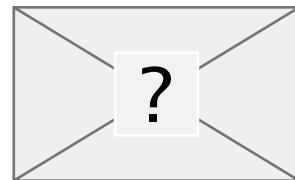


Pravděpodobnost a matematická statistika

Doc. RNDr. Gejza Dohnal, CSc.

dohnal@nipax.cz



Pravděpodobnost a matematická statistika

2010

1. týden (20.09.-24.09.) Data, typy dat, variabilita, frekvenční analýza (histogramy, četnosti absolutní, relativní, prosté, kumulativní), základní statistické charakteristiky (průměr, výběr. rozptyl, minimum, maximum, medián, kvartily, boxplot), sešikmenná rozdělení (vzájemná poloha mediánu a střední hodnoty), chvosty, kvantily
2. týden (27.09.-01.10.) Princip statistické indukce, výběr, vlastnosti výběru, experiment. Náhodná veličina, rozdělení pravděpodobnosti a jeho souvislost s histogramem. Pravděpodobnost, pravidla pro počítání s pravděpodobnostmi, podmíněná pravděpodobnost, závislost náhodných veličin.
3. týden (04.10.-08.10.) Využití závislosti při stanovení pravděpodobnosti - věta o úplné pravděpodobnosti a Bayesova věta
4. týden (11.10.-15.10.) Rozdělení chyb měření - normální rozdělení a počítání s ním. Odhady parametrů normálního rozdělení. Intervaly spolehlivosti pro normální data. Jednovýběrové testy o střední hodnotě
5. týden (18.10.-24.10.) Výběrový poměr jako odhad pravděpodobnosti sledovaného jevu. Alternativní rozdělení, binomické rozdělení. Intervalový odhad výběrového poměru. Výběry s vracením a bez vracení (binomické a hypergeometrické rozdělení)
6. týden (25.10.-29.10.) odpadá
7. týden (01.11.-05.11.) Poruchy v čase (Poissonův proces). Poissonovo rozdělení, exponenciální rozdělení, jeho výhody a nevýhody, modelování doby do poruchy pomocí Weibullova rozdělení, lognormálního rozdělení, případně useknuté normální rozdělení.
- 8. týden (08.11.-12.11.) Testy dobré shody, Q-Q graf (pouze vysvětlení), testy normality. Některé neparametrické testy**
9. týden (15.11.-19.11.) Dvě náhodné veličiny - srovnání dvou výběrů (dvouvýběrové testy)
10. týden (22.11.-26.11.) Dvě náhodné veličiny. Dvourozměrné četnosti jako odhad dvourozměrného rozdělení, frekvenční tabulka. Marginální rozdělení (vše pouze diskrétně s tabulkou)
11. týden (29.11.-03.12.) Závislost náhodných veličin, míry závislosti (kovariance, korelace), test významnosti korelačního koeficientu
12. týden (06.12.-10.12.) Regrese, lineární regresní model (přímková, kvadratická, polynomická regrese), analýza reziduí, pásy spolehlivosti
13. týden (13.12.-17.12.) Více výběrů, jednoduché třídění, ANOVA.
14. týden (20.12.-22.12.) Rezerva, opakování, testy normality (náhrada za 28.10.)

Pravděpodobnostní modely

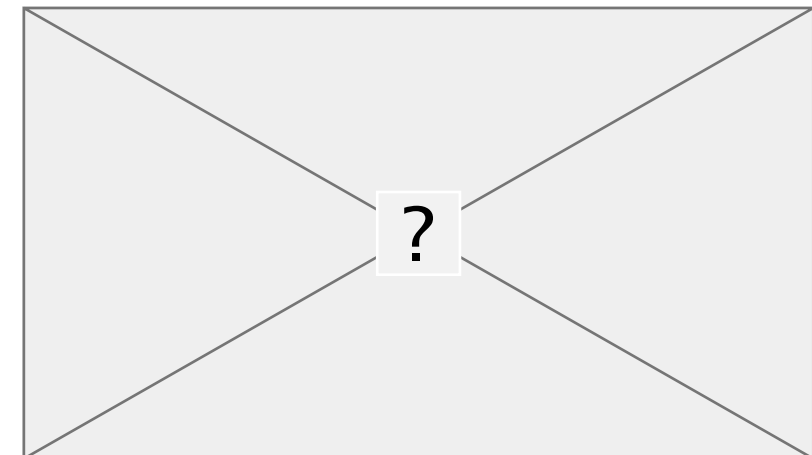
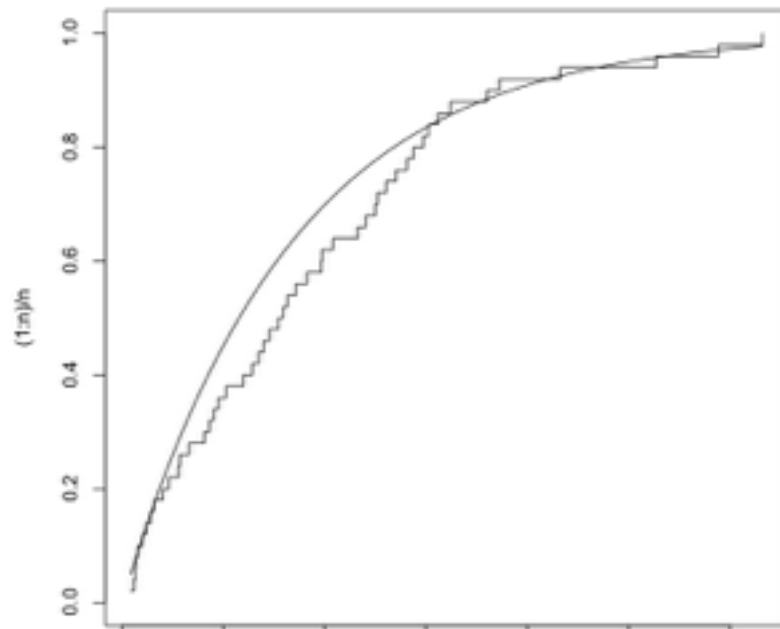
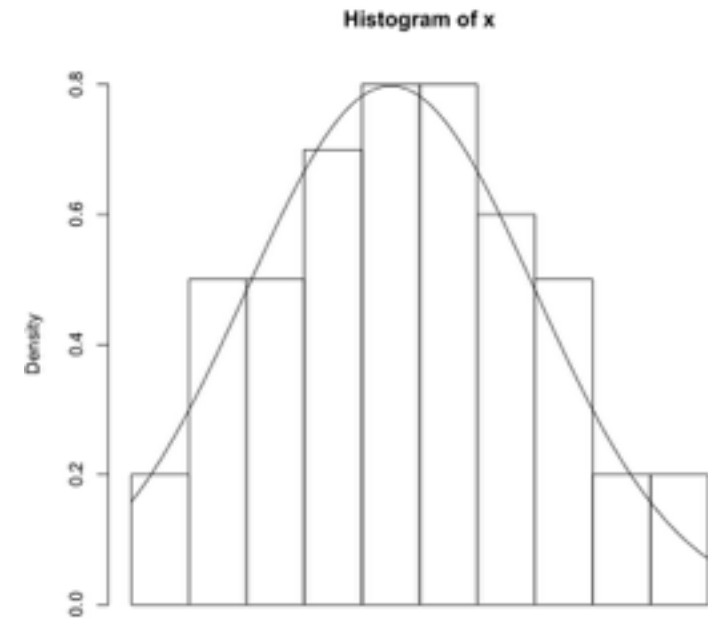
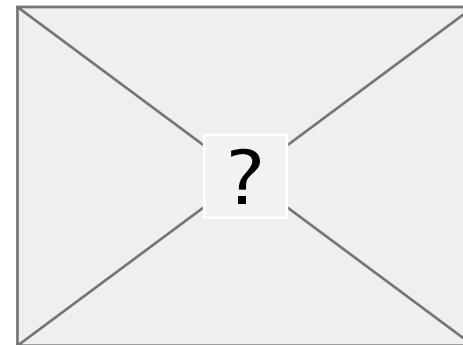
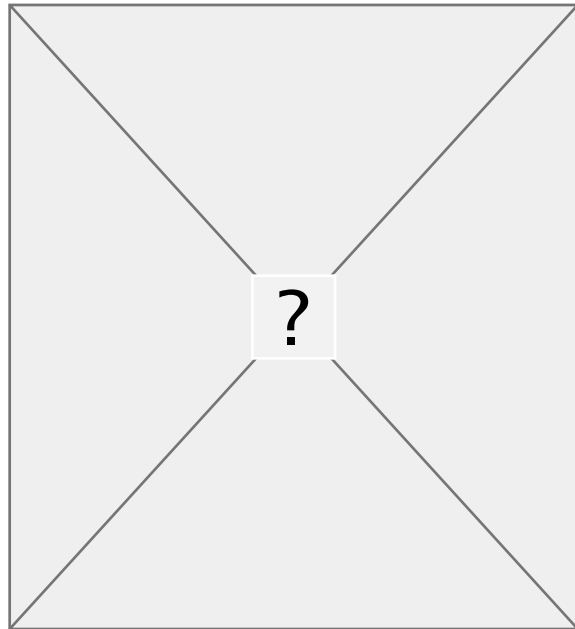
1) Diskrétní:

- Rovnoměrný $\Omega = \{1, 2, \dots, N\}$
- Alternativní $\Omega = \{0, 1\}$
- Binomický $\Omega = \{0, 1, \dots, n\}$
- Hypergeometrický $\Omega = \{\max(0, n + M - N), \dots, \min(n, M)\}$
- Geometrický $\Omega = \{0, 1, 2, \dots\}$
- Poissonův $\Omega = \{0, 1, 2, \dots\}$

2) Spojité:

- Rovnoměrný $\Omega = \langle a, b \rangle$
- Normální $\Omega = (-\infty, \infty)$
- Exponenciální $\Omega = \langle 0, \infty \rangle$
- Weibullův
- Logaritmicko-normální

