Práce s daty

Peter Spáč

6.10.2016

Práce s daty

- Analýza dat jako klíčová část výzkumné práce
- Aplikace vhodného modelu na data
- Ne všechna data jsou vhodná pro všechny možné operace
- Předpoklady použitelnosti dat

Data a testy

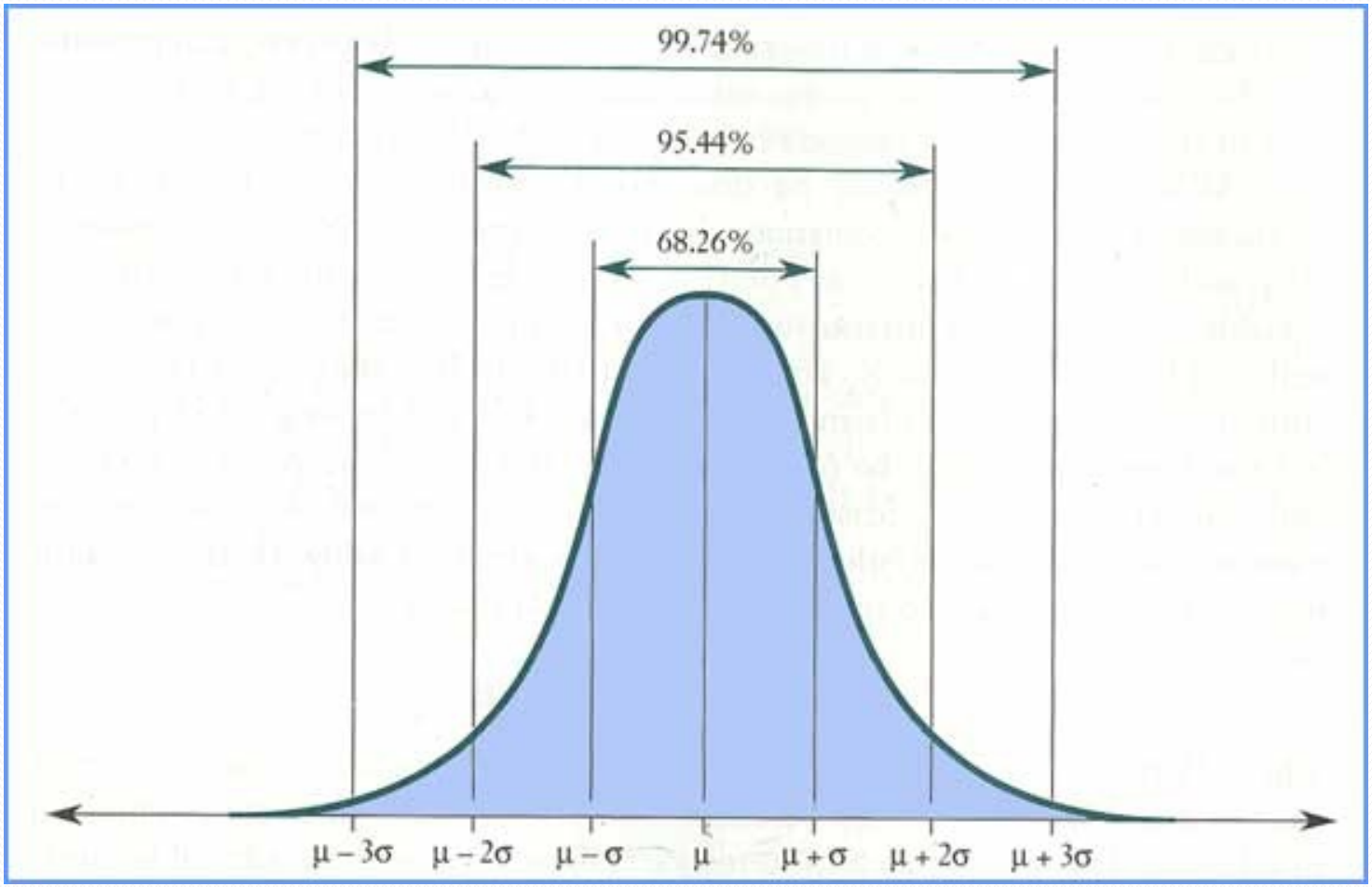
- Různé statistické testy mají odlišné nároky na vstupní data
- Použití nesprávných dat může vést k nepřesným výsledkům
- Druhy testů – parametrické a neparametrické
- Potřebná kontrola dat před samotnou analýzou

Parametrická data

- Základní předpoklady (ne pro každý parametrický test):

