

## Übungsblatt 12

### $\chi^2$ -Tests, Varianzanalyse, lineare Regression

Abgabetermin: **Mittwoch, 26. Mai 2010**, bzw. **Freitag, 28. Mai 2010**, bei der Semesterassistentin oder beim Semesterassistenten in der jeweiligen Übungsstunde.

#### $\chi^2$ -Test auf gegebene Verteilung

##### **Aufgabe 126** °:

In einem grossen Wald mit gleichartigem Gelände und Baumbestand wird das Vorkommen der Heidelbeere untersucht. Auf quadratischen Flächen von je 1 m<sup>2</sup> Grösse erhält man durch Auszählung der Pflanzen:

Anzahl der Planzen pro m <sup>2</sup>	0	1	2	3	4	5	6	≥ 7
Anzahl der Flächenstücke	15	18	11	4	1	0	1	0

Darf eine Poissonverteilung mit Erwartungswert 1.2 angenommen werden?

#### $\chi^2$ -Test auf Unabhängigkeit

##### **Aufgabe 127** (4 Punkte):

Am 27. Januar 1987 berichtete die New York Times auf der Titelseite von den Resultaten einer Studie über die präventive Wirkung von Aspirin gegen Herzinfarkte bei Männern mittleren Alters. Für die Studie wurden 22071 Männer mittleren Alters zufällig je einer von zwei Gruppen zugeordnet. Der einen Gruppe wurde Aspirin verabreicht, der anderen ein Placebo. Von 11037 Personen, die Aspirin eingenommen hatten, bekamen 104 einen Herzinfarkt; von den 11034 Personen, welchen ein Placebo verabreicht wurde, erlitten 189 einen Herzinfarkt. Besteht ein signifikanter Unterschied ( $\alpha = 0.001$ )?

#### Einfache Varianzanalyse

**Hinweis.** Gegeben sind  $k$  Gruppen, in der  $i$ -ten Gruppe  $n_i$  Beobachtungen  $y_{i1}, y_{i2}, \dots, y_{in_i}$ . Wir betrachten das Modell

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{für } i = 1, \dots, k \text{ und } j = 1, \dots, n_i,$$

wobei die Fehler  $\varepsilon_{11}, \dots, \varepsilon_{kn_k}$  unabhängig und normalverteilt mit Erwartungswert 0 und gleicher Varianz  $\sigma^2$  sind. Zu testen ist die Nullhypothese

$H_0$ : „Die Gruppen unterscheiden sich bezüglich der Erwartungswerte nicht, d.h.  $\mu_1 = \dots = \mu_k$ .“

gegen die Alternative

$H_1$ : „Es gibt Unterschiede bei den Erwartungswerten der Gruppen, d.h. es existieren  $r$  und  $s$ ,  $1 \leq r, s \leq k$ , so dass  $\mu_r \neq \mu_s$ .“

Für eine geeignete Teststatistik benötigen wir folgende Notationen. Wir schreiben

$$\bar{Y}_{i.} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \quad i = 1, \dots, k$$

für die Gruppenmittelwerte, und

$$\bar{Y}_{..} := \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}, \quad \text{wobei } n := n_1 + \dots + n_k,$$

für den Gesamtdurchschnitt (grand mean). Dann ist

$$MS_G := \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

ein Mass für die Streuung der Gruppenmittelwerte (mean square of groups), und

$$MS_E := \frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

stellt ein Mass für die Streuung der Beobachtungen innerhalb der einzelnen Gruppen dar (mean square of errors). Ist  $MS_G$  einiges grösser als  $MS_E$ , so können wir annehmen, dass sich die Gruppen hinsichtlich der Erwartungswerte unterscheiden. Wir verwenden als Teststatistik

$$V := \frac{MS_G}{MS_E}.$$

Unter der Nullhypothese ist  $V$   $F$ -verteilt mit  $k-1$  und  $n-k$  Freiheitsgraden.

### Aufgabe 128 °:

Zwei Gruppen von männlichen Ratten erhielten stark bzw. schwach proteinhaltiges Futter. Die Gewichtszunahmen in Gramm waren bei der ersten Gruppe

148, 143, 161, 125, 141, 151, 166, 158, 149, 148

und bei der zweiten Gruppe

142, 135, 134, 126, 148, 142, 148, 127.

Prüfen Sie die Hypothese  $H_0$  zum Niveau  $\alpha = 5\%$ , dass die beiden Gruppen gleiche Erwartungswerte haben gegen die Alternative  $H_1$ , dass die Erwartungswerte verschieden sind mit Hilfe der Varianzanalyse.

