Willkommen

in der Welt von

Mag. Helmut Grillenberger - über 25 Jahre Kompetenz im analytischen Bereich

Statistik ● Data Mining ● Maschinelles Lernen ● Simulation ● Software

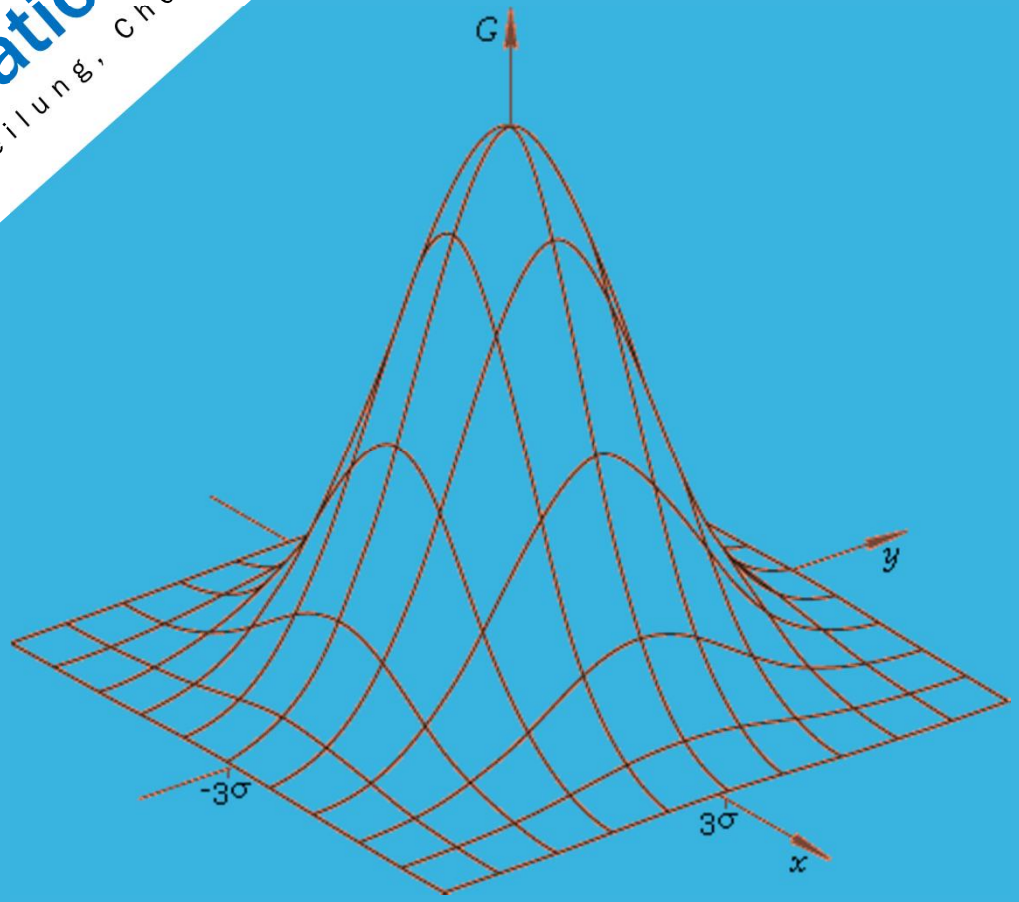
www.usedata.com

Kundenliste

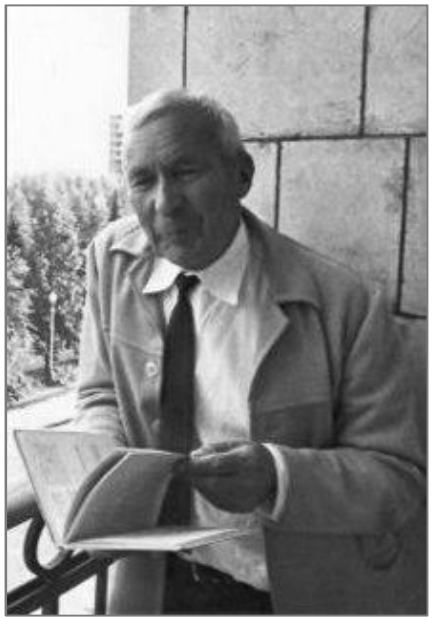
3 Pagen
Brau Union
Breuninger, Stuttgart
Doc Morris
Deutsche Post
Donauuniversität Krems
Institut für Bildung im Gesundheitsdienst
Institut f. höhere Studien
Jelmoli Versand
Klingel, Pforzheim
La Redoute
Landesstatistik OÖ
Landesstatistik SBG
Landesstatistik STMK
Landesstatistik VBG
Magistrat Salzburg
Österr. Institut für
KH Betriebsführung
Peter Hahn GmbH.
Ulla Popken
Universal Versand
Universitätsklinik Graz
Vamed, Wien
Voest, Linz
WEKA Verlag Augsburg
Wiener Kranken-anstaltenverbund
Yves Rocher Deutschland

