Übersicht Statistik für Bioinformatiker

I) Wahrscheinlichkeitsrechnung

II) Statistik

II.a Schätzer

II.a.0 Allgemeines zu Schätzern
II.a.1 Schätzer für die Erfolgswahrscheinlichkeit
II.a.2 Schätzer für den Erwartungswert
II.a.2.1 Stichprobenmittel (empirisches Mittel) $\overline{X}$
II.a.2.2 gestutztes Mittel
II.a.2.3 Median
II.a.3 Schätzer für die Varianz
II.a.3.1 empirische Varianz $S(x)^2$

II.b Konfidenzintervalle

II.b.0 Allgemeines zu Konfidenzintervallen
II.b.1 Fall I: „kleine“ n (≤ 29)
II.b.2 Fall II: „große“ n

II.c Tests

II.c.0 Allgemeines zu Tests
II.c.1 Einstichprobenfall
II.c.1.1 Test auf den Erwartungswert μ_0; δ^2 bekannt
II.c.1.2 Test auf den Erwartungswert μ_0; δ^2 unbekannt
II.c.1.3 Test auf die Varianz δ_0^2
II.c.2 Zweistichprobenfall
II.c.2.1 Test zum Vergleich zweier Erwartungswerte μ_1 und μ_2
II.c.2.2 Test zum Vergleich zweier Varianzen δ_1^2 und δ_2^2
II.c.3 Weitere Tests
II.c.3.1 Test auf eine bestimmte Wahrscheinlichkeit p_0
II.c.3.2 Test auf die Häufigkeitwahrscheinlichkeiten von r verschiedenen Ausgängen
II.c.3.3 Test auf die Unabhängigkeit zweier Ereignisse

II.d Lineare Regression

II.d.0 Allgemeines zur Linearen Regression
II.d.1 Bestimmung der Regressionsgekoeffizienten $\hat{\beta}_0 + \hat{\beta}_1$

I.1 Kombinatorik

Gesucht: Zahl der Möglichkeiten für eine bestimmte Ziehung oder Verteilung

Wahrscheinlichkeit: $IP = \frac{\#\text{günstige Fälle}}{\#\text{alle Fälle}}$

(1) Urnenmodelle

Modell: Ziehe k Kugeln aus einer Urne, die n unterscheidbare Kugeln enthält

U1: mit Zurücklegen und Beachtung der Reihenfolge

$\Rightarrow n^k$ Möglichkeiten

- bei jeder Ziehung wird aus der gleichen Menge gezogen
- ➢ Nachrichten der Länge *k* über einem Alphabet mit *n* Zeichen
- ➢ **Würfeln** mit *k* Würfeln ($n = 6$)

U2: ohne Zurücklegen, mit Beachtung der Reihenfolge

$\Rightarrow n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$ Möglichkeiten

- 1. Ziehung aus *n* Kugeln, 2. Ziehung aus *n-1*, k-te Ziehung aus *n-k+1* Kugeln
- Kugeln werden aufgeteilt in *k* gezogene und *n-k* nicht gezogene – nur bei den gezogenen werden verschiedene Reihenfolgen unterschieden

U3: ohne Zurücklegen, ohne Beachtung der Reihenfolge

$\Rightarrow \frac{n!}{(n-k)! \cdot k!} = \binom{n}{k}$ Möglichkeiten

- häufig: Kugeln werden erst gezogen, dann geordnet, so daß die ursprüngliche Reihenfolge verloren geht
- ➢ **Lotto (** $n = 49, k = 6$)
- ➢ alle k-elementigen Teilmengen einer n-elementigen Menge

U4: mit Zurücklegen, ohne Beachtung der Reihenfolge

$\Rightarrow \binom{n+k-1}{k}$ Möglichkeiten

- jedesmal wird aus der gleichen Menge gezogen
- kommt am seltensten vor
- Bsp.: Möglichkeiten für *k* nicht unterscheidbare Spatzen, sich auf *n* Telegraphendrähte zu verteilen

(2) als Allokationsmodelle

Modell: Verteile k Murmeln auf n verschiedene Schachteln (Schubladen), bestimmte Schachteln bleiben leer.

