sloučit některé (obvykle sousední) kategorie***

Regrese – jednoduchá regrese

- Cíl jednoduché (simple) regrese: najít model funkční závislosti (spojité) veličiny *Y* na jedné (spojité) veličině (na tzv. regresoru) *X*
- model lineární $Y=\beta_0+\beta_1 X$

  kvadratický $Y=\beta_0+\beta_1 X+\beta_2 X^2$

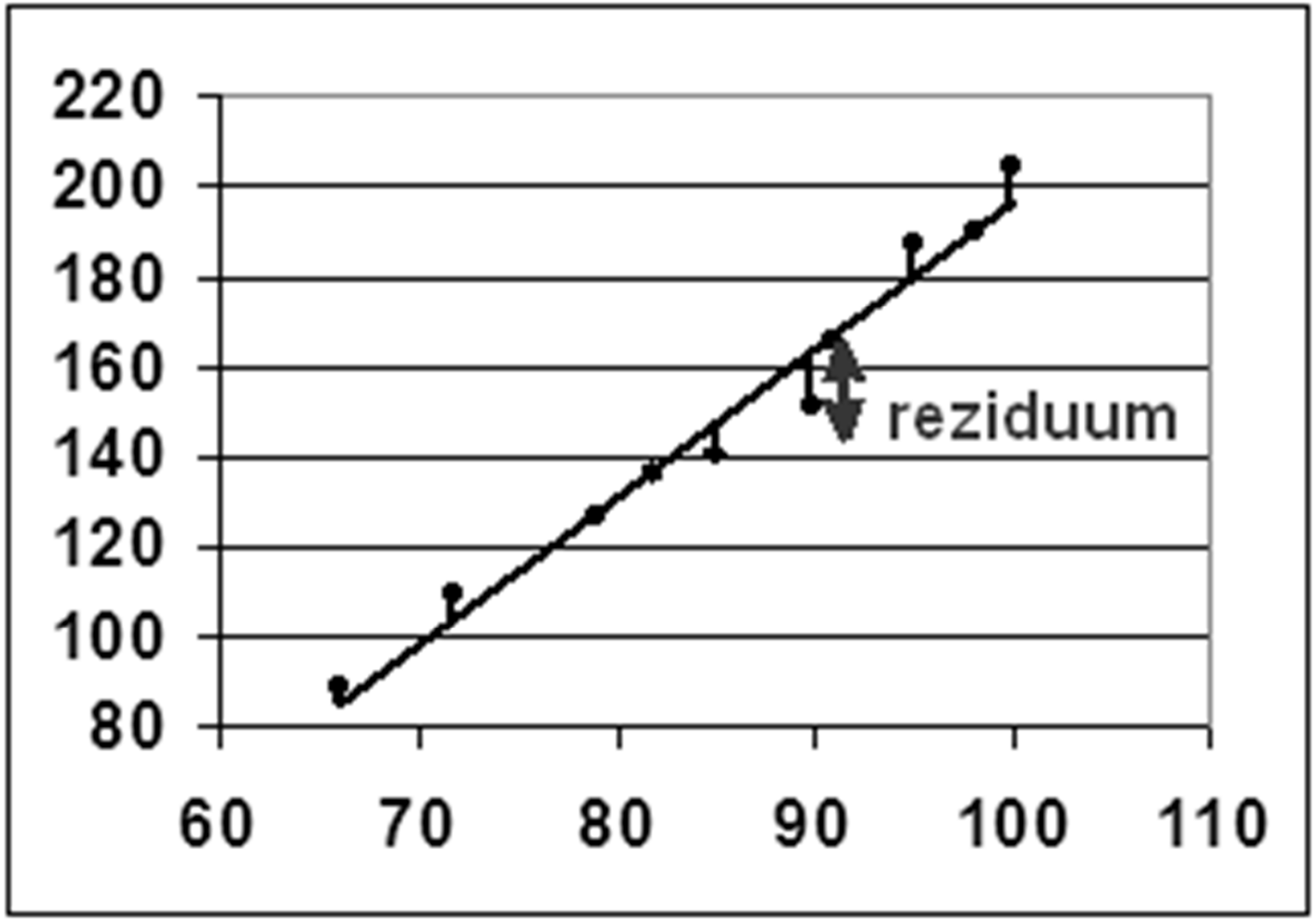
  (*tvar často napoví bodový graf dat*)
- *Příklad: závislost hmotnosti jedince na jeho tělesné výšce*

Jednoduchá regrese

- Značení:

$(x_i ; y_i)$	$i=1,\ldots,n$	data
Y_i	$i=1,\ldots,n$	model
$e_i = y_i - Y_i$	$i=1,\ldots,n$	reziduum

Regrese – bodový graf



Jednoduchá lineární regrese

$y_1 = Y_1 + e_1 = (\beta_0 + \beta_1 \cdot x_1) + e_1$

$y_2 = Y_2 + e_2 = (\beta_0 + \beta_1 \cdot x_2) + e_2$

...

$y_n = Y_n + e_n = (\beta_0 + \beta_1 \cdot x_n) + e_n$

β_0 parametr – prostý člen
(průsečík grafu přímky s o_Y)

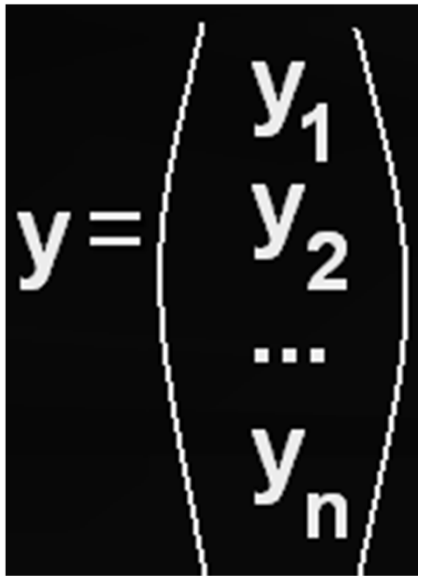
β_1 parametr – lineární člen
(směrnice grafu přímky)

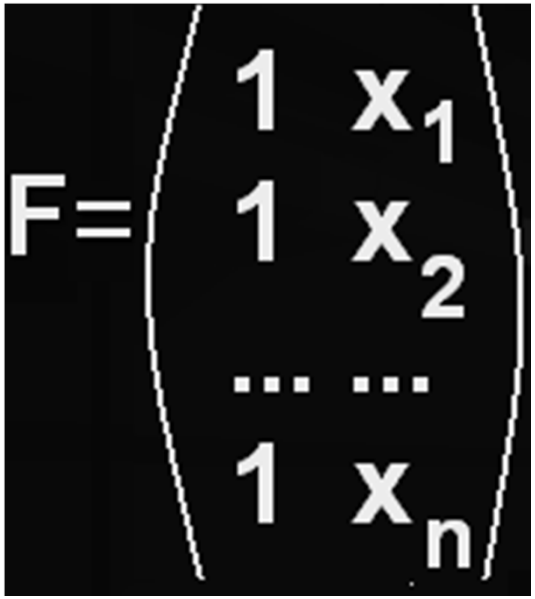
Jednoduchá lineární regrese

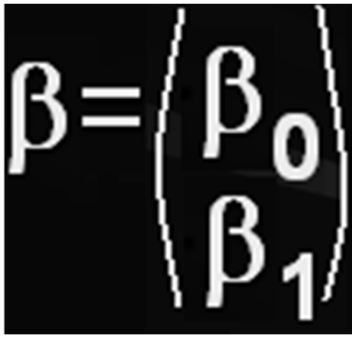
aneb MATICOVĚ:

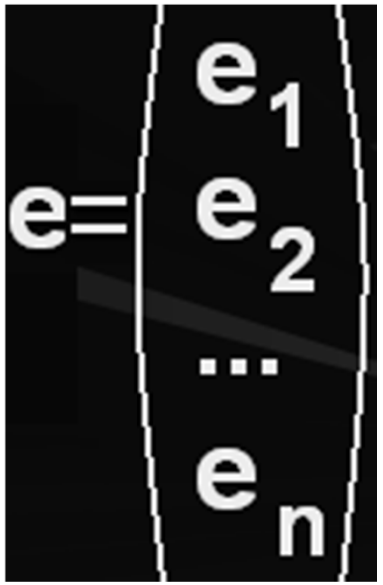
$$\mathbf{y} = \mathbf{F} \cdot \beta + \mathbf{e}$$

kde









Jednoduchá lineární regrese

Nechť $\mathbf{e=0}$, pak: $\mathbf{F \cdot b} = \mathbf{y} \longrightarrow \mathbf{b} = ?$

Pozor: **F** je matice, nelze s ní dělit!

„Trikové“ úpravy (vlastnosti matic):

$$\mathbf{F^T F \cdot b} = \mathbf{F^T y}$$

$$(\mathbf{F^T F})^{-1} \mathbf{F^T F \cdot b} = (\mathbf{F^T F})^{-1} \mathbf{F^T y}$$

$$\underline{\mathbf{b} = (\mathbf{F^T F})^{-1} \mathbf{F^T y}}$$

Regrese (dokonce každý typ)

$$\mathbf{b} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y}$$

u modelu jednoduché lineární regrese:

$$\mathbf{F} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$$

$$\mathbf{F}^T\mathbf{F} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

Regrese (dokonce každý typ)

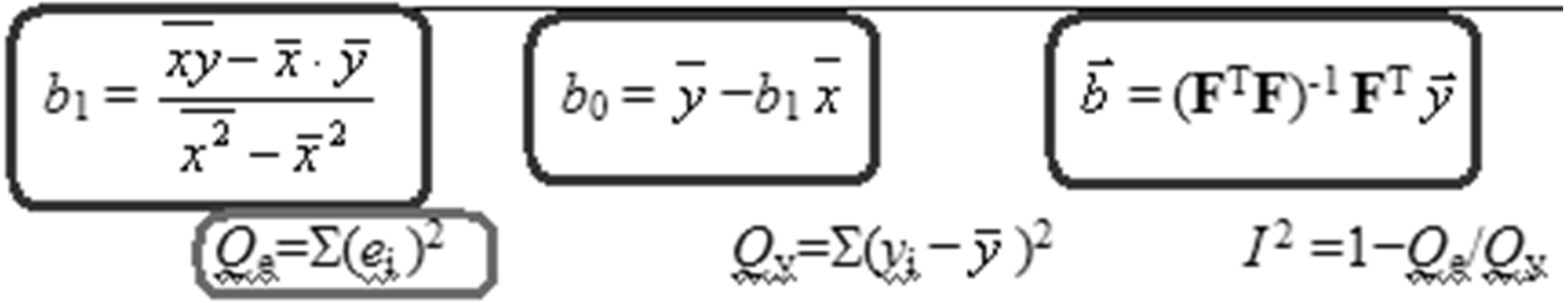
$$\mathbf{b} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y}$$

jde o univerzální (pro každý regresní model!) vzorec odhadu parametrů **b**, modely se liší jen konkr. tvarem **F**;

b = tzv. odhad metodou nejmenších čtverců (MNČ); zaručuje $\min\sum(e_i)^2$

Jednoduchá lineární regrese

- Přehled vzorců (verze s průměry):



k čemu Q_e (součet reziduálních čtverců)?

Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company

(*Neter-Wasserman-Kutner, USA, 1990*)

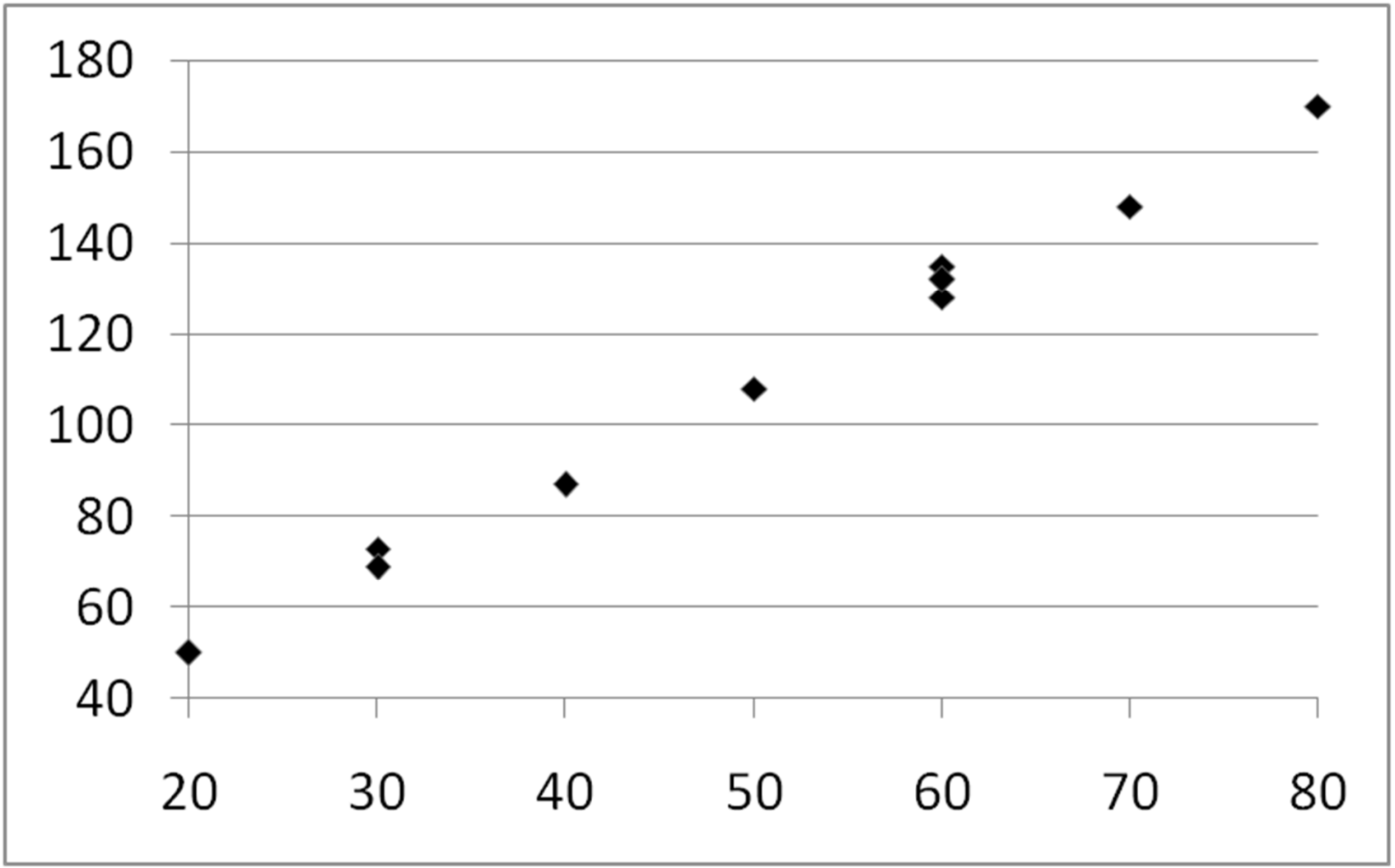
X…velikost staveniště

Y…počet hodin, odpracovaných dělníky

x_i	30	20	60	80	40	50	60	30	70	60
y_i	73	50	128	170	87	108	135	69	148	132

Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company



Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company

$\overline{x}=50$, $\overline{y}=110$, $\overline{x^2}=2840$, $\overline{y^2}=13466$,

