JAK OPTIMÁLNĚ VYUŽÍT STATISTIKY PŘI ZPRACOVÁNÍ DAT

PREZENTACE PRO KURZ ZÁKLADŮ VĚDECKÉ PRÁCE V AKADEMII VĚD ČR

Doc. RNDr. Zdeněk Karpíšek, CSc.

Centrum pro jakost a spolehlivost výroby (CQR) MŠMT ČR (www.cqr.cz)
Odbor statistiky a optimalizace, Ústav matematiky
Fakulta strojního inženýrství, Vysoké učení technické v Brně
(www.mat.fme.vutbr.cz/home/karpisek)
Katedra aplikovaných disciplín, Akademie Sting v Brně

E-mail: karpisek@fme.vutbr.cz, karpisek@sting.cz

POPISNÁ STATISTIKA (DESKRIPTIVNÍ STATISTIKA):

Základní atribut: prvky pozorovaného statistického souboru nemají náhodný charakter

Popis souborů:

1. **Grafy**
2. **Číselné charakteristiky**

Nedostatek: **neúplné informace o pozorovaných statistických znacích**
→ vyvozené závěry mají subjektivní charakter

TEORIE PRAVDĚPODOBNOSTI = matematický model náhody

1. **Náhodné jevy**
2. **Pravděpodobnost náhodných jevů, podmíněná pst, nezávislé náhodné jevy**
3. **Náhodné veličiny, jejich funkční a číselné charakteristiky**
4. **Náhodné vektory, jejich funkční a číselné charakteristiky**
5. **Rozdělení psti pro aplikace**
6. **Náhodné procesy**
7. **Teorie spolehlivosti**
8. **Teorie hromadné obsluhy**

a další

MATEMATICKÁ STATISTIKA (INDUKČNÍ STATISTIKA, INFERENČNÍ METODY):

Základní atribut: prvky pozorovaného statistického souboru mají náhodný charakter

→ popis vychází ze spojení metod popisné statistiky a teorie pravděpodobnosti

→ model je založen na pojmu a vlastnostech tzv. náhodného výběru

Úlohy matematické statistiky:

1. **Odhady:**
   - **(a) parametrů rozdělení pravděpodobnosti – bodové a intervalové**
   - **(b) rozdělení pravděpodobnosti**

2. **Testování hypotéz:**
   - **(a) o parametrech a vlastnostech rozdělení pravděpodobnosti**
   - **(b) o rozdělení pravděpodobnosti**

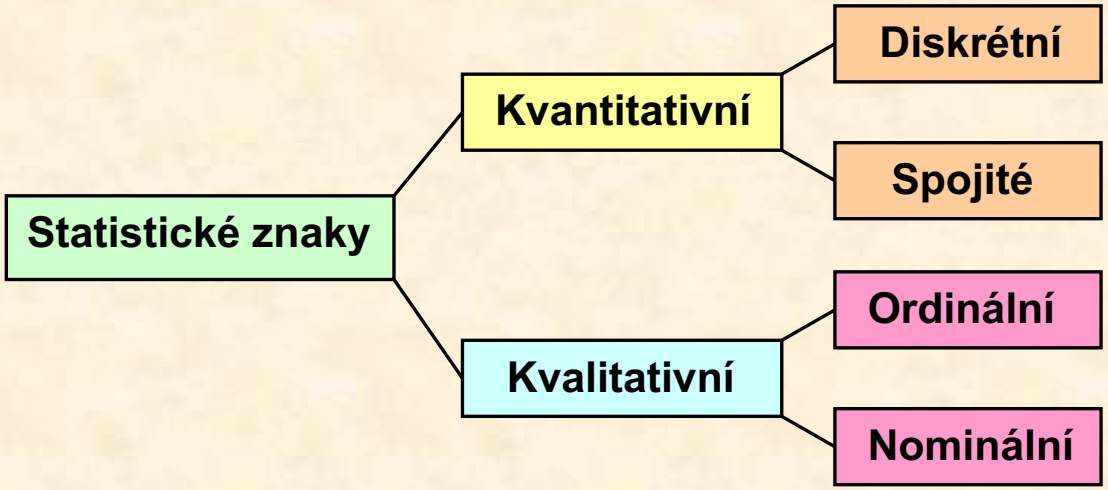
Odhady a testy se dle potřeby a požadavků provádějí současně: regresní analýza, ANOVA, kategoriální analýza aj.

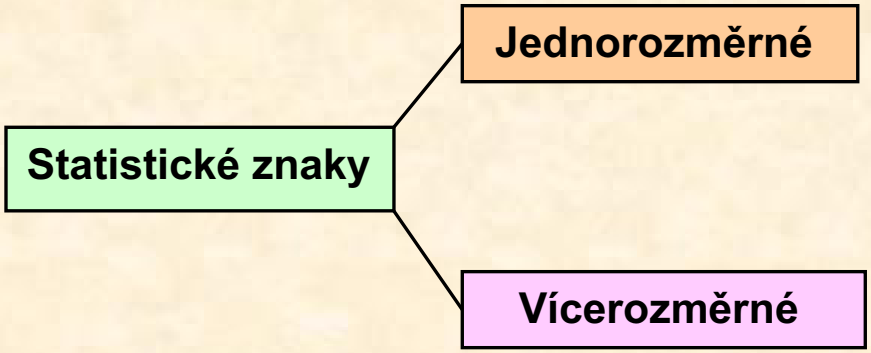
Průzkumová (exploratorní) analýza = spojení vybraných metod popisné a indukční statistiky

Data mining = hledání hodnotných informací ve velkých objemech dat

Obecný statistický model:

základní soubor (populace) = souhrn statistických jednotek
statistické jednotky → statistické znaky → hodnoty





Stochastický model:

- **diskrétní kvantitativní znak ~ diskrétní náhodná veličina a její rozdělení psti**
- **spojitý kvantitativní znak ~ spojitá náhodná veličina a její rozdělení psti**
- **ordinální kvalitativní (kategoriální) znak ~ multinomické rozdělení psti četností**
- **nominální kvalitativní (kategoriální) znak ~ multinomické rozdělení psti četností**
- **jednorozměrný statistický znak ~ náhodná veličina**
- **vícerozměrný statistický znak ~ náhodný vektor**

Základní soubor → výběrový soubor, rozsah

Výběry podle rozsahu:

