Statistik Formelsammlung (gekürzt für 5 Abende Veranstaltung)

Formeln zum Kurs "Statistik" an der VWA Essen von Prof. Dr. Peter von der Lippe

Inhaltsverzeichnis: (gestrichene Abschnitte sind grau markiert)

1. Häufigkeitsverteilungen

Einzelwerte : x_i bedeutet $x_1,x_2,...,x_n$ (i = 1, ..., n)

Gruppierte Daten : x_j bedeutet j-te Ausprägung des Merkmals X (j=1,...,m)

n_j absolute Häufigkeit der j-ten Merkmalsauspägung

h_j relative Häufigkeit (j=1,...,m)

Summenhäufigkeiten : absolut $N_k = n_1+n_2+...+n_k$

relativ $H_k = h_1+h_2+...+h_k$ Treppenfunktion ($H_1 = 0$, $H_m = 1$)

klassierte Daten : x_{ju} = Untergrenze.., x_{jo} = Obergrenze der j-ten Größenklasse

$\overline{x}_j$ = wahrer Klassenmittelwert, $\hat{\overline{x}}_j = \frac{1}{2}\left(x_{ju} + x_{jo}\right)$ = geschätzter Klassenmittelwert.

2. Mittelwerte (geometrisches Mittel weniger relevant)

	ungewogen (Berechnung aus Einzelwerten)	gewogen (Berechn. aus gruppierten/klassierten Daten)
arithmetisch $\overline{x}$	$\overline{x} = \frac{1}{n}\sum x_i = \frac{x_1+...+x_n}{n}$	$\overline{x} = \frac{1}{n}\sum x_j n_j = \sum x_j h_j$
geometrisch $\overline{x}_G$	$\overline{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n} = \left(\prod x_i\right)^{1/n}$	$\overline{x}_G = \left(x_1^{n_1} x_2^{n_2} ... x_m^{n_m}\right)^{\frac{1}{n}} = \prod x_j^{h_j}$
harmonisch $\overline{x}_H$	$\frac{1}{\overline{x}_H} = \frac{1}{n}\sum \frac{1}{x_i}$	$\frac{1}{\overline{x}_H} = \frac{1}{n}\sum \frac{1}{x_j} \cdot n_j = \sum \frac{1}{x_j} \cdot h_j$

Zentralwert (Median) : der in der Größe nach (von kleinen zu großen Merkmalswerten) geordneten Reihe von n Elementen in der Mitte (an der $\frac{n+1}{2}$ ten Stelle) stehende Wert, bzw. der x-Wert bei der 50% Linie der Summenhäufigkeitskurve H_k.

3. Streuung

Varianz ungewogen : $s^2 = \frac{1}{n}\sum\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum x_i^{\,2} - \bar{x}^2$

gewogen : $s^2 = \sum\left(x_j - \bar{x}\right)^2 h_j = \sum x_j^{\,2} h_j - \bar{x}^2$

Standardabweichung : $s = +\sqrt{s^2}$ Variationskoeffizient: $V = \frac{s}{\bar{x}}$

5. Verhältniszahlen, Wachstumsraten

Strukturbeschreibung : Quoten, Beziehungszahlen

Beschreibung der Dynamik: Messzahlen $m_{0t} = \frac{y_t}{y_0}$ ($\rightarrow$ Indizes als Mittelwerte von Messzahlen) sowie Wachstumsraten und –faktoren

6. Indexformeln

a) historische

Dutot: Messzahl von Mittelwerten	Carli: Mittelwert von Messzahlen*
$P_{0t}^D = \frac{\bar{p}_t}{\bar{p}_0}$ mit $\bar{p}_t = \frac{1}{n}\sum p_{it}$ und $\bar{p}_0$ analog	$P_{0t}^C = \frac{1}{n}\sum \frac{p_{it}}{p_{i0}}$