Testy dobré shody

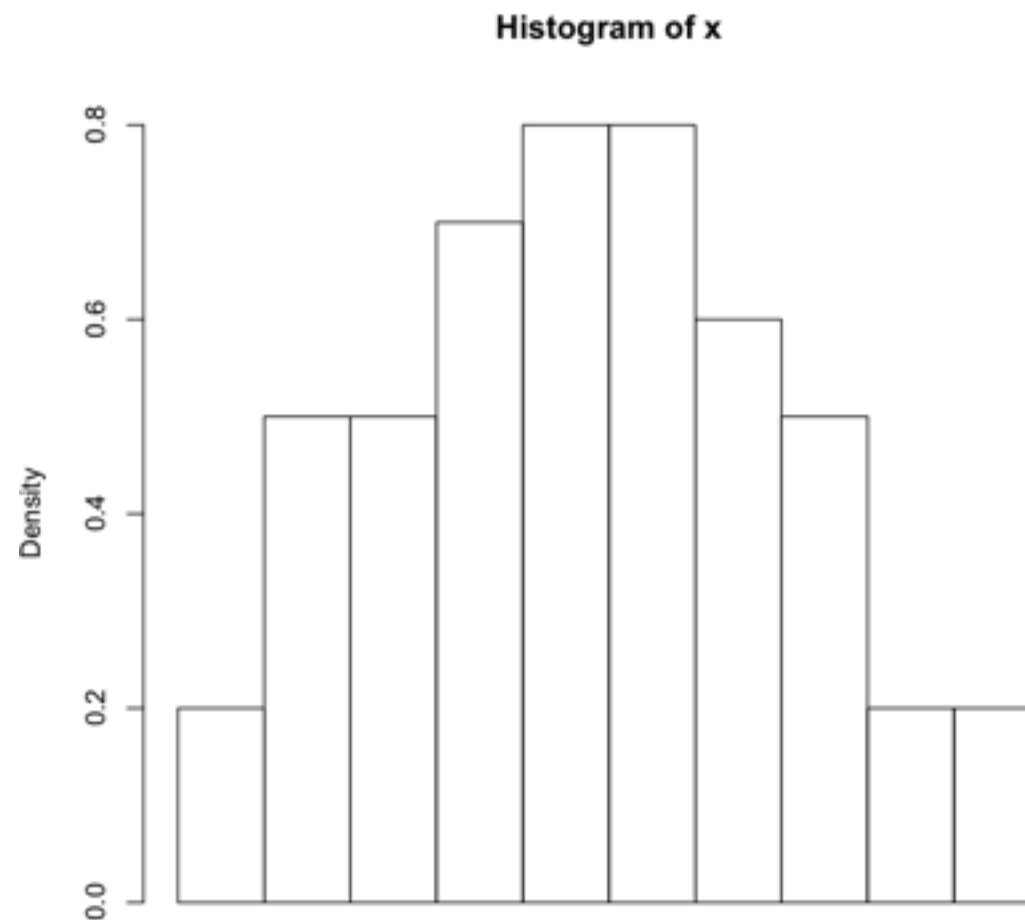


Jaká je shoda pozorovaného experimentu s teoretickým modelem?

Testy dobré shody

Co máme k dispozici?

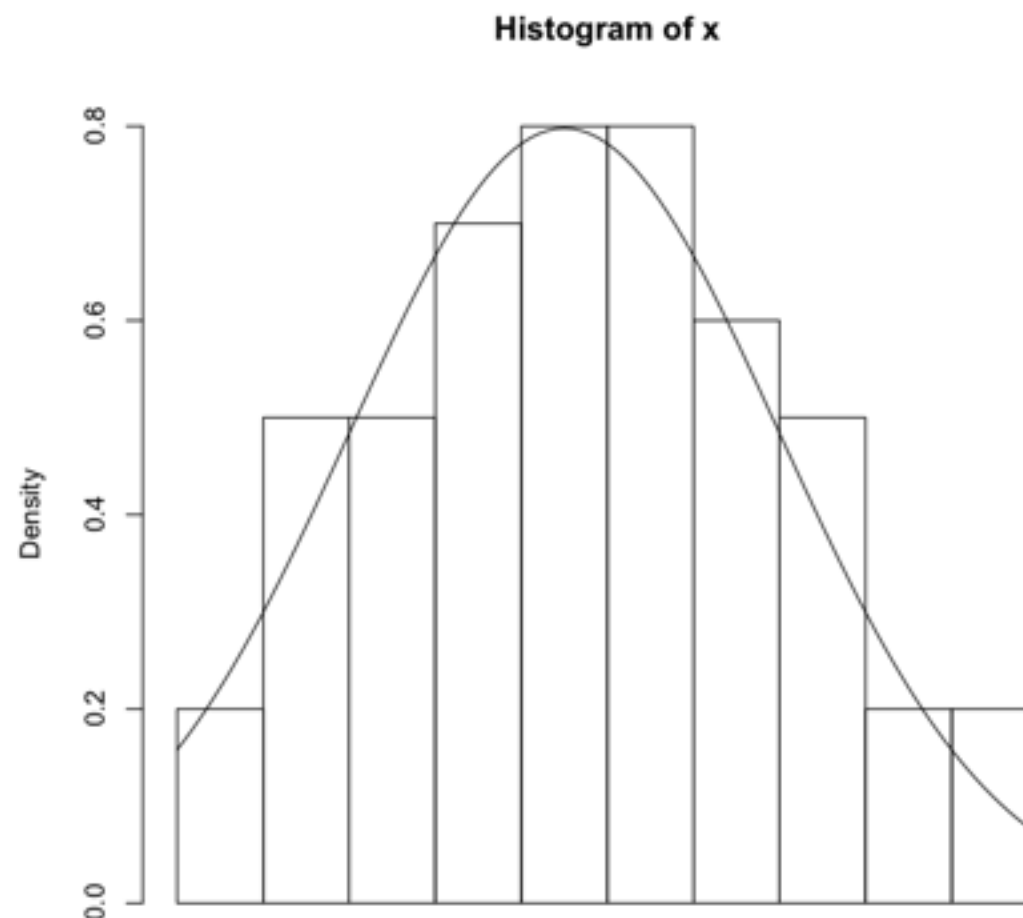
1) Pozorování výsledků experimentu (měření) = data



Testy dobré shody

Co máme k dispozici?

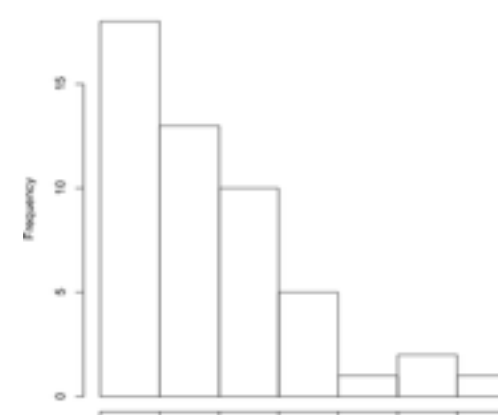
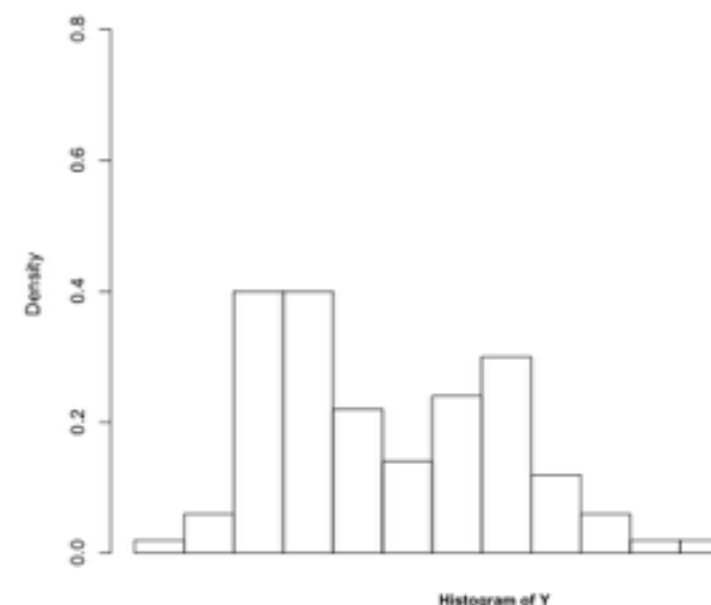
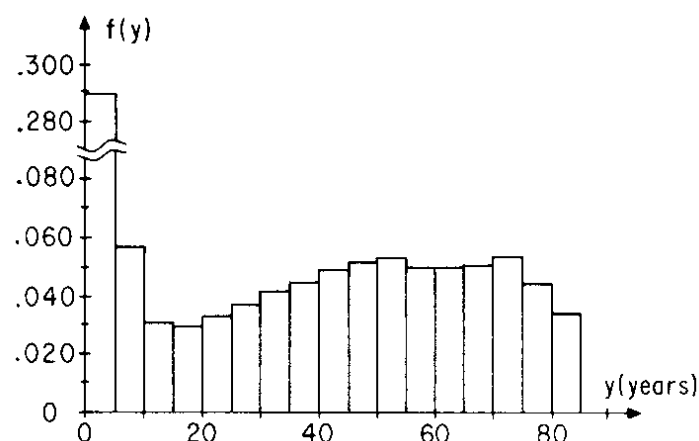
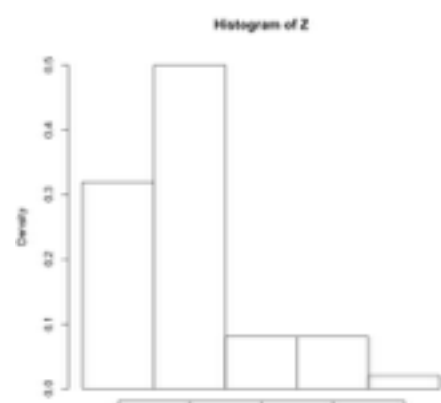
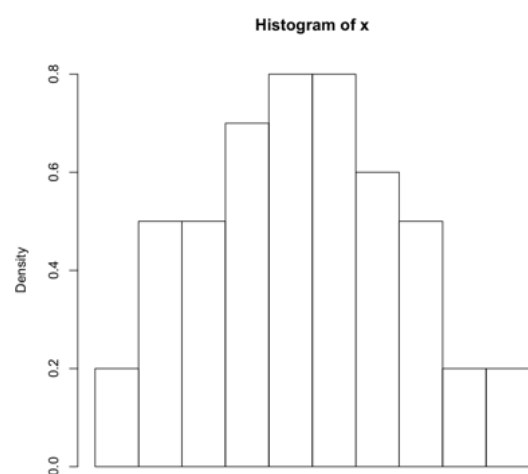
- 1) Pozorování výsledků experimentu (měření) = data
- 2) Představu o hypotetickém (teoretickém) rozdělení pozorované veličiny



Testy dobré shody

Co s tím?

