

OBSAH

Charakteristika a význam distribučnej analýzy

Vzorové údaje 1.3

Praktické použitie programu SAS (SAS Enterprise Guide)

Príklad 1.9 (SAS)

Praktické použitie programu EXCEL (Microsoft 365)

Príklad 1.9 (Excel)

Praktické použitie programu R (R Studio)

Príklad 1.9 (Program R)

Ďalšie príklady použitia distribučnej analýzy

Zdroje a zoznam použitej literatúry

Charakteristika a význam distribučnej analýzy

Distribúcia údajov (rozdelenie početnosti) zaznamenáva, s akou početnosťou nastávajú rôzne vlastnosti, javy a procesy. Každéj vlastnosti, javu, alebo procesu sa priradí početnosť (rozdelenie) jej a jeho výskytu. Rozdelenie početností (distribúcia) sa najčastejšie vizuálne znázorňuje pomocou histogramov alebo pomocou funkcie, ktorú nazývame empirická distribučná funkcia.

Distribučná analýza údajov vychádza z toho aké typy náhodných premenných analyzujeme. Všeobecne analyzovanými údajmi môžu byť údaje diskkrétne a údaje spojité. Diskkrétne náhodné premenné nadobúdajú iba hodnoty, ktorých počet je ohraničený, alebo majú spočítateľný počet možných hodnôt. Spojité náhodné premenné môže nadobudnúť ľubovoľnú hodnotu medzi dvoma rôznymi hodnotami.

Podľa typu údajov a distribučnej funkcie rozoznávajú rôzne (najznámejšie) pravdepodobnostné rozdelenia analyzovaných údajov (diskkrétne typy rozdelenia početnosti a spojité typy rozdelenia početnosti).

Diskkrétne typy rozdelenia početnosti (pravdepodobnosti):

- rovnomerné diskkrétne rozdelenie,
- alternatívne (Bernoulliho) rozdelenie,
- binomické rozdelenie,
- Poissonovo rozdelenie,
- hypergeometrické rozdelenie,
- geometrické rozdelenie.

Spojité typy rozdelenia početnosti (pravdepodobnosti):

- rovnomerné rozdelenie,
- exponenciálne rozdelenie,
- **normálne (Gaussovo) rozdelenie**,
- Studentovo t- rozdelenie,
- F-rozdelenie,
- Chí- kvadrát rozdelenie,
- beta rozdelenie,
- gama rozdelenie,
- lognormálne rozdelenie,
- Weibullovo rozdelenie,

Pri rozdeleniach pravdepodobnosti charakterizujeme najčastejšie uvedené vlastnosti a parametre: funkciu a parametre hustoty rozdelenia pravdepodobnosti, kumulatívnu distribučnú funkciu, strednú hodnotu, median, modus, rozptyl.

Hodnoty aritmetického priemeru, mediánu a modusu slúžia zo základných štatistických charakteristík na prvé číselné posúdenie pravdepodobnosti možného rozdelenia hodnotených údajov. Doplňujúcimi charakteristikami sú koeficient šikmosti a špicatosti.

Komplexná distribučná analýza obsahuje: meranie centrálnej tendencie (polohy), sumárnu štatistiku, testy polohy (t- Test, poradový test), grafické posúdenie rozdelenia početnosti (hlavne pri normálnom rozdelení početnosti), testy normality, výpočet percentíl resp. kvantíl, nájdenie a zobrazenie výskytu extrémnych hodnôt.

Kontrola normality

Väčšina štatistických analýz a metód predpokladá, že údaje, ktoré analyzujeme majú normálne rozdelenie. Pred, alebo v rámci skoro každej analýzy, overujeme často predpoklad normálneho rozdelenia početnosti jednotlivých vlastností a znakov. Normalita sa overuje na úrovni celého výberového súboru údajov, alebo v rámci jednotlivých triedení znakov (v rámci definovaných skupín hodnoteného údajového súboru). Normálne rozdelenie je rozdelenie pravdepodobnosti, ktoré je symetrické a vytvára zvonovú krivku. Pri normálnom rozdelení patrí 68% výsledkov do jednej štandardnej odchýlky a 95% do dvoch štandardných odchýlok.

Existuje niekoľko najčastejšie používaných spôsobov, ako skontrolovať normalitu rozdelenia početnosti:

- grafické zostavenie histogramov s jednoduchým vizuálnym posúdením normality (často využívané pri analýze biologických údajov),
- výpočet šikmosti (koeficient šikmosti je v skutočnosti mierou asymetrie rozloženia početnosti, ak je hodnota koeficienta šikmosti blízka číslu 0, znamená to, že údaje sú s najväčšou pravdepodobnosťou normálne),
- pozitívne skreslenie údajov znamená, že existuje niekoľko extrémne veľkých hodnôt, ktoré menia svoj priemer na kladné skreslenie (nazýva sa tiež vpravo skosené skreslenie, šikmost' > 0), ďalší spôsob je vzájomná poloha priemeru a mediánu. Priemer je väčší ako medián a medián je väčší ako modus,
- opakom je negatívne skreslenie (vľavo skosené skreslenie spôsobené výskytom extrémne malých hodnôt, šikmost' < 0 , priemer je menší ako medián a medián je menší ako modus),

- ak je šikmosť < -1 alebo $> +1$, distribúcia je veľmi skreslená, ak je šikmosť medzi -1 a $-0,5$ alebo medzi $0,5$ a $+1$, distribúcia je mierne skreslená, ak je šikmosť $> -0,5$ a $< 0,5$, rozdelenie je približne symetrické alebo normálne.

Väčšina štatistických hypotéz je založená na predpoklade, že je vopred známe pravdepodobnostné rozloženie analyzovaných údajov. Numerické testy normality sledujú, či údajový súbor pochádza z normálneho rozdelenia početnosti, alebo nepochádza. Testy normality je možné riešiť numericky ale aj graficky.

