12 TESTOVÁNÍ NEPARAMETRICKÝCH HYPOTÉZ

Dosud jsme se zabývali testováním parametrický hypotéz, což jsou hypotézy o parametrech rozdělení (populace). Statistickým hypotézám o jiných vlastnostech populace (tvar rozdělení, závislost proměnných…) se říká **neparametrické hypotézy.**

Zaměříme se na některé z tzv. **testů dobré shody**.

χ^2 – test dobré shody

Volba nulové hypotézy

Test dobré shody se používá nejčastěji pro ověřování těchto hypotéz:

a) H_0: Výběr pochází z populace, v níž jsou relativní četnosti jednotlivých variant rovny číslům $\pi_{0,1}; \pi_{0,2}; \ldots; \pi_{0,k}$ (populace musí být roztříditelná podle nějakého znaku do k skupin)

b) H_0: Výběr pochází z rozdělení určitého typu (např. normální), jehož parametry jsou dány (**úplně specifikovaný model**)

c) H_0: Výběrový soubor pochází z rozdělení určitého typu (např. normální)
(**neúplně specifikovaný model** – neověřujeme informace o parametrech rozdělení, parametry modelu odhadujeme)

Volba testové statistiky

Jako testovou statistiku volíme statistiku G, která má pro dostatečný rozsah výběru asymptoticky χ^2_{k-h-1} rozdělení:

$$T(\underline{X}) = G = \sum_{i=1}^{k} \frac{(n_i - n \cdot \pi_{0,i})^2}{n \cdot \pi_{0,i}} \rightarrow \chi^2_{k-h-1},$$

kde **n** je rozsah výběru, **k** je počet variant, **h** je počet odhadovaných parametrů modelu, $\mathbf{n_i}$ jsou skutečné četnosti jednotlivých variant a $\boldsymbol{\pi_{0,i}}$ jsou očekávané relativní četnosti (tj. relativní četnosti, jichž by měly nabýt jednotlivé varianty v případě, že je splněna nulová hypotéza).

$\mathbf{n.\pi_{0,i}}$ jsou tedy očekávané četnosti jednotlivých variant (tj. četnosti, jichž by měly nabýt jednotlivé varianty v případě, že je splněna nulová hypotéza) a $\mathbf{(n_i - n.\pi_{0,i})}$ pak jsou odchylky očekávaných četností od četností skutečných.

Za výběr dostatečného rozsahu považujeme výběr, pro nějž platí, že všechny očekávané četnosti jsou vyšší než 5 ($n \cdot \pi_{0,i} > 5$ (i = 1,2, …, k))

Výpočet p-value

Při tomto testu určujeme p-value jako: $p - value = 1 - F_0(x_{OBS})$

12.1. Hodilo se 6000 krát hrací kostkou a zaznamenaly se počty padlých ok...

x_i (číslo které padlo)	1	2	3	4	5	6
n_i (četnost jeho výskytu)	979	1002	1015	980	1040	984

Je možné na základě příslušného testu na hladině významnosti 5% spolehlivě tvrdit, že kostka je "falešná", tj. že pravděpodobnosti všech čísel na kostce nejsou stejné?

Řešení:

Musíme testovat, zda rozdělení „počtu ok“ padlých na kostce je takové, že pravděpodobnosti všech možných hodnot jsou 1/6.

Pro tento test dobré shody doporučujeme použít χ^2 test dobré shody (H_0 je ve tvaru a)):

Volba nulové a alternativní hypotézy

H_0: Pravděpodobnost „počtu ok“ na kostce je dána následující tabulkou:

x_i (číslo které může padnout)	1	2	3	4	5	6
$\pi_{0,i}$ (nulová pravděpodobnost jeho výskytu)	1/6	1/6	1/6	1/6	1/6	1/6

H_A: $\overline{H_0}$ **,** tj. pravděpodobnost „počtu ok“ na kostce je jiná než je uvedeno ve výše uvedené tabulce

Volba testové statistiky

Rozsah výběru: n = 6000
Počet variant: k = 6
Počet odhadovaných parametrů: h = 0

$$\pi_{0,1} = \pi_{0,2} = \ldots = \pi_{0,6} = 1/6 \quad \Rightarrow n \cdot \pi_{0,1} = n \cdot \pi_{0,2} = \ldots = n \cdot \pi_{0,6} = 1000 \Rightarrow 1000 > 5 \Rightarrow$$

Rozsah výběru je dostatečný proto, abychom mohli použít testovou statistiku G

$$T(\underline{X}) = G = \sum_{i=1}^{k} \frac{(n_i - n \cdot \pi_{0,i})^2}{n \cdot \pi_{0,i}} \rightarrow \chi^2_{k-h-1}$$

Výpočet pozorované hodnoty x_{OBS}:

x_i (číslo které padlo)	1	2	3	4	5	6
n_i (četnost jeho výskytu)	979	1002	1015	980	1040	984
$n.\ \pi_{0,i}$ (očekávaná četnost jeho výskytu)	1000	1000	1000	1000	1000	1000

$$x_{OBS} = T(\underline{X})_{H_0} = G_{H_0} = \sum_{i=1}^{k} \frac{(n_i - n \cdot \pi_{0,i})^2}{n \cdot \pi_{0,i}} = \frac{(979-1000)^2}{1000} + \frac{(1002-1000)^2}{1000} + \ldots + \frac{(984-1000)^2}{1000} = 2{,}93$$

Výpočet p-value:

$$p - value = 1 - F_0(x_{OBS})$$
$$F_0(x_{OBS}) = F_0(2{,}93)$$

$0{,}250 < F_0(2{,}93) < 0{,}500$ (viz. Tabulka 3, počet stupňů volnosti je 5 (6-1))

$0{,}500 < 1 - F_0(2{,}93) < 0{,}750$

$0{,}500 < p - value < 0{,}750$

Rozhodnutí:

$p - value > 0{,}05 \Rightarrow$ Nezamítáme nulovou hypotézu, tj. nelze tvrdit, že kostka je „falešná“.