1) Histogram - poskytuje předběžnou představu o tvaru hustoty

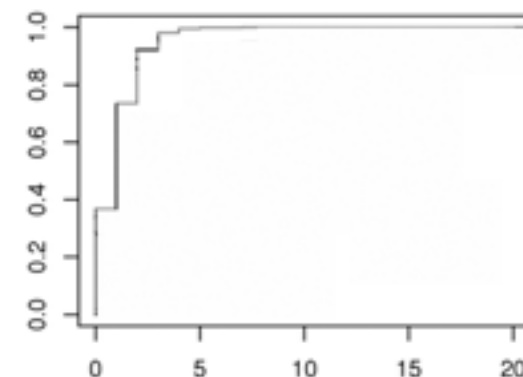
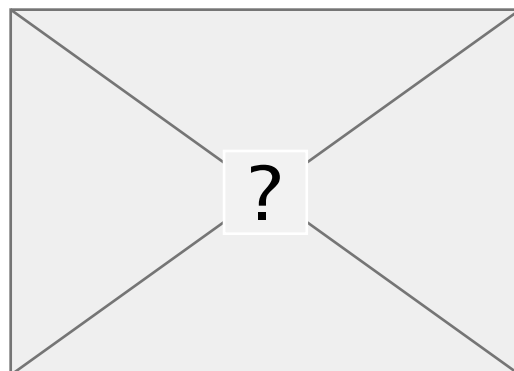
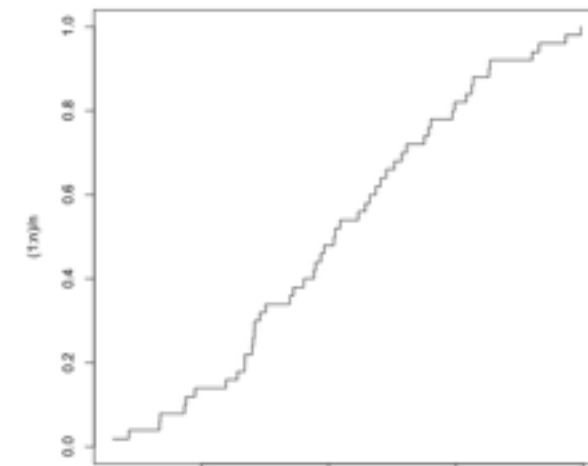
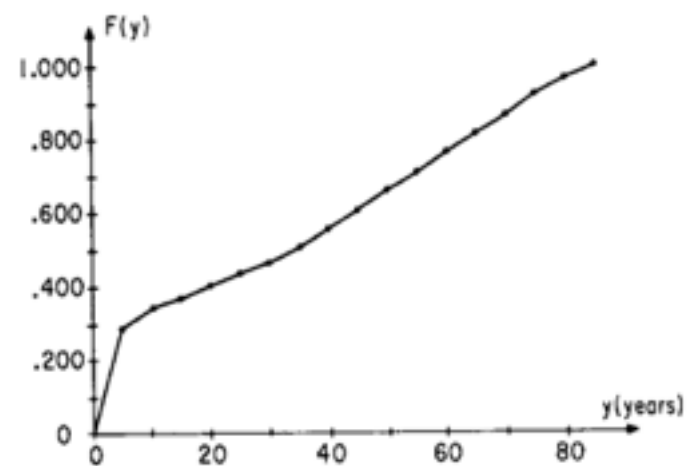
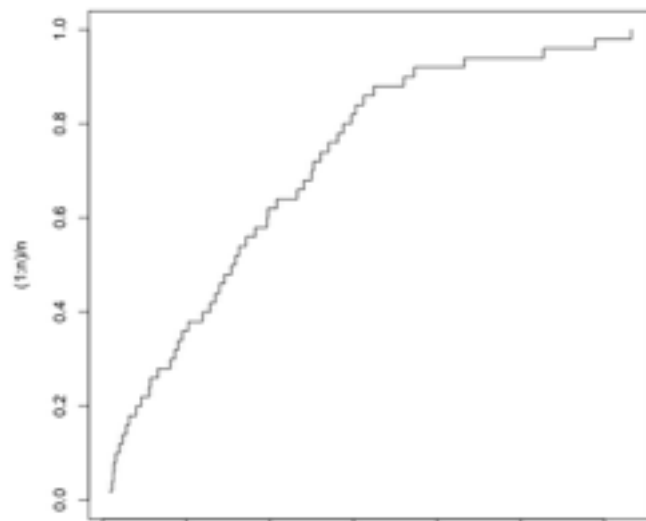


Lze použít například Sturgessovo pravidlo pro volbu počtu tříd: $k = 1 + \frac{1}{3} \log_{10}(k)$

Testy dobré shody

Co s tím?

2) Empirická distribuční funkce - poskytuje předběžnou představu o tvaru distribuční funkce



Testy dobré shody

Co s tím?

3) Informaci o rozdělení nám poskytují i výběrové momenty:

1. výběrový moment = aritmetický průměr:
(bodový odhad střední hodnoty) $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
2. výběrový centrální moment = výběrový rozptyl $m_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$
3. výběrový centrální moment
(bodový odhad koeficientu šikmosti: $S_{kew} = \frac{m_3}{\sqrt{m_2^3}}$) $m_3 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^3$
4. výběrový centrální moment
(bodový odhad koeficientu špičatosti: $K_{urt} = \frac{m_4}{m_2^2}$) $m_4 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^4$

Testy dobré shody

Co dál?

1) Grafická analýza

- histogram, boxplot, empirická distribuční funkce

- pravděpodobnostní papír

osa x: lineární

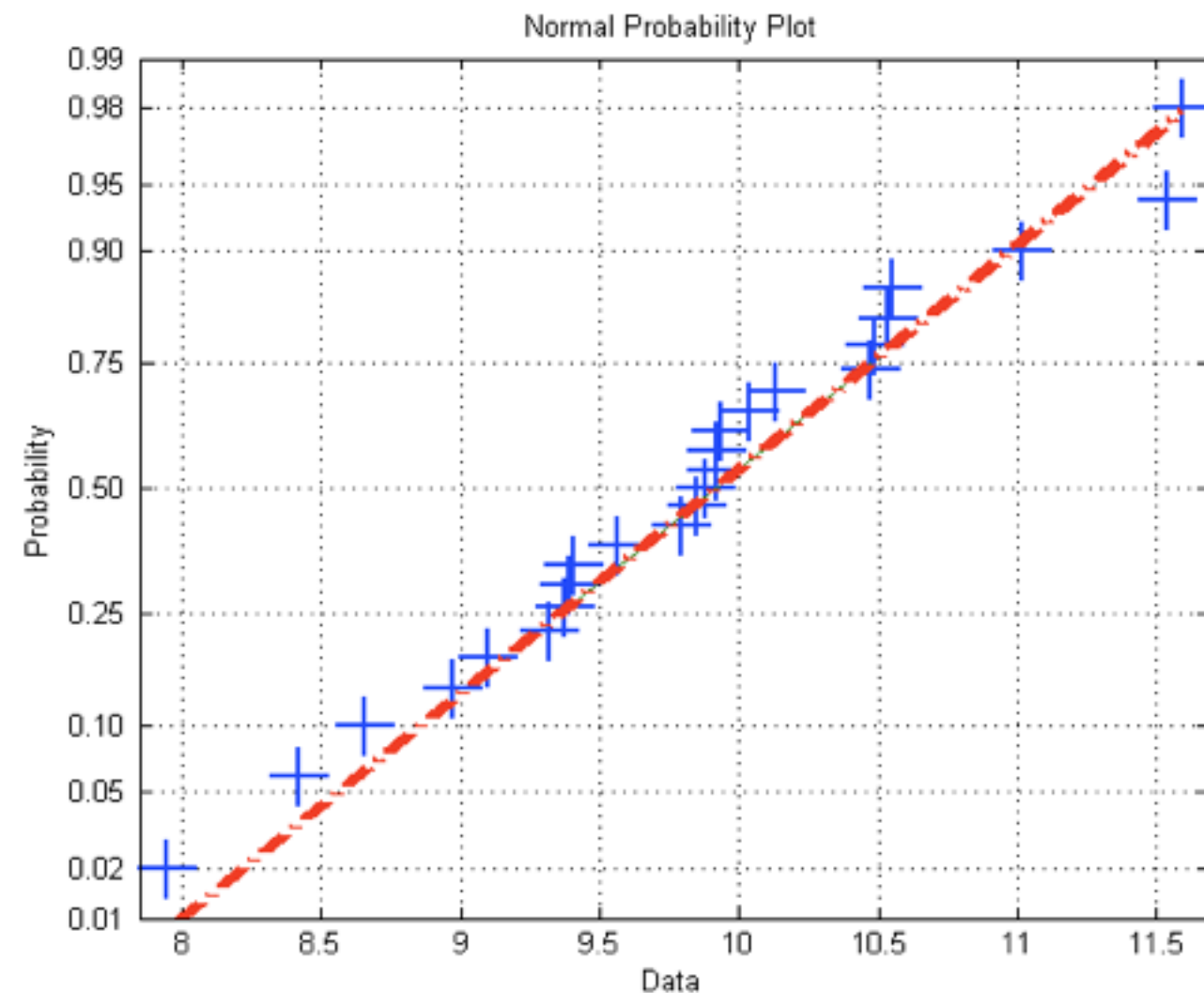
osa Y: transformované

“pravděpodobnostní”

měřítko

Zakreslujeme dvojice

$(x_{(i)}, i/n)$



Testy dobré shody

Co dál?

