

- Es soll hier erwähnt werden, dass neben einseitigen Tests auch zweiseitige Tests der Form

$$H : \{p = p_0\}$$

gegen

$$K : \{p \neq p_0\}$$

existieren.

- Das Testverfahren ist analog zum einseitigen Fall.
- Man konstruiert sich ein Intervall  $I \in p_0$ , so dass  $P_H\left(\frac{\sum_{i=1}^n X_i}{n} \in I\right) \geq 1 - \alpha$  und  $I$  dabei möglichst klein.
- Man entschließt sich  $H$  anzunehmen, falls  $\frac{\sum_{i=1}^n X_i}{n} \in I$  und entscheidet sich für  $K$ , falls  $\frac{\sum_{i=1}^n X_i}{n} \notin I$ .

- Wir werden nun ein Schätzproblem ansprechen, das sehr häufig in Anwendungen auftaucht.
- Zu diesem Zweck stellen wir uns vor, wir haben ein Experiment gemacht und im Laufe dieses Experiments Datenpaare  $(x_i, y_i)$ ,  $i = 1, \dots, n$  erhoben.
- Wir möchten die *funktionale Abhängigkeit* der  $y_i$  von den  $x_i$  quantifizieren; dabei gehen wir davon aus, dass bereits bekannt sei, dass die  $y_i$  (in guter Näherung) linear von den  $x_i$  abhängen, d.h.

$$y_i = f(x_i) = \alpha + \beta x_i, \quad \text{für alle } 1 \leq i \leq n.$$

- Dies kann entweder aufgrund theoretischer Überlegungen klar sein oder durch empirische Untersuchungen bestätigt sein.

- Zeichnen wir die Punkte  $(x_i, y_i)$  in ein cartesisches Koordinatensystem, so erwarten wir zwar, dass sie auf einer Geraden liegen.
- Jedoch werden sie – etwa bedingt durch Messfehler oder durch geringfügige Abweichungen von dem linearen Zusammenhang – nicht exakt eine Gerade beschreiben, sondern eher eine langgestreckte Punktwolke.
- Die Frage ist nun: Gegeben die Messwerte  $(x_i, y_i)$ ,  $1 \leq i \leq n$ , wie können wir eine sinnvolle Gerade durch diese Punkte legen, was sind die Parameter  $\hat{\alpha}$  und  $\hat{\beta}$  dieser sogenannten *Ausgleichsgeraden*.

- Dieses Problem unterscheidet sich von dem vorher behandelten Schätzproblem.
- Hier liegen ja nicht  $n$  durch den Zufall beeinflusste Beobachtungen ein und desselben Experiments vor.
- Stattdessen setzen wir voraus, die  $x_i$  genau zu kennen, und wollen durch eine geschickte Wahl von  $\hat{\alpha}$  und  $\hat{\beta}$  erreichen, dass die Daten

$$y_i - (\hat{\alpha} + \hat{\beta}x_i), \quad 1 \leq i \leq n,$$

möglichst wenig um den Wert 0 streuen.

- Diese Bedingung ist sicher sinnvoll, denn bei perfekter linearer Abhängigkeit der Daten und bei Kenntnis der Parameter hätte diese Differenz den Wert Null für alle  $1 \leq i \leq n$ .

- Wir wollen im Mittel die Gerade richtig schätzen.
- Dabei wollen wir im Sinne der Varianz möglichst wenig darum streuen.
- Wir werden also die Bedingungen

$$\sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x_i)) = 0 \quad \text{und} \quad \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x_i))^2 \quad \text{minimal} \quad (4)$$

für die Wahl von  $\hat{\alpha}$  und  $\hat{\beta}$  berücksichtigen.

- Diese Methode geht auf Carl Friedrich Gauß (1777–1855) zurück und wird wegen der zweiten Bedingung auch die *Methode der kleinsten Quadrate* genannt.

- Wir setzen zur Abkürzung

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i, \quad \overline{xx} := \frac{1}{n} \sum_{i=1}^n x_i^2, \quad \overline{xy} := \frac{1}{n} \sum_{i=1}^n x_i y_i, \quad (5)$$

und  $\bar{y}$  und  $\overline{yy}$  analog.

- Man beachten, dass  $\overline{xx} - (\bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 > 0$  ist, wenn nicht alle  $x_i = \bar{x}$  sind.
- Dies ist aber erfüllt, sobald wir zwei verschiedene Daten  $x_i$  und  $x_j$  haben.
- Wir berechnen den quadratischen Fehler

$$Q = \frac{1}{n} \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x_i))^2$$

## Definition

Die Gerade  $y = \hat{\alpha} + \hat{\beta}x$  mit

$$\hat{\beta} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{xx} - (\bar{x})^2} \quad \text{und} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

heißt *Ausgleichsgerade* der Datenpunkte  $(x_1, y_1), \dots, (x_n, y_n)$  und  $\hat{\beta}$  der *empirische Regressionskoeffizient*.