Verhaltenssimulation mit R

Multivariate Normalverteilung, Cholesky Decomposition, ...



Am Anfang ...



Wikipedia: Andrei Kolmogorov

Axiom 1: $0 \leq P(A) \leq 1$

Axiom 2: $P(\emptyset) = 0$

Axiom 3: $P(S) = 1$

Axiom 4: $P\bigcup_{i=1}^{\infty} A_i = \sum_{i=1}^{\infty} P(A_i)$

Volladditivität

Zu Beginn des 20 Jahrhunderts legte der Russe Andrei Kolmogorov mit nur vier einfachen Axiomen den Grundstein für die heutige Wahrscheinlichkeitsrechnung.

Axiome werden nicht hinterfragt. Es handelt sich um Aussagen, die dem Hausverstand nicht widersprechen. Die Wahrscheinlichkeitsrechnung ist eine Wissenschaft, die auf klaren Regeln aufbaut.

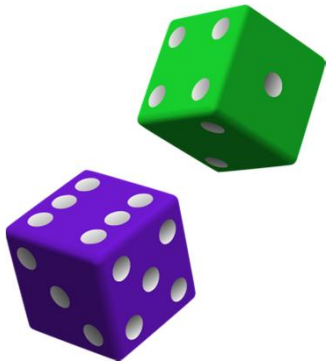
Jede Abbildung die dem Ereignis A einen Wert P(A) zuordnet und dabei die Axiome 1 bis 4 erfüllt ist eine Wahrscheinlichkeitsverteilung.

$$A \rightarrow P(A)$$

Diskrete Wahrscheinlichkeitsverteilung

Population, Stichprobenraum, Grundgesamtheit

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12



X = Summe aus beiden Würfen

random variable

X =	2	3	4	5	6	7	8	9	10	11	12
p =	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Wahrscheinlichkeitsverteilung

