Einführung in die induktive Statistik

Friedrich Leisch

Institut für Statistik
Ludwig-Maximilians-Universität München

SS 2009, Statistische Signifikanztests

Beispiel: Sonntagsfrage

Vier Wochen vor der österreichischen Nationalratswahl 1999 wurde 499 Haushalten die „Sonntagsfrage“ gestellt: Falls nächsten Sonntag Wahlen wären, welche Partei würden Sie wählen?

	SPÖ	ÖVP	FPÖ	Grüne	LIF	Sonst
Umfrage	38%	24%	25%	6%	4%	3%
Wahl	33.15%	26.91%	26.91%	7.4%	3.65%	1.98%

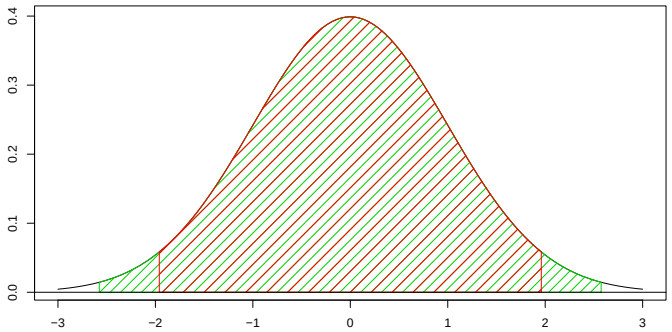
Frage 1: War das Ergebnis für die SPÖ überraschend?

Frage 2: Mit welcher Wahrscheinlichkeit mußte das LIF damit rechnen, den Wiedereinzug ins Parlament nicht zu schaffen? Mit welcher Wahrscheinlichkeit die Grünen?

Frage 3: War das Gesamtergebnis überraschend?

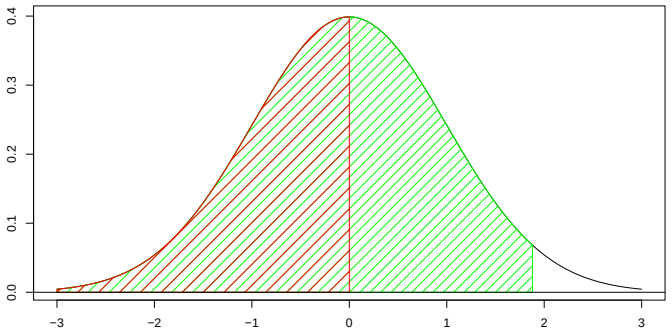
Statistische Tests

Beim Beantworten der Frage 1 mit Hilfe von Konfidenzintervallen haben wir nachgesehen, ob der z-Wert in einem 95% bzw. 99% KI liegt:



Statistische Tests

Frage 2: mit welcher Wahrscheinlichkeit ist der Parameter kleiner als eine vorgegebene Schranke? Aus 4% wurden $t = 1.88$ für die Grünen bzw. $t = 0$ für die Liberalen.



Statistische Tests

In beiden Fällen haben wir versucht—bei gegebener Stichprobe—eine Hypothese zu überprüfen und mit einer Alternative zu vergleichen:

1. **Hypothese:** Trotz Umfragewerten von 38% bekommt die SPÖ nur 33.15%.
   Alternative: Bei 38% in der Umfrage ist 33.15% kein wahrscheinliches Resultat für die SPÖ.

2. **Hypothese:** Die Grünen bzw. Liberalen bekommen (mindestens) 4%.
   Alternative: Die Grünen bzw. Liberalen bekommen weniger als 4%.

Statistische Tests

Alle klassischen statistischen Tests basieren auf diesem Grundprinzip:

1. Es wird eine sogenannte Nullhypothese H_0 und eine Alternative H_1 gebildet.
2. Mittels einer Teststatistik wird berechnet, wie wahrscheinlich die interessierende Eigenschaft der Stichprobe unter der Nullhypothese ist. Bildung eines Konfidenzintervalls für die Teststatistik (unter der Nullhypothese), Reduktion des Entscheidungsproblems auf den Wert einer einzigen Zahl.
3. Falls die Teststatistik außerhalb des Konfidenzintervalls liegt, wird die Nullhypothese verworfen (zu unwahrscheinlich gegeben die Stichprobe).
4. Die Art der verwendeten Teststatistik bestimmt, welche Eigenschaft der Verteilung untersucht wird (Lokation, Streuung, . . .).

Statistische Tests

Fehler 1. Art: Die Nullhypothese stimmt, aber der Test verwirft sie (=Größe oder Signifikanzniveau α des Tests).
Fehler 2. Art: Die Alternative stimmt, aber der Test akzeptiert die Nullhypothese.

	Realität	
Testresultat	Nullhypothese	Alternative
Nullhypothese	1− Größe	Fehler 2. Art
Alternative	Fehler 1. Art	Macht

Optimaler Test: Maximale Macht bei gegebener Größe, gleichzeitige Reduktion beider Fehlerarten nicht mehr möglich.

Die Gewichtung der beiden Fehlerarten hängt meist von der Anwendung ab.

Design von Experimenten

Bei der Planung von Experimenten, die mit Hilfe statistischer Tests ausgewertet werden sollen, spielt die Güte des Tests eine wesentliche Rolle:

1. Wahl des Testverfahrens (z.B. Gauß-Test)

2. Nullhypothese und Alternative festlegen

3. Wahl des Signifikanzniveaus α, typische Werte für α sind 0.05, 0.03, oder 0.01.

Falls die Alternative auch als Punkthypothese formuliert wird, kann die notwendige Stichprobengröße n für eine gewünschte Macht berechnet werden (z.B. aus Gütefunktion ablesbar).

Mittelwert und Varianz der Normalverteilung

Gauß-Test

Gegeben sei eine normalverteilte Stichprobe, die Varianz σ^2 ist bekannt.

Nullhypothese: Erwartungswert $\mu = \mu_0$
Alternative: Erwartungswert $\mu \neq \mu_0$

Die Teststatistik

$$z = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}}, \quad \hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

ist standardnormalverteilt, falls die Nullhypothese stimmt.
$\rightarrow$ Ablehnung der Nullhypothese falls $|z|$ „zu groß“.

Z.B. für $\alpha = 0.05$ lehnen wir H_0 ab, falls $|z| > z_{0.975} = 1.96$.

Anmerkung: Der Test kann wegen des zentralen Grenzwertsatzes auch für hinreichend große nicht normalverteilte Stichproben verwendet werden.

Gauß-Test für Proportionen

(approximativer Binomialtest)

Gegeben sei eine binomialverteilte Stichprobe. Da die Varianz $\sigma^2 = \pi(1-\pi)$ unter der Nullhypothese bekannt ist, können wir bei hinreichend großem n den Gauß-Test für Hypothesen über die Trefferwahrscheinlichkeit π verwenden.

