Biometrie und Epidemiologie - VO SS 2008
Zusammenfassung

Harald Köstinger
Student an der Technischen Universität Wien

Contents

1 Theoretische Grundlagen (S. 19) . . . 3
1.1 Merkmale (S. 21) . . . 3
1.1.1 Grundbegriffe (S. 20) . . . 3
1.1.2 Ziel- und Einflussgrößen (S. 22) . . . 3
2 Die Beschreibung eines Merkmals (S. 53) . . . 4
2.1 Lagemaße (S. 55) . . . 4
2.1.1 Mittelwert $\bar{x}$ (S. 55) . . . 4
2.1.2 Median / Rangliste $\tilde{x}$ (S. 57) . . . 4
3 Wahrscheinlichkeiten in der Medizin (S. 101) . . . 4
3.1 Wiederholung - Begriffe . . . 4
3.1.1 Zufallsexperiment . . . 4
3.1.2 Ereignisraum, Ereignisse, Elementarereignis . . . 5
3.1.3 Wahrscheinlichkeit . . . 5
3.1.4 Satz von der totalen Wahrscheinlichkeit . . . 5
3.1.5 Additionssatz . . . 5
3.1.6 Bedingte Wahrscheinlichkeit . . . 5
3.2 Wahrscheinlichkeiten in der Epidemiologie (S. 111) . . . 6
3.2.1 Prävalenz (Punktprävalenz) . . . 6
3.2.2 Periodenprävalenz . . . 6
3.2.3 Inzidenz . . . 6
3.2.4 Krankheitsspezifische Mortalität . . . 6
3.2.5 Letalität . . . 6
3.2.6 Kontagionsindex . . . 7
3.2.7 Manifestationsindex . . . 7
3.3 Bevölkerungsstatistiken (S. 114) . . . 7
3.3.1 Spezielle Wahrscheinlichkeiten (S. 114) . . . 7
3.3.2 Sterbetafeln (S. 115) . . . 7
3.4 Diagnostische Tests (S. 118) . . . 8
3.4.1 Gütekriterien . . . 8
3.4.2 Vorhersagewerte (S. 119) . . . 9
4 Diskrete Verteilungen (S. 125) . . . 9
4.1 Die Binomialverteilung (S. 129) . . . 9

4.2 Die Poisson-Verteilung . . . 10
4.3 Die Polynomialverteilung (S. 138) . . . 11
4.4 Die negative Binomialverteilung (S. 139) . . . 11
4.5 Die hypergeometrische Verteilung (S. 140) . . . 11
5 Stetige Verteilungen (S. 145) . . . 12
5.1 Die Normalverteilung (S. 148) . . . 12
5.1.1 Die Standardnormalverteilung (S. 149) . . . 12
5.1.2 σ-Bereiche, Referenzbereiche (S. 151) . . . 13
5.1.3 Normalisierende Transformationen (S. 152) . . . 13
5.2 Sätze der Wahrscheinlichkeitsrechnung (S. 155) . . . 14
5.2.1 Die Tschebyscheff'sche Ungleichung (S. 155) . . . 14
5.2.2 Das Gesetz der großen Zahlen (S. 157) . . . 14
5.2.3 Der zentrale Grenzwertsatz . . . 15
5.2.4 Bedeutung der NV . . . 15
5.3 Die Verteilung von Überlebenszeiten (S. 162) . . . 15
5.3.1 Wichtige Begriffe . . . 15
5.3.2 Die Exponentialverteilung (S. 164) . . . 18
5.3.3 Weibull-Verteilung (S. 165) . . . 18
5.4 Prüfverteilungen (S. 166) . . . 19
5.4.1 Die Chi^2 Verteilung (S. 168) . . . 19
5.4.2 Die F-Verteilung (S. 170) . . . 19
6 Schätzverfahren (S. 173) . . . 20
6.1 Punktschätzungen (S. 173) . . . 20
6.1.1 Kriterien zur Güte einer Schätzung (S. 174) . . . 20
6.1.2 Spezielle Schätzfunktionen (S. 175) . . . 21
7 Tests zum Vergleich von Häufigkeiten (S. 229) . . . 21
7.1 Der Binomialtest für eine Stichprobe (S. 229) . . . 21
7.2 Chi^2-Tests (S. 231) . . . 21
7.2.1 Der Chi^2-Vierfelder Test (S. 231) . . . 21
7.2.2 Der Chi^2-Test für $k * l$ Felder (S. 236) . . . 21
7.2.3 Assoziationsmaße für qualitative Merkmale (S. 237) . . . 21
7.2.4 Der McNemar-Test (S. 238) . . . 22
7.2.5 Der Logranktest (S. 242) . . . 22
7.3 Ausblick auf die logistische Regression (S. 245) . . . 23
8 Risikostudien (S. 261) . . . 23
8.1 Deskriptive Studien (S. 264) . . . 23
8.1.1 Fallberichte (S. 264) . . . 23
8.1.2 Fallserien (S. 265) . . . 23
8.1.3 Prävalenzstudien (S. 266) . . . 23
8.1.4 Populationsstudien (S. 266) . . . 23
8.2 Fall-Kontroll-Studien (S. 267) . . . 23
8.2.1 Matchen (S. 268) . . . 24
8.2.2 Biasquellen (S. 269) . . . 24
8.2.3 Odds-Ratio (S. 271) . . . 24
8.3 Kohortenstudien (S. 272) . . . 25
8.3.1 Effektmaße (S. 273) . . . 25

8.3.2 Biasquellen (S. 276) . . . 25
9 Studien zu Therapie und Prognose . . . 25
9.1 Prognosestudien (S. 306) . . . 25
9.1.1 Begriffe . . . 25
9.1.2 Die Kaplan-Meier-Methode (S. 307) . . . 26
10 Übersicht . . . 27
10.1 Verteilungen und deren Annäherungen . . . 27
10.2 Tests . . . 30

Bemerkung Die Seitenangaben beziehen sich auf das Buch 'Basiswissen Medizinische Statistik, Christel Weiß, 4. Auflage'

1 Theoretische Grundlagen (S. 19)

1.1 Merkmale (S. 21)

1.1.1 Grundbegriffe (S. 20)

Untersuchungseinheiten sind die Personen oder Objekte einer Stichprobe. (auch Merkmalsträger)

Beobachtungseinheiten sind die kleinsten Einheiten, an denen die einzelnen Beobachtungen registriert werden.

Merkmale Beobachtungseinheiten sind durch Merkmale charakterisiert - das sind Eigenschaften, die für die Fragestellung relevant sind.

Merkmalausprägungen darunter versteht man die Werte, die jedes Merkmal annehmen kann.

1.1.2 Ziel- und Einflussgrößen (S. 22)

Zweck einer Studie ist es, Erkenntnisse über eine oder mehrere Zielgrößen zu gewinnen. Merkmale, die in einem funktionalen Zusammenhang mit der Zielgröße stehen, werden als Einflussgrößen bezeichnet. Sie können aufgeteilt werden in:

- Faktoren, die erfasst und ausgewertet werden
- Störgrößen, die im Versuchsplan nicht berücksichtigt sind oder nicht erfasst werden
- Begleitmerkmale, die zwar erfasst, aber nicht ausgewertet werden (zB Nebenwirkungen bei einer klinisch-kontrollierten Studie)

2 Die Beschreibung eines Merkmals (S. 53)

2.1 Lagemaße (S. 55)

2.1.1 Mittelwert $\bar{x}$ (S. 55)

Ist das bekannteste Lagemaß. Errechnet sich aus

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

Sein Nachteil ist, dass er sich von Ausreißern in den Daten stark beeinflussen lässt und daher bei schiefen Verteilungen verzerrte Bilder bringt.

Merke Er darf nur für quantitative Merkmale verwendet werden. Vor allem bei symmetrischen, eingipfeligen Verteilungen sinnvoll. Nutzt alle Informatinen der Daten aus.

Er besitzt auch eine sog. **Minimumeigenschaft**.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - c)^2 \ \forall c \in \Re \setminus \{\bar{x}\}$$

2.1.2 Median / Rangliste $\tilde{x}$ (S. 57)

Ist der 'mittlerste' Wert in den Daten, nachdem diese sortiert wurden bzw das Mittel der beiden 'mittleren' Daten. Er ist robist gegen Ausreißer. Stimmen Mittelwert und Median ca. überein, so kann man auf eine symmetrische Verteilung schließen.

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

Merke Sinnvoll, wenn die Daten ordinal skaliert sind, bei quantitativen Merkmalen, die schief verteilt sind, bei Verdacht auf Ausreißer oder bei zensierten Daten.

Auch er hat eine **Minimumeigenschaft**

$$\sum_{i=1}^{n}|x_i - \tilde{x}| \leq \sum_{i=1}^{n}|x_i - c| \ \forall c \in \Re \setminus \{\tilde{x}\}$$

3 Wahrscheinlichkeiten in der Medizin (S. 101)

3.1 Wiederholung - Begriffe

3.1.1 Zufallsexperiment

Zufallsexperimente lassen sich wie folgt charakterisieren:

- es wird nach einer bestimmten Vorschrift durchgeführt
- es ist (zumindest prinzipiell) beliebig oft wiederholbar

- mehrere Ausgänge oder Ergebnisse sind möglich
- das Ergebnis eines einzelnen Experiments ist vorab ungewiss.

Beispiele hierfür: Würfeln, Werfen von Münzen, Erfassen der Blutgruppen, ...

3.1.2 Ereignisraum, Ereignisse, Elementarereignis

Stellt die Menge aller möglichen Ereignisse dar. Er wird mit Ω bezeichnet. Teilmengen davon nennt man Ereignisse, 1-elementige Teilmengen **Elementarereignisse**. Die leere Menge $\oslash$ wird als das unögliche Ereignis bezeichnet.

3.1.3 Wahrscheinlichkeit

Kurz ausgedrückt: die Anzahl der günstigen Ereignisse durch die Anzahl der möglichen.

$$P(A) = \frac{\text{Anzahl der günstigen Ereignisse}}{\text{Anzahl der möglichen Ereignisse}}$$

3.1.4 Satz von der totalen Wahrscheinlichkeit

Besagt, das das Ereignis A entweder zusammen mit Ereignis B oder $\bar{B}$ auftritt:

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$

3.1.5 Additionssatz

Für die Vereinigung zweier Ereignisse A und B gilt allgemein:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Wenn beide Ereignisse A und B disjunkt sind, dh $A \cap B = \oslash$, dann gilt folgendes:

$$P(A \cup B) = P(A) + P(B)$$

3.1.6 Bedingte Wahrscheinlichkeit

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Diese Formel quantifiziert die Wahrscheinlichkeit für das Eintreten des Ereignisses A eingeschränkt auf die Menge, die dem Ereignis B entspricht. Aus der bedingten Wahrscheinlichkeit folgt auch der **Multiplikationssatz**, mit dem sich die Wahrscheinlichkeit berechnen lässt, das zwei Ereignisse A und B gemeinsam eintreten:

$$P(A \cap B) = P(A|B)P(B)$$

Wenn A und B unabhängig sind, bedeutet dies, dass das Eintreten von B keinerlei Einfluss auf das Eintreten von A hat. ($P(A|B) = P(A)$) Damit erhält man den Multiplikationssatz und Additionssatz für unabhängige Ereignisse.

$$P(A \cap B) = P(A)P(B)$$
$$P(A \cup B) = P(A) + P(B) - P(A)P(B)$$

3.2 Wahrscheinlichkeiten in der Epidemiologie (S. 111)

Die Epidemiologie befasst sich mit dem Auftreten von Krankheiten in einer größeren Population. Aufgaben und Ziele sind zB das Erkennen von Ursachen und Risikofaktoren von Krankheiten, das Bestimmen deren Verbreitung in der Bevölkerung, ...

3.2.1 Prävalenz (Punktprävalenz)

Ist der relative Krankenbestand zu einem Zeitpunkt - also die Wahrscheinlichkeit $P(K_t)$ für eine beliebige Person aus der Population, zum Zeitpunkt t erkrankt zu sein. (eignet sich für Krankheiten die chronisch sind oder wiederholt auftreten)
Die Punktprävalenz muss immer mit einem Zeitpunkt angegeben werden, auf den sie sich bezieht.
Die Prävalenz informiert über die Auswirkungen einer Krankheit in der Gesamtpopulation.
Die Prävalenz ist eine a-priori Wahrscheinlichkeit.

3.2.2 Periodenprävalenz

Das Bestimmen von Punktprävalenzen kann bei Krankheiten mit kurzer Dauer schwierig sein. Die Periodenprävalenz bezieht sich auf einen längeren Zeitraum. Dabei werden alle Personen berücksichtigt, die zu Beginn, während oder am Ende des Beobachtungszeitraums erkrankt waren.

3.2.3 Inzidenz

Dies ist die Neuerkrankungsrate, also die Wahrscheinlichkeit $P(K)$ für eine beliebige Person, während einer Beobachtungszeit zu erkranken. Sie gibt im Prinzip an, wie groß das Erkrankungsrisiko für eine Person ist. Die Inzidenz hat nur Aussagekraft bei Krankheiten, die in einem Zeitraum maximal einmal auftreten.

$$\text{Prävalenz} = \text{Inziden} * \text{durchschnittliche Dauer}$$

3.2.4 Krankheitsspezifische Mortalität

Darunter versteht man die Todesrate - also die Wahrscheinlichkeit $P(K \cap T)$, während der Beobachtungszeit an der Krankheit K zu erkranken und daran zu versterben.

3.2.5 Letalität

Ist die Tödlichkeitsrate der Erkrankten - dh, die Wahrscheinlichkeit zu sterben, wenn man krank ist. Formal: $P(T|K)$.
Aus letzteren beiden lässt sich nun folgendes schließen (bei Anwendung des Multiplikatinssatzes):

$$P(K \cap T) = P(K)P(T|K)$$

oder in Worten

$$\text{Mortalität} = \text{Inzidenz} * \text{Letalität}$$

3.2.6 Kontagionsindex

Ist ein Maß für die Ansteckungsfähigkeit, dh die Wahrscheinlichkeit mit der sich eine nicht immune Person, die mit dem Erreger in Kontakt kommt, ansteckt.

3.2.7 Manifestationsindex

Ist die Wahrscheinlichkeit, mit der eine infizierte Person manifest erkrankt (die Krankheitsbereitschaft).

3.3 Bevölkerungsstatistiken (S. 114)

3.3.1 Spezielle Wahrscheinlichkeiten (S. 114)

- Natalität
  Geburtenrate - also der Anteil geborener Kinder im Verhältnis zur Gesamtpopulation während eines Beobachtungszeitraums.

$$\text{Natalität} = \frac{\text{Anzahl lebend geborener Kinder / Jahr}}{\text{Populationsgröße}}$$

- Fertilitätsziffer
  Fruchtbarkeitsrate - das ist die Wahrscheinlichkeit, dass eine Frau im gebährfähigen Alter ein lebendes Kind zur Welt bringt.

$$\text{Fertilitätsziffer} = \frac{\text{Anzahl lebend geborene Kinder / Jahr}}{\text{Anzahl Frauen im gebährfähigem Alter}}$$

- Pearl-Index
  Ein Risikomass bezüglich der Sicherheit einer Verhütungsmethode. Er gibt an, wie viele von 100 Frauen in einem Jahr ungewollt schwanger werden, obwohl sie eine bestimmte Verhütungsmethode anwenden. Dieses Mass ist im Gegensatz zur Fertilitätsziffer keine Wahrscheinlichkeit!

- Sterbeziffer
  Gesamtmortalität - Der Anteil der im Beobachtungszeitraum Verstorbenen.

$$\text{Sterbeziffer} = \frac{\text{Anzahl gestorbene pro Jahr}}{\text{Populationsgröße}}$$

3.3.2 Sterbetafeln (S. 115)

Sie beschreibt die Verteilung von Lebensdauern und basiert auf folgenden Häufigkeiten:

- l_0: Anzahl von Lebendgeborenen innerhalb eines Beobachtungszeitraums
- l_x: Anzahl der Personen, die ihren x-ten Geburtstag erleben und danach noch unbestimmte Zeit leben.

- $d_x = l_x - l_{x+1}$: die Anzahl der lebend Geborenen, die zwischen dem x-ten und (x+1)-ten Geburtstag sterben.

Der Einfachheit halber werden Lebensdauern typischerweise als diskretes Merkmal aufgefasst mit den Ausprägungen x (Anzahl der erreichten Lebensjahre) und den absoluten Häufigkeiten d_x.

Die Sterbeziffern sind die altersspezifischen Mortalitätsraten mit

$$q_x = \frac{d_x}{l_x}(x = 0, ..., \omega)$$

q_x gibt also die Wahrscheinlichkeit an, mit der man zwischen dem x-ten und (x+1)-ten Geburtstag stirbt. ω gibt hierbei das letzte in der Sterbetafel berücksichtigte Alter bekannt. Man nimmt an: $l_{\omega+1} = 0$ wobei ω typischerweise 100 gesetzt wird.

Somit lässt sich auch die Lebenserwartung eines x-jährigen bestimmen:

$$e_x = \frac{1}{2} + \frac{1}{l_x} \sum_{y=x+1}^{\omega} l_y$$

Zur Herleitung dieser Formel:

$$e_o = \frac{1}{2} + \text{Anzahl der erwarteten vollendeten Lebensjahre} =$$
$$\frac{1}{2} + [0\frac{d_0}{l_0} + 1\frac{d_1}{l_0} + ... + (\omega - 1)\frac{d_{\omega-1}}{l_0} + (\omega)\frac{d_\omega}{l_0}] =$$
$$\frac{1}{2} + \frac{1}{l_0}[0(l_0 - l_1) + 1(l_1 - l_2) + 2(l_2 - l_3) + ... + \omega(l_\omega - l_{\omega+1})] \text{ wobei } l_{\omega+1} = 0 =$$
$$\frac{1}{2} + \frac{1}{l_0} \sum_{x=0}^{\omega} l_x$$

Die Verteilungsfunktion $F(x)$ gibt den relativen Anteil der Lebendgeborenen an, deren Sterbealter kleiner als x ist.

$$F(x) = 1 - \frac{l_x}{l_0} \text{ für } 0 \leq x \leq \omega$$

3.4 Diagnostische Tests (S. 118)

3.4.1 Gütekriterien

Diagnostische Tests werden benutzt um größere Sicherheit bezüglich des Krankheitsstatus eines Patienten zu gewinnen. Es gibt im Wesentlichen 2 Testergebnisse:

- ein positives Ergebnis (T_+), wenn die Person erkrankt ist (K)
- ein negatives Ergebnis (T_-), wenn die Person nicht erkrankt ist ($\bar{K}$)

Die Güte wird nun quantifiziert durch:

Sensitivität Dies ist die bedingte Wahrscheinlichkeit $P(T_+|K)$ dass der Test bei einer kranken Person richtig (also positiv) reagiert. (Logisch gedacht: je höher dieser Wert, desto mehr positive Entscheidungen gibt es also, wenn die Person krank war - sprich, desto öfter ist man mit dem Testergebnis richtig gelegen).

Man legt Wert auf eine hohe Sensitivität wenn

- es sich um eine Krankheit mit schweren Folgen für den Patienten handelt
- eine Erfolg versprechende Therapie zur Verfügung steht
- falsch positive Befunde mit vertretbarem Aufwand und ohne allzu große Belastungen für die betreffende Person geklärt werden können.

Spezifität Gibt die Wahrscheinlichkeit $P(T_-|\bar{K})$ an, dass eine nicht erkrankte Person ein richtiges (also negatives) Testergebnis erhält. (selbige Überlegung wie oben).

Man legt Wert auf eine hohe Spezifität wenn

- keine Therapie mit Aussicht auf Besserung bekannt ist
- die Therapie zu hohen finanziellen Belastungen führt
- die Therapie schwere Nebenwirkungen hat
- die Nachfolgeuntersuchungen mit erheblichen Risiken und Belastungen verbunden sind

Im Idealfall sollten beide Werte 1 sein - sprich man liegt mit dem Ergebnis richtig.

3.4.2 Vorhersagewerte (S. 119)

Sie geben an, mit welcher Wahrscheinlichkeit das Testergebnis den richtigen Krankheitsstatus des Patienten anzeigen. Dabei versteht man unter dem **positiven Vorhersagewert** die bedingte Wahrscheinlichkeit $P(K|T_+)$ und unter dem **negativen Vorhersagewert** die bedingte Wahrscheinlichkeit $P(\bar{K}|T_-)$. Die Wahrscheinlichkeit $P(K|T_+)$bzw$P(K|T_-)$ liegt ja erst nach dem Testergebnis vor weshalb sie auch als 'a-posteriori' Wahrscheinlichkeit bezeichnet wird.

4 Diskrete Verteilungen (S. 125)

4.1 Die Binomialverteilung (S. 129)

Sie basiert auf dem Bernoulli-Experiment. Sie beschreibt, mit welcher Wahrscheinlichkeit das Ereignis A bei n unabhängigen Beobachtungen 0,1,2, ... oder n-mal gezählt wird. (Ziehen-mit-zurücklegen!)

- Das Experiment kann beliebig oft durchegführt werden
- jedex X_i nimmt mit der Wahrscheinlichkeit p den Wert 1 (bei Eintreten des Ereignisses A) und mit der Wahrscheinlichkeit $q = 1 - p$ den Wert 0 an
- dann quantifiziert die Zufallsvariable $X = X_1 + X_2 + ... + X_n$ wie häufig bei n Experimenten A eingetreten ist. X wird durch die Binomialverteilung beschrieben. $X\tilde{}B(n, p)$

$$E(X) = \sum_{i=1}^{n} EX_i = n * p$$

$$Var(X) = \sum_{i=1}^{n} Var(X_i) = n * p * (1-p)$$

$$\gamma_1 = \frac{q-p}{\sigma} \text{ mit } \sigma = \sqrt{n * p * (1-p)}$$

Die Verteilung ist rechts-schief wenn $\gamma_1 > 0$ und links-schief, wenn $\gamma_1 < 0$.

$$P(X = k) = \binom{n}{k} (p)^k (1-p)^{n-k}$$

4.2 Die Poisson-Verteilung

Die Poisson-Verteilung wird etwa bei epidemiologischen Untersuchungen eingesetzt, wo umfangreiche Populationen vorliegen, jedoch das Eintreten einer bestimmten Krankheit bei einem Individuum recht unwahrscheinlich ist. zB Geburten und Down-Syndrom ($n = 2000, p = 0,0001$).

$$X \tilde{} P(\lambda)$$

$$EX = n * p = \lambda = \text{const}, \; n \to \infty, \; p = \frac{\lambda}{n}$$

Die Binomialverteilung kann durch die Poisson Verteilung approximiert werden, wenn $n \geq 30 \wedge p \leq 0,1$. Wir wollen nun den Übergang von $B(n,p) \to P(\lambda)$ mit $p = \frac{\lambda}{n}, \lambda = const$ machen.

$$EX = \lim_{n\to\infty} EX = \lim_{n\to\infty} n * p = \lim_{n\to\infty} n * \frac{\lambda}{n} = \lambda$$

$$VarX = \lim_{n\to\infty} n * p * q = \lim_{n\to\infty} n * \frac{\lambda}{n} * (1 - \frac{\lambda}{n}) = \lambda \to \sigma = \sqrt{\lambda}$$

$$\gamma_1 = \lim_{n\to\infty} \frac{q-p}{\sigma} = \lim_{n\to\infty} \frac{(1-\frac{\lambda}{n}) - \frac{\lambda}{n}}{\sqrt{\lambda}} = \frac{1}{\lambda} > 0$$

Nun zur Herleitung der Poisson-Verteilung an sich.

$$P(X=k) = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} = \frac{n*(n-1)*...*(n-k+1)}{k!} \frac{\lambda^k}{n^k} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} =$$

$$\lim_{n\to\infty} \frac{\lambda^k}{k!} \frac{n*(n-1)*...*(n-k+1)}{n^k} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} = ...$$

$$\lim_{n\to\infty} \frac{n*(n-1)*...*(n-k+1)}{n^k} = \lim_{n\to\infty} 1*(1-\frac{1}{n})*(1-\frac{2}{n})*...*(1-\frac{k-1}{n}) = 1$$

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{-k} = 1^{-k} = 1$$

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = \lim_{n\to\infty} \left(1+\frac{-\lambda}{n}\right)^n = e^{-\lambda}$$

$$\Longrightarrow P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

4.3 Die Polynomialverteilung (S. 138)

Sie beschreibt eine Serie von Zufallsexperimenten, bei denen pro Beobachtung eines von k möglichen Ereignissen $A_1, A_2, ..., A_k$ mit den Wahrscheinlichkeiten $p_1, p_2, ..., p_k$ auftreten kann. (Anwendung zB Blutgruppen).

$$P(n_1, n_2, ..., n_k | p_1, p_2, ..., p_k) = \frac{(p_1)^{n_1} * ... * (p_k)^{n_k}}{n_1! * ... * n_k!} * n!$$

4.4 Die negative Binomialverteilung (S. 139)

Sie beschreibt, mit welcher Wahrscheinlichkeit das Ereignis A gerade bei der j-ten Beobachtung zum r-ten Mal eintritt. Diese Verteilung wird häufig zur Analyse von Wartezeiten verwendet.

$$X \tilde{} NB(r,p)$$

$$\text{spez: } X \tilde{} NB(1,p) \rightarrow \text{ geometr. Vert.}$$

$$P(X=j) = \binom{j-1}{r-1} q^{j-r} * p^r \text{ für } j = 1, 2, 3, ...$$

4.5 Die hypergeometrische Verteilung (S. 140)

Beschreibt n Beobachtungen, bei denen jeweils alternativ das Ereignis A oder $\bar{A}$ eintreten können wobei diese **nicht unabhängig** voneinander sind. (ziehen-**ohne**-zurücklegen!).

Die Zufallsvariable $X \tilde{} HG(n; N, M)$ gibt an, wie häufig das Merkmal A bei n Beobachtungen auftritt. Die Wahrscheinlichkeiten für $k = 0, ..., n$ sind:

Population	N	Personen
	M	Personen mit Eigenschaft A
	$N-M$	Personen mit Eigenscahft $\bar{A}$
Stichprobe	n	Personen
	k	Personen mit Eigenschaft A
	$n-k$	Personen mit Eigenscahft $\bar{A}$

$$P(X=k) = \frac{\binom{M}{k} * \binom{N-M}{n-k}}{\binom{N}{n}}$$

Der Quotient $p = M/N$ wird auch als Anteilswert bezeichnet. Aus ihm ergibt sich der Erwartungswert sowie die Varianz.

$$E(X) = np = n * \frac{M}{N}$$
$$Var(X) = \frac{N-n}{N-1} * n * p * (1-p)$$

Der Faktor $\frac{N-n}{N-1}$ entspricht der **Endlichkeitskorrektur**. Falls N im Vergleich zu n sehr groß ist, dann kann die hypergeometrische auch durch die Binomialverteilung angenähert werden.

5 Stetige Verteilungen (S. 145)

5.1 Die Normalverteilung (S. 148)

Normalverteilte Zufallsvariable X sind durch den Erwartungswert μ und der Standardabweichung σ eindeutig bestimmt.

- die Glockenkurve ist symmetrisch: $f(\mu + x) = f(\mu - x)$
- sie hat 2 Wendepunkte bei $x = \mu - \sigma$ und $x = \mu + \sigma$
- ihr Maximum liegt bei $x = \mu$

$$P(a \leq X \leq b) = F(b) - F(a)$$

5.1.1 Die Standardnormalverteilung (S. 149)

Jede normalverteilte Zufallsvariable $X \tilde{} N(\mu, \sigma^2)$ lässt sich in die Standardnormalverteilung $Z \tilde{} N(0,1)$ transformieren durch:

$$Z = \frac{X - \mu}{\sigma}$$
$$\Phi(-z) = P(Z \leq -z) = P(Z \geq z) = ... = 1 - \Phi(z)$$

Mit ϕ bezeichnen wir die Dichtefunktion f von Z und mit Φ die Verteilungsfunktion F von Z.

Transformation in Z: z informiert, um wie viele Standardabweichungen der entsprechende x-Wert vom Erwartungswert μ abweicht.

5.1.2 σ-Bereiche, Referenzbereiche (S. 151)

Die Normalverteilung wird häufig dafür verwendet, um quantitative, symmetrisch verteilte und eingipfelige Merkmale zu beschreiben. Referenzbereiche können nun vorgegeben werden, um zB Aktionen auszulösen, wenn Werte außerhalb dieser liegen: etwa pathologische Untersuchungen, etc.

Der $1 - \sigma$ Bereich wird zB angegeben durch:

$$[\mu - \sigma, \mu + \sigma]$$

Der 95%-Referenzbereich wird ausgerechnet durch:

$$0,95 = P(-z \leq Z \leq -z) = 2\Phi(z) - 1$$
$$\Phi(z) = \frac{1,95}{2} = 0,975 \leftrightarrow z = 1,96$$
$$[\mu - 1,96\sigma, \mu + 1,96\sigma]$$

5.1.3 Normalisierende Transformationen (S. 152)

Oft liegen keine symmetrischen bzw schiefe Verteilungen vor. Diese können aber mittels Transformationen in eine entsprechende Normalverteilung überführt werden.
Eine rechts-schiefe Verteilung (dh, Gipfel am linken Rand, rechts langsam weit abfallend) kann zB über $Y = lnX$ transformiert werden. Voraussetzung ist, dass alle $x_i > 0$ sind. Dann ist X log-normalverteilt. Die Art der Transformation, ob log oder sogar log-log, hängt von der Verteilung ab und muss empirisch bestimmt werden.

$$X \tilde{} LN(\mu, \sigma^2)$$
$$Y \tilde{} N(\mu, \sigma^2)$$
$$X = e^Y$$
$$P(X \leq d) = P(lnX \leq lnd) = P(Y \leq lnd) = \Phi(\frac{lnd - \mu}{\sigma})$$
$$p = \Phi(z_p) = P(\frac{Y - \mu}{\sigma} \leq z_p) = P(Y \leq z_p * \sigma + \mu) =$$
$$P(e^Y \leq e^{\mu + z_p * \sigma}) \rightarrow x_p = e^{\mu + z_p * \sigma}$$

Andere Transformationen für rechtsschiefe Verteilungen sind zB $Y = ln(X+a), a > 0$ oder $Y = lnlnX$.

links-schiefe Verteilungen lassen sich dann über Potenztransformationen in eine Normalverteilung überführen: $Y = X^{1,5}$.

5.2 Sätze der Wahrscheinlichkeitsrechnung (S. 155)

5.2.1 Die Tschebyscheff'sche Ungleichung (S. 155)

Sie erlaubt die Abschätzung der Wahrscheinlichkeit, mit der die Zufallsvariable X um mehr als eine feste Zahl vom Erwartungswert μ abweicht.

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}, \text{ für alle } k > 0 \ (\epsilon := k\sigma)$$

Die Ungleichung gilt für alle stetigen Verteilungen, sowohl symmetrische als auch schiefe. Für symmetrische, eingipfelige Verteilungen gibt es von Gauß eine schärfere Ungleichung:

$$P(|X - \mu| > k\sigma) \leq \frac{4}{9k^2}, \text{ für alle } k \geq 2/\sqrt{3} \approx 1,155$$

5.2.2 Das Gesetz der großen Zahlen (S. 157)

Es ist klar, dass sich der Erwartungswert einer Grundgesamtheit besser durch einen Mittelwert schätzen lässt, je größer der zugrunde liegende Stichprobenumfang ist. Das Gesetz der großen Zahlen besagt nun, dass sich ein Mittwelwert $\bar{x}$ mit wachsendem Stichprobenumfang dem Erwartungswert μ nähert. Formal bedeutet dies:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \rightarrow_{n \rightarrow \infty} \mu$$

Somit lässt sich folgendes herleiten:

$$E(\bar{X}) = \mu$$
$$Var(\bar{X}) = \frac{\sigma^2}{n}$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$\sigma_{\bar{x}}$ wird als **Standardfehler des Mittelwertes** bezeichnet und beschreibt im Prinzip die Variabilität der möglichen Mittwelwerte. (alle X_i sind unabhängige Zufallsvariable).

5.2.3 Der zentrale Grenzwertsatz

Sagt aus, das unter sehr allgemeinen Bedingungen die Summe einer großen Anzahl von Zufallsvariablen normalverteilt ist, wenn die X_i unabhängige und identisch verteilte Zufallsvariable sind.

$$\sum X_i \tilde{} N(n * \mu, n * \sigma^2)$$

asymptotisch für $n \to \infty$ gilt dann

$$Z_n = \frac{\sum X_i - n * \mu}{\sqrt{n} * \sigma} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tilde{} N(0, 1)$$

Daraus folgt

- dass eine Zufallsvariable normalverteilt ist, wenn zahlreiche Einflüsse additiv und unabhängig voneinander zusammenwirken
- dass die Verteilung von Mittelwerten $\bar{X} \tilde{} N(\mu, \frac{\sigma^2}{n})$ ist asymptotisch für $n \to \infty$. (in der Praxis: $n \geq 25$)
- dass eine Binomialverteilung für ein hinreichend großes n duch die Normalverteilung angenähert werden kann mit dem Erwartungswert $\mu = np$ und der Varianz $\sigma^2 = npq$, wenn $npq \geq 9$ ist.

5.2.4 Bedeutung der NV

- empirische VT: belebte Natur 'strebe' NV an. Dieser Ansatz stimmt nicht.
- approximative VT: eventuell Transformationen bzw Approximationen (B oder P)
- Verteilungen für statistische Kenngrößen insbesondere $\bar{X}$

5.3 Die Verteilung von Überlebenszeiten (S. 162)

5.3.1 Wichtige Begriffe

- Überlebenszeit
  Ist die Dauer zwischen einem definierten Anfangsereignis und dem Eintritt eines zufallsbedingten Endereignisses. (zB Geburt - Tod) Wenn ein Lebewesen von der Geburt bis zu seinem Tod beobachtet wird, so spricht man von **Lebensdauer**.

- Überlebensfunktion
  Sei T (oder X) eine Zufallsvariable zur Beschreibung der Überlebenszeit. Diese kann logischerweise nur positive Werte annehmen. Dann ist die Überlebensfunktion definiert als:

$$F(t) = P(X \leq t)$$
$$s(t) = P(X > t) = 1 - F(t)$$

  $s(t)$ beschreibt dabei die Überlebensfunktion (survival function) und gibt die Wahrscheinlichkeit an, mit der ein Individuum den Zeitpunkt t überlebt.

- Bedingte Überlebenswahrscheinlichkeit
  Sie gibt die Wahrscheinlichkeit an, dass ein Individuum, welches den Zeitpunkt t überlebt hat, eine weiter Zeitspanne Δt überleben wird.
  Die Bedingte Wahrscheinlichkeit allgmein setzt sich folgend zusammen:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  Daraus kann die bedingte Überlebenswahrscheinlichkeit hergeleitet werden:

$$P(X > t + \Delta t | X > t) = \frac{P(X > t + \Delta t \cap X > t)}{P(X > t)} =$$
$$\frac{P(X > t + \Delta t)}{P(X > t)} = \frac{s(t + \Delta t)}{s(t)}$$

- Bedingte Sterbewahrscheinlichkeit

$$P(t < X \leq t + \Delta t | X > t) = \frac{P(t < X \leq t + \Delta t \cap X > t)}{P(X > t)} =$$
$$\frac{P(t < X \leq t + \Delta t)}{P(X > t)} = \frac{F(t + \Delta t) - F(t)}{s(t)}$$

- Momentane Sterberate (Hazard-Rate)
  Ergibt sich aus folgender Beziehung:

$$r(t) = \frac{f(t)}{s(t)}$$

  wobei

$$f(t) = F'(t)$$

  Die momentane Sterberate hat den Vorteil gegenüber der Mortalität, das sie unabhängig vom Beobachtungszeitraum ist und für jeden Zeitraum t angegeben werden kann.

Sie ergibt sich aus der bedingten Sterbewahrscheinlichkeit

$$r(t) = lim_{\Delta t \to 0} \frac{P(t < X \leq t + \Delta t | X > t)}{\Delta t} =$$

$$lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{s(t)} = F'(t) \frac{1}{s(t)} = \frac{f(t)}{s(t)}$$

- Interpretation von r(t)
  In linearer Approximation ist (für ein hinreichend kleines Δt) bekanntlich

  $$F(t + \Delta t) \approx F(t) + F'(t)\Delta t$$

  Daher gilt, für die Wahrscheinlichkeit, dass ein Individuum, nachdem es den Zeitpunkt t erlebt hat, im darauf folgenden Zeitintervall der Länge Δt stirbt (=bed. Sterbewahrscheinlichkeit)

  $$P(t < X \leq t + \Delta t | X > t) = \frac{F(t + \Delta t) - F(t)}{s(t)}$$
  $$\approx \frac{F(t) + F'(t)\Delta t - F(t)}{s(t)} = \frac{f(t)\Delta t}{s(t)} = r(t)\Delta t$$

  Diese Wahrscheinlichkeit ist also in linearer Näherung das Produkt aus

  - momentaner Sterberate
  - Länge des Zeitintervalls

  und somit proportional zu Δt.

  Mit $\Delta t = 1$ ergibt sich daraus

  $$r(t) \approx P(t < X \leq t + 1 | X > t)$$

  Das bedeutet, dass die Ausfallrate $r(t)$ folgendermassen interpretiert werden kann:
  $r(t)$ ist die Wahrscheinlichkeit, dass ein Individuum, nachdem es den Zeitpunkt (das Alter) t erlebt hat, während der folgenden (kleinen) Zeiteinheit stirbt.