1) Grafická analýza

- histogram, boxplot, empirická distribuční funkce

- pravděpodobnostní papír

osa x: lineární

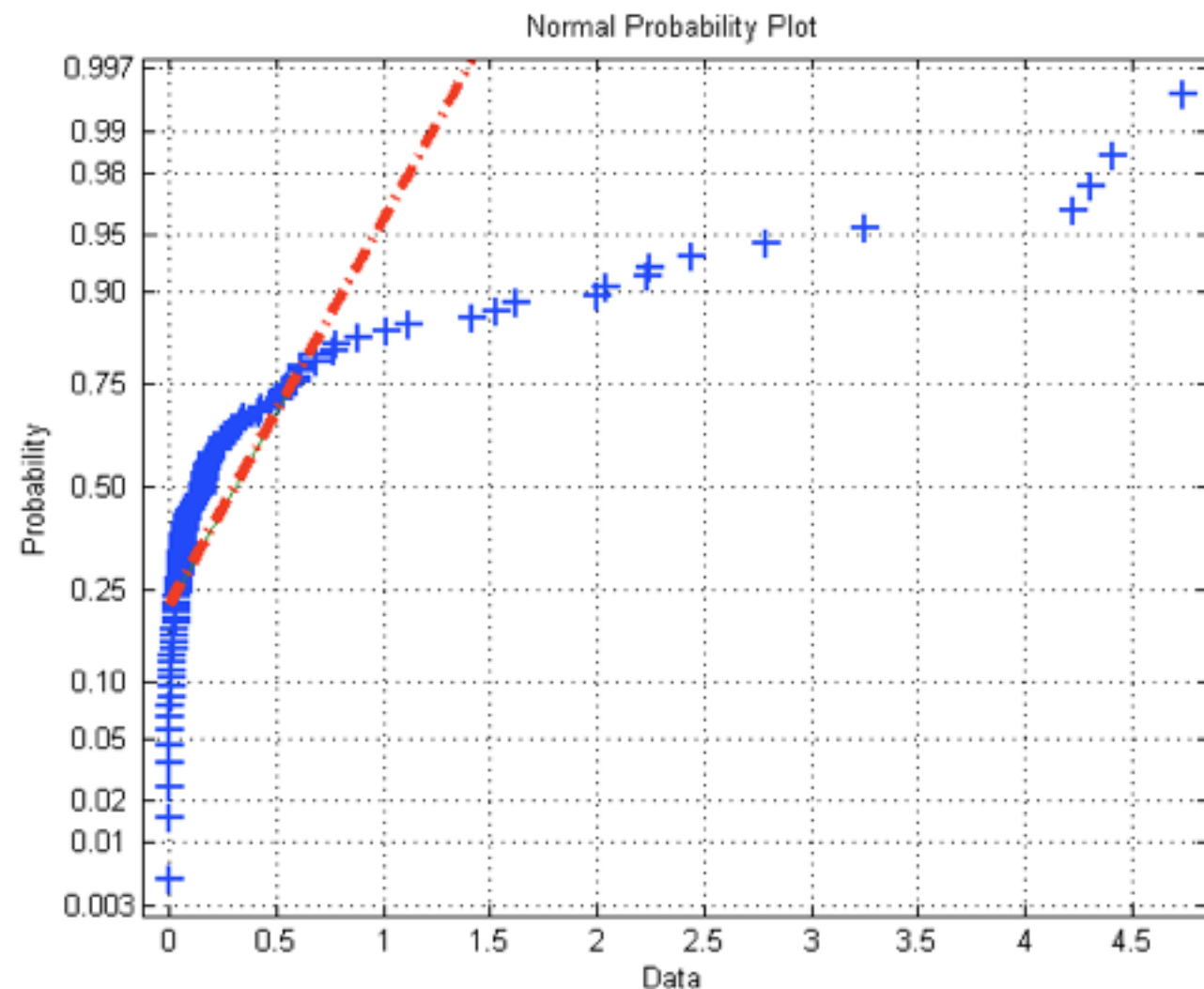
osa Y: transformované

“pravděpodobnostní”

měřítko

Zakreslujeme dvojice

$(x_{(i)}, i/n)$



Testy dobré shody

Co dál?

1) Grafická analýza

- histogram, boxplot, empirická distribuční funkce

- pravděpodobnostní papír

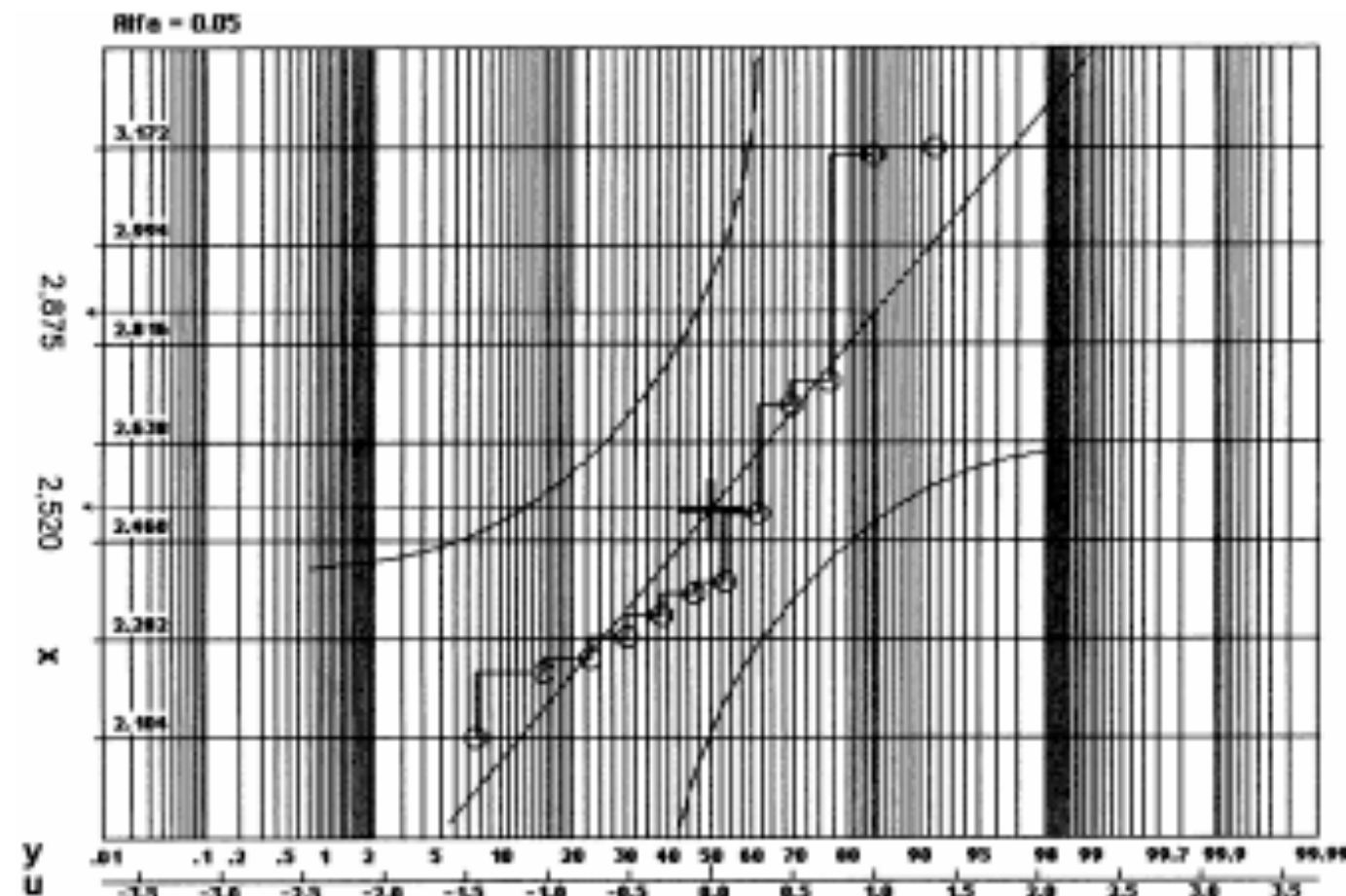
osa x: lineární

osa Y: transformované

“pravděpodobnostní”
měřítko

Zakreslujeme dvojice

$(x_{(i)}, i/n)$



Testy dobré shody

Co dál?

1) Grafická analýza

- histogram, boxplot, empirická distribuční funkce

- pravděpodobnostní papír

osa x: lineární

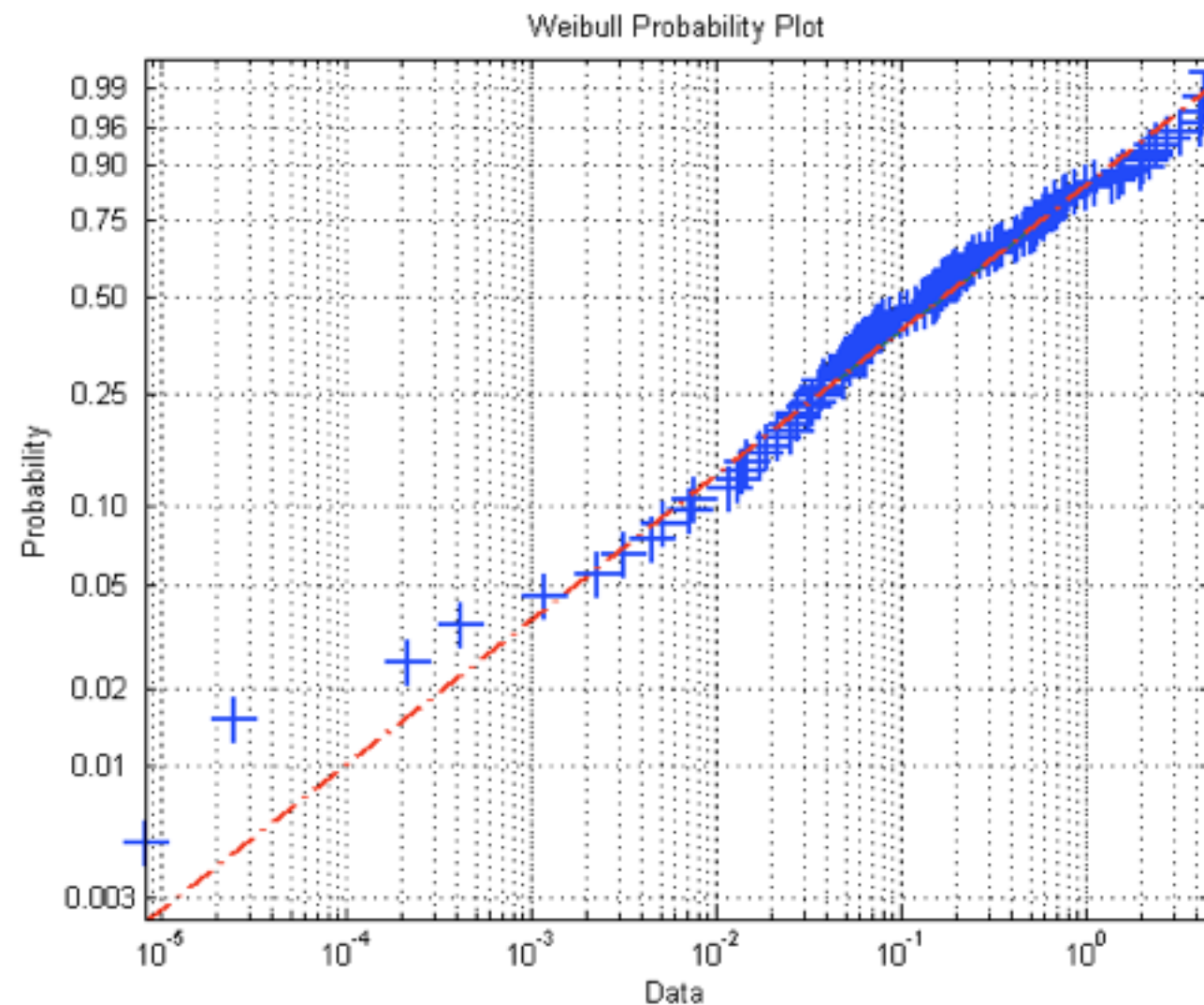
osa Y: transformované

“pravděpodobnostní”