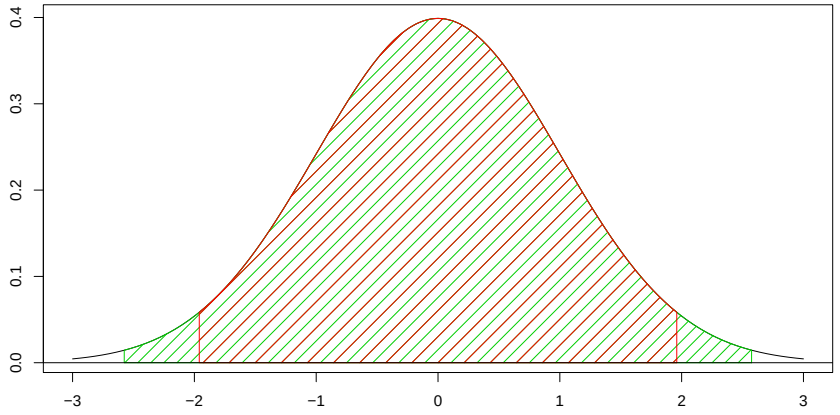
Nullhypothese: $\pi = \pi_0$
Alternative: $\pi \neq \pi_0$

Die Teststatistik

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim N(0,1)$$

ist für hinreichend großes n standardnormalverteilt.

Gauß-Test



Gauß-Test: Fortsetzung Frage 1

Korrekt formuliert muß der Test für Frage 1 lauten:

Nullhypothese: $\pi = 33.15\%$
Alternative: $\pi \neq 33.15\%$

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.38 - 0.3315}{\sqrt{0.3315(1-0.3315)/499}} \approx 2.30$$

Wir verwenden also die theoretische Varianz unter der Null statt der Stichprobenvarianz. Das Resultat verändert sich nicht wesentlich (Ablehnung der Null bei $\alpha = 0.05$, Annahme bei $\alpha = 0.01$).

Einseitige Gauß-Tests

In manchen Fällen hat man entweder eine Vermutung über die Richtung der Abweichung von der Nullhypothese oder eine Richtung ist irrelevant.

Dann wird ein sogenannter einseitiger Test formuliert (der größere Macht gegen die Alternative hat):

Nullhypothese: $\mu = \mu_0$
Alternative 1: $\mu < \mu_0$
Alternative 2: $\mu > \mu_0$

Die Teststatistik

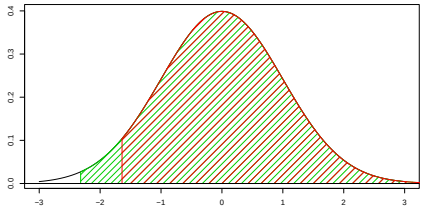
$$z = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}}, \quad \hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

bleibt gleich, wird aber mit einem einseitigen Konfidenzintervall verglichen.

Ein- & zweiseitige Gauß-Tests

$H_1 : \mu \neq \mu_0$

$H_1 : \mu < \mu_0$



Gauß-Test: Fortsetzung Frage 2

Korrekt formuliert muß der Test für Frage 2 lauten:

Nullhypothese: $\pi = 4\%$
Alternative: $\pi < 4\%$

$$z_G = \frac{\hat{\pi}_G - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.06 - 0.04}{\sqrt{0.04(1-0.04)/499}} \approx 2.28$$

$$z_L = \frac{\hat{\pi}_L - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.04 - 0.04}{\sqrt{0.04(1-0.04)/499}} = 0$$

In beiden Fällen können wir also den Verbleib im Parlament nicht ausschließen.

Achtung: Nullhypothese und Alternative werden nicht symmetrisch behandelt!!!

Einseitige Gauß-Tests

Was passiert bei einem einseitigen Test der Form

Nullhypothese: $\mu = \mu_0$
Alternative: $\mu < \mu_0$

wenn in Wahrheit $\mu > \mu_0$ ist? Es wird (mit großer Wahrscheinlichkeit) die Nullhypothese angenommen. Größere Macht hat einen Preis, der Test hat eine „blinden Fleck“ bekommen.

Einseitige Tests sind das häufigste Beispiel für Tests, wo die Alternative keine sogenannte Omnibus-Alternative („alles außer der Null“) ist. Jeder Test kann nur zwischen Null und Alternative unterscheiden, wenn die Wahrheit außerhalb dieser beiden Bereiche liegt, ist das Ergebnis unbestimmt, d.h., man kann noch nicht einmal die Wahrscheinlichkeit dafür angeben.

Kritische Werte

Der sogenannte kritische Wert eines Tests ist das Quantil der Verteilung der Teststatistik, mit dem die beobachtete Teststatistik bei gegebenem α verglichen werden muß. Für Gauß-Tests sind dies Quantile der Standardnormalverteilung:

Alternative	Ablehnung Nullhypothese
$\mu \neq \mu_0$	$\lvert z \rvert > z_{1-\alpha/2}$
$\mu < \mu_0$	$z < z_\alpha$
$\mu > \mu_0$	$z > z_{1-\alpha}$

Anmerkung: Für symmetrische Verteilungen wie die Normalverteilung gilt $z_\alpha = -z_{1-\alpha}$ und in Tabellen finden sich oft nur die $1-\alpha$-Quantile.

Der Bereich, in dem die Nullhypothese verworfen wird, heißt auch Ablehnungsbereich des Tests.

p-Werte

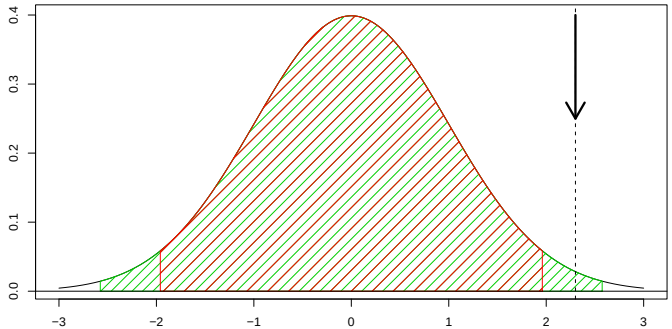
An Frage 1 (SPÖ) haben wir gesehen, daß das Signifikanzniveau α starken Einfluß auf den Ausgang des Tests haben kann, falls die Teststatistik z im Grenzbereich liegt: Für $\alpha = 0.05$ hätten wir H_0 verworfen, für $\alpha = 0.01$ akzeptiert.

Eine alternative Möglichkeit besteht darin, sich direkt die Wahrscheinlichkeit der Teststatistik unter der Nullhypothese zu betrachten.

Exakt formuliert: Als p-Wert eines Tests wird die Wahrscheinlichkeit bezeichnet, unter der Nullhypothese die vorliegende Stichprobe oder eine noch unwahrscheinlichere zu beobachten.

