Přípravný kurz - kvantitativní výzkum

FSS928

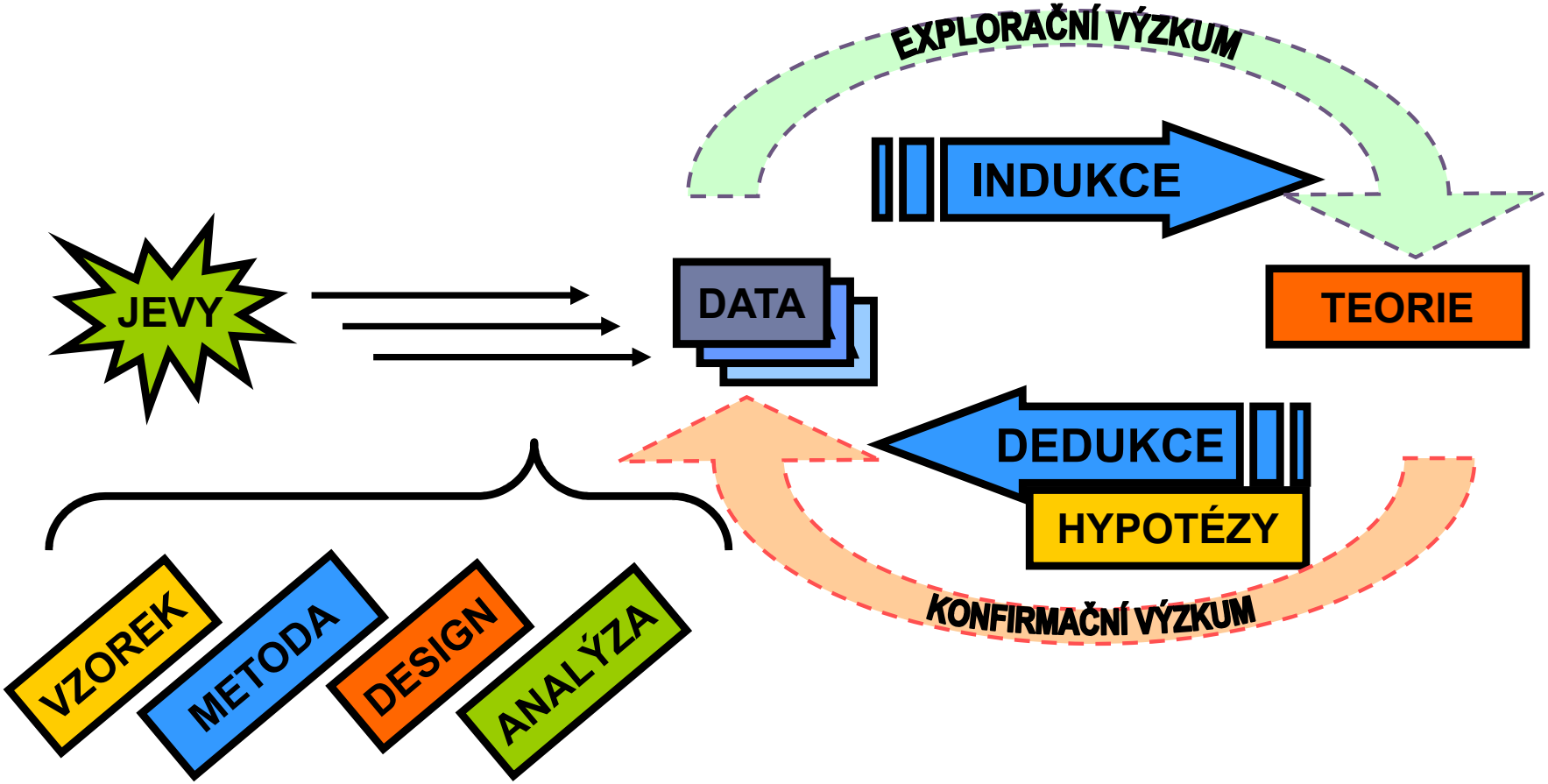
Celkový plán

- Setkání 1
  - Rekapitulace výzkumného kontextu
  - Rekapitulace základních principů statistické deskripce
  - Analýza rozptylu jako 1. obecná podoba lineárního modelu
- Setkání 2
  - Lineární regrese jako 2. obecná podoba lineárního modelu
- Setkání 3
  - Základy měření konstruktů
  - Explorační faktorová analýza jako model měření

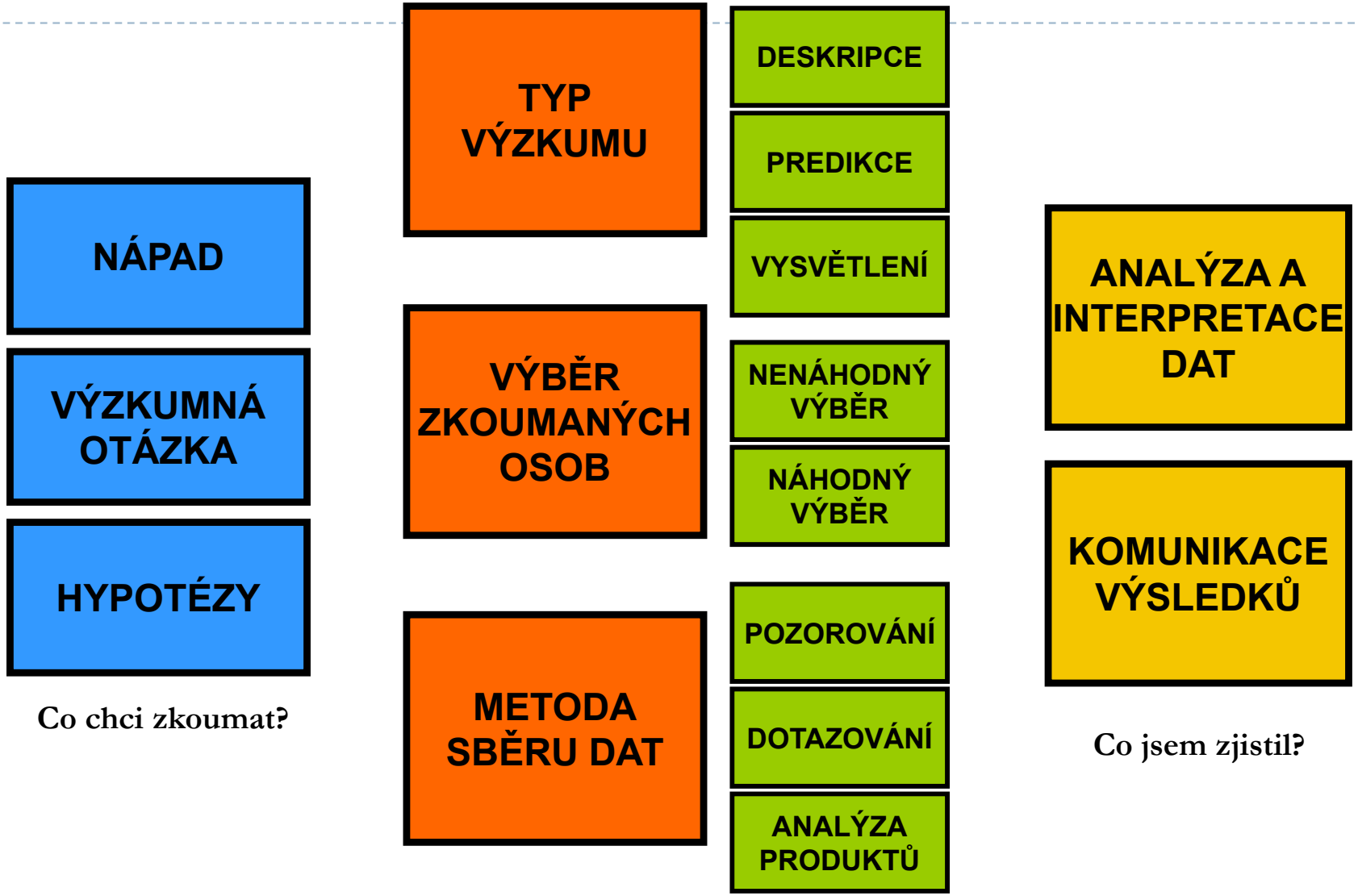
Hiles: 8 „faktů“ o výzkumu

- Výzkum je zkoumání, které vede k **novým poznatkům**.
- **Výzkumná otázka** a poznatky musí být diskutovány **v kontextu předchozích znalostí**.
- Každý výzkum spočívá na nějakých **předpokladech** – různí badatelé nevyhnutelně pracují s **různými paradigmaty**.
- Ve výzkumu děláme teoretická a praktická **rozhodnutí**, přičemž se držíme **etických** principů.
- Postupy ve výzkumu by měly být **systematické**, přísné a musí být natolik jasné, aby se daly **replikovat**.
- Výzkumná zjištění by měla být **jasná a přesvědčivá** pro druhé.
- Výzkum jen **zřídka úplně zodpoví** výzkumnou otázku; obvykle vede spíše k nápadům na další zkoumání.
- Do výzkumu se pouštíme, protože máme zájem a chceme něco změnit – **publikace** a sdílení s kolegy je přirozenou součástí výzkumného procesu.

JAK FUNGUJE VĚDA?



MAPA VÝZKUMU



KONCEPČNÍ A PRAKTICKÁ STRÁNKA VÝZKUMU

VÝZKUMNÁ OTÁZKA

HYPOTÉZA

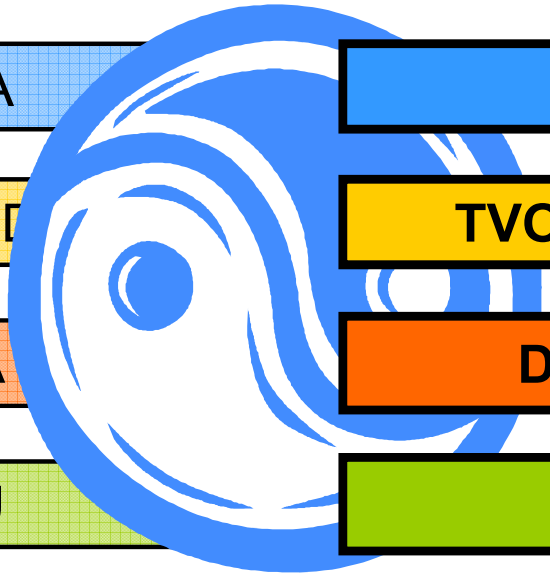
KONCEPČNÍ
STRÁNKA VÝZKUMU

EXTERNÍ VALIDITA

KONSTRUKTOVÁ VALID

INTERNÍ VALIDITA

VALIDITA ZÁVĚRŮ



PRAKTICKÁ
STRÁNKA VÝZKUMU

VZOREK

TVORBA DAT – MĚŘENÍ

DESIGN VÝZKUMU

ANALÝZA DAT

upraveno dle Trochim, Donnelly, www.socialresearchmethods.net

K teorii, konstruktům, hypotézám a modelům

- …
- Model je naše představa o nějakém aspektu fungování jevu.
- Jev nemůžeme přímo uchopit, již jeho vnímání a myšlení je tvůrčí čin, jehož výsledkem je konstrukt.
- Konstrukt je samozřejmě redukcí … abstrakcí, zprůměrováním v myšlenkovém prostoru i čase – konstrukt je model.
- Změřením (operacionalizací) nějakého aspektu konstruktu vzniká proměnná … a další redukce.
- Model se stává modelem fungování proměnné
- Při převodu modelu do statističtiny si často uvědomíme míru redukce, k niž zatím došlo a klesneme na mysli.
- Problém není v míře schematičnosti (ta je poplatná naší kognitivní kapacitě) ale v jejím uvědomění.
- Problém bývá také ve zpětném „překladu“ zpět k jevům.

K externí validitě - reprezentativnosti

- Vzorek jevů má dobře zastupovat populaci jevů, o které se chceme něco dozvědět – reprezenta-ce (-tivnost)
- Celá statistika stojí na jediné věci, která je jistá a to je náhoda (teorie pravděpodobnosti) … odtud náhodný výběr.
  - mnoho výběrových strategií
- Problematičnost náhodnosti výběru lidí >>
  - nutnost zpětné kontroly/úvahy, replikace
  - překvapivá robustnost při absenci výrazných zkreslujících faktorů

TVOŘÍME DATA - VALIDITA

KONSTRUKTOVÁ VALIDITA (v širším smyslu)

- měříme ten správný aspekt správného jevu?

- **OBSAHOVÁ:** názor odborníků, construct mapping, faktorování
- ***ZJEVNÁ (face, zdánlivá):*** *názor laiků, <u>není</u> validita*
- **EMPIRICKÁ - KRITERIÁLNÍ:** korelace mezi výsledky postupu a kritériem
  - SOUBĚŽNÁ(konvergentní), PREDIKTIVNÍ, INKREMENTÁLNÍ, DIFERENCIÁLNÍ (diskriminační)
- **KONSTRUKTOVÁ** (v užším smyslu) = unidimezionalita + struktura korelací se souvisejícími konstrukty odpovídá teorii
  - Nikdy nekončící proces ověřování

TVOŘÍME DATA - RELIABILITA

PROMĚNNÁ – odráží jeden aspekt jevu, výsledek měření

RELIABILITA

- jak přesně měříme? NIKDY zcela přesně, vždy s chybou

- STABILITA - korelace opakovaných měření - r_{tt}
- OBJEKTIVITA – shoda posuzovatelů, pozorovatelů, adminstrátorů >> standardizace
  - Cohenovo κ
- VNITŘNÍ KONZISTENCE
  - zjišťují všechny položky totéž?
  - Cronbachovo α - horní mez reliability

Reliabilita je podmínkou validity

2 SKUPINY METOD

OBSERVAČNÍ

- POZOROVÁNÍ
- ANALÝZA PRODUKTŮ
- MAPY, STOPY
- EXPERIMENTÁLNÍ METODY
- TESTY

DOTAZOVACÍ

- ROZHOVOR
- FOCUS GROUP
- DOTAZNÍK
  - POSTOJOVÉ ŠKÁLY
  - POSUZOVACÍ ŠKÁLY

Experimentování - základy

Když udělám *tohle*, stane se to, co myslím?

To, co se stalo, stalo se to **pouze** díky tomu, co jsem udělal?

- **manipulace** nezávislou proměnnou
  - komu + udělám
- **měření** závislé proměnné
- **kontrola** intervenujících (vnějších) proměnných
  - AKT :: fixace, párování, znáhodnění, měření, znalost :: PAS

Experimentování – praktické fičury

- **Kauzální** usuzování
  - kauzalita je přenositelnější než koincidence
- Ex je **interaktivní**, s důrazem na **kontext**
  - potenciál pro exploraci, kvalitativní analýzu.
- Ex je malý, flexibilní
- Ex je **nenáročný na vzorek**
  - teoretické a replikační zobecňování
- Ex umožňuje nejširší paletu **dg. metod**
- Ex je **náročný na vědění**
- Ex je **náročný na kontrolu**

Experimentování v „terénu“

Problémy

- Náhodné rozdělení (R:116)
- Manipulace NP
- Etika
- Kontrola spíše pasivní

Výhody

- Zobecnitelnost (EV)
- Méně reaktivity
- Dostupnost lidí

PRAVÉ EXPERIMENTY v terénu zřídka možné

Vždy o nich alespoň zauvažujeme. (R:123 box 5.5, 5.6)

KVAZIEXPERIMENTY (R:135), SINGLE CASE EX. (R:146)

EX-POST-FACTO (post hoc) – retrospektivní, přirozené ex. (R:154)

Výběrová šetření (surveys)

Jaký je aktuální stav něčeho?

Souvisí spolu výskyt různých jevů?

- V psychologii: korelační design, longitudinální
- Fíčury
  - cílem je popis
  - obvykle důraz na neintervenování
  - převaha dotazování
  - důraz reprezentativnost vzorku (R:161)
  - spíš jednorázovost, rozsáhlost
  - široký záběr

Výběrová šetření – praktické aspekty

- Samo šetření je intervencí
- Problémy self-reportů
  - lidé toho o sobě vědí méně, než si myslíme
  - mezi „říkáním“ a chováním je malá korelace
- Poskytuje info především o tom, jak se věci mají, než o tom, jak je změnit (budík).
- Obvykle třeba zjišťovat mnoho věcí
- Je-li „populace“ malá, snažíme se o *cenzus,* náhodný výběr není vhodný.
- Pilotáž otázek je obvykle nutností
  - cognitive interview
- Když nezasahovat, zvážit dokumenty, archivy, stopy.

Rekapitulace výzkumného kontextu

- Teorie
  - Jevy, pojmy, konstrukty
- Výzkumná otázka, cíl
- Hypotéza
  - Operacionalizace >> proměnné
- Data
  - Vzorek (reprezentace)
  - Design (vyloučení nežádoucích vlivů)
- Data vs. hypotéza … analýza
- Odpověď na výzkumnou otázku
- Začlenění zjištění do teorie

STATISTIKA

Co je to vlastně statistika?

- **Popis** získaných **dat** o **jevech**, které se vyskytují ve větších množstvích (>7)
  - Popis **proměnných**: jaké podoby jevu, jak časté?
  - Popis vztahů mezi proměnnými/jevy

- Statistické **usuzování** ze vzorku na populaci
  - Pravděpodobnostní usuzování
  - Konfrontace očekávání (modelů) se získanými daty
  - Testování hypotéz

K čemu je statistika jako taková?