$\overline{xy}=(30\cdot 73+\ldots+60\cdot 132)/10=6180$

$b_1=(6180-50\cdot 110)/(2840-50^2)=\underline{\underline{2,0}}$

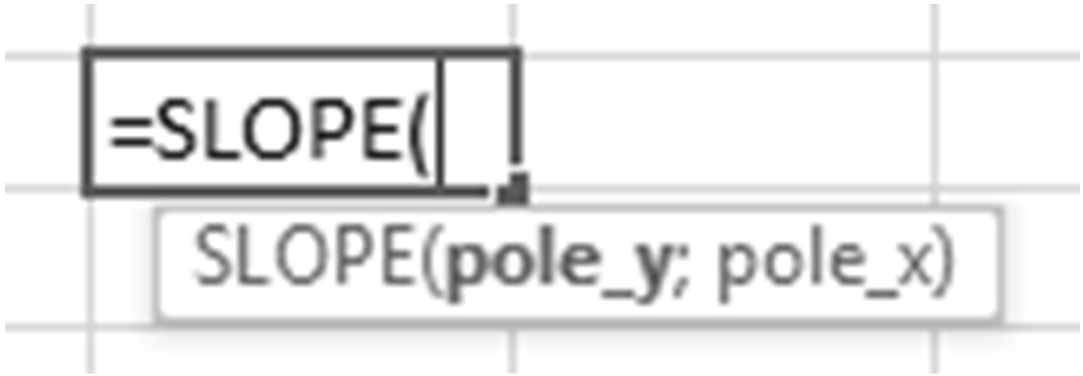
$b_0=110-2\cdot 50=\underline{\underline{10,0}}$ (*ne vždy celočíselně*)

Nalezený model: $\underline{\underline{Y=10+2X}}$

Jednoduchá lineární regrese-příklad

V Excelu:

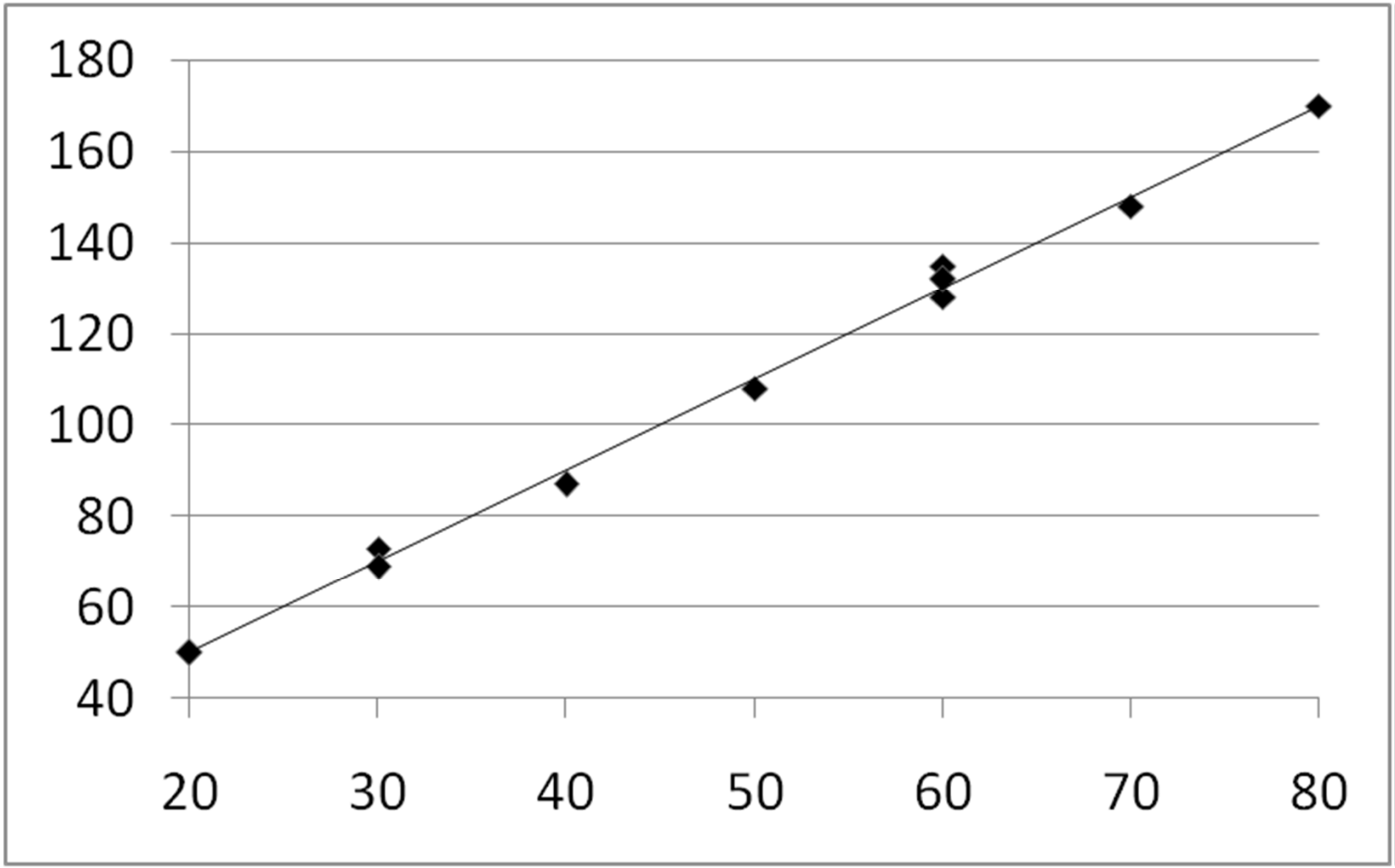
b_1=



b_0=

Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company



Jednoduchá lineární regrese-příklad

Př: Data - Westwood Company

Určete pro nalezený model Q_e:

$Y_1 = 10 + 2 \cdot 30 = 70, \quad e_1 = 73 - 70 = 3$

…

$Y_{10} = 10 + 2 \cdot 60 = 130,$

$e_{10} = 132 - 130 = 2$

$Q_e = 3^2 + 0^2 + (-2)^2 + \ldots + 2^2 = \underline{\underline{60}}$

A k čemu dál využít tuto hodnotu?

Korelovanost je obecně míra lineární závislosti

V každém typu regresního modelu lze určit tzv. index determinace:

$$I^2 = 1 - Q_e/Q_Y$$

kde $Q_Y = \Sigma(y_i - \bar{y})^2$

Korelovanost

Př: Data - Westwood Company

Určete pro nalezený model (pro nějž vyšlo Q_e=60) hodnotu I^2 :

$Q_Y=(73-110)^2+\ldots+(132-110)^2=$

$=1369+\ldots+484 =13660;$

$I^2 = 1-60/13660 = 1-0{,}004 = \underline{0{,}996}$

Korelovanost

Př: Data - Westwood Company

Interpretace: Nalezený model (Y=10+2X) vysvětluje z 99,6 % variabilitu proměnné Y
ANEB
jde o model velmi silné závislosti proměnné Y na proměnné X.