12.2. Výrobní firma odhaduje počet poruch určitého zařízení během 100 hodin pomocí Poissonova rozdělení s parametrem 1,2. Zaměstnanci zaznamenali pro kontrolu skutečné počty poruch celkem ve 150-ti 100 hodinových intervalech (výsledky jsou uvedeny v tabulce). Ověřte čistým testem významnosti, zda má počet poruch daného zařízení během 100 hodin skutečně Poissonovo rozdělení s parametrem λt=1,2.

x_i – počet poruch během 100 hodin provozu	0	1	2	3	4
n_i - počet pozorování	52	48	36	10	4

Řešení:

Musíme testovat, zda počet poruch daného zařízení během 100 hodin má skutečně Poissonovo rozdělení s parametrem 1,2. Pro tento test dobré shody doporučujeme použít χ^2 test dobré shody (H_0 je ve tvaru b) – tj. jde o **úplně specifikovaný model** (víme jaký má být parametr rozdělení)):

Definujme si náhodnou veličinu X jako počet poruch daného zařízení během 100 hodin provozu.

Volba nulové a alternativní hypotézy

H_0: Počet poruch daného zařízení během 100 hodin (náhodná veličina X) má Poissonovo rozdělení s parametrem 1,2

H_A: $\overline{H_0}$, tj. počet poruch daného zařízení během 100 hodin (náhodná veličina X) nemá Poissonovo rozdělení s parametrem λ=1,2

Volba testové statistiky

Rozsah výběru: n = 150
Počet variant: k = 5
Počet odhadovaných parametrů: h = 0

Pokud platí H_0, pak X (počet poruch během 100 hodin) má Poisoonovo rozdělení se střední hodnotou 1,2 (= λt). Na základě této informace můžeme určit nulové pravděpodobnosti $\pi_{0,i}$.

$$\pi_{0,i} = P(X = x_i) = \frac{(\lambda t)^{x_i}}{x_i!} \cdot e^{-\lambda t} = \frac{(1{,}2)^{x_i}}{x_i!} \cdot e^{-1{,}2}$$

Zároveň si určíme očekávané četnosti.

x_i – počet poruch během 100 hodin provozu	0	1	2	3	4
n_i – počet pozorování	52	48	36	10	4
$\pi_{0,i}$	0,301	0,361	0,217	0,087	0,034
$n.\pi_{0,i}$ - očekávané četnosti	45,2	54,2	32,6	13,1	5,1

Všechny očekávané četnosti jsou větší než 5, tudíž rozsah výběru je dostatečný proto, abychom mohli použít testovou statistiku G

$$T(\underline{X}) = G = \sum_{i=1}^{k} \frac{(n_i - n \cdot \pi_{0,i})^2}{n \cdot \pi_{0,i}} \rightarrow \chi^2_{k-h-1}$$

Výpočet pozorované hodnoty x_{OBS}:

$$x_{OBS} = T(\underline{X})_{H_0} = G_{H_0} = \sum_{i=1}^{k} \frac{(n_i - n \cdot \pi_{0,i})^2}{n \cdot \pi_{0,i}} =$$

$$= \frac{(52-45{,}2)^2}{45{,}2} + \frac{(48-54{,}2)^2}{54{,}2} + \ldots + \frac{(4-5{,}1)^2}{5{,}1} = 3{,}13$$

Výpočet p-value:

H_A: $p-value = 1 - F_0(x_{OBS})$

$F_0(x_{OBS}) = F_0(3{,}13)$

$0{,}250 < F_0(3{,}13) < 0{,}500$ (viz. Tabulka 3, počet stupňů volnosti = 5-0-1 = 4)

$0{,}500 < 1 - F_0(3{,}13) < 0{,}750$

$0{,}500 < p-value < 0{,}750$

Rozhodnutí:

$p-value > 0{,}05 \Rightarrow$

Nezamítáme nulovou hypotézu, tzn. nemáme námitek proti použití Poissonova rozdělení s parametrem 1,2 pro odhad počtu poruch daného zařízení během 100 hodin provozu (toto rozdělení je vhodným modelem pro počet poruch).

12.3. Na dálnici byly v průběhu několika minut měřeny časové odstupy [s] mezi průjezdy jednotlivých vozidel. Zjištěné hodnoty těchto odstupů jsou v další tabulce:

2,5	**6,8**	**5,0**	**9,8**	**4,0**	**2,3**	**4,2**	**1,9**	**8,7**	**7,7**	**5,9**	**5,3**	**8,4**	**3,6**	**9,2**
4,3	**2,6**	**13,0**	**5,4**	**8,6**	**4,2**	**2,9**	**1,5**	**1,8**	**1,6**	**5,9**	**8,3**	**5,2**	**6,9**	**5,1**
1,3	**6,4**	**6,5**	**5,7**	**3,6**	**4,8**	**4,0**	**7,3**	**24,9**	**10,6**	**15,0**	**5,3**	**4,0**	**3,3**	**6,0**
4,6	**1,6**	**1,9**	**1,5**	**11,1**	**4,3**	**5,5**	**2,1**	**2,9**	**3,0**	**3,8**	**1,0**	**1,5**	**8,6**	**4,4**
6,8	**5,2**	**3,0**	**8,0**	**4,0**	**4,7**	**7,3**	**2,3**	**1,9**	**1,9**	**4,6**	**6,4**	**5,3**	**3,9**	**2,4**
1,2	**6,2**	**4,3**	**2,6**	**2,7**	**2,0**	**0,8**	**3,7**	**6,9**	**2,8**	**4,3**	**4,9**	**4,1**	**4,5**	**4,4**
11,9	**9,0**	**5,6**	**4,8**	**2,8**	**2,1**	**4,3**	**1,0**	**1,6**	**2,5**	**2,2**	**1,3**	**1,8**	**1,6**	**3,8**
3,1	**1,6**	**4,9**	**1,8**	**3,9**	**3,4**	**1,6**	**4,5**	**5,8**	**6,9**	**1,8**	**2,6**	**6,8**	**2,5**	**1,9**
3,1	**10,8**	**1,6**	**2,0**	**4,9**	**11,2**	**1,6**	**2,2**	**3,8**	**1,1**	**1,8**	**1,4**			

Otestujte čistým testem významnosti, zda lze časové odstupy mezi vozidly považovat za náhodnou veličinu s normálním rozdělením.