Najznámejšie numerické testy normality sú:

1. Shapiro Wilk test, ktorý sa využíva najčastejšie pri veľkosti vzorky menšej ako 2000 pozorovaní. Nastavenie nulovej hypotézy je, že rozdelenie početnosti je normálne. V prípade, že pravdepodobnosť (p-hodnota) je menšia ako 0,05 nulovú hypotézu zamietame. Znamená to, že distribúcia nie je normálna. Ak je p-hodnota väčšia ako 0,05, rozdelenie početnosti je normálne.
2. Kolmogorov-Smirnovov test ktorý sa využíva najčastejšie pri veľkosti vzorky väčšej ako 2000 pozorovaní. Rovnako aj pri tomto teste nulová hypotéza predpokladá, že údaje majú normálne rozdelenie.

Ďalšou numerickou metódou posudzovania normality je výpočet percentíl, ktoré sa využívajú spolu aj grafickými posudzovania normality údajov.

Dôležitým prvkom kontroly normality je nájdenie, zobrazenie a možné odstránenie odľahlých (extrémnych, vybočujúcich pozorovaní resp. meraní). Ak napríklad chceme odstrániť takéto merania (ako príklad 10 % údajov) musíme poznať 10. a 90. percentil a odstránime hodnoty nad týmito percentilovými hodnotami. Zvyčajne potom vypočítavame nový priemer hodnôt (tzv. orezaný, trimmed priemer).

K numerickým metódam patrí aj výpočet t- Testu, pri ktorom testuje nulovú hypotézu, že priemer hodnoteného ukazovateľa sa rovná hodnote 0. Alternatívnou hypotézou je, že priemer sa nerovná 0. Posudzovanie testu je podobné ako pri testoch normality.

V rámci grafických metód je najviac využívané zostavenie histogramov, krabicových grafov a zostavenie grafu normálnej pravdepodobnosti.

Vzorové údaje 1.3 (zdroj: Databáza údajov projektu KEGA)

Súbor údajov **DATA-SB.xlsx** obsahuje nasledovné premenné:

CISLO	číslo dojnice
PL	poradie laktácie
DATUMO	datum otelenia
PLEMENO	plemenná príslušnosť (číselný kód plemena)
FARMA	farma (číselný kód)
SKUPINA	skupina zvierat podľa veku, plemena a sezóny telenia
MLIEKO	produkcia mlieka (kontrolný-denný nádoj, kg)
TUK	produkcia tuku (kontrolný-denný nádoj, kg)
TUKP	obsah tuku (kontrolný-denný nádoj, %)
BIELK	produkcia bielkovín (kontrolný-denný nádoj, kg)
BIELKP	obsah bielkovín (kontrolný-denný nádoj, %)
SB	počet somatických buniek v tisícoch (kontrolný-denný nádoj)
LDNI	laktačný deň kontrolného-denného nádoja

Vzor súboru údajov (10 záznamov):

Cislo	PI	Datumo	Plemeno	Farma	Skupina	Mlieko	Tuk	Tukp	Bielk	Bielkp	SB	LDNI
CZ000004310953	01	03.11.06	06	406512011	06	24.50	0.97	3.95	0.79	3.23	52	22
CZ000004310953	01	05.12.06	06	406512011	06	29.00	1.10	3.80	0.94	3.23	34	54
CZ000004310953	01	05.01.07	07	406512011	06	30.30	1.30	4.28	0.99	3.27	46	85
CZ000004310953	01	06.02.07	07	406512011	06	32.80	1.19	3.64	1.05	3.19	58	117
CZ000004310953	01	06.03.07	07	406512011	06	34.60	1.22	3.53	1.10	3.18	89	145
CZ000004310953	01	05.04.07	07	406512011	06	42.00	1.61	3.84	1.45	3.46	86	175
CZ000004310953	01	04.05.07	07	406512011	06	23.60	0.90	3.80	0.83	3.51	79	204
CZ000004310953	01	04.06.07	07	406512011	06	29.30	1.06	3.61	0.99	3.38	105	235
CZ000004310953	01	04.07.07	07	406512011	06	29.40	1.05	3.58	0.98	3.33	166	265

Praktické použitie programu SAS (SAS Enterprise Guide)

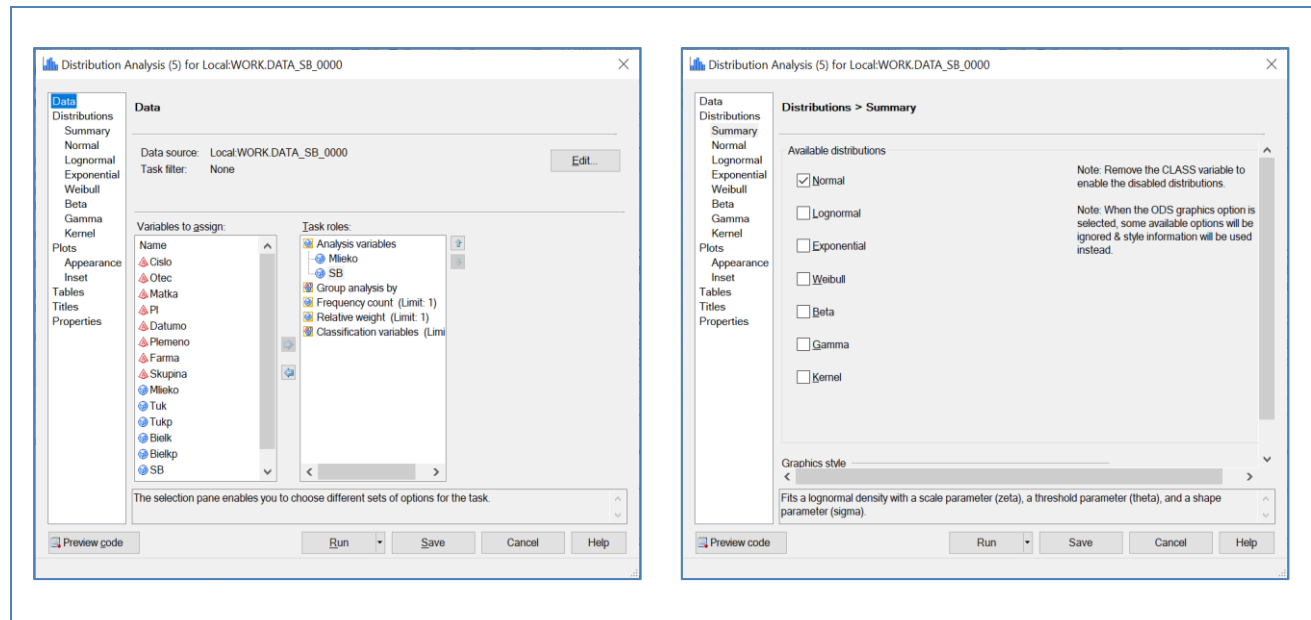
Úlohy (Tasks): Describe - Distribution Analysis, Analyze - Capability