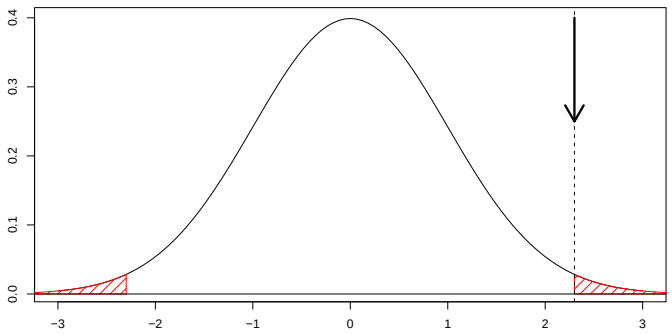
Die Nullhypothese wird abgelehnt, wenn der p-Wert sehr klein ist (0.05, 0.03, . . .).

p-Werte



Bsp SPÖ: $z = 2.30$

p-Werte



Bsp SPÖ: $p = 2 * (1 - \Phi(|z|)) = 2 * (1 - \Phi(2.30)) \approx 0.0214$

Güte von Tests

Nach Konstruktion gilt bei Tests mit Signifikanz α

$$P(\text{Fehler 1.Art}) = P(H_0 \text{ ablehnen} \mid H_0 \text{ richtig}) \leq \alpha$$

Die Güte des Tests definiert sich also primär über die Wahrscheinlichkeit für den Fehler 2.Art:

$$\begin{aligned} P(\text{Fehler 2.Art}) &= P(H_0 \text{ nicht ablehnen} \mid H_1 \text{ richtig}) \\ &= 1 - P(H_0 \text{ ablehnen} \mid H_1 \text{ richtig}) \end{aligned}$$

Güte von Tests

Gütefunktion $g(\mu)$ fasst $P(\text{Fehler 1.Art})$ und $P(\text{Fehler 2.Art})$ in einer Funktion zusammen:

$$g(\mu) = P(H_0 \text{ ablehnen} \mid \mu)$$

mit μ der unbekannte wahre Parameter.

Es gilt:

- $\alpha = P(\text{Fehler 1.Art}) = g(\mu_0)$
- $\beta = P(\text{Fehler 2.Art}) = 1 - g(\mu),\ \mu \neq \mu_0$

Güte von Tests

Beispiel: Mittelwert einer normalverteilten Stichprobe (Gauß-Test)

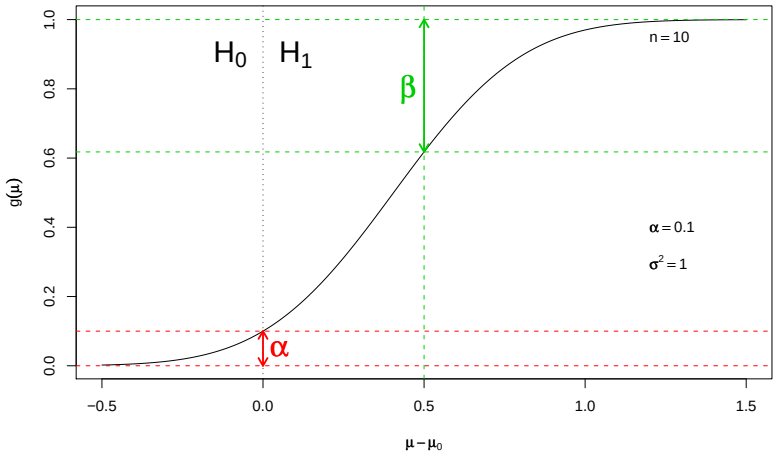
$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}}, 1\right)$$

Einseitiger Test:

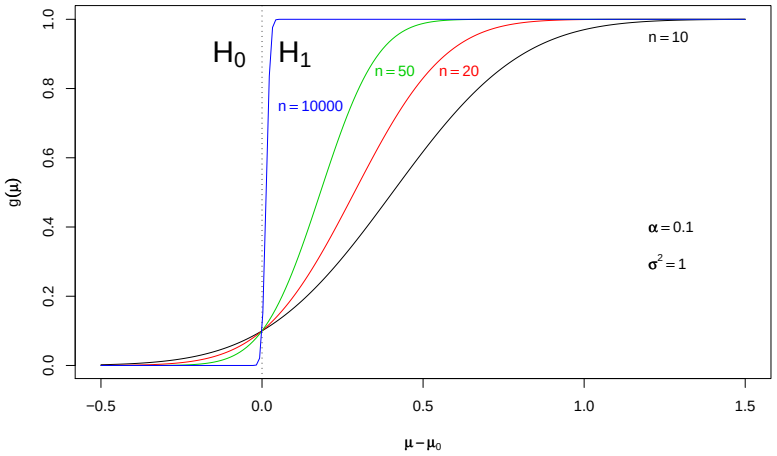
$$H_0 : \mu \leq \mu_0 \qquad H_1 : \mu > \mu_0$$

$$\begin{aligned} g(\mu) &= P(H_0 \text{ ablehnen} \mid \mu) = \\ &= P(Z \geq z_{1-\alpha} \mid \mu) = \\ &= 1 - \Phi\left(z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

Güte von Tests



Güte von Tests



Güte von Tests

Beispiel: Mittelwert einer normalverteilten Stichprobe (Gauß-Test)

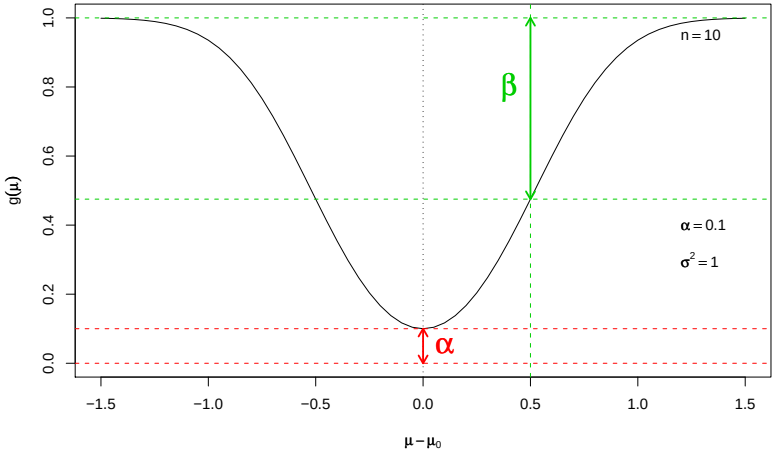
$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}}, 1\right)$$

Zweiseitiger Test:

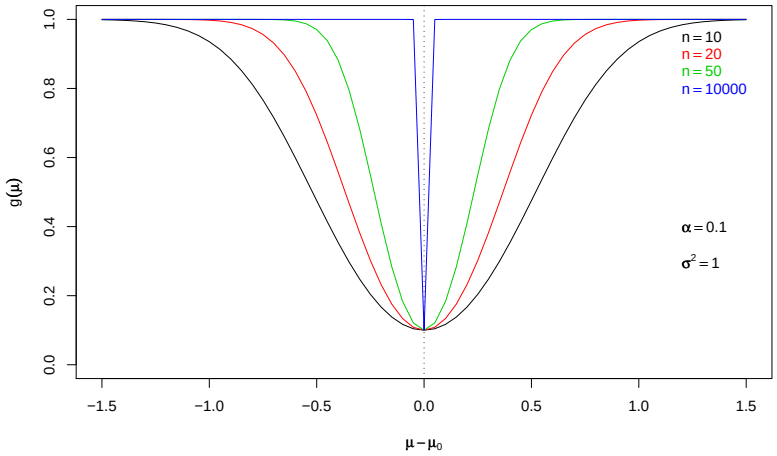
$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

$$\begin{aligned} g(\mu) &= P(H_0 \text{ ablehnen} \mid \mu) = \\ &= P(|Z| \geq z_{1-\alpha/2} \mid \mu) = \\ &= \Phi\left(-z_{1-\alpha/2} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

Güte von Tests



Güte von Tests



t-Test

Gegeben sei eine normalverteilte Stichprobe, die Varianz σ^2 ist jedoch nicht bekannt und muß durch $\hat{\sigma}^2$ geschätzt werden.