měřítko

Zakreslujeme dvojice

$(x_{(i)}, i/n)$

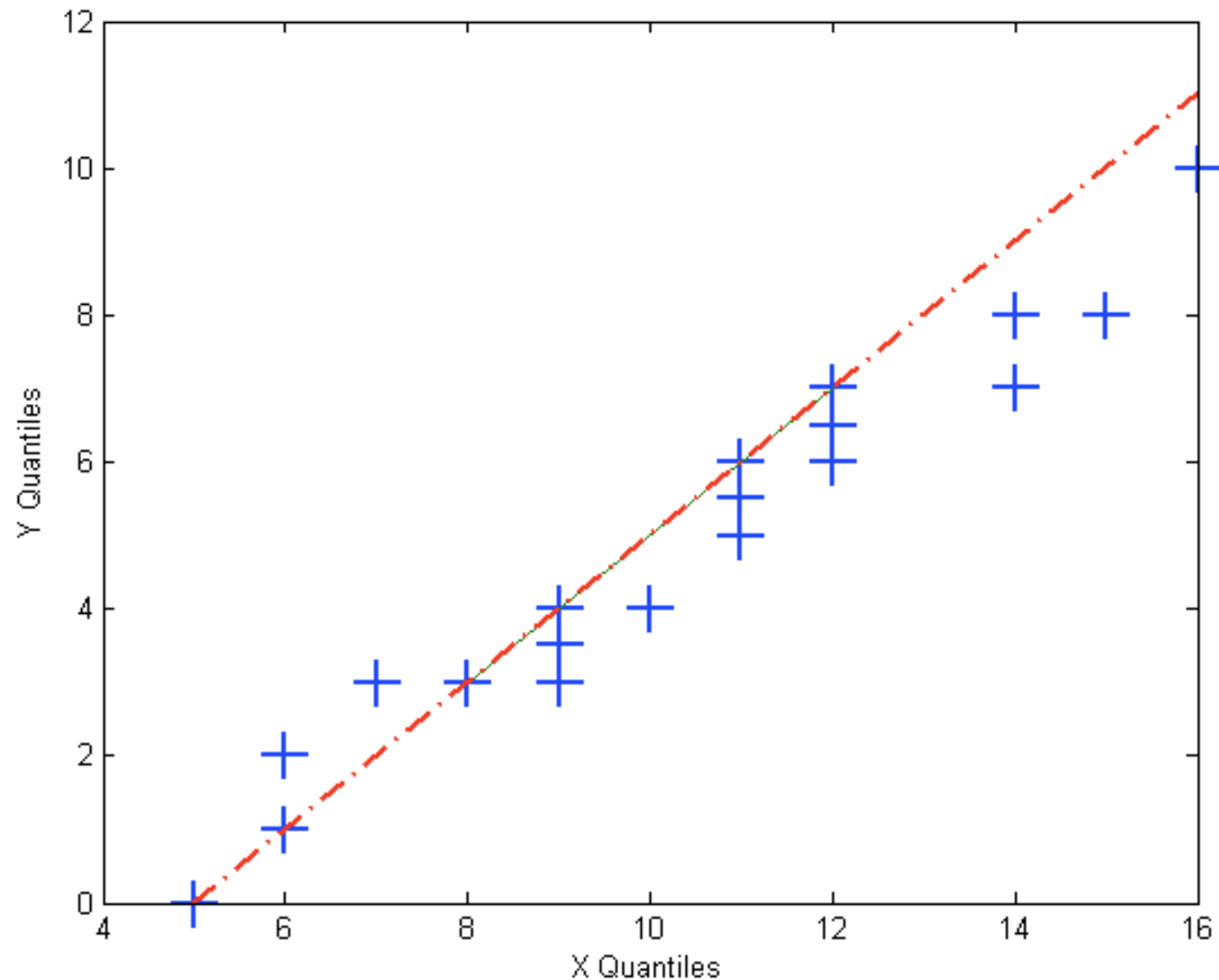


Testy dobré shody

Co dál?

1) Grafická analýza

- Q-Q graf
osa x: měření
osa y: kvantily
hypotetické d.f



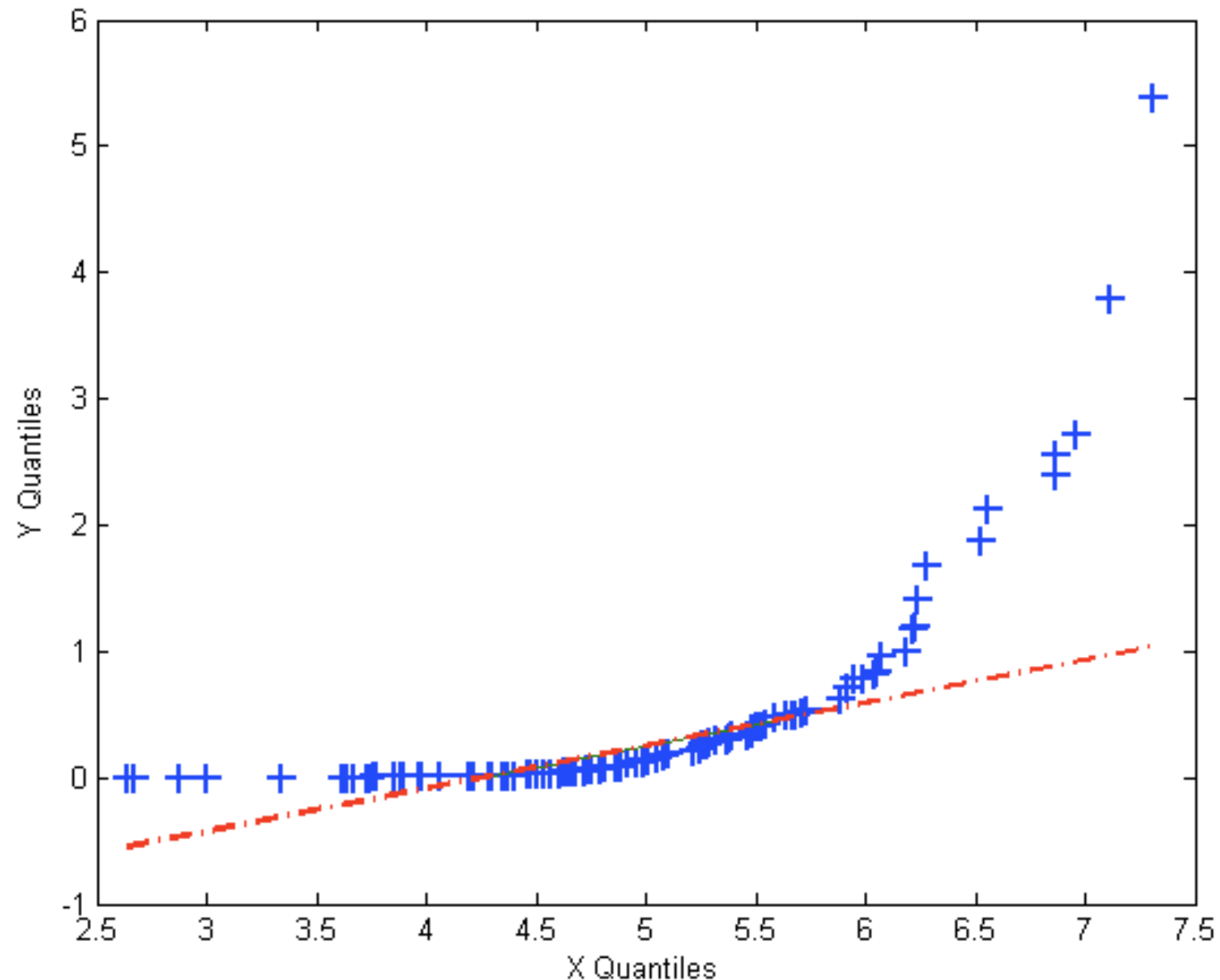
Zakreslujeme dvojice $(x_{(i)}, F^{-1}(i/n))$

Testy dobré shody

Co dál?

1) Grafická analýza

- Q-Q graf
osa x: měření
osa y: kvantily
hypotetické d.f.



Pomocí grafické analýzy můžeme metodou srovnání se standardními modely pouze odhadnout typ rozdělení

Testy dobré shody

Co dál?

2) Kvantitativní testy hypotézy o daném typu rozdělení

nulová hypotéza : $H_0 : F(x) = F_0(x)$

alternativní hypotéza: $H_A : F(x) \neq F_0(x)$

testová statistika : $T(X_1, X_2, \dots, X_n)$

hladina významnosti: α

chyba 1. druhu: zamítneme hypotézu, která platí

chyba 2. druhu: nezamítneme hypotézu, která neplatí

hladina významnosti testu: pravděpodobnost chyby 1. druhu

síla testu: pravděpodobnost zamítnutí hypotézy, když neplatí

p-hodnota: nejmenší hladina významnosti, při které bychom ještě zamítli nulovou hypotézu.

Testy dobré shody

Co dál?

