

Arith. Mittel	\bar{x}	Kovarianz	C_{XY}
Median	\tilde{x}	Korrelation	r_{XY}
Varianz	σ^2	Chi Quadrat	χ^2
SDA	σ	KontingenzK	K
Sample Var.	S^2	Sample SDA	S
Bestimmtheit	R^2	Korrig. K	K^*
Adj. Best.	R_a^2	Erwartungswert	$E(X)$

1 B. Statistik

- Qualitative Merkmale:
 - Variieren nach Beschaffenheit
 - Bspw. Geschlecht
- Quantitative Merkmale:
 - Variieren nach Wert/Zahlen
 - Bspw. Alter, Einkommen
- Diskrete Merkmale:
 - abgestufte Werte
 - Bspw. Einkommensklasse
- Stetige Merkmale:
 - können im Intervall jeden reellen Wert annehmen
 - Bspw. Körpergröße

Skalenniveaus

- Nominal
 - nur Gleichheit oder Andersartigkeit feststellbar (keine Bewertung)
 - stets qualitativ (Religion, Beruf etc.)
- Ordinal
 - natürliche oder festzulegende Rangfolge
 - IQ, Schulnoten
- Kardinal/Metrisch
 - numerischer Art
 - Ausprägung und Unterschied sind messbar
 - verhältnisskaliert (Absoluter Nullpunkt vorhanden; Gewicht, Preis (Doppelt so viel.))
 - intervallskaliert (Kein Nullpunkt, nur Differenzen; Temperatur (10 Grad wärmer als gestern))

Werte

- Arithmetisches Mittel \bar{x}
 - $\bar{x} = \frac{1}{n} \sum_{i=1}^n a_i = \frac{a_1 + a_2 + \dots + a_n}{n}$

- Nur auf kardinale Merkmale
- Summe aller Abweichungen vom Mittel = 0
- Verschiebung um konstanten Wert $a + \bar{x}$
- Multiplikation mit konstantem Wert $a \cdot \bar{x}$
- Auch als gewichtetes arth. Mittel mit den relativen Häufigkeiten
- Median \tilde{x}
 - Mittleres Element der geordneten Liste
 - Bei gerader Anzahl, Durchschnitt der mittleren Elemente
 - Ordinal und Kardinale Merkmale
- Modus
 - Meist auftretendes Element
 - Alle Skalenniveaus
- Quartile (sortieren & ablesen)
 - Unteres Quartil $\tilde{x}_{0,25}$
 - Oberes Quartil $\tilde{x}_{0,75}$
 - GENAU 25% der Werte
- Varianz σ^2
 - Populations Varianz $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$
 - Sample Varianz $S^2_{n-1} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
 - Altn. Formel $\sigma^2 = \overline{x^2} - \bar{x}^2$
 - Eigenschaften:
 - Immer ≥ 0
 - Addition mit a, Varianz unverändert
 - Multiplikation mit b, $Varianz * b^2$
- Standardabweichung σ
 - $\sigma = \sqrt{\sigma^2}$
 - StichprobenSDA $S = \sqrt{S^2_{n-1}}$
- (Inter-) Quartilsabstand $\tilde{x}_{0,75} - \tilde{x}_{0,25}$

Zweidimensionale Häufigkeitstabellen

- Statistische Variablen X und Y mit versch. Ausprägungen
- Spaltensummen sowie Zeilensummen = n
- Relative Häufigkeit $h_{ij} = \frac{n_{ij}}{n}$
- Randverteilung = Betrachtung einer einzigen Variable
- $Z = X + Y$; $\bar{z} = \bar{x} + \bar{y}$;

Kovarianz

- Arithmetisches Mittel des Produkts der Abweichung der einzelnen Beobachtungen

von ihrem Mittel

- $C_{XY} := \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y}) = \overline{xy} - \bar{x} * \bar{y}$
- $C_{XY} > 0$ "große X-Werte zu großen Y-Werten"
- $C_{XY} < 0$ "große Werte zu kleine Werten"
- Sind zwei Variablen statistisch unabhängig ist die Kovarianz = 0

