Univariate Statistik — Prüfung: falls mit dem TR gerechnet, immer dokumentieren!!!

Grundlagen und Begriffe

Skala	Eigenschaft	
Nominal	Codierung	qualitativ
Ordinal	plus Reihenfolge	
Intervall	plus Abstände	quantitativ, metrisch
Verhältnis	plus absoluter Nullpunkt	

Diskret: bestimmte Werte (z.B. Alter in ganzen Jahren => nur Ganzzahlen)

Stetig: alle Werte (auch „unterjähriges“ Alter möglich)

- null (nichts)
0 mehr als „nichts“, aber kleiner als die Hälfte der verwendeten Werte
. Zahlenangabe nicht möglich
… Daten nicht erhältlich oder ohne Bedeutung oder weggelassen

Histogramm

$$\text{Dichte} = \frac{\text{relative Häufigkeit}}{\text{Klassenbreite}}$$

Univariate Statistik – Mittelwerte

Arithmetische Mittel (Durchschnitt)

Additiv/durchschnittlich stetige Rendite
Das arithmetische Mittel ist die Summe der Messwerte x_i dividiert durch die Anzahl der Messwerte.

$\dot{x}$: („x-quer“) arithmetisches Mittel
n: Anzahl der Werte
W_i: absolute Häufigkeit
w_i: relative Häufigkeit

1-Var Stats -> $\dot{x}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Gewichteter Durchschnitt:

Arithmetische Mittel der Werte x_i mit den absoluten Häufigkeiten W_i (bzw. den relativen Häufigkeiten w_i), k = Anzahl verschiedene Merkmalswerte oder Anzahl Klassen.

Klassenmitte: muss selber berechnet werden

$$\sum_{i=1}^{k} x_i \cdot w_i$$

Median

Zentralwert, Teilung der die Grössen in zwei Hälften. Links und rechts des Median liegen 50% der geordneten Werte.

Bei Klassen interpolieren!
TR: LinReg mit L2 (relative Häufigkeiten kumuliert), L1 (Klassenobergrenzen), ONE, YES; dann f(0.5) oder f(50)

Perzentile, Dezile, Quartile oder allgemein Quantile

Der Median ist das 0.5-Quantil, das 2. Quartil (2/4) (bei Streuung), das 5. Dezil (5/10) oder das 50. Perzentil (50/100).

$$\text{Median} = \text{Untergrenze der Medianklasse} + \frac{j}{f} \cdot \text{Medianklassenbreite}$$

mit j = Anzahl (Anteil) Werte in der Medianklasse bis zur Mitte
f = Anzahl (Anteil) Werte in der Medianklasse

Modus

Kommt am Häufigsten vor / Klasse mit der grössten Dichte
Modalklasse: Klasse mit der grössten Dichte

$$\text{Dichte} = \frac{\text{relative Häufigkeit}}{\text{Klassenbreite}}$$

Rechtsschief
Modus < Median < arithmetisches Mittel
Linksschief
Arithmetisches Mittel < Median < Modus
Symmetrisch
Arithmetisches Mittel = Median = Modus

Geometrisches Mittel

Multiplikativ/durchschnittlich effektive Rendite
Bei zeitlich durchschnittlichen Wachstumsraten (zeitlich aufeinanderfolgend)

$\dot{R}$: durchschnittliche Wachstumsrate

Stetige Rendite: r = ln(1+R)
Durchschnittliche effektive Rendite: $\dot{R} = e^r - 1$

$$\bar{R} = \sqrt[n]{\prod_{t=1}^{n}(1+R_t)} - 1$$

Sind absolute Zahlen bekannt, dann gilt:

$$\bar{R} = \sqrt[n]{\frac{\text{Endwert}}{\text{Anfangswert}}} - 1$$

Faktoren in L1
L2 = ln(L1)
1-Var Stats mit L2 (und ggf. Gewichten in L3)
Resultat: $e^{\dot{x}}$
(vgl. stetige Renditen)
FRQ (Frequency = Gewichtung)

Univariate Statistik – Streuungsmasse

Spannweite

Spannweite = Maximum – Minimum

Eliminierung von Ausreissern durch Quartilsabstand

Quartilsabstand = 3. Quartil – 1. Quartil
Maximum – 3. Quartil – Median – 1. Quartil - Minimum

Standardabweichung und Varianz

Abweichung vom Mittelwert der Messwerte berechnen. Dann Durchschnitt dieser Abweichungen bestimmen.