Nullhypothese: Erwartungswert $\mu = \mu_0$
Alternative: Erwartungswert $\mu \neq \mu_0$

Die Teststatistik

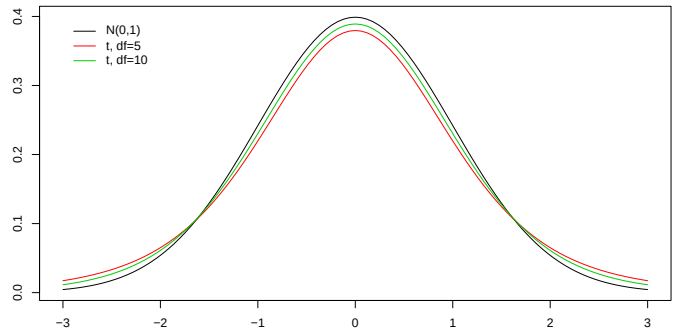
$$t = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

ist t-verteilt mit $n-1$ Freiheitsgraden.

t-Test

Bei wachsenden Freiheitsgraden konvergiert die t-Verteilung gegen die N(0,1), da die Varianz immer besser geschätzt werden kann:



t-Test

Verwendung einer t-Tabelle: Ganz analog zur Normalverteilungstabelle, allerdings müssen die Freiheitsgrade noch zusätzlich berücksichtigt werden.
Bei zweiseitigem Test müssen Quantile zum Niveau $1-\alpha/2$ verwendet werden, z.B., bei $\alpha = 0.05$ Quantile für 0.975.

df	0.95	0.975
1	6.314	12.706
2	2.920	4.303
3	2.353	3.182
4	2.132	2.776
5	2.015	2.571
6	1.943	2.447
7	1.895	2.365
8	1.860	2.306
9	1.833	2.262
10	1.812	2.228
⋮	⋮	⋮

t-Test

Beispiel 1a: Einer Lieferung Dioden mit gewünschtem Durchlaßwiderstand von $100m\Omega$ wird eine zufällige Stichprobe der Größe 10 entnommen und vermessen:

```
114.62 110.10 106.31  99.30 107.28
108.35 113.64 117.92 130.15 102.74
```

Wir erhalten $\hat{\mu} = 111.04$, $\hat{\sigma} = 8.71$ und somit $t = 4.0066$ bei 9 Freiheitsgraden. Wir verwerfen daher die Nullhypothese: der kritische Wert aus der Tabelle beträgt 2.262 bei $\alpha = 0.05$, der p-Wert des Tests ist 0.00308.

Der mittlere Durchlaßwiderstand entspricht mit großer Wahrscheinlichkeit nicht den Herstellerangaben.

t-Test: zwei Stichproben

Gegeben seien zwei unabhängige normalverteilte Stichproben X und Y vom Umfang n_x und n_y, deren Varianzen σ_x^2 und σ_y^2 nicht bekannt sind.

Nullhypothese: Differenz der Erwartungswerte $\mu_x - \mu_y = \mu_0$
Alternative: Differenz der Erwartungswerte $\mu_x - \mu_y \neq \mu_0$

Die Teststatistik

$$t = \frac{(\hat{\mu}_x - \hat{\mu}_y) - \mu_0}{\hat{\sigma}}$$

ist t-verteilt, wobei die genaue Form des Varianzschätzers $\hat{\sigma}^2$ und die Freiheitsgrade df noch davon abhängen, ob die Varianzen σ_x^2 und σ_y^2 als gleich angenommen werden dürfen.

t-Test: zwei Stichproben

Varianzen gleich:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{(n_x + n_y)[(n_x - 1)\hat{\sigma}_x^2 + (n_y - 1)\hat{\sigma}_y^2]}{n_x n_y (n_x + n_y - 2)} \\ df &= n_x + n_y - 2 \end{aligned}$$

Varianzen ungleich:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y} \\ df &= \frac{(\hat{\sigma}_x^2/n_x + \hat{\sigma}_y^2/n_y)^2}{\hat{\sigma}_x^4/[n_x^2 * (n_x - 1)] + \hat{\sigma}_y^4/[n_y^2 * (n_y - 1)]} \end{aligned}$$

t-Test: zwei Stichproben

Beispiel 1b: Wir haben zwei Lieferungen A und B von Dioden und wollen wissen, ob der Durchlaßwiderstand der beiden Lieferungen ident ist ($\mu_0 = 0$). Jeder Lieferung wird eine Stichprobe vom Umfang 10 entnommen und vermessen:

```
A: 114.62 110.10 106.31  99.30 107.28
   108.35 113.64 117.92 130.15 102.74
B: 101.77 109.86 131.41 105.29 104.49
   118.62 108.60 139.09 113.72 114.91
```

Wir erhalten $\hat{\mu}_A = 111.04$, $\hat{\sigma}_A = 8.71$, $\hat{\mu}_B = 114.77$ und $\hat{\sigma}_B = 12.06$. Unter der Annahme gleicher Varianzen ergibt sich $t = -0.7934$ bei 18 Freiheitsgraden. Die Nullhypothese wird angenommen: der entsprechende kritische Wert aus der t-Tabelle ist 2.100922, der p-Wert des Tests ist 0.4379. Die beiden Lieferungen haben mit großer Wahrscheinlichkeit denselben mittleren Durchlaßwiderstand.

t-Test: verbundene Stichproben

Gegeben sei eine Stichprobe $\{(x_1, y_1), \ldots (x_n, y_n)\}$ aus abhängigen Paaren (X, Y) normalverteilter Größen, deren Varianz unbekannt ist.

