Formelsammlung Statistik

Lukas Warode

Maße der zentralen Tendenz

Modus

- Nominales Skalenniveau
- Häufigster Wert

$$x_{mod}$$

Median

- Ordinales Skalenniveau
- Mittlere Ausprägung bei Anordnung der Variable

Ungerade Anzahl an Fällen (n):

$$\tilde{x} = x_{(\frac{n+1}{2})}$$

```
##  [1]  1  7 10 18 20 24 25 26 36 37 40
```

```
## [1] 24
```

Gerade Anzahl an Fällen (n):

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

```
##  [1]  1  7 10 18 20 24 25 26 36 37
```

```
## [1] 22
```

Arithmetisches Mittel

- Metrisches Skalenniveau
- Summe aller Fälle durch Anzahl der Fälle teilen

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=i}^{n} x_i$$

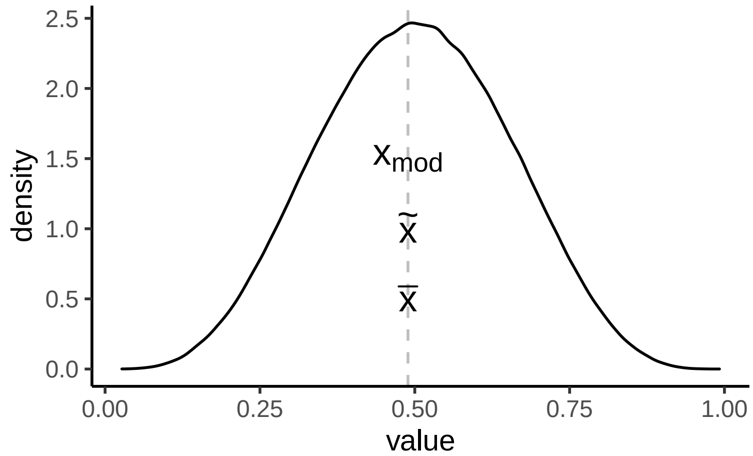
```
##  [1] 37  1 25 10 36 18 24  7 20 26
```

```
## [1] 20.4
```

Verteilungsformen

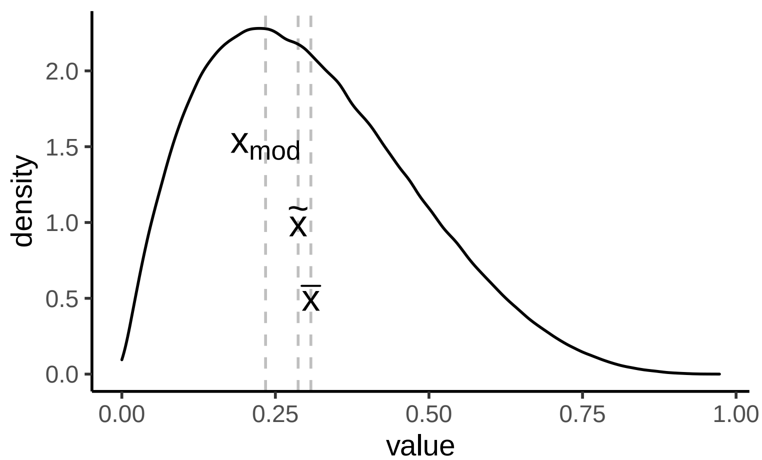
Symmetrisch (Normalverteilung)