Varianz: durchschnittliche quadratische Abweichung; Quadrate der Abweichungen berechnen
Standardabweichung: Wurzel aus der Varianz

σ^2: (sigma Quadrat) Varianz $\frac{n}{n-1} \cdot \sigma^2$

σ: (sigma) Standardabweichung $\sqrt{\frac{n}{n-1}} \cdot \sigma$

Variantskoeffizient

$$V_\sigma = \frac{\sigma}{\bar{x}} \approx \frac{s}{\bar{x}}$$

s: Streuungsmass

x-Werte in L1 und optional W-Werte (Gewichte) in L2

Varianzschätzer s^2: 1-Var Stats > Sx^2

Standardabweichung: 1-Var Stats > Sx oder σx ablesen
(falls rel. Häufigkeiten in L2: nur σ) (σ ggf. umrechnen in s)

Univariate Statistik – Indexzahlen

Indexzahlen

1 Grösse: einfacher Index
X Grössen: zusammengesetzter Index
t: aktuelle Periode

$p_0^{(i)}$ Preis des Gutes i in der Basisperiode
$p_t^{(i)}$ Preis des Gutes i in der Periode t
$q_0^{(i)}$ Menge des Gutes i in der Basisperiode
$w_0^{(i)} = p_0^{(i)} \cdot q_0^{(i)}$ Wert des Gutes i in der Basisperiode

$$I_t = \frac{\sum_{i=1}^{n} p_t^{(i)} \cdot q_0^{(i)}}{\sum_{i=1}^{n} p_0^{(i)} \cdot q_0^{(i)}} \cdot 100 \quad I_t = \frac{\sum_{i=1}^{n} \text{Preisindex}_i \cdot \text{Wert}_i}{\sum_{i=1}^{n} \text{Wert}_i}$$

Umbasierung

Auf gleichen Basiszeitpunkt umberechnen.

Monat	LIK Dez. 2005 = 100	LIK Dez. 2010 = 100
Aug 10	103.4	*99.2*
Sep 10	103.4	*99.2*
Okt 10	103.9	*99.7*
Nov 10	104.2	*100.0*
Dez 10	**104.2**	**100.0**
Jan 11	*103.8*	99.6
Feb 11	*104.2*	100.0
Mrz 11	*104.9*	100.7
Apr 11	*105.0*	100.8
Mai 11	*105.0*	100.8
Jun 11	*104.7*	100.5
Jul 11	*103.9*	99.7

103.4/1.042=99.2

=> Umrechnungsfaktor

99.7*1.042=103.9

Kursive Zahlen = berechnete Zahlen

Bivariate Statistik - Korrelation

Daten zentrieren: Kovarianz (Vorzeichen)

2 metrische Merkmale X und Y.
Positive Korrelation (Kennzahlen):
- Je grösser X, desto grösser Y. + * + = +
- Je kleiner X, desto kleiner Y. - * - = +

Negative Korrelation (Kennzahlen):
- Je grösser X, desto kleiner Y. + * - = -
- Je kleiner X, desto grösser Y. - * + = -

$$s_{XY} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

r = 1: positiver funktionaler Zusammenhang
r = -1: negativer funktionaler Zusammenhang
r ≈ 0.9: starker positiver Zusammenhang
r ≈ -0.6: mittlerer negativer Zusammenhang
r ≈ 0: kein Zusammenhang / Unabhängigkeit
r = 0: kein linearer Zusammenhang

Daten standardisieren: Korrelationskoeffizient (Interpretation)

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{s_{XY}}{s_X \cdot s_Y}$$

x-Werte in L1, y-Werte in L2:
Korrelationskoeffizient r =
LinReg ax+b „r ablesen“

-1≤r≤1
In der Nähe von 1: starker Zusammenhang
In der Nähe von 0: schwacher Zusammenhang

r (lin. Zusammenhang)	Stärke des funktionalen Zusammenhangs
0	kein nachweisbarer funkt. Zusammenhang (sehr schwach)
Ca. +- 0.25	
	schwach
+-0.5	
	mittel
Ca. +- 0.75	
	stark
+- 1	

Bivariate Statistik - Regression

Methode der kleinsten Quadrate

$\hat{y}$ = ax + b ($\hat{y}$ befindet sich auf der Gerade)
e^2_i: senkrechter Abstand im Quadrat (Nähe) zwischen dem Datenpunkt (x_i / y_i) und dem Punkt auf der Gerade (x_i / $\hat{y}_i$). $e_i = y_i - \hat{y}_i$

Gleichung 1	Gleichung 2
$a \cdot \sum_{i=1}^{n} x_i^2 + b \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$	$a \cdot \sum_{i=1}^{n} x_i + b \cdot n = \sum_{i=1}^{n} y_i$

sys-solv
a = x / b = y (entspricht nicht r!)

Lineares Modell	$y = ax + b$	(TR: *Lin*Reg ax+b)
Polynom-Modell	$y = \sum_{i=0}^{n} a_i x^i$	(Excel)
	$y = ax^2 + bx + c$	(TR: QuadraticReg)
	$y = ax^3 + bx^2 + cx + d$	(TR: CubicReg)
Logarithmisches Modell	$y = a + b \cdot \ln(x)$	(TR: *Ln*Reg a+blnx)
Potenz-Modell	$y = a \cdot x^b$	(TR: PwrReg ax^b)
Exponentielles Modell	$y = a \cdot b^x$	(TR: ExpReg ab^x)
	$y = a \cdot e^{c \cdot x} = a \cdot (e^c)^x$	(Excel)

Für Trendrechnungen schreibt man t anstatt x.

Das Bestimmheitsmass R^2

totale Varianz = erklärebare Varianz + nicht erklärbare Varianz
erklärbare Varianz = R^2
nicht erklärbare Varianz = 1 – R^2

Interpretation:
61.5% der Varianz der Investitionstätigkeit ist durch die Zinsentwicklung erklärbar, 38.5% sind nicht erklärbar. Es gibt noch andere Gründe für die Schwankung der Investitionen.
Je grösser R^2 desto besser passt sich die Regressionskurve der Punktewolke an.
R^2 = 1 => der vermutete funktionale Zusammenhang trifft absolut zu
R^2 = 0 => der vermutete funktionale Zusammenhang trifft absolut nicht zu

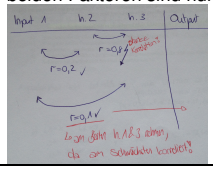
Lineare Mehrfachregression (nur Interpretation)

Die Einflussfaktoren dürfen nicht korreliert sein (in der Praxis nur schwach korreliert).
Input: Werbung und Preis
Output: Umsatz

x-Werte in L1, y-Werte in L2:
y=ax+b LinReg ax+b => a & b ablesen
Bestimmheitsmass r^2 oder R^2 ablesen

$$X_{st} = \frac{X - \bar{x}}{s} \qquad s = \text{Standardabweichung von } X$$

In unserem Beispiel beträgt die Korrelationskoeffizient zwischen Werbeeinsatz und Preis 0.21, d.h. die beiden Faktoren sind nur geringfügig korreliert (Inputfaktoren).



Zeitreihenanalyse (Mischung aus Univariate und Bivariate Statistik)

Deskriptive Zerlegung einer Zeitreihe

Y_t: Zeitreihen
F_t: Trend / Konjunkturkomponente zusammengefasst = Trendkomponente / glatte Komponente
S_t: Saisonkomponente
E_t: Restkomponente (einmalig/zufällig)

Ermittlung der Trendkomponente F_t

Methode der kleinsten Quadrate (Regression)
Für Trendfunktion wird Regression benötigt
F_t = a*x+b x: Nummer des Zeitpunktes
Nummeriert wird selber

Additiv: $Y_t = F_t + S_t + E_t$
Multiplikativ: $Y_t = F_t * S_t * E_t$

Methode des gleitenden Durchschnitts

Benachbarte Werte werden geglättet
- ungerade Ordnung: symmetrisch um Zahl
- gerade Ordnung: unsymmetrisch um Zahl, deshalb 0.5!

F = (1.Z + 2.Z + 3.Z)/3
F = (1.Z*.05 + 2.Z + 3.Z + 4.Z + 5.Z*0.5)/4
Um Trendänderungen sichtbar zu machen!

Ermittlung der Saisonkomponente S_t

S_t: durchschnittliche Abweichung

additiv
L1: Nr.
L2: Anzahl (Y_t)
L3: Trend (F_t)
S_t: durchschnittliche Abweichung

$Y_t = F_t + S_t$

multiplikativ Prognosewerte: f(Nr.)
L1: Nr.
L2: Umsatz
L3: Trendwert
„L1=L2/L3“: Quartalsquotient / durchschn. Quartalsquotient (geom. Mittel)

$Y_t = F_t * S_t$

Saisonbereinigung

additiv (aktueller Wert / vorgegangener Wert)
L1 (L2-L3): Anzahl Saisonbereinigt
L2: Anzahl
L3: durchschn. Abweichung (Saisonkomponente)

$Yber_{,t} = Y_t - \dot{S}_t$
L1 = L2 – L3

multiplikativ
L1: Nr. (1. Zahl löschen)
L2: Umsatz (letzte Zahl löschen)
L3: durchschn. Abweichung

$Yber_{,t} = Y_t / \dot{S}_t$
L1 = L2 / L3
L3 = (L1 / L2) - 1