- **velmi malé (do cca 20)**
- **malé (obvykle do cca 30 až 50)**
- **velké (řádově stovky)**
- **velmi velké (řádově tisíce a více)**

Požadavky na výběr:

- **reprezentativní (informace bez omezení)**
- **homogenní (bez vlivu dalších faktorů)**

- **náhodný**

Neurčitost výběru = zkreslení informací o základním souboru

Druhy výběrů: **bez opakování, s opakováním, záměrný, oblastní (stratifikovaný), mechanický a další**

Statistický soubor = soubor pozorovaných hodnot (x_1, x_2,…, x_n) znaku, resp. náhodné veličiny X na vybraných statistických jednotkách, resp. z jednotlivých pozorování (analogicky pro náhodný vektor)

POPISNÁ STATISTIKA

Zpracování statistického souboru = příprava + grafické znázornění + výpočet číselných charakteristik

Roztříděný soubor: $(x_1^*, f_1),\ldots,(x_m^*, f_m)$ … třídy, střed a četnost

Uspořádaný statistický soubor: $(x_{(1)},\ldots,x_{(n)})$, $x_{(i)} \leq x_{(i+1)}$

Grafy = vizuální informace o poloze, variabilitě, symetrii, modalitě, …: krabicový graf, histogram, sloupcový graf, výsečový graf, …

Číselné (empirické) charakteristiky = číselné informace o poloze, variabilitě, symetrii, modalitě, …:

1. **Průměr (aritmetický, geometrický, …), kvantily (medián, kvartily, …), modus, polosuma, uřezaný průměr, …**
2. **Rozptyl, směrodatná odchylka, rozpětí, mezikvartilová odchylka, mutabilita, entropie, …**
3. **Koeficient šikmosti (asymetrie), koeficient špičatosti (excesu), …**
4. **Kovariance, korelační koeficient, pořadové korelační koeficienty, koeficienty asociace, …**

a další

Některé vlastnosti aritmetického průměru:

- **poměrně citlivý na změnu hodnot souboru**
- **citlivý na extrémně odchýlené hodnoty**
- **u kladně (záporně) asymetrických souborů je průměr větší (menší) než medián**
- **konvergence s rostoucím rozsahem souboru k průměru celé populace**
- **obvykle rychlá konvergence rozdělení pravděpodobnosti průměru k normálnímu rozdělení**

Poznámky k číselným charakteristikám:

- **geometrický průměr nelze nahradit aritmetickým průměrem**
- **míry variability se v aplikacích bohužel často opomíjí**
- **nezjišťuje se asymetrie souboru**
- **netestují se extrémně odchýlené hodnoty**
- **koeficient korelace je pouze mírou linearity vztahu mezi X a Y**
- **r = 0 nemusí znamenat nezávislost X a Y**
- **r ≠ 0 neprokazuje kauzalitu**
- **regresní analýza = "jemnější" vyjádření závislosti mezi X a Y a umožňuje predikci**

TEORIE PRAVDĚPODOBNOSTI

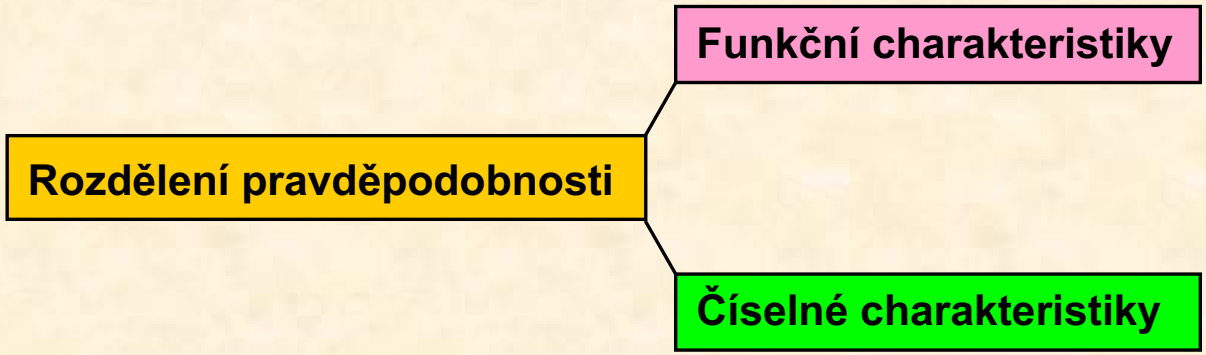
Pravděpodobnost P(A) je teoretická míra možnosti nastoupení náhodného jevu A.

Klasická definice: P(A) = m/n

- **m = počet příznivých případů jevu A**
- **n = počet všech možných případů**

Axiomatická definice - založená na teorii množin

Náhodná veličina (proměnná):



Funkční charakteristiky: distribuční funkce, hustota aj.

