Prüfungsordner Biometrie und Epidemiologie

Stand Juni 2005

Gesammelt und ausgearbeitet von Murmel (Murmel.vienna@gmx.at)
Fragen, die nicht wirklich gekommen sind, aber den gestellten Fragen ähnlich sind oder die von Prof. Hasibeder während der LVA genannt wurden und deshalb potentielle Prüfungsfragen darstellen, wurden mit einem * gekennzeichnet.
Unterlagen sind nicht erlaubt, einfache Taschenrechner schon. Wichtig! Geodreieck mitnehmen für Fragen zur Kaplan-Meier Methode! Neben diesem Prüfungsordner sollten besonders die Fragen aus Christel Weißs Fragenkatalog der Kapitel 9-11 wiederholt werden, 6 daraus machen 30% der Prüfung aus!

Kapitel 3

1. (5.10.2004) (25.01.2005) Geometrisches und harmonisches Mittel (S. 51)

a) Wann wird das geometrische Mittel statt dem arithmetischen Mittel verwendet?

Das geometrische Mittel wird bei relativen Änderungen verwendet, bei denen sich der Unterschied zweier Merkmalswerte besser durch einen Quotienten als durch eine Differenz beschreiben lässt. Dies ist der Fall bei Wachstumserscheinungen (z.B. die Zunahme der Bevölkerung in konstanten, aufeinander folgenden Zeiträumen oder die Zunahme der Unterhaltskosten in einer Klinik) sowie bei Verdünnungsreihen (z.B. Antikörperthiter).

b) Wie lautet die Formel für das geometrische Mittel?

Wenn x_i die relativen Änderungen bezeichnen (x_i >0 und dimensionslos), berechnet sich das geometrische Mittel als:

$$\bar{x}_G = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}$$

c*) Wann wird das harmonische Mittel statt dem arithmetischen Mittel verwendet?

Das harmonische Mittel dient als Lagemaß, wenn die Beobachtungswerte x_i Verhältniszahlen (also Quotienten) sind. Dieses Maß wird verwendet, wenn sich die x_i in ihren Nennern unterscheiden. Damit lassen sich etwa eine Durchschnittsgeschwindigkeit, eine Durchschnittsleistung oder eine mittlere Dichte berechnen.

d) Wie lautet die Formel für das harmonische Mittel?

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

e) Wie kann das geometrische Mittel mithilfe eines arithmetischen Mittels geeignet transformierter Beobachtungswerte angeschrieben und berechnet werden?

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

Beweis:

$$\ln \bar{x}_G = \frac{1}{n} \sum_{i=1}^{n} \ln x_i \Leftrightarrow e^{\ln \bar{x}_G} = e^{\frac{1}{n} \sum_{i=1}^{n} \ln x_i}$$

$$\Leftrightarrow \bar{x}_G = \sqrt[n]{e^{\sum_{i=1}^{n} \ln x_i}} = \sqrt[n]{e^{\ln x_1} \cdot \ldots \cdot e^{\ln x_n}} = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}$$

f*) Wie kann das harmonische Mittel mithilfe eines arithmetischen Mittels geeignet transformierter Beobachtungswerte angeschrieben und berechnet werden?

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

Beweis:

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

Kapitel 4

2*. Welche Arten von Scheinkorrelation gibt es? (S. 79)

Man unterscheidet:

- **Formale Korrelation**. Sie entsteht, wenn relative Häufigkeiten verglichen werden. Ein Beispiel dafür wäre der Vergleich von Grippeerkrankungen und chronischen Herz- und Kreislauferkrankungen. Da während einer Grippeepidemie die relative Häufigkeit der Grippeerkrankungen steigt, bewirkt dies automatisch, dass die relative Häufigkeit der Herz- und Kreislauferkrankungen sinkt. Der so ermittelte Korrelationskoeffizient ist deshalb kein geeignetes Maß für den tatsächlichen Zusammenhang.
- **Selektionskorrelation**. In der Stichprobe muss die gesamte Variationsbreite der zu untersuchenden Merkmale vertreten sein. Wenn man jedoch bei der Wahl der Beobachtungseinheiten selektiert, ergibt sich eine Korrelation, die nicht die Verhältnisse in der Grundgesamtheit widerspiegeln.
- **Korrelation durch Ausreißer**. Ein Ausreißer kann einen betragsmäßig hohen Korrelationskoeffizienten verursachen. Dies kann durch einen Blick auf die Punktwolke vermieden werden.
- **Inhomogenitätskorrelation**. Sie ergibt sich, wenn für zwei inhomogene Gruppen ein gemeinsamer Korrelationskoeffizient berechnet wird. Die graphische Darstellung würde dabei aus zwei Punktwolken bestehen, die sich nicht überlappen.
- **Gemeinsamkeitskorrelation**. Wenn zwei Merkmale durch ein drittes beeinflusst werden, liegt eine Gemeinsamkeitskorrelation vor.

3. (14.10.2003) (23.11.2004) (S. 87)
Regression: erklärte Varianz, Residualvarianz, Gesamtvarianz
a) Geben Sie die Formeln für jede dieser Größen an.

Erklärte Varianz = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Residualvarianz = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Gesamtvarianz = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

b) Wie hängen diese zusammen?

Es gilt:
Gesamtvarianz = Residualvarianz + erklärte Varianz

c) Was ist das Bestimmtheitsmaß?
Das Bestimmtheitsmaß (auch Determinationskoeffizient) r² ist der Anteil der erklärten Varianz an der Gesamtvarianz. Man berechnet es demnach mittels der Formel:

$$r^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

d) Welche Extremwerte kann es annehmen und was bedeutet dies?
Da die erklärte Varianz mindestens gleich 0 und höchstens so groß wie die Gesamtvarianz ist, erstreckt sich der Wertebereich des Bestimmtheitsmaßes r² zwischen 0 und 1. Im Extremfall r²=1 ist die Residualvarianz gleich 0. Im anderen Extremfall r² = 0 ist die Beschreibung mittels einer Regressionsgeraden sinnlos.