1. Kardinální data (interval)
2. Nezávislost
3. Normální distribuce dat
4. Homogenita rozptylu

Normální distribuce



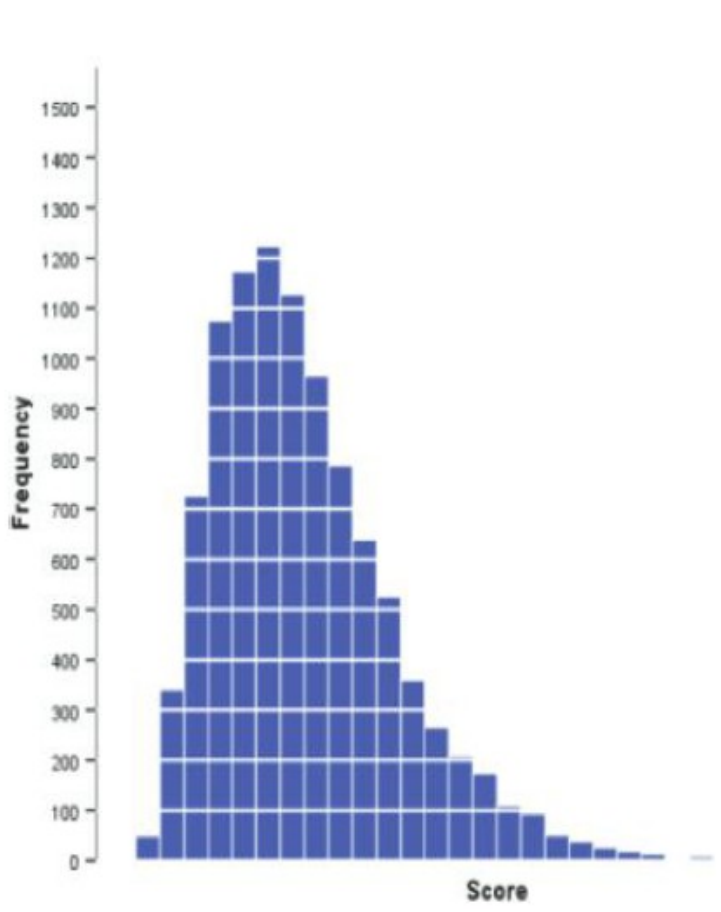
Normální distribuce

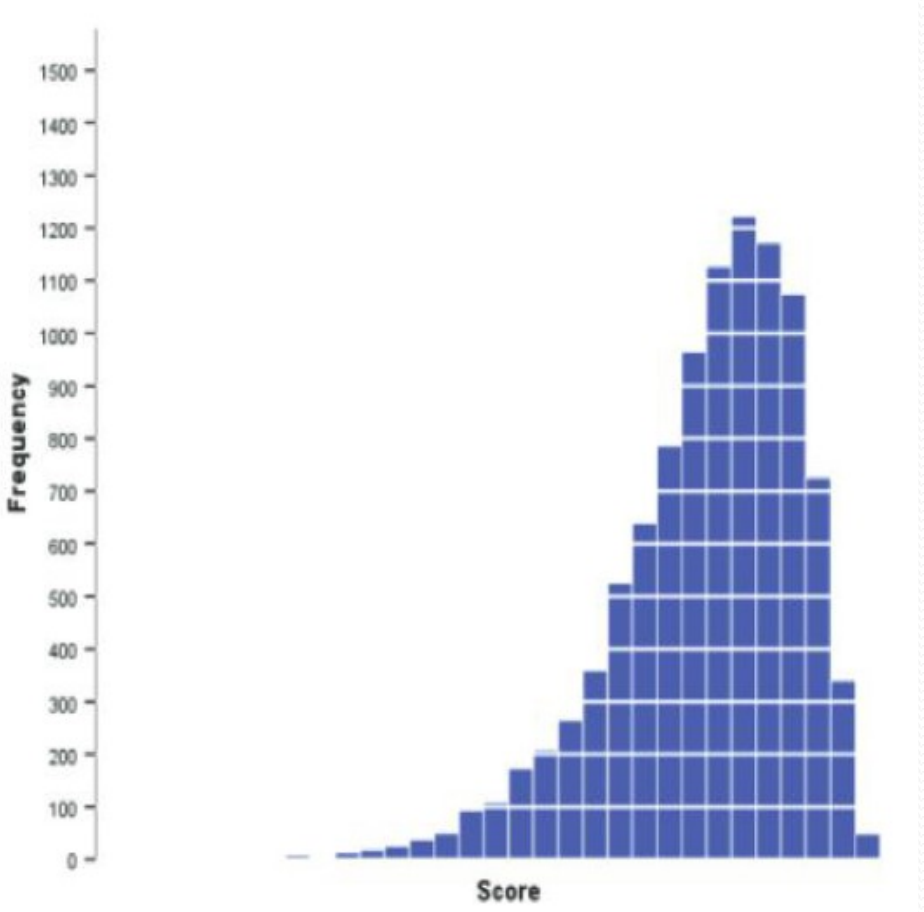
- Specifické uspořádání dat
- Důležitá pro lineární modely
- Vícero způsobů jejího posouzení
  - Vizuálně
  - Numerické hodnoty
  - Testy