Korelace

- Korelace obecně je míra kvality (vhodnosti, těsnosti) nalezeného regresního modelu pro daná data; vychází z hodnot reziduí
- V každém typu regresního modelu lze použít index determinace I^2

  (0 až 1, resp. 0 % až 100 %);

  vyjadřuje, z kolika % je variabilita závisle proměnné (*Y*) vysvětlena daným modelem

Korelace spec. pro model jednoduché lineární regrese

- Korelační koeficient (verze s průměry):

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\left(\overline{x^2} - \bar{x}^2\right) \cdot \left(\overline{y^2} - \bar{y}^2\right)}}$$

←„*kovariance*“

Korelace spec. pro model jednoduché lineární regrese

- Korelační koeficient
- vždy v rozmezí -1 až +1 (NE % !)
- záporný při “klesající regresní přímce”
- kladný při “rostoucí regresní přímce”
- čím DÁL od 0, tím silnější je lineární závislost („korelovanost“) mezi X a Y
- platí: $r^2 = I^2$

Korelace spec. pro model jednoduché lineární regrese

Př: Data - Westwood Company

$$r = (6180 - 50 \cdot 110) / \sqrt{(2840 - 50^2) \cdot (13466 - 110^2)} =$$

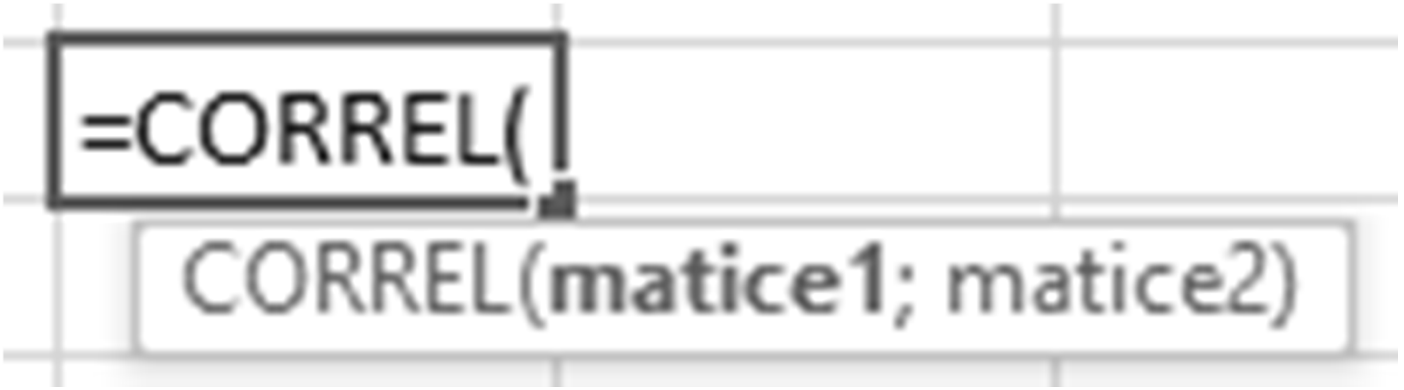
$$= \underline{\underline{0{,}998}} \quad (\textit{platí: } 0{,}998^2 = I^2 = 0{,}996)$$

Silná přímá* lineární závislost počtu prac. hodin na velikosti staveniště.

* *tj. dle „rostoucí přímky“ (nepřímá=?)*

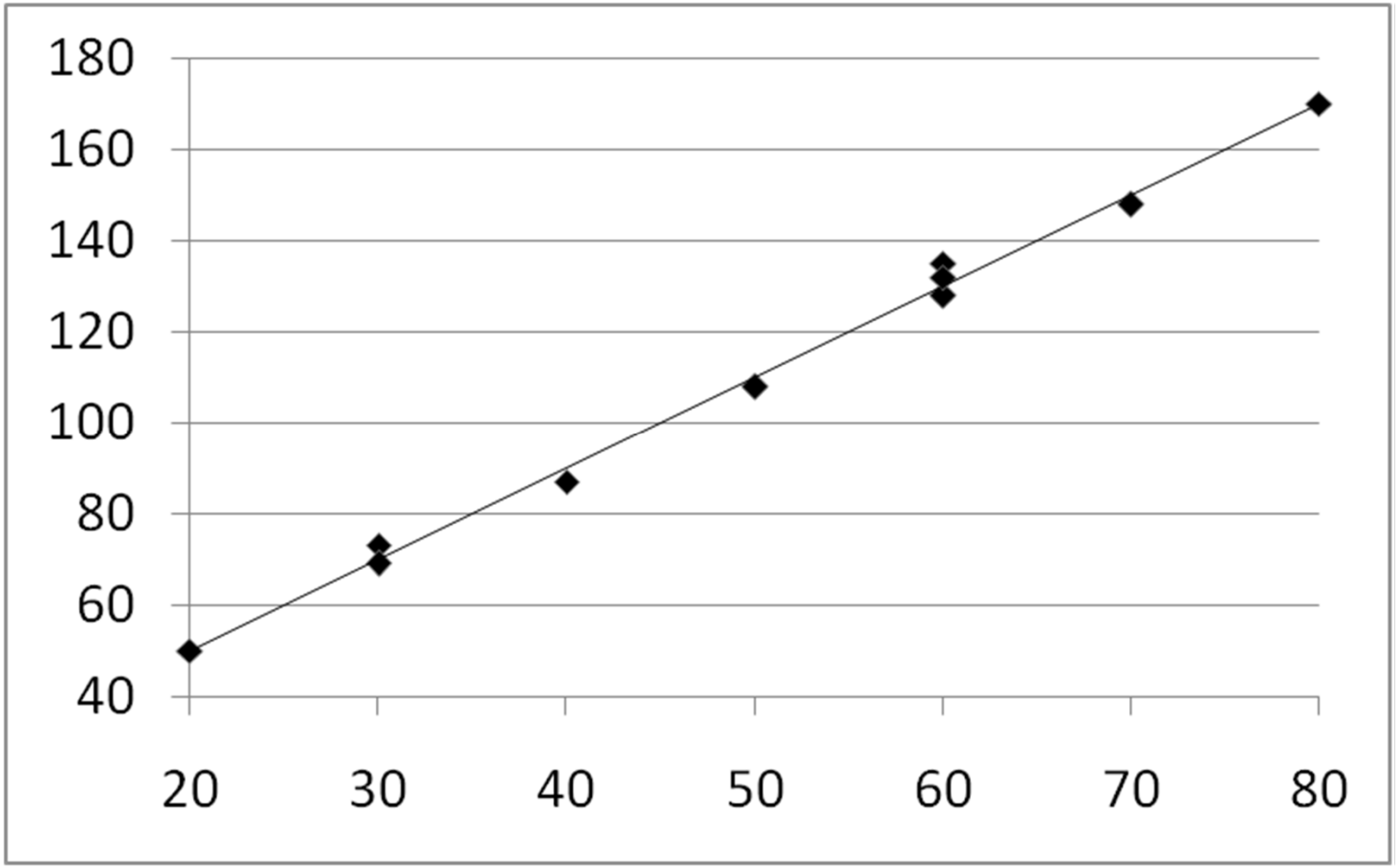
Jednoduchá lineární regrese-příklad

V Excelu:

r = 

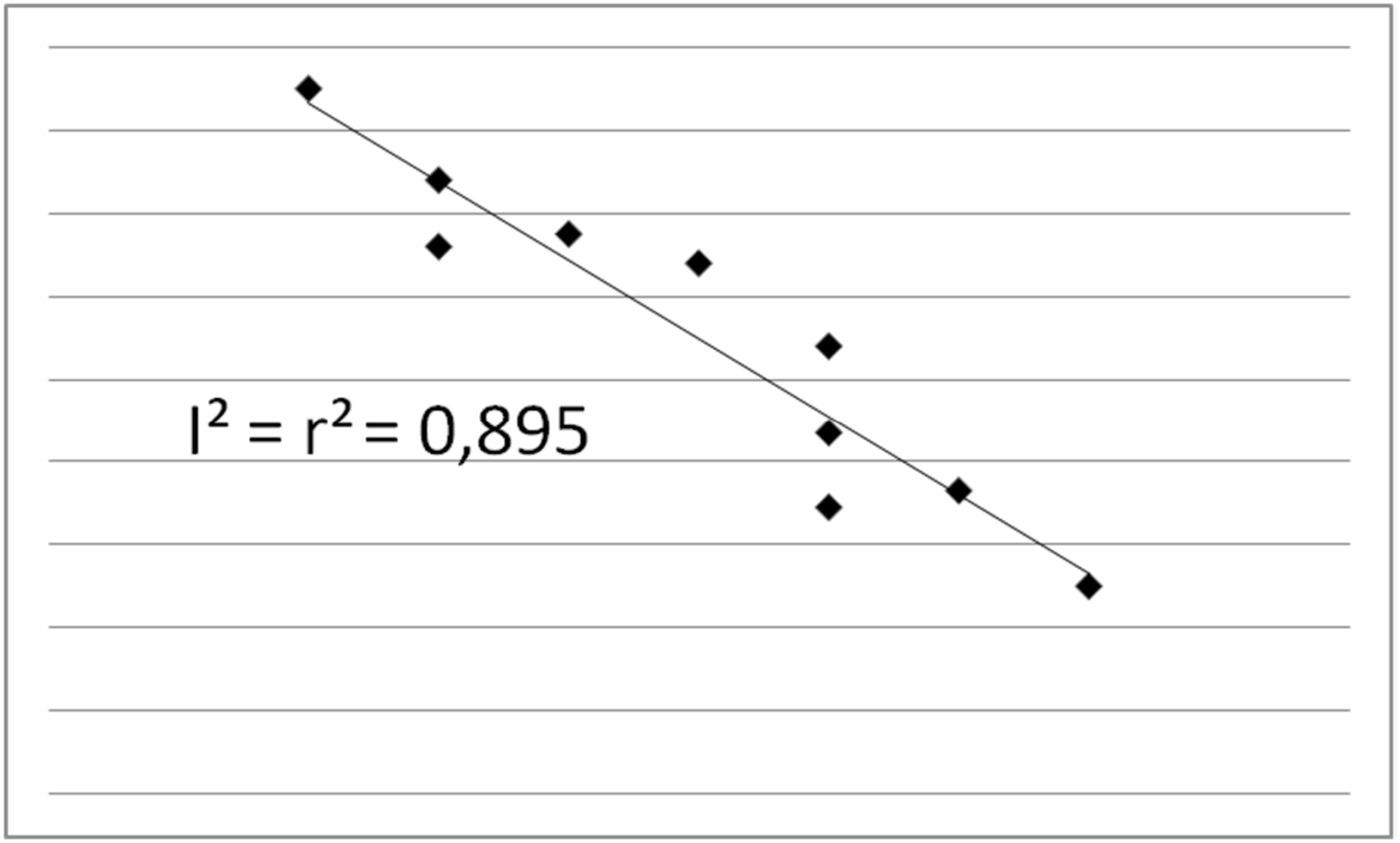
Korelace spec. pro model jednoduché lineární regrese

Př: Data Westwood Company (r=0,998)



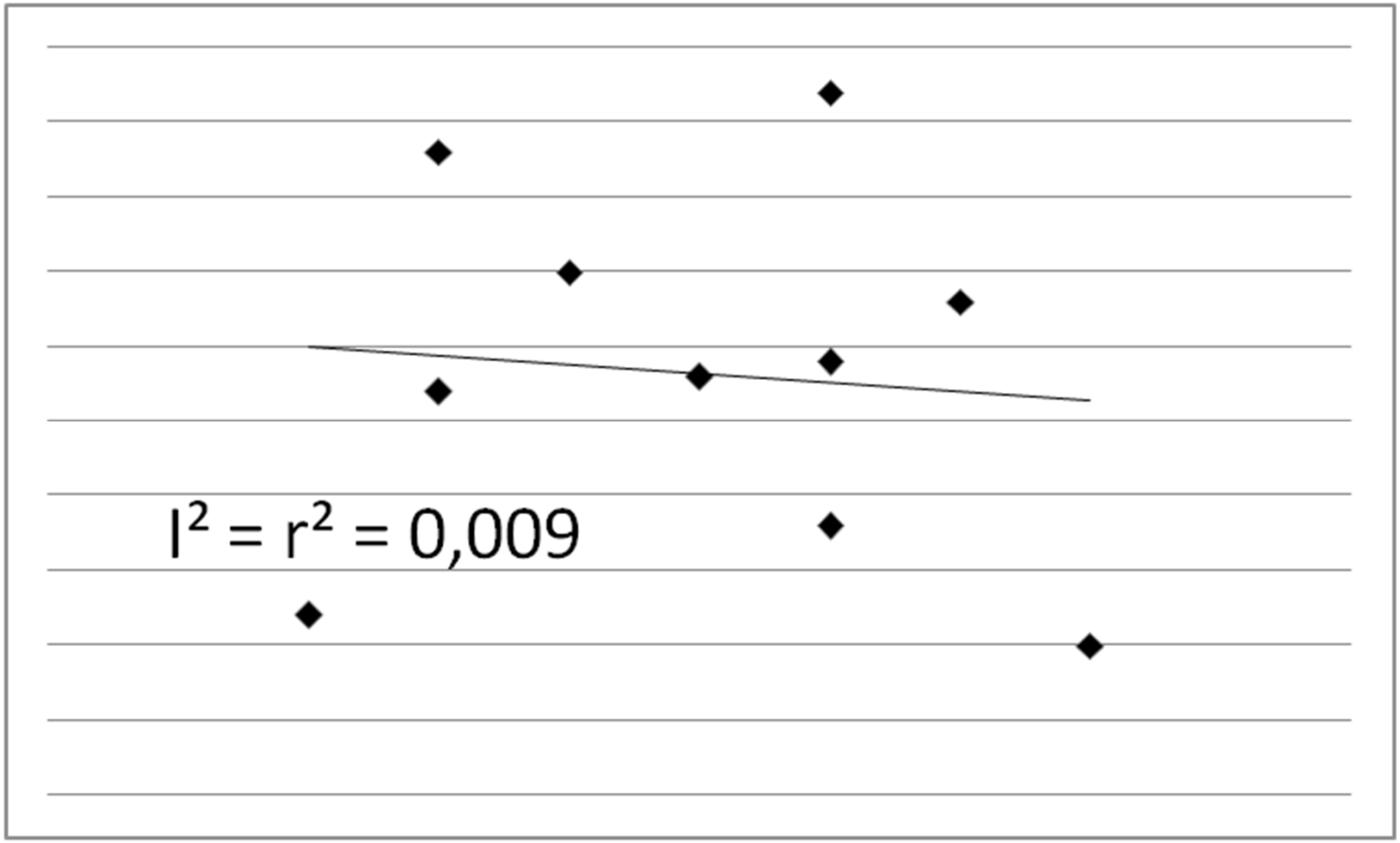
Korelace spec. pro model jednoduché lineární regrese

