

Tests im Binomialmodell

Martin Kolb

Universität Paderborn

14. Januar 2022



Tests im Binomialmodell

Es sei X_1, \dots, X_n eine mathematische Stichprobe einer zum Parameter p Bernoulliverteilten Grundgesamtheit X vom Umfang n . Wir veranschaulichen mögliche Konstruktionserfahren statistischer Hypothesentests für den Parameter $p \in (0, 1)$.



Tests im Binomialmodell

Es sei X_1, \dots, X_n eine mathematische Stichprobe einer zum Parameter p Bernoulliverteilten Grundgesamtheit X vom Umfang n . Wir veranschaulichen mögliche Konstruktionserfahren statistischer Hypothesentests für den Parameter $p \in (0, 1)$.

Erinnerung:

Die Anzahl der Erfolge

$$Y = \sum_{i=1}^n X_i$$

ist binomialverteilt mit Parametern n und p , d.h. insbesondere gilt für $k = 0, \dots, n$

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

sowie

$$\mathbb{E}[Y] = n \cdot p \quad \text{und} \quad \text{Var}(Y) = n \cdot p \cdot (1 - p).$$

Tests im Binomialmodell

Vorgehen:

Das Niveau $\alpha \in (0, 1)$ sei nun fixiert und im folgenden interessieren wir uns für die Hypothesen

$$H_0 : p \leq p_0 \in (0, 1) \quad \text{gegen} \quad H_1 : p > p_0$$



Tests im Binomialmodell

Vorgehen:

Das Niveau $\alpha \in (0, 1)$ sei nun fixiert und im folgenden interessieren wir uns für die Hypothesen

$$H_0 : p \leq p_0 \in (0, 1) \quad \text{gegen} \quad H_1 : p > p_0$$

Unseren Verwerfungsbereich wollen wir in der Form

$$R := \{Y \geq k\}$$

für geeignetes k wählen. Die Intuition dahinter ist natürlich, dass zu viele Erfolge nicht gut mit der Nullhypothese verträglich sind.

Tests im Binomialmodell

Wir müssen sicherstellen, dass die Wahrscheinlichkeit für einen Fehler erster Art nicht α überschreitet, d.h. dass

$$\forall p \leq p_0 : \beta(p) = \mathbb{P}_p(Y \geq k) \leq \alpha$$

gilt.

Tests im Binomialmodell

Wir müssen sicherstellen, dass die Wahrscheinlichkeit für einen Fehler erster Art nicht α überschreitet, d.h. dass

$$\forall p \leq p_0 : \beta(p) = \mathbb{P}_p(Y \geq k) \leq \alpha$$

gilt. Wir haben gesehen, dass die Abbildung

$$(0, 1) \ni p \mapsto \mathbb{P}_p(Y \geq k)$$

monoton wachsend ist. Somit genügt es sicherzustellen, dass

$$\mathbb{P}_{p_0}(Y \geq k) \leq \alpha$$

erfüllt ist.



Tests im Binomialmodell

Zur Festlegung des Verwerfungsbereichs R definieren wir also

$$k = k(n, \alpha) = \min \left\{ l = 0, 1, \dots, n \mid \sum_{j=l}^n \binom{n}{j} p_0^l (1 - p_0)^{n-l} \leq \alpha \right\}.$$

Sollte

$$\left\{ l = 0, 1, \dots, n \mid \sum_{j=l}^n \binom{n}{j} p_0^l (1 - p_0)^{n-l} \leq \alpha \right\} = \emptyset$$

gelten, dann setze formal $k = \infty$. In diesem letzten Fall ist der Verwerfungsbereich also die leere Menge und wir verwerfen H_0 nie.



Tests im Binomialmodell

Bemerkung:

Bei den Hypothesen

$$H_0 : p = p_0 \in (0, 1) \quad \text{gegen} \quad H_1 : p \neq p_0$$

sprechen sowohl zu viele als auch zu wenige Erfolge gegen H_0 .

Tests im Binomialmodell

Bemerkung:

Bei den Hypothesen

$$H_0 : p = p_0 \in (0, 1) \quad \text{gegen} \quad H_1 : p \neq p_0$$

sprechen sowohl zu viele als auch zu wenige Erfolge gegen H_0 .
Man wird also als Ablehnungsbereich R einen Ansatz

$$R = \{Y \leq k\} \cup \{Y \geq l\}$$

als sinnvoll erachten können. Die Grenzen l und k müssen dann in Analogie zu obigem Vorgehen ermittelt werden.