Normální distribuce

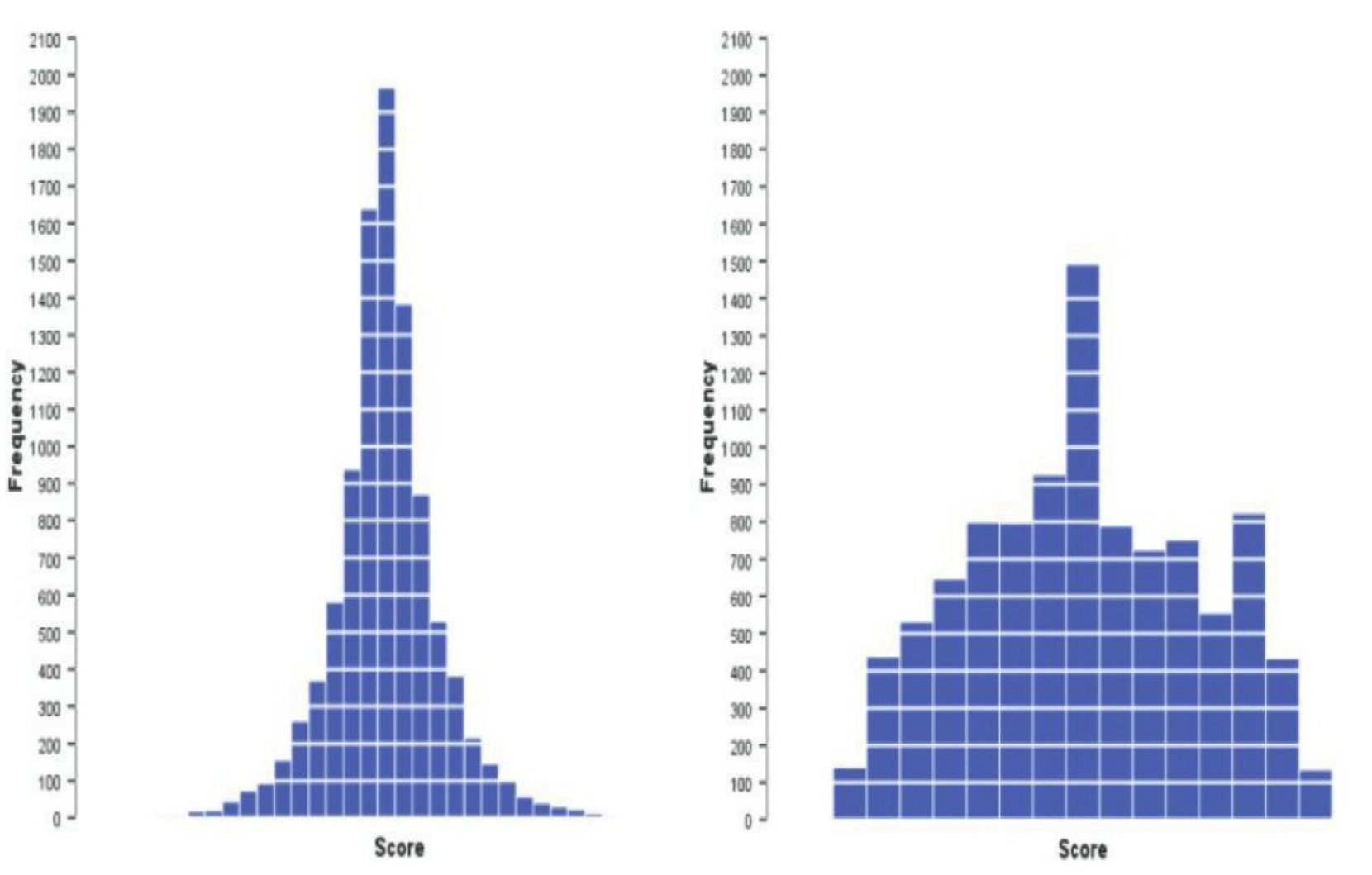
- Odchylky od normální distribuce
- Šikmost:
  - Vrchol křivky je posunutý doleva (doprava)
- Špičatost:
  - Ploché nebo naopak strmé rozložení
- Při dokonale normální distribuci mají šikmost i špičatost hodnotu nula

Pozitivně a negativně sešikmená distribuce (Field 2009: 20)





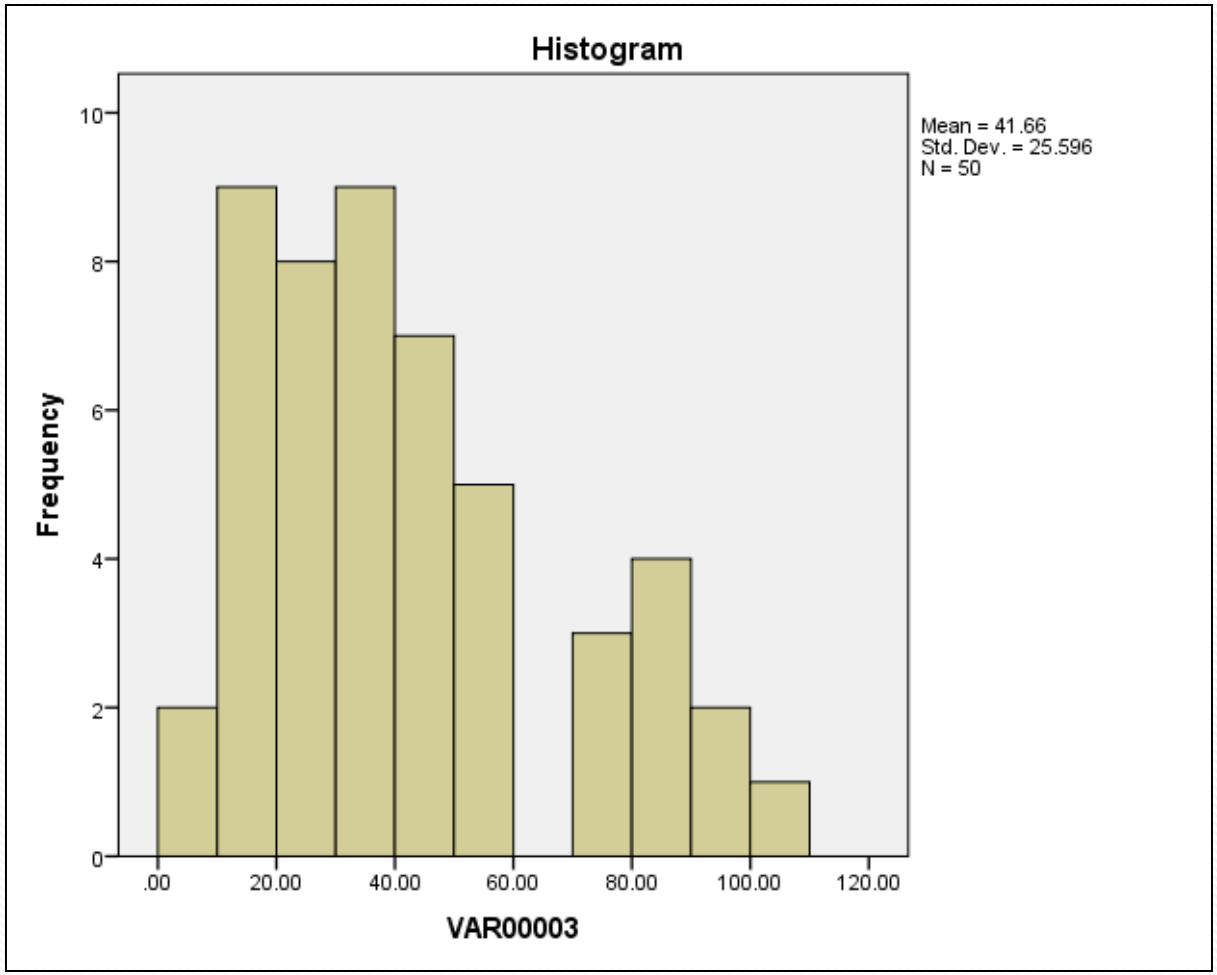
Pozitivně a negativně špičatá distribuce (Field 2009: 20)