Für die Hilfsvariable $z = x - y$ haben wir die Stichprobe

$$\{z_1 = x_1 - y_1, \ldots, z_n = x_n - y_n\},$$

diese dürfen wir als unabhängig identisch normalverteilt betrachten.
$\rightarrow$ Für Hypothesen über den Mittelwert der Differenz von X und Y können wir den normalen t-Test auf die neue Stichprobe Z anwenden.

t-Test: verbundene Stichproben

Nullhypothese: Erwartungswert der Differenzen $\mu_z = \mu_0$
Alternative: Erwartungswert der Differenzen $\mu_z \neq \mu_0$

Die Teststatistik

$$t = \frac{\hat{\mu}_z - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

$$\hat{\mu}_z = \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad \hat{\sigma}_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$$

ist t-verteilt mit $n - 1$ Freiheitsgraden.

t-Test: verbundene Stichproben

Beispiel 2: In einem Betrieb werden die Reißlasten von Drähten mit einer Maschine A untersucht. Da der Bedarf an solchen Untersuchungen steigt, wird eine weitere Zerreißmaschine B angeschafft. Um die Gleichwertigkeit der beiden Maschinen zu prüfen, werden 12 Drahtproben geteilt und jede Hälfte an einer Maschine getestet. Es ergeben sich die Reißlast-Messungen von

```
A: 35  46  34  27  37  59  52  61  21  31  37  27
B: 39  51  32  23  41  53  51  55  19  36  37  26

Z: -4  -5   2   4  -4   6   1   6   2  -5   0   1
```

mit einer mittleren Differenz von $\hat{\mu}_z = 0.33$ bei $\hat{\sigma}_z = 4.03$. Wir erhalten $t = 0.2865$ bei 11 Freiheitsgraden und akzeptieren die Nullhypothese ($p = 0.7798$).

Die Maschinen haben gleichwertige mittlere Reißlast-Messungen.

F-Test: Vergleich von Varianzen

Gegeben seien zwei unabhängige normalverteilte Stichproben X und Y vom Umfang n_x und n_y, deren Mittelwerte und Varianzen nicht bekannt sind.

Nullhypothese: $\sigma_x^2 = \sigma_y^2$
Alternative: $\sigma_x^2 \neq \sigma_y^2$

Wir nehmen an, daß $\hat{\sigma}_x^2 > \hat{\sigma}_y^2$ ist (sonst Umbenennung). Dann hat die Teststatistik

$$F = \frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2}$$

eine F-Verteilung mit $n_x - 1$ und $n_y - 1$ Freiheitsgraden.

F-Test: Vergleich von Varianzen

Beispiel 1c: Haben die beiden Dioden-Lieferungen dieselbe Varianz? Wir hatten geschätzte Standardabweichungen von $\hat{\sigma}_A = 8.71$ und $\hat{\sigma}_B = 12.06$ und bilden damit die Teststatistik

$$F = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_A^2} = \frac{12.06^2}{8.71^2} = 1.918239$$

Der kritische Wert der F-Verteilung mit 9 und 9 Freiheitsgraden liegt bei 4.025994, wir nehmen daher die Nullhypothese an ($p = 0.346$), obwohl die Varianz der Stichprobe B fast doppelt so groß wie die der Stichprobe A ist.

Nichtparametrische Tests

Vorzeichen-Test

- Annahmen: $X_1, \ldots, X_n$ unabhängige Wiederholungen, X besitzt stetige Verteilungsfunktion
- Hypothesen:

(a)	$H_0 : x_{med} = \delta_0$	$H_1 : \ x_{med} \neq \delta_0$
(b)	$H_0 : x_{med} \geq \delta_0$	$H_1 : \ x_{med} < \delta_0$
(c)	$H_0 : x_{med} \leq \delta_0$	$H_1 : \ x_{med} > \delta_0$

- Teststatistik: $A =$ Anzahl der Stichprobenvariablen mit einem Wert kleiner als δ_0
- Verteilung unter $x_{med} = \delta_0$: $B(n, 0.5)$, für $n \geq 25$ approximativ $N(0.5n, 0.25n)$
- Ablehnungsbereiche: Für $n \geq 25$ wie beim approximativen Binomialtest mit $\pi_0 = 0.5$. Für $n < 25$ exakter Binomialtest nötig.

Vorzeichen-Test

- Keine Annahmen über Verteilungstyp notwendig; nur: stetige Verteilungsfunktion.

  Deshalb: verteilungsfreier bzw. nonparametrischer Test
- Unter $x_{med} = \delta_0$ gilt $P(X_i < \delta_0) = 0.5$; $\Rightarrow$ $A \sim B(n, 0.5)$.

  D.h.: Vorzeichen-Test ist spezieller Binomialtest auf $\pi_0 = 0.5$.
- Falls X normalverteilt: Effizienzverlust, d.h. geringere Güte als Student-Test

Wilcoxon-Vorzeichen-Rang-Test

- Annahmen: $X_1, \ldots, X_n$ unabhängig und identisch verteilt wie X.

  X metrisch skaliert und symmetrisch verteilt. Verteilungsfunktion stetig.

- Hypothesen:

$$\begin{array}{lll} (a) & H_0 : x_{med} = \delta_0 & H_1 : \; x_{med} \neq \delta_0 \\ (b) & H_0 : x_{med} \geq \delta_0 & H_1 : \; x_{med} < \delta_0 \\ (c) & H_0 : x_{med} \leq \delta_0 & H_1 : \; x_{med} > \delta_0 \end{array}$$

Wilcoxon-Vorzeichen-Rang-Test

- Teststatistik: $W = \sum_{i=1}^{n} rg|D_i| Z_i$

  mit $D_i = X_i - \delta_0, \quad Z_i = \begin{cases} 1 & D_i > 0 \\ 0 & D_i < 0 \end{cases}$.

Für $n > 20$ ist W approximativ verteilt nach $N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$.

- Ablehnungsbereich:

$$\begin{array}{lll} (a) & W < w_{\alpha/2} & \text{oder} \quad W > w_{1-\alpha/2} \\ (b) & W < w_{\alpha} & \\ (c) & W > w_{1-\alpha}, & \end{array}$$

  wobei w_α das tabellierte α-Quantil der Verteilung von W ist.

Wilcoxon-Vorzeichen-Rang-Test

- Keine Annahmen über Verteilungstyp notwendig; nur: stetige und symmetrische Verteilungsfunktion. Deshalb: verteilungsfreier/nonparametrischer Test.

- Wegen Symmetrie: $x_{med} = E(X)$.

  $\Rightarrow$ Hypothesenpaare (a), (b), (c) identisch zum Gauß- und Student-Test

  $\Rightarrow$ Alternative zum Student-Test; keine Normalverteilungsannahme notwendig.

Wilcoxon-Vorzeichen-Rang-Test

- Zur Teststatistik W:
  1. Berechne die Differenzen $D_i = X_i - \delta_0, \; i = 1, \ldots, n$.
  2. Bilde die zugehörigen betragsmäßigen Differenzen $|D_1|, \ldots, |D_n|$.
  3. Ordne diesen betragsmäßigen Differenzen Ränge zu, d.h. der kleinste Betrag erhält den Rang 1, der zweitkleinste Betrag den Rang 2, usw..

  Bezeichnet $rg|D_i|$ den Rang von $|D_i|$, ergibt sich die Teststatistik als die Summe

$$W = \sum_{i=1}^{n} rg|D_i| Z_i \quad \text{mit} \quad Z_i = \begin{cases} 1 & \text{wenn } D_i > 0 \\ 0 & \text{wenn } D_i < 0. \end{cases}$$

  W stellt damit die Summe über alle Ränge dar, die zu Beobachtungen gehören, für die $X_i > \delta_0$, d.h. $D_i > 0$ gilt.