4*. Was bedeutet es, wenn der Assoziationskoeffizient nach Yule Q kleiner, gleich oder größer als 0 ist? (S. 93)
Folgende 4Feldertafel:

	B_1	B_2	$\sum$
A_1	n_{11}	n_{12}	$n_{1.}$
A_2	n_{21}	n_{22}	$n_{2.}$
$\sum$	$n_{.1}$	$n_{.2}$	n

$$Q = \frac{n_{11} * n_{22} - n_{12} * n_{21}}{n_{11} * n_{22} + n_{12} * n_{21}}$$

Für Q=0 ist kein Zusammenhang zwischen den Merkmalen nachzuweisen.
Im Fall Q>0 ist in der Stichprobe:
- die Merkmalsart B2 unter den Individuen mit Merkmalsart A2 häufiger als unter den Individuen mit Merkmalsart A1
- die Merkmalsart A2 unter den Individuen mit Merkmalsart B2 häufiger als unter den Individuen mit Merkmalsart B1

Im Fall Q<0 ist in der Stichprobe:
- die Merkmalsart B2 unter den Individuen mit Merkmalsart A2 seltener als unter den Individuen mit Merkmalsart A1
- die Merkmalsart A2 unter den Individuen mit Merkmalsart B2 seltener als unter den Individuen mit Merkmalsart B1

Kapitel 5

5. (22.6.2004) Tschebyscheff'sche Ungleichung (S. 122)
a) Geben Sie die Formel dafür an.

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2} \forall k > 0$$

b) Für welche Verteilungen hat Gauß eine schärfere Ungleichung nachgewiesen?
Gauß hat für symmetrische, eingipfelige Verteilungen eine schärfere Ungleichung nachgewiesen.

c) Geben Sie die Formel für diese schärfere Ungleichung an.

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

Kapitel 6

6. (5.10.2004) (14.10.2003) Sterbetafel (S. 132)

a) Was stellen die folgenden Größen dar: l0, lx, dx. Geben Sie auch die entsprechenden Formeln an.

l0 ist die Anzahl der Lebendgeborenen innerhalb eines Beobachtungszeitraums (z.B. in einem bestimmten Jahr). (Keine Formel, Konstante)

lx ist die Anzahl der Personen, die ihren x-ten Geburtstag erleben und danach noch unbestimmte Zeit leben. (Keine Formel, Konstante)

dx ist die Anzahl der Lebendgeborenen, die zwischen ihrem x-ten und (x+1)ten Geburtstag sterben.

$$d_x = l_x - l_{x+1}$$

b) Wie kann man qx, e0 und ex interpretieren? Geben Sie auch die entsprechenden Formeln an.

qx ist die Sterbeziffer, das ist die Wahrscheinlichkeit, dass jemand, der seinen x-ten Geburtstag erlebt hat, vor seinem (x+1)ten Geburtstag stirbt.

$$q_x = \frac{d_x}{l_x} \quad \text{x} = 0,\ldots,\text{w}$$

e0 ist die durchschnittliche Lebenszeit (oder Lebenserwartung) eines Neugeborenen.

$$e_0 = \frac{1}{2} + \frac{1}{l_0} \sum_{x=1}^{\omega} l_x$$

ex ist demnach die durchschnittliche (verbleibende) Lebenszeit eines x-Jährigen.

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

c) Leiten Sie die Formel für ex her. (s. auch Aufgabe 6)

Wie e0 auch ist ex als Erwartungswert die Summe des Produkts der erreichbaren Alter mit der Wahrscheinlichkeit, in diesem Jahr zu sterben (also der Anzahl der im spezifischen Jahr [von x bis Omega] Sterbenden durch die im Jahr x Geborenen) + ½ (weil man nicht wissen kann ob die Person am Anfang oder am Ende des entsprechenden Jahres stirbt und so einfach den Mittelwert nimmt).

Hier zieht man allerdings nochmals x ab, denn das Alter x wurde bereits erreicht (wir wollen die Anzahl Jahre, die die Person noch zu leben hat, nicht das wahrscheinliche Alter, das sie erreicht). Aus diesen Überlegungen ergibt sich:

$$e_x = x * \frac{d_x}{l_x} + (x+1) * \frac{d_{x+1}}{l_x} + (x+2) * \frac{d_{x+2}}{l_x} + \ldots + \omega * \frac{d_\omega}{l_x} + \frac{1}{2} - x = \frac{1}{l_x} \sum_{y=x}^{\omega} \left(y * d_y\right) + \frac{1}{2} - x$$

Laut Definition ist $d_x = l_x - l_{x+1}$ und daher gilt:

$$e_x = \frac{1}{l_x}\sum_{y=x}^{\omega}\left(y*(l_y - l_{y+1})\right)+\frac{1}{2}-x = \frac{1}{l_x}*\left[x*(l_x - l_{x+1})+(x+1)*(l_{x+1}-l_{x+2})+...+\omega*(l_\omega - l_{\omega+1})\right]+\frac{1}{2}-x$$