Řešení:

Nechť: náhodná veličina X je definována jako časový odstup mezi průjezdy jednotlivých vozidel.

Volba nulové a alternativní hypotézy:

H_0: Časové odstupy mezi průjezdy jednotlivých vozidel mají normální rozdělení.
H_A: Časové odstupy mezi průjezdy jednotlivých vozidel nemají normální rozdělení.

Volba testové statistiky:

Pokud se nám podaří splnit předpoklady pro χ^2 test dobré shody ($n \cdot \pi_{0,i} > 5$), můžeme řešit daný problém pomocí tohoto testu (H_0 bude vyjádřená ve verzi c) – **neúplně specifikovaný model**).

- Nejdříve **odhadneme parametry rozdělení** (μ odhadneme průměrem, σ odhadneme výběrovou směrodatnou odchylkou (nejlepší nestranné bodové odhady)):

  Rozsah výběru: n = 132

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{132} x_i}{132} = 4{,}6 \qquad \hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}} = 3{,}3$$

- V dalším kroku musíme **rozdělit data do „rozumného“ počtu intervalů a najít očekávané četnosti pro příslušné intervaly**. Na jejich základě rozhodneme, zda můžeme pro řešení daného problému použít χ^2 test dobré shody.

  Intervaly se volí většinou pouze na základě vlastní úvahy. Snažíme se však dodržovat několik pravidel:

  - Pokud je to možné, dodržujeme konstantní šířku intervalu (třídy)
  - Počet intervalů v „rozumných“ mezích. Obvykle se považuje za vhodné volit 5 až 15 intervalů. Počet intervalů nemá být ani příliš malý (vede k hrubému, zjednodušenému pohledu na rozdělení pravděpodobnosti), ani příliš velký (který dělá rozdělení pravděpodobnosti nepřehledným).
  - Intervaly nemusí mít stejnou šířku, avšak proto, abychom mohli použít χ^2 test dobré shody, musí být očekávané četnosti pro příslušné intervaly větší než 5.

Pokusíme se tedy rozdělit data do „rozumného“ počtu intervalů, najdeme očekávané četnosti pro příslušné intervaly a pak data přerozdělíme tak, aby byla splněna podmínka pro použití χ^2 testu dobré shody.

Jak spočítat očekávané četnosti?

Očekávané četnosti: $n \cdot \pi_{0,i}$

Očekávané relativní četnosti: $\pi_{0,i}$ určíme jako pravděpodobnosti výskytu náhodné veličiny X na příslušném intervalu (předpokládáme-li platnost H_0, známe rozdělení X (parametry tohoto rozdělení jsme odhadli). Pravděpodobnost, že náhodná veličina s normálním rozdělením ($N(\hat{\mu}; \hat{\sigma}^2)$) leží v i-tém intervalu je:

$$\pi_{0,i} = F(x_i) - F(x_{i-1}),$$

kde x_i je horní hranice intervalu a $x_0 = -\infty$.

Rozdělení do intervalů, příslušné očekávané relativní četnosti a očekávané četnosti

i	Časový interval [s]	Počet pozorování v časovém intervalu	Očekávané relativní četnosti $\pi_{0,i}$	Očekávané četnosti $n.\pi_{0,i}$
1	$(-\infty; 1,5\rangle$	11	0,174	22,9
2	$(1,5; 1,8\rangle$	13	0,024	3,2
3	$(1,8; 2,0\rangle$	7	0,017	2,3
4	$(2,0; 2,5\rangle$	10	0,047	6,2
5	$(2,5; 2,9\rangle$	8	0,041	5,4
6	$(2,9; 3,6\rangle$	8	0,078	10,3
7	$(3,6; 4,0\rangle$	10	0,047	6,2
8	$(4,0; 4,4\rangle$	10	0,048	6,3
9	$(4,4; 4,9\rangle$	10	0,060	8,0
10	$(4,9; 5,8\rangle$	12	0,106	14,0
11	$(5,8; 6,8\rangle$	10	0,106	13,9
12	$(6,8; 8,7\rangle$	12	0,145	19,2
13	$(8,7; \infty)$	11	0,107	14,1
Součet	x	132	1,000	x

Protože normální náhodná veličina může nabývat libovolné hodnoty z množiny reálných čísel, volíme jsou dva krajní intervaly pro potřeby testu rozšířeny na: $(-\infty; 1,5\rangle$, $(8,7; \infty)$.

- Platí-li H_0: $X \to N(4,6; (3,3)^2)$

$$\pi_{0,1} = \mathrm{P}(\mathrm{X} \in (-\infty; 1,5)) = \mathrm{P}(\mathrm{X} < 1,5) = \mathrm{F}(1,5) = \Phi\left(\frac{1,5-4,6}{3,3}\right) = \Phi(-0,94) = 1 - \Phi(0,94) =$$

$$= 1 - 0,826 = 0,174$$

$$\vdots$$

$$\pi_{0,13} = \mathrm{P}(\mathrm{X} \in (8,7; \infty)) = \mathrm{P}(\mathrm{X} > 8,7) = 1 - \mathrm{F}(8,7) = 1 - \Phi\left(\frac{8,7-4,6}{3,3}\right) = 1 - \Phi(1,24) =$$

$$= 1 - 0,893 = 0,107$$

- Pohledem na očekávané četnosti zjistíme, že jsme intervaly zvolili poměrně dobře – pouze 2. a 3. intervalu přísluší očekávané četnosti nižší než 5 (to odporuje použitelnosti χ^2 testu dobré shody). Tento nedostatek snadno napravíme tím, že tyto intervaly sloučíme.