U1: mit Mehrfachbesetzung, Murmeln sind unterscheidbar

- pro Schachtel ist mehr als eine Murmel erlaubt
- man kann entscheiden, *welche* Murmel in welcher Schachtel ist

U2: ohne Mehrfachbesetzung, Murmeln sind unterscheidbar

- pro Schachtel ist max. eine Murmel erlaubt

U3: ohne Mehrfachbesetzung, Murmeln sind nicht unterscheidbar

- man kann nur entscheiden, *wieviele* Murmeln in welcher Schachtel sind

U4: mit Mehrfachbesetzung, Murmeln sind nicht unterscheidbar

Übersicht (blau: Urnenmodelle, gelb: Allokationsmodelle)

Ziehe k Kugeln aus einer Urne mit n Kugeln	mit Zurücklegen	ohne Zurücklegen	
in Reihenfolge	n^k	$\frac{n!}{(n-k)!}$	Murmeln unterscheidbar (z.B. numeriert)
ohne Reihenfolge	$\binom{n+k-1}{k}$	$\binom{n}{k}$	Murmeln nicht unterscheidbar
	mit Mehrfachbelegung	ohne Mehrfachbelegung	**Verteile k Murmeln in n Schachteln**

I.2 Bedingte Wahrscheinlichkeiten und Unabhängigkeit

I.2.1 Definition bedingte Wahrscheinlichkeit

$IP(A \mid B) = \frac{IP(A \cap B)}{IP(B)}$ (W-keit, daß *A* eintritt, unter der Voraussetzung, daß *B* eingetreten ist)

I.2.2 Formel von der totalen Wahrscheinlichkeit

Sei $\Omega = \bigcup_{i=1}^{\infty} B_i$ eine Zerlegung in paarweise disjunkte Ereignisse. (Ersetze ∞ durch *n*, falls Ω nur in endlich viele B_i zerlegt wird.) Dann gilt: $IP(A) = \sum_{i=1}^{\infty} IP(A \mid B_i) \cdot IP(B_i)$.

I.2.3 Bayes'sche Formel

Sei $\Omega = \bigcup_{i=1}^{\infty} B_i$ eine Zerlegung in paarweise disjunkte Ereignisse. (Ersetze ∞ durch *n*, falls Ω nur in endlich viele B_i zerlegt wird.) Dann gilt: $IP(B_i \mid A) = \frac{IP(A \mid B_i) \cdot IP(B_i)}{IP(A)}$.

I.2.4 Unabhängigkeit

Zwei Ereignisse *A* und *B* heißen unabhängig $\Leftrightarrow IP(A \cap B) = IP(A) \cdot IP(B)$.

n Ereignisse $A_1, \ldots, A_n$ heißen unabhängig, wenn für jede Auswahl $A_{i1}, A_{i2}, \ldots, A_{ik}$ die Produktformel gilt:

$IP(A_{i_1} \cap \ldots \cap A_{i_k}) = IP(A_{i_1}) \cdot \ldots \cdot IP(A_{i_k}).$ $\quad (k \le n \text{ und } 1 \le i_1 < i_2 < \ldots < i_k \le n)$

I.3 Zufallsvariable

Eine Funktion $X : \Omega \to IR$ heißt Zufallsvariable. Zufallsvariablen bringen charakteristische Größen eines Zufallsexperiments zum Ausdruck (z.B. Gewinn/Verlust beim Glücksspiel).

– wichtige ZV sind Indikatorvariablen: $I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$ ($A \subset \Omega$ ist ein Ereignis), I_A ist Bernoulli-verteilt, d.h. $IP(I_A = 1) = p$, $IP(I_A = 0) = 1\text{-}p$, $IP(I_A = x) = 0$ für alle $x \notin \{0,1\}$. $IE(I_A) = IP(A)$ (Erwartungswert der Bernoulli-Verteilung)

I.3.1 Verteilung von Zufallsvariablen

Verteilungsfunktion: $F_X(x) = IP(X \in (-\infty, x])$

Eigenschaften von F:

– $\forall x : F_X(x) \in [0,1]$

– F_X ist monoton wachsend

– $\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1$

Verteilung der Summe zweier unabhängiger ZV X, Y:

1) X, Y diskret verteilt mit $IP(X = x_i) = p_i$ und $IP(Y = y_j) = q_j$

$$\Rightarrow IP(X + Y = k) = \sum_{i=0}^{k} p_i q_{k-i}$$

2) X, Y absolut stetig verteilt mit Dichten f, g

$\Rightarrow X + Y$ hat die Dichte $h(t) = (f * g)(t) = \int_{-\infty}^{\infty} f(t-u) g(u)\, du$ ($*$ – „Faltung“)

I.3.2 Unabhängigkeit von Zufallsvariablen

$X_1, \ldots, X_n$ heißen unabhängig, wenn für alle Intervalle $I_1, \ldots, I_n$ gilt:

$\{X_1 \in I_1\}, \ldots, \{X_n \in I_n\}$ sind unabhängig.

Bei diskreter Verteilung genügt Betrachtung „einpunktiger“ Intervalle.

I.3.3 Erwartungswert einer Zufallsvariablen

– X diskret verteilt, Werte $x_1, x_2, \ldots$; W-keiten $p_1, p_2, \ldots$ ($p_i = IP(X = x_i)$): $IE(X) = \sum_{i=1}^{\infty} p_i x_i$

– X absolut stetig verteilt mit Dichte f: $IE(X) = \int_{-\infty}^{\infty} t\, f(t)\, dt$

– $IE(cX) = c \cdot IE(X)$

– $IE(X + Y) = IE(X) + IE(Y)$

– X, Y unabhängig $\Rightarrow IE(X \cdot Y) = IE(X) \cdot IE(Y)$

– X diskret verteilt mit Werten $x_1, x_2, \ldots \Rightarrow IE(h(X)) = \sum_{i=1}^{\infty} h(x_i) IP(X = x_i)$

– X absolut stetig verteilt mit Dichte $f \Rightarrow IE(h(X)) = \int_{-\infty}^{\infty} h(t)\, f(t)\, dt$

– **Poincaré’sche Formel:** $IP(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{\substack{\{i_1,\ldots,i_n\} \\ \subset \{1,\ldots,n\}}} IP(A_{i_1} \cap \ldots \cap A_{i_k})$

  - Bsp.: n Briefe werden zufällig auf die dazugehörigen n Umschläge verteilt ➔ Wahrscheinlichkeit, daß mind. ein Brief im richtigen Umschlag landet.

I.3.3 Varianz, Standardabweichung und Kovarianz

Varianz: $Var(X) = V(X) = IE\left(\left|X - IE(X)\right|^2\right)$

praktischer: $V(X) = IE(X^2) - \left(IE(X)\right)^2$

Standardabweichung (Streuung) $\sigma_X = \sqrt{V(X)}$

Eigenschaften der Varianz:

- $V(aX+b) = a^2 V(X)$
- $X_1, \ldots, X_n$ unabhängig $\Rightarrow V(X_1 + \ldots + X_n) = V(X_1) + \ldots + V(X_n)$

Kovarianz: $Cov(X,Y) = IE\left(\left(X - IE(X)\right)\cdot\left(Y - IE(Y)\right)\right)$ (X, Y ZV mit endlicher Varianz)

- ***praktischer***: $Cov(X,Y) = IE(XY) - IE(X)\cdot IE(Y)$
- $Cov(X,Y) = 0 \Rightarrow$ X, Y „unkorreliert“

Eigenschaften der Kovarianz:

- $Cov(X,Y) = Cov(Y,X)$
- $Cov(a_1X_1 + a_2X_2, Y) = a_1Cov(X_1,Y) + a_2Cov(X_2,Y)$
- $\left|Cov(X,Y)\right| \le \sigma_X\sigma_Y$

Korrelationskoeffizient: $\rho_{X,Y} = \dfrac{Cov(X,Y)}{\sigma_X\sigma_Y}$

Je näher $\rho_{X,Y}$ bei ±1 liegt, umso besser liegen die Punktepaare $\left(X(\omega), Y(\omega)\right)$ auf einer Geraden.