- Formalizované **zpracování zkušenosti**, když
  - počet zkušeností, výskytů jevu přesáhne 7±2 (automat)
  - hledané je malé (mikroskop)
  - záludnosti naší kognice představují problém
- „Objektivní“ (=v komunitě srozumitelný) popis výskytu jevů
- Hledání společného, typického, normálního i jedinečného
- Hledání vztahů mezi jevy (v precizně stanovených podobách, modelech)
- Statistika je nástroj - na rozdíl od privátního myšlení je „open-source“
- Trénuje myšlení
  - kritické myšlení
  - myšlení o variabilitě jevů
  - uvědomění si všudypřítomnosti chyby měření (vnímání)
  - **pravděpodobnostní myšlení**

Data, proměnné

- Data vznikají měřením (aplikací metod) **jevů**
- Proměnné tvoříme z dat
  - Proměnné vznikají **kódováním dat**
  - Z jedněch dat můžeme udělat více proměnných
- Proměnné reprezentují *znaky, charakteristiky, atributy, vlastnosti* zkoumaných jevů či objektů, popř. jejich kombinace
- Proměnné nabývají různých hodnot, pokud ne, jsou to **konstanty**

Vznik dat - měření

- Standardizovaný postup, procedura
- M= přiřazování čísel z nějaké množiny čísel podle pravidla
- **Procedura dává číslům smysl**
- Tato procedura je **vždy** zatížena chybou
- Y = T + E
  - Naměřená hodnota = skutečná hodnota + chyba

Chyby měření

- (ne)přesnost
  - Měříme-li vícekrát tentýž objekt, střední hodnota všech měření odpovídá skutečné hodnotě.
  - ≈ náhodná chyba
  - ≈ ≈ přibližně odpovídá pojmu reliabilita
- (ne)správnost
  - Měříme-li vícekrát tentýž objekt, střední hodnota je systematicky vyšší nebo nižší než je skutečná hodnota
  - ≈ systematická chyba
- Tyto chyby se mohou kombinovat

AJ: accuracy, bias, random error, systematic error, reliability

Úrovně měření (typy měřítka, škály)

Úroveň	***Operace***	***Příklady***
Nominální	= ≠	pohlaví, tramvaj, hodnota
Ordinální	= ≠ > <	známky, souhlasení
Intervalová	= ≠ > < + −	°C, IQ, „dobré“ psychotesty
Poměrová	= ≠ > < + − × ÷	K, váha, počty, frekvence

1+2: kategorické, 2:pořadová; 1: kvalitativní ☹

3+4: metrické, kardinální;

viz extrakt z Urbánka v ISu

AJ: level of measurement, nominal, ordinal, interval, ratio

☹: intenzivní, pořadová/kategorická ordinální

Další typy proměnných

- Diskrétní vs. spojité
- Dichotomické (alternativní) vs. polytomické

AJ: discrete, continuous, dichotomous, alternative, polytomous


Google
is it normal
is it normal to talk to yourself
is it normal for your period to be brown
is it normal to miss a period
is it normal to be sexually attracted to numbers
is it normal to bleed during intercourse
is it normal to get your period late
is it normal to poop green
is it normal to have headaches everyday
is it normal to have hair on your bum
is it normal to spot during pregnancy
Google Search
I'm Feeling Lucky
Advanced search
Language tools
FAIL
failblog.org

Jaké hodnoty máme v datech?

- Jaké hodnoty proměnné/ých se v datech vyskytují? – *třídění, kódování*
  - Jaké různé odpovědi jsme získali na tu kterou otázku dotazníku?
  - Jaké různé počty sledovaných chování se při pozorování vyskytly?
- Kolik kterých hodnot máme? – *četnosti*
  - Je některých víc, jiných míň?
  - Zdá se být v četnostech jednotlivých hodnot nějaký řád?

Mám rád(a), když je všechno v mém životě jasné a přehledné.

		četnost	rel. četnost	rel. četnost platných hodnot	kumulativní relativní četnost
Valid	velmi souhlasím	11	15,3	15,5	15,5
	středně souhlasím	19	26,4	26,8	42,3
	spíš souhlasím	25	34,7	35,2	77,5
	spíš nesouhlasím	7	9,7	9,9	87,3
	středně nesouhlasím	8	11,1	11,3	98,6
	velmi nesouhlasím	1	1,4	1,4	100,0
	Total	71	98,6	100,0	
Missing	System	1	1,4		
Total		72	100,0		

Přibližně kolik hodin týdně strávíte sportováním? (Binned)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<= 2,0	34	47,2	47,9	47,9
	2,1 - 4,0	24	33,3	33,8	81,7
	4,1 - 6,0	9	12,5	12,7	94,4
	8,1 - 10,0	2	2,8	2,8	97,2
	10,1+	2	2,8	2,8	100,0
	Total	71	98,6	100,0	
Missing	System	1	1,4		
Total		72	100,0		

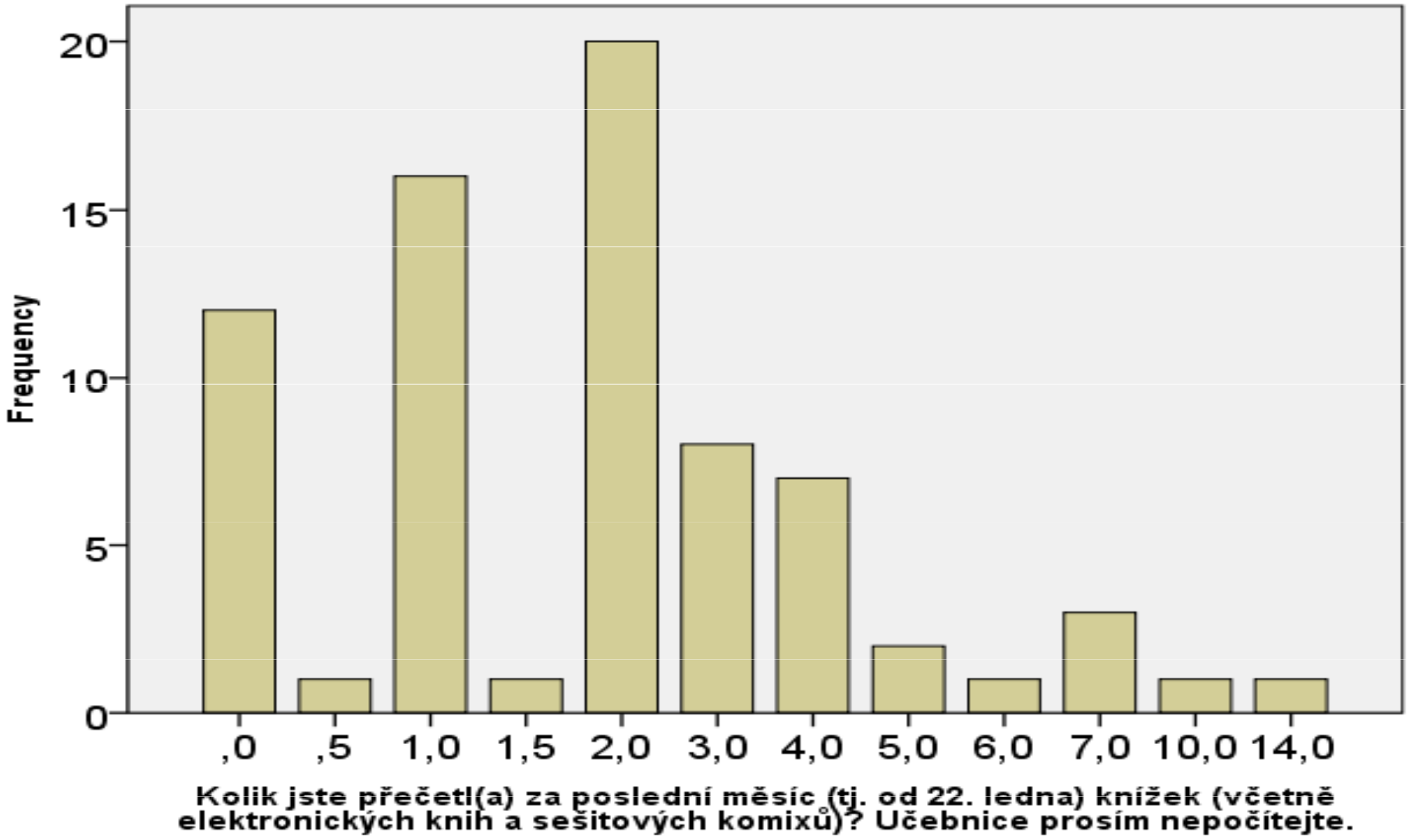
SPSS intervalové četnosti samo nedělá. Je třeba rekódovat hodnoty do intervalů (nová proměnná). K tomu např. fce Transform - Visual Binning.

Grafické podoby tabulky četností

- Kategorické proměnné
  - sloupcový graf (diagram)
  - koláčový diagram – zřídkakdy, neukazuje rozložení
- Metrické proměnné
  - histogram / stem-and-leaf – rozdělení hodnot do intervalů

AJ: bar chart, histogram, pie chart, frequency distribution

Sloupcový diagram



Histogram

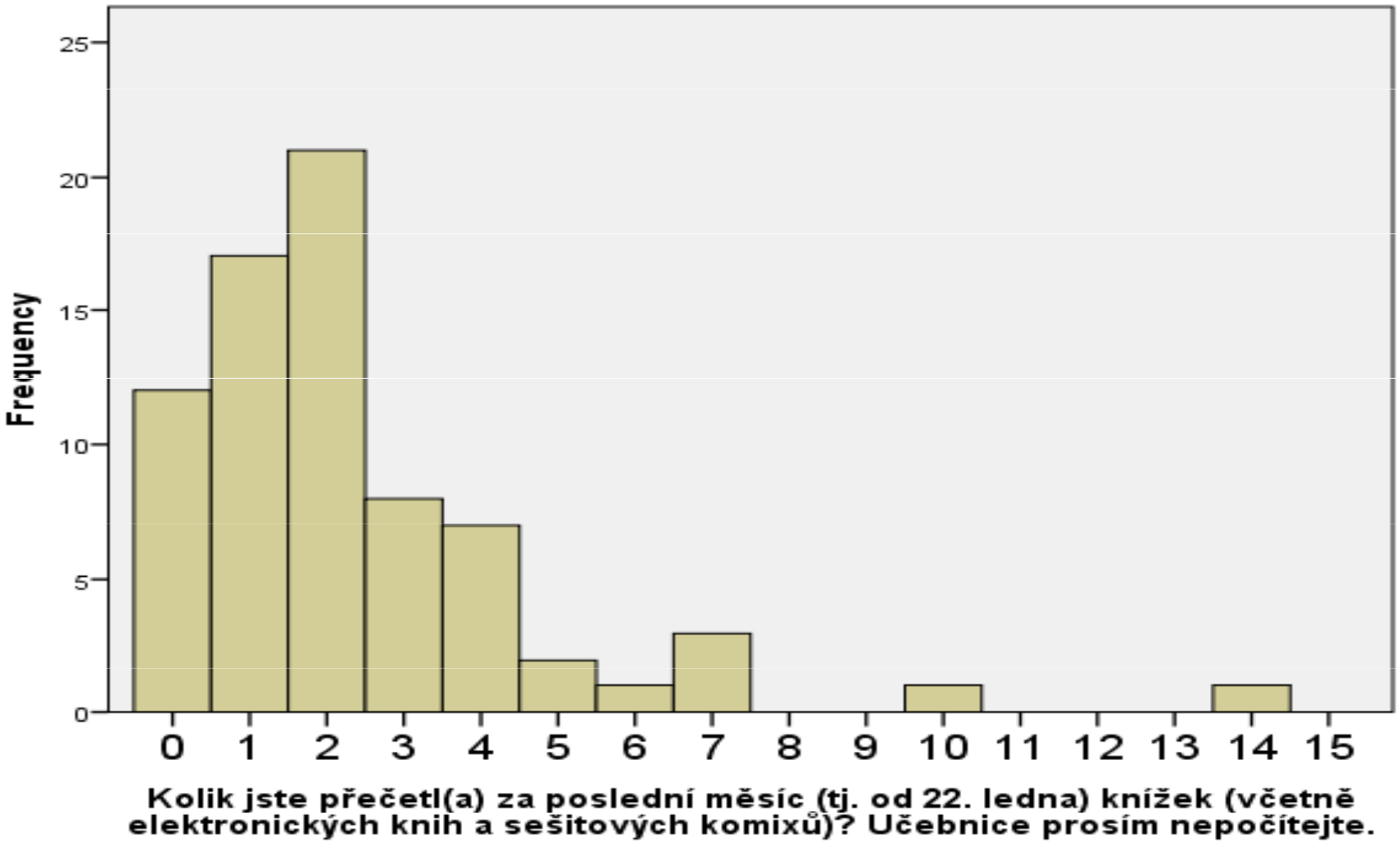


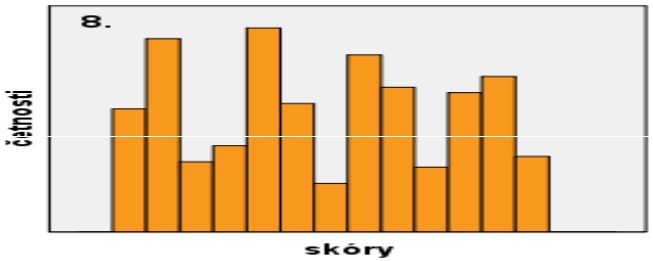
Diagram „stonek a list“

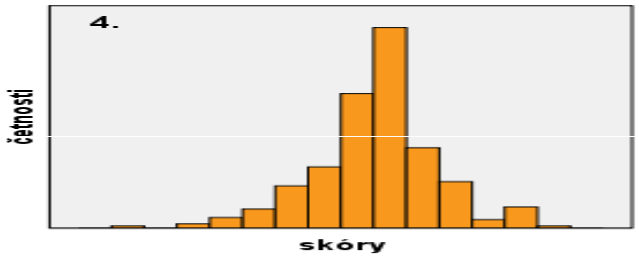
```
 Frequency    Stem &  Leaf
   32,00        0 .  00000000000000000000000000000000
   18,00        1 .  0000000000000000
   14,00        2 .  00000000000000
    7,00        3 .  0000000
    2,00        4 .  00
    4,00        5 .  0000
    1,00        6 .  0
    1,00        7 .  0
   10,00 Extremes    (>=8,0)

Stem width:         1
Each leaf:       1 case(s)
```

Větev°	list (jed. list1,000000, např. 6°5 = 6,500000)			
0°	00000000000000000000000000000000			.
1°	0000000000000000	.	.	.
2°	00000000000000	.	.	.
3°	0000000	.	.	.
4°	00	.	.	.
5°	0000	.	.	.
6°	0	.	.	.
7°	0	.	.	.
8°	0	.	.	.
9°		.	.	.
10°	000000	.	.	.
11°		.	.	.
12°		.	.	.
13°		.	.	.
14°		.	.	.
15°	000	.	.	.
min = 0,000000	max = 15,00000			Celk. N:

Rozložení *rozdělení, distribuce* četností

- Měřené jevy jsou nějak rozděleny do kategorií (intervalů) a tyto kategorie jsou různě „populární“ – četné.
- Četnosti u reálných ordinálních a vyšších proměnných obvykle nebývají **distribuovány** nahodile – jejich rozdělení zobrazené histogramem má popsatelný tvar.

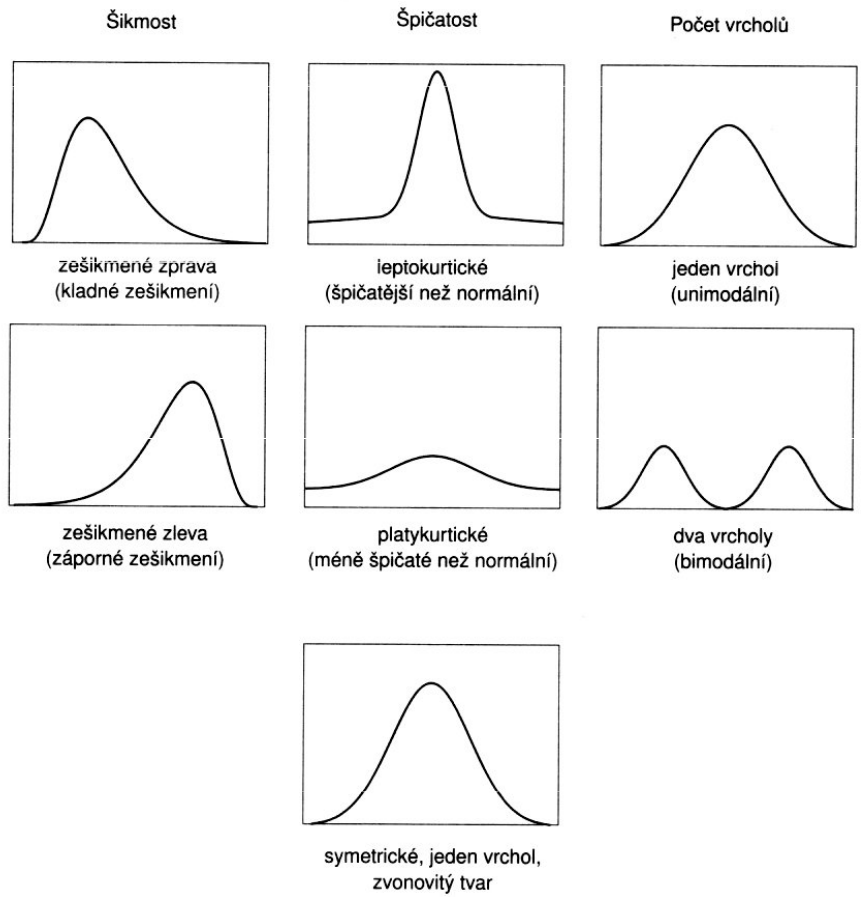




- **Rozdělení** četností je tedy to, kolik relativně (či absolutně) máme kterých hodnot měřené proměnné.
  - Typicky lze přibližně popsat slovy, např.: vyskytlo se hodně středních hodnot a relativně málo extrémních hodnot.
  - Toto **rozložení** jevů na měřené škále je nejlépe vidět na grafech.
  - Obvykle nějaké konkrétní rozložení očekáváme.