$$x_{mod} = \tilde{x} = \bar{x}$$



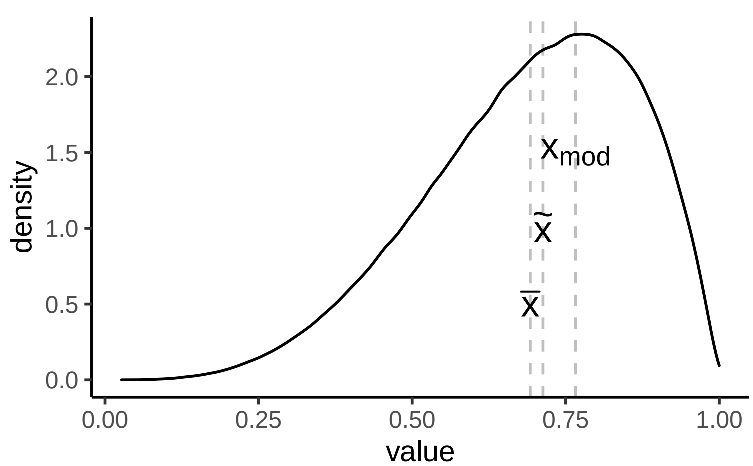
Linkssteil / Rechtsschief

$$x_{mod} < \tilde{x} < \bar{x}$$



Rechtssteil / Linksschief

$$x_{mod} > \tilde{x} > \bar{x}$$



Streuungsmaße

Spannweite

- Ordinales Skalenniveau
- Differenz zwischen größter und kleinster Ausprägung

$$R = x_{max} - x_{min}$$

Interquartilsabstand (IQR)

- Ordinales Skalenniveau
- Intervall der mittleren 50% der Stichprobe

$$IQR = Q_{0.75} - Q_{0.25}$$

Variation (Summe der Abweichungsquadrate)

- Metrisches Skalenniveau
- Englisch: *Sum of squares / sum of squared deviations*

$$SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Varianz

- Metrisches Skalenniveau
- Standardisierte Variation

$$Var(x) = s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standardabweichung

- Metrisches Skalenniveau
- Quadratwurzel der Varianz
- Durchschnittliche Abweichung von Werten zum Arithmetischen Mittel

$$\sigma_{x_{Population}} = \sqrt{\sigma_x^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$s_{x_{Stichprobe}} = \sqrt{s_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Variationskoeffizient (Abweichungskoeffizient)

- Metrisches Skalenniveau
- Relatives Streuungsmaß, d.h. nicht abhängig von der Maßeinheit der Variable
- Standardbweichung in Relation zum Arithmetischen Mittel

$$V_x = \frac{s_x}{\bar{x}}$$

Standardfehler

- (Durchschnittliche) Abweichung von Stichprobenkennwerten zu Populationskennwerten, z.B. vom Mittelwert

Standardfehler des Arithmetischen Mittelwertes (*Standard error of the mean*)

- Symbol des Populationsmittelwertes: μ
- Symbol des Stichprobenmittelwertes $\bar{x}$

$$\sigma_{\bar{x}_{Population}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

$$s_{\bar{x}_{Stichprobe}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$$

$$\hat{\sigma}_{\bar{x}_{Population,\ geschätzt}} = \frac{s}{\sqrt{n-1}} = \sqrt{\frac{s^2}{n-1}}$$

Standardfehler des Anteilswertes

- Symbol des Populationsanteilswertes: π
- Symbol des Stichprobenanteilswertes: p_x

$$\sigma(p_x)_{Population} = \sqrt{\frac{\pi_x \cdot (1-\pi_x)}{n}}$$

Schätzung der Populationsvarianz: $\pi_x \cdot (1-\pi_x)$ aus der Stichprobenvarianz: $p_x \cdot (1-p_x)$

$$\hat{\sigma}(p_x)_{Population} = \sqrt{\frac{p_x \cdot (1-p_x)}{n}}$$

Konfidenzintervall

- Generalisierbarkeit von Parametern aus der Stichprobe (auf die Population)
- Geschätzter Intervallbereich, in dem Parameter der Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit liegen
- Z-Standardisierung: $z = \frac{x-\mu}{\sigma}$

Konfidenzintervall des Populationsmittelwertes (μ_x)

- Kleine Stichproben: t-Verteilung
- Große Stichproben: Standardnormalverteilung

Bestimmung der Intervallgrenzen:

$$\bar{x} - \frac{s_x}{\sqrt{n-1}} \cdot z_{(1-\frac{\alpha}{2})} < \mu_x < \bar{x} + \frac{s_x}{\sqrt{n-1}} \cdot z_{(1-\frac{\alpha}{2})}$$