i	Časový interval [s]	Počet pozorování v časovém intervalu	Očekávané relativní četnosti $\pi_{0,i}$	Očekávané četnosti $n.\pi_{0,i}$
1	$(-\infty; 1{,}5\rangle$	11	0,174	22,9
2	$(1{,}5; 2{,}0\rangle$	20	0,041	5,4
3	$(2{,}0; 2{,}5\rangle$	10	0,047	6,2
4	$(2{,}5; 2{,}9\rangle$	8	0,041	5,4
5	$(2{,}9; 3{,}6\rangle$	8	0,078	10,3
6	$(3{,}6; 4{,}0\rangle$	10	0,047	6,2
7	$(4{,}0; 4{,}4\rangle$	10	0,048	6,3
8	$(4{,}4; 4{,}9\rangle$	10	0,060	8,0
9	$(4{,}9; 5{,}8\rangle$	12	0,106	14,0
10	$(5{,}8; 6{,}8\rangle$	10	0,106	13,9
11	$(6{,}8; 8{,}7\rangle$	12	0,145	19,2
12	$(8{,}7; \infty)$	11	0,107	14,1
Součet	X	132	1,000	x

- Nyní jsou splněny předpoklady pro použití χ^2 testu dobré shody. Jako testovou statistiku tedy volíme:

$$T(\underline{X}) = G = \sum_{i=1}^{k} \frac{(n_i - n \cdot \pi_{0,i})^2}{n \cdot \pi_{0,i}} \rightarrow \chi^2_{k-h-1}$$

Výpočet pozorované hodnoty x_{OBS}:

$$x_{OBS} = T(\underline{X})_{H_0} = G_{H_0} = \sum_{i=1}^{k} \frac{(n_i - n \cdot \pi_{0,i})^2}{n \cdot \pi_{0,i}} =$$

$$= \frac{(11-22{,}9)^2}{22{,}9} + \frac{(20-5{,}4)^2}{5{,}4} + \ldots + \frac{(11-14{,}1)^2}{14{,}1} = 59{,}7$$

Výpočet p-value:

Počet variant: k = 12
Počet odhadovaných parametrů: h = 2

$p - value = 1 - F_0(x_{OBS})$
$F_0(x_{OBS}) = F_0(59{,}7)$
$F_0(59{,}7) >>> 0{,}999$ (viz. Tabulka 3, počet stupňů volnosti = 12-2-1 = 9)
$1 - F_0(59{,}7) <<< 0{,}001$

$p\text{-}value <<< 0{,}001$

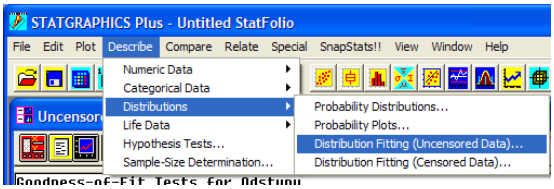
Rozhodnutí:

$p-value <<< 0{,}001 \Rightarrow$ Zamítáme nulovou hypotézu, tzn. že naměřené časové odstupy nelze považovat za výběr z normálního rozdělení.

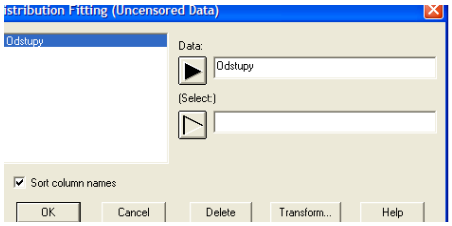
Řešení ve Statgraphicsu:

Nejdříve data zadáme do Statgraphicsu (pod názvem Odstupy), resp. použijeme již vytvořený soubor Dalnice.sf3.

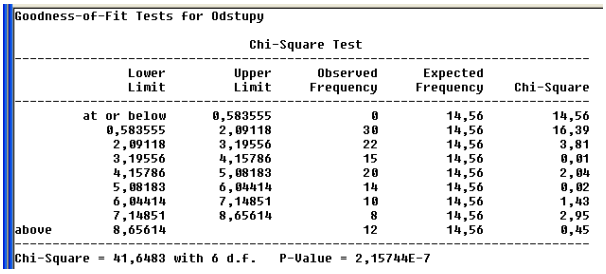
Chceme-li ověřit, zda data podléhají normálnímu rozdělení (nejčastěji se vyskytující požadavek na test dobré shody), zvolíme menu **Describe\Distributions\Distribution Fitting (Uncensored Data) …**



Jako Data zadáme testované hodnoty, tj. Odstupy.



V levém dolním textovém okně nalezneme výsledky testu χ^2 dobré shody (Pearsonova testu).



Goodness-of-Fit Tests for Odstupy

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	0,583555	0	14,56	14,56
0,583555	2,09118	30	14,56	16,39
2,09118	3,19556	22	14,56	3,81
3,19556	4,15786	15	14,56	0,01
4,15786	5,08183	20	14,56	2,04
5,08183	6,04414	14	14,56	0,02
6,04414	7,14851	10	14,56	1,43
7,14851	8,65614	8	14,56	2,95
above 8,65614		12	14,56	0,45

Chi-Square = 41,6483 with 6 d.f. P-Value = 2,15744E-7

Zjištěné výsledky se liší od výsledků, které jsme získali při „ručním“ výpočtu, neboť ve Statgraphicsu bylo zvoleno jiné rozčlenění do tříd. Konečný výsledek je však stejný.

Rozhodnutí:

$p-value <<< 0{,}001 \Rightarrow$ Zamítáme nulovou hypotézu, tzn. že naměřené časové odstupy nelze považovat za výběr z normálního rozdělení.

Kolmogorovův – Smirnovův test pro 1 výběr

Kolmogorovův – Smirnovův test se používá k ověření hypotézy, že pořízený **výběr pochází z rozdělení se spojitou distribuční funkcí F(x)**. F(x) musí být **úplně specifikovaná**.

V případě **výběru malého rozsahu**, dáváme tomuto testu přednost před χ^2 testem dobré shody.