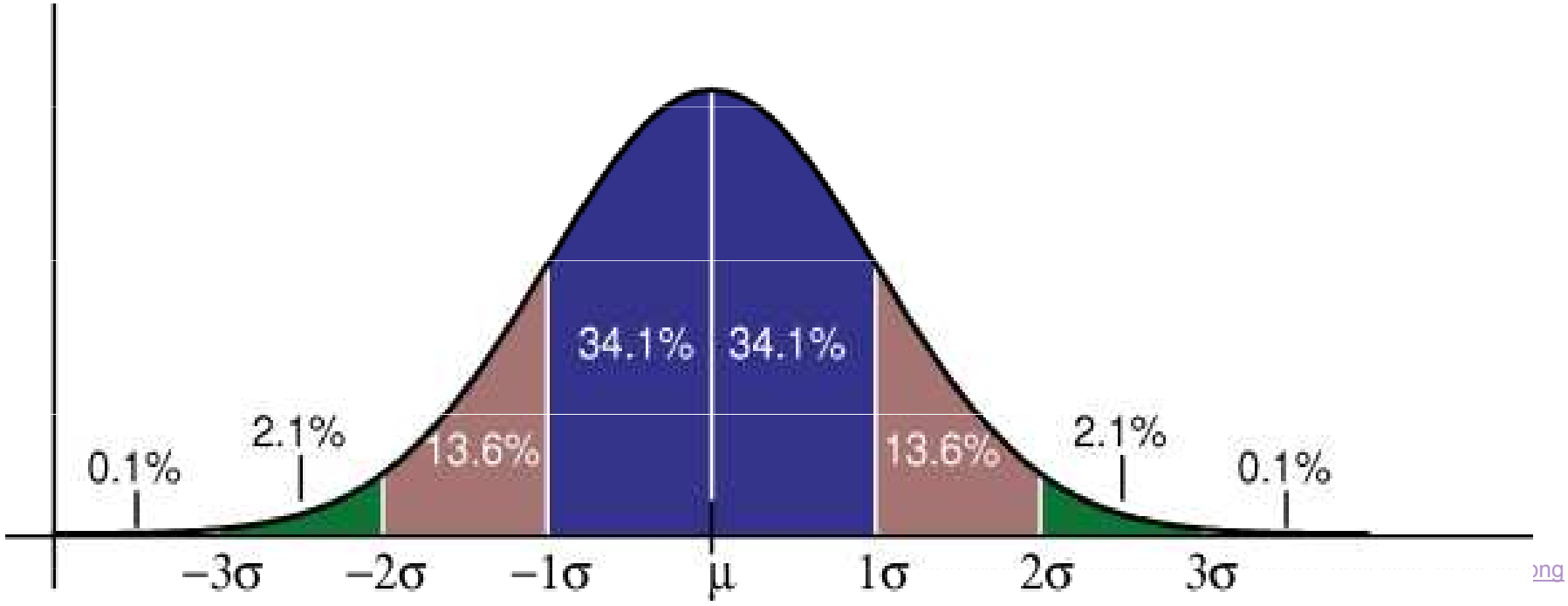
Tvar rozložení četností

- Normální
- Uniformní
- Počet vrcholů
  - Unimodální, bimodální, multimodální
- Zešikmení
  - Zešikmené zprava (pozitivně), efekt podlahy
  - Zešikmené zleva (negativně), efekt stropu
- Strmost
  - Leptokurtické, platykurtické



AJ: frequency distribution, normal, rectangular, unimodal, bimodal, positively/negatively skewed, lepto(platy)kurtic, floor/ceiling effect

Normální (Gaussovo) rozložení



- „Normální“ ve smyslu „velmi běžné“
- Tam, kde se setkává mnoho nezávislých vlivů.
- Ne vždy, nesouvisí s „kvalitou“ dat.

AJ: normal distribution, bell curve

Pravděpodobnostní rozložení náhodné proměnné

Je-li **proměnná náhodná** (tj. její hodnoty lze považovat za výsledek náhodných pokusů)… …jaká je *P* výskytu jednotlivých možných hodnot?

- Vzpomeňme si, že $P(A) = n / m$, blíží-li se počet pokusů ∞ (populaci)
- Máme-li tedy dost velký, náhodně vybraný vzorek, pak *P* výskytu jednotlivých hodnot ≈ jejich relativní četnost

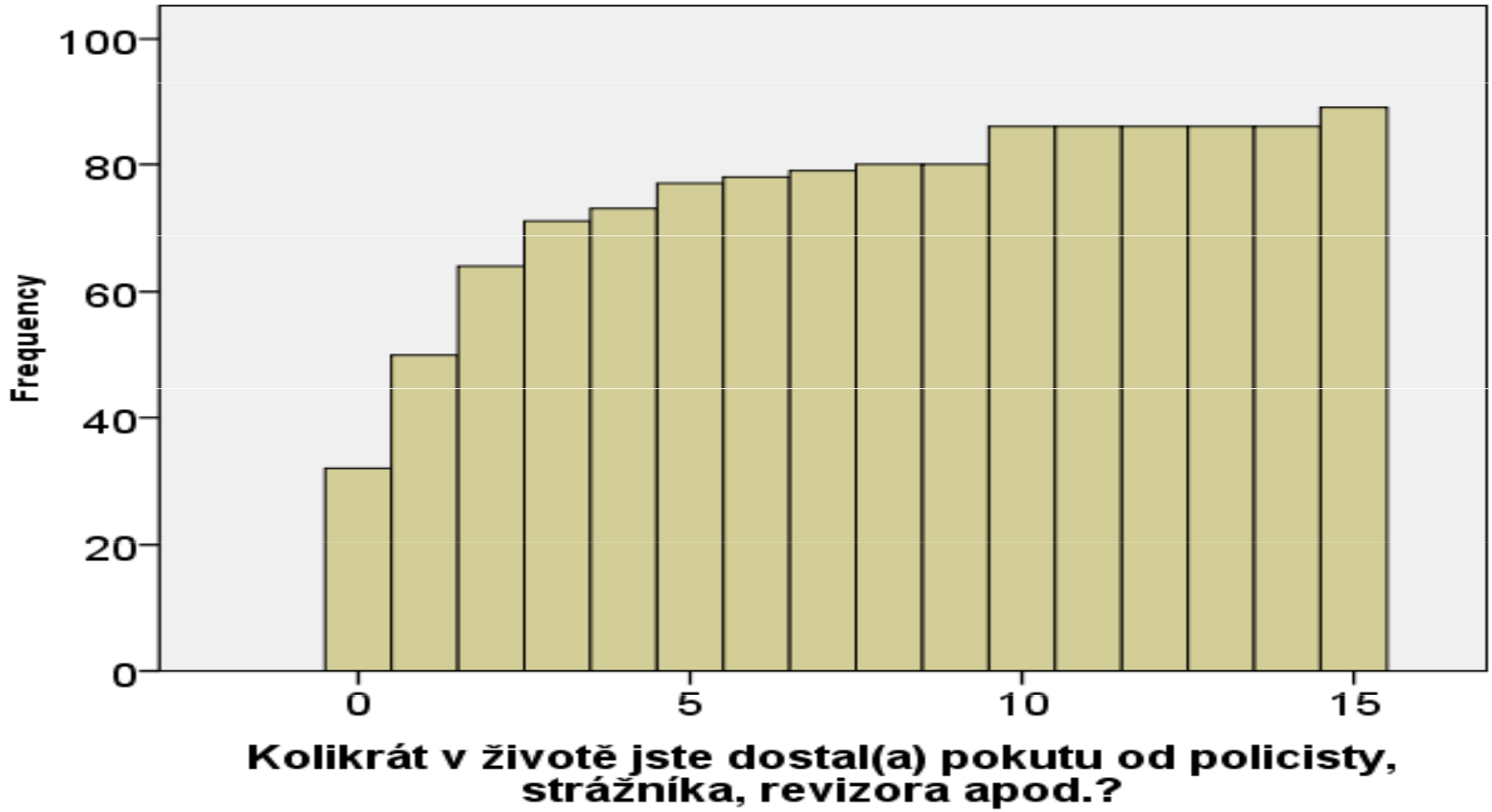
Pravděpodobnostní rozložení = teoretické rozložení rel. četností

- U diskrétních proměnných uvažujeme o *P* výskytu jednotlivých hodnot.
- U spojitých proměnných neuvažujeme o *P* výskytu jednotlivých hodnot (∞), ale spíše o *p* výskytu hodnot v intervalech – **hustota pravděpodobnosti**

- P-nostní rozložení je popsáno **distribuční funkcí**
  - $F(x) = P(X \leq x)$ tj. *P* výskytu hodnot ≤ x
  - Tato *P* je rovna „ploše oblasti pod křivkou hustoty pravděpodobnosti“

AJ: random variable, probability distribution, distribution function, probability density

Kumulativní histogram



Popis rozložení pomocí percentilů

- *X*-tý percentil
  - hodnota, pro kterou platí, že *X* % lidí (jevů) ve vzorku má/získalo tuto nebo menší hodnotu
  - lze snadno odečíst z kumulativního histogramu či patřičného sloupce tabulky četností
- Typicky rozložení popisujeme
  - 10., 20., …, 80.,90. percentilem – obecně
  - min, 25., 50., 75., max – nejčastěji
  - min., 1., 5., 10., 25., 50., 75., 90., 95., 99. – v normách

Shrnutí

- První informací (*statistikou*), která nás zajímá je **četnost** výskytu jednotlivých hodnot (resp. hodnot uvnitř jednotlivých intervalů)
- Konfiguraci **četností** nazýváme **rozložení (rozdělení)**.
- Rozložení popisujeme (=komunikujeme je)
  - tabulkou četností
  - graficky – histogram, sloupcový diagram
  - pomocí percentilů
- O typu, tvaru **rozložení** hodnot proměnné uvažujeme většinou graficky – **histogram, sloupcový diagram**.
- Nej… rozložením je tzv. **normální rozložení**.
- Byť tohle je 5. třída ZŠ – už tady se **šidí**.

Nedalo by se rozložení hodnot proměnné popsat úsporněji než pomocí tabulky četností, histogramu?

Kde na měřené škále se data nalézají?

UKAZATEL CENTRÁLNÍ TENDENCE

Jak moc jsou na ní rozptýlená?

UKAZATEL VARIABILITY

Centrální tendence (=střední hodnoty, umístění)

- CT je jeden údaj, jímž se snažíme popsat rozložení četností jedné proměnné
- Jeho kouzlo i zrádnost je právě v tom, že je to právě jeden údaj.
- CT udává průměrnou, typickou, reprezentativní, *očekávanou* hodnotu
  - Co se tím míníme, záleží na tom, jakou míru CT se rozhodneme použít

AJ: measures of central tendency, of location

Modus, medián a průměr

Modus - kategoriální typická hodnota

$\hat{X}$, Mo

- nejčastější hodnota, h. s nejvyšší četností
- jediná možnost u nominálních dat, u vyšších úrovní často užitečnou volbou

Medián – pořadová střední hodnota

$\tilde{X}$, Md

- hodnota prvku uprostřed uspořádaného souboru, 50. percentil (P_{50})
- při sudém počtu prvků je mediánem kterékoli číslo z intervalu mezi nejbližší vyšší a nejbližší nižší hodnotou (konsensuálně střed intervalu)
- pořadová data a výše

Aritmetický průměr – deviační, ochylková, momentová střední h.

$\overline{X}$, M, m

- jak ho znáte ze školy
- pouze intervalová a poměrová data
- velmi citlivý na extrémní hodnoty

AJ: mode, median, mean

Tcelkem Stem-and-Leaf Plot

```
 Frequency    Stem &  Leaf

     3,00         0 .  011
    15,00         0 .  222333333333333
    24,00         0 .  444444444444555555555555
    15,00         0 .  666666666777777
     3,00         0 .  889
     8,00         1 .  00011111
     2,00         1 .  23
     5,00 Extremes    (>=1388)

Stem width:      1000
Each leaf:       1 case(s)
```

N	Platných	75
	Chybějících	0
Průměr		683
Medián		557
Modus		49[a]

a. Multiple modes exist. The smallest value is shown

Míry variability (rozptýlenosti)

- Druhé číslo, jímž popisujeme rozložení hodnot proměnné
- Udává, jak moc či málo jsou data na škále rozptýlená.
  - Malá variabilita = většina hodnot v souboru je stejných nebo velmi blízkých
  - Vysoká variabilita = hodnoty jsou velmi rozmanité (n. rozložení je bimodální)

Rozpětí, rozptyl, směrodatná ochylka

Nominální – entropie – nepoužívá se

Pořadové

- (variační) rozpětí = $X_{max} - X_{min}$ (extrémně roste s velikostí vzorku)
- (inter)**kvartilové rozpětí** = $Q_3 - Q_1$, IQR

Odchylkové (deviační, momentové) ukazatele

- založené na odchylkách od průměru: $x = X - m$
- průměrná absolutní odchylka ($\Sigma|x| / n$) – nepoužívá se
- průměrná odchylka na druhou – **rozptyl**
  - populační ($\Sigma x^2 / n$) vs. výběrový ($\Sigma x^2 / (n-1)$)
  - součet odchylek na druhou = **suma čtverců**
- **směrodatná odchylka** (standardní odchylka)
  - odmocnina rozptylu - návrat k původní jednotce

AJ: measures of variability, entropy, rank-order, range, interquartile range, variance, standard deviation, sum of squares, square, square root

Tcelkem Stem-and-Leaf Plot

```
Frequency    Stem &  Leaf

     3,00    0 .  011
    15,00    0 .  222333333333333
    24,00    0 .  444444444445555555555555
    15,00    0 .  666666666777777
     3,00    0 .  889
     8,00    1 .  00011111
     2,00    1 .  23
     5,00 Extremes (>=1388)

Stem width:      1000
Each leaf:       1 case(s)
```

Směrodatná odch.		440
Rozptyl		193673
Variační rozpětí		2313
Percentiles	25	402
	50	557
	75	787

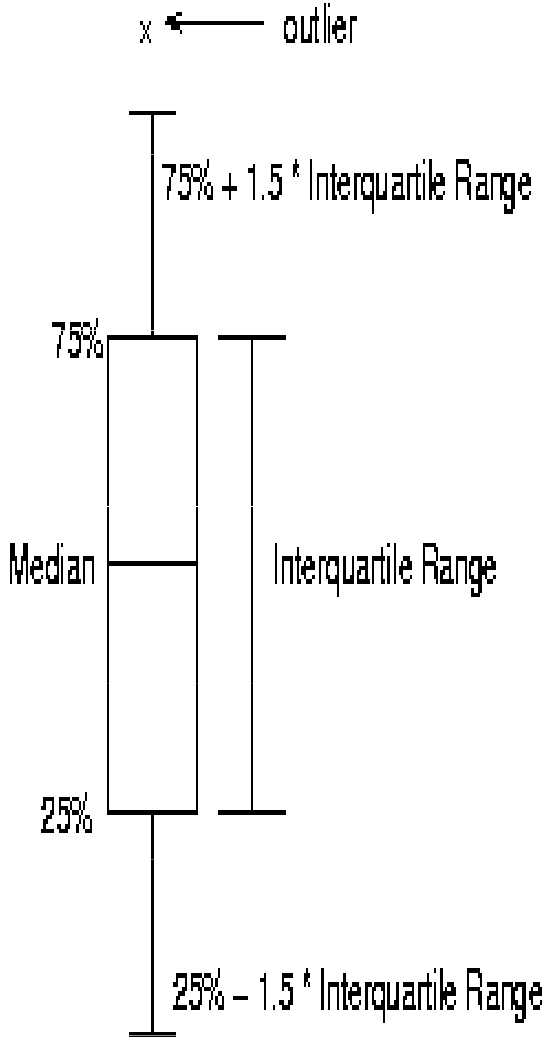
Ukazatele centrální tendence a variability - poznámky

- je třeba je umět spočítat ručně (a zopakovat si práci se sumačním symbolem Σ)
- i vážený průměr
- jak je ovlivní datové transformace přičtení konstanty a násobení konstantou
- vhodnost použití ukazatelů centrální tendence (Hendl s.95)

AJ: weighted mean, add, multiply

Boxplot – krabicový graf s anténami

- krabice je od Q_1 do Q_3
- v krabici se značí medián
- antény jsou X_{min} do X_{max}, **maximálně!** však 1,5x délka krabice (kvartilového rozpětí)
- hodnoty vzdálenější se značí jako body – odlehlé hodnoty
- hodnoty ještě vzdálenější (více než 3x délka krabice od Q_1 nebo Q_3) jsou někdy označovány jako extrémně odlehlé hodnoty



Volba popisných statistik

- Zvažujeme
  - úroveň měření
  - tvar rozložení – symetrie, normalita
  - cíl studie – pouze popis X usuzování, porovnávání
- Podle komunikačních cílů…
  - Je-li cílem především deskripce dat(=rozložení), pak použijeme **POŘADOVÉ** ukazatele. Připojíme-li i odchylkové, nic nezkazíme.
    - ***N, min, Q_1, Md, Q_3, max***
    - **boxplot**
    - pro individuální skóry **percentily**
  - Je-li cílem další usuzování, porovnávání apod., používáme **ODCHYLKOVÉ** ukazatele … pokud to úroveň měření dovoluje
    - ***N, m, s*** (*N, M, SD*)
    - popis rozložení
    - pro individuální skóry **z-skóry**

Myšlenka korelace

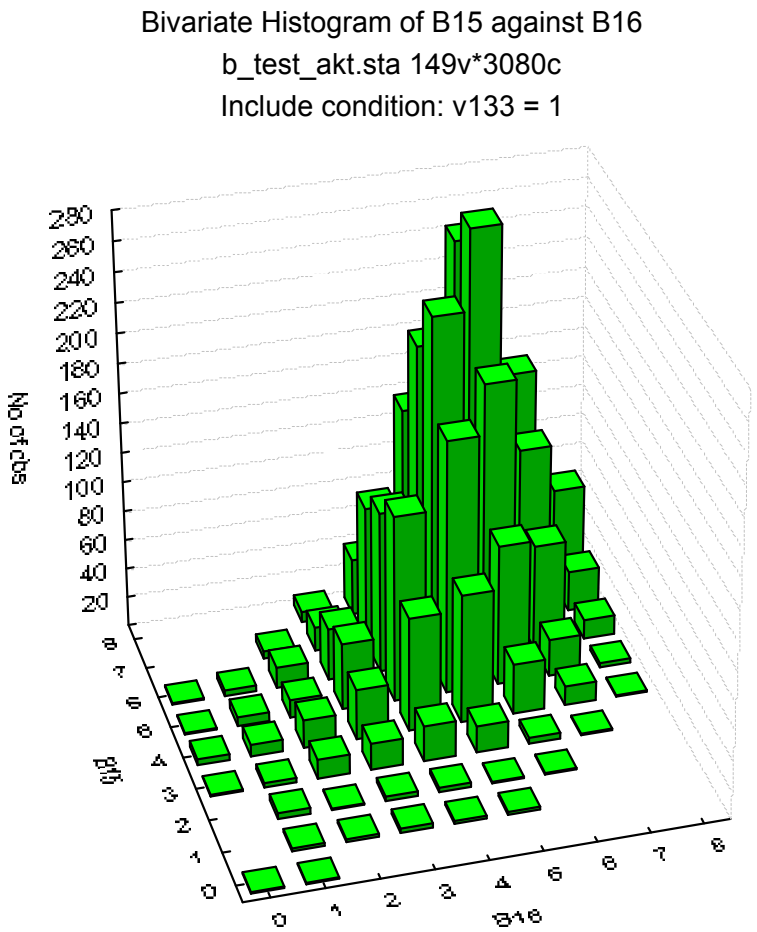
- Hypotézy o vzájemné souvislosti jevů:
  - Predikuje intelekt akademický úspěch?
  - Mají dobří češtináři i dobré známky z matematiky?
  - Existuje souvislost mezi mírou depresivní a anxiózní symptomatiky?
  - Jsou měsíční příjem a délka pracovní doby dobrými prediktory životní spokojenosti?
  - Jsou různá umělecká nadání specifická, nebo vycházejí ze stejného „všeobecného“ talentu?