1. Vizuální posouzení

- Nejjednodušší způsob posouzení normality (a také subjektivní)
- Posouzení tvaru volným okem
- Histogram – graf zobrazující četnosti
- P-P plot – graf srovnávající očekávané (normální) a reálné rozložení dat

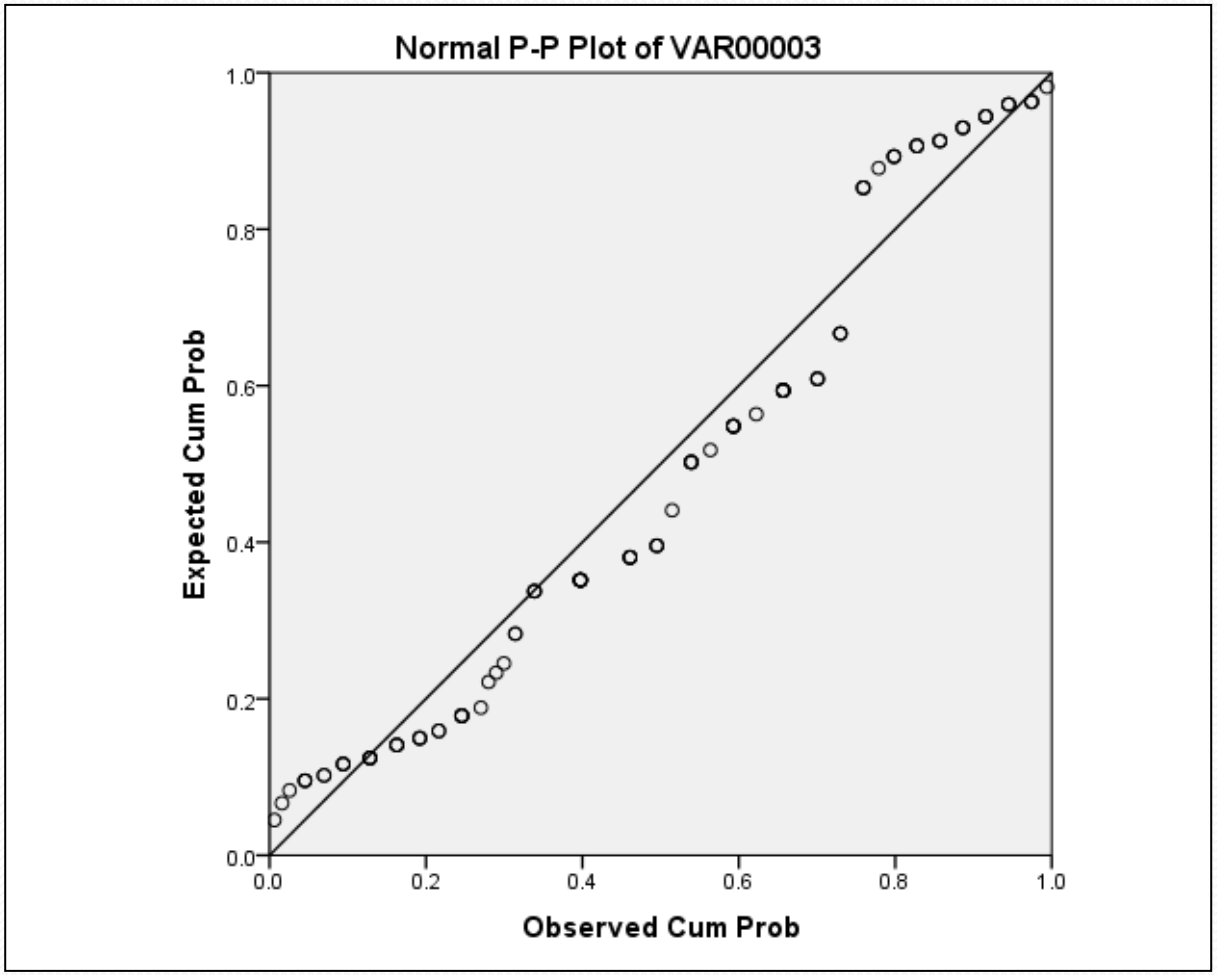
Histogram


Histogram
Mean = 41.66
Std. Dev. = 25.596
N = 50
Frequency
0
2
4
6
8
10
.00
20.00
40.00
60.00
80.00
100.00
120.00
VAR00003