### Aufgabe 129 (3 Punkte):

Welches Resultat liefert ein  $t$ -Test für zwei unabhängige Stichproben in der Situation der vorhergehenden Aufgabe ( $\alpha = 5\%$ )?

### Aufgabe 130 (6 Punkte):

Wir messen die Durchmesser der Häuschen einer Schneckenart auf drei verschiedenen Wiesen. Pro Wiese vermessen wir vier Schnecken und notieren die erhaltenen Werte (in cm) in die folgende Tabelle:

Wiese 1	Wiese 2	Wiese 3
1.67	1.11	1.74
1.23	1.61	1.62
1.44	1.42	1.55
1.53	1.52	1.89

Unterscheiden sich die Schnecken dieser Art hinsichtlich der Häuschendurchmesser auf den verschiedenen Wiesen? Führen Sie zur Beantwortung dieser Frage eine Varianzanalyse durch. Wählen Sie als Signifikanzniveau  $\alpha = 5\%$ .

### Einfache lineare Regression

**Hinweis.** Wir betrachten das Modell

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad \text{für } i = 1, \dots, n,$$

wobei die Fehler  $\varepsilon_1, \dots, \varepsilon_n$  unabhängig und normalverteilt mit Erwartungswert 0 und gleicher Varianz  $\sigma^2$  sind. Die Methode der kleinsten Quadrate führt zu folgenden Schätzern der Steigung  $\beta$  und des Achsenabschnitts  $\alpha$ :

$$\hat{\beta} := \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{und} \quad \hat{\alpha} := \bar{Y} - \hat{\beta} \bar{x}.$$

Wir bezeichnen die  $y$ -Werte unserer Regressionsgeraden  $y = \hat{\alpha} + \hat{\beta}x$  an den Datenpunkten  $x_i$  mit  $\hat{Y}_i$ , also

$$\hat{Y}_i := \hat{\alpha} + \hat{\beta}x_i \quad \text{für } i = 1, \dots, n.$$

Ein Schätzer der Varianz  $\sigma^2$  der  $\varepsilon_i$  ist gegeben durch

$$\hat{\sigma}^2 := \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

Wir testen, ob die Nullhypothese  $H_0: \beta_1 = b$  verworfen werden kann oder nicht. Die dazu geeignete Teststatistik

$$T := \frac{\hat{\beta} - b}{\sqrt{\hat{\sigma}^2 / \sum_{i=1}^n (x_i - \bar{x})^2}}$$

folgt unter  $H_0$  der  $t$ -Verteilung mit  $n - 2$  Freiheitsgraden. Meistens ist die Frage interessant, ob die  $X$ -Grösse überhaupt Einfluss auf die  $Y$ -Grösse hat, d.h. wir testen die Nullhypothese  $H_0: \beta = 0$  („kein Einfluss“).

### Aufgabe 131 (5 Punkte):

Zeigen Sie, dass

$$SS_Y = SS_R + SS_E,$$

wobei

$$SS_Y := \sum_{i=1}^n (y_i - \bar{y})^2, \quad SS_R := \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad SS_E := \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

In Worten: Die Variation in den  $y$ -Daten ( $SS_{yy}$ , die totale Quadratsumme) lässt sich aufspalten in einen Teil der durch die Regression erklärt wird ( $SS_R$ , die Quadratsumme der Regression) und eine Summe der Fehler ( $SS_E$ , die Quadratsumme der Fehler).

**Aufgabe 132** (5 Punkte):

Zeigen Sie, dass die Schätzer  $\hat{\beta}$  und  $\hat{\alpha}$  erwartungstreu sind:

$$E(\hat{\beta}) = \beta \quad \text{und} \quad E(\hat{\alpha}) = \alpha.$$

*Hinweis:* Benutzen Sie die Eigenschaft, dass der Erwartungswert linear ist und der Erwartungswert einer Konstanten gerade die Konstante selbst ist:

$$E(c_1 + c_2 Y) = c_1 + c_2 E(Y),$$

wobei  $c_1, c_2$  konstant sind und  $Y$  eine Zufallsgröße ist. Allenfalls können Sie auch  $E(\hat{\beta}) = \beta$  annehmen und dann zeigen, dass  $E(\hat{\alpha}) = \alpha$  gilt. Dies ist weniger aufwendig als der Nachweis von  $E(\hat{\beta}) = \beta$ .