I.4 diskrete Verteilungen

I.4.1 Hypergeometrische Verteilung

- Urne mit N Kugeln, davon R rote und N-R schwarze
- ziehen n-mal ohne Zurücklegen
- das Ergebnis „genau r rote und n-r schwarze Kugeln gezogen“ hat die W-keit

$$IP = h(r; n, N, R) = \frac{\binom{R}{r}\cdot\binom{N-R}{n-r}}{\binom{N}{n}}$$

- ZV X *h(•; n, N, R)*-verteilt $\Rightarrow IE(X) = n\dfrac{R}{N}$
- Bsp.1 **Skat**: N = 32 Karten, davon z.B. R = 4 Asse, n = 10 Karten kriegt jeder – wie wahrscheinlich ist es, daß ich genau r = 3 Asse auf der Hand habe?
- Bsp.2 **Lotto**: N = 49 Kugeln, davon sind R = 6 von mir getippt, n = 6 werden gezogen – wie wahrscheinlich ist es, daß genau r = 4 getippte gezogen werden?
- Bsp.3 capture-recapture-method: N **Fische im Teich** (N unbekannt), davon R markierte, n werden gefangen, darunter r markierte. Dann gilt: $N \approx n\cdot\frac{R}{r}$.

I.4.2 Binomialverteilung

- Experiment mit zwei möglichen Ausgängen (1 = Erfolg, 0 = Mißerfolg, p = W-keit für Erfolg) wird n-mal (voneinander unabhängig) nacheinander ausgeführt
- W-keit für genau k Erfolge und n-k Mißerfolge: $IP = b(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$
- ZV X *b(•; n, p)*-verteilt $\Rightarrow IE(X) = np,\ V(X) = np(1-p)$
- Gesamtexperiment läßt sich als Binärbaum der Höhe n darstellen (z.B. gehe nach links bei Erfolg, nach rechts bei Mißerfolg)
- Bsp.: n-maliges Ziehen mit Zurücklegen aus einer Urne mit N Kugeln, davon R rote $\left(\Rightarrow p = \dfrac{R}{N}\right)$ – mit welcher W-keit sind unter den gezogenen Kugeln genau k rote?
- Bsp.: n-maliger Münzwurf (p = 0,5) – mit welcher W-keit genau k-mal Kopf?

I.4.3 Poisson-Verteilung

- Approximation für die Binomialverteilung für kleines *p* und großes *n* (tritt bei **seltenen Ereignissen** auf)
- Bsp.: Rosinen pro Brötchen; Druckfehler pro Seite; gleichzeitig geführte Telefonate innerhalb einer Firma
- $IP(X = j) = \frac{\lambda^j}{j!} e^{-\lambda},\ \lambda \in IN$
- ZV *X* Poisson-verteilt $\Rightarrow IE(X) = \lambda$

I.5 absolut stetige Verteilungen

I.5.0 Definition

Eine absolut stetige Wahrscheinlichkeitsverteilung auf *IR* ist gegeben durch

$$IP([a,b]) = \int_a^b f(x)\,dx \text{ mit } f(x) \geq 0 \text{ für alle } x \text{ und } \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

f – Wahrscheinlichkeitsdichte

Verteilungsfunktion $F(x) = \int_{-\infty}^{x} f(t)\,dt \Rightarrow IP([a,b]) = F(b) - F(a)$

I.5.1 Gleichverteilung

- z.B. auf einem Intervall $[a,b] \subseteq IR$
- jeder Punkt aus $\Omega = [a,b]$ wird mit der gleichen W-keit getroffen
- W-keit, in einem Teilintervall $[c,d] \subseteq \Omega$ zu landen: $IP([c,d]) = \frac{d-c}{b-a}$
- $f(x) = \begin{cases} 1 & x \in [a,b] \\ 0 & \text{sonst} \end{cases}$
- weitere typische Ωs: ganz *IR* oder IR^2 Flächen (Rechtecke, Kreise (z.B. Dart)) in IR^2

I.5.2 Normalverteilung

- tritt vor allem bei Zufallsvariablen auf, die von vielen verschiedenen Einflüssen geprägt sind
- Approximation für die Binomialverteilung für große *n* (ab ca. 30), *p* darf nicht zu klein sein
- Standardnormalverteilung *N*(0,1): $F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt$
- Normalverteilung $N(\mu, \sigma^2)$ zu den Parametern μ und σ: $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt$
- ZV *X* $N(\mu, \sigma^2)$-verteilt $\Rightarrow IE(X) = \mu,\ Var(X) = \sigma^2$
- ZV *X* $N(\mu, \sigma^2)$-verteilt $\Rightarrow X^* = \frac{X - \mu}{\sigma}$ ist *N*(0,1)-verteilt