P-P plot

- Využívá standardizaci proměnných (tzv. z-skóre)
- Pravděpodobnost výskytu hodnoty
- Pomocí z-skóre graf srovnává skutečnou a normální distribuci
- Překrytí vyjadřuje normální distribuci našich dat

P-P plot


Normal P-P Plot of VAR00003
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

2. Numerické hodnoty

- Vyčíslení šikmosti a špičatosti
- Odchylky od nuly (kladné i záporné) jsou vychýlením od normální distribuce
- Samotné naměřené hodnoty jsou informativní, pro interpretaci se dělí svou standardní chybou (počítá SPSS)
- Přijatelné hodnoty (z):
  - Malý vzorek: do 1,96 (- 1,96)
  - Velký vzorek: do 2,58 (-2,58)
  - Velmi velký vzorek - nepoužívat

Práce v SPSS

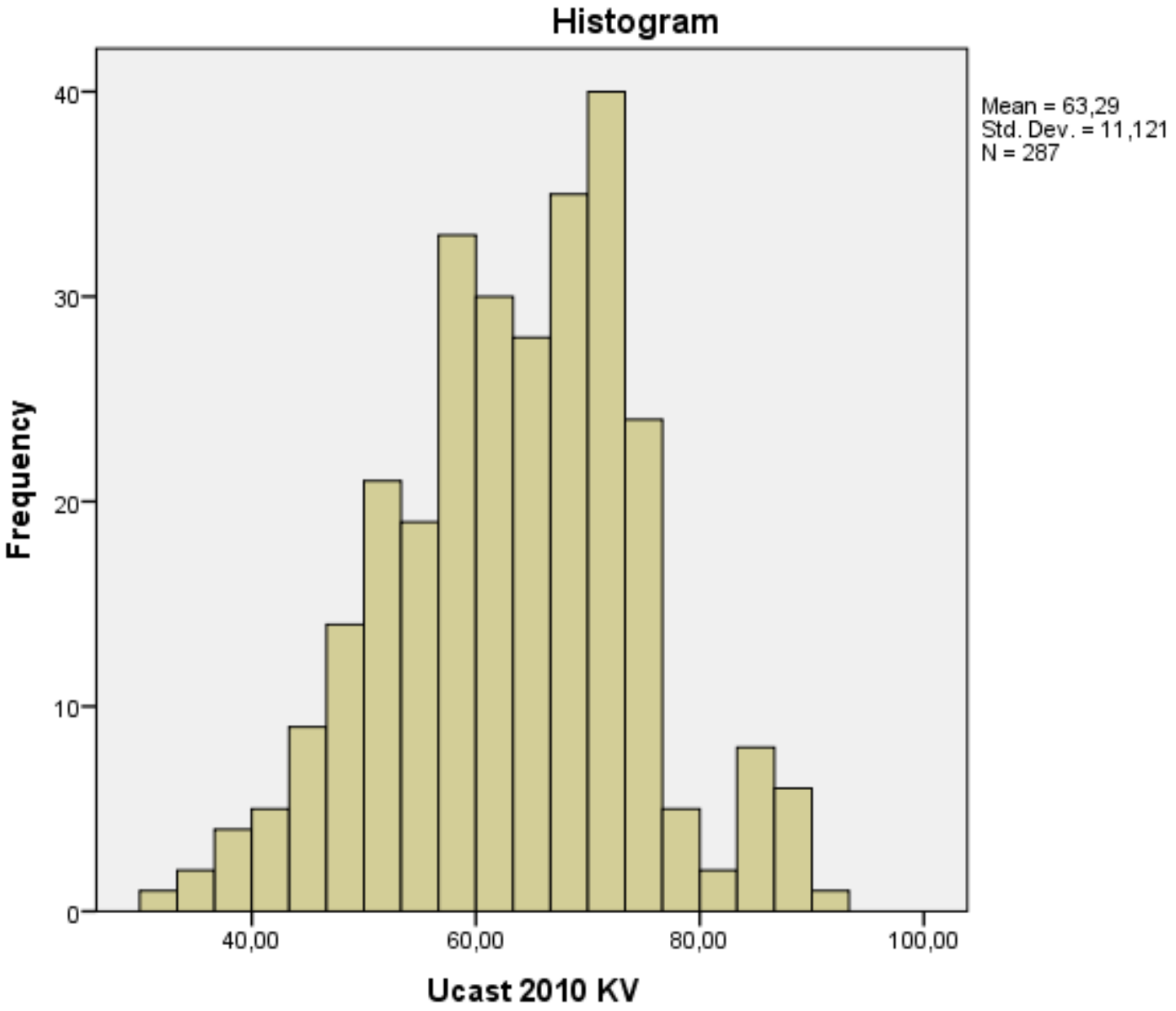
- Histogram
  - Analyze → Descriptive Statistics → Frequencies
  - Charts – Histograms + Show normal curve on histogram