Wir wissen, dass $l_{\omega+1}=0$ (denn laut Definition ist ω das maximal erreichbare Alter [ca. 120 Jahre], die Anzahl der Personen, welche das Alter ω+1 erreichen, ist daher gleich 0). Dies eingesetzt und die Gleichung ausmultipliziert ergibt:

$$e_x = \frac{1}{l_x}*\left[x*l_x - x*l_{x+1}+x*l_{x+1}-x*l_{x+2}+l_{x+1}-l_{x+2}+...+\omega*l_{\omega-1}-\omega*l_{\omega-1}-\omega*l_\omega+l_\omega+\omega*l_\omega\right]+\frac{1}{2}-x$$

Wie man sieht, heben sich hier viele Terme auf, sodass nach dem Wegstreichen nur folgendes übrig bleibt:

$$e_x = \frac{1}{l_x}*\left[x*l_x+l_{x+1}+...+l_\omega\right]+\frac{1}{2}-x = x+\frac{1}{l_x}\left[l_{x+1}+...+l_\omega\right]+\frac{1}{2}-x = \frac{1}{l_x}\sum_{y=x+1}^{\omega}l_y+\frac{1}{2}$$

7. (26.6.2003) (5.10.2004) (14.10.2003 -> positiv) (23.11.2004) (25.01.2005 -> P(K-|T+)) (S. 136)

a) Geben Sie die formelle (symbolische) Schreibweise für den positiven/negativen Vorhersagewert / für die Wahrscheinlichkeit, dass eine Krankheit nicht vorhanden ist, obwohl das Testergebnis positiv ist an. (S.139)

Ereignis	Bezeichnung der Wahrscheinlichkeit	formelle Schreibweise
Krankheit liegt vor	Prävalenz (a-priori-Wahrscheinlichkeit)	$P(K)$
Testergebnis richtig positiv	Sensitivität	$P(T_+ \mid K)$
Testergebnis falsch negativ	---	$P(T_- \mid K)$
Testergebnis richtig negativ	Spezifität	$P(T_- \mid \overline{K})$
Testergebnis falsch positiv	---	$P(T_+ \mid \overline{K})$
Krankheit liegt vor, falls Testergebnis positiv	positiver Vorhersagewert (a-posteriori-Wahrscheinlichkeit)	$P(K \mid T_+)$
Krankheit liegt nicht vor, falls Testergebnis negativ	negativer Vorhersagewert	$P(\overline{K} \mid T_-)$

b) Geben Sie die Formel zu dessen Berechnung mit dem Bayes'schen Theorem an. (6.12)
(Beispiel positiver Vorhersagewert, andere analog)

$$P(K \mid T_+) = \frac{P(K \cap T_+)}{P(T_+)} = \frac{P(T_+ \cap K)}{P(T_+ \cap K)+P(T_+ \cap \overline{K})} = \frac{P(T_+ \mid K)\cdot P(K)}{P(T_+ \mid K)\cdot P(K)+P(T_+ \mid \overline{K})\cdot P(\overline{K})}$$

c) Wie sieht die Formel aus, wenn man sie nur mehr mit Prävalenz, Spezifität und Sensitivität darstellt?

Die wichtigen Umformungsregeln für diese Berechnung sind:

$P(T_- \mid K) = 1 - P(T_+ \mid K)$

$P(T_+ \mid \overline{K}) = 1 - P(T_- \mid \overline{K})$

$P(K) = 1 - P(\overline{K})$

Durch Einsetzen ergibt sich:

$$P(K \mid T_+) == \frac{P(T_+ \mid K) \cdot P(K)}{P(T_+ \mid K) \cdot P(K) + [1 - P(T_- \mid \overline{K})] \cdot [1 - P(K)]}$$

8*. Wann legt man Wert auf eine hohe Sensitivität/Spezifität? (S. 140-141)

Auf eine hohe Sensitivität legt man Wert, wenn

- es sich um eine Krankheit mit schlimmen (oder gar lebensbedrohlichen) Folgen für den Patienten handelt
- eine erfolgsversprechende Therapie zur Verfügung steht
- falsch-positive Befunde mit vertretbarem Aufwand und ohne allzu große Belastungen für die betreffende Person geklärt werden können

Eine hohe Spezifität ist anzustreben, wenn

- keine Therapie mit Aussicht auf Besserung oder Heilung bekannt ist,
- die Therapie zu unverhältnismäßig hohen finanziellen Belastungen für den Patienten oder das Gesundheitswesen führt,
- die Therapie mit schweren Nebenwirkungen behaftet ist,
- die Nachfolgeuntersuchungen mit erheblichen Risiken oder psychischen Belastungen für den Patienten verbunden sind

Kapitel 7

9. (26.6.2003) (S.151)

a) Wie nennt man diese Verteilung: $P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$ **?**

Dies ist die Poissonverteilung.

b*) Leiten Sie diese Verteilung her.

$$P = \frac{\lambda}{n}$$

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} \cdot \frac{\lambda^k}{n^k} \cdot (1 - \frac{\lambda}{n})^{n-k}$$

Wenn $n \to \infty$

$$P(X = k) = \frac{1}{k!} \cdot \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{n^k} \cdot \lambda^k \cdot (1 - \frac{\lambda}{n})^n \cdot (1 - \frac{\lambda}{n})^{-k}$$

und

$$\lim \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{n^k} = \lim 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \ldots \cdot \left(1 - \frac{k-1}{n}\right) = 1$$

$n \to \infty$ $n \to \infty$

$$\lim \left(1 - \frac{\lambda}{n}\right)^n = \lim \left(1 + \frac{-\lambda}{n}\right)^n = e^{-\lambda} \text{ (s. Mathe 1)}$$

$n \to \infty$ $n \to \infty$

$$\lim \left(1 - \frac{\lambda}{n}\right)^{-k} = 1 \Rightarrow \lim P(x = k) = \frac{1}{k!} \cdot 1 \cdot \lambda^k \cdot e^{-\lambda} \cdot 1 = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

$n \to \infty$ $n \to \infty$

c) Geben Sie die Formeln für den Erwartungswert μ und die Varianz s^2 der Verteilung an.

$$E(X) = \lambda = n \cdot p$$

$$Var(X) = s^2 = n \cdot p \cdot (1-p) = \lambda \cdot \left(1 - \frac{\lambda}{n}\right) \to \lambda$$
$$n \to \infty$$

d) Geben Sie die Formel für die Schiefe γ_1 an.