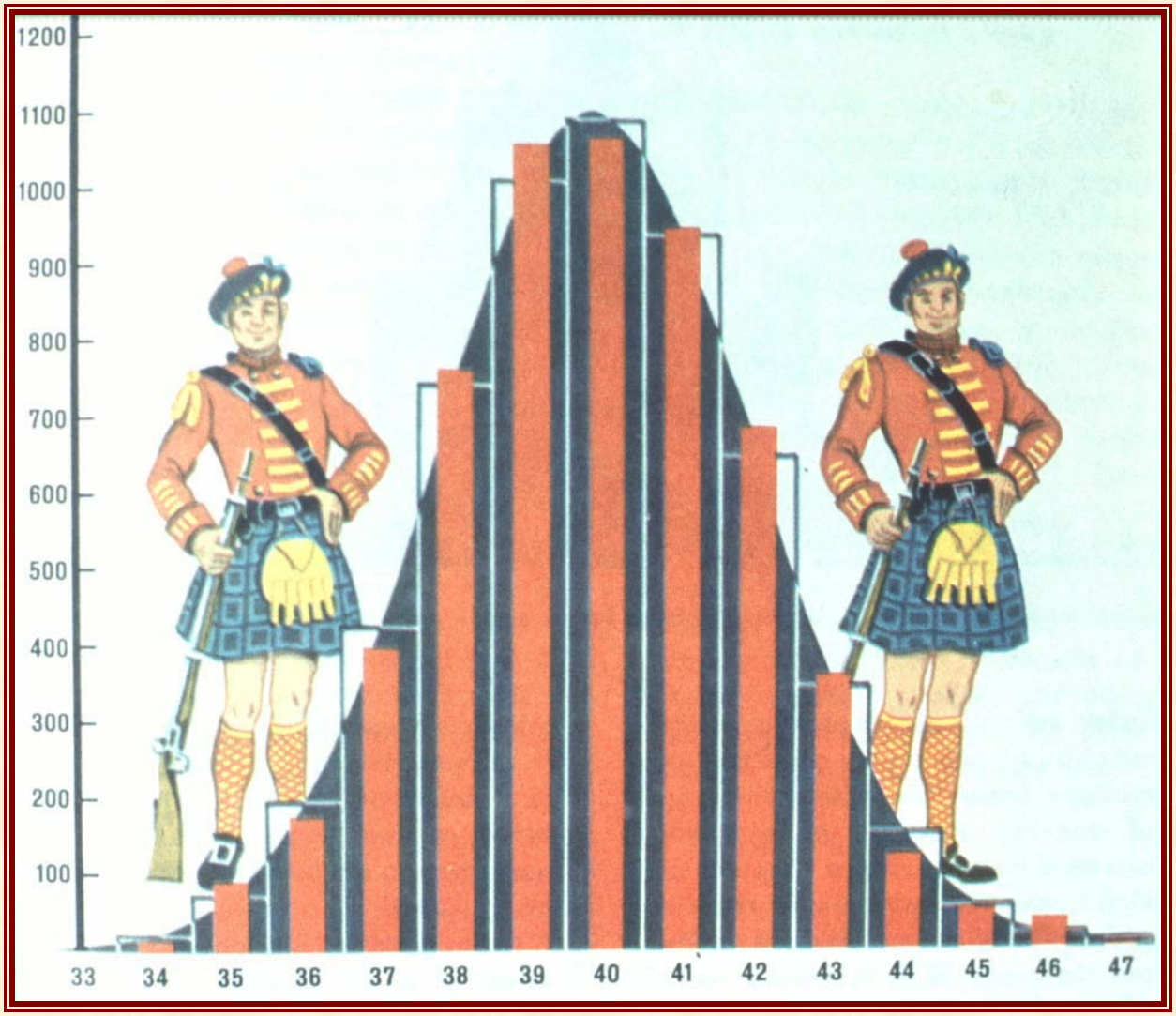
Číselné charakteristiky: střední hodnota, rozptyl aj.

Rozdělení pravděpodobnosti pro modelování reálných jevů:

binomické, hypergeometrické, Poissonovo, rovnoměrné, normální (Gaussovo), exponenciální, Weibullovo aj., aj.

Bernoulliův zákon velkých čísel - asymptotické chování relativní četnosti

Normální rozdělení - významné postavení při modelování reálného světa:

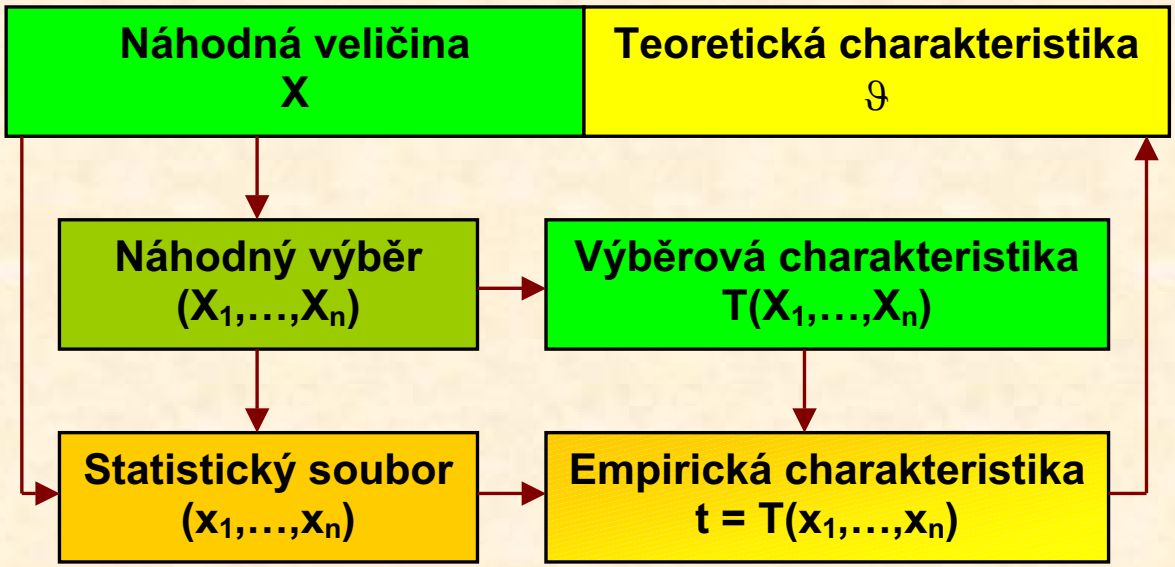


MATEMATICKÁ STATISTIKA

Principy matematické statistiky:

- **hodnoty získané výběrem ze základního souboru jsou náhodné**
- **získaný statistický soubor je hodnotou náhodného výběru**

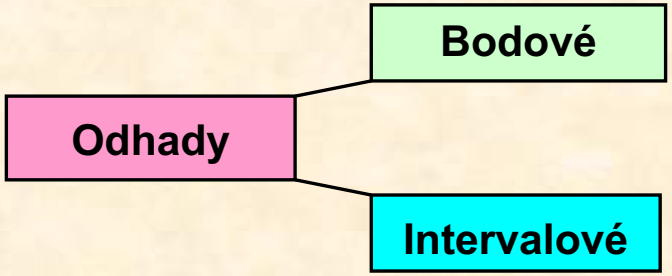
Statistická indukce:



Střední hodnota výběrového průměru = střední hodnota pozorované veličiny ("průměru" populace) a rozptyl výběrového průměru $\to 0$ pro $n \to \infty$, takže pro dostatečně velké n je takřka jistě průměr souboru blízký neznámé střední hodnotě; avšak tento rozptyl $\to 0$ s rychlostí $n^{1/2}$. Velmi často však rozdělení výběrového průměru konverguje k rozdělení normálnímu.