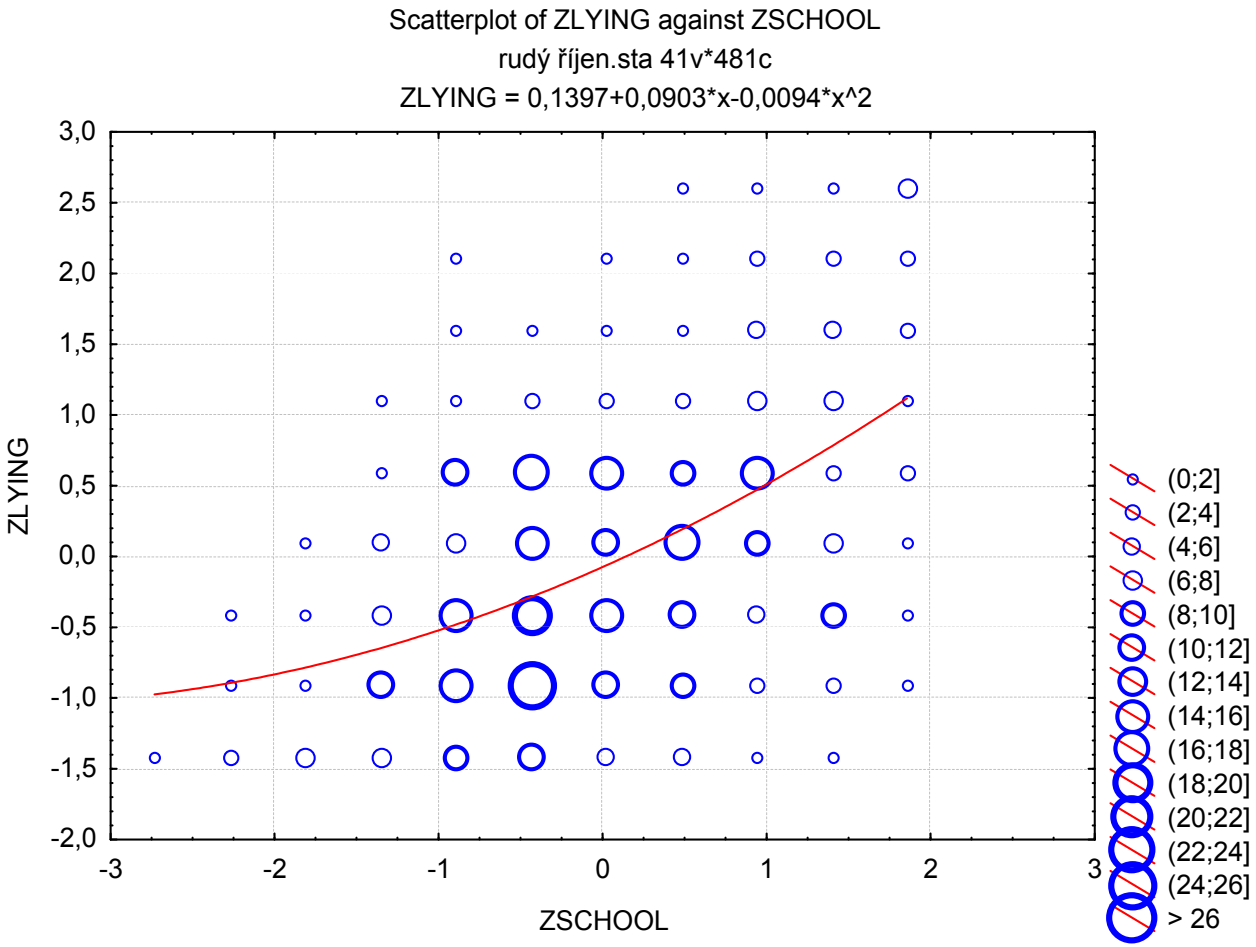
Kontingenční tabulka

		známka z matematiky					celkem
		1	2	3	4	5	
známka z čj	1	82	40	8	1	0	131
	2	71	200	73	17	0	361
	3	4	75	109	25	0	213
	4	1	7	23	24	1	56
	5	0	0	2	1	2	5
celkem		158	322	215	68	3	766

- Kontingenční tabulka…
  - Hodnoty je třeba přehledně uspořádat (stejně jako u tabulky četností)
  - Pro data všech úrovní měření, nejvhodnější pro diskrétní prom. s málo hodnotami
  - Buňky mohou obsahovat absolutní četnosti, rel. četnosti (řádkové, sloupcové, celkové)
  - Poslední sloupec/řádek obsahuje tzv. sloupcové/řádkové marginální (relativní) četnosti
  - Je grafickou podobou je trojrozměrného sloupcový diagramu či histogramu (může obsahovat i intervaly)
  - Relativně vysoké četnosti v jedné z diagonál naznačují lineární provázanost proměnných

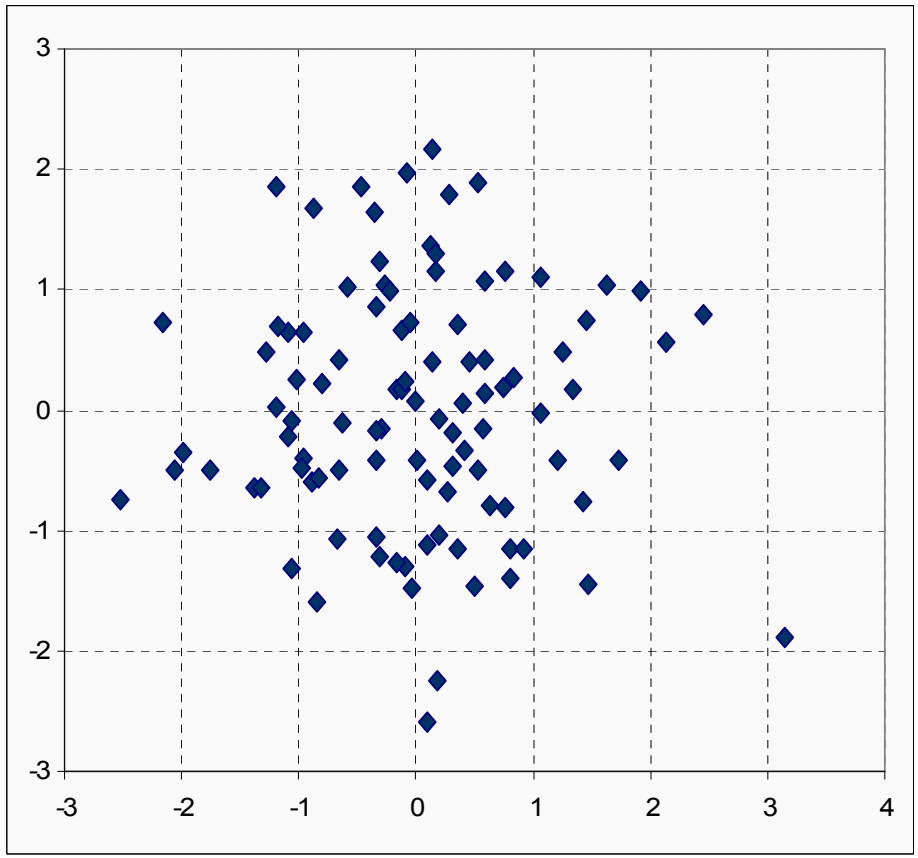
Grafická zobrazení dvourozměrného rozdělení



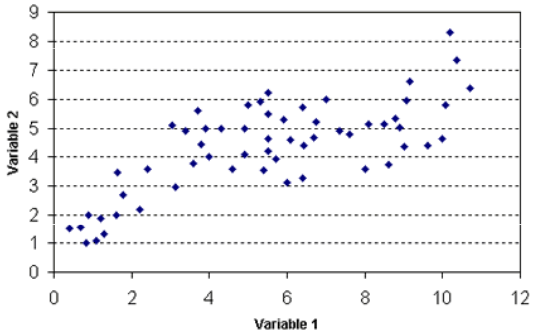


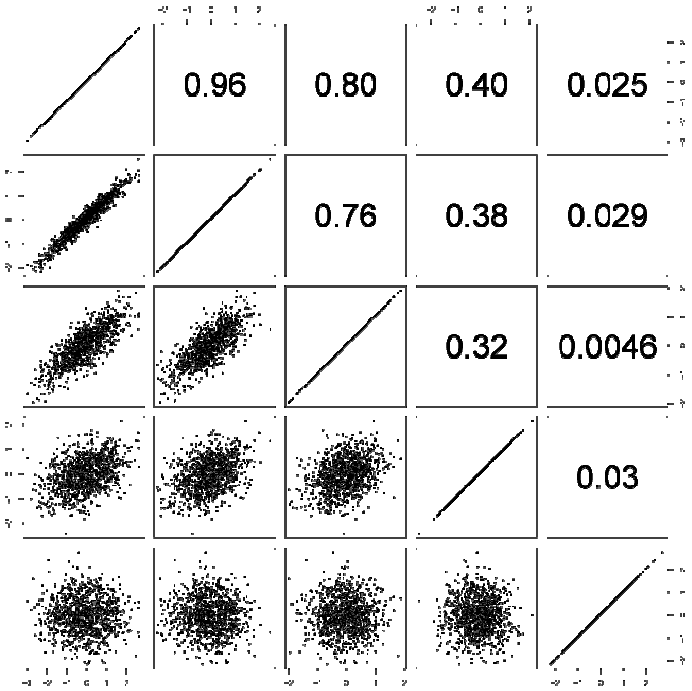
Bodový graf - scatterplot

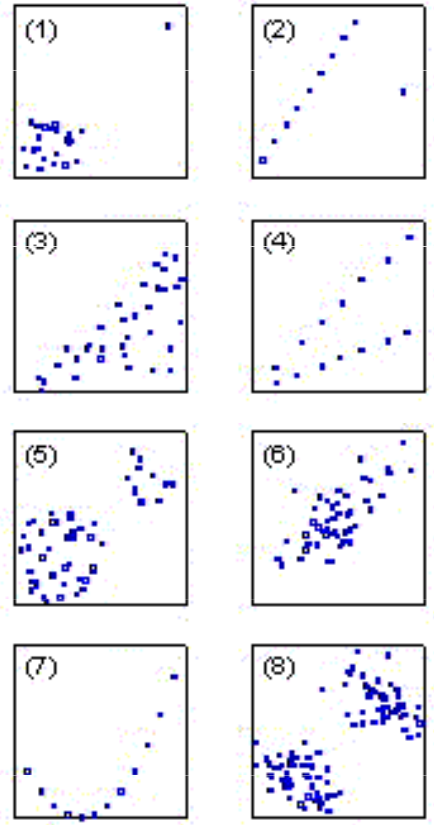
- Bodový graf – scatterplot
- Nahrazuje kontingenční tabulku, jsou-li obě proměnné spojité; pro proměnné s málo body měření nemá smysl
- Každá osa reprezentuje jednu proměnnou, každý bod je jedna zkoumaná osoba (jednotka)
- Poskytuje tím lepší evidenci o vztahu dvou proměnných…
  - …čím více měření jsme provedli
  - …čím přesnější jednotlivá měření byla
- Parametrem počtu měření může být např. velikost bodu
  - Frequency scatterplot

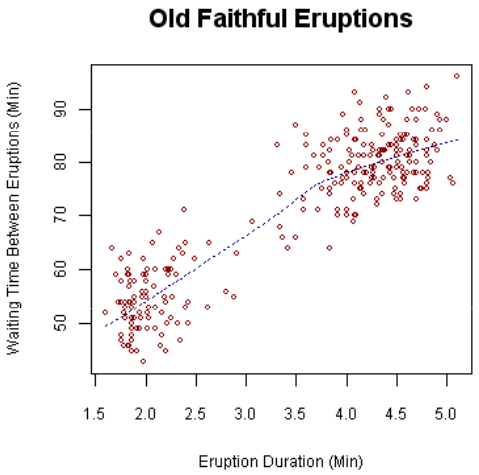


Různé podoby/druhy vztahu









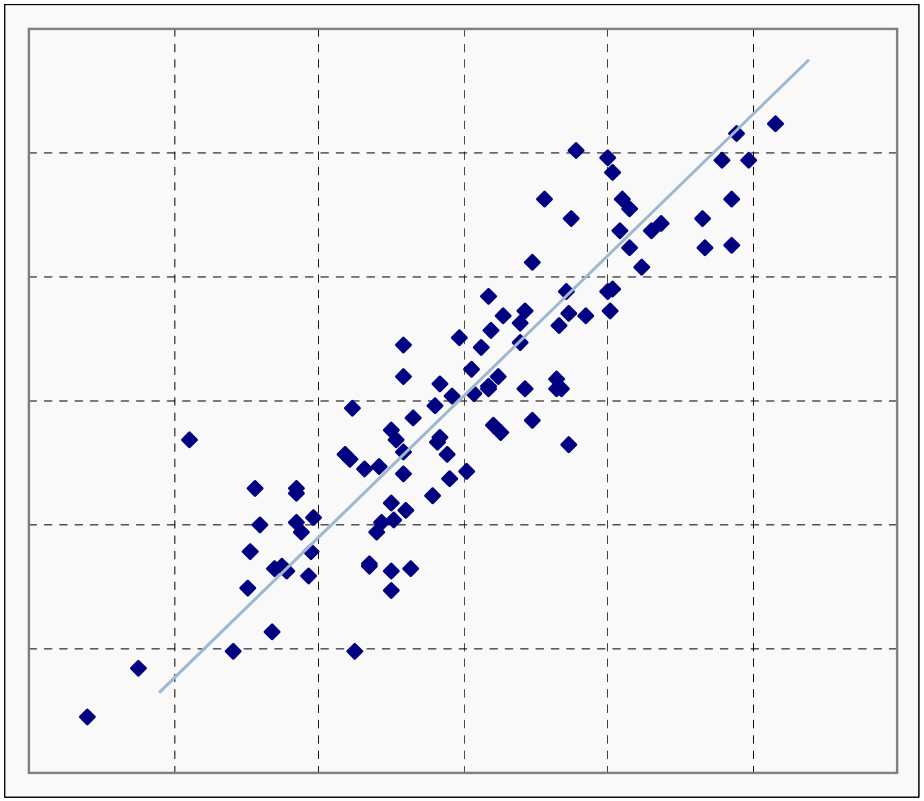
Pouze takto vypadající scattery zobrazují vztah mezi 2 proměnnými, který je lineární a dobře (=smysluplně, výstižně) popsatelný pomocí Pearsonova korelačního koeficientu. U ostatních jde buď o vztahy nelineární, nebo je problém v heterogenitě, outlierech…

Lineární souvislost, vztah

- Lineární vztah je to, co se obvykle míní slovem korelace.
- Je to monotónní vztah, který se dá popsat slovy čím více X, tím více/méně Y.
- Projevuje se tak, že scatterplot se dá proložit „ideální“ přímkou
  - y = ax + b

    Tato funkce/přímka popisuje strmost vztahu.

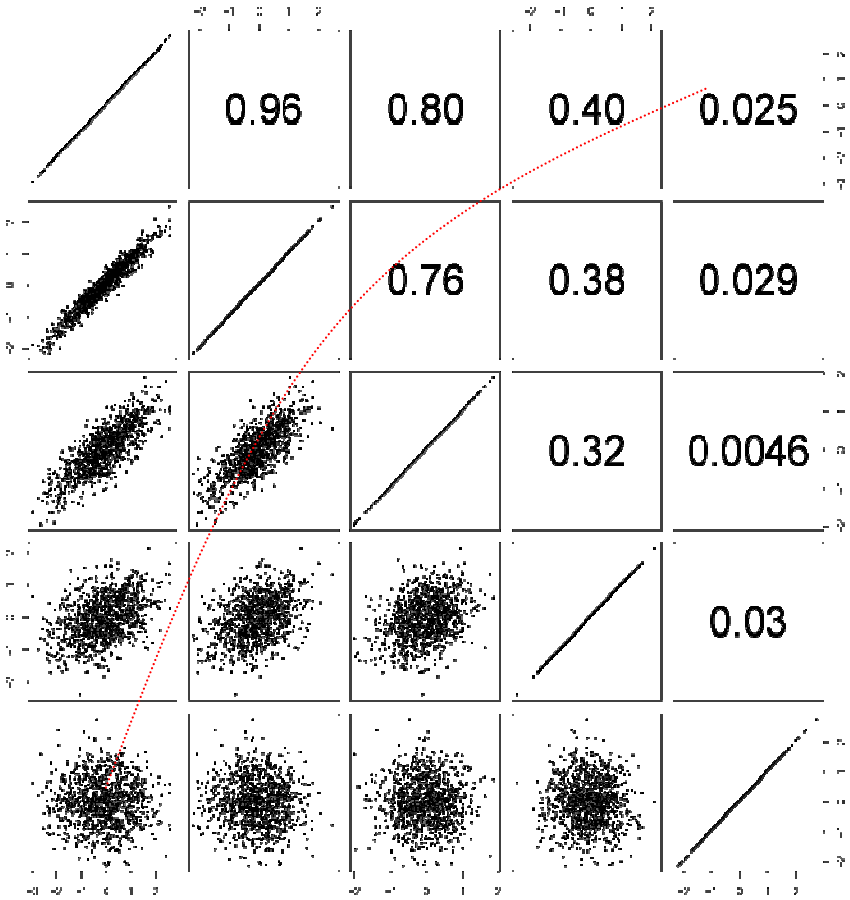
    Korelace popisuje **těsnost** vztahu.