Stetige Wahrscheinlichkeitsverteilung

Population, Stichprobenraum, Grundgesamtheit

Körpergröße:	*192 cm*
Gewicht	*94 kg*
Umsatz	*400 €*

random variables

168 cm
64 kg
670 €

182 cm
85 kg
300 €

164 cm
58 kg
720 €

176 cm
70 kg
820 €

180 cm
75 kg
150 €

164 cm
59 kg
1.040 €

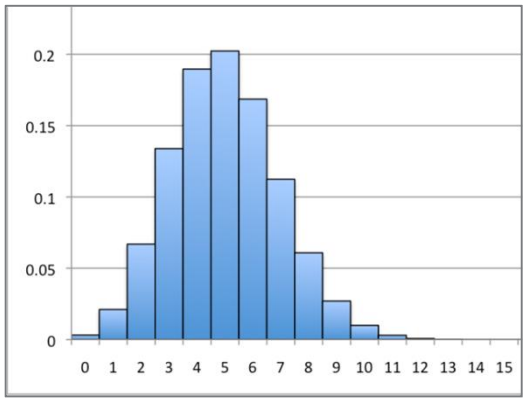
174 cm
65 kg
680 €

182 cm
74 kg
410 €

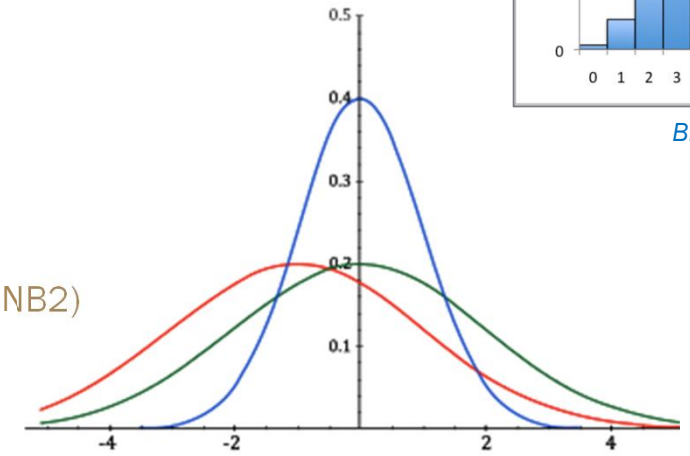
178 cm
84 kg
220 €

Wahrscheinlichkeitsverteilungen in R

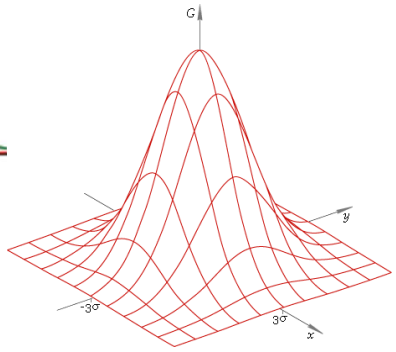
- Beta Verteilung
- Binomialverteilung
- Cauchy Verteilung
- χ^2 Verteilung
- Exponentialverteilung
- Poissonverteilung
- F-Verteilung
- Γ-Verteilung
- Geometrische Verteilung
- Hypergeometrische Verteilung
- Log-Normal Verteilung
- Multinomiale Verteilung
- Negative Binomial Verteilung (Pascal, NB2)
- Normalverteilung
- Poisson Verteilung
- t-Verteilung
- Gleichverteilung
- Weibull Verteilung
- diverse Testverteilungen
- ...



Binomial Verteilung



(eindimensionale) Normalverteilung



bivariate Normalverteilung

Normalverteilung in *R*

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

d → Dichte (density)
p → Wahrscheinlichkeit (probability)
q → Quantil (quantile)
r → Zufall (random)

Multivariate Normalverteilung

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$f(\boldsymbol{x}) = \frac{1}{(2\,\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2}$$

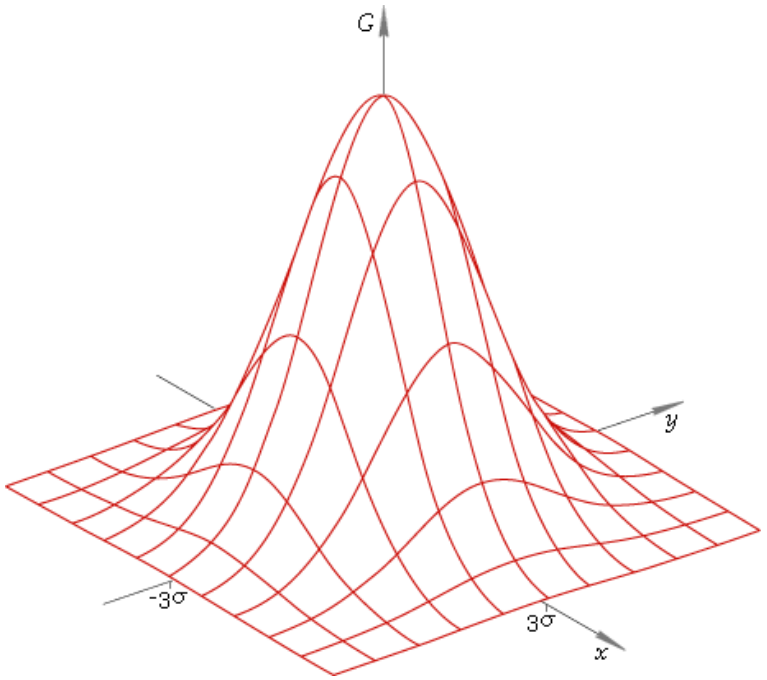
am Beispiel einer zweidimensionalen Normalverteilung:

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}$$

am Beispiel einer dreidimensionalen Normalverteilung:

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\\ \mu_3\end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{pmatrix}$$

usw.



$$\boldsymbol{x} \sim \mathbb{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Fragestellung

DOB	Heimtex	Wäsche
91.98111	92.73000	94.90000
121.14430	85.90000	105.20000
72.72758	38.90000	50.66667
65.50214	51.00000	144.00000
105.72602	21.27500	84.92941
87.68437	87.42500	49.33000
97.15212	91.41667	45.42500
96.56613	70.57200	31.50000
87.68655	115.45000	47.41000
72.91667	39.90000	52.90000
121.12256		117.58125