I.5.3 Exponentialverteilung

- tritt bei Ereignissen auf, die über einen bestimmten Zeitraum stets die gleiche Wahrscheinlichkeit, einzutreten, haben
- Verteilung der Wartezeit auf das erste Eintreffen von Ereignissen wie Vulkanausbrüche, Flugzeugabstürze, radioaktiver Zerfall, Polizeikontrollen, …
- „gedächtnislos“, d.h. $IP(X > x_1 + x_2 \mid X > x_2) = IP(X > x_1)$
- $f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$

I.6 Grenzwertsätze

$X_1, \ldots, X_n$ unabhängig mit derselben Verteilung wie X; $S_n = X_1 + \ldots + X_n$

Schwaches Gesetz der großen Zahl

Für jedes $\varepsilon > 0$ gilt: $IP\left(\left|\frac{S_n}{n} - IE(X)\right| \geq \varepsilon\right) \xrightarrow{n\to\infty} 0$.

- Klartext: das arithm. Mittel der Versuche konvergiert gegen den Erwartungswert

mit Tschebyschew genauer: $IP\left(\left|\frac{S_n}{n} - IE(X)\right| \geq \varepsilon\right) \leq \frac{V(X)}{n\varepsilon^2}$

Tschebyschew-Ungleichung

Version I: Y Zufallsvariable, $\varepsilon > 0 : IP\left(|Y| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} IE(Y^2)$

Version II: Z Zufallsvariable, $\varepsilon > 0 : IP\left(|Z - \mu| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} V(Z)$

- dient dem Abschätzen von W-keiten
- häufig nutzlos, weil Abschätzung > 1

Starkes Gesetz der großen Zahl

$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i(\omega) \xrightarrow{n\to\infty} IE(X)$ „fast sicher“

- die Menge der ω, wo das nicht stimmt, hat W-keit 0

Satz von de Moivre-Laplace

X Bernoulli-ZV (Indikatorvariable) mit Erfolgsw-keit p, $X_1, \ldots, X_n$ unabh. Wiederholungen

$\Rightarrow S_n = X_1 + \ldots + X_n$ binomialverteilt:

Dann ist S_n „ungefähr normalverteilt“ mit $\mu = np$ und $\sigma^2 = np(1-p)$

$\Rightarrow S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}}$ ist ungefähr $N(0,1)$-verteilt, und es gilt:

$IP(a \leq S_n^* \leq b) \xrightarrow{n\to\infty} \Phi(a) - \Phi(b)$

Achtung: $\{a \leq S_n \leq b\} = \left\{\frac{a - np}{\sqrt{np(1-p}} \leq S_n^* \leq \frac{b - np}{\sqrt{np(1-p}}\right\}$

bessere Approximation durch

±½-Korrektur: Verwende $IP(a - 0{,}5 \leq S_n \leq b + 0{,}5)$ statt $IP(a \leq S_n \leq b)$.

Zentraler Grenzwertsatz

Die Normalverteilung approximiert nicht nur die Binomialverteilung sondern **jede** Verteilung unabhängiger identisch verteilter X_i.

II) Statistik

a) Schätzer

0. Allgemeines zu Schätzern

Ein Schätzer für einen Parameter υ einer Verteilung ist eine Abbildung T: $\mathbb{R}^n \rightarrow \mathbb{R}$, so dass für unabhängige identisch verteilte $X_1,...,X_n$ der aus der Beobachtung $(x_1,...x_n) = (X_1(w),...X_n(w))$ berechnete Wert $T(x_1,...,x_n)$ als geschätzter Wert für den Parameter υ gelten kann.