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

Ist die Verteilung symmetrisch, linkschief, rechtschief oder hat sie unterschiedliche Schiefe in Abhängigkeit von q?
Wie obige Formel beweist, ist die Schiefe immer > 0, die Poissonverteilung ist also immer rechtsschief.

10. (14.10.2003) Verteilung (S. 154)
a) Welche Verteilung beschreibt, mit was für einer Wahrscheinlichkeit nach j Beobachtungen das Ereignis A r-mal eingetreten ist?
Die negative Binomialverteilung beschreibt dies.

b) Wie ist deren formale Schreibweise?
X: NB(r,p)

c) Ist r = 1, handelt es sich um einen Spezialfall. Wie nennt man diesen? Schreiben Sie dessen Formel auf.
Dieser Spezialfall ist die geometrische Verteilung. Seine Formel ist:
$P(X = j) = q^{j-1} \cdot p = (1-p)^{j-1} \cdot p$

d) Geben Sie die Formel für die Verteilung an, was muss für j in Bezug auf r gelten?

$$P(X = j) = \binom{j-1}{r-1} \cdot q^{j-r} \cdot p^r$$

mit $j \geq r$

11. (5.10.2004) Geben Sie die Intervalle und Wahrscheinlichkeiten bei der Normalverteilung für 1-sigma-bereich, 2-sigma-bereich und 3-sigma-bereich an (S. 163)

Intervallgrenzen für X: N(μ,σ²)	Intervallgrenzen für Z: N(0,1)	Bezeichnung des Intervalls	Wahrscheinlichkeit P
$\mu-\sigma \leq X \leq \mu+\sigma$	$-1 \leq Z \leq 1$	1σ-Bereich	0,6827
$\mu-2\sigma \leq X \leq \mu+2\sigma$	$-2 \leq Z \leq 2$	2σ-Bereich	0,9545
$\mu-3\sigma \leq X \leq \mu+3\sigma$	$-3 \leq Z \leq 3$	3σ-Bereich	0,9973

12. (26.6.2003) Geben Sie die für medizinische Fragenstellungen wichtigen Referenzbereiche der Normalverteilung:
a) 95%
b) 99%
an. (S.163)

Intervallgrenzen für X: N(μ,σ²)	Intervallgrenzen für Z: N(0,1)	Bezeichnung des Intervalls	Wahrscheinlichkeit P
μ-1,96σ ≤ X ≤ μ+1,96σ	-1,96 ≤ Z ≤ 1,96	95%-Referenzbereich	0,95
μ-2,58σ ≤ X ≤ μ+2,58σ	-2,58 ≤ Z ≤ 2,58	99%-Referenzbereich	0,99

13. (26.6.2003)
Man kann nach dem zentralen Grenzwertsatz eine Binomialverteilung für hinreichend großes n durch eine Normalverteilung X ~ N(μ, s²) approximieren. (S.168)
a) Unter welcher Bedingung gilt dies (Faustregel)?
Als Faustregel gilt, dass die Ungleichung npq ≥ 9 erfüllt sein muss.

b) Berechnen Sie die Parameter μ und s² bei der approximierten Normalverteilung.

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

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

Daher X ~ N(np, np(1-p))

14. (26.6.2003) (23.11.2004) Lebensdauern (S. 175)
a) Was stellt r(t) dar und was ist dessen Formel? (7.31)
r(t) ist die momentane Sterberate (auch Ausfallrate). Ihre Formel ist: $r(t) = \frac{f(t)}{S(t)}$

b) Leiten sie r(t) mit den Definitionen von S(t) und f(t) her. Stellen Sie es jeweils alleine durch f(t), F(t) und S(t) dar. Wie können S(t) bzw. F(t) bzw f(t) alleine durch r(t) ausgedrückt werden?

$r(t) = \frac{f(t)}{S(t)}$ wobei f(t)=F'(t) und S(t) = 1 – F(t)

S(t) ist hierbei die Survivalfunktion P(X>t), F logischerweise die Wahrscheinlichkeit, mit der ein Individuum vor t stirbt: P(X≤t).
Daraus folgt: f(t) = -S'(t) und wegen F(0) = 0 bzw S(0) = 1 gilt

$$F(t) = \int_0^t f(\tau)d\tau \text{ und } S(t) = 1 - \int_0^t f(\tau)d\tau$$

Darstellung von r(t) durch s(t) oder F(t) oder f(t): Es ergibt sich

$$r(t) = \frac{-S'(t)}{S(t)} = -(\ln S(t))'$$

$$r(t) = \frac{F'(t)}{1-F(t)}$$

$$r(t) = \frac{f(t)}{1 - \int_0^t f(\tau)d\tau}$$

Ausdrücken der Parameter in Funktion von r(t):
Mit der Notation τ statt t ist

$-r(\tau) = \frac{S'(\tau)}{S(\tau)}$ Daraus folgt (mit S(t)>0)

$$-\int_0^t r(\tau)d\tau = \int_0^t \frac{S'(\tau)}{S(\tau)}d\tau = \ln S(t) - \ln S(0) = \ln S(t) - \ln 1 = \ln S(t)$$

Damit gilt:

$$S(t) = e^{-\int_0^t r(\tau)d\tau}$$

$$F(t) = 1 - e^{-\int_0^t r(\tau)d\tau}$$

$$f(t) = r(t) \cdot e^{-\int_0^t r(\tau)d\tau}$$

c*) Schreiben Sie auf, wie die Sterberate durch die bedingte Sterbewahrscheinlichkeit hergeleitet wird.