* Meßzahlenmittelwert

b) aktuelle

Preisindex von	Messzahlenmittelwertformel	Aggregatformel
Laspeyres	$P_{0t}^L = \sum_{i=1}^{n} \frac{p_{it}}{p_{i0}} \frac{p_{i0}q_{i0}}{\sum p_{i0}q_{i0}}$	$P_{0t}^L = \frac{\sum p_t q_0}{\sum p_0 q_0}$
Paasche	$P_{0t}^P = \sum_{i=1}^{n} \frac{p_{it}}{p_{i0}} \frac{p_{i0}q_{it}}{\sum p_{i0}q_{it}} = \left(\sum \frac{p_{i0}}{p_{it}} \frac{p_{it}q_{it}}{\sum p_{it}q_{it}}\right)^{-1}$	$P_{0t}^P = \frac{\sum p_t q_t}{\sum p_0 q_t}$

c) Anwendungen

- Mengenindizes Q_{0t}^L, Q_{0t}^P entstehen aus Preisindizes P_{0t}^L, P_{0t}^P durch Vertauschen von Mengen und Preisen
- Wertindex (Lebenshaltungskosten): $W_{0t} = \frac{\sum p_t q_t}{\sum p_0 q_0}$
- Preisbereinigung (Deflationierung) $W_{0t} = P_{0t}^P Q_{0t}^L = P_{0t}^L Q_{0t}^P$

7. Zweidimensionale Häufigkeitsverteilungen

a) Randverteilungen $h_{i.}$ und $h_{.j}$

	y_1 ... y_j	Σ
x_1	h_{11} ... h_{1j} ...	$h_{1.}$
...	... ...	...
xi	h_{i1} ... h_{ij} ...	$h_{i.}$
Σ	$h_{.1.}$ ... $h_{.j}$ ...	1

Randverteilung von x (Summen über alle Spalten) $h_{i.} = \sum_{j=1}^{j=k} h_{ij} = \sum_j h_{ij} = h\left(X = x_i\right)$

Randverteilung von y: $h_{.j} = \sum_{i=1}^{i=m} h_{ij} = \sum_i h_{ij} = h\left(Y = y_j\right)$

b) bedingte Verteilungen und bedingte Mittelwerte

bedingte* Verteilung von x	bedingte Verteilung von y
$h_{i\mid j} = \frac{h_{ij}}{h_{.j}} = \frac{n_{ij}}{n_{.j}} = h\left(x \mid Y = y_j\right)$	$h_{j\mid i} = \frac{h_{ij}}{h_{i.}} = \frac{n_{ij}}{n_{i.}} = h\left(y \mid X = x_i\right)$
$\overline{x}\mid y = \overline{x}\,(y_j) = \sum_{i=1}^{i=m} x_i h_{i\mid j}$	$\overline{y}\mid x = \overline{y}\,(x_i) = \sum_{j=1}^{j=k} y_j h_{j\mid i}$

* bedingt heißt: *wenn* y = ...

c) Unabhängigkeit: $n_{ij} = \frac{n_{i.} n_{.j}}{n}$ bzw. $h_{ij} = h_{i.} h_{.j}$

d) Kovarianz

bei Einzelbeobachtungen (="ungewogen")	bei gruppierten Daten
$s_{xy} = \frac{1}{n}\sum_{i=1}^{i=n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$ oder nach Verschiebungssatz $= \frac{1}{n}\sum x_i y_i - \overline{x}\cdot\overline{y}$	$s_{xy} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{k}\left(x_i - \overline{x}\right)\left(y_j - \overline{y}\right)\cdot n_{ij}$ $= \sum\sum\left(x_i - \overline{x}\right)\left(y_j - \overline{y}\right)\cdot h_{ij}$ $= \sum\sum x_i y_j h_{ij} - \overline{x}\cdot\overline{y}$

e) Korrelationskoeffizent: $r_{xy} = \frac{s_{xy}}{s_x s_y}$ $\qquad -1 \le r_{xy} \le +1$

8. Regressionsanalyse (lineare einfache Regression)

$y_i = f\left(x_i\right) + u_i = a + b\cdot x_i + u_i$, (i=1,...,n) bei linearer Regressionsfunktion $\hat{y}_i = a + b\cdot x_i$