Príklad 1.9 (SAS)

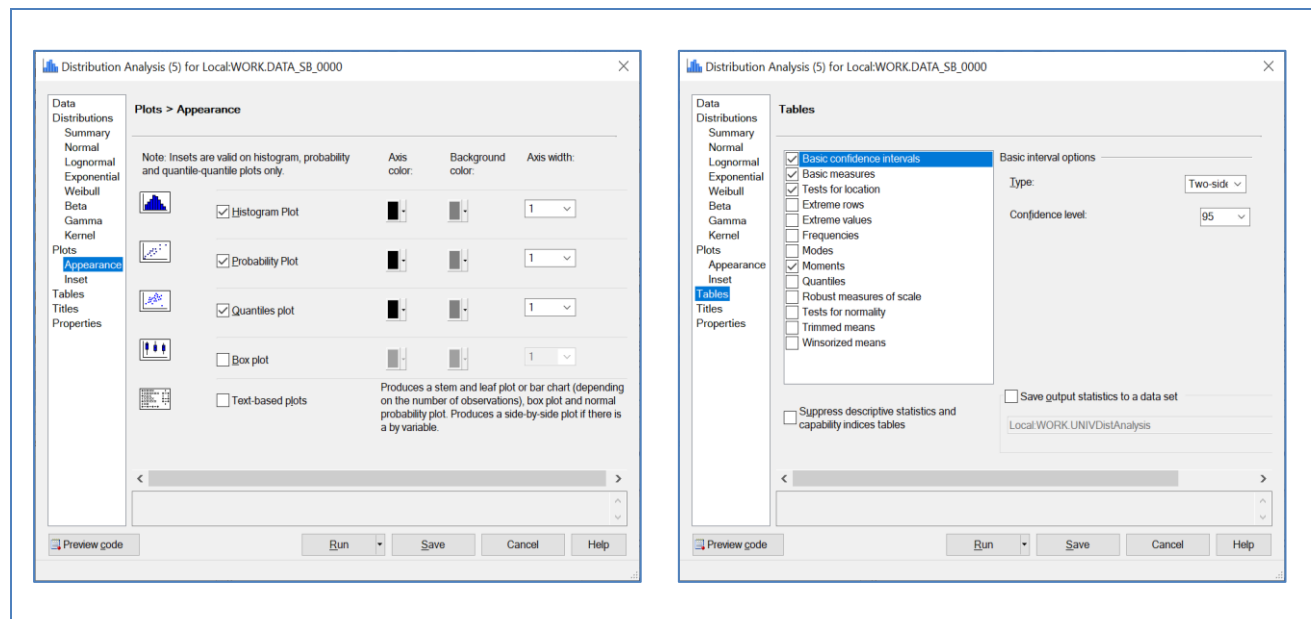
Zistite či ukazovateľ denný nádoj mlieka v kilogramoch (Mlieko) a ukazovateľ somatické bunky (SB) majú normálne rozdelenie počtosti. V prípade že niektorý z ukazovateľov nemá normálne rozdelenie počtosti uskutočnite logaritmickú transformáciu (normalizáciu) ukazovateľa.

Distribučná analýza (Príklad 1.9) má nasledovné základné nastavenia a parametre:

Tab 3.1a Nastavenie analýzy - príklad 1.9 (zdroj: projekt KEGA, Candrák, 2021)



Tab 3.1b Nastavenie analýzy - príklad 1.9 (zdroj: projekt KEGA, Candrák, 2021)



Výsledky (ukazovateľ Mlieko)

Tab 3.2 Výsledky analýzy – sumárna štatistika

Moments			
N	300	Sum Weights	300
Mean	29.1353333	Sum Observations	8740.6
Std Deviation	9.47688006	Variance	89.8112557
Skewness	0.22776211	Kurtosis	-0.0480214
Uncorrected SS	281513.86	Corrected SS	26853.5655
Coeff Variation	32.5271036	Std Error Mean	0.54714793

Tab 3.3 Výsledky analýzy – základné štatistické charakteristiky

Basic Statistical Measures			
Location		Variability	
Mean	29.13533	Std Deviation	9.47688
Median	29.35000	Variance	89.81126
Mode	22.40000	Range	52.30000
		Interquartile Range	13.05000