2) Kvantitativní testy hypotézy o daném typu rozdělení

Chí-kvadrát test dobré shody

Kolmogorov-Smirnovův test

Testy normality (Shapiro-Wilkův test, testy na základě šikmosti a špičatosti, Lilieforsův, Anderson-Darlingův test)

Kvantitativní statistické testy nám poskytnou objektivní míru shody dat s teoretickým modelem

Chí-kvadrát test dobré shody

Test srovnává empirické a teoretické četnosti při zadaném třídění:

- i) provedeme roztrídění naměřených hodnot do k tříd
- ii) napočítáme empirické četnosti n_1, n_2, \dots, n_k
- iii) napočítáme pravděpodobnosti tříd p_1, p_2, \dots, p_k při hypotetickém rozdělení (kde $p_j = F(x_{j+1}) - F(x_j)$)
- iv) napočítáme teoretické četnosti np_1, np_2, \dots, np_k
- v) pokud pro všechna $j=1, 2, \dots, k$ platí $np_j > 5$, spočítáme hodnotu testové statistiky

$$\chi^2 = \sum_{j=1}^k \frac{(n_j - np_j)^2}{np_j}$$

- vi) neplatí-li podmínka v bodě (v), provedeme úpravu třídních intervalů (nemusejí být stejně velké)

Chí-kvadrát test dobré shody

Test srovnává empirické a teoretické četnosti při zadaném třídění pomocí testové statistiky

$$\chi^2 = \sum_{j=1}^k \frac{(n_j - np_j)^2}{np_j}$$

- vii) známe-li parametry hypotetického rozdělení předem, bude mít testová statistika rozdělení $\chi^2(k-1)$ a nulovou hypotézu zamítneme, pokud bude $\chi^2 \geq \chi_{1-\alpha}^2(k-1)$, kde $\chi_{1-\alpha}^2(k-1)$ je $(1-\alpha)$ -kvantil chí-kvadrát rozdělení o $(k-1)$ stupních volnosti.
- viii) pokud neznámé parametry hypotetického rozdělení odhadujeme z naměřených dat, bude mít testová statistika chí-kvadrát rozdělení o $(k-r-1)$ stupních volnosti, kde r je počet odhadovaných parametrů. Nulovou hypotézu v tomto případě zamítneme, pokud bude
- $$\chi^2 \geq \chi_{1-\alpha}^2(k-r-1)$$

Kolmogorov-Smirnovův test dobré shody

Test srovnává empirickou a teoretickou distribuční funkci pomocí maximálního rozdílu hodnot.

- i) seřadíme n naměřených hodnot podle velikosti od nejmenší do největší
- ii) pro každou hodnotu $x_{(i)}$ spočteme rozdíly

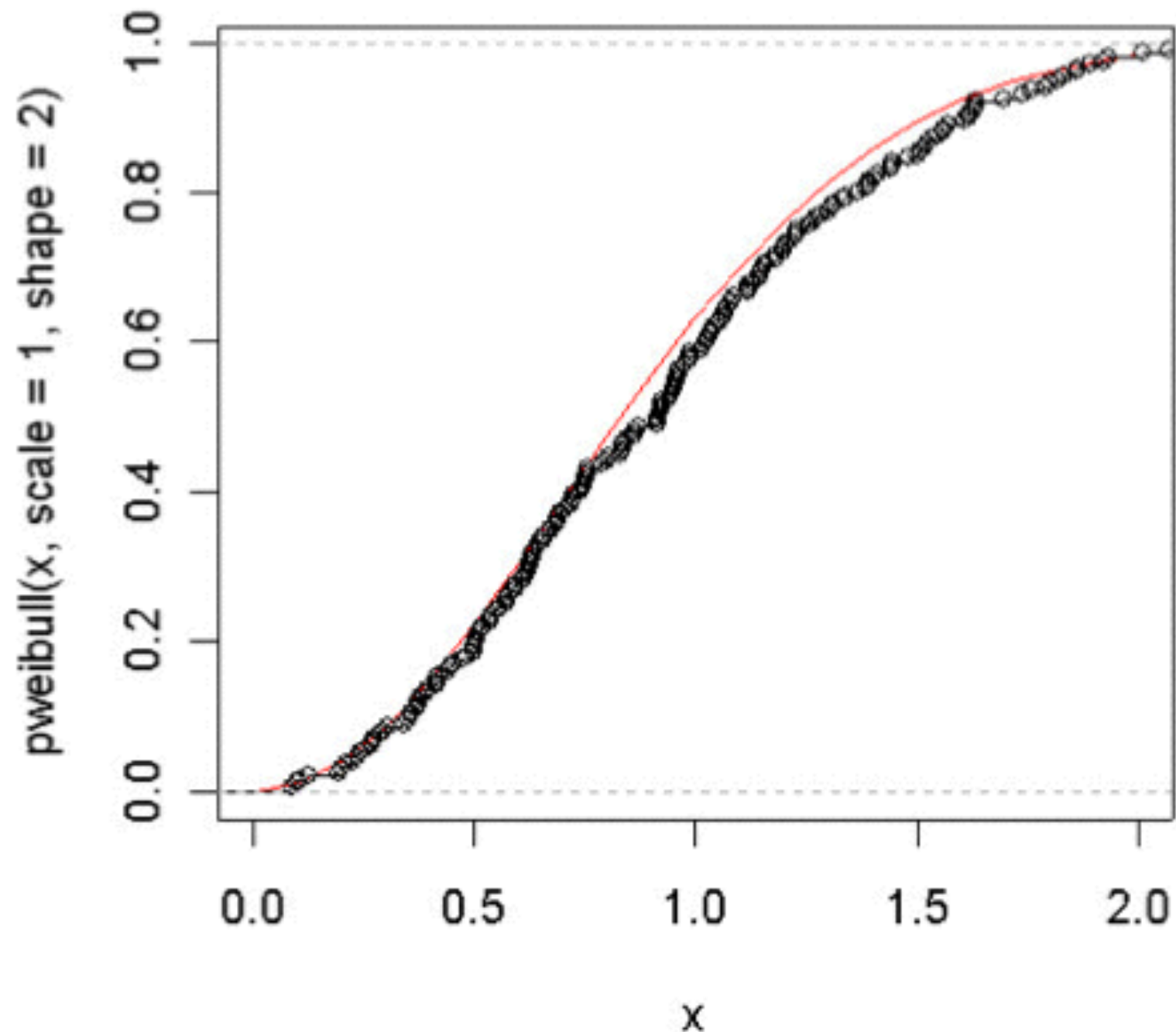
$$\left| F_0(x_{(i)}) - \frac{i}{n} \right|, \quad \left| F_0(x_{(i)}) - \frac{i-1}{n} \right|$$

- iii) největší z těchto rozdílů je hodnota testové statistiky $D(n)$
- iv) pokud je hypotetické rozdělení známé včetně parametrů, použijeme krok (v). Jinak musíme použít některou z modifikací K-S testu (Liliefors, Anderson-Darling)
- v) pro malá n tuto hodnotu porovnáme s tabulkovou kritickou hodnotou $d_{1-\alpha}(n)$ pro K-S-test. Pro velká n můžeme použít aproximaci $d_{1-\alpha}(n) = \sqrt{(1/2n) \ln(2/\alpha)}$

Pokud je $D(n) \geq d_{1-\alpha}(n)$, nulovou hypotézu zamítáme.

Kolmogorov-Smirnovův test dobré shody

```
> x<-seq(0,2,0.1)
> plot(x,pweibull(x,scale=1,shape=2),type="l",col="red")
> plot(ecdf(x.wei),add=TRUE)
```



Kolmogorov-Smirnovův test dobré shody

```
> ks.test(x.wei, "pweibull", shape=2, scale=1)
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: x.wei D = 0.0623, p-value = 0.4198
```

```
alternative hypothesis: two.sided
```

Testy normality

Testy na základě šikmosti a špičatosti

Za předpokladu, že výběr pochází z normálního rozdělení, platí pro index šikmosti:

$$E(S_{kew}^{norm}) = 0$$

$$Var(S_{kew}^{norm}) = \frac{6(n-2)}{(n+1)(n+3)}$$

a pro index špičatosti:

$$E(K_{urt}^{norm}) = 3 - \frac{6}{n+1}$$

$$Var(K_{urt}^{norm}) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

Máme-li dostatečný počet pozorování (řádově stovky), mají statistiky

$$T_3 = \frac{S_{kew}^{norm}}{\sqrt{Var(S_{kew}^{norm})}} \quad T_4 = \frac{K_{urt}^{norm} - E(K_{urt}^{norm})}{\sqrt{Var(K_{urt}^{norm})}}$$

přibližně standardní normální rozdělení pravděpodobnosti.

Testy normality

Testy na základě šikmosti a špičatosti

Tedy hypotézu o normalitě na základě šikmosti zamítáme, pokud bude platit $|T_3| \geq u_\alpha$, nebo pokud bude $p \leq \alpha$, kde $p = 2 \min\{\Phi(T_3), 1 - \Phi(T_3)\}$

Hypotézu o normalitě na základě špičatosti zamítáme, pokud bude platit $|T_4| \geq u_\alpha$, nebo pokud bude $p \leq \alpha$, kde $p = 2 \min\{\Phi(T_4), 1 - \Phi(T_4)\}$

Oba testy by se měly používat současně, proto se často používá kombinovaný test s testovou statistikou $T_{34} = T_3^2 + T_4^2$, která má χ^2 -rozdělení o 2 stupních volnosti. Hypotézu o normalitě potom zamítáme, když $T_{34} \geq \chi_\alpha^2(2)$

Testy normality

Shapirův-Wilkův test

Jeden z nejsilnějších testů normality

$$SW = \frac{\left[\sum_{i=1}^n a_{(i)} x_{(i)} \right]^2}{\sum_{i=1}^n a_{(i)}^2 \sum_{i=1}^n (x_{(i)} - \bar{x})^2}$$

kde $a_{(i)} = \Phi^{-1}\left(\frac{8i-3}{8n+2}\right)$ a kritické hodnoty jsou tabelovány.

=> pro aplikaci tohoto testu potřebujete tabulky a počítač, případně specializovaný statistický software.

```
> shapiro.test(x.norm)
```

```
Shapiro-Wilk normality test
```

```
data: x.norm W = 0.9938, p-value = 0.5659
```

Testy normality

Shapirův-Wilkův test

Jeden z nejsilnějších testů normality

$$SW = \frac{\left[\sum_{i=1}^n a_{(i)} x_{(i)} \right]^2}{\sum_{i=1}^n a_{(i)}^2 \sum_{i=1}^n (x_{(i)} - \bar{x})^2}$$

kde $a_{(i)} = \Phi^{-1} \left(\frac{8i - 3}{8n + 2} \right)$ a kritické hodnoty jsou tabelovány.