ODHADY PARAMETRŮ

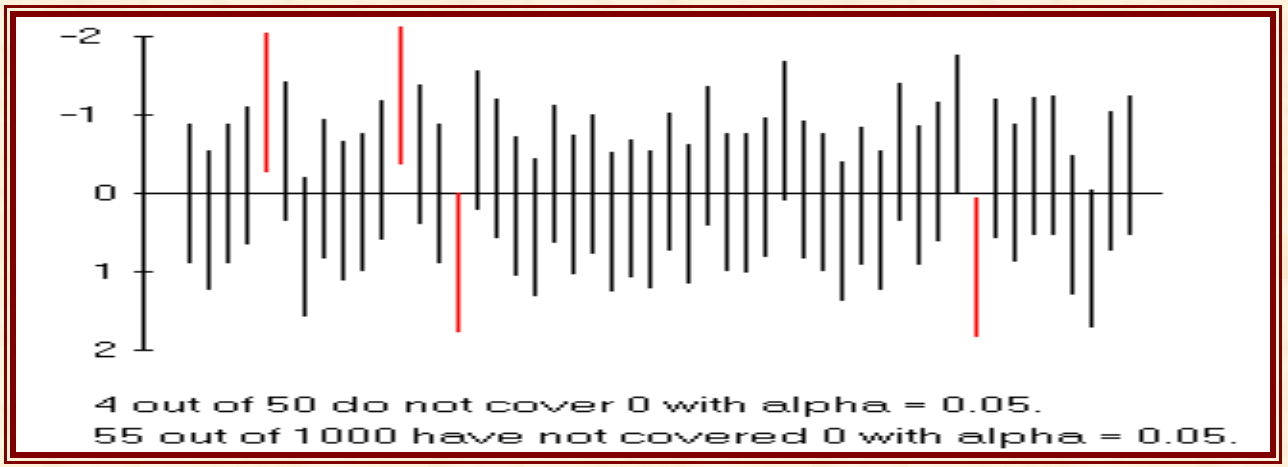
Odhad parametru ϑ = výběrová charakteristika $T(X_1,\ldots,X_n)$



Bodový odhad ϑ : $t = T(x_1,..,x_n)$

Intervalový odhad ϑ se spolehlivostí $1 - \alpha$: konfidenční interval $\langle T_1;T_2\rangle \Rightarrow \langle t_1;t_2\rangle$

Spolehlivost $1 - \alpha$ = pst úspěšnosti odhadu, konvence 0,95 a 0,99
Riziko chybného odhadu = α



Příklad:

Při průzkumu názoru z dotázaných n osob řeklo "ano" x osob. Pro spolehlivost 0,95:

n	x	Bodový odhad (%)	Intervalový odhad (%)	
			Od	Do
400	80	20	16,08	23,92
1600	320	20	18,04	21,96
6400	1280	20	19,02	20,98

TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

Statistické hypotéza = tvrzení o vlastnostech pozorované náhodné veličiny (vektoru)

Nulová hypotéza H_0 $\leftrightarrow$ Alternativní hypotéza H_A

Druhy hypotéz:

- **parametrické a neparametrické**
- **jednoduché a složené**
- **jednostranné a oboustranné**
- **sdružené**

Algoritmus testování hypotézy pomocí statistického souboru:

1. **Stanovení hypotéz H_0 a H_A.**
2. **Volba testového kritéria $T(X_1,\ldots,X_n)$.**
3. **Výpočet hodnoty testového kritéria $t = T(x_1,..,x_n)$.**
4. **Stanovení hladiny významnosti α a kritického oboru W_α.**
5. **Rozhodnutí o hypotézách H_0 a H_A.**

Hladina významnosti:
α = obvykle 5% anebo 1%

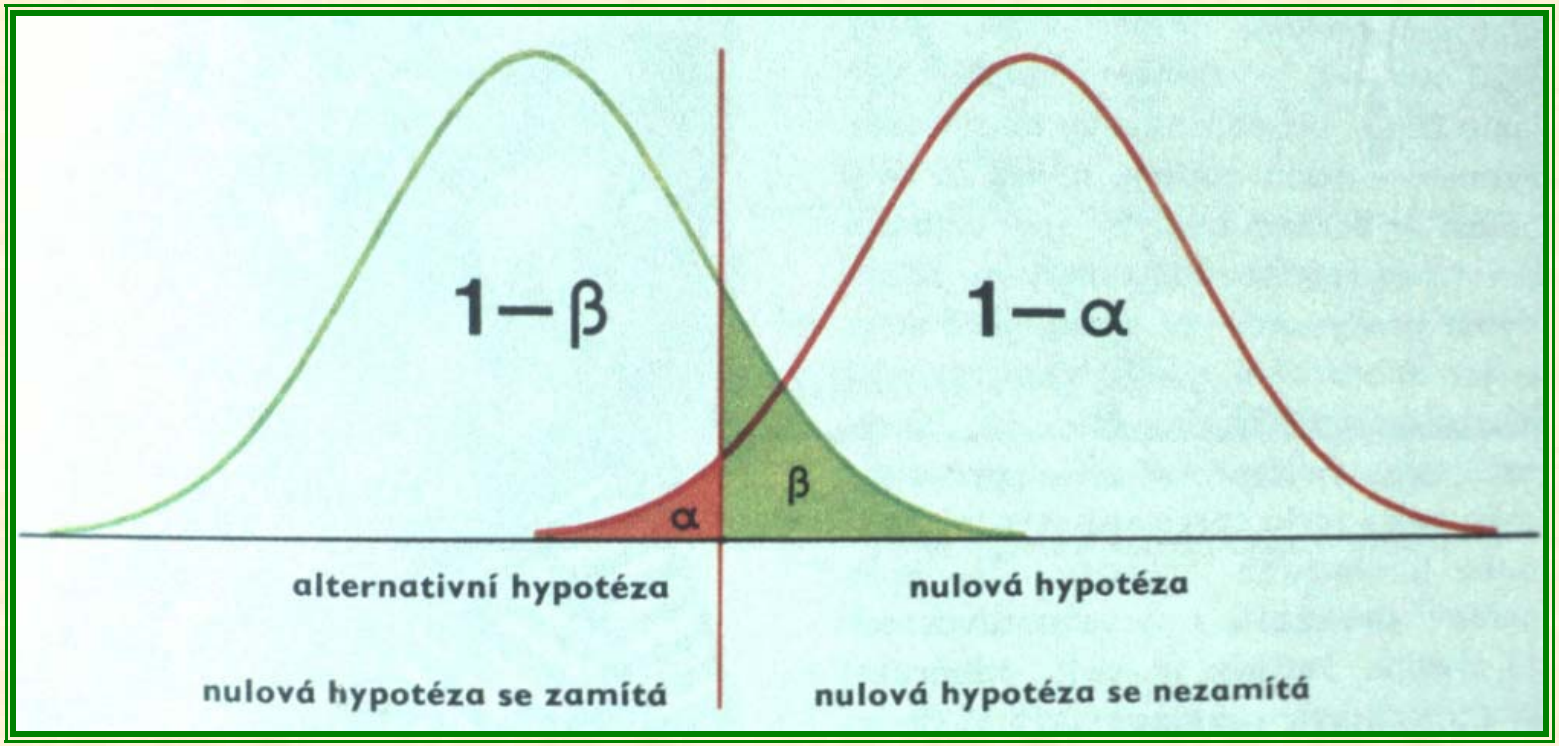
Rozhodnutí:

- $t \in W_\alpha \Rightarrow H_0$ zamítáme a H_A nezamítáme
- $t \notin W_\alpha \Rightarrow H_0$ nezamítáme a H_A zamítáme

H_0	PLATÍ	NEPLATÍ
ZAMÍTÁME	CHYBA 1. DRUHU	-------
NEZAMÍTÁME	-------	CHYBA 2. DRUHU

Rizika:

- pravděpodobnost chyby 1. druhu = hladina významnosti α
- pravděpodobnost chyby 2. druhu β snižujeme (stanovujeme) zvýšením rozsahu n



Aspekty:

- **nezamítnutí hypotézy neznamená vždy její přijetí ⇒ zvětšíme rozsah výběru a znovu testujeme**
- **nezamítnutí nebo přijetí hypotézy není potvrzení její platnosti**

Aplikace P-hodnoty a intervalových odhadů

DOPORUČENÝ POSTUP APLIKACE STATISTICKÝCH METOD VE VÝZKUMU:

1. Stanovení úkolu a pracovních hypotéz.
2. Vytvoření rigorózního a realizovatelného plánu experimentu, pozorování, průzkumu apod.
3. Realizace bodu 2, tj. získání statistických souborů.
4. Verifikace statistických souborů v rámci dané vědní disciplíny.
5. Výběr adekvátních statistických metod pro řešení.
6. Realizace statistických výpočtů pomocí modulů adekvátního profesionálního softwaru (Statistica, Minitab, Statgraphics, Systat, QCExpert, …, Excel aj.).
7. Analýza získaných výsledků a jejich aplikace pro řešení stanovených úkolů a ověření pracovních hypotéz.
8. Dle potřeb a nutností návrat k předcházejícím bodům uvedeného algoritmu.
9. Publikace nezbytných informací a výsledků získaných statistickou analýzou.
10. ???

UKÁZKA APLIKACE STATISTICKÝCH METOD č. 1

ROZDĚLENÍ PRAVDĚPODOBNOSTI KONCENTRACE LEGOVACÍHO PRVKU Ni

Rozdělení koncentrace X [% hmotnostního obsahu] legovacího (přísadového) prvku ve struktuře oceli má určující vliv na její materiálové vlastnosti: pevnost, tažnost, tvrdost aj. Hodnoty obsahu jednotlivých prvků v oceli byly získány energiově RTG mikroanalýzou na lineárním úseku v délce 1000 μm. Vzhledem k náhodnému charakteru obsahu a způsobu jeho měření je vhodné modelovat koncentraci X jako spojitou náhodnou veličinu. Pro statistické zpracování byl vybrán prvek Ni. Naměřené hodnoty obsahu tvoří statistický soubor a naším úkolem je:

- **zpracovat tento soubor metodami popisné statistiky,**
- **stanovit tvar pozorovaného rozdělení pravděpodobnosti,**
- **určit bodové a intervalové odhady jeho parametrů a charakteristik.**

V materiálovém inženýrství se nejčastěji používá normální (Gaussovo) rozdělení pravděpodobnosti $\mathrm{N}\left(\mu,\sigma^2\right)$ **s hustotou pravděpodobnosti**

$$f\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right],\ x\in(-\infty,+\infty),$$

a základními číselnými charakteristikami

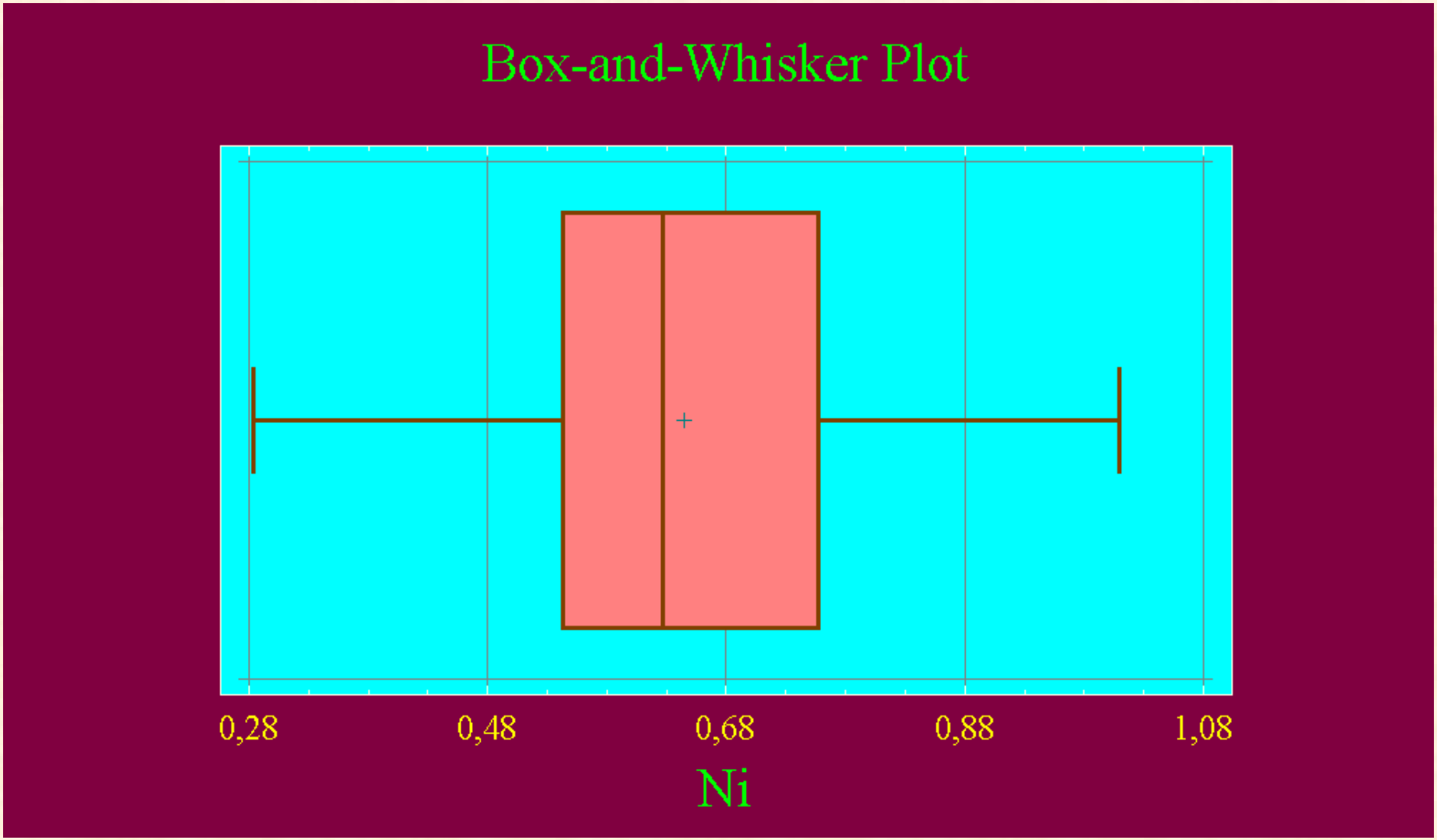
$$E(X)=x_{0,5}=\hat{x}=\mu,\ \ D(X)=\sigma^2,$$