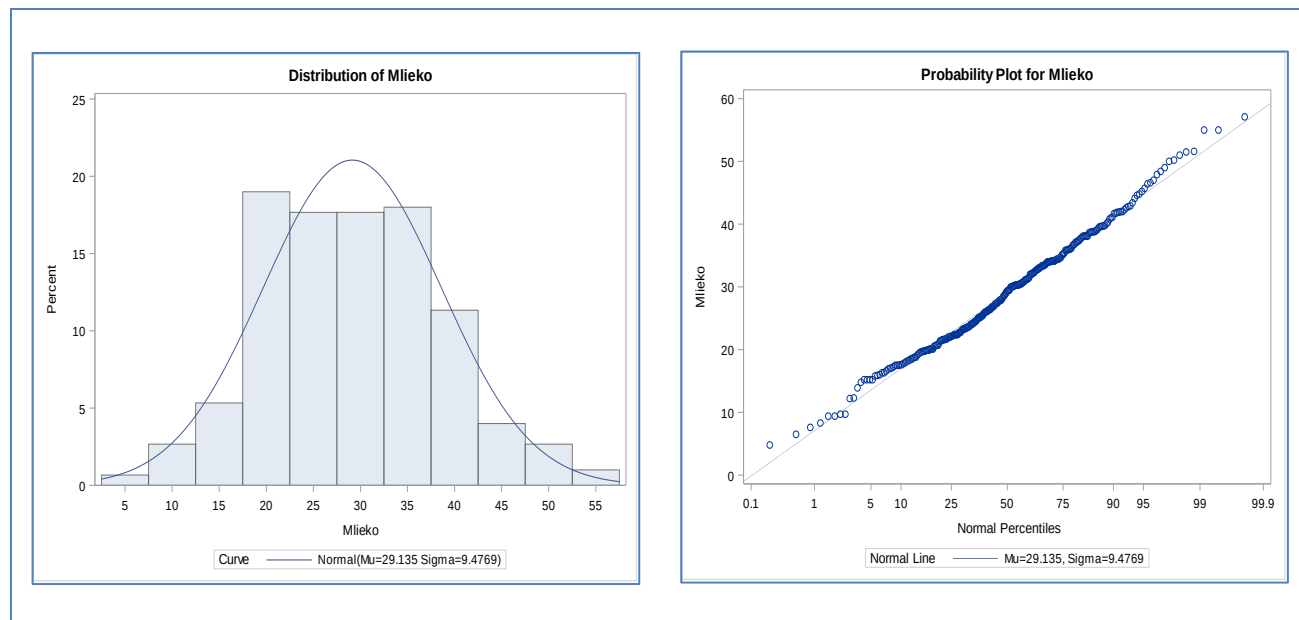
Tab 3.4 Výsledky analýzy – konfidenčné intervaly (intervaly spoľahlivosti)

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	29.13533	28.05858	30.21208
Std Deviation	9.47688	8.77438	10.30261
Variance	89.81126	76.98968	106.14374

Na základe vypočítaných sumárnych štatistík a základných štatistických charakteristík môžeme konštatovať, že ukazovateľ mlieko má iba mierne skreslené normálne rozdelenie početnosti (koeficient šikmosti dosiahol hodnotu +0.22776211. Hodnoty priemeru a mediánu sú približne rovnaké, modus je posunutý do ľava.

Grafické výsledky (ukazovateľ Mlieko)

Tab 3.5 Grafické výsledky kontroly normality – mlieko (zdroj: projekt KEGA, Candrák, 2021)



Na základe grafického znázornenia distribučnej analýzy môžeme konštatovať, že hodnotený ukazovateľ Mlieko má normálne rozdelenie. Potvrdzuje to aj graf pravdepodobnosti normality percentíl. Jednotlivé body ležia v približnej priamke, čo naznačuje, že údaje sú s najväčšou pravdepodobnosťou normálne rozložené (veľmi malá odchylka je iba na začiatku a konci percentíl).

Numerické výsledky testov normality (ukazovateľ Mlieko)

Tab 3.6 Testy normality – mlieko (zdroj: projekt KEGA, Candrák, 2021)

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.993035	Pr < W	0.1771
Kolmogorov-Smirnov	D	0.043526	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.088601	Pr > W-Sq	0.1632
Anderson-Darling	A-Sq	0.582795	Pr > A-Sq	0.1334

Testy normality jednoznačne potvrdili, že ukazovateľ Mlieko má normálne rozdelenie početnosti (hodnoty $p > 0,05$). Stanovená hypotéza H_0 o dodržaní normality ukazovateľa je platná.

Výsledky (ukazovateľ počet somatických buniek, SB)

Tab 3.7 Výsledky analýzy – sumárna štatistika

Moments			
N	300	Sum Weights	300
Mean	1005.06	Sum Observations	301518
Std Deviation	1945.46847	Variance	3784847.57
Skewness	3.72503273	Kurtosis	15.8871944
Uncorrected SS	1434713106	Corrected SS	1131669425
Coeff Variation	193.567396	Std Error Mean	112.321675

Tab 3.8 Výsledky analýzy – základné štatistické charakteristiky

Basic Statistical Measures			
Location		Variability	
Mean	1005.060	Std Deviation	1945
Median	329.000	Variance	3784848
Mode	0.000	Range	14426
		Interquartile Range	730.50000