Versandkunden kaufen Produkte aus unterschiedlichen Sortimenten:

- *Damenoberbekleidung*
- *Heimtextilien*
- *Wäsche*

Die Merkmale DOB, Heimtex und Wäsche genügen einer multivariaten Normalverteilung mit den Parametern $\boldsymbol{\mu}$ *und* $\boldsymbol{\Sigma}$.

$$\mu = \begin{pmatrix} 91{,}95 \\ 93{,}50 \\ 62{,}16 \end{pmatrix}$$

Mittelwertsvektor μ

$$\Sigma = \begin{pmatrix} 1669{,}08 & 317{,}16 & 170{,}33 \\ 317{,}16 & 1819{,}09 & 140{,}05 \\ 170{,}33 & 140{,}05 & 1041{,}41 \end{pmatrix}$$

Varianz-Kovarianzmatrix Σ

die Verteilung (das Kaufverhalten) dieser Merkmale soll nun simuliert werden.

Lösung: Cholesky Decomposition

obere Dreiecksmatrix U

$$U'U = \Sigma$$

Cholesky Decomposition

$$\begin{pmatrix} c_{11} & c_{12} & c_{13} \\ 0 & c_{22} & c_{23} \\ 0 & 0 & c_{33} \end{pmatrix}$$

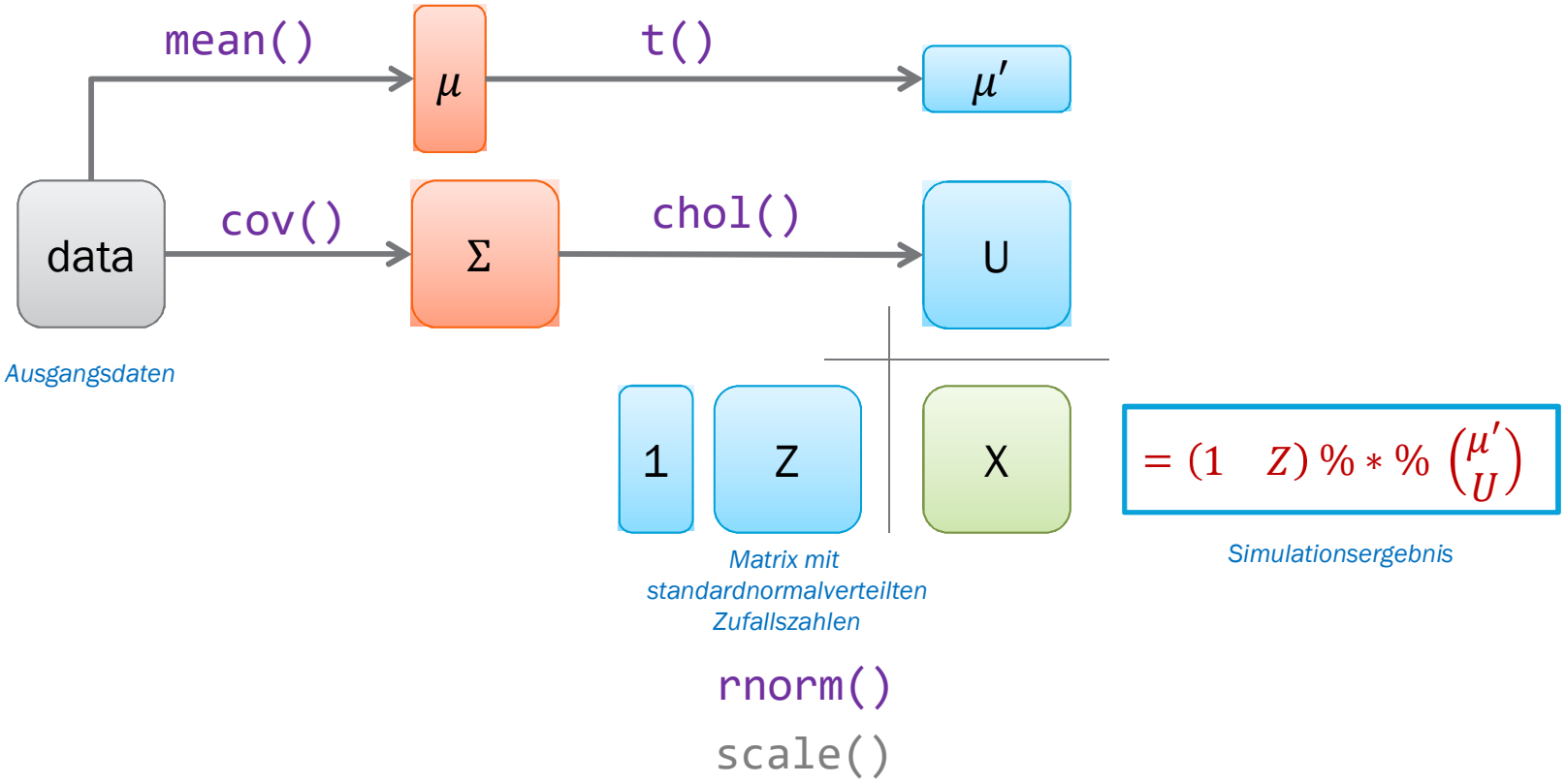
$$\begin{pmatrix} c_{11} & 0 & 0 \\ c_{12} & c_{22} & 0 \\ c_{13} & c_{23} & c_{33} \end{pmatrix} \quad \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad = \Sigma = \begin{pmatrix} 1669{,}08 & 317{,}16 & 170{,}33 \\ 317{,}16 & 1819{,}09 & 140{,}05 \\ 170{,}33 & 140{,}05 & 1041{,}41 \end{pmatrix}$$