kde μ je střední hodnota, $x_{0,5}$ je medián, $\hat{x}$ je modus, σ^2 je rozptyl a σ je směrodatná odchylka.
Pro statistické výpočty byl použit profesionální software Statgraphics Centurion XV.I. Zpracováním statistického souboru 100 naměřených hodnot koncentrace přísadového prvku Ni v nízkolegované oceli byly získány následující číselné a grafické výsledky.

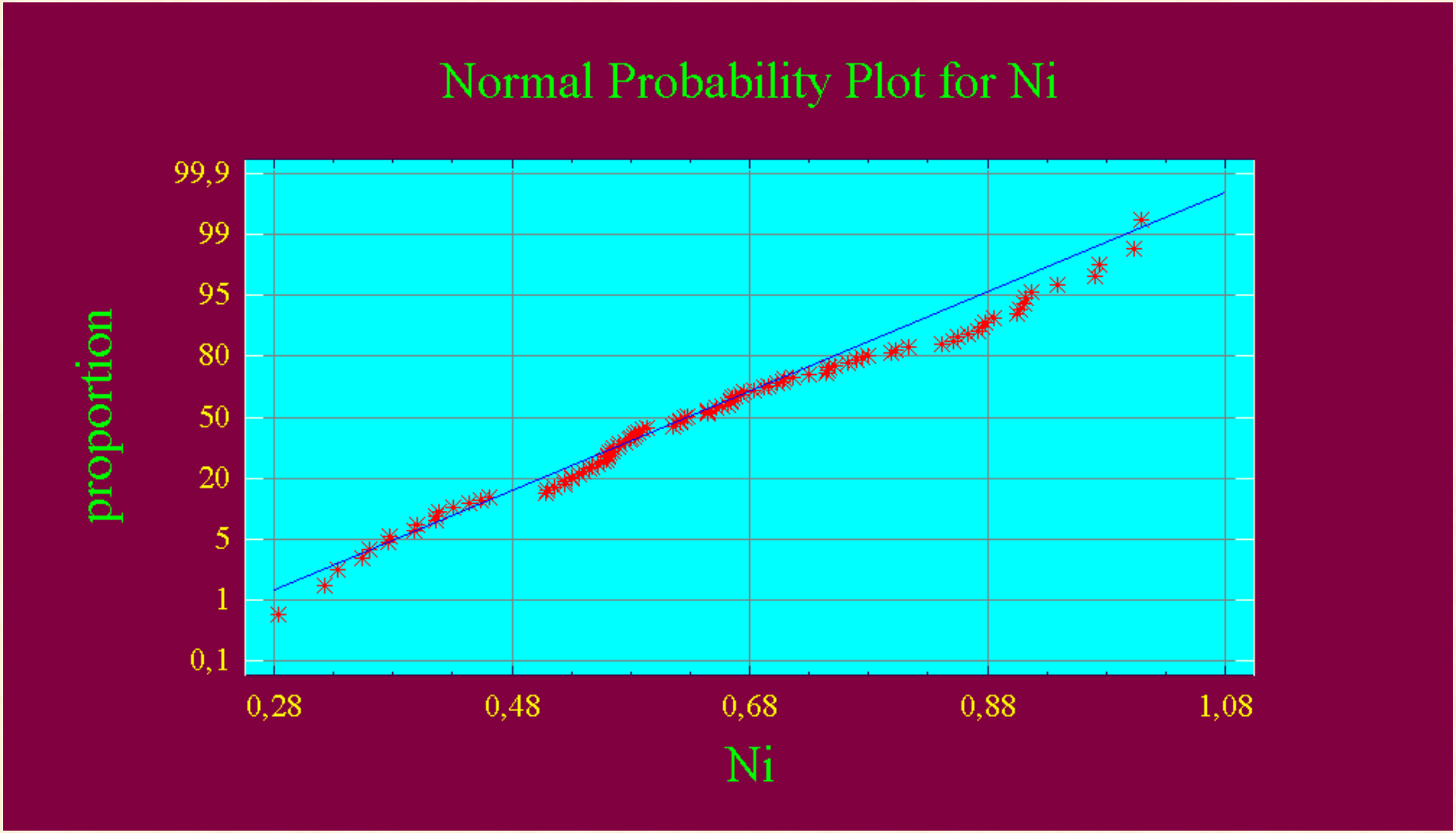
POPISNÉ CHARAKTERISTIKY

Summary Statistics for Ni	
Count = 100	Lower quartile = 0,542912
Average = 0,645077	Upper quartile = 0,756923
Median = 0,626583	Interquartile range = 0,214011
Variance = 0,0287817	Skewness = 0,165103
Standard deviation = 0,169652	Stnd. skewness = 0,674032
Minimum = 0,284121	Kurtosis = –0,52586
Maximum = 1,00947	Stnd. kurtosis = –1,07341
Range = 0,725349	Coeff. of variation = 26,2994%

KRABICOVÝ GRAF



EMPIRICKÝ ODHAD NORMÁLNÍHO ROZDĚLENÍ PRAVDĚPODOBNOSTI



ZÁVĚR: Z grafu odhadujeme, že jde o normální rozdělení.