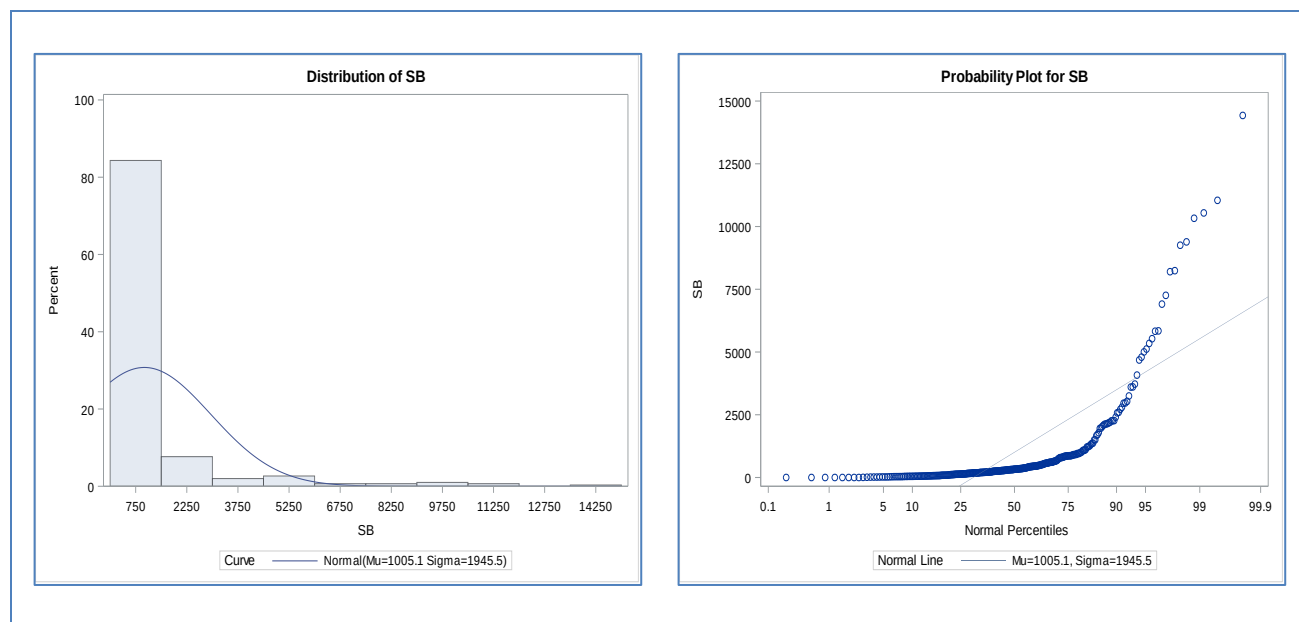
Tab 3.9 Výsledky analýzy – konfidenčné intervaly (intervaly spoľahlivosti)

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	1005	784.01884	1226
Std Deviation	1945	1801	2115
Variance	3784848	3244518	4473135

Na základe vypočítaných sumárnych štatistík a základných štatistických charakteristík môžeme konštatovať, že ukazovateľ somatické bunky má výrazné skreslené normálne rozdelenie početnosti (koeficient šikmosti dosiahol hodnotu +3.72503273 a koeficient špicatosti hodnotu až +15.8871944. Obrovské rozdiely sú aj pri hodnotách priemeru, modusu a mediánu. Podľa číselných hodnôt je rozdelenie početnosti výrazne skosené do prava.

Grafické výsledky (ukazovateľ počet somatických buniek)

Tab 3.10 Grafické výsledky kontroly normality – somatické bunky (zdroj: projekt KEGA, Candrák, 2021)



Histogram aj graf pravdepodobnosti percentíl ukazujú, že ukazovateľ počet somatických buniek nemá normálne rozdelenie.

Numerické výsledky testov normality (ukazovateľ počet somatických buniek)

Tab 3.11 Testy normality – somatické bunky (zdroj: projekt KEGA, Candrák, 2021)

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.51037	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.30298	Pr > D	<0.0100
Cramer-von Mises	W-Sq	9.99515	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	50.52059	Pr > A-Sq	<0.0050

Testy normality potvrdzujú rovnaký výsledok, ktorý iba potvrdzuje výraznú odlišnosť rozdelenia početnosti od normálneho rozdelenia. Pre takýto prípad je potrebné uskutočniť odporúčanú transformáciu počtu somatických buniek a následne vykonať novú distribučnú analýzu takto transformovaných (normalizovaných údajov).

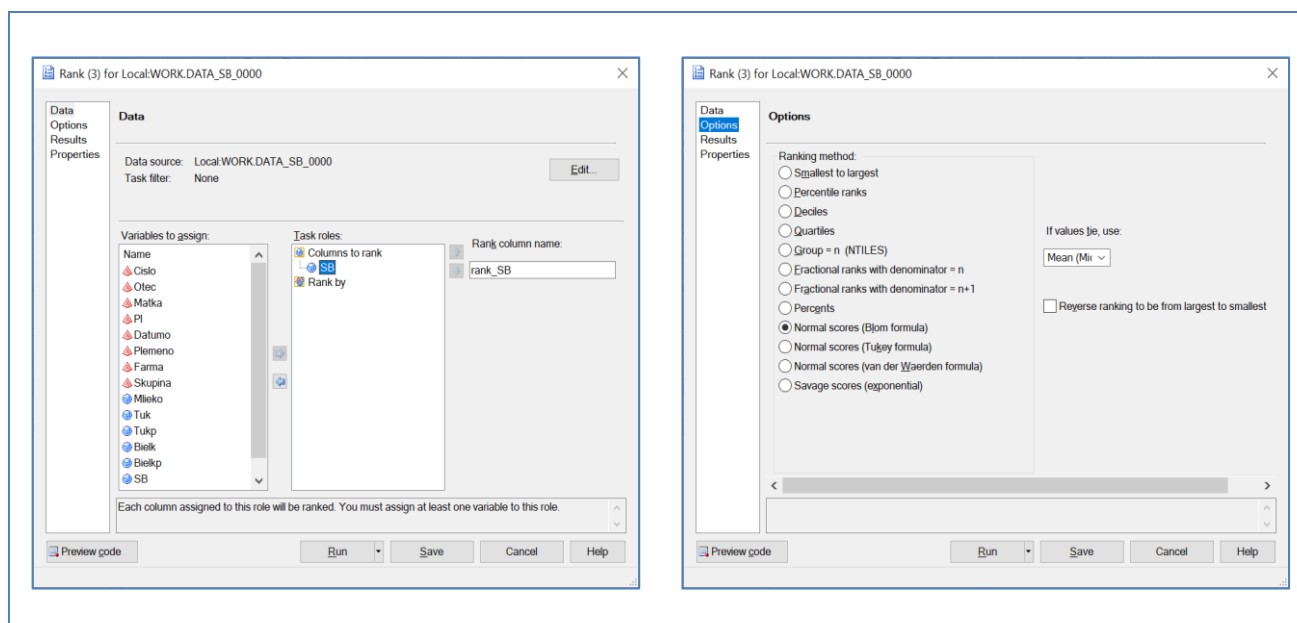
Postup transformácie (normalizácie) ukazovateľa počet somatických buniek