Schätzung der Koeffizenten / Parameter mit der Methode der kleinstenQuadrte

1. Normalgleichung	$an + b\sum x = \sum y$
2. Normalgleichung	$a\sum x + b\sum x^2 = \sum xy$

Berechnung der Regressionsgerade in zwei Schritten

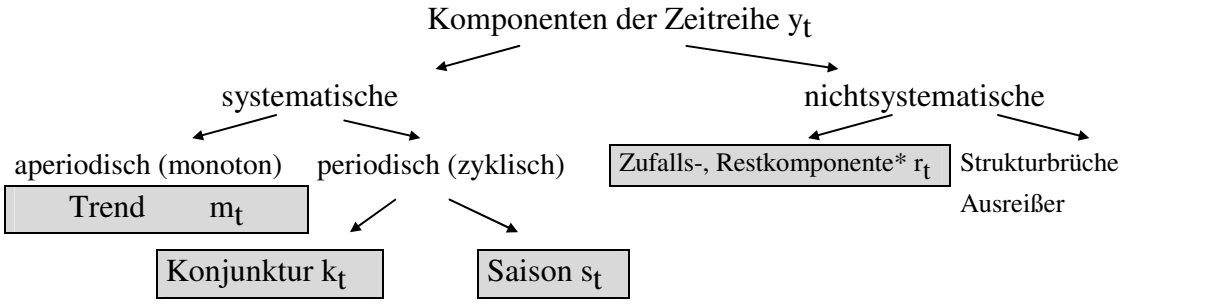
1.) $b = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \dfrac{s_{xy}}{s_x^2} = \dfrac{\text{Kovarianz}}{\text{Varianz}}$

2.) $a = \bar{y} - b\bar{x}$ das heißt: die Regressionsgrade geht durch den Schwerpunkt

Varianzzerlegung	$\frac{1}{n}\sum (y_i - \bar{y})^2 = \frac{1}{n}\sum (\hat{y}_i - \bar{y})^2 + \frac{1}{n}\sum (y_i - \hat{y}_i)^2$		
	Gesamtvarianz s_y^2	erklärte Varianz $s_{\hat{y}}^2$	Residualvarianz s_u^2

9. Zeitreihenanalyse

a) Komponentenmodell



glatte Komponente: $m_t + k_t$ *additive* Überlagerung $y_t = m_t + k_t + s_t + r_t$

* oder irreguläre Komponente

b) Bestimmung des Trends, bzw. der glatten Komponente

1. Methode der kleinsten Quadrate - linearer Trend

$\hat{y}_t = m_t = a + b \cdot t$ (Bestimmung von a und b wie in der Regressionsanalyse)

Nichtlineare Trends nach Variablensubstitution oder linearisierender Transformation

2. Methode der gleitenden Mittelwerte

p ungerade Zahl p = 2k + 1

$$\tilde{y}_t = \frac{1}{p}\left(y_{t-k} + y_{t-k+1} + \ldots + y_t + \ldots + y_{t+k-1} + y_{t+k}\right)$$

p gerade $(p = 2k)$ zentrierter gleitender Durchschnitt

$$\tilde{y}_t^z = \frac{1}{p}\left(\frac{1}{2}y_{t-k} + y_{t-k+1} + \ldots + y_t + \ldots + y_{t+k-1} + \frac{1}{2}y_{t+k}\right)$$

c) Exponential Smoothing (exponentielles Glätten) gestrichen

10. Wahrscheinlichkeitsrechnung

a) Binomialkoeffizient

Anzahl der Kombinationen ohne Wiederholungen $K = \binom{n}{i} = \dfrac{n!}{i!(n-i)!}$

Die Entwicklung des Binoms $(a+b)^n$ führt zu folgender Summe

$$\binom{n}{0}a^0b^n+\binom{n}{1}a^1b^{n-1}+\ldots+\binom{n}{i}a^ib^{n-i}+\ldots+\binom{n}{n-1}a^{n-1}b+\binom{n}{n}a^nb^0$$

Nach Definition gilt $\binom{n}{0}=\binom{n}{n}=1$ und $\binom{n}{1}=n$, Symmetrie $\binom{n}{i}=\binom{n}{n-i}$

b) Additionssatz $P(A\cup B)$ ***Multiplikationssatz*** $P(A\cap B)=P(AB)$ ***bei zwei Ereignissen***

Sätze	Additionssätze	Multiplikationssätze
a) allgemein	(1) $P(A\cup B)=P(A)+P(B)-P(AB)$	(3) $P(AB)=P(A)\,P(B\mid A)=P(B)\,P(A\mid B)$
b) speziell	(2) $P(A\cup B)=P(A)+P(B)$ wenn A und B disjunkte Ereignissen sind, also $A\cap B=\varnothing$	(4) $P(AB)=P(A)\,P(B)$ wenn je zwei Ereignisse A und B paarweise unabhängig sind

c) bedingte Wahrscheinlichkeit und (stochastische)Unabhängigkeit

P(A|B) ist die Wahrscheinlichkeit des Eintreffens des Ereignisses A unter der Voraussetzung, daß Ereignis B eingetreten ist $P(A\mid B)=\frac{P(A\cap B)}{P(B)}$ entsprechend $P(B\mid A)=\frac{P(A\cap B)}{P(A)}$.