untere Dreiecksmatrix U' | *positive definite Matrix A* | *Varianz-Kovarianzmatrix* Σ

positive definite Matrix A
$\mathbb{R}$, *symetrisch*, $x'Ax > 0$

U = chol(Σ) **U' = t(chol(Σ))**

Ablauf der Simulation



Zahlenbeispiel

```
v1  = rep(1, 10000)
dob = rnorm(10000)
htx = rnorm(10000)
wae = rnorm(10000)
```
scale()

	DOB	Heimtex	Wäsche
M	91.95353	93.498089	62.155193
DOB	40.85440	7.763083	4.169074
Heimtex	0.00000	41.938393	2.567610
Wäsche	0.00000	0.000000	31.897308

← *Mittelwertsvektor*

Cholesky Faktor obere Dreiecksmatrix

μ' U

1 Z

V1	dob	htx	wae
1	0.85948496	-0.246279031	-1.262942373
1	0.70757437	-1.451560271	0.327135852
1	0.50505217	0.692536944	0.331867482
1	-0.42524473	-0.670944382	0.112941116
1	-0.46285189	-0.865918838	1.533033943
1	-1.47145085	1.139795413	0.596967445
1	0.69399915	2.156900342	-0.278610990
1	1.29293180	1.852655234	1.481522877
1	0.71655413	0.259305194	0.048896249
1	-0.85783953	0.710960836	1.109472903
1	-1.00849413	-0.478812025	-1.038160329
1	1.48009401	1.743415638	0.530024410

Dummy-Vektor

standardnormalverteilte Zufallszahlen

DOB	Heimtex	Wäsche
127.067267	89.841796	24.821640
120.861051	38.114943	71.812835
112.587128	126.462738	76.624637
74.580409	62.058550	62.262111
73.043991	53.589687	106.901843
31.838288	129.876282	77.988811
120.306443	189.342597	61.699666
144.775476	181.232610	119.559009
121.227914	109.935602	67.368014
56.907009	116.655164	95.793466
50.752106	65.588458	23.606785
152.421875	178.104233	88.708578

Simulationsergebnis

$X = (1 \quad Z)$ %*% $\begin{pmatrix} \mu' \\ U \end{pmatrix}$

Multiplikations-Operator in R

Kontrolle

	Mittelwertsvektor μ	*Varianz-Kovarianzmatrix* Σ
Ausgangsdaten	$\mu = \begin{pmatrix} 91{,}95 \\ 93{,}50 \\ 62{,}16 \end{pmatrix}$	$\Sigma = \begin{pmatrix} 1669{,}08 & 317{,}16 & 170{,}33 \\ 317{,}16 & 1819{,}09 & 140{,}05 \\ 170{,}33 & 140{,}05 & 1041{,}41 \end{pmatrix}$
Simulationsergebnis	$\mu = \begin{pmatrix} 91{,}95 \\ 93{,}50 \\ 62{,}16 \end{pmatrix}$	$\Sigma = \begin{pmatrix} 1669{,}08 & 322{,}79 & 175{,}78 \\ 322{,}79 & 1821{,}24 & 134{,}09 \\ 175{,}78 & 134{,}09 & 1041{,}60 \end{pmatrix}$