  Wichtig ist auch, wie man $r(t)$ dazu verwenden kann, um $s(t)$, $F(t)$ und $f(t)$ auszudrücken! (siehe Skriptum)

5.3.2 Die Exponentialverteilung (S. 164)

Im einfachsten Fall lässt sich die Überlebenswahrscheinlichkeit modellieren als

$$s(t) = P(X > t) = e^{-\lambda t}$$

wobei $\lambda > 0$. Daraus folgt weiters:

$$F(t) = 1 - s(t) = P(X \leq t) = 1 - e^{-\lambda t}$$
$$f(t) = F'(t) = \lambda e^{-\lambda t}$$

Man könnte auch die Exponentialverteilung als Spezialfall der Sterberate sehen mit $r(t) = \lambda$ (also konstant).

$$\int_0^t r(\tau) d\tau = \int_0^t \lambda d\tau = \lambda t$$

Die Exponentialverteilung hat einige bemerkenswerte Eigenschaften. Die Wahrscheinlichkeit, noch eine Zeitspanne der Länge Δt zu leben, ist also unabhängig vom Alter. Deshalb wird die Exponentialverteilung auch gedächtnislose Verteilung genannt. Wegen dieser Eigenschaft ist die Sterberate kostant. Die Exponentialverteilung wird deshalb zur Beschreibung von Lebensdauern nicht alternder Objekte oder von Überlebenszeiten bei Individuen herangezogen. Beispiel: das Überleben nach einer schweren Krankheit mit kurzer Lebenserwartung.
Wichtige Kenngrößen sind noch der Median, der Erwartungswert und die Varianz. Diese sind umso größer je kleiner die momentane Sterberate λ ist. Die Schiefe beträgt grundsätzlich 2 - die Verteilung ist also rechtsschief.

$$\widetilde{\mu} = \frac{1}{\lambda} ln2$$
$$\mu = \frac{1}{\lambda}$$
$$\sigma^2 = \frac{1}{\lambda^2}$$

5.3.3 Weibull-Verteilung (S. 165)

Sie kann als Spezialfall der Exponentialverteilung gesehen werden bzw umgekehrt.

$$Exp(\lambda) = WB(\lambda, 1)$$

Eine Zufallsvariable X heißt Weibull-verteilt mit den Parametern $\lambda > 0$ und $\gamma > 0$ wenn für ihre Verteilungsfunktion gilt

$$F(t) = 1 - e^{-\lambda t^\gamma}$$

für $t > 0$.

$$s(t) = e^{-\lambda t^\gamma}$$

$$f(t) = \lambda\gamma t^{\gamma-1} e^{-\lambda t^{\gamma}}$$

$$r(t) = \lambda\gamma t^{\gamma-1}$$

5.4 Prüfverteilungen (S. 166)

5.4.1 Die Chi^2 Verteilung (S. 168)

Beschreibt die Verteilung des Quadrats einer standardnormalverteilten Zufallsvariable. Sie ist immer rechtschief. Für den Erwartungswert von $\chi_1^2 = Z^2$ gilt:

$$EZ^2 = VarZ + (EZ)^2 = 1 \rightarrow EZ^2 = VarZ + (EZ)^2 = 1 + 0^2 = 1$$

Voraussetzungen:

1. $Z_1, ...Z_n$ sind unabhängig
2. $Z_i \tilde{} N(0,1)$ für $i = 1, ...n$

$$E(\sum_1^n Z_i^2) = \sum_1^n EZ_i^2 = \sum_1^n 1 = n$$

$$Var(\sum_1^n Z_i^2) = 2n$$

$$\gamma_1 = \sqrt{\frac{8}{n}}$$

$$\sum_1^n Z_i^2 \tilde{} \chi_n^2$$

$$\sum_1^n (\frac{X_i - \mu}{\sigma})^2 \tilde{} \chi_n^2$$

$$\sum_1^n (\frac{X_i - \bar{X}}{\sigma})^2 \tilde{} \chi_{n-1}^2$$

$$\sum_1^n (\frac{X_i - \bar{X}}{\sigma})^2 = \frac{n-1}{n-1}\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{\sigma^2} * S^2 \tilde{} \chi_{n-1}^2$$

5.4.2 Die F-Verteilung (S. 170)

Seien S_1^2 und S_2^2 die Varianzen von 2 unabhängigen Stichproben aus 2 normalverteilten Grungesamtheiten mit **der gleichen Varianz**. n_1, n_2 sind dabei die Stichprobenumfänge.

$$F := \frac{S_1^2}{S_2^2} = F_{n_1-1;n_2-1}$$

Mit der vereinfachten Bezeichnung χ_f^2 für eine Zufallsvariable welche χ_f^2 verteilt ist, gilt auch dass diese Zufallsvariable

$$F_{n_1-1;n_2-1} = \frac{S_1^2}{S_2^2} = \frac{\frac{\sigma^2}{n_1-1}\chi_{n_1-1}^2}{\frac{\sigma^2}{n_2-1}\chi_{n_2-1}^2} = \frac{\chi_{n_1-1}^2}{\chi_{n_2-1}^2}\frac{n_2-1}{n_1-1}$$

ist.

6 Schätzverfahren (S. 173)

Hier bedient man sich der Methoden der induktiven (schließenden bzw anayltischen) Statistik um aus einer zufälligen Stichprobe, die repräsentativ für ein übergeordnetes Kollektiv ist (die Grundgesamtheit) und deren Verteilung bekannt ist, Schätzungen bezüglich bestimmter Charakteristika zu machen.

6.1 Punktschätzungen (S. 173)

Ein Verfahren, bei dem ein unbekannter Parameter durch einen einzigen Wert geschätzt wird, einen Punktschätzer. Die Schätzfunktion oder auch Schätzer genannt ist eine Vorschrift, nach der aus den Daten einer Stichprobe des Umfangs n ein angenäherter Wert für den unbekannten Parameter berechnet wird.

6.1.1 Kriterien zur Güte einer Schätzung (S. 174)

1. Erwartungstreue
   Fordert, dass der Durchschnitt (genauer: der Erwartungswert) aller theoretisch denkbaren Schätzwerte aus den Stichproben des Umfangs n mit dem unbekannten Parameter übereinstimmt. Eine erwartungstreue Schätzung heißt unverzerrt.

2. Konsistenz
   Varianz der Schätzfunktion soll mit n gegen unendlich gegen 0 gehen.

3. Effizienz
   Die Varianz soll also möglichst gering sein. Eine hohe Effizienz sagt aus, dass der Schätzer auch bei geringer Stichprobengröße n brauchbare Werte liefert.

4. Exhaustivität
   Er ist erschöpfend, wenn er möglichst alle Informationen berücksichtigt.

6.1.2 Spezielle Schätzfunktionen (S. 175)

Erwartungswert

$$E(\bar{X}) = \mu \text{ ... erwartungstreu}$$
$$Var(\bar{X}) = \frac{n * \sigma^2}{n^2} = \frac{\sigma^2}{n} \to 0 \text{ für } n \to \infty \text{ ... konsistent}$$

Median

$$\tilde{X} \text{ ist erwartungstreu: } E\tilde{X} = \tilde{\mu}$$
$$Var\tilde{X} = \frac{\pi}{2}\frac{\sigma^2}{n}$$

Voraussetzungen: Verteilung ist stetig und symmetrisch!

Varianz

7 Tests zum Vergleich von Häufigkeiten (S. 229)

7.1 Der Binomialtest für eine Stichprobe (S. 229)

7.2 Chi^2-Tests (S. 231)

Sie dienen zur Analyse von Häufigkeitsunterschieden. Da sich Häufigkeiten bei jeder Merkmalsart und jede Skalenniveau ermitteln lassen, sind diese Tests sehr vielseitig anwendbar.

7.2.1 Der Chi^2-Vierfelder Test (S. 231)

7.2.2 Der Chi^2-Test für $k * l$ Felder (S. 236)

7.2.3 Assoziationsmaße für qualitative Merkmale (S. 237)

Sie beschreiben die Stärke eines Zusammenhangs.

Phi-Kooeffizient Beschreibt den Zusammenhang zwischen 2 Alternativmerkmalen. Er weißt bei 0 auf vollkommene Unabhängigkeit hin. Er ist signifikant größer als 0 wenn das Ergebnis des Vierfeldertests signifikant ist.

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

Assoziationskoeffizient nach Yule (S. 52) Er nimmt den Wert 0 an, wenn $ad = bc$ (vollkommene Unabhängigkeit). Ansonsten erstreckt sich Q zwischen -1 und 1.

$$Q = \frac{ad - bc}{ad + bc}$$

Odds Ratio (S. 51) Dies ist ein Assoziationsmaß, das den Grad eines Zusammenhangs zwischen 2 Alternativmerkmalen quantifiziert. Es wird gebildet, indem man aus den Häufigkeiten im Inneren der Vierfeldertafel das Kreuzprodukt bildet. Ein Odds Ratio mit dem Wert 1 zeigt, dass kein Zusammenhang zwischen den beiden Merkmalen besteht.

$$OR = \frac{ad}{bc}$$

7.2.4 Der McNemar-Test (S. 238)

7.2.5 Der Logranktest (S. 242)

Er dient zum Vergleich von Verteilungen von Überlebenszeiten bei 2 unverbundenen Stichproben. Seine Besonderheit ist, dass er auch zensierte Daten angemessen berücksichtigt. Mit ihm wird beispielsweise überprüft, ob sich eine Therapie oder ein prognostischer Faktor auf die Überlebenszeit oder allgemein auf die Zeit bis zum Eintreten eines bestimmten Endereignisses auswirkt.

$$\begin{array}{lcl} H_0: & S_1 = S_2 & \Leftrightarrow \quad S_1(x) \equiv S_2(x) \\ H_1: & S_1 \neq S_2 & \Leftrightarrow \quad S_1(x) \neg \equiv S_2(x) \end{array}$$

Tab. 1: Hypothesen des Logranktests

Seien die Stichproben gekennzeichnet durch $j = 1, 2$ und die Zeitpunkte t_i für $i = 1...k$ (wobei $t_0 = 0$), so gilt folgendes:

d_{ji}	Anzahl der Endereignisse in Stichprobe j zur Zeit t_i
d_j	Summe aller Todesfälle in Stichprobe j dh $\sum_{i=1}^{k} d_{ji}$
c_{ji}	Anzahl der zensierten Daten in Stichprobe j zur Zeit $[t_{i-1}, t_i)$
$n_{j0} := n_j$	Anzahl der Beobachtungseinheiten in Stichprobe j

Somit ist

$$n_{ji} := n_{j;i-1} - d_{j;i-1} - c_{ji}$$

bzw.

$$n_{ji} - d_{ji} - c_{j;i+1} = n_{j;i+1}$$

$$e_j := \sum_{i=1}^{k} (d_{1i} + d_{2i}) \frac{n_{ji}}{n_{1i} + n_{2i}}$$

für $j = 1, 2$. Daraus folgt weiter

$$e_1 + e_2 = ... = \sum_{i=1}^{k} (d_{1i} + d_{2i}) \left(\frac{n_{1i}}{n_{1i} + n_{2i}} + \frac{n_{2i}}{n_{1i} + n_{2i}} \right) =$$

$$\sum_{i=1}^{k} (d_{1i} + d_{2i}) = d_1 + d_2$$

$$e_2 = d_1 + d_2 - e_1$$

Die Teststatistik für den Logranktest lautet

$$\chi^2 = \frac{(d_1 - e_1)^2}{e_1} + \frac{(d_2 - e_2)^2}{e_2}$$

mit $f = 1$ Freiheitsgraden.

Die H_0 Hypothese $H_0 : S_1 = S_2$ wird abgelehnt $\leftrightarrow \chi^2 > \chi^2_{1;1-\alpha}$.

7.3 Ausblick auf die logistische Regression (S. 245)

Die logistische Regression ist ein multiples Verfahren, mit dem es möglich ist, die Wahrscheinlichkeit für das Auftreten eines bestimmten Endereignisses basierend auf mehreren Einflussgrößen zu modellieren.

8 Risikostudien (S. 261)

8.1 Deskriptive Studien (S. 264)

8.1.1 Fallberichte (S. 264)

Ist eine ausführliche Beschreibung eines interessanten Einzelfalls oder einiger weniger Fälle.

8.1.2 Fallserien (S. 265)

Unterscheidet sich von Fallberichten durch die Anzahl der involvierten Patienten. Ist eine einfache deskriptive Studie an einer größeren Gruppe von Personen, die an einer bestimmten Krankheit leiden und darüber hinaus einige Besonderheiten aufweisen.

8.1.3 Prävalenzstudien (S. 266)

Sind Querschnittstudien, in denen bei jedem Teilnehmer erfasst wird, ob er an einer bestimmten Erkrankung leidet und ob er exponiert ist.

8.1.4 Populationsstudien (S. 266)

Unterscheiden sich von anderen Studien, dass nicht Individuen untersucht, sondern Gruppen oder Länder zugrunde gelegt werden.

8.2 Fall-Kontroll-Studien (S. 267)

Hier werden Fälle (Patienten, die an einer bestimmten Krankheit leiden - K) und Kontrollen (Personen, die von dieser Krankheit nicht betroffen sind - $\bar{K}$) bezüglich eines oder mehrerer Faktoren miteinander verglichen. Sie sind retrospektiv und analytisch. Die Fälle werden meist aus Kliniken oder aus Arztpraxen rekrutiert. Es ist sinnvoll, eine gute Verteilung von neu diagnostizierten Fällen (Inzidenzfälle) und bereits länger erkrankten Personen (Prävalenzfälle) zu nehmen. Die Wahl der Kontrollen muss wohl überlegt werden (nicht erkrankt an K, repräsentativ, vergleichbar, ...). Möglichkeiten für die Wahl der

Kontrollen: Populationsbasierter Ansatz, Krankenhausbasierter Ansatz, Kontrollen aus dem Umfeld der Fälle, Mehrere Kontrollgruppen.

8.2.1 Matchen (S. 268)

Problem bei Fall-Kontroll-Studien: es kann sein, dass Fälle nur junge Individuen sind und Kontrollen nur alte oder umgekehrt.

1. **Paarweises Matching**
   Dabei wird für jeden Einzelfall eine passende Kontrolle gesucht, die mit dem Fall in einigen relevanten Merkmalen übereinstimmt. So erhält man strukturgleiche Gruppen bezüglich der gematchten Merkmale.

2. **Gruppen-Matching**
   Hier wird versucht eine Kontrollgruppe zu finden, in der die Häufigkeitsverteilung der Merkmale ähnlich der Verteilung in der Fallgruppe ist.

8.2.2 Biasquellen (S. 269)

Es kann diverse systematische Fehler (Bias) geben:

1. **Selektionsbias**
   Wenn sich Fall- und Kontrollgruppen in wesentlichen Eigenschaften unterschieden. Lösung: Matchen

2. **Informations-Bias 1**
   Dies ist der Recall-Bias - Fälle, die von der Krankheit betroffen sind, können sich an Ereignisse besser erinnern als Kontrollen.

3. **Informations-Bias 2**
   Man verschweigt manchmal 'negative' Eigenschaften des Falles.

4. **Bias durch Confounder**
   Besonders groß. Confounder sind Einflussfaktoren, die zwar auf ein erhöhtes Risiko hinweisen, aber selbst nicht für das Entstehen der Krankheit verantwortlich sind.

8.2.3 Odds-Ratio (S. 271)

Zusammenhang zwischen 2 Alternativmerkmalen. zB Chi^2-4-Felder-Test. Das Odds-Ratio ist als Annäherung für das relative Risiko zu sehen.

a/c und b/d als Schätzwerte für die Odds (Chancen).

$$Odds = \frac{P(R|K)}{P(\bar{R}|K)}, \frac{P(R|\bar{K})}{P(\bar{R}|\bar{K})}$$

R ... Risikofaktor liegt vor. Schätzwert für die

$$OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc} \approx RR \text{ s.S. } 274$$

Bei paarweisem Matchen sollte der McNemar Test angewandt werden!

8.3 Kohortenstudien (S. 272)

Eine Kohortenstudie ist eine prospektive, longitudinale Studie, bei der eine große Gruppe (Kohorte) von Personen, die in unterschiedlicher Weise exponiert und nicht erkrankt sind, eine Zeit lang beobachtet werden.