Výhody Kolmogorovova - Smirnovova test oproti χ^2 testu dobré shody:

- větší síla testu $(1-\beta)$
- nemá omezující podmínky
- vychází z jednotlivých pozorování a nikoliv u údajů setříděných do skupin (nedochází ke ztrátě informace obsažené ve výběru)

Volba nulové a alternativní hypotézy

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

H_A: $\overline{H_0}$

kde F(x) je distribuční funkce rozdělení, z něhož náhodný výběr pochází (teoretická distribuční funkce)

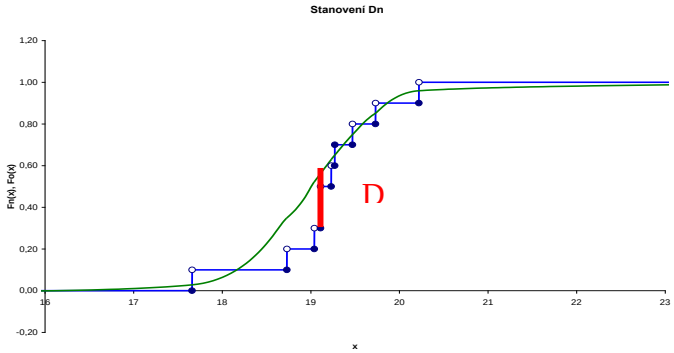
Volba testové statistiky $T(\underline{X})$ (včetně nulového rozdělení)

Uvažujme vzestupně uspořádaný náhodný výběr ze spojitého rozdělení: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

Jako testové kritérium použijeme statistiku D_n, jejíž význačné kvantily jsou tabelovány. Testová statistika D_n je definována jako maximální odchylka teoretické a empirické distribuční funkce.

$$T(\underline{X}) = D_n = \sup_x |F_n(x) - F_0(x)| = \max\left(D_1^*, D_2^*, \ldots, D_n^*\right),$$

kde $D_i^* = \max\left\{\left|F_0(x_i) - \frac{i-1}{n}\right|, \left|\frac{i}{n} - F_0(x_i)\right|\right\}$ *pro* $i = 1,2,\ldots,n$



Dále postupujeme standardně podle čistého testu významnosti.

Výpočet p-value

Při tomto testu určujeme p-value jako: $p - value = 1 - F_0(x_{OBS})$

12.4. V tabulce je 10 čísel generovaných jako hodnoty rozdělení N (19; $0{,}7^2$). Ověřte Kolmogorovovým – Smirnovovým testem, zda generované hodnoty pocházejí z předpokládaného rozdělení.

Generované hodnoty x_i	**19,732**	**19,108**	**19,234**	**19,038**	**19,270**	**19,105**	**19,473**	**17,660**	**20,219**	**18,727**

Řešení:

Volba nulové a alternativní hypotézy:

H_0: $F(x) = F_0(x)$, kde $F_0(x)$ je distribuční funkce normálního rozdělení o parametrech $\mu = 19$, $\sigma = 0{,}7$. (neboli: data pocházejí z N (19; $0{,}7^2$))

H_A: Data nepocházejí z N (19; $0{,}7^2$)

Volba testové statistiky:

$$T(\underline{X}) = D_n = \sup_x |F_n(x) - F_0(x)| = \max\left(D_1^*, D_2^*, \ldots, D_n^*\right)$$

$$\text{kde } D_i^* = \max\left\{\left|F_0(x_i) - \frac{i-1}{n}\right|, \left|\frac{i}{n} - F_0(x_i)\right|\right\} \qquad pro\ i = 1,2,\ldots,n$$

Výpočet pozorované hodnoty x_{OBS} (MS Excel):

Seřazené hodnoty $x_{(i)}$	**Pořadí (i)**	**(i-1)/n**	**i/n**	**$F_0(x_{(i)})$**	**D_i pro i/n**	**D_i pro (i-1)/n**	**D_i^***
17,660	1	0,00	0,10	0,03	0,07	0,03	0,07
18,727	2	0,10	0,20	0,35	0,15	0,25	0,25
19,038	3	0,20	0,30	0,52	0,22	0,32	0,32
19,105	4	0,30	0,40	0,56	0,16	0,26	0,26
19,108	5	0,40	0,50	0,56	0,06	0,16	0,16
19,234	6	0,50	0,60	0,63	0,03	0,13	0,13
19,270	7	0,60	0,70	0,65	0,05	0,15	0,15
19,473	8	0,70	0,80	0,75	0,05	0,05	0,05
19,732	9	0,80	0,90	0,85	0,05	0,05	0,05
20,219	10	0,90	1,00	0,96	0,04	0,06	0,06

x_{OBS} = 0,32

Výpočet p-value:

$p - value = 1 - F_0(x_{OBS})$

$F_0(x_{OBS}) = F_0(0{,}32)$

$F_0(0{,}32) < 0{,}9$ (viz. Tabulka 5, n = 10)

$1 - F_0(0{,}32) > 0{,}1$

$p - value > 0{,}1$

Rozhodnutí:

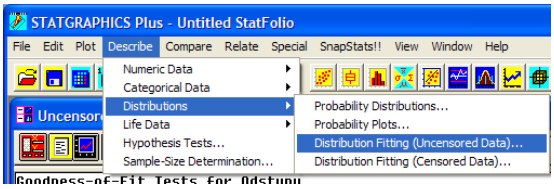
$p-value > 0{,}1 \Rightarrow$ Nezamítáme nulovou hypotézu, tzn. nelze tvrdit, že získaná data nepodléhají normálnímu rozdělení s parametry $\mu = 19$, $\sigma = 0{,}7$.

Řešení ve Statgraphicsu:

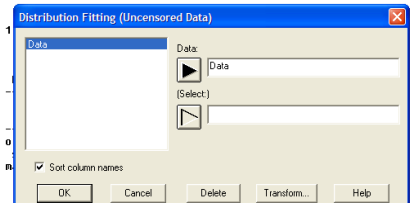
Statgraphics používá Kolmogorovův Smirnovův test automaticky pro neúplně specifikovaný výběr, tj. neumožňuje zadat požadované parametry teoretického rozdělení.