Př: Jiná data (r = -0,946)



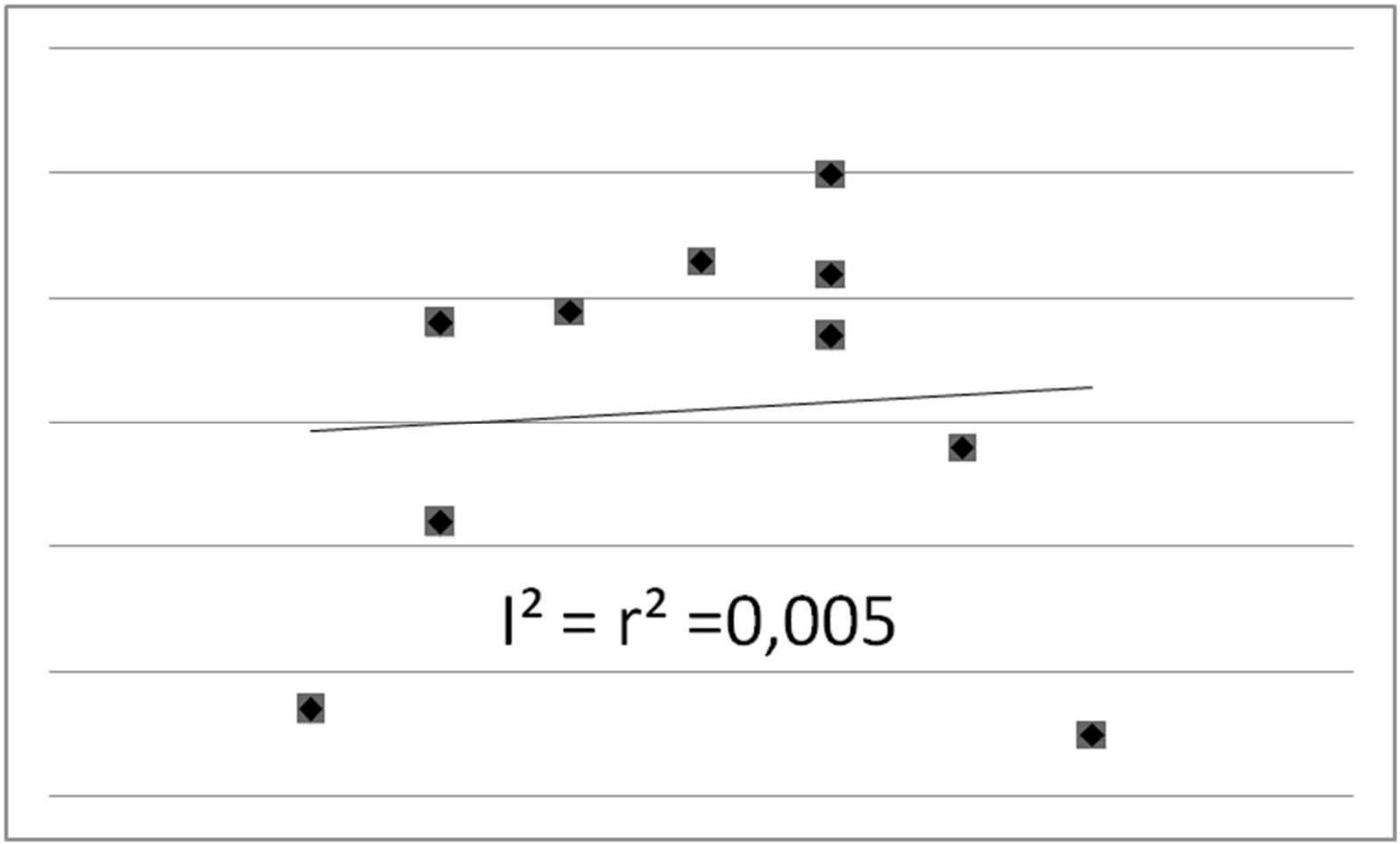
Korelace spec. pro model jednoduché lineární regrese

Př: Jiná data (r = -0,098)



Korelace spec. pro model jednoduché lineární regrese

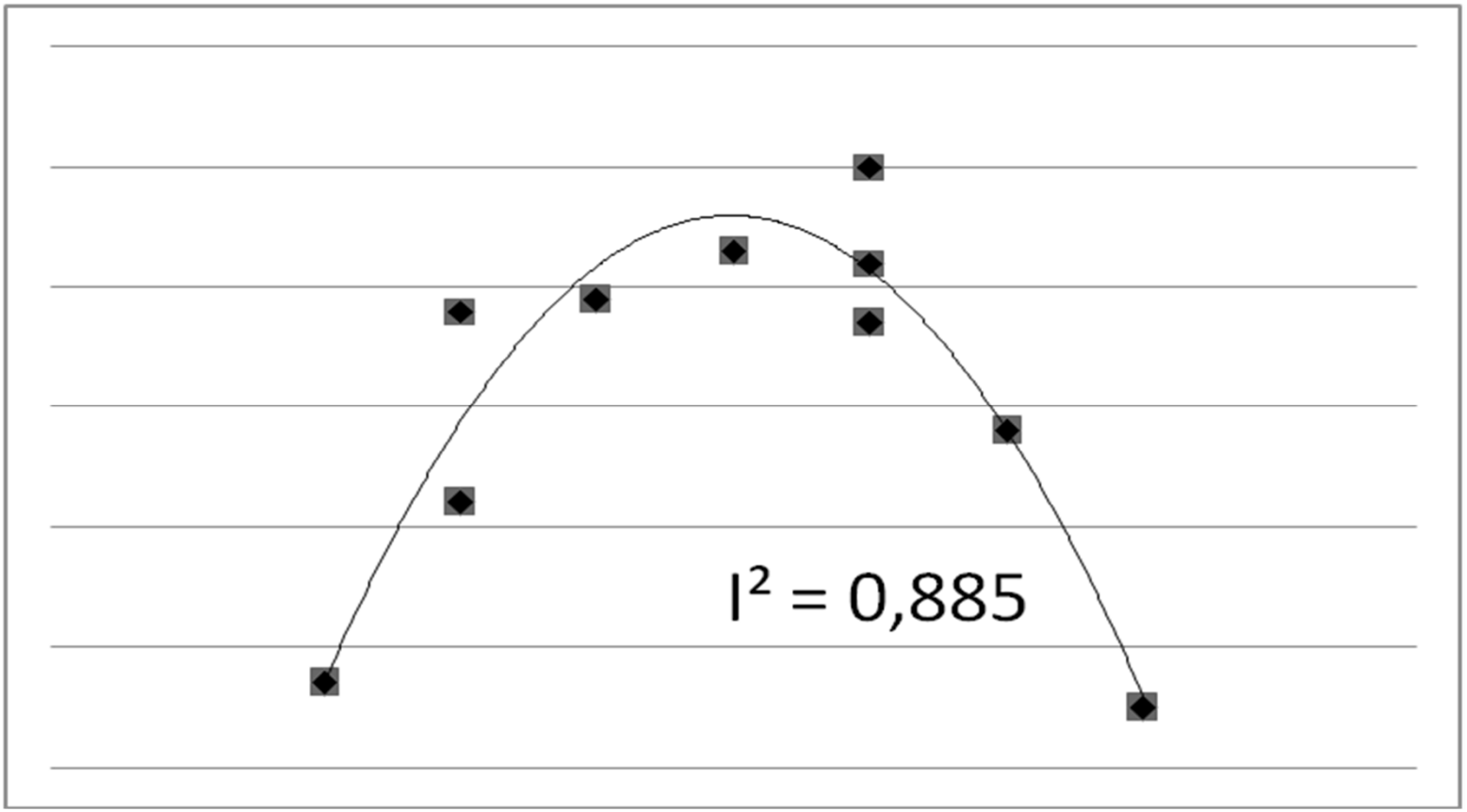
Př: Jiná data (r = 0,075)