R-Code Teil 1

```
# Bestimmung von Mittelwert, Kovarianz und Correlation --------------------

M <- colMeans(dt)
V <- cov(dt)
C <- cor(dt)
printStat(dt, "Ausgangsdaten")
```

```
# Simulation von standardnormalverteilten Zufallszahlen -------------------

set.seed(15675949)
dob <- rnorm(10000)
htx <- rnorm(10000)
wae <- rnorm(10000)
Z <- cbind(dob, htx, wae)
Z <- scale(Z)

zM <- colMeans(Z)
zV <- cov(Z)
zC <- cor(Z)
printStat(Z, "standardnormalverteilte Zufallszahlen")

Z <- cbind(1, Z)
printData(Z, "standradnormalverteilte Zufallszahlen")
```

R-Code Teil 2

```
# Cholesky Decomposition ---------------------------------------------------

U = chol(V)
U = rbind(M, U)
printData(U, "Cholesky Decomposition")


# Erstellung normalvert. Zufallszahlen mit Mw Vektor und Varianzmatrix ----

X <- Z %*% U
printData(X, "Simulation")
```

```
# Überprüfung des Simulationsergebnisses ----------------------------------

xM <- colMeans(X)
xV <- cov(X)
xC <- cor(X)

printStat(dt, "Ausgangsdaten")
printStat(X, "Simulation")
```

R-Code Helper Funktions

```r
printStat <- function(X, label = NULL, d = 2) {

  M <- colMeans(X)
  V <- cov(X)
  C <- cor(X)

  if (!is.null(label)) {
    cat("--- Statistiken für", label, "---")
  } else {
    print("--- Statistiken ---")
  }

  cat("\n", "Mittelwertsvektor:", "\n")
  print(round(M, digits = d))
  cat("\n", "Kovarianzmatrix:", "\n")
  print(round(V, digits = d))
  cat("\n", "Korrelationsmatrix:", "\n")
  print(round(C, digits = d))

}
```

Woher weiß man, welche Verteilungen man für eine Simulation nutzen soll?

Für diese Frage gibt es mehrere Antworten:

Zunächst gibt es Verteilungen, die sich aus theoretischen Überlegungen ergeben. So genügen die Mittelwerte einer Stichprobe - auf Grund des zentralen Grenzwertsatzes – einer Normalverteilung. Diese Aussage gilt, wenn die Population normalverteilt ist. Mit zunehmender Größe der Stichprobe (> 30) gilt die Aussage aber auch für andere Verteilungen.

Ich habe in meinem Vortrag die negative Binomialverteilung beschrieben. Sie ergibt sich, wenn man z. B. eine Münze solange wirft, bis das erste Mal *Kopf* erscheint. Die Anzahl der dafür benötigten Versuche unterliegt einer negativen Binomialverteilung. Diese Verteilung ist auch unter dem Begriff *Pascal Verteilung* bekannt.

Man schätzt, dass es rund 13 Formen der negativen Binomialverteilung gibt (*Boswell and Patil, 1970, Chance mechanisms generating negative binomial distributions in Random Counts in Models and Structures, Volume 1*). Die bekannteste ist die Verteilung mit der Bezeichnung *NB2*. Sie ist eine Mischung aus einer Poisson- und einer Gammaverteilung.

Die NB2 wird heute genutzt, um die Anzahl Besucher auf einer Webseite in einem Onlineshop zu simulieren. Dass es beim Besuch dieser Webseite zu einem Kauf kommt, kann mittels Binomialverteilung beschrieben werden. Mit Hilfe der Exponentialverteilung lässt sich z. B. die Distanz eines Kunden zum Geschäft beschreiben.

Fachspezifische Bücher zu bestimmten Themen enthalten ebenfalls Hinweise zu Verteilungstypen. Letztendlich ist es aber die Erfahrung, die erste Hinweise dazu gibt. Vermutet man eine Verteilung, kann man zumindest mittels statistischen Tests prüfen, ob diese auch tatsächlich vorliegt.

Warum sind häufig Tests in der Werbung ihre Kosten nicht wert?