Konfidenzintervall des Populationsanteilswertes (π_x)

- Standardnormalverteilung (wenn Stichprobe ausreichend groß)

Bestimmung der Intervallgrenzen:

$$p_x - \sqrt{\frac{p_x \cdot (1-p_x)}{n}} \cdot z_{(1-\frac{\alpha}{2})} < \pi_x < p_x + \sqrt{\frac{p_x \cdot (1-p_x)}{n}} \cdot z_{(1-\frac{\alpha}{2})}$$

t-Test

t-Test: Mittelwert

t-Test für einen Mittelwert

- H_0: $\mu_1 = \mu$
- H_A: $\mu_1 \neq \mu$

Teststatistik:

$$Z = \frac{\bar{x} - \mu}{\frac{s_x}{\sqrt{n-1}}}$$

- Kritische Testwerte bei z_α und $z_{(1-\alpha)}$

t-Test für 2 Mittelwerte

- H_0: $\mu_1 - \mu_2 = 0$
- H_A: $\mu_1 - \mu_2 \neq 0$

Teststatistik:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_{x_1}^2}{n_1 - 1} - \frac{s_{x_2}^2}{n_2 - 1}}}$$

- Kritische Testwerte bei $z_{\frac{\alpha}{2}}$ und $z_{(1-\frac{\alpha}{2})}$

t-Test: Populationsanteil

t-Test für einen Populationsanteil

- H_0: $\pi_1 = \pi$
- H_A: $\pi_1 \neq \pi$

Teststatistik:

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}}$$

- Kritische Testwerte bei z_α und $z_{(1-\alpha)}$

t-Test für 2 Populationsanteile

- H_0: $\pi_1 - \pi_2 = 0$
- H_A: $\pi_1 - \pi_2 \neq 0$

Teststatistik:

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_2 \cdot (1-p_2)}{n_2}}}$$

- Kritische Testwerte bei $z_{\frac{\alpha}{2}}$ und $z_{(1-\frac{\alpha}{2})}$

Chi-Quadrat-Unabhängigkeitstest (χ^2)

- Bivariater Test auf stochastische Unabhängigkeit
  - H_0: Beide Zufallsvariablen sind stochastisch *unabhängig* voneinander
  - H_A: Beide Zufallsvariablen sind stochastisch *nicht unabhängig* voneinander

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- i: *"Zeilen"*
- j: *"Spalten"*
- n_{ij}: Beobachtete Häufigkeiten
- e_{ij}: Erwartete Häufigkeiten
  - $e_{ij} = \frac{n_i \cdot n_j}{n}$
    - n_i: Zeilenhäufigkeit
    - n_j: Spaltenhäufigkeit

Berechnung der Freiheitsgrade (*degrees of freedom*): $df = (I-1) \cdot (J-1)$

Zusammenhangsmaße auf Basis von χ^2

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

$$Cramer‘s\ V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}}$$

$$Kontingenzkoeffizient\ C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

$$C_{korrigiert} = \frac{C}{\sqrt{\frac{k-1}{k}}}$$

- $k = Kleinste\ Zeilenzahl\ oder\ Spaltenzahl$

F-Test – Einfaktorielle Varianzanalyse

- F-Test testet den Anteil erklärter Varianz an unerklärter Varianz zwischen mehreren Gruppen

- x_{ij}: Beobachtung i in der Gruppe j

- $\bar{x}_j$: Mittelwert der Gruppe j
- Varianz zwischen den Gruppen: $\sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2$
  - $df_1 : j - 1$
- Varianz innerhalb der Gruppen: $\sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$
  - $df_2 : n - j$