Jednou z možností transformácie ukazovateľov, ktoré nemajú normálne rozdelenie, je často krát používaná logaritmická transformácia údajov. Program SAS Enterprise Guide poskytuje možnosti veľmi efektívnej logaritmickkej transformácie.

Úlohy (Tasks): Data – Rank

Nastavenie parametrov transformácie počtu somatických buniek:

Tab 3.12 Nastavenie transformácie (zdroj: projekt KEGA, Candrák, 2021)



Pre transformáciu údajov sme použili transformáciu na tzv. normálne skóre (Blom formula, vzorec). Transformáciou sa pôvodný hodnotený súbor údajov rozšíri o nový ukazovateľ (rank_SB), ktorý následne môžeme použiť pri opakovanej distribučnej analýze, aby sme overili, či uskutočnená transformácia údajov bola úspešná.

Uvedený postup transformácie sa dá ale požiť aj prepočtom ukazovateľa počet somatických buniek podľa vhodne zvoleného logaritmického vzorca. V prípade hodnotenia somatických buniek v rámci kontroly mliekovej úžitkovosti hovädzieho dobytku v Slovenskej republike sa používa práve takýto postup. Využíva sa nasledovný vzorec na prepočet skóra za somatické bunky:

$$\text{skóre} = \ln(\text{SB}/100) + 3 \text{ (požitý prirodzený logaritmus).}$$

Výsledky (transformovaný ukazovateľ počtu somatických buniek, rank_SB)

Tab 3.13 Výsledky analýzy – sumárna štatistika

Moments			
N	300	Sum Weights	300
Mean	0	Sum Observations	0
Std Deviation	0.99445219	Variance	0.98893515
Skewness	0.01598122	Kurtosis	-0.1594861
Uncorrected SS	295.691611	Corrected SS	295.691611
Coeff Variation	.	Std Error Mean	0.05741472

Tab 3.14 Výsledky analýzy – základné štatistické charakteristiky

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	0.99445
Median	0.00000	Variance	0.98894
Mode	-2.28140	Range	5.14693
		Interquartile Range	1.34769

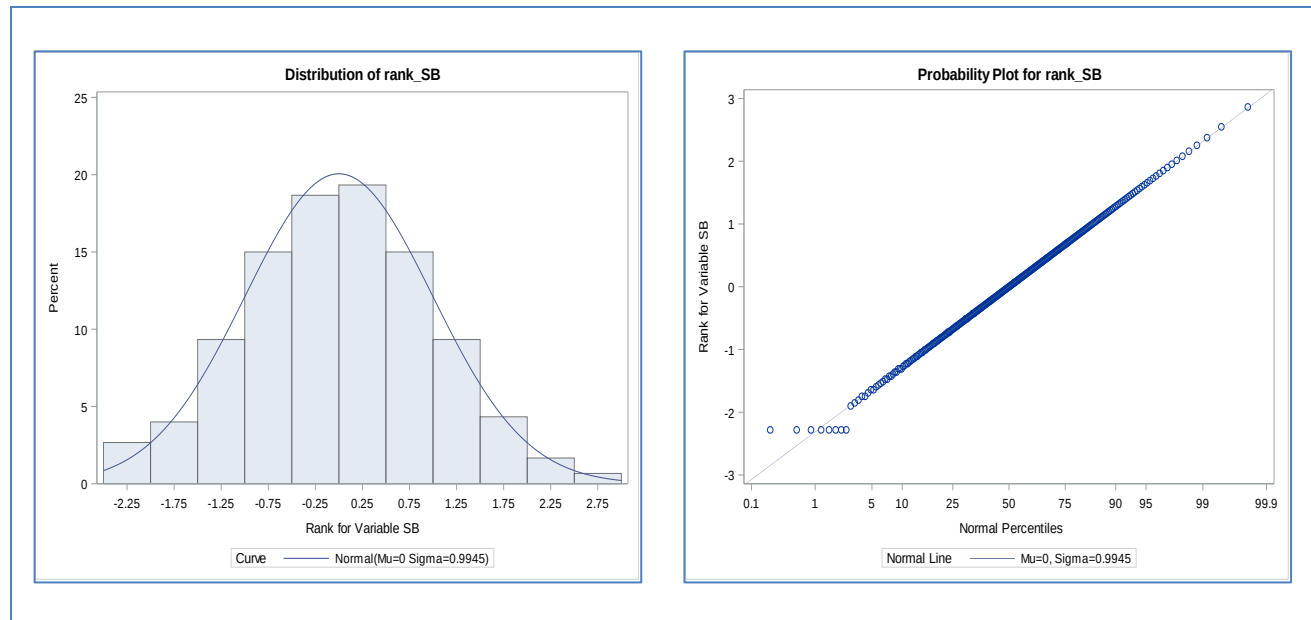
Tab 3.15 Výsledky analýzy – konfidenčné intervaly (intervaly spoľahlivosti)

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	0	-0.11299	0.11299
Std Deviation	0.99445	0.92074	1.08110
Variance	0.98894	0.84775	1.16878

Na základe vypočítaných sumárnych štatistík a základných štatistických charakteristík môžeme jednoznačne tvrdiť, že transformovaný ukazovateľ má skoro ideálne normálne rozdelenie početnosti (koeficient šikmosti dosiahol hodnotu +0.01598122, čo je prakticky 0). Hodnoty priemeru a mediánu sú po transformácii rovné 0, iba modus má mierne zápornú hodnotu.