AJ: linear association, correlation, monotonous relationship

Těsnost vztahu

- Čím těsnější (=intenzivnější, silnější) vztah 2 proměnných je, tím jsou body více nahuštěny okolo nějaké přímky
- Těsnost nesouvisí se sklonem té přímky, ale pouze s tím, jak moc se scatterplot podobá přímce.
- Těsnost se udává bezrozměrným číslem od 0 do 1, kde 0=žádný vztah(těsnost) a 1= maximální vztah
  (data na diagonále v obrázku napravo)
- Znaménko udává, zda jde o vztah čím víc, tím víc (+) nebo o vztah čím víc, tím míň (-)
- Rozsah je tedy od -1 do 1
- Těsnost -> kovariance



AJ: strength of association/relationship/correlation, positive relationship, negative(inverse) relationship

Kovariance (=sdílený rozptyl)

- Míru těsnosti lineárního vztahu dvou proměnných lze vyjádřit číselně
- Kovariance vypovídá o míře „sdíleného rozptylu“

$$c_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} x_i y_i$$

Vzpomeňte si na výpočet rozptylu. Ten byl $\Sigma x^2 / (n - 1)$. Tohle je $\Sigma xy / (n - 1)$. Místo x*x je tu x*y, proto je to ko-variance

Tato suma je tím vyšší čím máme v sadě dat více dvojic xy, u nichž je hodnota x i y nadprůměrná nebo podprůměrná. Sumu naopak snižují dvojice, kde je jedna hodnota nadprůměrná a druhá podprůměrná.

- kde *x, y* jsou deviační skóry, tj. odchylky od průměru
- Kovariance je stejně jako rozptyl nepraktická – výsledek je v jakýchsi „jednotkách na druhou“

AJ: covariance, shared variance

Korelace (=standardizovaný sdílený rozptyl)

- Chceme-li se zbavit obtížně interpretovatelných jednotek u kovariance, dosáhneme toho podobně jako při výrobě z-skórů – podělením deviačního skóru příslušnou směrodatnou odchylkou (=standardizace)

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_1 - m_x}{s_x}\right)\left(\frac{y_1 - m_y}{s_y}\right)$$

- Zakroužkovanou část vzorce už ale známe – to je transformace na z-skór. Korelace jednodušeji je tedy:

$$r_{xy} = \frac{\sum z_x z_y}{n-1}$$

AJ: correlation

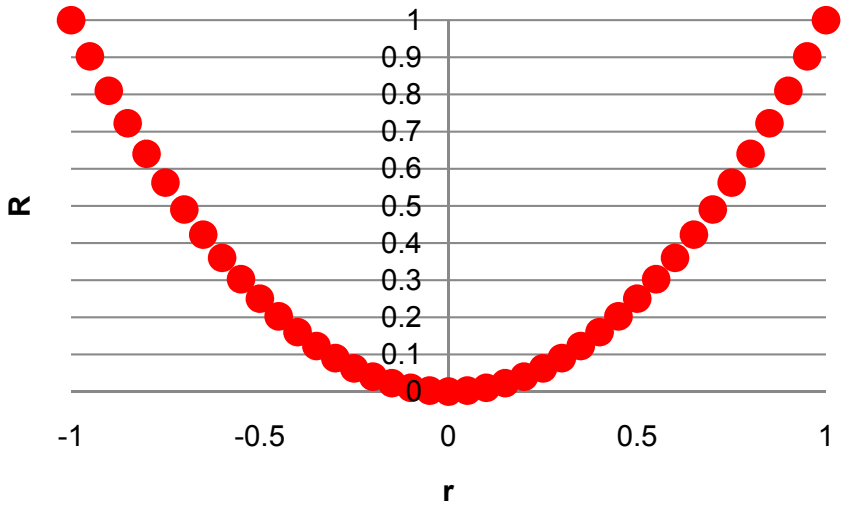
Vlastnosti popsaného koeficientu korelace I.

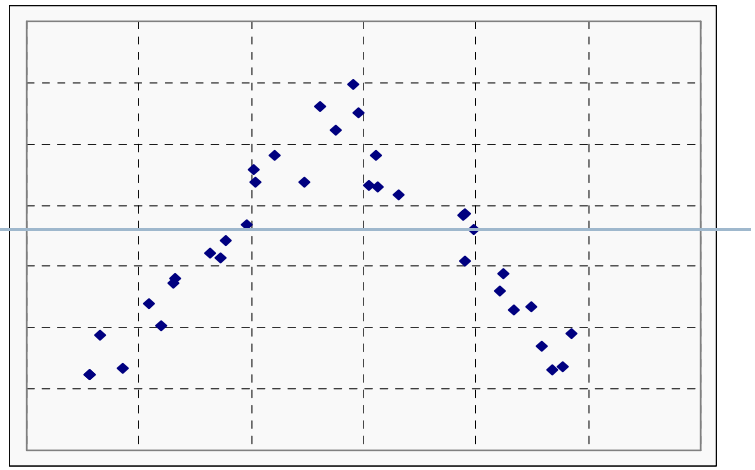
- Jde o tzv. Pearsonův součinový, momentový koeficient korelace
  - patří tedy do kategorie momentových ukazatelů (viz předchozí přednáška) a platí pro něj podobné věci:
    - nutná intervalová a vyšší úroveň měření
    - velký vliv odlehlých hodnot na výsledek
    - je vhodný pro popis normálně rozložených proměnných
    - vyjadřuje pouze sílu(těsnost) lineárního vztahu
- Nabývá hodnot v rozmezí -1 až 1
  - 0 = žádný vztah
  - 1(-1) = dokonalý kladný (záporný) vztah; identita proměnných
- Korelace nepopisuje funkční vztah dvou proměnných, ale pouze jeho směr a těsnost.

Vlastnosti Pearsonova koeficientu korelace II.

- Je vázán na homogenitu souboru
- Není aditivní
- r^2 = R = koeficient determinace (někdy D)
  - = proporce sdíleného rozptylu
  - V důsledku toho:

    0,3-0,1 ≠ 0,7-0,5
- r = 0 neznamená, že mezi rozděleními proměnných není žádná souvislost, znamená pouze, že mezi nimi není *lineární* vztah.

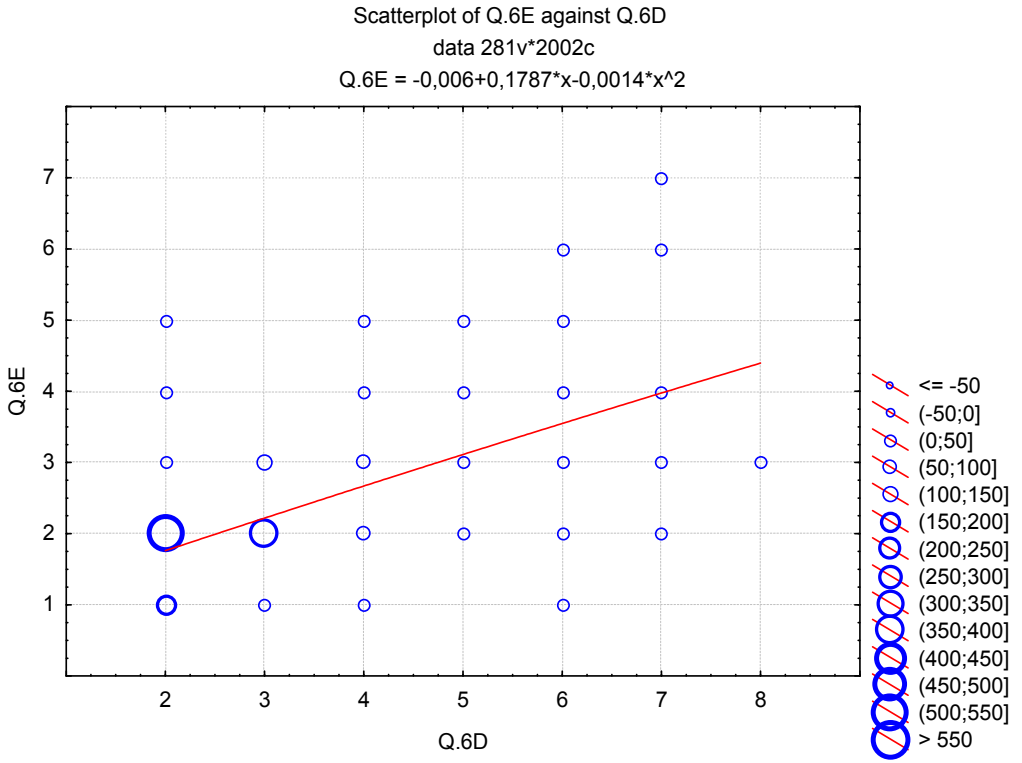




AJ: sample/population homogeneity, additivity, coefficient of determination

Vlastnosti Pearsonova koeficientu korelace III.

- Kdy nemá korelace smysl?
  - V1: Kolik hodin denně sledujete televizi?
  - V2: Kolik hodin denně sledujete televizní zpravodajství?
  - Proč? ☺
- Korelace proměnných se společnou příčinnou:
  - Swoboda: platy kněžích a ceny vodky v průběhu doby korelují!
  - IQ dětí a velikost a jejich výška prý také…
  - … kovariance proměnných se společnou příčinnou je základem dalších metod analýzy dat v psychologii: analýzy reliability a faktorové analýzy.



Korelační koeficienty pro pořadová data

- vhodné nejen pro pořadová data, ale i pro intervalová, která mají rozložení výrazně odlišné od normálního
- zachycují i nelineární monotónní vztahy (viz Hendl, s260)
- ukazatele toho, nakolik jsou pořadí podle korelovaných dvou proměnných stejná
- Spearmanův koeficient rhó – ρ, r_s
  - založený na velikosti rozdílů v pořadí
  - ekvivalentem Pearsonova koeficientu na pořadových datech
  - lze interpretovat r^2
- Kendallův koeficient tau – τ (s variantami „b“ nebo „c“)
  - založený na počtu hodnot (prvků výběrového souboru) mimo pořadí
  - vyjadřuje spíše pravděpodobnost, že se prvky výběrového souboru uspořádají podle obou proměnných do stejného pořadí

AJ: Spearman (rank correlation) rho, Kendall tau (-b,-c), rank

Korelační koeficienty další

- korelačních koeficientů existuje velké množství
- specifická užití – např. ϕ
- zjednodušení ručních výpočtů – např. r_{pb}
- ještě budeme mluvit o vztazích mezi nominálními proměnnými…

!! Korelace neznamená kauzalitu, jde spíše o koincidenci !!

AJ: phi, point-biserial correlation

Využití korelací v konstrukci psychologických testů

- Položky lze sčítat, pokud spolu korelují.
- Položky korelují, existuje-li společný důvod pro určitý způsob odpovídání na ně – měřená charakteristika.

Jak moc spolu musí korelovat?

$$r_{tt} = \frac{k r_M}{1 + (k-1) r_M} \qquad r_{tt} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_t^2}\right)$$

r_{tt} je vnitřní konzistence, r_M je průměrná korelace mezi položkami, k je počet položek

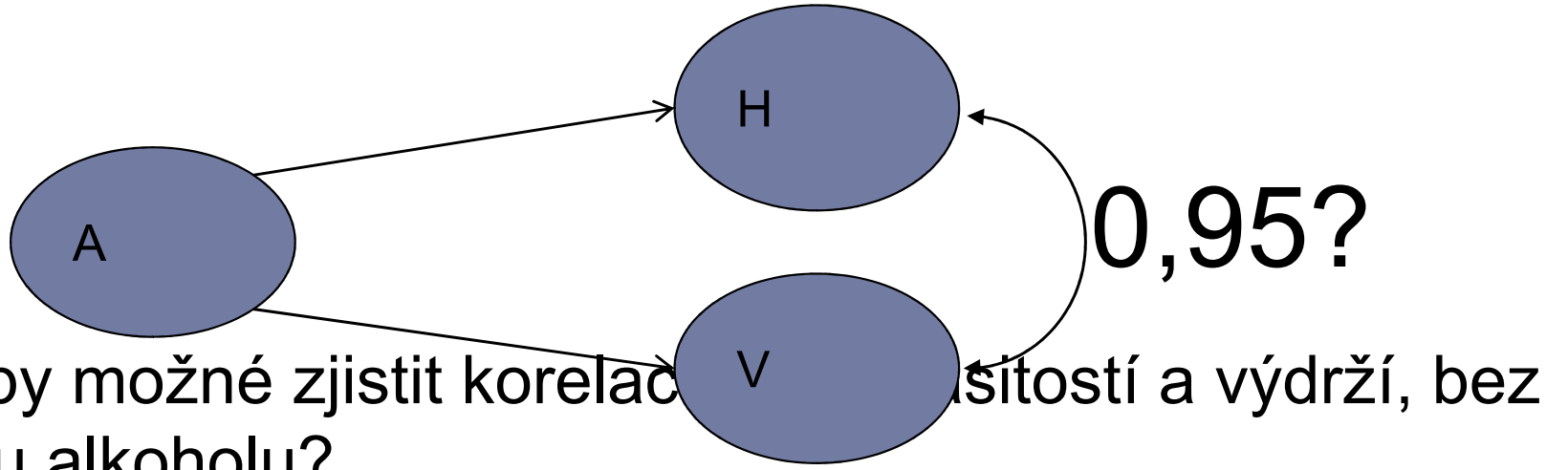
- při 10 položkách stačí průměrná korelace 0,2

Vnitřní konzistence – **Cronbachovo** α – horní mez reliability

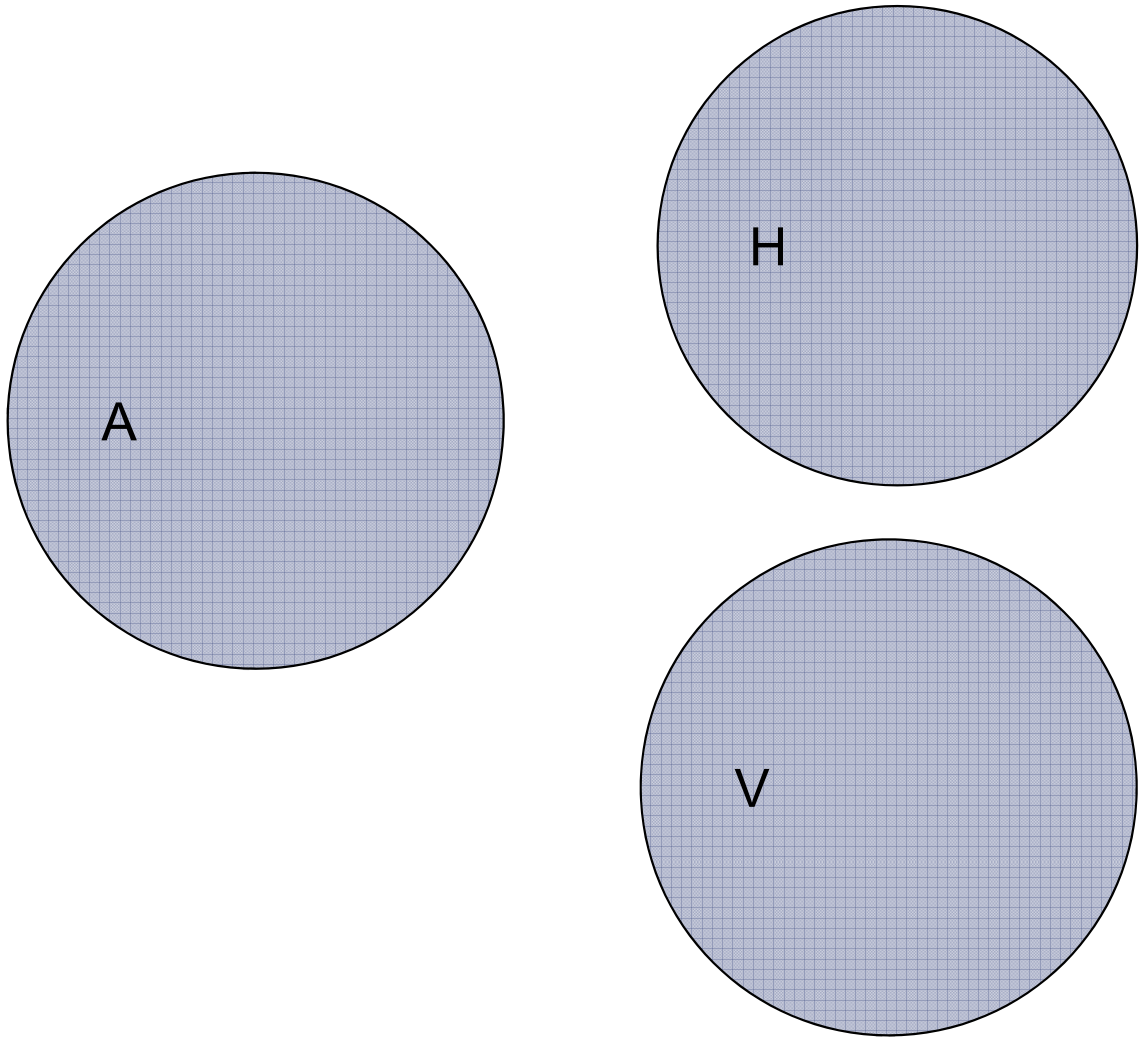
- minimálně 0,7 pro výzkum, 0,9 pro diagnostiku

Vztah mezi třemi proměnnými
Parciální a semiparciální korelace

Zjistili jsme, že účastníci našeho experimentu se nám opili.
To nám vadí, protože opilost snižuje citlivost na podněty a zvyšuje obě naše proměnné.