Wilcoxon-Vorzeichen-Rang-Test

Bei Bindungen (ties): Durchschnittsränge vergeben.

- Idee der Teststatistik:
  - Unter $x_{med} = \delta_0 \Rightarrow$ (wegen symmetrischer Verteilung) Summe der Ränge mit $D_i > 0 \approx$ Summe der Ränge mit $D_i < 0$
    $\Rightarrow E(W) = (rg(D_1) + \ldots + rg(D_n))/2 = (1 + \ldots + n)/2 = \frac{n(n+1)}{4}$
  - Ist $x_{med} < \delta_0$ bzw. $x_{med} > \delta_0$: Anzahl der i mit $X_i > \delta_0$ bzw. $X_i > \delta_0$ wird kleiner.

Wilcoxon-Vorzeichen-Rang-Test

- Verteilung von W unter $x_{med} = \delta_0$ hängt nicht von der wahren Verteilung von X ab: verteilungsfreier Test.
  Exakte Herleitung für endliches n schwierig.
  $\Rightarrow$ Tabellen für Quantile bzw. Normalverteilungsapproximation

- Geringer Effizienzverlust gegenüber Student-Test, falls X tatsächlich normalverteilt.

Wilcoxon-Rangsummen-Test

Verallgemeinerung des Vorzeichen-Rang-Tests für den Vergleich der Mediane zweier Stichproben.

Annahme: Verteilungsfunktionen F und G von X bzw. Y haben gleiche Form, sind aber möglicherweise um ein Stück gegeneinander verschoben.

Idee: Unter $H_0 : x_{med} = y_{med}$ sind F und G identisch, d.h. x- und y-Werte kommen aus der gleichen Verteilung.

$\Rightarrow$ Bilde gepoolte Stichprobe $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ und zugehörige Ränge

$$rg(X_1), \ldots, rg(Y_m)$$

(Bei Bindungen: Durchschnittsränge vergeben.)

Teststatistik: T_W = Summe der Ränge, die zu x-Werten gehören. Falls $F \neq G$: T_W zu groß und/oder zu klein.

Wilcoxon-Rangsummen-Test

Annahmen:
$X_1, \ldots, X_n$ unabhängig und identisch verteilt wie X
$Y_1, \ldots, Y_m$ unabhängig und identisch verteilt wie Y
$X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ unabhängig
X und Y besitzen stetige Verteilungsfunktion F bzw. G,
Verteilung von $X - Y$ ist symmetrisch.

Hypothesen:

$$\begin{array}{lllll} (a) & H_0 : x_{med} = y_{med} & \text{vs.} & H_1 : x_{med} \neq y_{med} \\ (b) & H_0 : x_{med} \geq y_{med} & \text{vs.} & H_1 : x_{med} < y_{med} \\ (c) & H_0 : x_{med} \leq y_{med} & \text{vs.} & H_1 : x_{med} > y_{med} \end{array}$$

Wilcoxon-Rangsummen-Test

- Teststatistik:

$$T_W = \sum_{i=1}^{n} rg(X_i)$$

- Ablehnungsbereiche:

  (a) $T_W < w_{\alpha/2}(n,m)$ oder $T_W > w_{1-\alpha/2}(n,m)$
  (b) $T_W < w_{\alpha}(n,m)$
  (c) $T_W > w_{1-\alpha}(n,m)$

  wobei $w_{\tilde{\alpha}}$ das tabellierte $\tilde{\alpha}$-Quantil der Verteilung von T_W ist.

- Für $m > 25$ oder $n > 25$ ist die Teststatistik approximativ

$$N\left(\frac{n(n+m+1)}{2}, \frac{nm(n+m+1)}{12}\right)$$

  verteilt, sonst Tabelle.

Der χ^2-Test

Der χ^2-Test

Der χ^2-Test ist sehr flexibel und wird in vielen Bereichen eingesetzt: Immer wenn beobachtete diskrete Ereignisse mit theoretischen Wahrscheinlichkeiten verglichen werden sollen, kann nach dem Prinzip

$$X^2 = \sum \frac{(\text{beobachtet} - \text{erwartet})^2}{\text{erwartet}}$$

verfahren werden. Nach einer Einteilung in Klassen (wie in einem Histogramm) kann er auch zum Vergleich kontinuierlicher Größen verwendet werden.

Sind $X_1, \ldots, X_k$ unabhängig standardnormalverteilt, so ist

$$\chi^2 = X_1^2 + X_2^2 + \ldots + X_k^2$$

χ^2-verteilt mit k Freiheitsgraden.

χ^2-Test: Eine Stichprobe

(χ^2-Anpassungstest)

Stichprobe der Größe n einer Variablen mit k Merkmalen, beobachte Häufigkeiten $h_1, \ldots, h_k$. Test ob wahre Wahrscheinlichkeiten gleich $\pi_1, \ldots, \pi_k$:

$$X^2 = \sum_{i=1}^{k} \frac{(h_i - n\pi_i)^2}{n\pi_i}$$

Falls die π_i vorgegeben sind (oder aus anderen Daten ermittelt wurden), hat man $k-1$ Freiheitsgrade. Wurden die π_i mittels Maximum Likelihood geschätzt und dabei r Parameter verwendet, so hat man nur mehr $k-r-1$ Freiheitsgrade.

Beispiel: Zufallszahlen

Ein Computerprogramm soll gleichverteilte Zufallszahlen erzeugen, und zwar die natürlichen Zahlen 0 bis 9. Der Programmierer erhält nach einem Testlauf von 10000 Versuchen folgende Häufigkeiten:

0	1	2	3	4	5	6	7	8	9
956	998	1043	1059	968	985	1087	1042	967	895

$$X^2 = \frac{(956-1000)^2}{1000} + \ldots + \frac{(895-1000)^2}{1000} = 29.966$$

ergibt bei 9 Freiheitsgraden $p = 0.0004446$, das Programm funktioniert mit großer Wahrscheinlichkeit nicht.

Beispiel: Sonntagsfrage

Vier Wochen vor der Nationalratswahl 1999 wurde 499 Haushalten die „Sonntagsfrage“ gestellt: Falls nächsten Sonntag Wahlen wären, welche Partei würden Sie wählen?

	SPÖ	ÖVP	FPÖ	Grüne	LIF	Sonst
Umfrage	38%	24%	25%	6%	4%	3%
Wahl	33.15%	26.91%	26.91%	7.4%	3.65%	1.98%
Umfrage	190	120	125	30	20	14
Erwartet	165.41	134.28	134.28	36.92	18.21	9.88

Frage 3: War das Gesamtergebnis überraschend?

$$X^2 = \frac{(190-165.41)^2}{165.41} + \ldots + \frac{(14-9.88)^2}{9.88} = 9.0053$$

ergibt bei 5 Freiheitsgraden $p = 0.1089$.