Er heißt **erwartungstreu**, wenn $|E_\upsilon(T(X1,...Xn)) = \upsilon$ (in Worten etwa: „wenn der Erwartungswert über alle unabhängigen Werte mit dem wirklichen (zu schätzenden) Parameter übereinstimmt.“)

Er heißt **konsistent**, wenn $|P_\upsilon(|T(X_1,...X_n) - \upsilon| > \varepsilon) \xrightarrow{(n\rightarrow\infty)} 0$ (in Worten etwa: „wenn die der Betrag der Differenz zwischen dem wirklichen Parameter und seinem Schätzer für sehr große n quasi 0 wird“)

1. Schätzer für die Erfolgswahrscheinlichkeit

k = Anzahl der Erfolge n = Anzahl der Versuche

- **$p^\wedge$(Erfolg) = Schätzer für die Erfolgswkt. =** $\frac{k}{n}$

→ $p^\wedge$ ist ein „maximum-likelyhood- Schätzer“
→ erwartungstreu
→ konsistent

2. Schätzer für den Erwartungswert ($|E_\mu X = \mu$)

1. Stichprobenmittel (empirisches Mittel) $\overline{X}$

- $\overline{X} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} X_k$

→ **erwartungstreu**
→ **konsistent**

2. gestutztes Mittel (für nach Größe geordnete Werte)

→ arithmetisches Mittel der Zahlen bis auf die kleineste und die größte

- **gestutztes Mittel** $\frac{x_2 + \ldots + x_{n-1}}{n-2} = \frac{1}{n-2}\sum_{k=2}^{n-1} X_k$

→ „robuster“ als das Stichprobenmittel, da evtl. besonders weit vom wirklichen E-wert entfernte Werte wegfallen

3. Median (für nach Größe geordnete Werte)

→ mittlere Zahl

- **Median =** X_{m+1}**, wenn n = 2m+1** (ungerade Anzahl an Versuchen)
  = $\frac{1}{2}(X_m + X_{m+1})$**, wenn n = 2m** (gerade Anzahl an Versuchen)

3. Schätzer für die Varianz

1.empirische Varianz S(x)

- **S(x)²** = $\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2$

→ $S(x)=\sqrt{S(x)^2}$

→ n-1 wegen der Erwartungstreue

b) Konfidenzintervalle

0. Allgemeines zu Konfidenzintervallen

Da unser Schätzer $\hat{p}$ sogut wie nie mit der wirklichen Wkt. p übereinstimmen wird, wollen wir lieber einen Bereich $\hat{B} \subset \mathbb{R}$ berechnen, in dem man p vermuten darf. Solch eine Menge nennt man Konfidenzintervall und wir wollen solche Intervalle [g_{unten}(x), g_{oben}(x)] aus der Beobachtung berechnen. Mit dem „Niveau“ 1-α ∈[0,1] legen wir fest, „mit welcher Wkt. Der wahre Wert p im Intervall liegen soll“ (für Niveau = 100% ⇒ K-Intervall = [0,1]!). Typische Wahlen in Anwendungen sind α = 0.05 und α = 0.01.

1. Fall I: „kleine“ n (≤29)

→ binomial(b(,n,p))-verteilte ZV

Vorgehensweise: **1. α = 5% oder α = 1% ?**
2. n = Stichprobengröße = ?
3. z = Anzahl der „Erfolge“ = ?
4. Eintrag in der Tabelle (Tafel 8 & 8a) suchen
g_u = pz und g_o = $\overline{pz}$

2. Fall II: „große“ n

→ Approximation durch N(0,1)-verteilte ZV

$$\mathbf{p_u} = \frac{1}{n+4}\left(k+2-2\sqrt{\frac{k(n-k)}{n}+1}\right)$$

$$\mathbf{p_o} = \frac{1}{n+4}\left(k+2+2\sqrt{\frac{k(n-k)}{n}+1}\right)$$

c) Tests

0. Allgemeines zu Tests

Man macht einen Test zur Überprüfung eines Parameters auf seine Richtigkeit. Dazu stellt man eine Hypothese H_0 gegen eine Alternative H_1 auf, die dann nach dem Test entweder verworfen oder halt nicht verworfen werden kann (z.B. dass der Erwartungswert des Gewichts einer Kuh ≥ 100g ist – diese Hypothese wird wohl nie verworfen werden). Zu testenden Parameter sind u.a. Erwartungswerte, Varianzen, Wahrscheinlichkeiten, etc. Ein Test besteht aus mehreren unabhängigen Durchführungen eines Zufallsexperiments. Danach wird eine Test- oder Prüfgröße T gebildet, die mit einem (zumeist aus Tabellen abgelesenen) Wert verglichen wird. Führt