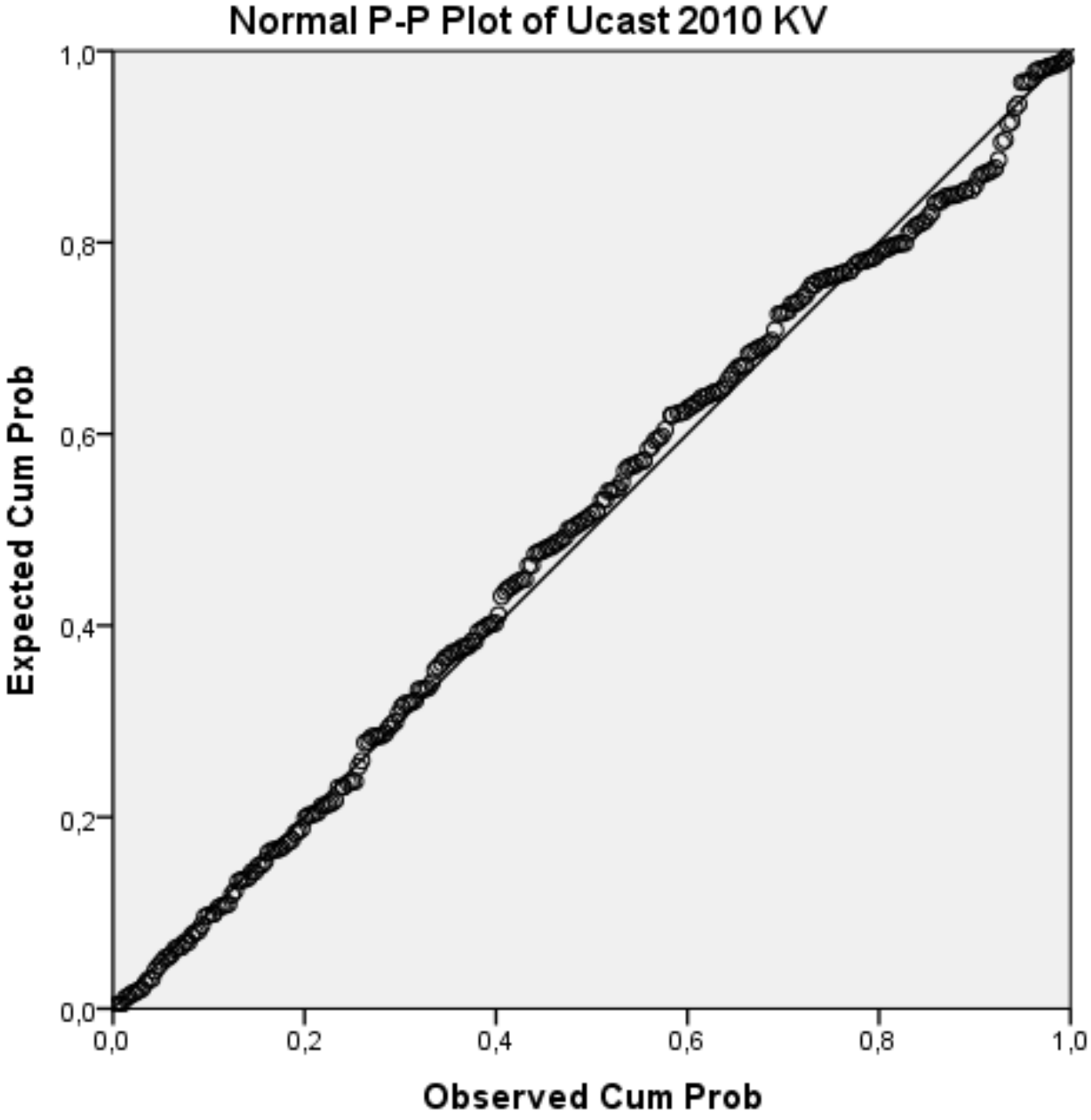
- P-P plot
  - Analyze → Descriptive Statistics → P-P Plots
  - Default nastavení („Test Distribution“ = Normal)

Práce v SPSS

- Šikmost a špičatost
  - Analyze → Descriptive Statistics → Frequencies
  - Statistics – Skewness, Curtosis

  - Kromě toho je pro názornost vždy vhodné nechat si spočítat i základní deskriptivní statistiky (průměr, rozpětí, sm. odchylku, kvartily atd.)


Histogram
Mean = 63,29
Std. Dev. = 11,121
N = 287
Frequency
40
30
20
10
0
40,00
60,00
80,00
100,00
Ucast 2010 KV


Normal P-P Plot of Ucast 2010 KV
Expected Cum Prob
Observed Cum Prob
0,0
0,2
0,4
0,6
0,8
1,0

Práce v SPSS

Statistics		
Ucast 2010 KV		
N	Valid	287
	Missing	0
Mean		63.2855
Median		63.7400
Mode		66.67
Skewness		-.113
Std. Error of Skewness		.144
Kurtosis		-.025
Std. Error of Kurtosis		.287
Sum		18162.93

3. Testy normálního rozložení

- Kolmogorov-Smirnov test, Shapiro-Wilk test
- Logika testů – srovnávají skutečné hodnoty s normální distribucí se stejným průměrem a směrodatnou odchylkou
- Statisticky signifikantní výsledky indikují nenormální rozložení dat
- **Při velkém počtu dat** mohou i malé odchylky od normality způsobit signifikantní výsledky