χ^2-Homogenitätstest

Ziel: Test auf Gleichheit der Verteilungen von zwei oder mehr Variablen $X_1, X_2, \ldots, X_k$. Meist: X_i Merkmal X in i-ter Population oder unter i-ter Versuchsbedingung.

X jeweils entweder kategorial mit m Kategorien oder klassiert in m Klassen.

		Merkmalsausprägungen			
		1	$\ldots$	m	
Population	1	h_{11}	$\ldots$	h_{1m}	n_1
	2	h_{21}	$\ldots$	h_{2m}	n_2
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	k	h_{k1}	$\ldots$	h_{km}	n_k
		$h_{\cdot 1}$	$\ldots$	$h_{\cdot m}$	

χ^2-Homogenitätstest: Bsp

Beispiel: Kreditwürdigkeit

X_1 Kontostand ($m = 3$) bei guten Krediten ($n_1 = 700$)

X_2 Kontostand bei Problemkrediten ($n_2 = 300$)

		Konto			
		nein	gut	mittel	
Kreditwürdigkeit	unproblematisch	139	348	213	700
	Problem	135	46	119	300
		274	394	332	1000

χ^2-Homogenitätstest

- Idee: Unter $H_0 : P(X_1 = j) = \ldots = P(X_k = j)$ für $j = 1, \ldots, m$ sind die Verteilungen identisch.

  $\Rightarrow \frac{h_{\cdot j}}{n}$ Schätzer für $P(X_i = j)$.

  Da $h_{ij} \sim B(n_i, P(X_i = j))$ und $E(h_{ij}) = n_i P(X_i = j)$
  $\Rightarrow \tilde{h}_{ij} = n_i \frac{h_{\cdot j}}{n}$ erwartete Häufigkeit von h_{ij} unter H_0

- Teststatistik χ^2 vergleicht h_{ij} und $\tilde{h}_{ij}$ für alle i, j.

χ^2-Homogenitätstest

Definition: χ^2-Homogenitätstest für k Stichproben

- Annahmen: Unabhängige Stichprobenziehung in den k Populationen
- Hypothesen: $H_0 : \; P(X_1 = j) = \cdots = P(X_k = j), \quad j = 1, \ldots, m$

  $H_1 : \; P(X_{i_1} = j) \neq P(X_{i_2} = j)$ für mindestens ein Tupel (i_1, i_2, j)
- Teststatistik: $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(h_{ij} - \frac{n_i h_{\cdot j}}{n}\right)^2}{\frac{n_i h_{\cdot j}}{n}}$
- Verteilung unter H_0: approximativ $\chi^2((k-1)(m-1))$
- Ablehnungsbereich: $\chi^2 > \chi^2_{1-\alpha}((k-1)(m-1))$

χ^2-Homogenitätstest: Bsp

Tabelle der zu erwartenden Häufigkeiten $\tilde{h}_{ij}$

		Konto			
		nein	gut	mittel	
Kreditwürdigkeit	unproblematisch	191.80	275.80	232.40	700
	Problem	82.20	118.20	99.60	300
		274	394	332	1000

$\Rightarrow \quad \chi^2 = 116.851 > \chi^2_{0.95}(2) = 5.99 \quad \Rightarrow \quad H_0$ ablehnen

Zusammenhangsanalyse

Zusammenhangsanalyse

Gegeben: Paare (X_i, Y_i), $i = 1, \ldots, n$, i.i.d. wie (X, Y)

Möglichkeiten:

1. X und Y kategorisch (oder metrisch und klassiert):
   Test auf Unabhängigkeit

2. X und Y metrisch: Test auf Korrelation

3. X metrisch und Y kategorisch: Verallgemeinerung t-Test bzw. Wilcoxon-Rangsummentest für mehr als 2 Gruppen
   $\rightarrow$ Varianzanalyse, Kruskal-Wallis-Test

Zusammenhangsanalyse in 1. und 2. unterstellt keine Wirkungsrichtung, d.h. X und Y werden symmetrisch behandelt.

Zusammenhangsanalyse

Beispiel: Sonntagsfrage

	CDU/CSU	SPD	FDP	Grüne	Rest	
Männer	144	153	17	26	95	435
Frauen	200	145	30	50	71	496
insgesamt	344	298	47	76	166	931

Frage: Geschlecht und Parteipräferenz abhängig?

χ^2-Unabhängigkeitstest

Gegeben: Stichprobenvariablen (X_i, Y_i), $i = 1, \ldots, n$
Hypothesen:

$H_0 : P(X = i, Y = j) = P(X = i) \cdot P(Y = j)$ für alle i, j
$H_1 : P(X = i, Y = j) \neq P(X = i) \cdot P(Y = j)$ für mind. ein Paar (i, j)

		Y: 1	$\ldots$	m	
X	1	h_{11}	$\ldots$	h_{1m}	$h_{1\cdot}$
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	k	h_{k1}	$\ldots$	h_{km}	$h_{k\cdot}$
		$h_{\cdot 1}$	$\ldots$	$h_{\cdot m}$	n

unter H_0 $\longrightarrow$

		Y: 1	$\ldots$	m	
X	1	$\frac{h_1 h_1}{n}$	$\ldots$	$\frac{h_1 h_m}{n}$	$h_{1\cdot}$
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	k	$\frac{h_k h_1}{n}$	$\ldots$	$\frac{h_k h_m}{n}$	$h_{k\cdot}$
		$h_{\cdot 1}$	$\ldots$	$h_{\cdot m}$	n

Teststatistik: $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$ mit $\tilde{h}_{ij} = \frac{h_{i\cdot} h_{\cdot j}}{n}$

Ablehnungsbereich: $\chi^2 > \chi^2_{1-\alpha}((k-1) \cdot (m-1))$

χ^2-Unabhängigkeitstest

Beispiel: Sonntagsfrage

Berechnung von χ^2 analog zur deskriptiven Statistik:
$\chi^2 = 20.065$

$(k-1)(m-1) = 1 \cdot 4 = 4$
$\chi^2_{0.95}(4) = 9.488$

$20.065 > 9.488 \Rightarrow H_0$ bei $\alpha = 5\%$ ablehnen, d.h. signifikanter Zusammenhang zwischen Geschlecht und Parteipräferenz.

Korrelationstest

Annahmen: Unabhängige gemeinsam normalverteilte Stichprobenvariablen (X_i, Y_i), $i = 1, \ldots, n$ (sonst Rangkorrelation, komplizierter).

Hypothesen:

$$(a) \quad H_0 : \rho_{XY} = 0 \quad \text{vs.} \quad H_1 : \rho_{XY} \neq 0$$
$$(b) \quad H_0 : \rho_{XY} \geq 0 \quad \text{vs.} \quad H_1 : \rho_{XY} < 0$$
$$(c) \quad H_0 : \rho_{XY} \leq 0 \quad \text{vs.} \quad H_1 : \rho_{XY} > 0$$

Teststatistik:
$$T = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sqrt{n-2}$$

Ablehnungsbereiche:

$$(a) \quad |T| > t_{1-\frac{\alpha}{2}}(n-2)$$
$$(b) \quad T < -t_{1-\alpha}(n-2)$$
$$(c) \quad T > t_{1-\alpha}(n-2)$$

Regelkarten und wiederholte Tests

Beispiel: Fertigungsüberwachung

Beispiel 4: Bei der Fertigung von Bildröhren soll der Kathodenstrom überwacht werden. Bei ungestörter Fertigung ist der Kathodenstrom normalverteilt mit $\mu = 25mA$ und $\sigma = 1mA$. In regelmäßigen Abständen wird eine Röhre entnommen und vermessen.