Bylo by možné zjistit korelac sitostí a výdrží, bez vlivu alkoholu?


A
H
V

Jak ale rozdělovat ty rozptyly?

Regrese dělí proměnnou na sdílený rozptyl a reziduální rozptyl….

Parciální korelace $r_{HV.A}$

- Uděláme regresi výdrže na alkohol – reziduum výdrže bez alkoholu
- Uděláme regresi hlasitosti na alkohol – reziduum hlasitosti bez alkoholu
- Korelace dvou reziduí je PARCIÁLNÍ KORELACE

$$r_{HV.A} = \frac{r_{HV} - r_{HA}r_{VA}}{\sqrt{1-r_{HA}^2}\sqrt{1-r_{VA}^2}}$$

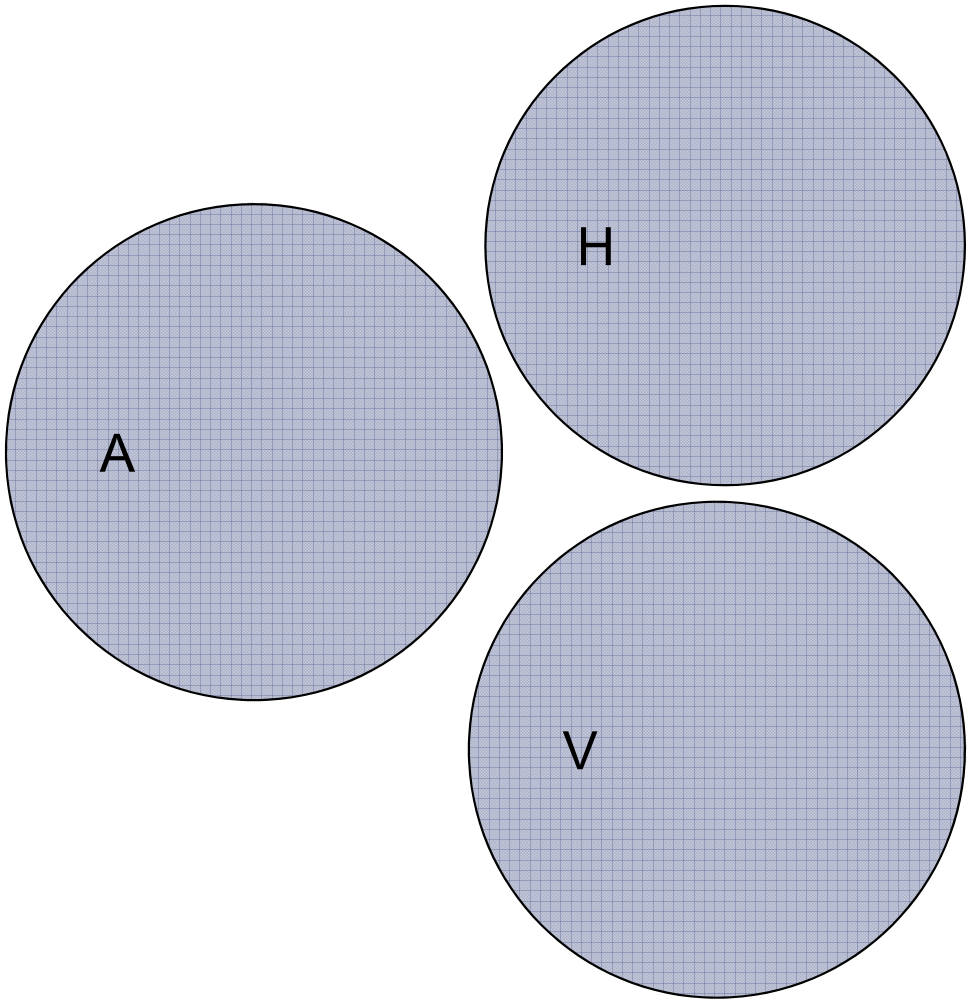
Semiparciální korelace $r_{H(V.A)}$

- Korelace rezidua (V.A) se závislou proměnnou (H)

$$r_{H(V.A)} = \frac{r_{HV} - r_{HA}r_{VA}}{\sqrt{1-r_{VA}^2}}$$

AJ: partial correlation, part (semi-partial) correlation

Korelace mezi hlasitostí a výdrží , **kontrolujeme-li statisticky*** alkohol je…



	hlasitost	vydrz	alkohol
hlasitost	1,000	,949**	,864**
vydrz	,949**	1,000	,902**
alkohol	,864**	,902**	1,000

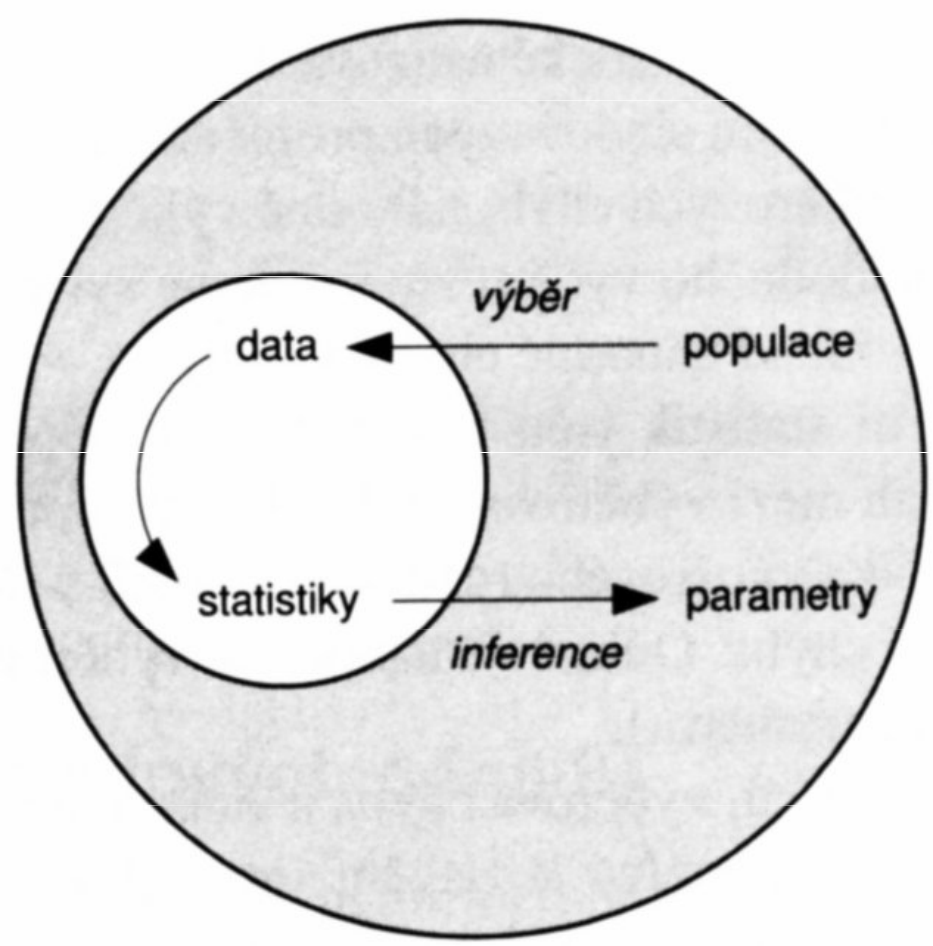
$r_{HV.A} = 0,78$

* Též, „pokud by alkohol byl konstantní“

Statistické usuzování, odhady

Věci, které můžeme přímo pozorovat, jsou téměř vždy pouze vzorky.
Alfred North Whitehead

Výběr – od deskripce k indukci



- Deskripce dat, odhad parametrů
- Usuzování = inference = indukce
- Počítá se s náhodným výběrem
  - tj. výběr jedince splňuje podmínky náhodného pokusu
  - není-li výběr v pravém slova smyslu náhodný, uvažujeme, v čem se p-dobně liší od náhodného

AJ: statistical description, inference, population, sample, data, statistics, inference, parameters, random sample (sampling)

Statistiky a parametry

- Na vzorku (datech) počítáme **statistiky**
- Hodnotě statistiky v celé populaci říkáme **parametr**.
  - Pro parametry používáme odpovídající písmena řecké abecedy
    - např. průměr: statistika *m*, parametr μ (mí)
    - další: *s* – σ (sigma), *r* – ρ (ró), *d* – δ (delta - rozdíl)
- Statistiky jsou **odhady** parametrů
  - tj. jsou vždy zatíženy chybou – **výběrovou chybou**
  - *chyby náhodné* – umíme spočítat, známe-li **výběrové rozložení**
  - *chyby systematické* – nevhodné statistiky, špatné měření, špatný způsob výběru vzorku (metodologie)

Jak dobré jsou tyto odhady?

AJ: estimates, sampling error. random error, systematic error, sampling distribution

Výběrové rozložení a sm. chyba

- Spočítáme-li tutéž statistiku na mnoha nezávislých náhodných vzorcích
  - získáme mnoho různých odhadů parametru
  - tyto odhady mají nějaké rozložení - **výběrové rozložení**

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

- **Výběrové rozložení** obvykle můžeme popsat
  - průměrem – ten se u dobrých statistik blíží hodnotě **parametru**
  - směrodatnou odchylkou – říkáme jí **směrodatná chyba** ((odhadu) parametru) nebo také střední chyba a obecněji i výběrová chyba
  - Čím je velikost vzorku/ů větší, tím je směrodatná chyba menší

AJ: sampling distribution, standard error (of the mean)

Výběrové rozložení (odhadu) průměru

Odhad průměru má přibližně **normální rozložení**,

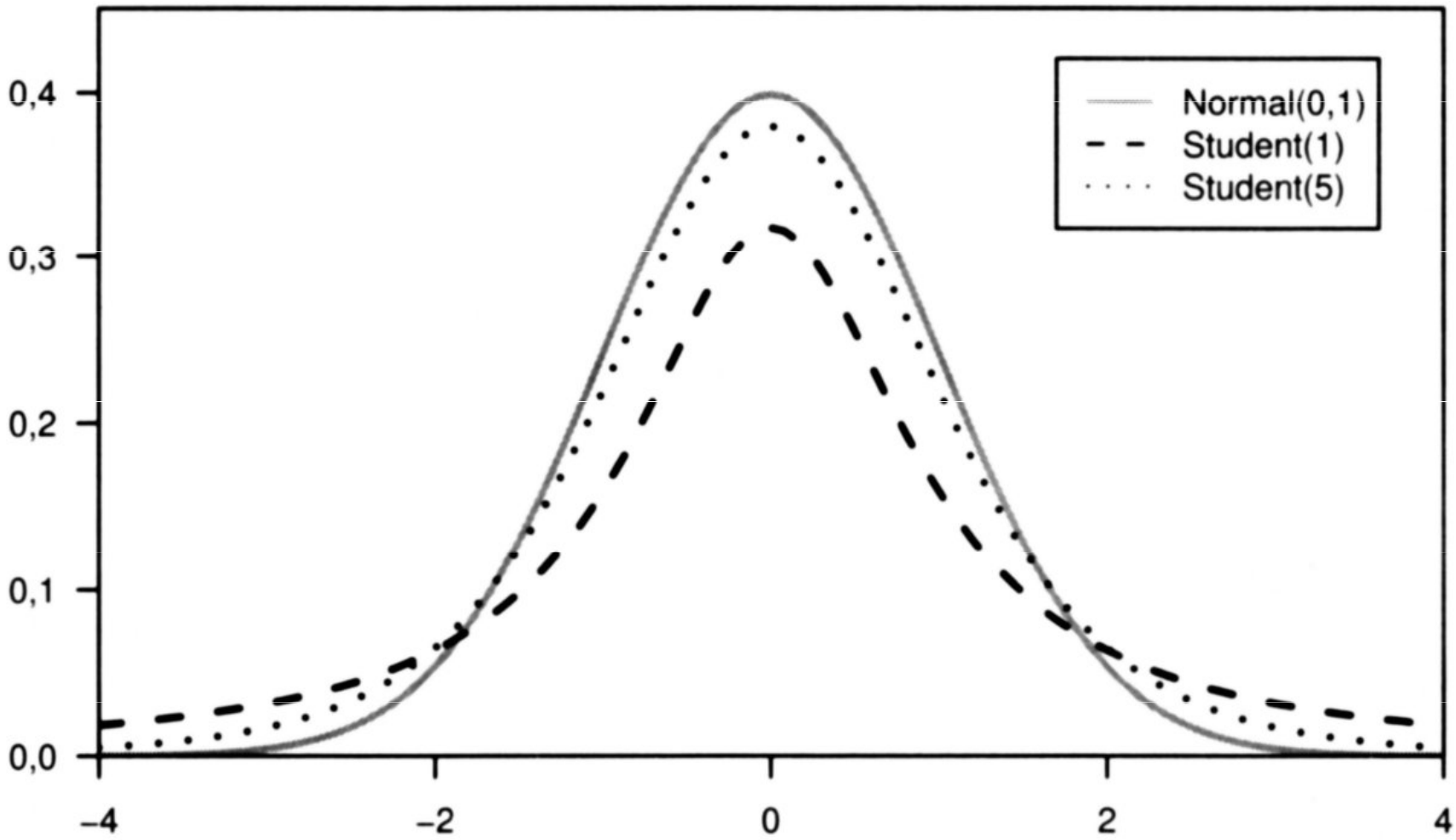
- jehož průměr je μ se směrodatnou chybou …….....$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$
- Platí to i tehdy, když rozložení proměnné není normální.
  - a to „díky“ **centrálnímu limitnímu teorému**
- Jenomže my obvykle neznáme σ…

Neznáme-li σ, musíme použít *s*

- průměr zůstává μ, směrodatná chyba je nyní …….. $s_{\bar{x}} = \frac{s}{\sqrt{N}}$
- výběrové rozložení není normální, jde o **Studentovo *t* -rozložení**
  - jako normální s těžšími konci (*t* je pro t-rozložení totéž, co *z* pro normální rozložení)
  - má různé tvary pro různá *n* : stupně volnosti – ν (ný)
    - zde $\nu = N-1$; čím vyšší *N*, tím se *t*-rozložení blíží normálnímu

AJ: central limit theorem, Student's t-distribution, degrees of freedom (d.f.)

Studentovo t -rozložení



Výběrové rozložení dalších statistik

Nyní je tedy třeba ke každé popisné statistice znát ještě další vlastnost – její teoretické **výběrové rozložení**

- relativní četnost – přibližně normální - Hendl 156
- rozptyl – po transformaci χ^2-rozložení (chí kvadrát) - Hendl 159
- Pearsonova *r* – po Fisherově transformaci normální – Hendl 252

Teoretická výběrová rozložení různých statistik jsou různá

- Statistika je obvykle transformována do podoby, která má jedno z běžných teoretických rozložení: normální, chí-kvadrát rozložení (Pearsonovo), *t*-rozložení (Studentovo), *F*-rozložení (Fisherovo, Snedecorovo)
- Netřeba je znát z hlavy, programy je používají za vás, ale stojí za to vědět, že existují přehledy – např. Receptář Oseckých nebo Sheskin ISBN 1584884401
- Pro interpretační potřeby si obvykle vystačíme s představou výběrového rozložení průměru
- Pozor, centrální limitní teorém se týká pouze výběrového rozložení průměru!