Rezept 3: Binomialtest

Annahmen:

Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und Bernoulli mit unbekannten Parameter p . Folglich gilt

$$Y := \sum_{i=1}^n X_i \sim \text{Bin}(n, p).$$



Rezept 3: Binomialtest

Annahmen:

Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und Bernoulli mit unbekannten Parameter p . Folglich gilt

$$Y := \sum_{i=1}^n X_i \sim \text{Bin}(n, p).$$

Hypothesen:

- a) $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$
- b) $H_0 : 0 \leq p \leq p_0$ gegen $H_1 : 1 \geq p > p_0$
- c) $H_0 : 0 \leq p \leq p_0$ gegen $H_1 : 1 \leq p > p_0$



Rezept 3: Binomialtest

Annahmen:

Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und Bernoulli mit unbekannten Parameter p . Folglich gilt

$$Y := \sum_{i=1}^n X_i \sim \text{Bin}(n, p).$$

Hypothesen:

- a) $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$
- b) $H_0 : 0 \leq p \leq p_0$ gegen $H_1 : 1 \geq p > p_0$
- c) $H_0 : 0 \leq p \leq p_0$ gegen $H_1 : 1 \leq p > p_0$

Teststatistik:

$$Y := \sum_{i=1}^n X_i,$$



Rezept 3: Binomialtest

Annahmen:

Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und Bernoulli mit unbekanntem Parameter p . Folglich gilt

$$Y := \sum_{i=1}^n X_i \sim \text{Bin}(n, p).$$

Hypothesen:

- a) $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$
- b) $H_0 : 0 \leq p \leq p_0$ gegen $H_1 : 1 \geq p > p_0$
- c) $H_0 : 0 \leq p \leq p_0$ gegen $H_1 : 1 \leq p > p_0$

Teststatistik:

$$Y := \sum_{i=1}^n X_i,$$

Ablehnungskriterium für H_0 bei Niveau α :

- a) $\{0, \dots, k, l, \dots, n\}$ mit $\mathbb{P}_{p_0}(X \leq k), \mathbb{P}_{p_0}(X \geq l) \leq \alpha/2$
- b) $\{l, \dots, n\}$ mit $\mathbb{P}_{p_0}(X \geq l) \leq \alpha$
- c) $\{0, \dots, k\}$ mit $\mathbb{P}_{p_0}(X \leq k), \mathbb{P}_{p_0}(X \geq l) \leq \alpha$.



Abschließend halten wir noch folgende Beobachtung fest. Unter den Annahmen dieses Abschnittes zur Tests im Binomialmodell gelten die folgenden Aussagen:



Abschließend halten wir noch folgende Beobachtung fest. Unter den Annahmen dieses Abschnittes zur Tests im Binomialmodell gelten die folgenden Aussagen:

- Ist der Umfang n sehr groß und $p = p_0$, dann ist die Zufallsvariable

$$Z := \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}}$$

als Folgerung aus dem Zentralen Grenzwertsatz approximativ standard normalverteilt.



Abschließend halten wir noch folgende Beobachtung fest. Unter den Annahmen dieses Abschnittes zur Tests im Binomialmodell gelten die folgenden Aussagen:

- Ist der Umfang n sehr groß und $p = p_0$, dann ist die Zufallsvariable

$$Z := \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}}$$

als Folgerung aus dem Zentralen Grenzwertsatz approximativ standard normalverteilt.

- Ist der Umfang n sehr groß und $p \neq p_0$, dann ist die Zufallsvariable

$$Z := \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}}$$

als Folgerung aus dem Zentralen Grenzwertsatz approximativ normalverteilt mit Erwartungswert

$$\sqrt{n} \frac{p - p_0}{\sqrt{p_0(1 - p_0)}} \text{ und Varianz } \frac{p(1 - p)}{p_0(1 - p_0)}$$

Rezept: Approximativer Binomialtest

Annahmen:

Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und Bernoulli mit unbekannten Parameter p . Folglich gilt

$$Y := \sum_{i=1}^n X_i \sim \text{Bin}(n, p).$$



Rezept: Approximativer Binomialtest

Annahmen:

Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und Bernoulli mit unbekannten Parameter p . Folglich gilt

$$Y := \sum_{i=1}^n X_i \sim \text{Bin}(n, p).$$

Hypothesen:

- a) $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$
- b) $H_0 : p \geq p_0$ gegen $H_1 : p < p_0$
- c) $H_0 : p \leq p_0$ gegen $H_1 : p > p_0$



Rezept: Approximativer Binomialtest

Annahmen:

Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und Bernoulli mit unbekannten Parameter p . Folglich gilt

$$Y := \sum_{i=1}^n X_i \sim \text{Bin}(n, p).$$

Hypothesen:

- a) $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$
- b) $H_0 : p \geq p_0$ gegen $H_1 : p < p_0$
- c) $H_0 : p \leq p_0$ gegen $H_1 : p > p_0$

Teststatistik:

$$Z := \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}},$$



Rezept: Approximativer Binomialtest

Annahmen:

Die Zufallsvariablen X_1, \dots, X_n seien unabhängig und Bernoulli mit unbekanntem Parameter p . Folglich gilt

$$Y := \sum_{i=1}^n X_i \sim \text{Bin}(n, p).$$

Hypothesen:

- a) $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$
- b) $H_0 : p \geq p_0$ gegen $H_1 : p < p_0$
- c) $H_0 : p \leq p_0$ gegen $H_1 : p > p_0$

Teststatistik:

$$Z := \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}},$$

Ablehnungskriterium für H_0 bei Niveau α :

- a) $|Z| > z_{1-\frac{\alpha}{2}}$
- b) $Z < z_\alpha$
- c) $Z > z_{1-\alpha}$



Explorative Datenanalyse etc.

Visualisierung der Daten und Prüfen von Verteilungsannahmen sind wesentlich:

Ausblick:

Betrachte den Datensatz:

sites.google.com/site/chiharahesterberg/data2/NCBirths2004.csv



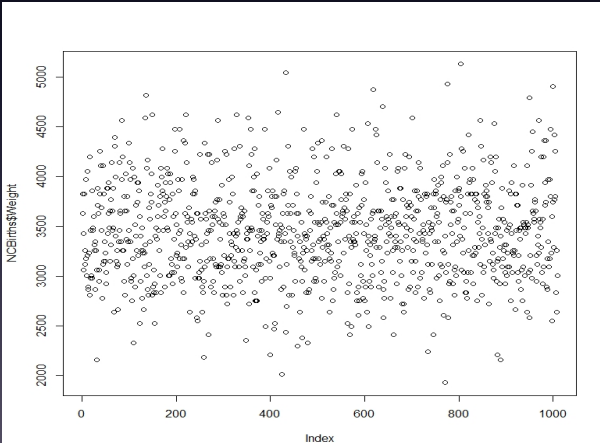
Explorative Datenanalyse etc.

Visualisierung der Daten und Prüfen von Verteilungsannahmen sind wesentlich:

Ausblick:

Betrachte den Datensatz:

sites.google.com/site/chiharahesterberg/data2/NCBirths2004.csv



Explorative Datenanalyse etc.

Es gelten:

Stichprobenmittel: 3448.26 Stichprobenvarianz: 237886.4

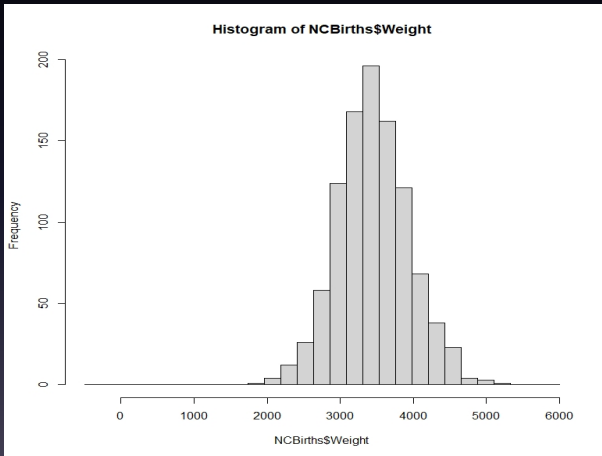


Abbildung: Histogramm: Geburtsgewichte

Explorative Datenanalyse etc.

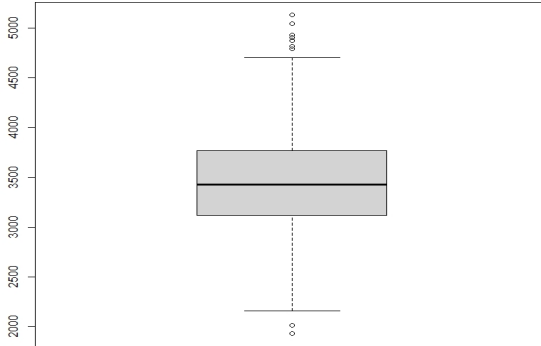


Abbildung: Boxplot: Geburtsgewichte



Explorative Datenanalyse etc.