Betrachten wir dazu eine Werbemaßnahme um den Umsatz steigern: Mittels eines Tests (dieser Test enthält die Werbemaßnahme) und einer Vergleichsgruppe können wir nach Ablauf des Testzeitraums für beide Gruppen den durchschnittlich erreichten Umsatz und damit die Differenz zwischen den beiden Gruppen berechnen. Es stellt sich nun die Frage, ob diese Differenz ausreicht, um die Wirkung der Werbemaßnahme zu bestätigen.

Um Mittelwerte zu vergleichen, können wir Gebrauch vom zentralen Grenzwertsatz machen. Dieser besagt, dass die Stichprobenmittelwerte einer Normalverteilung genügen und dass die Streuung der Stichprobenmittelwerte mit zunehmender Stichprobengöße abnimmt ($\sigma_{\bar{x}} = \sigma^2/n$).

Ist die Stichprobengröße zu klein, dann ist die Treffsicherheit des Tests gering. Der vorhandene Unterschied in den Stichprobenmittelwerten verschwindet im Rauschen beider Verteilungen. Die Streuungen in den Verteilungen machen einen vorhandenen Unterschied unsichtbar. Mit zunehmender Stichprobengröße nimmt die Streuung beiden Mittelwerte jedoch ab und auch noch so kleine Differenzen in den Stichprobenmittelwerten werden sichtbar. In der Statistik wird diese Eigenschaft mittels einer Powerfunktion (Gütefunktion) beschrieben.

In der Praxis werden aber oft – bewusst oder unbewusst - trennschwache Tests durchgeführt. Die vorhandenen Unterschiede werden dabei nicht erkannt. Es ist also wichtig, dass im Vorfeld von Tests über die Relevanz einer Werbemaßnahme gesprochen wird. Erst dann kann die mindestens erforderliche Stichprobengröße ermittelt werden, um mögliche Differenzen zu erkennen. Eine Kosten/Nutzen-Analyse entscheidet dann über den Einsatz.

Mit Hilfe der heute vorgestellten Simulation könnte man einen Parameter (z. B. den Mittelwert für DOB) erhöhen. Dann könnte man auf einfache Weise prüfen, ob diese Manipulation erkannt wird. Auch einem nicht mit der Statistik Vertrauten kann so anschaulich die Wirkung einer Werbemaßnahme näher gebracht werden.

Blog zur R-Initiative

- Helmut's R-Initiative
  Einführung der Begriffe Statistik, Informatik, Data Mining, maschinelles Lernen, Big Data und Data Science.
- Kredit Scoring, Teil 1
  mit Hilfe der logistischen Regression
- Kredit Scoring, Teil 2
  mit Hilfe von Entscheidungsbäumen

Mit Hilfe der R-Initiative biete ich Unternehmen die Möglichkeit die Programmiersprache R unter realen Aufgabestellungen zu testen.

Dienstleistungen

Kundenliste

3 Pagen
Brau Union
Breuninger, Stuttgart
Doc Morris
Deutsche Post
Donauuniversität Krems
Institut für Bildung im Gesundheitsdienst
Institut f. höhere Studien
Jelmoli Versand
Klingel, Pforzheim
La Redoute
Landesstatistik OÖ
Landesstatistik SBG
Landesstatistik STMK
Landesstatistik VBG
Magistrat Salzburg
Österr. Institut für KH Betriebsführung
Peter Hahn GmbH.
Ulla Popken
Universal Versand
Universitätsklinik Graz
Vamed, Wien
Voest, Linz
WEKA Verlag Augsburg
Wiener Kranken-anstaltenverbund
Yves Rocher Deutschland

Mag. Helmut Grillenberger
über 25 Jahre Kompetenz im analytischen Bereich

www.usedata.com
helmut.grillenberger@usedata.com
+43 6274 20804

- Aufbau und Weiterentwicklung eines **Analyseteams**
- Einrichtung und Nutzung eines **Data Warehouse**
- Zugriff auf Unternehmenskennzahlen mittels **Excel AddIn**
- Erstellen von **Scoringmodellen**.

- Implementierung von **Data Mining Algorithmen**
- Workshops zum **maschinellen Lernen**

- Workshops zur **Programmiersprache R**
- Blog zur **R-Initiative** auf usedata.com und LinkedIn

Vielen Dank
für Ihre Aufmerksamkeit!