Wegen der Normalverteilungsannahme sollten die gemessenen Kathodenströme mit

- 95% Wahrscheinlichkeit im Intervall 25 ± 1.96,
- 99% Wahrscheinlichkeit im Intervall 25 ± 2.57

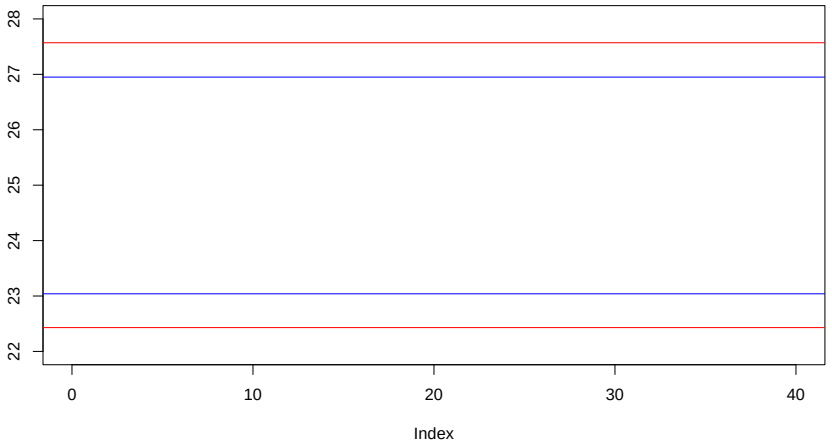
liegen.

Zur einfachen Handhabung derartiger Tests werden sogenannte Qualitätsregelkarten eingesetzt.

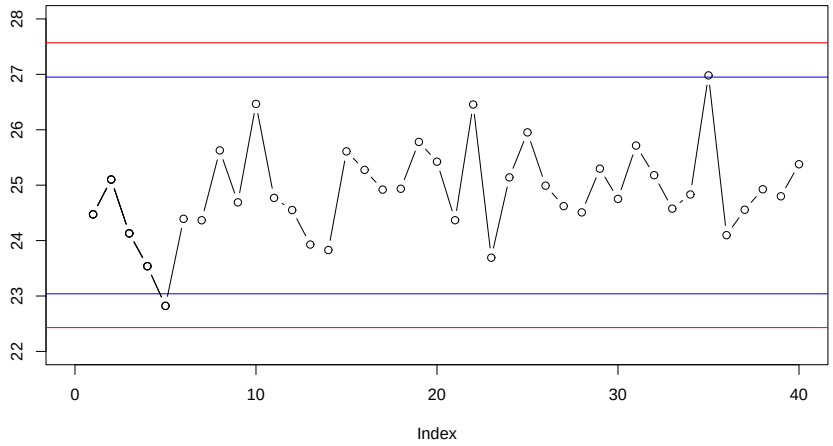
Kritische Werte

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	[illegible]	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	[illegible]	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	[illegible]	[illegible]	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	[illegible]	[illegible]	0.9964

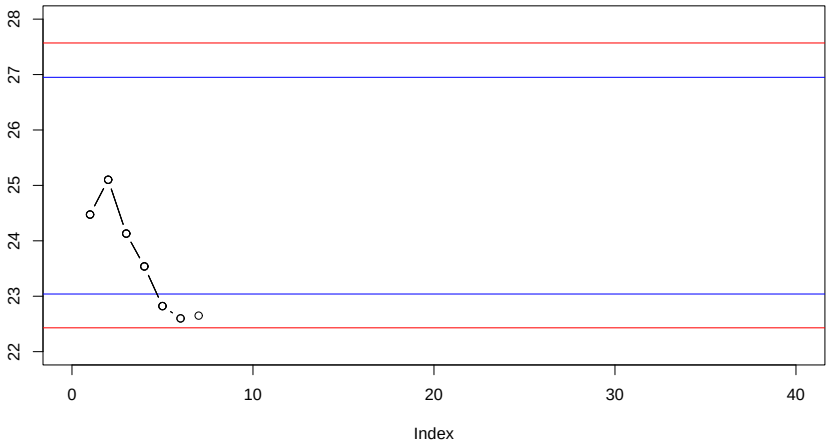
Regelkarte



Regelkarte



Regelkarte



Wiederholte Tests

Wie groß ist die Wahrscheinlichkeit, daß zwei aufeinander folgende Werte im blauen Bereich liegen?

Im Beispiel mit $\mu = 25mA$ und $\sigma = 1mA$:

$$\begin{aligned} \mathbb{P}\{|X_t - 25| > 1.96, |X_{t+1} - 25| > 1.96\} &= \\ \mathbb{P}\{|X_t - 25| > 1.96\}\mathbb{P}\{|X_{t+1} - 25| > 1.96\} &= \\ (1 - 0.95)(1 - 0.95) = 0.05^2 &= 0.0025 \end{aligned}$$

Drei aufeinanderfolgende Beobachtungen sollten nur mehr alle 8000 Messungen im blauen Bereich liegen:

$$0.05^3 = 0.000125$$

Wiederholte Tests

Ganz analog dazu wird auch wiederholtes Anwenden eines t-Tests oder χ^2-Tests irgendwann zu einem Fehler 1. Art führen. Falls k Tests gemeinsam eine Größe von α haben sollen, müssen wir das Niveau $\tilde{\alpha}$ der einzelnen Tests korrigieren.

$$\begin{aligned} \alpha &= 1-(1-\tilde{\alpha})^k \\ &= 1-\left(1-\binom{k}{1}\tilde{\alpha}+\binom{k}{2}\tilde{\alpha}^2-\ldots\right) \\ &\approx 1-1+k\tilde{\alpha}=k\tilde{\alpha} \end{aligned}$$

Bonferroni-Korrektur: $\tilde{\alpha}=\alpha/k$

Beispiel: Falls 4 Tests gemeinsam ein Niveau von $\alpha=0.05$ haben sollen, müssen die kritischen Werte zum Niveau $\tilde{\alpha}=0.0125$ verwendet werden.

Wiederholte Tests

Beispiel 1d: Getrennte t-Tests der beiden Dioden-Lieferungen A und B auf einen Mittelwert von 100 ergeben p-Werte von 0.003080 für A und 0.003780.

Zur Bonferroni-Korrektur multiplizieren wir die p-Werte einfach mit $k=2$ und erhalten korrigierte p-Werte von 0.00616 und 0.00756. Die Nullhypothese wird also auch nach der Korrektur für multiples Testen in beiden Fällen verworfen.