Korrelation

- Normal (Pearson) $r_{XY} = \frac{C_{XY}}{\sigma_x * \sigma_y}$
 - normiertes Maß für Strenge des linearen statistischen Zusammenhangs
 - r_{XY} hat das gleiche Vorzeichen wie C_{XY}
 - Bleibt unverändert bei linearer Transformation
 - $r_{XY} = r_{YX}$
 - $-1 \leq r_{XY} \leq +1$
- Rangkorrelation (Spearman) $r_{XY}^{Sp} = \frac{r_{rg(X),rg(Y)}}{r_{rg(X),rg(Y)}}$
 - für ordinale Variablen
 - misst linearen & monotonen Zusammenhangs
 - Ist unempfindlich gegenüber Ausreißern
 - Ränge müssen vorher berechnet werden
 - Berechnung mit TR über Ränge
 - $-1 \leq r_{XY}^{Sp} \leq +1$
- Kovarianz und Korrelation bedeuten nicht zwangsweise eine kausale Beziehung!

Kontingenzkoeffizient

- beschreibt die Stärke des Zusammenhangs zweier Merkmale, nicht deren Richtung
- Nur für nominale und ordinale Merkmale
- Chi-Quadrat $QK = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ji} - E_{ij})^2}{E_{ij}}$
 - $E_{ij} = \frac{1}{n} * n_i * n_j = \frac{1}{n} n(x_i) * n(y_j)$
 - Siehe Erweiterte Kontingenztafel
 - X und Y unabhängig: $QK = 0$
 - Sonst $QK > 0$
 - Für 2x2 Matrix: $QK = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$
 - a bis d sind Inhalte der Tabelle, Summen sind Randhäufigkeiten
- Kontingenzkoeffizient $K := \sqrt{\frac{QK}{QK+n}}$
 - normiertes Maß
 - X und Y unabhängig: $K = 0$
 - $0 \leq K \leq K_{max} = \sqrt{\frac{m-1}{m}} < 1$

- m = Minimum von Zeilenzahl und Spaltenzahl
- Korrigierter K.-koeffizient $K^* := \frac{K}{K_{max}} = \frac{\sqrt{\frac{QK * m}{(QK+n)(m-1)}}}{\sqrt{\frac{QK * m}{(QK+n)(m-1)}}}$
- $0 \leq K^* \leq 1$
 - Vergleichbar mit anderen K-Tabellen

2 Regression

- Lineare Regression $y(x) = a + bx$
 - $b = \frac{C_{XY}}{S_x^2}$ und $a = \bar{y} - b\bar{x}$
 - Interpret: b*x erhöht pro Einheit und a: Achsenabschnitt
 - Extrapolation (Punkte außerhalb der orig. Daten) nicht aussagekräftig
 - Regressionswerte $\hat{y}_i = y(x_i)$
 - Residuen (Fehler) $e_i = y_i - \hat{y}_i$
- Andere Regressionen:
 - $\hat{y} = a + bx + cx^2$ Quadr. Regr.
 - $\hat{y} = a + x^b$ Potenzfunkt.
 - $\hat{y} = ab^x$ Expo-funkt.
- Meth. kleinste Quadrate
- Varianzzerlegung $SSQ_{Total} = SSQ_{Reg} + SSQ_{Resi}$
 - $SSQ_{Reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ (Abweichung von Vorhersage und Mittelwert)
 - $SSQ_{Total} = \sum_{i=1}^n (y_i - \bar{y})^2$ (Gesamtabweichung)
 - $SSQ_{Resi} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ (Abweichung von Vorhersage und y)
- Bestimmtheitsmaß $R^2 = \frac{SSQ_{Reg}}{SSQ_{Total}} = \frac{S_y^2}{S_y^2} = r^2$
 - r^2 gilt nicht für Quadr. Reg. !!!
 - Schlecht $0 \leq R^2 \leq 1$ Gut
 - $R^2 \geq 0.8$ akzeptabel
- Multiple Regr.
 - Y wird durch mehrere Variablen erklärt
 - $\hat{y} = a + b_1 x_3 + b_2 x_3 + b_3 x_3$
- Adjustiertes Bestimmtheitsmaß $R_a^2 = R^2 - \frac{k}{n-k-1} * (1 - R^2)$
 - Hinzunahme von Params, erhöht den R^2 automatisch, auch wenn es nicht besser wird
 - $n = \text{Anzahl der Messwerte}$
 - $k = \text{Anzahl der Reg. Params}$