Die Box entspricht dem Bereich, in dem die mittleren 50 Prozent der Daten liegen. Die Länge der Box entspricht dem Interquartilsabstand. Der Median wird als durchgehender Strich in der Box eingezeichnet. Die Länge der Antennen sind John W. Tukey folgend auf maximal das 1,5-fache des Interquartilsabstands ($1,5IQR$) zu beschränken. Dabei endet der Whisker jedoch nicht genau nach dieser Länge, sondern bei dem Wert aus den Daten, der noch innerhalb dieser Grenze liegt. Werte außerhalb der Antennen werden separat in das Diagramm eingetragen und sind dann als Kandidaten für Ausreißer anzusehen.



Explorative Datenanalyse etc.

Normaler Quantil-Plot:

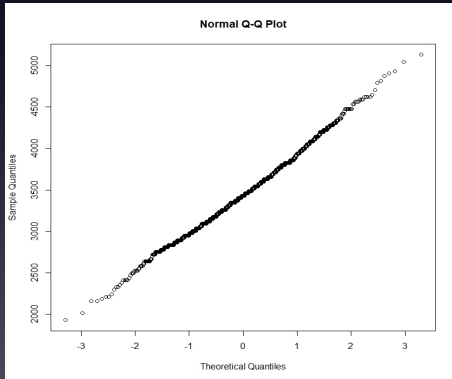
Angenommen, wir haben n Daten. Bezeichne mit q_k das $k/(n+1)$ -Quantil der Standardnormalverteilung und es seien $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ die geordneten Daten. Sind diese approximativ normalverteilt, so spricht dies zumindest für eine approximative Normalverteilung.



Explorative Datenanalyse etc.

Normaler Quantil-Plot:

Angenommen, wir haben n Daten. Bezeichne mit q_k das $k/(n+1)$ -Quantil der Standardnormalverteilung und es seien $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ die geordneten Daten. Sind diese approximativ normalverteilt, so spricht dies zumindest für eine approximative Normalverteilung.



Empirische Verteilung:

Für Daten x_1, \dots, x_n und $t \in \mathbb{R}$ definieren wir

$$\hat{F}_n(t) := \frac{1}{n} |\{x_i \mid x_i \leq t\}|$$

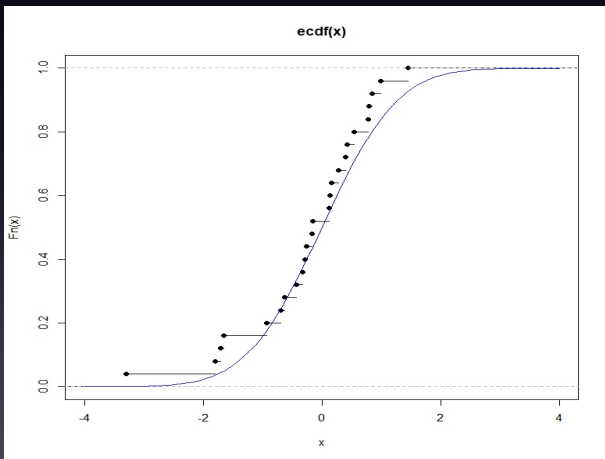


Empirische Verteilung:

Für Daten x_1, \dots, x_n und $t \in \mathbb{R}$ definieren wir

$$\hat{F}_n(t) := \frac{1}{n} |\{x_i \mid x_i \leq t\}|$$

Beispiel: Wir erzeugen 25 normal-verteilte Datenpunkte



Beispiel: Wir erzeugen 250 normal-verteilte Datenpunkte

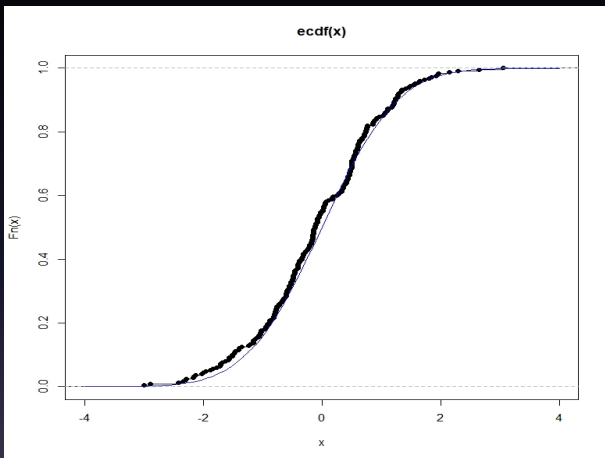


Abbildung: emp-Verteilung : x