Grafické výsledky (transformovaný ukazovateľ počtu somatických buniek)

Tab 3.16 Grafické výsledky kontroly normality – transformované somatické bunky



Grafické zobrazenia ukazujú jednoznačne, že transformovaný počet somatických buniek má priam ideálne normálne rozdelenie početnosti (minimálna odchylka od normality je iba na začiatku grafu, čo tvorí iba niekoľko málo pozorovaní).

Numerické výsledky testov normality (ukazovateľ transformovaný počet somatických buniek)

Tab 3.17 Testy normality – transformované somatické bunky

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.999542	Pr < W	1.0000
Kolmogorov-Smirnov	D	0.009190	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.001642	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.014594	Pr > A-Sq	>0.2500

Normálne rozdelenie početnosti pre transformovaný počet somatických buniek potvrdzujú aj jednotlivé testy normality (p hodnoty sú výrazne väčšie ako je definovaná hraničná hodnota 0,05).

Praktické použitie programu R (R Studio)**Príklad 1.9 (Program R)**

Zistite či ukazovateľ denný nádoj mlieka v kilogramoch (Mlieko) a ukazovateľ somatické bunky (SB) majú normálne rozdelenie početnosti. V prípade že niektorý z ukazovateľov nemá normálne rozdelenie početnosti uskutočnite logaritmickú transformáciu (normalizáciu) ukazovateľa.

Zdrojový kód potrebný na import a zobrazenie excelovského súboru "data_sb.xls", ktorý je možné použiť priamo v programe R, alebo R Studio ako súbor s názvom **data_sb**:

```
# Import a zobrazenie údajov (databáza projektu KEGA)
library(readxl)
data_sb <- read_excel("C:/Biometrics/Data/data_sb.xls",
  sheet = "data")
View(data_sb)
```

Zadanie analýzy (program R)

```
# Distribučná analýza
# Výpočet sumárnej štatistiky
descr(data_sb_rank, digits=4)

# Grafické výstupy
library("dplyr")
library("ggpubr")

hist(data_sb$Mlieko)
hist(data_sb$SB)

ggdensity(data_sb$Mlieko,
  main = "Density plot of Mlieko",
  xlab = "Mlieko")
ggdensity(data_sb$SB,
  main = "Density plot of SB",
  xlab = "SB")

ggqqplot(data_sb$Mlieko)
ggqqplot(data_sb$SB)

# Testovanie normality
shapiro.test(data_sb$Mlieko)
shapiro.test(data_sb$SB)
```

Výsledky analýzy (program R)

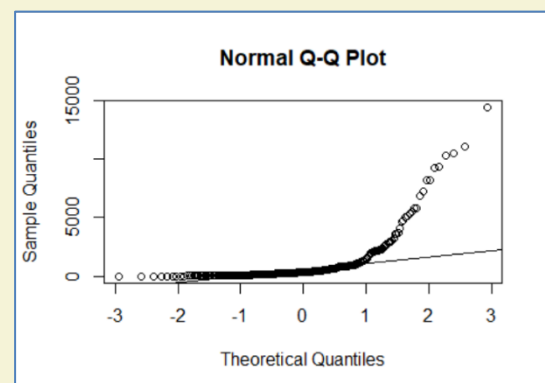
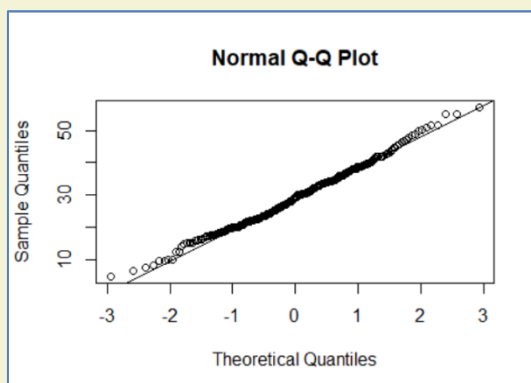
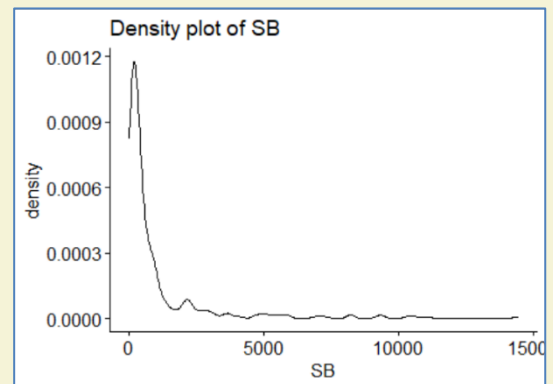
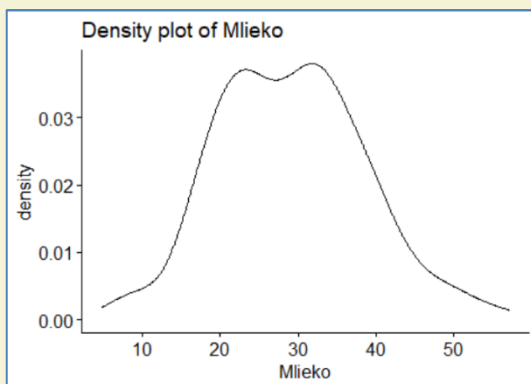
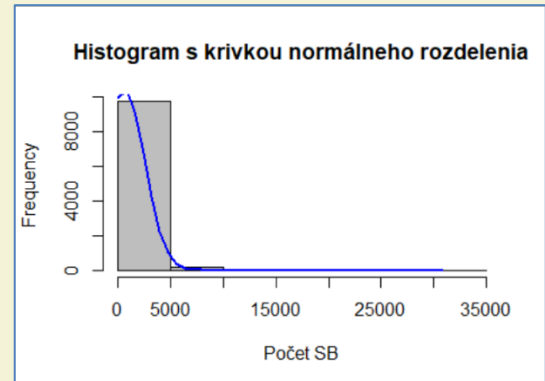
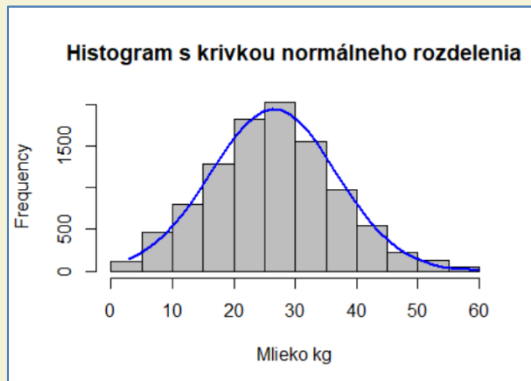
Výsledok distribučnej analýzy – sumárna štatistika (Mlieko, transformované SB, SB)