die Testgröße zu einem Wert, der „zu extrem“ und damit „zu unwahrscheinlich“ ist, verwirft man die Hypothese. Was „zu unwahrscheinlich“ heißt, muß natürlich vor der Durchführung feststehen. Gebräuchliche Werte für das Niveau des Tests sind u.a. 5% und 1%. Bewiesen oder wiederlegt ist nach einem Test nichts; selbst, wenn (fast) alles gegen die Hypothese spricht, kann sie richtig sein! Man kann daher nach Abschluß des Tests 2 Fehler machen:

H_0 ist richtig, wird aber verworfen ⇒ Fehler 1.Art (gravierender)

H_0 ist falsch, wird aber beibehalten ⇒ Fehler 2.Art (nicht so gravierend)

Es wird zusätzlich in einseitige Tests (entweder ein zu kleiner oder ein zu großer Wert von T führt zur Ablehnung) und zweiseitige Tests (zu kleine und zu große T führen zur Ablehnung) unterschieden.

1. EINSTICHPROBENFALL ($X_i \sim N(\mu,\delta^2)$-verteilt, n Versuchsdurchläufe)

1) Test auf den Erwartungswert μ_0; δ^2 bekannt ⇒ Gauß-Test

Vorgehensweise: **1. Errechnen von $\overline{X}$**

2. Errechen von T = $\frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$

3. Ablesen von N_α oder von $N_{1-\alpha}$ oder von $N_{\alpha/2}$ und $N_{1-\alpha/2}$

Auswertung:

- **H_0: $\mu = \mu_0$**

  falls $T \geq N_{1-\alpha/2}$ oder $T \leq N_{\alpha/2}$

  ⇒ H_0 ablehnen, ansonsten beibehalten

  Ablehnungs- bzw. Annahmebereich (zum Niveau 95%)

  $\left[\mu_0 - 1{,}96 * \sigma\sqrt{n}; \mu_0 + 1{,}96 * \sigma\sqrt{n}\right]$

  Liegt unser X außerhalb dieses Bereichs, wird H_0 abgelehnt

- **H_0: $\mu \geq \mu_0$**

  falls $T \leq N_\alpha$

  ⇒ H_0 ablehnen, ansonsten beibehalten

- **H_0: $\mu \leq \mu_0$**

  falls $T \geq N_{1-\alpha}$

  ⇒ H_0 ablehnen, ansonsten beibehalten

2) Test auf den Erwartungswert μ_0; δ^2 unbekannt ⇒ t-Test

Vorgehensweise: **1. Errechnen von $\overline{X}$**

2. Errechen von S(x)

3. Errechen von T = $\frac{X - \mu}{S(n)}\sqrt{n}$

⇒ t-Verteilung mit n-1 Freiheitsgraden

4. Ablesen von $t_{n-1;1-\alpha/2}$ oder von $t_{n-1;\alpha}$ oder von $t_{n-1;1-\alpha}$

Auswertung:
- **H_0: $\mu = \mu_0$**
  falls $|T| \geq t_{n-1;1-\alpha/2}$
  ⇒ H_0 ablehnen, ansonsten beibehalten

- **H_0: $\mu \geq \mu_0$**
  falls $T \leq t_{n-1;\alpha}$
  ⇒ H_0 ablehnen, ansonsten beibehalten

- **H_0: $\mu \leq \mu_0$**
  falls $T \geq t_{n-1;1-\alpha}$
  ⇒ H_0 ablehnen, ansonsten beibehalten

3) Test auf die Varianz δ_0^2 ⇒ χ^2-Test

Vorgehensweise: **1. Errechnen von $\overline{X}$**
2. Errechen von $S(x)^2$
3. Errechen von $T = (n-1)\frac{S(x)^2}{\sigma_0^2}$
⇒ χ^2-Verteilung mit n-1 Freiheitsgraden
4. