Wintersemester 2004 / 2005

Epidemiologie / Biometrie

Robert Hochstrat

14. März 2005

Zusammenschrift der Übung zur Vorlesung aus dem WS 04/05
Rückfragen, Ergänzungen und Korrekturen an robert_hochstrat@web.de
Natürlich übernehme ich keine Gewähr für Richtig- und Vollständigkeit!

Kenngrößen

- Lagemaße geben in geeigneter Weise ein Zentrum der Häufigkeitsverteilung an
- Streuungsmaße kennzeichnen die Variabilität der Häufigkeitsverteilung

	Zulässige Lagemaße	Zulässige Streuungsmaße
Nominal	Modus	keine
Ordinal	Modus Quantile	Spannweite Interquartilabstand
Metrisch	Modus Quantile Mittelwert	Spannweite Interquartilabstand Standardabweichung Varianz Variationskoeffizient

- Modus: Die Merkmalsausprägung, die am häufigsten beobachtet wird
- Quantile:

$$Q_p = X_{\lceil p \cdot n \rceil} \text{ für } p \cdot n \in \mathbb{N} \text{ und}$$
$$Q_p = \tfrac{1}{2} \cdot (X_{p \cdot n} + X_{(p \cdot n)+1}) \text{ für } p \cdot n \notin \mathbb{N}$$

  Das $p \cdot 100\%-$ Quantil teilt die Daten im Verhältnis $p \cdot 100\% : (1-p) \cdot 100\%$

- Mittelwert: S. Regressionsrechnung
- Median: 50% - Quantil
- Spannweite: $x_{max} - x_{min}$
- Interquartilabstand: $|Q_{\frac{3}{4}} - Q_{\frac{1}{4}}|$

Boxplott:

Wiskers ("Antennen") bis 1,5-facher Interquartilabstand, Box von $Q_{\frac{1}{4}}$ bis $Q_{\frac{3}{4}}$, Median wird durch Querlinie und Mittelwert durch Punkt gekennzeichnet.

Varianz

s. Rechnung zur Regressionsgeraden

Standardabweichung

Die Standardabweichung ist die Wurzel aus der Varianz.

Kovarianz

Die Kovarianz beschreibt gemeinsame Streuung zweier Merkmale Analog zur Varianz (S_{xx} bzw. S_{yy}):

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Korrelation

Korrelation ist ein Maß für die Stärke des Linearen Zusammenhangs zweier Merkmale.

Korrelationskoeffizient

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Lineare Regression

Wenn Größen Y linear von Größen X abhängen, kann man eine Regressionsgerade $Y = a \cdot X + b$ aufstellen.

a: Regressionskoeffizient

b: Achsenabschnitt.

Jeder Punkt (x_i, y_i) hat von der Geraden dem Abstand $d_i = (y_i - (a \cdot x_i + b))$

Idee: Bilde alle Differenzen, quadriere und minimiere ihre Summe bezüglich a und b. Die dadurch eindeutig bestimmte Gerade ist die Regressionsgerade von Y auf X.

Achtung: Nicht über Beobachtungsgrenzen hinaus Extrapolieren!

Rechnung zur Regression:

- Mittelwerte:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- Varianzen:

$$S_{xx} = \frac{1}{n-1} \cdot [\sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2] = \frac{1}{n-1} \cdot [\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}]$$

$$S_{yy} = \frac{1}{n-1} \cdot [\sum_{i=1}^{n} y_i^2 - \frac{1}{n} (\sum_{i=1}^{n} y_i)^2] = \frac{1}{n-1} \cdot [\sum_{i=1}^{n} y_i^2 - n \cdot \bar{y}]$$

- Kovarianz: (s. auch oben)

$$S_{xy} = S_{yx} = \frac{1}{n-1} \cdot [\sum_{i=1}^{n} x_i y_i - \frac{1}{n} (\sum_{i=1}^{n} x_i) \cdot (\sum_{i=1}^{n} y_i)] = \frac{1}{n-1} \cdot [\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}]$$

- Steigung der Regressionsgeraden:

$$b_{yx} = \frac{S_{yx}}{S_{xx}}$$

- Achsenabschnitt (Intercept) der Regressionsgeraden:

$$a_{yx} = \bar{y} - b_{yx} \bar{x}$$

- Regressionsgerade:

$$Y = b_{yx} \cdot X + a_{yx}$$

Güte der Regression:

$$B = r_{xy}^2$$

$0 \leq B \leq 1$ mit r_{xy} Korellationskoeffizient

r_{xy}^2 % der Varianz von Y ist durch X erklärbar.

Konfidenzintervalle

Ein KI ist ein Zufallsintervall, das mit einer vorgegebenen Wahrscheinlichkeit $(1-\alpha)$ einen festen, aber unbekannten Parameter enthält.
$(1-\alpha)$: Konfidenzniveau
α : Irrtumswahrscheinlichkeit

Konfidenzintervall für p

approximatives KI für für den Parameter p einer Binomialverteilung ist gegeben durch $[p_1, p_2]$ mit $p_{1/2} = \frac{k}{n} \pm u_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\frac{k}{n} \cdot (1-\frac{k}{n})}{n}}$ mit u, dem Quantil der Standardnormalverteilung.

Konfidenzintervall für μ

μ ist Erwartungswert einer Normalverteilung.

- bekannte Varianz: $\mu_{1/2} = \bar{x} \pm u_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}$
- geschätzte Varianz: $\mu_{1/2} = \bar{x} \pm t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{s}{\sqrt{n}}$
  mit s Standardabweichung

Statistische Tests

- Ein statistischer Test soll herausfinden, ob ein bestimmtes Modell gelten kann.
- Das zu prüfende Modell wird Nullhypothese genannt.
- Das Modell soll abgelehnt, wenn ein ein externes Ereignis beobachtet wird, das im Modell eine genügend kleine Wahrscheinlichkeit hat.
- Der Test ist eine Regel, die festlegt, wann diese Frage mit plausibel bzw. unplausibel beantwortet wird.

Signifikanztests

Ziel eines Signifikanztests: Etablierung einer neuen Hypothese, Nachweis eines Unterschieds zu einer bestehenden Hypothese.
Daher:
$-H_0$: Nullhypothese, Basis der Entscheidung
$-H_1$: Alternativhypothese
Vorgehen Annahme: H_1 unterscheidet sich nicht von H_0. Liegt dann jedoch die Beobachtung im unplausiblen Bereich, gilt das als Widerspruch und statistischer Beweis für den Unterschied zwischen H_0 und H_1.

	H_0	H_1
H_0	Richtige Entscheidung	Fehler 2. Art Wahscheinlichkeit β
H_1	Fehler 1. Art Wahrscheinlichkeit α	Richtige Entscheidung

1 - β : Power = Wahrscheinlichkeit, mit der ein vorhandener Unterschied aufgedeckt wird.

Fehler 1. und 2. Art

- Wahrscheinlichkeit für α kann frei gewählt werden. Üblich sind $5\%, 1\%$ und $0,1\%$
- Wahl $\alpha = 0,05$ bedeutet, dass man bereit ist, 5% der richtigen H_0 zu Gunsten von H_1 zu verwerfen.
- α heißt auch Signifikanzniveau bzw. Irrtumswahrscheinlichkeit.
- $\beta-$Fehler läßt sich nicht abschätzen, da Verteilung unter H_1 unbekannt ist.
- Wird H_0 zugunsten der Alternativhypothese verworfen, so ist das Testergebnis, dass der Unterschied signifikant ist.

Ablaufschema "Statistische Signifikanztests"

1. Formulierung der Hypothese
2. Festlegung des Signifikanzniveaus α
3. Zusammenfassung der Stichproben zu einem Wert $T = T(x_1, x_2, .., x_n)$
4. Ablesen eines Schwellenwertes c aus einer Tabelle (Quantil)
5. Entscheidung:
   - falls $T > c$ bzw. $p \leq \alpha : H_1$
   - falls $T \leq c$ bzw. $p < \alpha$: Nichtablehnung von H_0

χ^2-Test

Überprüfung der (Un-)Abhängigkeit von zwei Merkmalen
H_0 :Unabhängig
H_1 :Abhängig
Ablauf:

1. beobachtete Häufigkeiten

	$X=0$	$X=1$	
$H=0$	n_{11}	n_{12}	$n_{1\cdot}$
$G=1$	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

2. Erwartungen unter H_0 : $\frac{n_{11}}{n_{\cdot 1}} = \frac{n_{1\cdot}}{n_{\cdot\cdot}} \Leftrightarrow n_{11} = \frac{n_{\cdot 1} \cdot n_{1\cdot}}{n_{\cdot\cdot}}$
3. erwartete Häufigkeiten:

	$X=0$	$X=1$	
$H=0$	m_{11}	m_{12}	$n_{1\cdot}$
$G=1$	m_{21}	m_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

mit $m_{ij} = \frac{n_{j\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$

4. bestimme für jede Zelle i: $q_i = \frac{(Beobachteter\ Wert - Erwarteter\ Wert)^2}{Erwarteter\ Wert}$
5. Bestimme Teststatistik $Q = \sum_{Zellen} q_i$
6. Entscheidung: $Q > \chi^2_{(z-1)(s-1);(1-\alpha)} \Rightarrow H_0$ ablehnen.
   Mit z = Anzahl der Zeilen und s = Anzahl der Spalten in der zugehörigen Kontingenztafel (ohne Summen!) und $(1-\alpha)$ Signifikanzniveau.

t-Test

Zwei Stichproben heißen unverbunden, wenn sowohl die Daten innerhalb einer Stichprobe als auch die Daten aus beiden Stichproben zusammen alle unabhängig voneinander sind.
Zwei Stichproben heißen verbunden, wenn es zu jedem x aus der einen Stichprobe genau ein y aus der anderen Stichprobe gibt, mit dem es inhaltlich ein Paar bildet. Verbundene Stichproben müssen daher stets den gleichen Stichprobenumfang haben. Man nennt zwei verbundene Stichproben auch paarige Stichproben.

- t-Test für unverbundene Stichproben:
  - Zwei unabhängige Stichproben, die Realisationen von normalverteilten Zufallsvariablen mit unbekannter Varianz darstellen.
  - Hypothese $H_0 : \mu_1 = \mu_2$ gegen $H_1 : \mu_1 \neq \mu_2$
  - Teststatistik:
    $$T = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s^2}{n}+\frac{s^2}{m}}}$$
    mit der "gepoolten Varianz" $s^2 = \frac{1}{n+m-2} \cdot [(n-1) \cdot S_{xx} + (m-1) \cdot S_{yy}]$
  - Entscheidungsregel: H_1 annehmen, falls $|T| > t_{(n+m-2;1-\frac{\alpha}{2})}$
- t-Test für verbundene Stichproben:
  - Zwei Stichproben von Meßwertpaaren (x_i, y_i), die aus Grundgesamtheiten mit den Erwartungswerten μ_1 und μ_2 mit gleichen Varianzen stammen.
  - Die Differenzen $d_i = x_i - y_i$ sind Realisationen einer normalverteilten Zufallsvariablen D mit unbekannter Varianz
  - Hypothesen: $H_0 : \mu_1 = \mu_2$ gegen $H_1 : \mu_1 \neq \mu_2$
  - Teststatistik:
    $$T = \sqrt{n} \cdot \frac{|\bar{d}|}{\sqrt{S_{dd}}}$$
  - Entscheidungsregel: H_1 annehmen, falls $T > t_{(n-1;1-\frac{\alpha}{2})}$

Diagnostische Tests

	Realität	
Testentscheidung lautet:	krank	gesund
positiv (krank)	richtige Entscheidung	falsch-positiv: "Fehler 1. Art"
negativ (gesund)	falsch-negativ: "Fehler 2. Art"	richtige Entscheidung

bekannt sind:

- Prävalenz: P(krank) = P(Diagnose positiv), kurz: $P(D^+) = \frac{Summe\ der\ Erkrankten}{Gesamtzahl\ aller\ Versuchsobjekte}$
- Sensitivität("%-Zahl der Kranken, die als krank erkannt wird"):
  $P(T^+|D^+) = \frac{Anzahl\ der\ Erkrankten\ mit\ positivem\ Test}{Gesamtzahl\ der\ Erkrankten}$
- Spezifität("%-Zahl der Gesunden, die als gesund erkannt wird"):
  $P(T^-|D^-) = \frac{Anzahl\ der\ Gesunden\ mit\ negativem\ Test}{Gesamtzahl\ der\ Gesunden}$

Bedingung ist, dass die Diagnose bekannt ist.
Vorhersagewerte:

- Positiver Vorhersagewert: $P(D^+|T^+) = \frac{Positiv\ getestete\ kranke}{Gesamt\ positiv\ getestete}$
- Negativer Vorhersagewert $P(D^-|T^-) = \frac{Negtiv\ getestete\ Gesunde}{Gesamt\ negativ\ getestete}$

Hier ist der positive Test die Bedingung.

Von Interesse sind auch:

- A-priori-Odds = $\frac{Praevalenz}{1-Praevalenz}$
- A-posteriori-odds = a-priori-odds $\cdot LR^+$
- Likelihood Quotienten
  - $LR^+ = \frac{Sensitivitaet}{1-Spezifitaet}$
    "Ein positives Testergebnis ist LR^+ mal wahrscheinlicher, wenn eine Erkrankung vorliegt, als wenn keine vorliegt."
  - $LR^- = \frac{1-Sensitivitaet}{Spezifitaet}$
- In prospektiven Studien ist das Risiko bzw. das relative Risiko auch interessant: So ist zum Beispiel das Risiko interessant, Krank zu werden, wenn man einer Gefahr ausgesetzt wurde. $R_1 = \frac{Krank\ und\ Exposition}{Expositiongesamt}$ oder das Risiko, krank zu werden, ohne Exposition: $R_2 = \frac{Krank\ und\ keine\ Exposition}{keine\ Exposition\ gesamt}$
- Das relative Risiko ($\frac{R_1}{R_2}$) sagt aus, wie viel wahrscheinlicher es ist, zu erkranken, wenn man der Exposition ausgesetzt ist/war.
- Odds (=Chancen): $O_1 = \frac{Krank\ und\ Exposition}{Gesund\ und\ Exposition}$ $O_2 = \frac{Krank\ und\ keine\ Exposition}{Gesund\ und\ keine\ Exposition}$
- Odds-Ratio: $\frac{O_1}{O_2}$

ROC-Kurven

"Receiver Operation Characteristic" - Kurven helfen bei der Wahl eines Schwellenwertes für das Testverfahren, das zwischen "krank" und "gesund" entscheidet.
Bei der ROC-Kurve werden auf der X-Achse die Werte von (1-Spezifität) und auf der Y-Achse die zugehörigen Werte der Sensitivität für alle zu betrachtenden Variationen des Schwellenwertes. Diese Punkte verbindet man nun untereinander und die äußeren jeweils mit dem Ursprung bzw. dem Punkt (1,1). Der Schwellenwert ist als der Punkt auf dieser Kurve zu wählen, der am weitesten von der Winkelhalbierenden des Koordinatensystems entfernt ist.

Studientypen

Fall-Kontroll-Studie Unter einer Fall-Kontroll-Studie versteht man eine retrospektive Erhebung. Aus einer definierten Grundgesamtheit wird eine Stichprobe von Personen mit der interessierenden Erkrankung (Fälle) gezogen. Aus der gleichen Grundgesamtheit wird eine Stichprobe von Personen ohne diese Erkrankung (Kontrollen) gezogen. Die Exposition in der Vergangenheit gegenüber potentiellen Risikofaktoren wird ermittelt. Das wichtigste Risikomaß in Fall-Kontroll-Studien ist die Odds Ratio.
Unter einer Kohortenstudie versteht man eine prospektive Erhebung. Aus einer definierten Grundgesamtheit wird eine Stichprobe (Kohorte) gezogen, in der Risikofaktoren erhoben und Krankheiten erfaßt werden. Risikomaße in Kohortenstudien sind die Inzidenz, das Relative Risiko und die Risikodifferenz.