#

N: 300

	Mlieko	rank_SB	SB
-----	-----	-----	-----
Mean	29.1353	0.0000	1005.0600
Std.Dev	9.4769	0.9945	1945.4685
Min	4.8000	-2.2814	0.0000
Q1	22.1500	-0.6738	135.5000
Median	29.3500	0.0000	329.0000
Q3	35.2000	0.6738	866.0000
Max	57.1000	2.8655	14426.0000
MAD	10.0076	0.9990	375.8391
IQR	12.9750	1.3451	730.2500
Skewness	0.2255	0.0158	3.6879
SE .Skewness	0.1407	0.1407	0.1407
Kurtosis	-0.0867	-0.1956	15.4799

Výsledok distribučnej analýzy – grafické výstupy (Mlieko, SB)



Výsledok distribučnej analýzy – test normality (Mlieko, SB)

#

Shapiro-Wilk normality test

data: data_sb\$Mlieko

W = 0.99304, **p-value = 0.1771** # potvrdená normalita

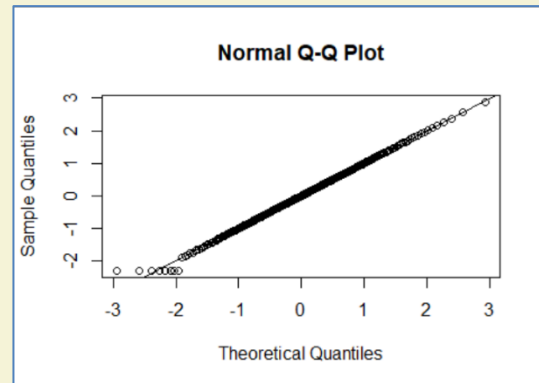
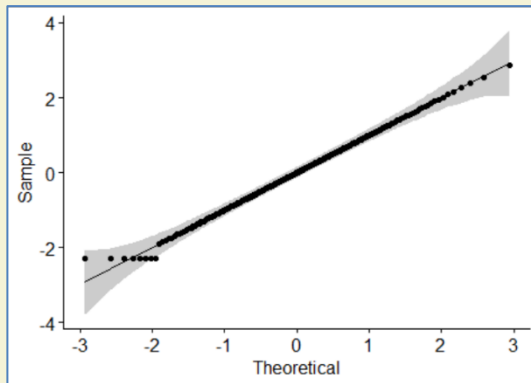
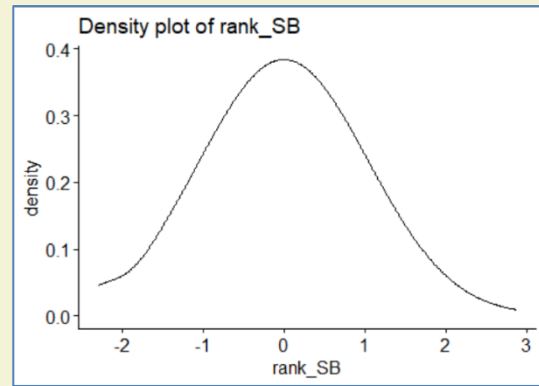
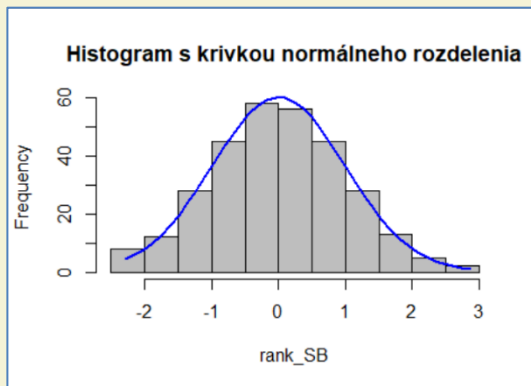
#

Shapiro-Wilk normality test

data: data_sb\$SB

W = 0.51037, **p-value < 2.2e-16** # nepotvrdená normalita

Výsledok distribučnej analýzy – grafické výstupy (transformované SB)



Výsledok distribučnej analýzy (transformované SB)

#

Shapiro-Wilk normality test

data: data_sb_rank\$rank_SB

W = 0.99655, p-value = 0.7657 # potvrdená normalita