- R_a^2 kann auch kleiner/negativ werden –> Variable nicht aufnehmen
- Anmerkungen:
 - Residualplot: Gutes Modell, wenn kein Muster erkennbar!
 - Optimum finden: 1. Ableitung = 0 setzen
 - "Faktor Größe" hat nichts mit Einfluss zu tun, nur bei standardisierten Daten

3 Wahrsch. Rech.

- Zufallsvariable $X: \Omega \rightarrow R$ mit $X(\omega) = x$
 - Funktion, die jedem Möglichen Ergebnis eine reelle Zahl zuordnet
 - Wahrscheinlichkeits-/ Dichtefunktion $f: P(X = x)$
 - Verteilungsfunktion $F: P(X \leq t)$
 - F ist Stammfunktion für f aber muss mit $+C$ angepasst werden
- Diskrete
 - $f: R \rightarrow [0, 1]$ mit $f(x) = P(X = x)$
 - $P(X = X)$ Wahrscheinlichkeit mit der X die Realisation x annimmt
 - $F(t) = P(X \leq t) = \sum_{x_i \leq t} P(X = x_i)$
- Stetige
 - Zufallsvariable ist stetig, wenn Wahrscheinlichkeit durch Dichtefunktion abbilden lässt
 - Dichtefunktion, wenn $\int_{-\infty}^{+\infty} f(x) dx = 1$ und $f(x) \geq 0$
 - $F(t) = P(X \leq t) = \int_{-\infty}^t f(x) dx$
- Erwartungswert
 - Diskret: $E(X) = \sum_{i=1}^n x_i * f(x_i)$
 - Stetig: $E(X) = \int_{x_{min}}^{x_{max}} x * f(x) dx$
- Varianz ($Var(X) = \sigma^2$) & SDA ($\sigma = \sqrt{\sigma^2}$)
 - Es gilt: $\sigma^2 = E((X - E(X))^2) = E(X^2) - (E(X))^2$
 - Diskret: $Var(X) = \sum_{i=1}^n (x_i - E(X))^2 * f(x_i)$
 - Stetig: $Var(X) = \int_{x_{min}}^{x_{max}} (x - E(X))^2 * f(x) dx$
- Rechenregeln
 - $E(a + b * X) = a + b * E(X)$
 - $Var(a + b * X) = b^2 * Var(X)$

- $E(X + Y) = E(X) + E(Y)$
- Stichprobe:
 - Stichprobenmittel von unabhängigen Variablen $\bar{X} := \frac{1}{n}(X_1 + \dots + X_n)$
 - $E(\bar{X}) = \mu$
 - SDA von \bar{X} $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
 - Wahrscheinlichkeit, dass etwas größer/kleiner bei Stichprobe n ist: Einfach Gaußtest.
- Normalverteilung
 - SD-normalverteilung mit $\mu = 0$ und $\sigma = 1$
 - z-Transformation $z = \frac{x - \mu}{\sigma}$
- Zentr.Grenz.Satz: Für hinreichend großes n jeder Verteilung gilt $\bar{X}_n \tilde{N}(\mu, \frac{\sigma^2}{n})$ "normalverteilt"

4 Schl. Statistik

Anmerkungen

- α meist 5% oder 1%
- "Mindestens" meint meist beidseitigen Test
- "Maximal" meint meist rechtsseitigen Test

Mittelwerttest

- GG ist norm. verteilt oder $n > 30$
- Stichprobenmittel \bar{x} und ggf. Stichprobenvarianz s^2 bekannt
- σ der GG bekannt
 - $z = \sqrt{n} \frac{\bar{x} - \mu_0}{\sigma}$
 - -> Tabelle Norm. Verteilung
- σ der GG unbekannt
 - $t = \sqrt{n-1} \frac{\bar{x} - \mu_0}{s_n}$
 - > t Tabelle!
- Gleiches gilt für t-1