Práce v SPSS

- Kolmogorov-Smirnov test, Shapiro-Wilk test
  - Analyze → Descriptive Statistics → Explore
  - V „Plots“ zvolit Normality plots with tests
  - Příslušné proměnné vložit do „Dependent list“
  - Možnost samostatné analýzy jednotlivých vymezených částí proměnných (pomocí jiné proměnné)

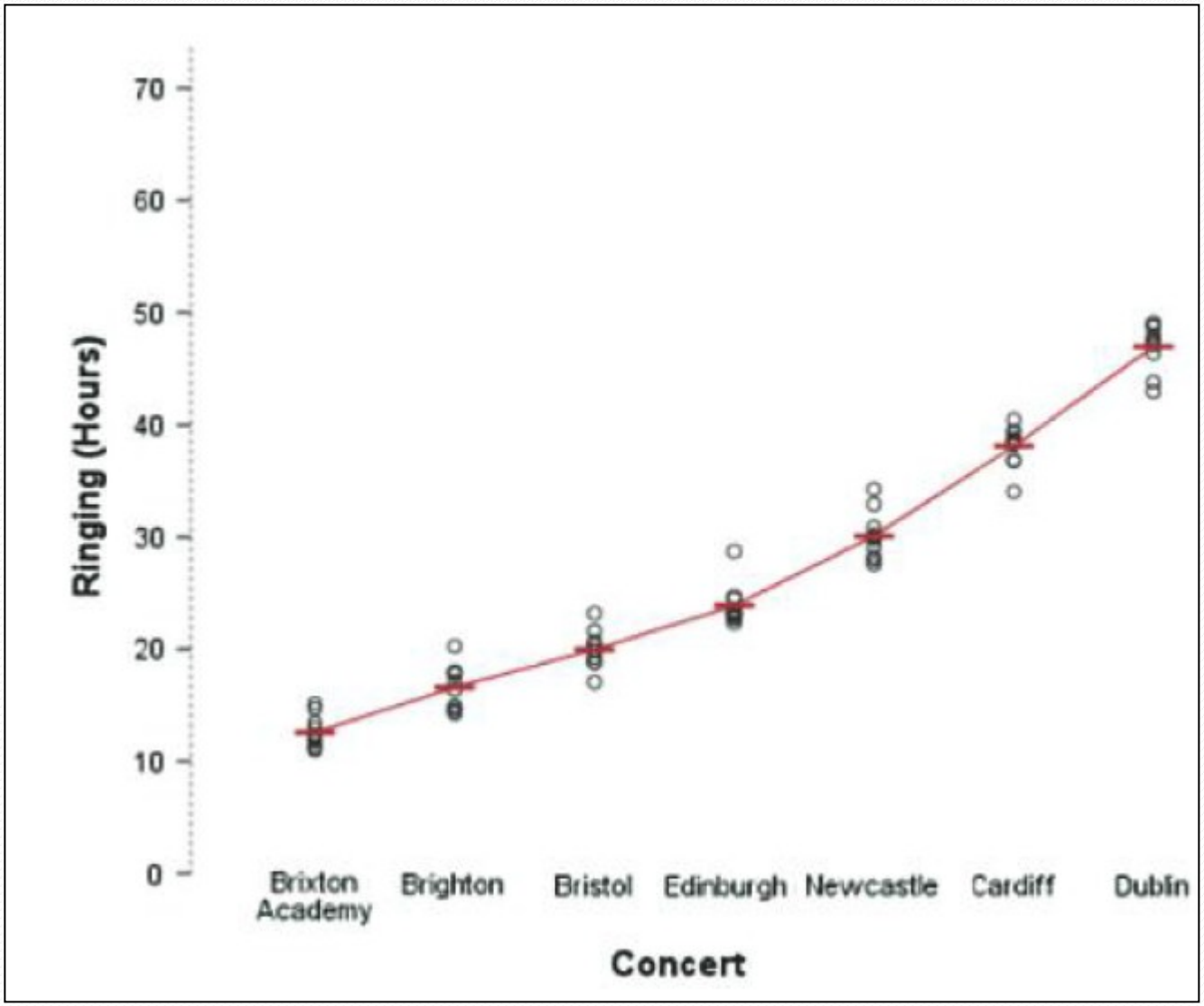
Práce v SPSS

Tests of Normality						
	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Ucast 2010 KV	.046	287	.200	.992	287	.155