**Aufgabe 133** (5 Punkte):

Beim Bau eines Strassentunnels zur Unterfahrung einer Ortschaft müssen Sprengungen durchgeführt werden. Die Erschütterung der Häuser darf einen bestimmten Wert nicht überschreiten. Man misst daher die Erschütterung  $y_i$  im Keller eines gefährdeten Hauses. Diese wird, bei konstanter Sprengung, im wesentlichen von der Distanz  $x_i$  zum Sprengort der  $i$ -ten Sprengung abhängen. Aus physikalischen Gründen kann angenommen werden, die Erschütterung  $y_i$  sei umgekehrt proportional zur quadrierten Distanz  $x_i^2$ , also

$$y_i \approx \alpha x_i^\beta, \quad \text{mit } \beta = -2.$$

Durch logarithmieren auf beiden Seiten erhalten wir eine lineare Abhängigkeit

$$\log(y_i) \approx \log(\alpha) + \beta \log(x_i).$$

Den logarithmierten Daten  $x'_i := \log(x_i)$  und  $y'_i := \log(y_i)$  können wir also ein Modell der Form

$$y'_i = \log(\alpha) + \beta x'_i + \varepsilon_i$$

zugrundelegen, die Theorie der linearen Regression anwenden und  $\beta$  schätzen.

Wir nehmen an, es seien 13 Sprengungen durchgeführt worden, und aus den Datenpaaren  $(x'_i, y'_i)$  wurde

$$SS_E = 0.1442, \quad SS_X = \sum_{i=1}^{13} (x'_i - \bar{x}')^2 = 0.4122 \quad \text{und} \quad \hat{\beta} = -1.9235$$

bereits berechnet. Wir wollen die physikalische Annahme prüfen, dass die Erschütterung umgekehrt proportional zur quadrierten Distanz ist. Testen Sie dazu die Nullhypothese  $H_0: \beta_1 = -2$  gegen die Alternative  $H_1: \beta_1 \neq -2$  zum Signifikanzniveau 5%.

**Aufgabe 134** (8 Punkte):

Gegeben seien folgende Messwerte:

$x_i$	1	1.5	3	3.5	4	6	7.5	8	9.5	10
$y_i$	2	2.5	4	0	1.5	7	6	10	5.5	9.5

- Skizzieren Sie diese Punkte in einem geeigneten Koordinatensystem.
- Bestimmen Sie die Gleichung der Regressionsgeraden  $y = \hat{\alpha} + \hat{\beta}x$  durch diese Punktpaare und zeichnen Sie diese im Koordinatensystem ein.
- Die Schätzungen  $\hat{\alpha}$  und  $\hat{\beta}$  (und damit auch die Gleichung der Regressionsgeraden) basieren auf der Methode der kleinsten Quadrate. Zeichnen Sie diese „kleinsten Quadrate“ im Bild ein.
- Offenbar hat die  $X$ -Grösse Einfluss auf die Zielgrösse  $Y$ . Was sagt hier der Test von  $H_0: \beta = 0$  gegen  $H_1: \beta \neq 0$  zum Signifikanzniveau 5%?

**Aufgabe 135** (4 Punkte):

Wir haben bei einer einfachen linearen Regression die  $x$ -Werte (Längenmessungen) in Metern angegeben. Nun möchte man die Masseinheit der  $x$ -Werte in Zentimeter ändern. Die Masseinheit der  $y$ -Werte lässt man gleich. Wie verändern sich nun die Schätzungen  $\hat{\alpha}$ ,  $\hat{\beta}$  und der Test, ob  $\beta = 0$ ?

*Die letzte Aufgabe*

**Achtung: Diese Aufgabe hat nichts mit Varianzanalyse oder linearer Regression zu tun. Lösen Sie sie mit einem Baumdiagramm!**

**Aufgabe 136** (4 Punkte):

In einer Fernseh-Show kann ein aus dem Publikum ausgewählter Kandidat auf folgende Art einen Preis im Wert von 100'000 Fr. gewinnen: Er hat drei geschlossene Türen zur Auswahl, wobei hinter genau einer der Preis versteckt worden ist. Nun darf er sich für eine Tür die er öffnen will entscheiden. Bevor diese geöffnet wird, teilt er seine Entscheidung dem Showmaster mit. Dieser, der natürlich weiss, hinter welcher Türe sich der Preis verbirgt, öffnet nun eine der beiden Türen die der Kandidat nicht ausgewählt hat. Er öffnet jedoch sicher nicht diejenige hinter der sich der Preis befindet.

Der Kandidat hat nun die Möglichkeit bei seiner Entscheidung zu bleiben oder zur anderen noch geschlossenen Türe zu wechseln, um diese dann öffnen zu lassen.

Wie soll er sich entscheiden? Soll er die Türe wechseln oder bei seiner ersten Entscheidung bleiben? Spielt es überhaupt eine Rolle wie er sich entscheidet?