Opět zadáme data do Statgraphicsu, tentokrát pod obecným názvem Data, resp. použijeme již vytvořený soubor K_S_test.sf3.

Opět zvolíme menu **Describe\Distributions\Distribution Fitting (Uncensored Data) …**



Jako Data zadáme testované hodnoty, tj. Data.

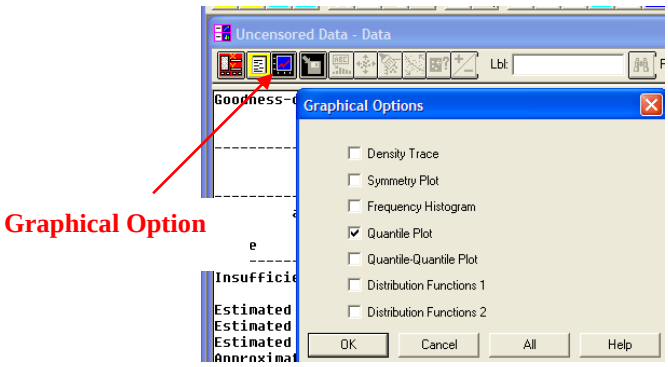


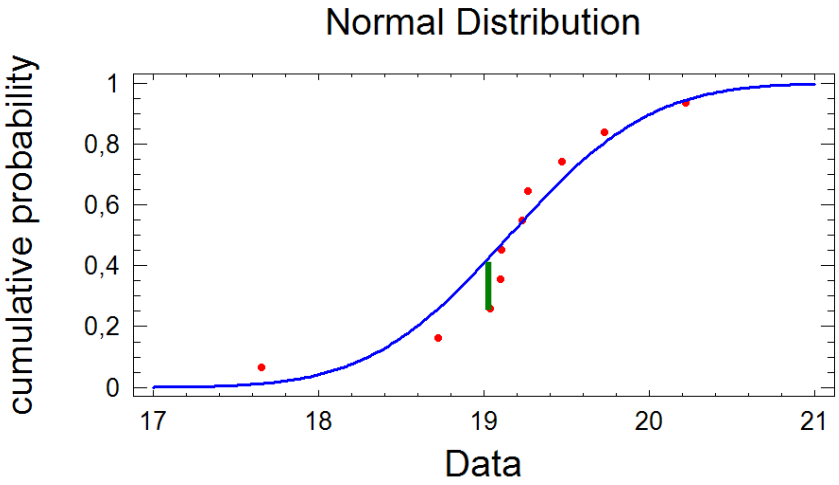
V levém dolním rohu najdeme v textovém výstupu výsledky Kolmogorovova-Smirnovova testu (všimněme si, že Statgraphics identifikoval nízký počet pozorování v souboru a tudíž nevygeneroval χ^2 test dobré shody).

```
Goodness-of-Fit Tests for Data

                          Chi-Square Test
-------------------------------------------------------------------------------
             Lower         Upper      Observed      Expected
             Limit         Limit     Frequency     Frequency     Chi-Square
-------------------------------------------------------------------------------
        at or below       18,9872             2          4,00          1,00
            18,9872        19,326             5          2,00          4,50
above        19,326                           3          4,00          0,25
-------------------------------------------------------------------------------
Insufficient data to conduct Chi-Square test.

Estimated Kolmogorov statistic DPLUS = 0,132648
Estimated Kolmogorov statistic DMINUS = 0,229591
Estimated overall statistic DN = 0,229591
Approximate P-Value = 0,667576

EDF Statistic          Value          Modified Form   P-Value
-------------------------------------------------------------------------
Kolmogorov-Smirnov D   0,229591        0,785449        >=0.10*
Anderson-Darling A^2   0,476895        0,523392        0,1827*
-------------------------------------------------------------------------
*Indicates that the P-Value has been compared to tables of critical values
specially constructed for fitting the currently selected distribution.
Other P-values are based on general tables and may be very conservative.
```

Kolmogorovovu-Smirnovovu testovou statistiku lze vidět na grafu, který srovnává skutečnou a teoretickou distribuční funkci.

Tento graf vygenerujeme klikneme-li na ikonu **Graphical Option** a zaškrtneme položku **Quantile Plot**.





Rozhodnutí:

$p-value > 0{,}1 \Rightarrow$ Nezamítáme nulovou hypotézu, tzn. nelze tvrdit, že získaná data nepodléhají normálnímu rozdělení.

Test nezávislosti v kontingenční tabulce

Testy nezávislosti v kontingenční tabulce řadíme mezi tzv. analýzu kategoriálních dat. Kontingenční tabulka vzniká setříděním prvků populace podle variant dvou kategoriálních znaků. Grafickou obdobou kontingenční tabulky je mozaikový graf. Tento graf se skládá z obdélníků, jejichž strany jsou úměrné příslušným marginálním relativním četnostem.

Pro ověření nezávislosti náhodných veličin X a Y (nezávislosti v kombinační tabulce) používáme test, který je založen na porovnávání empirických (pozorovaných) četností s četnostmi teoretickými, tj. takovými, které bychom očekávali v případě nezávislosti.

Test	Testová statistika	Nulové rozdělení
X^2 test nezávislosti v kontingenční tabulce	$G = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(n_{ij} - n_{ij}^{*}\right)^2}{n_{ij}^{*}}$	$\chi^2_{(m-1)(n-1)}$

Yatesova korekce (pro nízké očekávané četnosti)	$G = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(n_{ij} - n_{ij}^{*} - 0{,}5\right)^2}{n_{ij}^{*}}$	$\chi^2_{(m-1)(n-1)}$
McNemarův test (test shody rozdělení v čtyřpolní tab.)	$G = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}$	$\chi^2_{(1)}$

12.5. Pro diferencovaný přístup v personální politice potřebuje vedení podniku vědět, zda spokojenost v práci závisí na tom, jedná-li se o pražský závod či závody mimopražské. Výsledky šetření jsou v následující tabulce. Zobrazte data pomocí mozaikového grafu a na základě testu nezávislosti v kombinační tabulce rozhodněte o závislosti spokojenosti v zaměstnání na umístnění podniku.