Korelace spec. pro model jednoduché nelineární regrese

Př: Stejná data, ale jiný, kvadratický model

(*kde už tedy nepočítáme r, jen I^2 !*)



Jednoduchá regrese – různé modely

? Lze tedy říct, že parabola je vždy LEPŠÍ model než přímka ?

NE: Parabola je vždy VÝSTIŽNĚJŠÍ,

ale výhodou přímky je její

JEDNODUCHOST

Každý model = kompromis mezi výstižností a jednoduchostí

Jednoduchá regrese – různé modely

Adjustovaný index determinace R_{adj}^2

$$R_{adj}^2 = 1 - (1-I^2)\cdot(n-1)/(n-p)$$

= % kvalita modelu při zohlednění počtu parametrů (p), i ten slouží k porovnání různých modelů pro tatáž data,

a to dle hesla „čím větší (je R_{adj}^2), tím lepší (je pro daná data příslušný model)“

Testování - možnosti

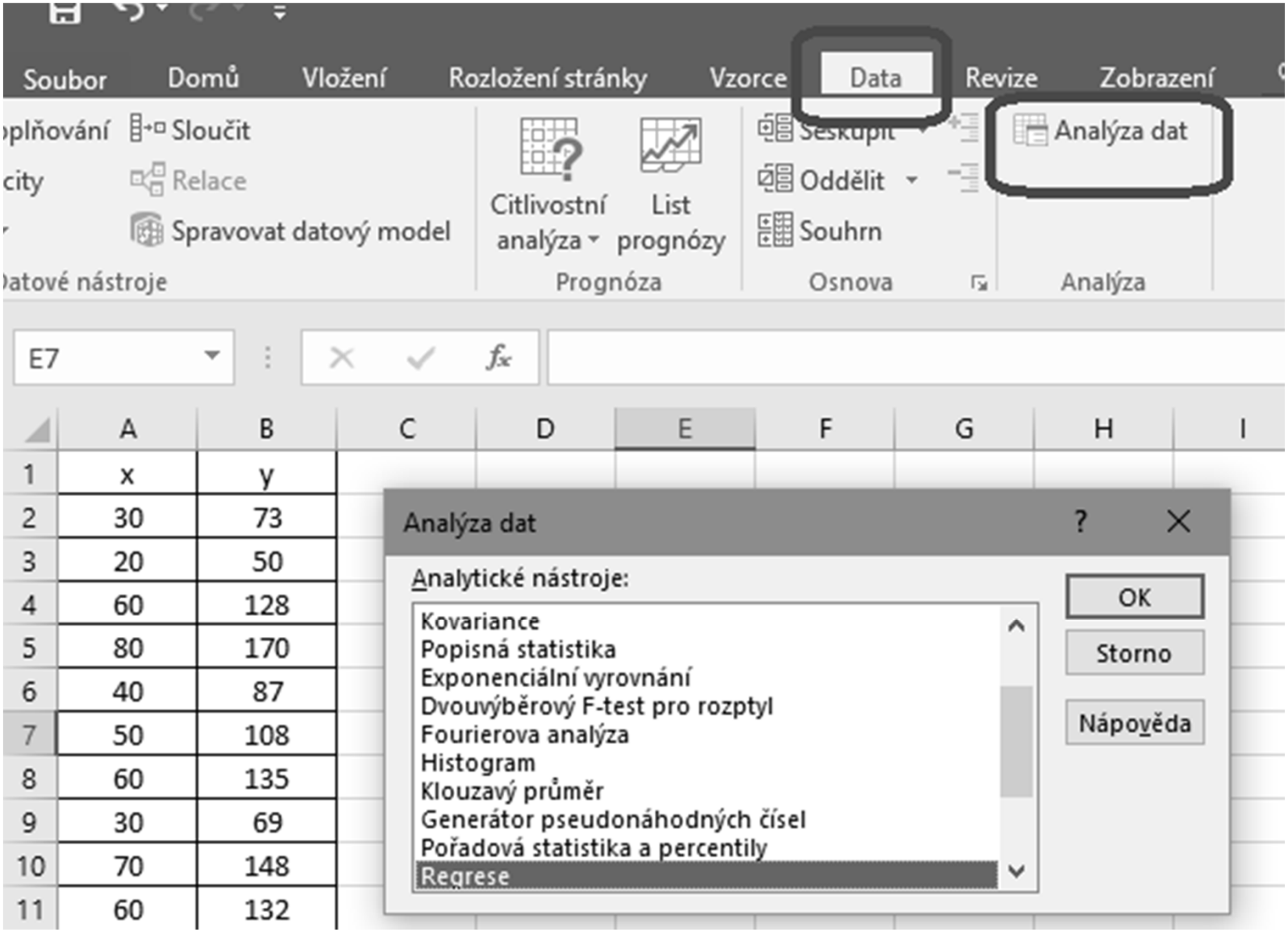
jednoduchá lineární regrese (p=2):

$H_0: \beta_1=0$ versus $H_1: \beta_1 \neq 0$

H_0….místo lineární funkce by jako model „bývala stačila“ funkce konstantní ($Y=\beta_0$) aneb „přímka s nulovou směrnicí“;