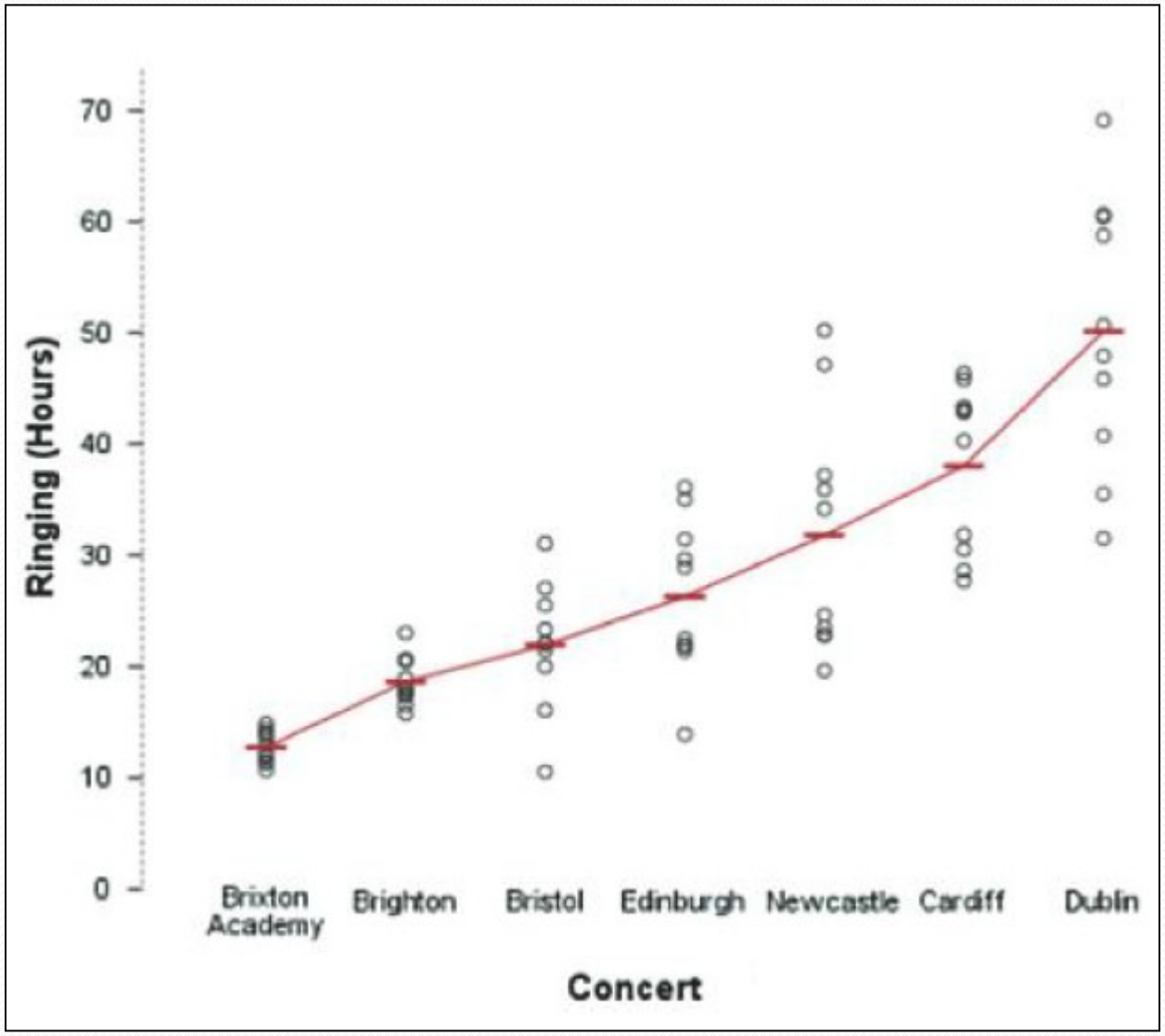
Homogenita rozptylu

- Předpoklad stejných rozptylů hodnot v jednotlivých skupinách případů
- Skupiny případů jsou vymezeny prediktorem (druhou proměnnou)
- Rozptyl výšky mzdy mezi věkovými skupinami obyvatel státu

Homogenita rozptylu (Field 2009: 146)



Homogenita rozptylu (Field 2009: 146)



Homogenita rozptylu

- Levenův test
- Testuje nulovou hypotézu, že rozptyly v různých skupinách jsou stejné
- Pokud test vyjde jako statisticky signifikantní, je předpoklad homogenity rozptylů narušený
- Při velkém počtu hodnot můžou i malé odlišnosti mezi rozptyly vést k signifikantním výstupům

Homogenita rozptylu

- Poměr rozptylů - kontrola Levenova testu
- Poměr největšího a nejmenšího rozptylu a srovnání výsledku s tabulkovými hodnotami
- Tabulková hodnota daná počtem skupin a počtem případů v nich

Práce v SPSS

- Levenův test
  - Analyze → Descriptive Statistics → Explore
  - Příslušné proměnné vložit do „Dependent list“ a „Factor list“
  - V „Plots“ si zvolit jednu z možností v „Spread vs Level with Levene Test“ (untransformed)

Práce v SPSS

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Ucast 2010 KV	Based on Mean	.785	2	284	.457
	Based on Median	.643	2	284	.527
	Based on Median and with adjusted df	.643	2	281.210	.527
	Based on trimmed mean	.759	2	284	.469

Když data nejsou parametrická

- Několik možností:
  - Transformace dat
  - Neparametrické testy
  - Navzdory všemu použití parametrických testů (ne každý test je imunní vůči porušení předpokladů dat)

Úprava dat

- Transformace za konkrétním účelem (např. snaha přiblížit se k normální distribuci dat)
- Různé možnosti - umocnění, odmocnění, logaritmus, 1/x
- Výběr techniky často systémem pokus – omyl
- SPSS někdy ulehčuje práci (Levenův test s volbou „transformed“)

Úprava dat

- Praktická úprava proměnných a jejich hodnot
- Překódování proměnných
- Vznik nových proměnných za pomoci existujících proměnných

Úprava dat v SPSS

- Vytvoření proměnné:
  - Transform → Compute Variable
- Překódování v rámci stejné proměnné:
  - Transform → Recode into Same Variable
- Překódování do jiných proměnných:
  - Transform → Recode into Different Variable