Stupeň spokojenosti	Místo	
	Praha	Venkov
Velmi spokojen	15	40
Spíše spokojen	50	130
Spíše nespokojen	25	10
Velmi nespokojen	10	20

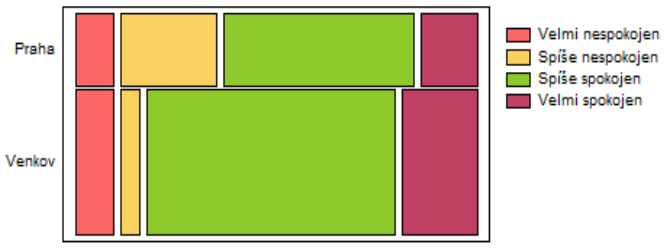
Řešení:

Nejdříve si data znázorníme pomocí mozaikového grafu, k čemuž potřebujeme znát marginální relativní četnosti:

	Velmi nespok	Spíše nespok	Spíše spokoj	Velmi spokoj	Row Total
Praha	10	25	50	15	100
	3,33%	8,33%	16,67%	5,00%	33,33%
Venkov	20	10	130	40	200
	6,67%	3,33%	43,33%	13,33%	66,67%
Column Total	30 10,00%	35 11,67%	180 60,00%	55 18,33%	300 100,00%

Cell contents:
Observed frequency
Percentage of table

Nyní můžeme sestrojit mozaikový graf. Na svislou osu budeme vynášet nezávisle proměnnou – tj. umístění podniku. Mozaikový graf proto bude tvořen dvěma řadami obdélníků (Praha, Mimo Prahu), přičemž řada odpovídající hodnotě „Praha“ bude mít šířku odpovídající 33,33% a řada odpovídající hodnotě „Mimo Prahu“ bude mít šířku odpovídající 66,67%. (Tzn., z celkové výšky mozaikového grafu bude řada odpovídající hodnotě „Praha“ zabírat 33,33%, …). Závisle proměnná (Stupeň spokojenosti) nabývá 4 hodnot, proto bude každý řádek mozaikového grafu tvořen čtyřmi obdélníky příslušných délek (např. obdélník odpovídající řádku „Praha“ a stupni spokojenosti – velmi spokojen bude mít délku odpovídající 15% celkové délky mozaikového grafu).



Všimněte si, že členitost grafu je způsobena zejména odlišný procentem „spíše nespokojených“ zaměstnanců.

Rozhodnutí o závislosti provedeme na základě testu nezávislosti v kombinační tabulce.

Volba nulové a alternativní hypotézy:

H_0: Spokojenost v práci **nezávisí** na umístění závodu.
H_A: Spokojenost v práci závisí na umístění závodu.

Volba testové statistiky:

$$T(\underline{X}) = G = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(n_{ij} - n_{ij}^{*}\right)^2}{n_{ij}^{*}} \rightarrow \chi^2_{(m-1)(n-1)}$$

Předpoklady testu:

Nutno ověřit, zda očekávané četnosti neklesly pod 2 a zda alespoň 80% z nich je větších než 5. Nejdříve si tedy z pozorovaných četností určíme četnosti marginální a pomocí nich pak četnosti očekávané.

Výpočet marginálních a očekávaných četností:

Stupeň spokojenosti	**Místo**		**Σ**
	Praha	**Venkov**	
Velmi spokojen	15	40	55
Spíše spokojen	50	130	180
Spíše nespokojen	25	10	35
Velmi nespokojen	10	20	30
Σ	100	200	300

n_i

$n_.$

n

Očekávané četnosti n_{ij}^{*}:

Stupeň spokojenosti	**Místo**	
	Praha	**Venkov**
Velmi spokojen	$\frac{55\cdot 100}{300} = 18{,}3$	$\frac{55\cdot 200}{300} = 36{,}6$
Spíše spokojen	$\frac{180\cdot 100}{300} = 60{,}0$	$\frac{180\cdot 200}{300} = 120{,}0$
Spíše nespokojen	$\frac{35\cdot 100}{300} = 11{,}7$	$\frac{35\cdot 200}{300} = 23{,}4$
Velmi nespokojen	$\frac{30\cdot 100}{300} = 10{,}0$	$\frac{30\cdot 200}{300} = 20{,}0$

Všechny očekávané četnosti jsou větší než 5.

Výpočet pozorované hodnoty:

$$x_{OBS} = T(\underline{X})_{H_0} = G = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(n_{ij} - n_{ij}^{*}\right)^2}{n_{ij}^{*}} = \frac{(15-18{,}3)^2}{18{,}3} + \frac{(50-60{,}0)^2}{60{,}0} + \ldots + \frac{(20-20{,}0)^2}{20{,}0} = 27{,}0$$

Výpočet p-value:

$m = 4, \quad n = 2 \quad \Rightarrow \quad$ počet stupňů volnosti $= (4-1)\cdot(2-1) = 3$

$p - value = 1 - F_0(x_{OBS})$

$F(27{,}0) >>> 0{,}999$ (viz. Tabulka 3, počet stupňů volnosti = 3)

$1 - F(27{,}0) <<< 0{,}001$

$p - value <<< 0{,}001$

Rozhodnutí:

P- value $< 0{,}01$, proto zamítáme nulovou hypotézu ve prospěch alternativy, tj. spokojenost v práci závisí na umístění závodu.

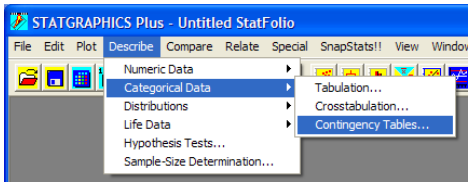
Řešení ve Statgraphicsu:

Nejdříve data zadáme do Statgraphicsu, resp. použijeme vytvořený datový soubor Spokojenost.sf3. **Pozor**, nezávisle proměnnou zadáváme jako kategoriální, závisle proměnnou zadáváme jako hlavičky sloupců.