=> pro aplikaci tohoto testu potřebujete tabulky a počítač, případně specializovaný statistický software.

Lilieforsův test

Testová statistika je totožná s Kolmogorov-Smirnovovým testem, parametry hypotetického rozdělení odhadujeme z dat a kritické hodnoty hledáme v tabulkách

Testy normality

Lilieforsův test

Testová statistika je totožná s Kolmogorov-Smirnovovým testem, parametry hypotetického rozdělení odhadujeme z dat a kritické hodnoty hledáme v tabulkách

Anderson-Darlingův test

Test, který je modifikací Kolmogorovova-Smirnovova testu (používá empirickou distribuční funkci a uspořádaný výběr) s testovou statistikou

$$AD = - \frac{\sum_{i=1}^n (2i - 1) (\ln F_0(x_{(i)}) + \ln(1 - F_0(x_{(n-i+1)})))}{n} - n$$

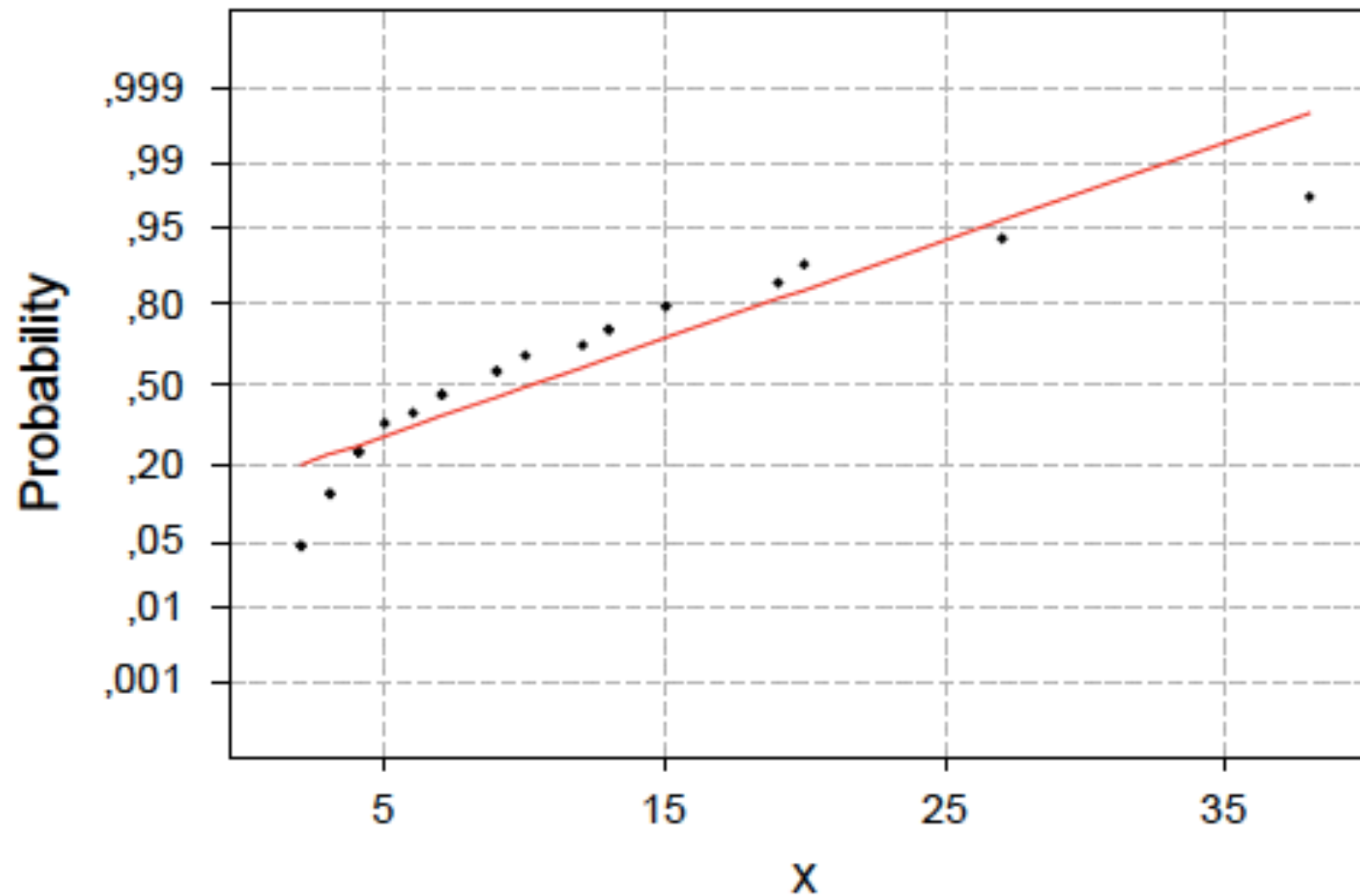
Kritické hodnoty jsou pro malá n tabelovány, pro velká n lze použít aproximaci $ad_{0,95} = 1,0348(1 - 1,013/n - 0,93/n^2)$

=> pro aplikaci tohoto testu potřebujete tabulky a počítač, případně specializovaný statistický software.

Testy normality

Anderson-Darlingův test

Normal Probability Plot



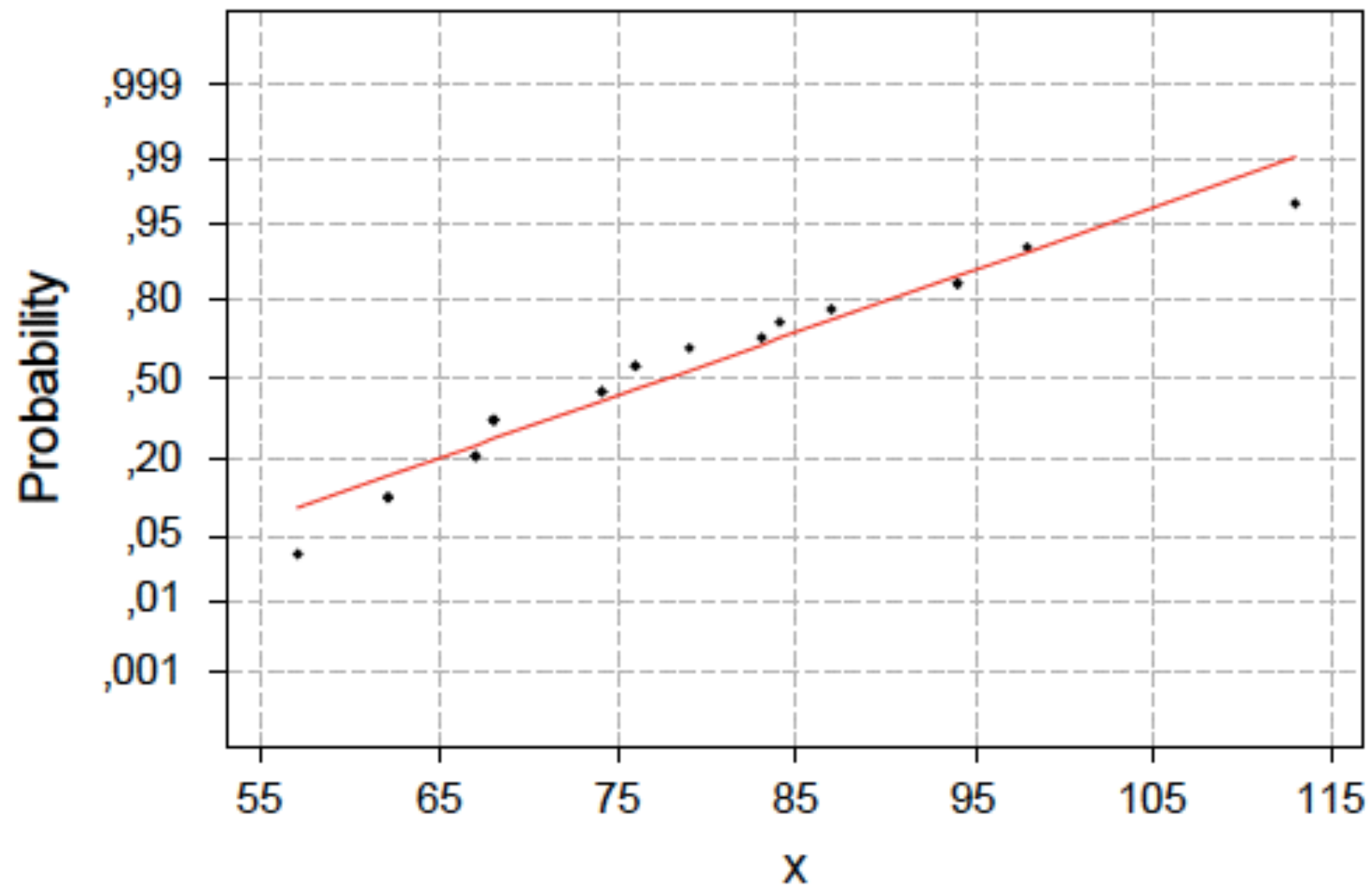
Average: 10,32
StDev: 8,57185
N: 25

Anderson-Darling Normality Test
A-Squared: 1,276
P-Value: 0,002

Testy normality

Anderson-Darlingův test

Normal Probability Plot



Average: 77,55
StDev: 14,1625
N: 20

Anderson-Darling Normality Test
A-Squared: 0,455
P-Value: 0,240

Testy dobré shody

24.52586 24.17119 24.54486 24.44240 23.93455 24.20389 24.19974 24.34851 23.94024 24.21022
24.87474 25.06155 25.48924 25.32572 23.71721 24.61622 25.06676 24.90055 24.36213 24.98580
24.80591 24.20853 24.72623 24.64437 24.70405 23.97645 25.29837 24.46910 24.99453 25.42994
24.66147 24.75773 25.03970 24.44901 25.13285 24.40205 24.78721 23.83656 24.17186 23.65390
24.48244 24.68550 24.22988 23.83956 24.09777 24.52098 24.89240 24.25332 24.14259 25.12906