Teststatistik:

$$F = \frac{erklärte\ Varianz}{unerklärte\ Varianz} = \frac{\frac{Varianz\ zwischen\ den\ Gruppen}{df_1}}{\frac{Varianz\ innerhalb\ der\ Gruppen}{df_2}} = \frac{\frac{\sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2}{j-1}}{\frac{\sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n-j}}$$

Metrische Zusammenhangsmaße

Kovarianz

- Kovarianz: Gemeinsame Varianz von 2 Variablen
  - $[-\infty,\ \infty]$

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Korrelation (Pearson-Korrelation: r)

- Korrelationskoeffizient: Standardisierte Kovarianz
  - $[-1,\ 1]$

$$r_{x,y} = Corr(x, y) = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Lineare Regression

Einfache lineare Regression

Lineare Modellfunktion: $y = \alpha + \beta x$

- y: Zielgröße
- α: Konstante (y-Achsenabschnitt)
- β: Steigung

Lineare Regressionsgleichung: $y_i = \alpha + \beta x_i + \varepsilon_i$

- Ziel: Finden der geschätzten Werte $\hat{\alpha}$ und $\hat{\beta}$ für die Parameter α und β, die die Regressionsgleichung am besten nachbilden können (den besten *fit* ermöglichen)
- x_i: Beobachtung i der unabhängigen Variable x

- ε_i: Fehlerterm (Residuum): Abweichung der empirischen Beobachtung y_i und dem geschätzten Wert der Regressionsgleichung $\hat{y}_i$
- "Geschätzte" Abweichung der Regressionsgerade von der empirischen Beobachtung: $\hat{\varepsilon_i} = y_i - \hat{y}_i = y_i - \alpha - \beta_{x_i}$

Ziel der geschätzten Regressiongleichung (Regressionsgerade): Minimierung der Summe der residualen Quadrate $\hat{\varepsilon_i}^2$ - Minimierung von: $\sum_{i=1}^n \hat{\varepsilon_i}^2 = \sum_{i=1}^n (y_i - \alpha - \beta_{x_i})^2$

Schätzung des Regressionskoeffizienten β: $\hat{\beta}$

$$\hat{\beta} = \frac{Cov(x,y)}{Var(x)} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= r_{x,y} \cdot \frac{s_y}{s_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} \cdot \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y}_i)^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_i)^2}}$$

Schätzung der Regressionskonstanten α: $\hat{\alpha}$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Determinationskoeffizient R^2

- "Erklärungsleistung" eines Regressionsmodells: Anteil erklärter Varianz (an der abhängigen Variablen) an Gesamtvarianz
- In einfacher (bivariater) linearer Regression: Quadrierter Pearson-Korrelationskoeffizient r

$$R^2 = \frac{erklärte\ Varianz}{Gesamtvarianz} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$$

$$R^2 = 1 - \frac{unerklärte\ Varianz}{Gesamtvarianz} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$$

Adjustiertes (auch: korrigiertes) R^2 (Korrigierter Determinationskoeffizient)

- "Korrektur" des R^2 für Anwendung innerhalb von Stichproben
- Abhängig von Fallzahl n und Anzahl unabhängiger Variablen p

$$R^2_{adj} = R^2 - \frac{p \cdot (1 - R^2)}{n - p - 1} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

Multiple lineare Regression

Multiple lineare Regressionsgleichung: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$

- Schätzung der linearen Regression anhand k unabhängiger Variablen für $i = 1, \ \ldots, \ n$

Matrixschreibweise der Multiplen linearen Regressionsgleichung:

$$
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{pmatrix}
$$

$$
\mathbf{y} = \mathbf{X}\beta + \varepsilon
$$

- Schätzung von β anhand Kleinste-Quadrate-Schätzung (OLS):

$$
\hat{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = (\mathbf{X'X})^{\mathbf{-1}}\mathbf{X'y}
$$

- $\mathbf{X'}$: Transponierte Matrix von $\mathbf{X}$
- $(\mathbf{X'X})^{\mathbf{-1}}$: Inverse Matrix von $\mathbf{X'X}$