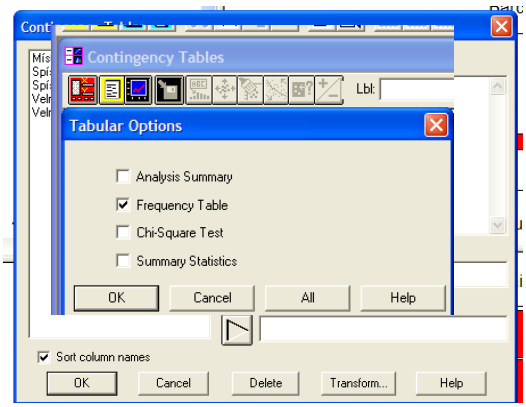
Spokojenost.sf3

	Místo	Velmi spokojen	Spíše spokojen	Spíše nespokojen	Velmi nespokojen
1	Praha	15	50	25	10
2	Venkov	40	130	10	20
3					
4					

Pro testování závislosti v kontingenční tabulce použijeme proceduru **Describe\Categorical Data\Contingency Tables …**



Jako **Columns** zadáme závisle proměnnou, tj. hodnoty zadané jako hlavičky sloupců. Nezávisle proměnnou zadáme jako **Labels.**

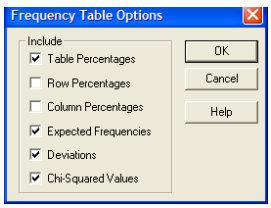


Grafický výstup této procedury, **mozaikový graf**, najdeme v pravém dolním rohu. Klikneme-li na ikonu **Tabular Options**, můžeme zaškrtnutím pole **Frequency Tables** získat příslušnou kontingenční tabulku.

	Spíše nespok	Spíše spokoj	Velmi nespok	Velmi spokoj	Row Total
Praha	25	50	10	15	100
	8,33%	16,67%	3,33%	5,00%	33,33%
Venkov	10	130	20	40	200
	3,33%	43,33%	6,67%	13,33%	66,67%
Column Total	35	180	30	55	300
	11,67%	60,00%	10,00%	18,33%	100,00%

Cell contents:
Observed frequency
Percentage of table

V kontingenční tabulce najdeme sdružené četnosti, sdružené relativní četnosti, marginální četnosti a marginální relativní četnosti. Provedeme-li RC na kontingenční tabulku, zvolíme menu **Pane Options** a zaškrtnutím příslušných polí můžeme tabulku doplnit o očekávané četnosti (Expected Frequencies), rozdíly mezi pozorovanými a očekávanými četnostmi (Deviations) a sčítance testové statistiky χ^2 (Chi-Squared Values).

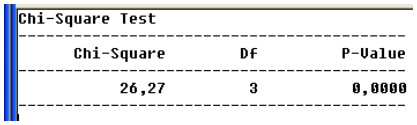


	Spíše nespok	Spíše spokoj	Velmi nespok	Velmi spokoj	Row Total
Praha	25	50	10	15	100
	8,33%	16,67%	3,33%	5,00%	33,33%
	11,67	60,00	10,00	18,33	
	13,33	-10,00	0,00	-3,33	
	15,24	1,67	0,00	0,61	
Venkov	10	130	20	40	200
	3,33%	43,33%	6,67%	13,33%	66,67%
	23,33	120,00	20,00	36,67	
	-13,33	10,00	0,00	3,33	
	7,62	0,83	0,00	0,30	
Column Total	35	180	30	55	300
	11,67%	60,00%	10,00%	18,33%	100,00%

Cell contents:
Observed frequency
Percentage of table
Expected frequency
Observed - expected frequency
Contribution to chi-squared

V rozšířené kontingenční tabulce **ověříme předpoklady testu**. V našem případě jsou všechny očekávané četnosti (expected frequency) větší než 5, tzn. že předpoklady testu jsou splněny.

Výsledky testu závislosti v kontingenční tabulce (hodnotu testové statistiky, p-value) najdeme v textových výstupech v části Chi-Square Test:



Chi-Square Test

Chi-Square	Df	P-Value
26,27	3	0,0000

Rozhodnutí:

P- value $< 0{,}01$, proto zamítáme nulovou hypotézu ve prospěch alternativy, tj. spokojenost v práci závisí na umístění závodu.

12.6. Byla vybrána skupina 100 řidičů, kteří měli za úkol projet se svými vozidly náročnou uzavřenou trať. Potom po požití alkoholu dostali stejný úkol. Má se zjistit, zda požití alkoholu ovlivňuje pravděpodobnost správného projetí trati. Je tedy třeba rozhodnout, zda se počet úspěšných řidičů před podáním alkoholu (jichž bylo 80) významně liší od počtu úspěšných řidičů po požití alkoholu (jichž pak bylo jen 60). Výsledky experimentu jsou shrnuty v následující tabulce:

Před požitím alkoholu	Po požití alkoholu		Celkem
	Bez chyby	Chybně	
Bez chyby	45	35	80
Chybně	15	5	20
Celkem	60	40	100

Řešení:

Jde o závislé proměnné (stejné osoby prováděly pokus „před“ a „po“), použijeme tedy MCNemarův test.

Nulová hypotéza: Procento „úspěšných“ řidičů nezávisí na podání alkoholu.
Alternativní hypotéza: Procento „úspěšných“ řidičů závisí na podání alkoholu.

Ověření předpokladu testu: $(n_{12} + n_{21})/2 = (45+5)/2 = 25 \geq 4$

Výpočet pozorované hodnoty: $x_{OBS} = \frac{(45-5)^2}{(45+5)} = 8$

Výpočet p-value:

$$0{,}995 < \chi^2_{(1)}(8) < 0{,}999$$
$$p - value = 1 - \chi^2_{(1)}(8)$$
$$0{,}001 < p - value < 0{,}005$$

Rozhodnutí: Zamítáme nulovou hypotézu, alkohol ovlivňuje „úspěšnost“ řidičů.