AJ: chi-square distribution, F-distribution

Bodové vs. intervalové odhady

α je p-nost chyby a proto je hladina spolehlivosti 1-α, tj. 95% spolehlivost znamená 5% chybovost: (1-0,05)

Parametr se můžeme snažit odhadnout…

- **bodovým odhadem** – tj. odhadujeme přímo hodnotu parametru, např. průměr. Kvalita bodového odhadu viz Hendl 169.
- **intervalovým odhadem** – tj. odhadnutím intervalu, který parametr s určitou p-ností zahrnuje
  - výsledkem intervalového odhadu je **interval spolehlivosti**
  - interval spolehlivosti tvoříme z bodového odhadu a znalosti jeho výběrového rozložení, tj. (bod±odchylka)
  - intervalový odhad lepší - více informací
  - té p-nosti se v tomto kontextu říká **hladina spolehlivosti** (1-α)
    - typicky se používá 95% a 99% hladina spolehlivosti
    - pak říkáme, že hledaný parametr je s 95% p-ností v intervalu spolehlivosti

$$(1-\alpha)\ CI = \bar{X} \pm z_{1-\alpha/2}\sigma_{\bar{X}}$$

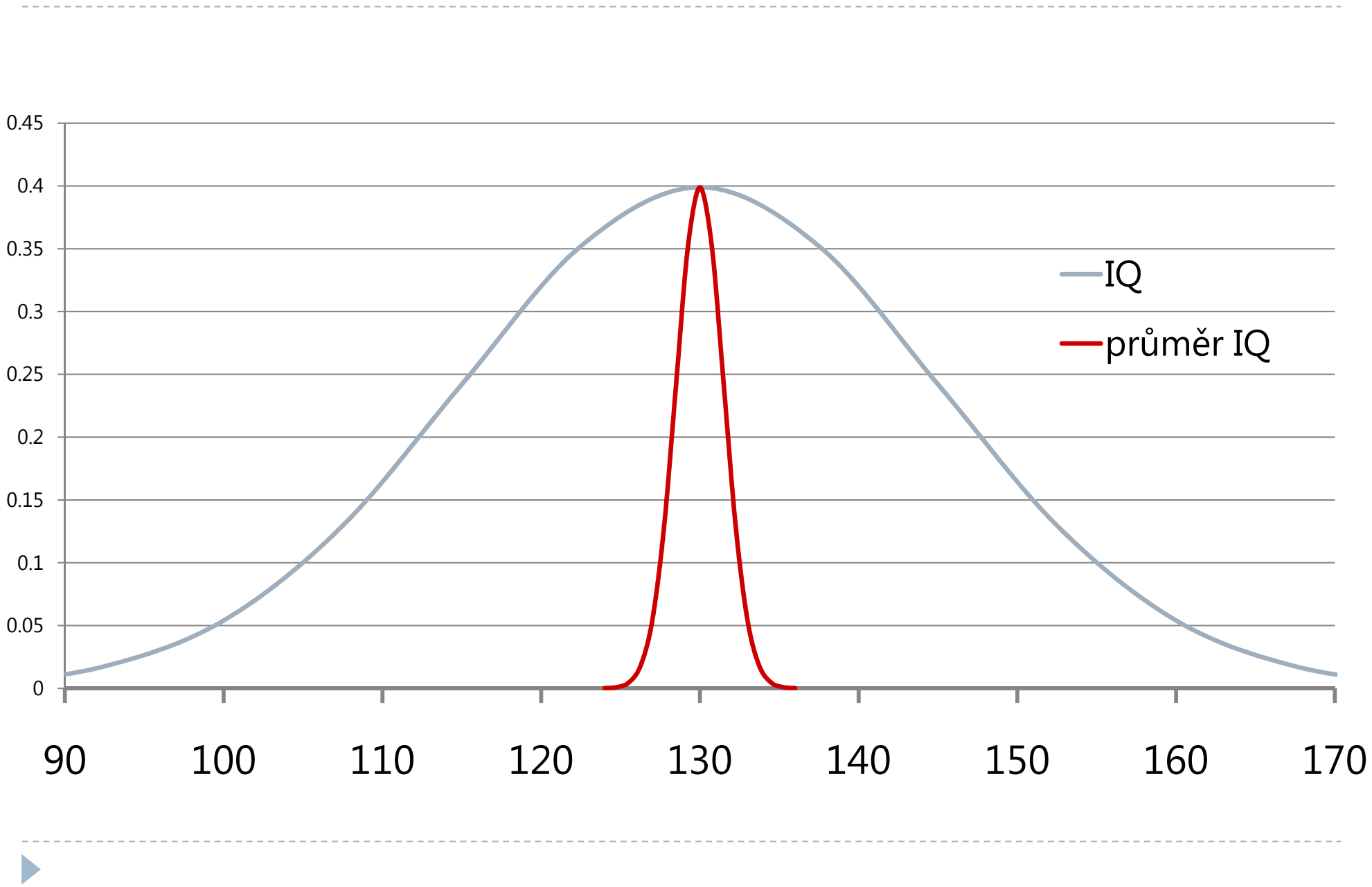
Zkuste si sami: http://onlinestatbook.com/stat_sim/conf_interval/index.html

AJ: point estimate, interval estimate, confidence interval (CI), level of confidence, consistency, unbiasedness, relative efficiency, resistence

Příklad konstrukce intervalu spolehlivosti pro průměr 1

Na vzorku dětí (*N*=100) s různobarevnýma očima jsme spočítali průměrné IQ 130, přičemž víme, že σ=15.

- **bodový odhad** průměrného IQ v populaci dětí s různobarevnýma očima (tj. parametru, μ) je 130
- **intervalový odhad**
  - Známe-li σ, výběrové rozložení průměru má **normální rozložení…**
  - …se středem v μ. μ neznáme, a tak použijeme bodový odhad $m = 130$
  - … se směrodatnou chybou odhadu průměru $s_m = \sigma/\sqrt{N} = 15/\sqrt{100} = 1{,}5$.
  - Zvolíme-li hladinu spolehlivosti $1\text{-}\alpha = 95\%$,
  - pak v tabulkách/Excelu zjistíme, že 95% normálního rozl. je mezi hodnotami $z = -1{,}96$ a $1{,}96$,tj. $_{1-\alpha/2}z = {}_{0{,}975}z = 1{,}96$, Excel: =NORMSINV(0,975)
  - interval spolehlivosti: $(m - 1{,}96s_m;\ m + 1{,}96s_m) = (127{,}1\ ;\ 132{,}9)$,
  - **tj. s 95% pravděpodobností 127,1 ≤ μ ≤ 132,9**


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
90
100
110
120
130
140
150
160
170
IQ
průměr IQ

Příklad konstrukce intervalu spolehlivosti pro průměr 2

Na vzorku dětí (*N*=100) s různobarevnýma očima jsme spočítali průměrné IQ 130 a *s* =15.

- **bodový odhad** průměrného IQ v populaci dětí s různobarevnýma očima (tj. parametru, μ) je 130
- **intervalový odhad**
  - střed intervalu spolehlivosti bude na bodovém odhadu, tj. $m = 130$
  - víme, že výběrové rozložení průměru má *t*–rozložení se stupni volnosti $\nu = N-1 = 99$
  - zvolíme-li hladinu spolehlivosti $1-\alpha = 95\%$,
  - pak v tabulkách (Excelu) zjistíme, že 95% *t*-rozložení je mezi hodnotami t=-1,98 a 1,98 (tj. ${}_{1-\alpha/2}t\,(\nu) = {}_{0,975}t\,(99) = 1,98$ excel: TINV(0,05;99))
  - směrodatná chyba odhadu průměru $s_m = s/\sqrt{n} = 15/\sqrt{100} = 1,5$
  - interval spolehlivosti: $(m - 1,98s_m;\ m + 1,98s_m) = (127,0\ ;\ 133,0)$,
  - **tj. s 95% pravděpodobností $127,0 \le \mu \le 133,0$**

pozor na tento rozdíl: ve středu intervalu je *m*, někde v intervalu je v 95% případů μ

Interpretace intervalu spolehlivosti

- … je prostá, avšak zrádná
- 95% interval spolehlivosti znamená, že sestrojujeme-li tento interval dle výše uvedených instrukcí, **v 95% případů sestrojení intervalu tento interval zahrnuje odhadovaný parametr**, tj. v 95% případů je závěr, že μ je mezi čísly *a* a *b*, správný.
- V tomto smyslu to také znamená, že máme subjektivní 95% jistotu, že parametr je v námi určeném intervalu.
- V konkrétním případě, kdy jsme spočetli konkrétní interval spolehlivosti ($127 \leq \mu \leq 133$), to neznamená, že v 95% případech je μ v intervalu od 127 do 133.
  - To proto, že μ je konstanta; při opakovaných výzkumech se nemění. Díky omylnému výběru v každém výzkumu vychází poněkud jiný interval sestrojený podle jiného výběrového průměru. Jinými slovy, trefujeme se obručí na kolík a ne kolíkem do obruče.
- O čem tohle slovíčkaření je? O rozdílu mezi četnostním a subjektivním (Bayesovským) pojetím pravděpodobnosti.

Od vzorku k populaci a zpět

Vzhledem k tomu, jaká nám na vzorku vyšla statistika, jaký je odpovídající populační parametr?

interval spolehlivosti

Pokud předpokládáme, že v populaci je hodnota parametru X, co si myslet o své hypotéze poté, co nám na vzorku vyšlo Y?

statistický test hypotézy

Hypotézy

- Příklady (statistických) hypotéz
  - H: $\mu = 100$: Populační průměr IQ je roven 100.
  - H: $\sigma = 10$: Populační směrodatná odchylka je 10.
  - H: $\mu_1 - \mu_2 = 0$: Populační průměry μ_1 (psychotici) a μ_2 (zdraví) jsou stejné.
  - H: $\rho_{xy} = 0$: Proměnné X (pití piva) a Y (dominance) spolu nekorelují
- Vezměme si tu první hypotézu konfrontujme s daty:
  - Na vzorku 1000 náhodně vybraných dospělých jsme zjistili průměrné IQ rovné 105 (s =14).

AJ: statistical hypotheses testing, hypothesis, hypothesis supported by data

Statistický test hypotézy

Statistické testování založeno na p-nosti

- Známe-li pravděpodobnostní rozložení statistik můžeme usuzovat, **jak pravděpodobná je určitá výběrová statistika vzhledem k hypotéze: $P(D|H)$**
  - D : např. m=9,78
  - H : např. μ =10, $P(D|H)$ je $P(m=9{,}78 \mid \mu=10)$
- Je-li $P(D|H)$ vysoká, je tím hypotéza podpořena.
- Je-li $P(D|H)$ nízká, je tím hypotéza „činěna méně p-nou“

- Jak „vysoká$_{\text{nízká}}$“ je vysoká$_{\text{nízká}}$ pravděpodobnost, abychom hypotézu podpořili$_{\text{vyvrátili}}$?

Jak vysoká $P(D \mid H)$ je nutná k přijetí H?

- Bayesovský přístup – otázka není relevantní
  - s H je spojena určitá p-nost a ta se díky $P(D \mid H)$ zvyšuje či snižuje
  - Bayesův teorém – $P(H \mid D) = P(H) * P(D \mid H) / P(D)$

- Fisher, Pearson, Neyman – otázka je relevantní
  - Popper – princip falzifikace – H nelze potvrdit, pouze vyvrátit
  - My ale nechceme své hypotézy vyvracet, spíš potvrzovat
  - P-N: princip vzájemně se doplňujících konkurenčních hypotéz
    - Vytvořme takovou H, kt. bude logickou negací naší vědecké hypotézy a říkejme jí **nulová H**. Když se nám podaří nulovou H vyvrátit, znamená to **jakousi podporu** pro naší vědeckou hypotézu.
  - Vyvrácení H_0: $P(D \mid H_0) <$ **0,05**; **0,01**; 0,001; 0,0001 podle zvyku

Terminologická vložka

H_0 : **nulová (statistická) hypotéza**

- logická negace (doplněk) vědecké hypotézy

H_1 : **alternativní (vědecká, výzkumná)hypotéza**

- ta, o kterou nám primárně jde; $P(H_0 \cup H_1) = 1$

$P(D \mid H_0)$, kdy H_0 zamítáme:

- značí se i ***p*** nebo ***Sig.***
- p-nost chybného zamítnutí H_0 - **chyba prvního typu**
- Je-li stanovena dopředu: **úroveň/hladina statistické významnosti** (průkaznosti), α, udává se často v procentech: 5%, 1%
  - chyba, jejíž velikost jsme ochotni tolerovat

Jednostranné vs. **oboustranné** hypotézy

- jednostranné, směrové: $\mu \geq 23$, $\mu \leq 0$, z různých důvodů se jim vyhýbáme
- oboustranné: $\mu = 23$

AJ: null hypothesis, scientific/alternative hypothesis, level of statistical significance, type I error, one-tailed, two-tailed, directional

Postup testování statistické hypotézy

1. Formulujte **statistickou hypotézu**, kterou budete testovat (vyvracet) (H_0: $\mu = 0$)
2. Zvolte **hladinu statistické významnosti**, tj. míru rizika, že dojde k chybě 1. typu (např. $\alpha = 0{,}05$)
3. Hledáme p-nost získání naší výběrové statistiky nebo extrémnější hodnoty, za předpokladu, že H_0 je pravdivá: $P(D|H_0)$, p, Sig.
   - cesta vede přes znalost výběrového rozložení statistiky
   - např. $m = 0{,}5$. $P(|m|=0{,}5|\mu=0)$
   - obvykle je nutný přepočet na tzv. testovou statistiku, např. *t, z…*
4. Vyneseme rozhodnutí o H_0: zamítnutí či přijetí
   - je-li $P(D|H_0) < \alpha$, pak H_0 zamítáme
   - je-li $P(D|H_0) \geq \alpha$, pak H_0 nezamítáme

Příklad – jednovýběrový *t*-test

Terapie nevhodného chování.

- Rozdíl před-po: m=2,7; s=3,5; N=10
- H : Terapie má efekt. ($\mu \neq 0$)

1. H_0 : Terapie nemá efekt: $\mu = 0$
2. V sociálních vědách běžně α=0,05
3. $P(|m| \geq 2{,}7 | \mu = 0) = ?$
   - s_m=3,5/odm(10)=1,1
   - $t = (m-\mu)/s_m$=2,7/1,1= 2,45
   - $P(|t| \geq 2{,}45 | \tau = 0)$ = TDIST(2,45;9;2) = 0,04
4. $P(|m| \geq 2{,}7 | \mu = 0) < 0{,}05$ >> zamítáme H_0

Protože při m =2,7 je velmi málo pravděpodobné, že by rozdíl byl 0, tak připouštíme, že nějaký rozdíl je.

Dichotomizace výsledků výzkumu

- Výsledek výzkumu je testováním zredukován na ano-ne

	H_0 přijata	**H_0 zamítnuta**
H_0 pravdivá (žádný efekt)	OK	chyba 1. typu α *(její pravděpodobnost)*
H_0 nepravdivá (efekt)	chyba 2. typu β	OK Síla (1-β)

Čím nižší je α, tím vyšší je β. Přesná podoba vztahu závisí na použitém testu. α i β mohou být nízké pouze při vysokých *n*.

Síla testu viz Hendl 401-411.

AJ: type-I error, type-II error, (statistical) power

Problémy statistického testování H

- Největší problém: dichotomizace
  - stejná velikost efektu dává při různých N jiné rozhodnutí o H_0
  - komplikuje až znemožňuje kumulativní budování znalostní báze
- Problém interpretace
  - p= $P(D|H_0)$ a nikoli $P(H|D)$
- Jak z jich ven?
  - VŽDY udávat velikost efektu (Cohenovo d, r, R^2, η^2, ω^2)
  - používat intervalové odhady
  - testování hypotéz používat pouze doplňkově

Základní výzkumné otázky/hypotézy

1. Stanovení hodnoty parametru v populaci
   - **stanovení intervalu spolehlivosti** na μ, σ, ρ, b…
   - srovnání statistiky s hypotetickou hodnotou – konstantou
     - Korelace mezi proměnnými
     - korelace, regrese, chí-kvadrát
     - H_1: ρ <>0 … H_0: ρ =0
     - např. Mezi věkem a počtem návštěv lékaře za rok existuje lineární korelace.
2. Rozdíl mezi skupinami/vzorky - populacemi
   - mezi průměry, korelacemi, rozptyly, pravděpodobnostmi, pořadími….
   - lze srovnávat 2 i více skupin-populací
   - např. H_1: $\mu_1 - \mu_2$<>0 … H_0: $\mu_1 - \mu_2$=0
   - např. Muži a ženy se liší v míře úzkostnosti.

- Rozdíl průměrů lze převést na korelaci a naopak - obecně mluvíme o **velikosti efektu/účinku**

AJ: difference, association, effect size, two-tailed, one-tailed (directional)

Přehledy statistických testů

- **receptář Oseckých** třídění podle
  - počtu výběrů(skupin) – 1, 2, nebo více
  - úrovně měření – alternativní, nominální, , intervalová
  - typu procedury – interval spolehlivosti, test hypotézy, velikost potřebného výběru
- **Hendl – kapitola 12 a str. 235**
- **online**
  - http://www.graphpad.com/www/book/Choose.htm
  - http://www.whichtest.info
  - http://www.socialresearchmethods.net/selstat/ssstart.htm
  - česky: http://meloun.upce.cz/metody/
- Sheskin, D.J.: *Handbook of parametric and nonparametric statistical procedures*. CRC press, 2004.
- Kanji, G.K.*: 100 statistical tests*. Sage, 2006.

Př.: Testy na rozdíly 2 středních hodnot

Intervalová závislá – rozdíly průměrů

- *párový test*: párový *t*-test
- *nezávislé skupiny*:
  - známý rozptyl v populaci: *z*-test
  - neznámý rozptyl v populaci: *t*-test pro nezávislé skupiny
    - varianta pro stejné a nestejné rozptyly mezi skupinami

Ordinální závislá – rozdíly mediánů, průměrného pořadí

- *párový test*: binomický znaménkový test, Wilcoxonovo *T* (int)
- *nezávislé skupiny*: Mann-Whitney *U*

Nominální závislá – shoda rozložení

- *párový test*: McNemarův test (dichotomie), Bowkerův test symetrie
- *nezávislé skupiny*: chí-kvadrát

AJ: sign test, chi-square, Wilcoxon *T*, Mann-Whitney *U*, paired(-samples) *t*-test (dependent, repeated measures), one-sample *t*-test,

Srovnání 2 nezávislých průměrů: *t* -test

Předpoklady použití … jsou-li výrazně porušeny, volíme raději neparametrický test

- proměnná je v populaci **normálně rozložená** - neřeší se, pokud je $n_1, n_2 > 30$
- homogenita rozptylů (**homoscedascita**), pokud $n_1 \neq n_2$
  - řeší modifikace t-testu pro nestejné rozptyly (6.2.3)
  - testuje se Levenovým testem (od oka $\mathbf{s_1^2/s_2^2<2}$)
- **nezávislost pozorování** - řeší párový t-test (pro závislé výběry) (6.2.4)

- H_0: $\mu_1 - \mu_2 = 0$ (nebo roven konstantě, nebo >/< 0 či *c*) a zvolíme $\alpha = 1\%, 5\%$, nebo 10%
- **Rozdíl průměrů *d*** má

  s^2_{pooled}

  výběrovou chybu s_d= $\sqrt{\{[((n_1-1)s_1^2+(n_2-1)s_2^2)/(n_1+n_2-2))]*[1/n_1+1/n_2]\}}$

  ***t*-rozložení** s n_1+n_2-2 **stupni volnosti** (ν)
- Spočítáme **testovou statistiku *t*** = $(m_1 - m_2)/s_d = d/s_d$
- Zjistíme jaká je ***p* (*t* ≥ |zjištěná hodnota|)** - tabulky, TDIST(t , ν)
- Je-li $p \geq \alpha$, pak H_0 zůstává platná, je-li $p < \alpha$, H_0 zamítáme (a konstatujeme existenci statisticky významného rozdílu).
- Spočítáme Cohenovo *d* a interval spolehlivosti pro rozdíl průměrů.

Velikost účinku/efektu

- Možnost srovnání mezi studiemi zkoumajícími tutéž výzkumnou otázku pomocí různě operacionalizovaných proměnných
- Možnost srovnání velikosti efektu vyjádřeného různými koeficienty
- Snadnější interpretace

Pro rozdíly středních hodnot

- **Cohenovo $d = |m_1 - m_2|/s_{pooled}$** ; $s_{pooled} = \sqrt{[((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2))]}$
- varianta $d' = |m_1 - m_2|/s_{con}$; s_{con} = s kontrolní skupiny

Pro těsnost vztahu (korelace)

- r a r^2, R^2, η^2(eta), ω^2 – podíl vysvětleného rozptylu závislé proměnné

Indikátory velikosti efektu lze mezi sebou navzájem převádět

- Cohenovo d na r : $r = \sqrt{(d^2/(d^2 + 4))}$
- r na Cohenovo d : $d = 2r/\sqrt{(1 - r^2)}$

AJ: effect size, Cohen's d, strength of association, explained variance

Síla testu

Síla testu (1-β) je pravděpodobnost, že existující rozdíl bude detekován, zjištěn jako statisticky významný.

Záleží na

- skutečné velikosti účinku (δ, ρ…)
- variabilitě proměnné(ých) – s, σ
- velikosti vzorku n
- zvoleném riziku chyby I. typu, α: čím nižší je α tím nižší je síla
- zvoleném testu (parametrické mají vyšší sílu)

Obvykle toužíme po co nejvyšší síle testu, cca 0,8 a výše.

- Bojujeme o ni především velikostí vzorku a kontrolou intervenujících proměnných (snižuje s).

Publikace výsledků testování hypotéz

- Primárně udáváme velikost efektu, nejlépe intervalem spolehlivosti
- Sekundárně udáváme výsledek statistického testování
  - udáváme získanou hodnotu *p* (Sig.)
  - uvádíme i testovou statistiku (i se stupni volnosti) – r, $t(\nu)$, $F(\nu_1, \nu_2)$, χ^2, M-W U…
- Interpretujeme nejlépe interval spolehlivosti. Výsledek statistického testování interpretujeme vzhledem k použité nulové hypotéze.

Testy normality rozložení

- Kolmogorov-Smirnov s Lillieforsovou korekcí, Shapiro-Wilk, D'Agostino-Pearson a jiné
- Testují H_0, že rozložení proměnné se neliší od normálního rozložení
  - jsou to jedny z tzv. **testů dobré shody** (goodness-of-fit tests)
  - testovaná H_0 je shoda; tj. $p<\alpha$ = příliš velká odchylka od normality
- **Jejich užívání je kontroverzní**
  - na malých vzorcích nenormalitu nedetekují (při n=20, 1-β < 0,5)
  - na velkých vzorcích (n > 1000) jsou naopak extrémně přísné
  - *t*-testy a ANOVA jsou proti narušení normality robustní, takže nám obvykle stačí konstatovat unimodalitu bez extrémního zešikmení
  - pro rozhodování mezi použitím parametrických a neparametrických testů volíme spíše **úroveň měření** a velikost vzorku

AJ: tests of (univariate) normality, goodness-of-fit tests

Více: http://www.psy.surrey.ac.uk/cfs/p8.htm, http://www.graphpad.com/library/BiostatsSpecial/article_197.htm

χ^2 – test dobré shody

- Liší se empirické četnosti nějakých jevů od teoreticky očekávaných četností?
  - Házení kostkou – kolikrát padne 1,2,…
  - Preference politických stran ve volbách…
  - Tedy jedna nominální proměnná, jeden výběr
- Testujeme pravděpodobnost daného rozdílu mezi empirickými a očekávanými hodnotami v rámci jednoho výběru
- H_0: $F(x) = F_0(x)$ vs. H_1: $F(x) \neq F_0(x)$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

  - k je počet kategorií, n velikost vzorku, n_i pozorovaná četnost v kat. i, p_i teoretická pravděpodnost jevu v kategorii (0 až 1); $\sum n_i = \sum np_i$
- Rozdělení χ^2; stupně volnosti df = k-1
- Překoná-li hodnota χ^2 kritickou mez, H_0 zamítáme.
- Pro získání pravděpodobnosti χ^2 CHIDIST(x,volnost); CHIINV(prst, volnost)
- Očekávané četnosti… při uniformním rozložení 1:1:1…; nebo libovolně teoreticky odvozené (10:24:32…)
- N_i i NP_i vždy jako četnosti; nikdy ne procenta = relativní četnosti (ztráta informace o velikosti vzorku.

AJ: Chi-square goodness-of-fit test, observed (empirical) frequency vs. expected frequency

Ve kterém městě by jste žili nejraději?

Kategorie	n	p	np	(n-np)^2/np
Paříž	38	0,2	28	3,57
New York	37	0,2	28	2,89
Londýn	22	0,2	28	1,29
L.A.	25	0,2	28	0,32
Tokio	18	0,2	28	3,57
Celkem	140	1	140	11,64
Chi2	**11,64**		**p**	**0,02**

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Analýza rozptylu

Srovnávání více než dvou průměrů

If your experiment needs statistics, you ought to have done a better experiment.
Ernest Rutherford

Omezení *t*-testu (i jeho nPar alternativ)

t-test umožňuje srovnání pouze dvou průměrů

- Více skupin (*j*) >> mnoho porovnání: *j* (*j* -1)/2

Více srovnání způsobuje strmý růst pravděpodobnosti chyby I. typu

- např. při α=0,05 a 20 testech *p*=0,64 (1 nebo více chyb)
  - aplikace binomického rozložení
- Platí to pro jakékoli statistické testy (zejm. korelace)

Je *nevhodné* provádět velké množství testů na jedněch datech (cca >5)

- Zneužití se označuje jako rybaření v datech – capitalizing on chance
- Lze kompenzovat korekcí hladiny α (Bonferroniho korekce), avšak za cenu značného snížení síly testu (1-β).
  - Místo α testujeme na hladině $\alpha' = \alpha / N$, kde N je počet prováděných testů.

AJ: multiple tests, capitalizing on chance, Bonferroni correction, statistical power

Řešení = Analýza rozptylu (ANOVA)

Testuje na více skupinách jen jednu hypotézu:

- Je někde mezi skupinovými průměry někde rozdíl?
  - Je mezi Pražáky, Brňáky a Ostraváky rozdíl v průměrné lakotě?
  - H_0: $\mu_{\text{Pražáci}} = \mu_{\text{Brňáci}} = \mu_{\text{Ostraváci}}$

- Je-li odpověď „**ano**“ ($p < \alpha$), pak se můžeme podívat na jednotlivé rozdíly detailněji (post-hoc testy)
- Je-li odpověď „**ne**“ ($p > \alpha$), pak bychom neměli (rybaření)

Terminologická vložka - ANOVA

- ANOVA = ANalysis Of Variance = analýza rozptylu
  - i přes svůj název jde o srovnávání **průměrů**
- ANOVA zjišťuje vztah mezi **kategoriální nezávislou** a **intervalovou závislou**.
  - kategoriální nezávislá = **faktor** (factor, „-way“)
  - hodnoty kategoriální nez. = **úrovně** (level, treatment)
- Zjištěný rozdíl = efekt, účinek (effect)

Princip ANOVY 1.

- rozptyl = *MS* = mean square
- MS_{within} : variabilita uvnitř skupin ($MS_{e,\ error}$)
  - $MS_{within} = SS_{within}/n - j$
- $MS_{between}$: s^2 spočítaný ze skupinových průměrů, variabilita uvnitř skupiny je ignorována (též $MS_{A,\ B,\ treatment}$)
  - $MS_{between} = SS_{between}/j - 1$

Platí-li H_0, jaký čekáme vztah mezi $Ms_{between}$ a Ms_{within} ?

	sk1	sk2	sk3	Celk.
čl1	2	4	6	
čl2	2	4	6	
čl3	2	4	6	
čl4	2	4	6	
čl5	2	4	6	
m	2	4	6	4
s^2	0	0	0	2,9
			MS_{bg}	20
			MS_w	0

	sk1	sk2	sk3	Celkem
čl1	0	6	2	
čl2	4	2	10	
čl3	0	6	2	
čl4	4	2	10	
čl5	2	4	6	
m	2	4	6	4
s^2	4,0	4,0	16,0	9,7
			MS_{bg}	20
			MS_w	8

	sk1	sk2	sk3	Celkem
čl1	1	4	2	
čl2	3	5	5	
čl3	5	1	3	
čl4	4	2	1	
čl5	2	3	4	
m	3	3	3	3
s^2	2,5	2,5	2,5	2,1
			MS_{bg}	0
			MS_w	2,5

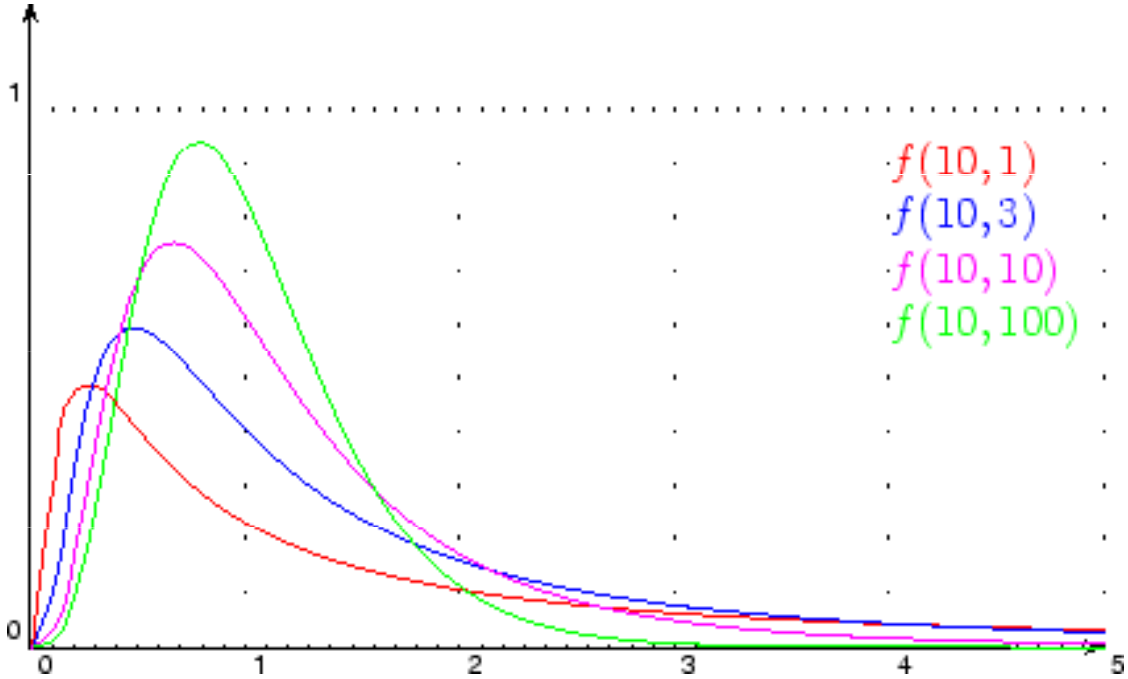
		F	2,5
		$_{0,95}F(2,12)$	3,8853
		p	0,1237

Princip ANOVY – *F*-test

- Čím jsou si průměry podobnější, tím je rozptyl mezi skupinami nižší (Platí-li H_0, $MS_{between}$ se blíží 0)
- Čím nižší je rozptyl uvnitř skupin (MS_{within} se blíží 0), tím průkaznější se průměry mezi skupinami zdají být.
- Důležitý je **poměr těchto dvou odhadů rozptylu:** $$F = \frac{MS_{between}}{MS_{within}}$$
- Čím vyšší je *F*-poměr, tím průkaznější jsou rozdíly mezi průměry (rozsah je 0 až ∞)
- *F* -poměr má při platnosti H_0 jako výběrová statistika ***F* – rozložení** s *(df1,df2)*, které má průměr přibližně 1

Fisherovo-Snedecorovo *F*-rozložení

- Podobně jako *t*-rozložení, je *F*-rozložení vlastně rodina mnoha rozložení mírně se lišící svým tvarem ($F(1; \nu) = t(\nu)^2$)
- Tato rozložení se liší tentokrát dvěma parametry – stupni volnosti
  - ν_1 = *počet skupin – 1* : stupně volnosti čitatele - $MS_{between}$
  - ν_2 = *počet lidí – počet skupin* : stupně volnosti jmenovatele - MS_{within}
  - na pořadí ZÁLEŽÍ



http://www.econtools.com/jevons/java/Graphics2D/FDist.html

Princip ANOVY – dělení rozptylu.

- Dělení variability (rozptylu) podle zdrojů **jako u lineární regrese**

$$X_{ij} = \mu + \alpha_j + e_{ij}$$

- X_{ij} = skóre jedince (*i*-tý jedinec v *j*-té skupině)
- μ = průměr populace
- α = vliv příslušnosti ke skupině (vliv úrovně faktoru)
- e_{ij}= chyba (vše, s čím nepočítáme, individuální prom.)

$$Y_i = a + b_1 X_1 + b_2 X_2 + \ldots + b_{j-1} X_{j-1} + e_i$$

$$X_{ij} - m = (m - m_j) + (X_{ij} - m_j)$$

odchylka od celkového průměru = odchylka od skupinového průměru + odchylka skupinového průměru od celkového průměru

- … odchylky umocněné na druhou = cesta k rozptylu

$$SS_{\text{Total}} = SS_{\text{Between (A, treatment)}} + SS_{\text{Within(Error)}}$$

$$MS_{\text{Total}};\ MS_{\text{Error}};\ MS_{\text{A}}$$

Velikost účinku (efektu)

- Podobně jako u regrese chceme vědět, jaká část rozptylu závislé je vysvětlená nezávislou
- Ekvivalentem R^2 je u anovy η^2 (eta)
  - $\eta^2 = SS_{Between}/SS_{Total}$
  - Poněkud přesnější je $\omega^2 = (SS_{Between} - df_{Between} . MS_{Within})/(SS_{Total} + MS_{Within})$
- Pro konkrétní rozdíl průměrů $d_{Coh} = m_1 - m_2/\sqrt{MS_{Within}}$

- Velikost účinku je <u>vždy</u> třeba uvádět

Předpoklady použití ANOVY

- normální rozložení uvnitř skupin
  - při n_j>30 a $n_1=n_2=\ldots=n_j$ je ANOVA robustní
- stejné rozptyly uvnitř skupin: homoskedascita
  - do $s_{max}/s_{min}<3$ je ANOVA robustní, zvláště při $n_1=n_2=\ldots=n_j$
- nezávislost všech pozorování
  - při opakovaných měřeních je třeba použít ANOVU pro opakovaná měření

- One-way ANOVA
  - kontrasty a post-hoc testy **ONEWAY**
- Faktoriální (two-way, three-way...) ANOVA
  - interakce **UNIANOVA**
- Analýza kovariance – ANCOVA
  - kontrola intervenující proměnné
- ANOVA s více závislými - MANOVA

Životní spokojenost a rodina

- Domníváme se, že kompletní rodina je základ životní spokojenosti.
  - H_1: $M_{komplet} > M_{nekomplet}$
- Zajímá nás, zda se liší chybění otce a jeho nahražení nevlastním otcem
  - H_2: $M_{bez\ otce} \neq M_{nevlastní\ otec}$

Kontrasty

- I když můžeme srovnat všechny průměry se všemi ostatními, platíme za to velkou ztrátou síly
- Řešením jsou předem plánovaná srovnání – **KONTRASTY**
- Lze srovnat kterékoli 2 skupiny nebo skupiny skupin
  - např. 1. skupinu se průměrem všech ostatních, kontrolní skupinu se každou ze zbývajících skupin zvlášť
- Realizuje se zvláštním kódováním
  - při platnosti nulové hypotézy je součet vážených průměrů 0
- H_1: 1. vs (2. a 3.) ..... -2 1 1
- H_2: 2. vs 3. ..... 0 -1 1

Post-hoc testy (simultánní porovnávání)

- Po (a pouze po) prokázání „nějakých“ rozdílů mezi průměry obvykle chceme vědět, mezi kterými skupinami konkrétně rozdíly jsou: **post-hoc testy**
- Srovnáváme každou skupinu s každou způsobem, který nezpůsobí nárůst α.
- Je-li důležité udržet α pod kontrolou, je správnou volbou **Scheffe**ho test nebo **Tukey**ho HSD – volba pro *rybaření*
- Máte-li stejně velké skupiny (balanced design) - **REGWQ**
- Pokud to α kritická a máte-li pár *kvazi*-hypotéz na mysli, pak je volbou **Student-Neuman-Keuls** (**S-N-K**)
- Extrémně „dajný“ a nepříliš vhodný pro více než 3 skupiny je **LSD** a proto se nedoporučuje.
- Při nesplnění homoscedascity – **Games-Howell**

Field 339

Faktoriální ANOVA

- více faktorů ... možnost **interakce** mezi nimi
- **fixed** vs. **random** faktory

Liší se výkonová motivace podle věku a pohlaví?

- INT: Jsou případné genderové rozdíly shodné v obou kohortách?

Liší se výkonová motivace mezi školami a podle pohlaví?

- INT: Liší se genderové rozdíly škola od školy?

Analýza kovariance

Velká variabilita závislé může zastírat rozdíly.
Dokážeme-li část její variability vysvětlit nějakým prediktorem, můžeme hledat rozdíly pouze ve zbývající části rozptylu závislé.

- statistická kontrola – jako parciální korelace a regrese
- Proměnnou, jejíž vliv chceme kontrolovat, vkládáme jako **kovariát**

MANOVA

- Máme-li více závislých
- Opatrně.

Shrnutí

- ANOVA je pro situace s intervalovou závislou a více kategorickými nezávislými – porovnávání mnoha průměrů
- Faktory mohou být **fixní** nebo **náhodné**
- ANOVA je podobná regresi – pro interpretaci je dobré si vyžádat „**parametry**“, tj. regresní váhy
- Lze testovat konkrétní hypotézy – **kontrasty**
- Lze testovat všechny možné rozdíly průměrů – **post hoc**
- Lze uvažovat o kombinovaném vlivu faktorů – **interakce**
- Lze kontrolovat vliv intervenujících proměnných – **kovariáty - ANCOVA**
- Lze mít i více závislých najednou – **MANOVA** - opatrně