Unabhängigkeit bedeutet $P(A\mid B)=P(A\mid\overline{B})=P(A)$ und damit auch $P(AB)=P(A)P(B)$.

11. Einige Wahrscheinlichkeitsverteilungen

a) Diskrete Verteilungen

Name	Wahrscheinlichkeits- und Verteilungsfunktion	Momente
Binomialverteilung	$f(x)=\binom{n}{x}\pi^x(1-\pi)^{n-x}$	$E(X)=n\pi$ $V(X)=n\pi(1-\pi)$
hypergeometrische Verteilung	$f(x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	$E(X)=n\pi$ $V(X)=n\pi(1-\pi)\frac{N-n}{N-1}=n\pi(1-\pi)F$ F = finite multiplier

geometr. Verteilung gestrichen

***b) stetige Verteilung*:** Normalverteilung (Tabelle siehe Seite 8)

Wahrscheinlichkeitsfunktion (Dichte)	$f_N(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\quad -\infty<x<+\infty$
Verteilungsfunktion (Tabellierung)	wenn $X\sim N(\mu,\sigma^2)$, dann ist $Z=\frac{X-\mu}{\sigma}\sim N(0,1)$ (Standardnormalverteilung). Die N (0,1)-verteilung ist tabelliert (siehe unten).
Parameter	μ (zugleich E(X)) und σ^2 (zugleich V(X))

12. Grenzwertsätze

Zentraler Grenzwertsatz (von Ljapunoff) :

Unter sehr allgemeinen, praktisch immer erfüllten Bedingungen sind Summen und Durchschnitte von unabhängigen Zufallsvariablen für große n angenähert normalverteilt.

13. Schätztheorie

a) Arten der Schätzung

Fragestellungen (Begriffspaare)

homograd / heterograd	Schlussweisen (direkt / indirekt)
Arten statistischer Inferenz: Schätzen[1] / Testen	mit Zurücklegen (m.Z.) / ohne Zurücklegen (o.Z.)

	direkter Schluss von der Grundgesamtheit auf die Stichprobe	indirekter Schluss von der Stichprobe auf die Grundgesamtheit
heterograd	$\mu \rightarrow \bar{x}$	$\bar{x} \rightarrow \mu$
homograd	$\pi \rightarrow p$	$p \rightarrow \pi$

b) Konfidenzintervalle (indirekter Schluß)

Intervalle für μ bzw. π bei einer Sicherheitswahrscheinlichkeit von $1-\alpha$

	heterograd	homograd
Ziehen mit Zurücklegen (Z.m.Z.)	$\bar{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}}$	$p \pm z_\alpha \sqrt{\frac{p\,q}{n-1}}$
Ziehen ohne Zurücklegen (Z.o.Z.)	$\bar{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$	$p \pm z_\alpha \sqrt{\frac{p\,q}{n-1} \frac{N-n}{N}}$

14. Durchführung und Interpretation eines Hypothesentests

a) Arten von Tests

Beispiele für die Fragestellung bei Tests (Parametertests)

a) Abnahmekontrolle (Tests mit **einer** Stichprobe; einseitiger Test des Ausschußanteils [homograd])
b) Qualitätskontrolle (Tests mit **einer** Stichprobe; zweiseitiger Test, Füllmenge [heterograd])
c) Kontrollgruppenexperiment (Tests mit **zwei** Stichproben)

b) Arbeitssschritte

1. Festlegung von Nullhypothese H_0 und Alternativhypothese H_1 (und damit Entscheidung ob ein- oder zweiseitig zu testen ist)
2. Festlegung des Signifikanzniveaus α und damit der Signifikanzschranke z_α
3. Berechnung der Prüfgröße z
4. Entscheidungsregel: Vergleich von z mit z_α

c) Schema der Fehlerarten

[1] weiter unterteilt in Punkt- und Intervallschätzung

	state of nature (wahrer Zustand)	
action (Entscheidung)	H_0 ist richtig	H_0 ist falsch
H_0 ablehnen	Fehler erster Art (α-Fehler)	kein Fehler
H_0 annehmen	kein Fehler	Fehler zweiter Art (ß-Fehler)

d) Prüfgrößen

(nur parametrische Tests, nur mit Normalverteilung, Prüfgrößen ohne Endlichkeitskorrektur