8.3.1 Effektmaße (S. 273)

Erkrankungsrisiko: bei Vorliegen eines Faktors R = $P(K|R)$. Demnach ist $P(K|\bar{R})$ die Wahrscheinlichkeit zu erkranken, wenn der Faktor R nicht vorliegt.

Absolute Risikorediktion

$$ARR = P(K|R) - P(K|\bar{R})$$

Number Needed to Treat (NNT)

$$NNT = 1/ARR$$

Relatives Risiko / Odds Ratio

$$RR = \frac{P(K|R)}{P(K|\bar{R})}$$

Relative Risikorediktion

$$RRR = \frac{P(K|R) - P(K|\bar{R})}{P(K|R)}$$

8.3.2 Biasquellen (S. 276)

Studienabbrecher (Drop Outs) können zu einem Selektionsbias führen. Ein Informationsbias kann auftreten, wenn Teilnehmer die exponiert sind, häufiger untersucht werden als andere Personen.

9 Studien zu Therapie und Prognose

9.1 Prognosestudien (S. 306)

9.1.1 Begriffe

- Kohortenstudien
  Gruppe von Personen, die gleiche Voraussetzungen haben und beobachtet wird über einen zeitlichen Verlauf.
- Natürlicher Verlauf
  Was passiert, wenn ich die Person nicht behandle? (Die Prognose ohne medizinische Intervention)
- Klinischer Verlauf
  Bezeichnet die Prognose, wenn eine entsprechende Behandlung erfolgt.

- Letalität
  Wahrscheinlichkeit zu sterben, wenn man erkrankt.
- Mortalität
  Wahrscheinlichkeit zu erkranken **und** zu sterben.
- Rezidivrate
  Rückfall von Krankheit
- mediane Überlebenszeit
  Die Hälfte der Personen lebt weniger lang, die Hälfte länger
- Remission
  vorübergehendes Nachlassen von Krankheitserscheinungen
- 5-Jahres-Überlebensrate
  $s(5)$ oder $1 - F(5)$

9.1.2 Die Kaplan-Meier-Methode (S. 307)

Sie dient der Schätzung von Überlebenszeiten - bzw zum Schätzen von $s(t)$. Die Schätzung wir dann als $\hat{s}(t)$ bezeichnet. Üblicherweise kommt es bei Studien zu Studienabbrechern (Drop-Outs), die die Studienergebnisse verfälschen. Die Kaplan-Meier-Methode arbeitet aber mit einer dynamischen Stichprobe, die erst verringert wird, wenn Personen aussteigen. Die Information wird also nicht 'verschenkt'.

- Die Studie startet mit n Patienten. Diese Anzahl reduziert sich im Laufe der Zeit.
- Endereignisse werden mit $t_1 < t_2 < ... < t_k$ bezeichnet. Die Anzahl der Patienten zum Zeitpunkt jeweils mit $d_1, d_2, ..., d_k$.
- Anzahl der Patienten, die unmittelbar vor einem Zeipunkt t_i noch in der Studie involviert sind, werden mit n_i bezeichnet.

Beispiel: Annahme bei $t = 25$ leben noch 79 Personen. Ab $t = 28$ scheidet 1 Person aus; bei $t = 30$ scheiden 2 Personen aus; bei $t = 31$ stirbt die nächste Person.

$$\hat{s}(25) = 0,79 = \frac{79}{100}$$

s muss zwischen dem 25ten und 31ten Tag nicht verändert werden. Ab $t = 30$ sind noch 76 Personen lebend in der Studie.

$$\hat{s}(31) = \hat{s}(25) = \frac{75}{76}$$

Daher ist

$$n_i = (n_{i-1} - d_{i-1}) - c_i, i = 1...k$$
$$n_0 := n, d_0 := 0, (n_1 = n - c_1)$$

$t_0 := 0$	Beginn des Beobachtungszeitraums
$t_{k+1} := t_\omega$	Ende des Beobachtungszeitraums
c_i	Anzahl zensierter Beobachtungen (censured date) $[t_{i-1}, t_i)$
d_i	Anzahl kritischer Endereignisse zum Zeitpunkt t_i (deaths)
n_i	Anzahl Personen, die unmittelbar vor t_i noch beobachtet werden

Dann gilt:

$$\hat{s}(0) = 1$$

$$\hat{s}(t_i) = \hat{s}(t_{i-1})\frac{n_i - d_i}{n_i}$$

für $i = 1, ...k$ und

$$\hat{s}(t) = \hat{s}(t_i)$$

für $t \in [t_i, t_{i+1}), i = 0, 1, ...k$.

$$s(t_i) = P(x > t_i)$$

10 Übersicht

10.1 Verteilungen und deren Annäherungen

VT	Anwendung/Anmerkung	Charakteristika	Annäherung
$X \tilde{} B(n,p)$	Ziehen mit zurücklegen	$E(X) = n * p$ $Var(X) = n * p * (1-p)$ $\gamma_1 = \frac{q-p}{\sigma}$ $\sigma = \sqrt{n * p * (1-p)}$ $P(X = k) = \binom{n}{k}(p)^k(1-p)^{n-k}$	kann durch die Normalveteilung angenähert werden. $\mu = np$, $\sigma^2 = npq$ wenn $npq \geq 9$
$X \tilde{} P(\lambda)$	Die Poisson-Verteilung wird etwa bei epidemiologischen Untersuchungen eingesetzt, wo umfangreiche Populationen vorliegen, jedoch das Eintreten einer bestimmten Krankheit bei einem Individuum recht unwahrscheinlich ist. zB Geburten und Down-Syndrom ($n = 2000, p = 0,0001$)	$E(X) = n * p = \lambda$ $Var(X) = n * p * (1-p) = \lambda$ $\gamma_1 = \frac{1}{\sqrt{\lambda}}$ $\sigma = \sqrt{\lambda}$ $p = \frac{\lambda}{n}$ $P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$	Approx. der Binomial-Verteilung durch die Poisson-Verteilung wenn: $n \geq 30 \wedge p \leq 0,1$
Polynomial	Blutgruppen	$P(n_1, ..., n_k \vert p_1, ..., p_k) = \frac{(p_1)^{n_1} * ... * (p_k)^{n_k}}{n_1! * ... * n_k!} * n!$	-
$X \tilde{} NB(r,p)$	Diese Verteilung wird häufig zur Analyse von Wartezeiten verwendet.	$P(X = j) = \binom{j-1}{r-1}q^{j-r} * p^r$ $E(X) = np = n * \frac{M}{N}$ $Var(X) = \frac{N-n}{N-1} * n * p * (1-p)$	-

Tab. 2: Übersicht zu den Verteilungen

VT	**Anwendung/Anmerkung**	**Charakteristika**	**Annäherung**
$X \tilde{} HG(n; N, M)$	ziehen-**ohne**-zurücklegen!	$P(X = k) = \frac{\binom{M}{k} * \binom{N-M}{n-k}}{\binom{N}{n}}$ $E(X) = np = n * \frac{M}{N}$ $Var(X) = \frac{N-n}{N-1} * n * p * (1-p)$	Falls N im Vergleich zu n groß $\rightarrow$ Binomialverteilung

Tab. 3: Übersicht zu den Verteilungen

10.2 Tests

Hier soll nun kurz eine Zusammenfassung stehen, wann welcher Test zum Einsetzen ist und wie seine Teststatistik lautet.

Test	**Anwendung**	**Hypothesen**	**Teststatistik**
Logranktest	Vergleich der Verteilungen von Überlebenszeiten bei 2 unverbundenen Stichproben. Besonderheit: Berücksichtigung von zensierten Daten.	$H_0: \quad S_1 = S_2$ $H_1: \quad S_1 \neq S_2$	$\chi^2 = \frac{(d_1-e_1)^2}{e_1} + \frac{(d_2-e_2)^2}{e_2}$ mit $f = 1$ Freiheitsgraden. H_0 ablehnen wenn $\chi^2 > \chi^2_{1;1-\alpha}$

Tab. 4: Übersicht zu den Tests