TEST NORMÁLNÍHO ROZDĚLENÍ PRAVDĚPODOBNOSTI

Goodness-of-Fit Tests for Ni

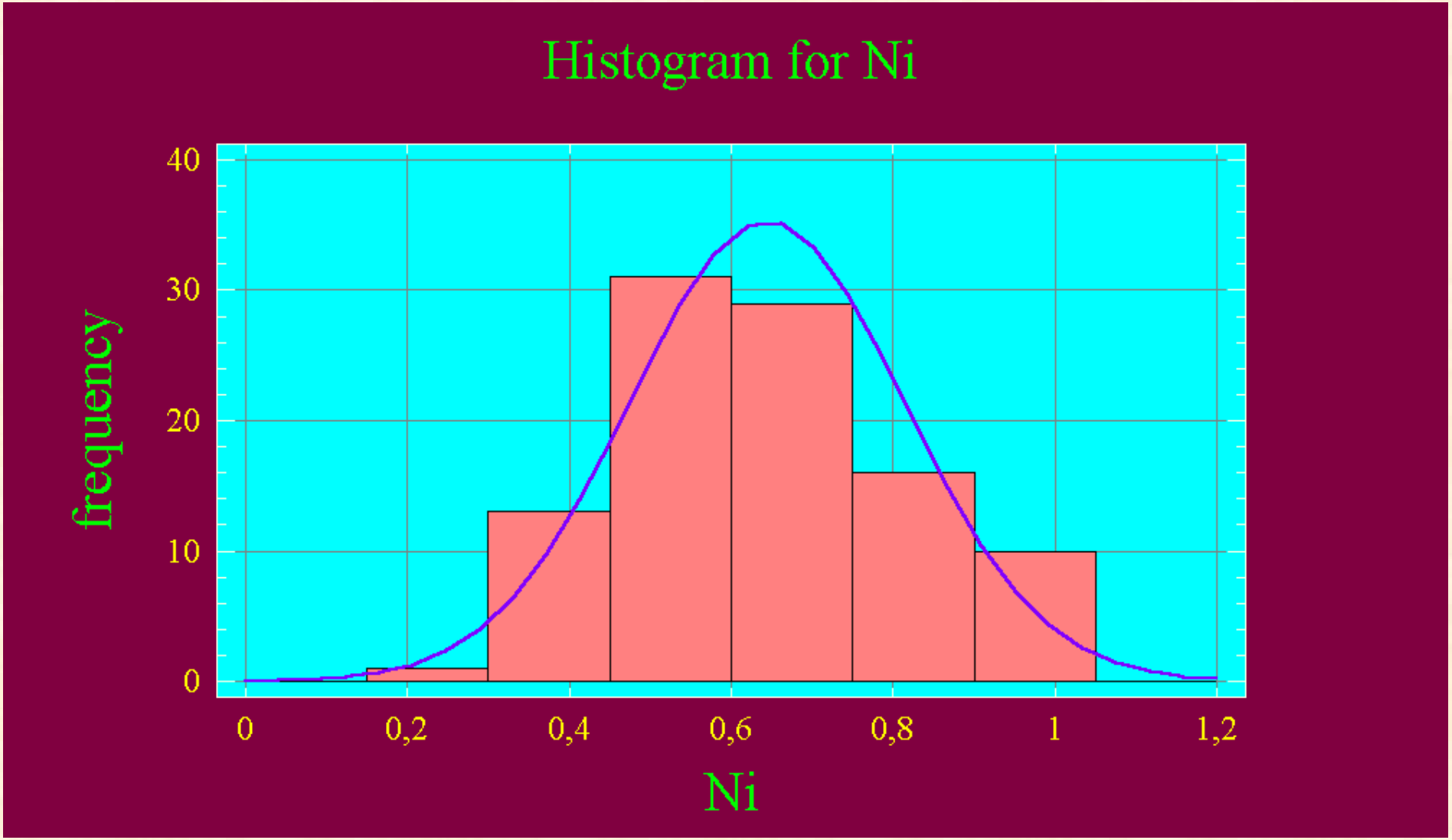
Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chisquare
at or below	0,45	14	12,51	0,18
0,45	0,6	31	27,01	0,59
0,6	0,75	29	33,66	0,65
0,75	0,9	16	20,17	0,86
above	0,9	10	6,65	1,69

