Ein kleinste–Quadrate–Schätzer

Definition

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

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

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

Der empirische Korrelationskoeffizient

- Es ist keine schwere Aufgabe zu zeigen, dass stets
$$-1 \leq \varrho \leq 1$$
gilt
- Es ist $\varrho = 1$ bzw. $\varrho = -1$ ist, wenn sämtliche Punkte (x_i, y_i) auf einer Geraden mit positiver bzw. negativer Steigung liegen.
- Daher wählt man den empirischen Korrelationskoeffizienten als ein Maß für die lineare Abhängigkeit der Daten.
- Allerdings bleibt zu beachten, dass der Korrelationskoeffizient nur einen Hinweis auf einen *linearen* Zusammenhang der Daten liefert.
- Andere, z.B. quadratische Abhängigkeiten können mit ihm nicht untersucht werden.
- Umgekehrt bedeutet ein ϱ nahe Null jedoch keineswegs, dass die Daten auch unabhängigen Zufallsvariablen entstammen.

Der exakte Test von Fisher und der χ^2–Test

- Wir haben mit dem empirischen Korrelationskoeffizienten ϱ und dem Regressionskoeffizienten $\hat{\beta}$ zwei Kennziffern kennengelernt, die lineare Abhängigkeit der Daten y_i von den Daten x_i, $1 \leq i \leq n$, aufzeigen und quantifizieren.
- Es ist sicher wünschenswert, auch statistische Tests zu Verfügung zu haben, welche die Hypothese untersuchen, dass die Daten y_i unabhängig von den Daten x_i sind.
- Dies liefert der sogenanntea *exakte Test von Fisher*.
- Ronald Aylmer Fisher (1890–1962) gilt als Pionier der statistischen Datenanalyse.
- Der exakte Test von Fisher ist ein Test auf Unabhängigkeit der Einträge der Vierfeldertafel.

Der exakte Test von Fisher und der χ^2–Test

- Im Allgemeinen haben wir folgende Situation:
- Wir unterteilen eine Grundmenge Ω von N Individuen anhand zweier Merkmale A und B mit je zwei Merkmalsausprägungen 0 und 1 und beobachten die folgenden Häufigkeiten.
-

	B_1	B_0	Summe
A_1	N_{11}	N_{10}	$N_{11} + N_{10}$
A_0	N_{01}	N_{00}	$N_{01} + N_{00}$
Summe	$N_{11} + N_{01}$	$N_{10} + N_{00}$	N

Der exakte Test von Fisher und der χ^2–Test

- Wir interpretieren die Beobachtung als Ausgang eines Zufallsexperiments, bei dem die Ausprägung der beiden Merkmale zufällig jedem Individuum zugeordnet wird.
- Vom zugrunde liegenden Zufallsmechanismus wissen wir gar nichts und wir fragen uns, ob die Ausprägungen voneinander unabhängig zugeordnet werden.
- Dies ist genau dann der Fall, wenn gilt

$$\mathbb{P}\Big(\{\omega \in \Omega \mid \omega \text{ hat } B_1\} \Big| \{\omega \in \Omega \mid \omega \text{ hat } A_1\}\Big) = \mathbb{P}(\{\omega \in \Omega \mid \omega \text{ hat } B_1\})$$

- Dies ist unsere Nullhypothese.
- Allerdings können wir sie nicht ohne weiteres mit der Binomialverteilung testen, denn die Zuordnung von Ausprägungen bei konstant gehaltenen Randsummen ist ja kein Ziehen mit Zurücklegen.

Der exakte Test von Fisher und der χ^2–Test

- Aus Beispielen wird deutlich, dass dieser Test für einen großen Stichprobenumfang auch die Berechnung unhandlicher Binomialkoeffizienten erfordert.
- Daher wird in der Praxis häufig ein anderer Test verwendet, der sogenannte χ^2–Test.
- Er besitzt darüber hinaus den Vorteil, auch dann anwendbar zu sein, wenn mehr als zwei Merkmale mit je zwei möglichen Ausprägungen zugeordnet werden.
- Dazu führen wir für $n \in \mathbb{N}$ die Funktion $g_n : (0, \infty) \to \mathbb{R}$ mit

$$g_n(x) = x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$$

ein.

- Das Integral $c_n := \int_0^\infty g_n(x)\,dx$ hat einen endlichen Wert
- Dieser Wert ist $c_n = 2^{n/2}\Gamma(n/2)$, wobei Γ die *Gammafunktion* bezeichnet.
- Daher ist

$$\gamma_n(x) := \frac{1}{c_n} g_n(x) = \frac{1}{c_n} x^{\frac{n}{2}-1} e^{-x^2}\ , \ x > 0, \qquad (6)$$

eine Wahrscheinlichkeitstsdichte.

Definition

Eine Zufallsvariable X heißt χ^2*–verteilt (lies:„chi–Quadrat“) mit* n *Freiheitsgraden*, wenn die Verteilung $\mathbb{P}_X$ die Dichte aus Gleichung (3) besitzt, wenn also gilt

$$\mathbb{P}(\{a \leq X \leq b\}) = \frac{1}{c_n} \int_a^b x^{\frac{n}{2}-1} e^{-\frac{x}{2}}\, dx \ , \quad \text{für } 0 < a < b.$$

- Genau wie bei der Normalverteilung, so ist auch bei der χ^2–Verteilung die Stammfunktion nicht elementar darstellbar, jedoch findet sich auch für sie eine Approximation in jedem guten Programm zur Tabellenkalkulation.