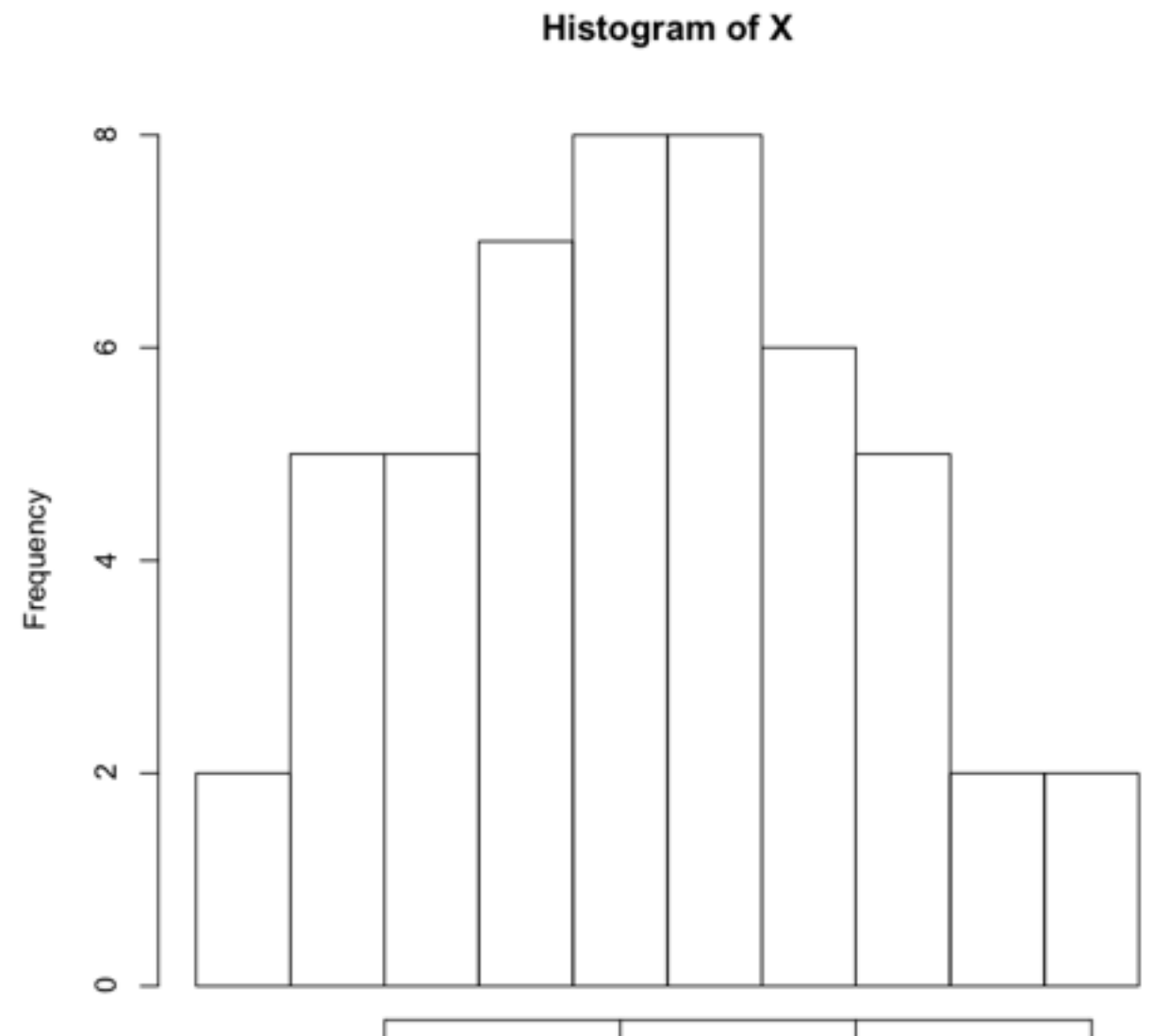
$$H_0 : F(x) = F_{N(24,55;0,21024)}(x)$$

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i = 24.54689$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = 0,2102477$$

$$s = \sqrt{0,2102477} = 0,4585$$

$$N(24,55;0,2102)$$

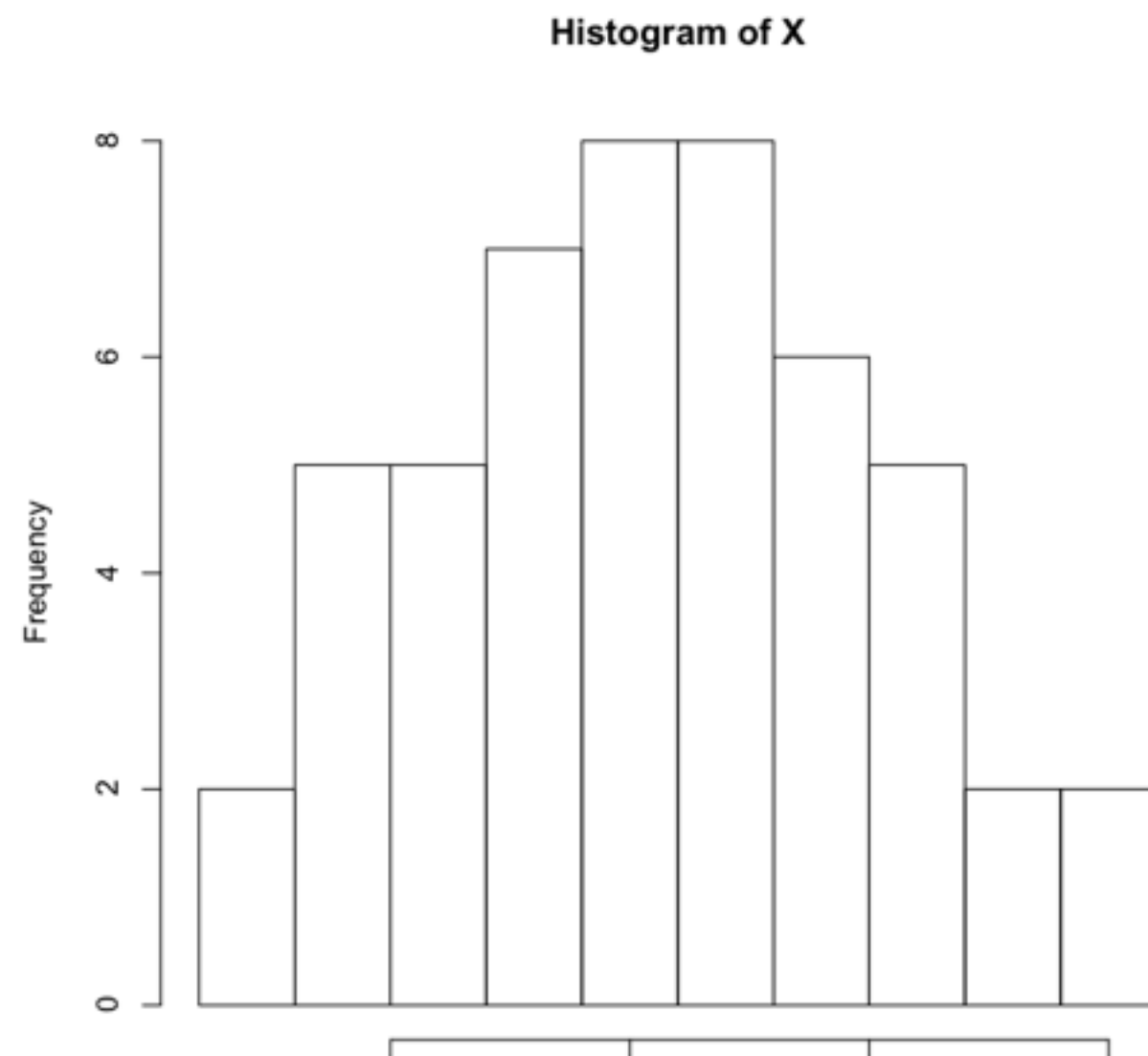


Testy dobré shody

24.52586 24.17119 24.54486 24.44240 23.93455 24.20389 24.19974 24.34851 23.94024 24.21022
 24.87474 25.06155 25.48924 25.32572 23.71721 24.61622 25.06676 24.90055 24.36213 24.98580
 24.80591 24.20853 24.72623 24.64437 24.70405 23.97645 25.29837 24.46910 24.99453 25.42994
 24.66147 24.75773 25.03970 24.44901 25.13285 24.40205 24.78721 23.83656 24.17186 23.65390
 24.48244 24.68550 24.22988 23.83956 24.09777 24.52098 24.89240 24.25332 24.14259 25.12906

$$H_0 : F(x) = F_{N(24,55;0,21024)}(x)$$

	i	ni	pi	npi	(ni-npi) ² /npi
23,6	23,8	2	0,0634	3,17	0,4323
23,8	24	5	0,0743	3,72	0,4433
24	24,2	4	0,1187	5,94	0,6312
24,2	24,4	8	0,1572	7,86	0,0025
24,4	24,6	8	0,1727	8,63	0,0463
24,6	24,8	8	0,1572	7,86	0,0025
24,8	25	6	0,1187	5,94	0,0007
25	25,2	5	0,0743	3,72	0,4433
25,2	25,4	2	0,0386	1,93	0,0026
25,4	26	2	0,0248	1,24	0,4636
suma		50	1,0000	50,00	2,4684



Testy dobré shody

24.52586 24.17119 24.54486 24.44240 23.93455 24.20389 24.19974 24.34851 23.94024 24.21022
 24.87474 25.06155 25.48924 25.32572 23.71721 24.61622 25.06676 24.90055 24.36213 24.98580
 24.80591 24.20853 24.72623 24.64437 24.70405 23.97645 25.29837 24.46910 24.99453 25.42994
 24.66147 24.75773 25.03970 24.44901 25.13285 24.40205 24.78721 23.83656 24.17186 23.65390
 24.48244 24.68550 24.22988 23.83956 24.09777 24.52098 24.89240 24.25332 24.14259 25.12906

$$H_0 : F(x) = F_{N(24,55;0,21024)}(x)$$

	i	ni	pi	npi	(ni-npi) ² /npi
23,6	24	7	0,1377	6,89	0,0018
24	24,2	4	0,1187	5,94	0,6312
24,2	24,4	8	0,1572	7,86	0,0025
24,4	24,6	8	0,1727	8,63	0,0463
24,6	24,8	8	0,1572	7,86	0,0025
24,8	25	6	0,1187	5,94	0,0007
25	26	9	0,1377	6,89	0,6482
	suma	50	1,0000	50,00	1,3332

$$\chi^2 = 1,3332 \leq \chi_{0,95}(47) = 32,3$$