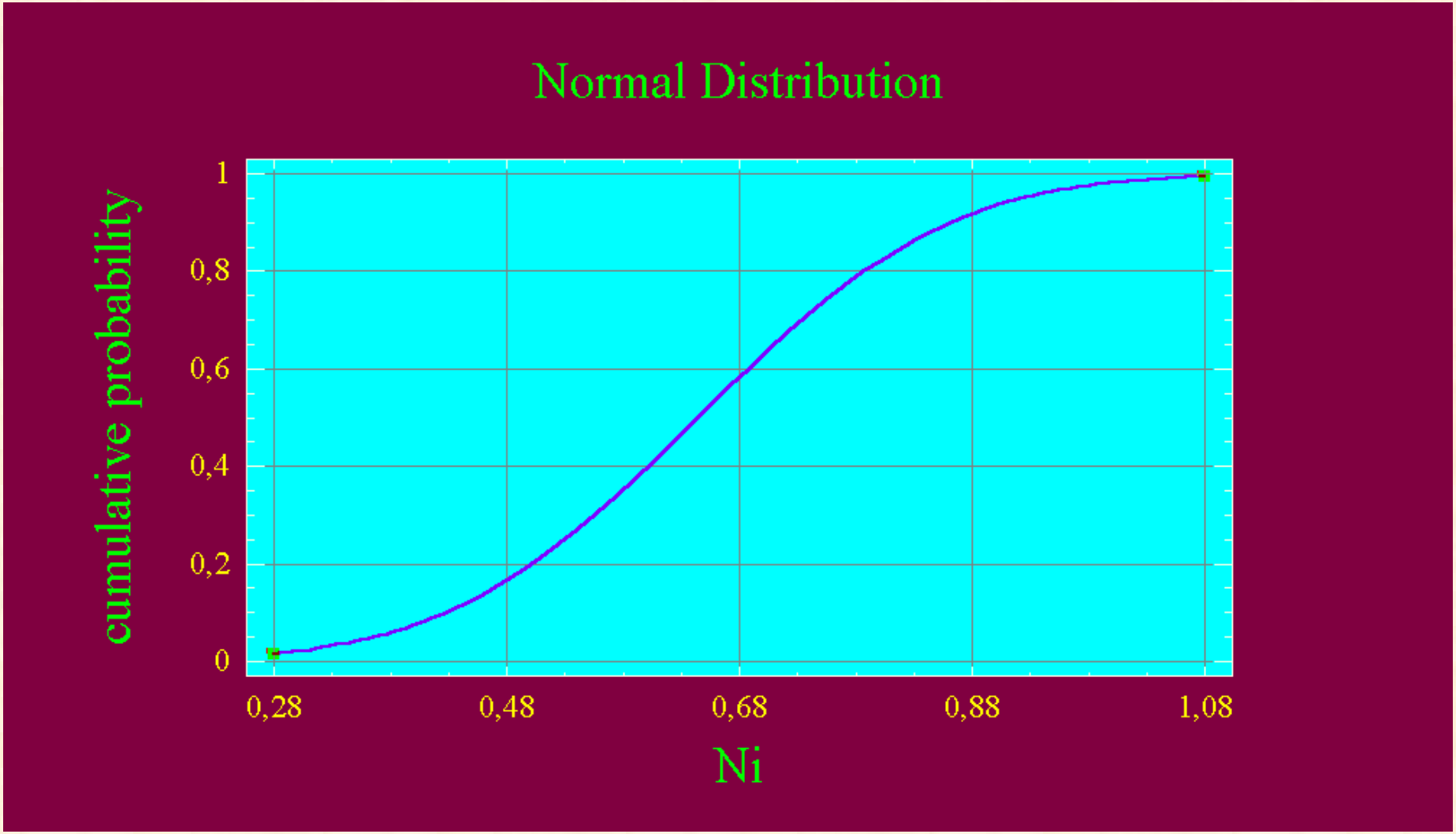
Chi-Square = 3,96445 with 2 d.f. P-Value = 0,137762
Estimated Kolmogorov statistic DPLUS = 0,0698738
Estimated Kolmogorov statistic DMINUS = 0,0579959
Estimated overall statistic DN = 0,0698738
Approximate P-Value = 0,713335

ZÁVĚR: Na základě obou testů nezamítáme hypotézu o normálním rozdělení na hladině významnosti 0,05.

HISTOGRAM A HUSTOTA PRAVDĚPODOBNOSTI



DISTRIBUČNÍ FUNKCE



BODOVÉ A INTERVALOVÉ ODHADY

Estimate of mean: 0,645077
Estimate of standard deviation: 0,169652
95,0 % confidence interval for mean: 0,645077 +/– 0,0336626
[0,611415;0,67874]
95,0 % confidence interval for standard deviation: [0,148955;0,19708]

****

CELKOVÉ ZÍSKANÉ VÝSLEDKY

- **obsah Ni v dané oceli má normální rozdělení pravděpodobnosti**
- **bodový odhad středního obsahu Ni je 0,645 % a bodový odhad směrodatné odchylky obsahu Ni je 0,1687 %**
- **se spolehlivostí 95 % je střední obsah Ni od 0,611 % do 0,679 % a směrodatná odchylka obsahu Ni od 0,1490 % do 0,1971 %**

UKÁZKA APLIKACE STATISTICKÝCH METOD č. 2

Šokující zjištění: Ženy jsou opravdu chytřejší než muži! (Super.cz --- 24. 2. 2010)

A jakže se na tuto převratnou pravdu přišlo? Jednoduše - z vědomostního internetového souboje milionů mužů a žen z národů devíti různých jazyků. Výsledky hovořily jasně - ženy si prostě vedly lépe než „pánové tvorstva“!

Průzkum probíhal na internetu a v jeho rámci padlo celkem patnáct milionů otázek! Testování probíhalo od října minulého roku a bylo od počátku velmi vyrovnané. V závěru ale přece jen se slabou převahou zvítězily ženy. Ty přitom odpověděly správně na 4 088 139 otázek a muži na 4 077 596 otázek. Dotazy byly kladeny z několika oblastí, přičemž nejoblíbenějším byl obor showbyznysu a zábavy, následovaný vědou, sportem, historií a uměním. Poslední byla kategorie lidé a místa. Ženy nejlépe odpovídaly právě v kategorii showbyznys + zábava a muži zase v kategorii věda + příroda. „Internetová bitva pohlaví nalákala obrovské množství lidi z celého světa“, přiznala jedna z autorek výzkumu Katreena Linesová. Jak vidno, boj mezi pohlavími je opravdu věčným tématem...

Pracovní hypotéza: Ženy jsou chytřejší než muži.

Statistická nulová hypotéza $H_0 : p_1 = p_2$ … alternativní hypotéza $H_A : p_1 > p_2$
Test statistické hypotézy:

Test hypotézy $H : p_1 = p_2$. Pozorovaná hodnota testového kritéria za předpokladu $n_1 > 50$ a $n_2 > 50$ je

$$t = \frac{\frac{x}{n_1} - \frac{y}{n_2}}{\sqrt{\bar{f}(1-\bar{f})}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}}$$

pro $\bar{f} = \frac{x+y}{n_1+n_2}$ a $\overline{W}_\alpha = \langle -u_{1-\alpha/2}; u_{1-\alpha/2} \rangle$, kde $u_{1-\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$-kvantil normálního rozdělení N(0; 1),

Počet otázek: $n_1 = n_2 = 15\,000\,000$
Počet správných odpovědí: ženy ... x = 4 088 139, muži ... y = 4 077 596
f_bar = 0,272191167 n_bar = 7500000
t = 4,324719102 $u_{0,95}$ = 1,644853 (P jednostr. = 7,64197E-06)

Závěr: Hypotézu H_0 zamítáme a hypotézu H_A nezamítáme, resp. přijímáme.

Přijímáme pracovní hypotézu, že ženy jsou chytřejší než muži! Opravdu?

DĚKUJI ZA POZORNOST!