Prüfgrößen bei	**homograd**	**heterograd**
einer Stichprobe	$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$	$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
zwei Stichproben	zuerst berechnen $p = \frac{1}{n}(n_1 p_1 + n_2 p_2)$ dann $z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	zuerst berechnen $\hat{\sigma}^2 = \frac{1}{n-2}(n_1 s_1^2 + n_2 s_2^2)$ dann $z = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Bei allen Prüfgrößen z ist mit der Normalverteilung zu rechnen

Ein häufig mißverstander Begriff ist "**signifikant**":

so unwahrscheinlich, daß es bei Geltung der Nullhypothese kaumzu erwarten ist; nicht mehr durch Zufall zu erklären; "überzufällig"

15. Notwendiger Stichprobenumfang

Um π (homograd) oder μ (heterograd) mit einem

- absoluten Fehler von e (das ist die halbe Länge des Konfidenzintervalls) und
- einer Sicherheit 1-α (damit ist auch z_α gegeben)

zu schätzen sind

$n \geq \frac{z_\alpha^2 \sigma^2}{e^2}$	bzw.	$n \geq \frac{z_\alpha^2 \pi(1-\pi)}{e^2}$
heterograd		homograd

Personen zu befragen.

Beispiel:
Man will das Durchschnittseinkommen μ mit einer Genauigkeit (ausgedrückt durch den Fehler) von ± 8 DM und einer Sicherheit von 90 % (also 1 - α = 0,9 und $z_\alpha = \pm 1{,}6449$) schätzen. Aus früheren Untersuchungen ist zu vermuten, daß die Standardabweichung der Einkommen 200 DM beträgt. Es sind dazu mindestens $n \geq \frac{(1{,}6449)^2 (200)^2}{8^2} = 1691{,}06$ also mindestens 1691 Personen zu befragen.
Bei einer geringeren Genauigkeit, etwa e = $\pm$ 15 DM (statt 8 DM) würde es schon ausreichen, mindestens 481 Personen zu befragen.

Standardnormalverteilung N(0,1)

Zur Erklärung der Funktionen F(z) und Φ(z) vgl. Vorlesung

	Dichtefunktion	**Verteilungsfunktion**	Symmetrische Intervallwahrscheinl.
z	$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$	$F(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{u^2}{2}}du$	$\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-z}^{z}e^{-\frac{u^2}{2}}du$
0	0.3989	0.5000	0.0000
0.1	0.3970	0.5398	0.0797
0.2	0.3910	0.5793	0.1585
0.3	0.3814	0.6179	0.2358
0.4	0.3683	0.6554	0.3108
0.5	0.3521	0.6915	0.3829
0.6	0.3332	0.7257	0.4515
0.7	0.3122	0.7580	0.5161
0.8	0.2897	0.7881	0.5763
0.9	0.2661	0.8159	0.6319
1.0	0.2420	0.8413	0.6827
1.1	0.2178	0.8649	0.7287
1.2	0.1942	0.8849	0.7699
1.3	0.1714	0.9032	0.8064
1.4	0.1497	0.9192	0.8385
1.5	0.1295	0.9332	0.8664
1.6	0.1109	0.9452	0.8904
1.7	0.0940	0.9594	0.9109
1.8	0.0790	0.9641	0.9281
1.9	0.0656	0.9713	0.9426
2.0	0.0540	0.9772	0.9545
2.1	0.0440	0.9821	0.9643
2.2	0.0355	0.9861	0.9722
2.3	0.0283	0.9893	0.9786
2.4	0.0224	0.9918	0.9836
2.5	0.0175	0.9938	0.9876
2.6	0.0136	0.9953	0.9907
2.7	0.0104	0.9963	0.9931
2.8	0.0079	0.9974	0.9949
2.9	0.0060	0.9981	0.9963
3.0	0.0044	0.9987	0.9973

Wichtige Signifikanzschranken und Wahrscheinlichkeiten

a) z-Werte für gegebene Wkt. 1 - α

P=1-α	**einseitig** F(z)	**zweiseitig** ϕ(z)
90%	1,2816	± 1,6449
95%	1,6449	± 1,9600
99%	2,3263	± 2,5758
99,9%	3,0902	± 3,2910

b) Wkt. für gegebenes z

z	F(z)	ϕ(z)
0	0,5000	0,0000
1	0,8413	0,6827
2	0,9772	0,9545
3	0,9987	0,9973