Seite:	H_0 behalten	H_0 verwerfen
Beide	$ z \leq z[1 - \alpha/2]$	$ z > z[1 - \alpha/2]$
Rechts	$z \leq z[1 - \alpha]$	$z > z[1 - \alpha]$
Links	$z \geq z[\alpha]$	$z < z[\alpha]$

Varianztest

- GG ist normalverteilt, α und σ_0 bekannt
- μ von GG. bekannt

- In der χ^2 Tabelle nachschlagen!
- – $t_n = \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2$
- Siehe (II)
- μ von GG. unbekannt
 - $t_n = n * \frac{s_n^2}{\sigma_0^2}$
 - Siehe (III)

II

H_0	H_1	Krit.
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$t_n < \chi_n^2[\alpha/2]$ $t_n > \chi_n^2[1 - \alpha/2]$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$t_n < \chi_n^2[\alpha]$
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$t_n > \chi_n^2[1 - \alpha]$

III

H_0	H_1	Krit.
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$t_n < \chi_{n-1}^2[\alpha/2]$ $t_n > \chi_{n-1}^2[1 - \alpha/2]$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$t_n < \chi_{n-1}^2[\alpha]$
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$t_n > \chi_{n-1}^2[1 - \alpha]$

Differenztest

- GG ist normalverteilt
- σ_X^2 und σ_Y^2 gleich aber unbekannt
- δ_0 vorgegeben oder $\delta = \mu_X - \mu_Y$

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{1}{n} + \frac{1}{m}} * \sqrt{\frac{n * s_n^2 + m * s_m^2}{n + m - 2}}}$$

H_0	H_1	Krit.
$\delta = \delta_0$	$\delta \neq \delta_0$	$ t_n > t_{n+m-2}[1 - \alpha/2]$
$\delta \geq \delta_0$	$\delta < \delta_0$	$t_n < t_{n+m-2}[\alpha]$
$\delta \leq \delta_0$	$\delta > \delta_0$	$t_n > t_{n+m-2}[1 - \alpha]$

χ^2 Test

- E_{ij} immer ≥ 5
- $H_0 = X, Y$ sind unabhängig; $H_1 = X, Y$ sind abhängig
- Prüfgröße χ^2 (wie oben, mit erw. Kont.-Tabelle)
- Krit.Wert: $c = \chi_{(k-1)(l-1)}^2[1 - \alpha]$
- $\chi^2 \leq c$ H_0 behalten
- $\chi^2 > c$ H_0 verwerfen

P-Test

Excel Tests

- Koeffizienten für jede X_i –> Formel lässt sich daraus ableiten
- t-Statistik:
 - Test, ob x überhaupt y beeinflusst
 - Parameter wird nur im Modell behalten wenn $|t| := |\frac{\beta_j}{\sigma_j}| > 2$
 - Signifikanzniveau von ca. 5%
 - Alternativ: p-Werte $< \alpha$ werden behalten, p-Werte $> \alpha$ werden verworfen
- F-Test des Bestimmtheitsmaßes:
 - Testet ob, nicht auch alle Parameter = 0 sein könnten (Sinnhaftigkeit der Regression)
 - H_0 :
 - Prüfgröße F aus Excel
 - FWert: aus F-Verteilung oder gegeben
 - $F \geq FWert$ H_0 verwerfen, Regressionsansatz sinnvoll
 - $F < FWert$ H_0 behalten, Regressionsansatz schlecht
 - Einfacher: Über FKrit
 - $pWert < FKrit$ H_0 behalten, Regressionsansatz sinnvoll
 - $pWert > FKrit$ H_0 verwerfen, Regressionsansatz schlecht

Other

Integrationsregel - Fläche unter Funktion
 $\int_a^b x^n = [\frac{1}{n+1} x^{n+1}]_a^b = [\frac{1}{n+1} b^{n+1}] - [\frac{1}{n+1} a^{n+1}]$

$$\frac{n_{ij}}{n_{ij} - E_{ij}} \mid \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Test\Realität	H_0 richtig	H_1 richtig
H_0 behalten	ok (Spezifität)	β Fehler (FP)
H_0 verwerfen	α Fehler (FN)	ok (Sensitivität)