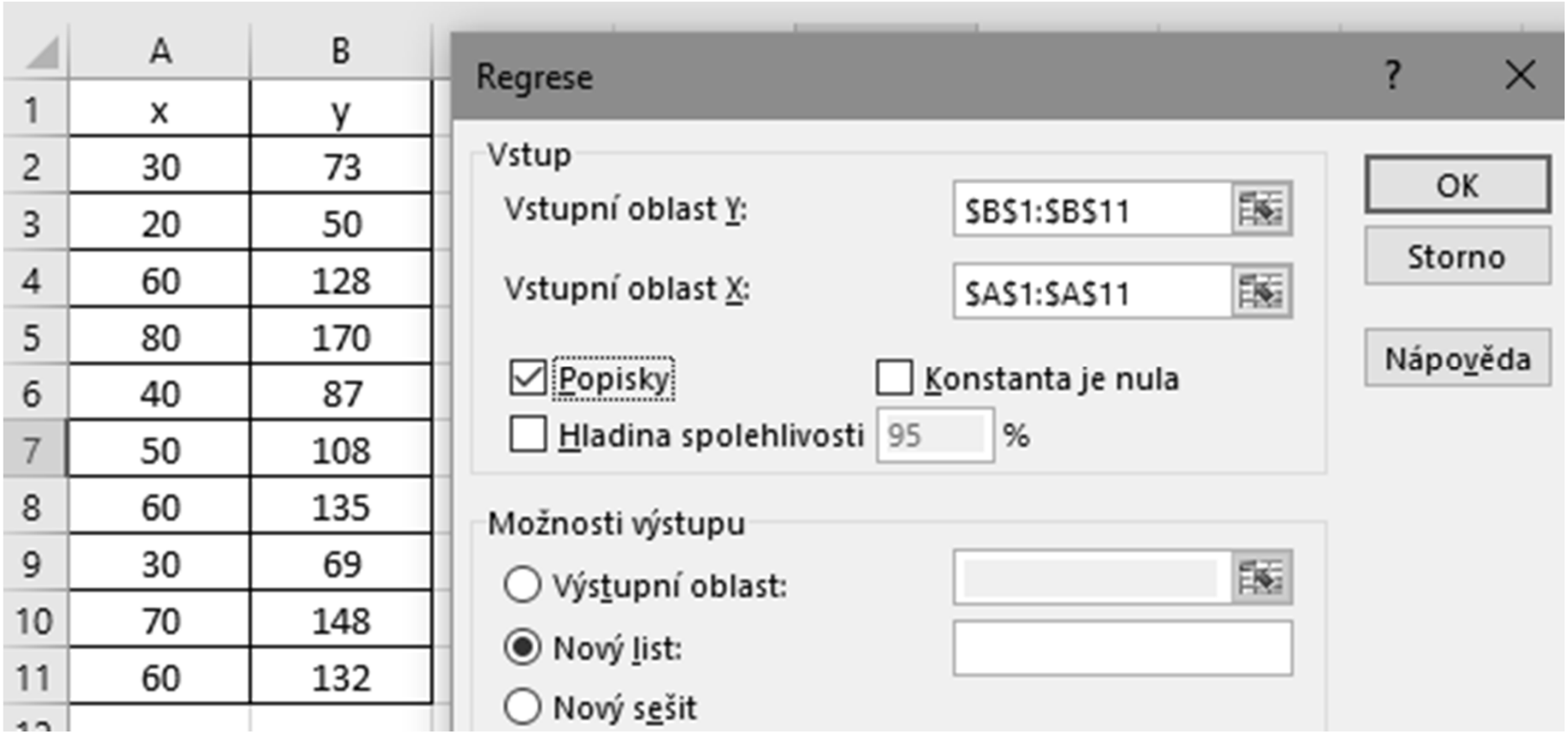
H_1….do vhodného modelu je potřeba zahrnout nenulovou „směrnici“

Testování regresních parametrů



	A	B
1	x	y
2	30	73
3	20	50
4	60	128
5	80	170
6	40	87
7	50	108
8	60	135
9	30	69
10	70	148
11	60	132

Testování regresních parametrů



	A	B
1	x	y
2	30	73
3	20	50
4	60	128
5	80	170
6	40	87
7	50	108
8	60	135
9	30	69
10	70	148
11	60	132

Testování regresních parametrů

VÝSLEDEK						
Regresní statistika						
Násobné R	0,997801	korel.koef.				
Hodnota spolehlivosti R	0,995608	index determinace				
Nastavená hodnota spole	0,995059	adjustovaný index determinace				
Chyba stř. hodnoty	2,738613					
Pozorování	10					
ANOVA	df				p-hodnota (celek)	
	Rozdíl	*SS*	*MS*	*F*	*Významnost F*	
Regrese	1	13600	13600	1813,333333	1,01959E-10	
Rezidua	8	60	7,5			
Celkem	9	13660				
	Koeficienty	*a stř. hoc*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	**10,00000**	2,5029	3,995302	0,00397576	4,228211282	15,77179
x	**2,00000**	0,047	42,58325	1,01959E-10	1,891694315	2,108306

Testování regresních parametrů

- Př: Data - Westwood Company (pokr.)

$p = 1{,}02 \cdot 10^{-10} < 0{,}05$

zamítáme $H_0 \rightarrow$ model JE VÝZNAMNÝ

Regrese pomocí R:

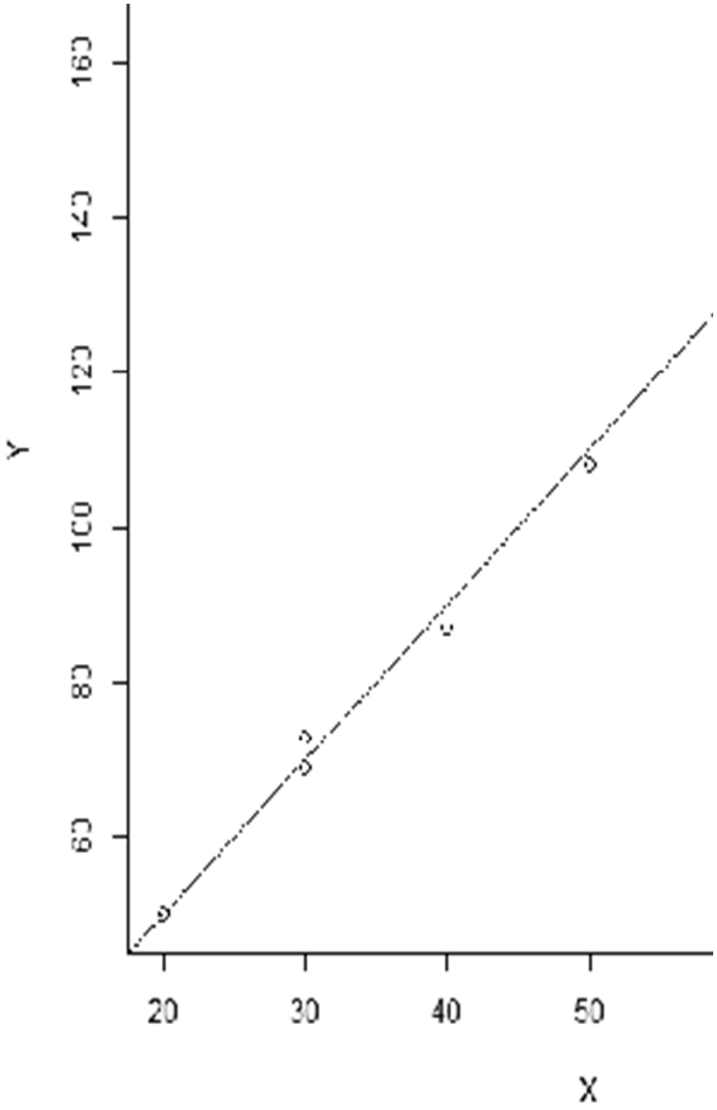
```
> X=c(30,20,60,80,40,50,60,30,70,60)
> Y=c(73,50,128,170,87,108,135,69,148,132)
> MODEL=lm(Y~X)
> summary(MODEL)
Call:
lm(formula = Y ~ X)

Residuals:
   Min     1Q Median     3Q    Max
  -3.0   -2.0   -0.5    1.5    5.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.00000    2.50294   3.995  0.00398 **
X            2.00000    0.04697  42.583 1.02e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.739 on 8 degrees of freedom
Multiple R-squared:  0.9956,    Adjusted R-squared:  0.9951
F-statistic:  1813 on 1 and 8 DF,  p-value: 1.02e-10

> R=cor(X,Y)
> R*R
[1] 0.9956076
> plot(Y~X)
> abline(MODEL)
```



GEOSTATISTIKA

Samostatně, viz prezentace
Přednáška 2018-05-03
ve studijních materiálech
(nepovinné, pro zájemce)