Bedingte Sterbewahrscheinlichkeit:

$$P(t < X < t + \Delta t \mid X > t) = \frac{P(t < X < t + \Delta t \cap X > t)}{P(X > t)} = \frac{P(t < X < t + \Delta t)}{P(X > t)} = \frac{F(t + \Delta t) - F(t)}{S(t)}$$

(vom 2ten auf den 3ten Schritt gilt weil die rechte Menge in der Vereinigung sowieso Teilmenge der linken ist)

Die momentane Sterberate lässt sich schreiben als:

$$r(t) = \lim_{\Delta t \to 0} \frac{P(t < X < t + \Delta t \mid X > t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{S(t)} \cdot \frac{1}{\Delta t} = \frac{1}{S(t)} \cdot \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t}$$

$$\Leftrightarrow r(t) = \frac{1}{S(t)} \cdot F'(t) = \frac{f(t)}{S(t)}$$

15. (14.10.2003) (23.11.2004) Weibull-Verteilung als Verallgemeinerung der Exponentialverteilung (S. 170 - 175)

a) Geben Sie die Formel der Dichtefunktion f(t) der Exponentialverteilung, sowie die Formeln des Medians $\tilde{\mu}$, des Erwartungswertes (= mittlere Lebensdauer) μ und der Varianz σ^2 an.

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

$$\tilde{\mu} = \frac{1}{\lambda} \cdot \ln 2$$

$$\mu = \frac{1}{\lambda}$$

$$\sigma^2 = \frac{1}{\lambda^2}$$

b) Geben Sie die Verteilungsfunktion F(t) der Weibull-Verteilung an.

$F(t) = 1 - e^{-\lambda \cdot t^{\gamma}}$ für t > 0

c) Welche Fälle unterscheidet man bezüglich der Sterberate und was beschreiben sie?

Man unterscheidet:

- Sterberate konstant (γ=1). Dieser Spezialfall ist die Exponentialverteilung.
- Sterberate monoton wachsend (γ>1). Eine Weibullverteilung mit γ>1 ist geeignet, um ein Altern zu beschreiben.
- Sterberate monoton fallend (0<γ<1). Hierbei nimmt mit wachsendem Alter die Sterberate ab. Diese Verteilung beschreibt ein Überleben mit Regeneration.

Kapitel 8

16. (22.6.2004) (25.01.2005) Erwartungstreue/Konsistenz (S. 187)

Schreiben Sie die folgenden Aussagen symbolisch auf und leiten Sie sie her:

a) S^2 ist erwartungstreuer Schätzer für σ^2

$E(S^2) = \sigma^2$

Herleitung:

$$E(S^2) = E\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}\right) = E\left(\frac{\sum_{i=1}^{n}(X_i - \mu)^2 - n\cdot(\overline{X} - \mu)^2}{n-1}\right)$$

Aus der Definition der Varianz folgt:

$$E(X_i - \mu)^2 = Var(X_i) = \sigma^2$$

und damit

$$E\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) = n\cdot\sigma^2$$

Außerdem gilt:

$$E(\overline{X} - \mu)^2 = Var(\overline{X}) = \frac{\sigma^2}{n}$$

Einsetzen ergibt:

$$E(S^2) = \frac{n\cdot\sigma^2 - \sigma^2}{n-1} = \sigma^2$$

b) S^2 ist konsistent

$$Var(S^2) = \frac{2\sigma^4}{n-1} \to 0$$

$$n \to \infty$$

Herleitung:

Wir wissen, dass:

$$\frac{n-1}{\sigma^2}\cdot S^2 \sim \chi^2_{n-1}$$

und

$$Var\left(\frac{n-1}{\sigma^2}\cdot S^2\right) = 2(n-1)$$

Daraus folgt:

$$Var\left(S^2\right) = \left(\frac{\sigma^2}{n-1}\right)^2 \cdot 2(n-1) = \frac{2\sigma^4}{n-1} \to 0$$

$$n \to \infty$$

c) S ist kein erwartungstreuer Schätzer für σ

$E(S) \neq \sigma$

Herleitung:

Wir wissen (nach 5.26), dass:

$$Var\left(S\right) = ES^2 - (ES)^2 = \sigma^2 - (ES)^2$$

Daraus folgt:

$$ES = \sqrt{\sigma^2 - Var(S)} < \sigma$$

weil $Var(S) > 0$ und damit $E(S) \neq \sigma$

17. (26.6.2003) (25.01.2005) (22.6.2004) Rechenbeispiel (S. 191)
Nach einer Organtransplantation wurden bei 20 Patienten die Überlebenszeiten in Tagen ermittelt. Bei 8 Patienten konnten zu folgenden Zeitpunkten des kritischen Endereignisses ermittelt werden: 12, 39, 75,12, 105, 85, 90, 151. 6 Patienten schieden während der Beobachtungszeit (170 Tage) zu folgenden Zeitpunkten aus der Studie aus: 20, 49, 75, 20, 88, 120. Die restlichen Patienten überlebten die Beobachtungszeit. Schätzen Sie S(t) und stellen Sie dies
a) tabellarisch mit 3 Spalten (Anfangszeitpunkt, Endzeitpunkt, Schätzwert)
Die wichtigen Formeln für die Berechnung der Schätzfunktion sind:

$\hat{S}(t) = 1$ für t € [0,t1)

$\hat{S}(t) = \frac{n_1 - d_1}{n_1}$ für t € [t1,t2)

$\hat{S}(t) = \frac{n_1 - d_1}{n_1} \cdot \frac{n_2 - d_2}{n_2}$ für t € [t2,t3)

…

$\hat{S}(t) = \frac{n_1 - d_1}{n_1} \cdot \frac{n_2 - d_2}{n_2} \cdot \ldots \cdot \frac{n_i - d_i}{n_i}$ für t € [ti,t(i+1)) und i = 1,2,…,k

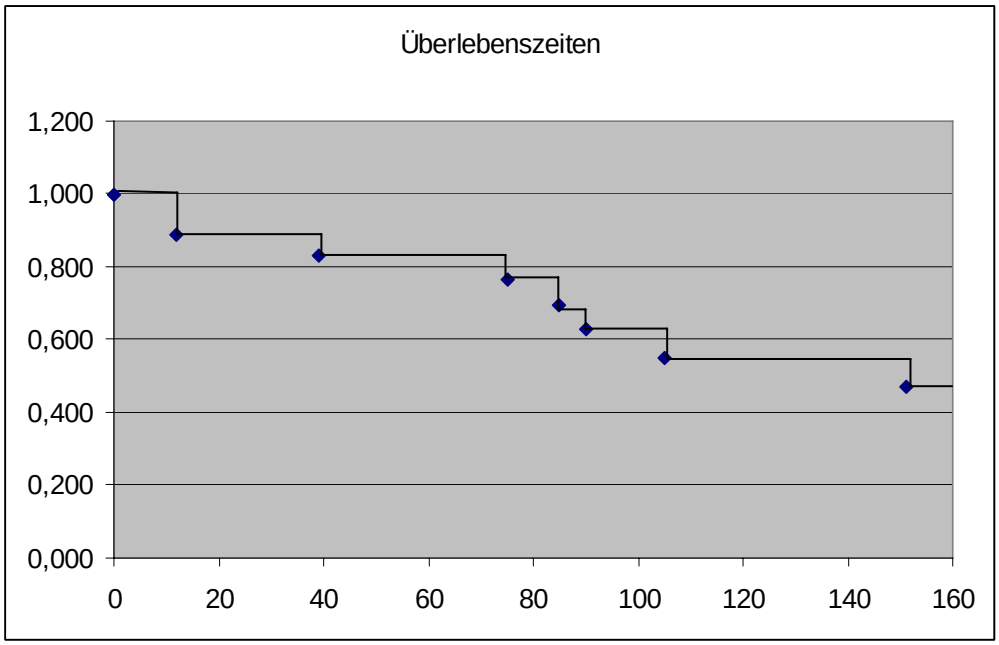
Wird eine Person zum Zeitpunkt t zensiert und ist t ein Intervallende, so wird die Person erst im Folgeintervall (also nach dem Zeitpunkt t) zensiert.
Hilfstabelle:

t_i	t_{i+1}	c	n	d	n-d	$\hat{S}(t)$
0	11	0	20	0	20	1,000
11	12	2	18	2	16	0,889
12	39	1	15	1	14	0,830
39	75	1	13	1	12	0,766
75	85	1	11	1	10	0,696
85	90	0	10	1	9	0,627
90	105	1	8	1	7	0,548
105	151	0	7	1	6	0,470
151	170	6	0	0	0	0,403

Daraus ergibt sich die folgende Tabelle zu drei Spalten:

t_i	t_{i+1}	$\hat{S}(t)$
0	11	1,000
11	12	0,889
12	39	0,830
39	75	0,766
75	85	0,696
85	90	0,627
90	105	0,548
105	151	0,470
151	170	0,403

b) grafisch (Einheit Abszisse: 1 mm, Einheit Ordinate: 100 mm) dar.
Anm: Auf die Einheiten wurde keine Rücksicht genommen. „Einheit Abszisse: 1mm“ bedeutet, dass eine Einheit auf der x-Achse (also ein Tag) einem Millimeter entsprechen soll, „Einheit Ordinate: 100mm“ dementsprechend, dass ein Schätzwert von 1 10cm einnimmt (also 0,1 auf der y-Achse einem cm entsprechen). 1,2 sollte natürlich nicht aufscheinen (Excel-Fehler).



Kapitel 9

18. (23.11.2004) (5.10.2004) (25.01.2005) Wann wird H0 auf einem Signifikanzniveau alpha beibehalten, wann wird H1 angenommen, wenn man den p-Wert kennt? *Was ist der p-Wert? (S. 211)
Der p-Wert quantifiziert die Wahrscheinlichkeit, dass das gefundene Ergebnis (oder ein noch extremeres Ergebnis) zustande kommt, wenn in Wirklichkeit die Nullhypothese richtig ist. Wenn p kleiner ist als das festgelegte Signifikanzniveau, wird die Alternativhypothese angenommen.

19. (5.10.2004) (26.6.2003) Zweiseitiger t-Test für eine Stichprobe mit
H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$
(S. 213)
a) In welchem Fall wird H0 beim t-Test beibehalten, in welchem Fall H1 angenommen?
Man bestimmt die Prüfgröße und den kritischen Wert für den Test und erhält dadurch ein Ergebnis:
Wenn die Prüfgröße im kritischen Bereich liegt, entscheidet man sich für H1.
Wenn die Prüfgröße im Annahmebereich liegt, entscheidet man sich für H0.
Dies bedeutet, dass H0 beibehalten wird, wenn der Absolutbetrag der Testgröße kleiner als der kritische Wert ist.

b) Erklären sie möglichst genau im Speziellen den Zusammenhang zwischen Testergebnis und Konfidenzintervall.
Die Prüfgröße liefert keinen Hinweis darauf, wie groß der unbekannte Parameter μ ist. Es bietet sich deshalb an, anhand des Mittelwerts ein auf der t-Verteilung basierendes Konfidenzintervall zu berechnen.

c) Leiten Sie das Konfidenzintervall für oben angegebenen Test her.
Für die zweiseitige Fragestellung wird angenommen: $\alpha 1 = \alpha 2 = \alpha/2$
Bei der einseitigen Fragestellung wäre es je nach Richtung $\alpha 1 = \alpha$ und $\alpha 2 = 0$ oder umgekehrt.

Es gilt:

$$t_{n-1;\alpha 1} \le T \le t_{n-1;1-\alpha 2}$$

$$\Leftrightarrow t_{n-1;\alpha 1} \cdot \frac{s}{\sqrt{n}} \le \overline{X} - \mu_0 \le t_{n-1;1-\alpha 2} \cdot \frac{s}{\sqrt{n}}$$

$$\Leftrightarrow \overline{X} - t_{n-1;1-\alpha 2} \cdot \frac{s}{\sqrt{n}} \le \mu_0 \le \overline{X} - t_{n-1;\alpha 1} \cdot \frac{s}{\sqrt{n}} \text{ (Achtung, durch multiplizieren mit -1 verdreht)}$$

$$\Leftrightarrow \overline{X} - t_{n-1;1-\alpha 2} \cdot \frac{s}{\sqrt{n}} \le \mu_0 \le \overline{X} + t_{n-1;1-\alpha 1} \cdot \frac{s}{\sqrt{n}} \text{ (weil } t_{n-1;\alpha 1} = -t_{n-1;1-\alpha 1})$$

$$\Leftrightarrow \mu_0 \in \left[\overline{X} - t_{n-1;1-\alpha 2} \cdot \frac{s}{\sqrt{n}}; \overline{X} + t_{n-1;1-\alpha 1} \cdot \frac{s}{\sqrt{n}} \right]$$

d) Wie lautet die Testgröße t?

$$t = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$$

20.(23.11.2004) t-Test für zwei unverbundene Stichproben (S. 218)
a) Was sind die Prämissen/Vorraussetzungen dieses Tests, wie ist insbesondere die Prüfgröße bei Zutreffen von H0 verteilt?
Die Prämissen des Tests sind folgende:

- Es liegen zwei unverbundene Stichproben der Umfänge n1 und n2 vor;
- Die Daten beider Stichproben entstammen normalverteilter Grundgesamtheiten mit derselben Varianz, also X: $N(\mu 1,\sigma^2)$ und Y: $N(\mu 2,\sigma^2)$

Beide Verteilungen sollten also dieselbe Form haben und sich höchstens bezüglich ihrer Erwartungswerte unterscheiden. Wichtig um gültige Resultate zu erhalten ist ebenfalls, dass:

- beide Stichprobenumfänge mindestens 10 (bei nicht symmetrischen Verteilungen 20) betragen und ähnlich groß sind und
- die Zufallsvariablen X und Y ungefähr denselben Verteilungstyp haben. Dies lässt sich über die empirischen Kenngrößen oder eine graphische Darstellung überprüfen.

Bei Zutreffen von H0 ($\mu 1 = \mu 2$) ist die Testgröße folgendermaßen verteilt:

$$\overline{X} - \overline{Y} \sim N\left(0, \sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

b*) Leiten Sie diese Verteilung und damit auch die Prüfgröße t her.

Es gilt $\overline{X} = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} X_i$ und $\overline{Y} = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} Y_i$

Wobei alle $X_i \sim N(\mu_1, \sigma^2)$, $Y_i \sim N(\mu_2, \sigma^2)$

und alle X_i, Y_i unabhängig

$$\Rightarrow \overline{X} \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right), \overline{Y} \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right)$$

$$\Rightarrow E(\overline{X} - \overline{Y}) = E(\overline{X}) + E((-1) \cdot \overline{Y}) = E(\overline{X}) + (-1) \cdot E(\overline{Y}) = \mu_1 - \mu_2$$

Für die Varianz gilt:

$$\Rightarrow Var(\overline{X}-\overline{Y}) = Var\overline{X} + Var((-1)\cdot\overline{Y}) = Var\overline{X} + (-1)^2 \cdot Var(\overline{Y}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Daher gilt:

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Ist H0 erfüllt, so gilt $\mu_1 = \mu_2$ und damit $\mu_1 - \mu_2 = 0$
Daraus ergibt sich:

$$\overline{X} - \overline{Y} \sim N\left(0, \sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Daraus ergibt sich auch die Prüfgröße:

$$t = \frac{\overline{x} - \overline{y} - 0}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

c) Was stellt s² in diesem Test dar? Schreiben Sie die Formeln für dieses s² und die Prüfgröße t an. In welchem Fall wird H0 <u>beibehalten</u>? Schreiben Sie die vereinfachten Formeln von t und s² an, wenn der Umfang der beiden Stichproben gleich ist (n=n1=n2).
S² ist die „gepoolte" Varianz, die aus den Werten beider Stichproben berechnet wird. Sie lässt sich mit den beiden empirischen Varianzen s_1^2 und s_2^2 schreiben als:

$$s^2 = \frac{(n_1 - 1)\cdot s_1^2 + (n_2 - 1)\cdot s_2^2}{n_1 + n_2 - 2}$$

Für die Prüfgröße t gilt:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

Hierbei wird H0 <u>beibehalten</u>, falls $|t| < t_{f;1-\alpha/2}$ (bei 2-seitiger Fragestellung) oder $|t| < t_{f;1-\alpha}$ (bei 1-seitiger Fragestellung). Hierbei ist f = n1 + n2 - 2 die Anzahl der Freiheitsgrade.

Bei gleichen Stichprobenumfängen n = n1 = n2 vereinfachen sich diese Formeln zu:

$$t = \frac{\overline{x} - \overline{y}}{s \cdot \sqrt{2/n}}$$

$$s^2 = \frac{s_1^2 + s_2^2}{2}$$

d*) Was sind die Grenzen des 2-seitigen Konfidenzintervalls?
Achtung! Hier ist ein Fehler im Buch. Die Grenzen des 2-seitigen Konfidenzintervalls für $\mu_1 - \mu_2$ sind:

$$\overline{x} - \overline{y} \pm t_{n_1+n_2-2;1-\alpha/2} \cdot s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Kapitel 11

21. (5.10.2004) Kohortenstudien (S. 277)

a) Zeichnen Sie die Vierfeldertafel auf, welche Zahlen sind darin vorgegeben?

	Ausgangssituation		
	R	$\overline{R}$	$\sum$
K	a	b	a+b
$\overline{K}$	c	d	c+d
$\sum$	a+c	b+d	n

R bezeichnet das Exponiertsein gegenüber einem bestimmten Risikofaktor und K das Auftreten der zu untersuchenden Krankheit.
Demnach kann man die Zahlen folgendermaßen interpretieren:
a ist die Anzahl Personen, die dem Risiko ausgesetzt werden und krank werden.
b ist die Anzahl Personen, die nicht dem Risiko ausgesetzt werden, aber krank werden.
c ist die Anzahl Personen, die dem Risiko ausgesetzt werden, aber nicht krank werden.
d ist die Anzahl Personen, die weder dem Risiko ausgesetzt werden, noch krank werden.

b) Was kann aus den Zahlen in dieser Tafel geschätzt werden? Schreiben sie diese Werte symbolisch und verbal auf.

Aus den Zahlen können folgende Werte geschätzt werden:

- die Inzidenz $\frac{a+b}{n}$

- die bedingte Wahrscheinlichkeit, bei Vorliegen des Risikofaktors zu erkranken:

$$P(K \mid R) = \frac{a}{a+c}$$

- das Risiko, krank zu werden, wenn der Risikofaktor nicht vorliegt:

$$P(K \mid \overline{R}) = \frac{b}{b+d}$$

c) Wie lautet die Formel für den Schätzwert?

Als Schätzwert für das relative Risiko ergibt sich:

$$\rho = \frac{P(K \mid R)}{P(K \mid \overline{R})} = \frac{a \cdot (b+d)}{b \cdot (a+c)} = \frac{ab+ad}{ab+bc}$$

22. (22.6.2004) (23.11.2004) (25.01.2005) Fall-Kontroll-Studie (S. 279)

a) Zeichnen Sie die Vierfeldertafel auf, welche Zahlen sind darin vorgegeben?

Anm: Bei der Fall-Kontroll-Studie empfiehlt es sich, die 4felder Tafel anders aufzuschreiben als bei der Kohortenstudie.

	Ausgangssituation		
	K	$\overline{K}$	$\sum$
R	a	c	a+c
$\overline{R}$	b	d	b+d
$\sum$	a+b	c+d	n

R bezeichnet das Exponiertsein gegenüber einem bestimmten Risikofaktor und K das Auftreten der zu untersuchenden Krankheit.

Demnach kann man die Zahlen folgendermaßen interpretieren:
a ist die Anzahl Personen, bei denen die Krankheit aufgetreten ist und die auch dem Risiko ausgesetzt waren.
b ist die Anzahl Personen, bei denen die Krankheit aufgetreten ist, die dem Risiko aber nicht ausgesetzt waren.
c ist die Anzahl Personen, bei denen die Krankheit nicht aufgetreten ist, die aber dem Risiko ausgesetzt waren.
d ist die Anzahl Personen, bei denen die Krankheit nicht aufgetreten ist, die aber auch nicht dem Risiko ausgesetzt waren.

b) Geben Sie die Formel für die Odds-Ratio an.
Die Odds-Ratio, das Maß für das Risiko (Chancenverhältnis), kann durch folgende Formel errechnet werden:

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

c) Wofür ist die Odds-Ratio ein guter Schätzer? Schreiben Sie dies verbal und symbolisch an.
Die Odds-Ratio ist ein guter Schätzer für das relative Risiko: $\hat{\omega} = \rho$???

d) Unter welcher Bedingung ist die Odds-Ratio ein guter Schätzer?
Die Odds-Ratio ist bei Krankheiten mit geringer Inzidenz ein guter Schätzer.

e) Leiten Sie die Odds-Ratio her.

$$\omega = \frac{P(R \mid K) / P(\overline{R} \mid K)}{P(R \mid \overline{K}) / P(\overline{R} \mid \overline{K})} = \frac{\frac{P(R \cap K)}{P(K)} \cdot \frac{P(K)}{P(\overline{R} \cap K)}}{\frac{P(R \cap \overline{K})}{P(\overline{K})} \cdot \frac{P(\overline{K})}{P(\overline{R} \cap \overline{K})}} = \frac{P(R \cap K) \cdot P(\overline{R} \cap \overline{K})}{P(R \cap \overline{K}) \cdot P(\overline{R} \cap K)} = \frac{ad}{bc}$$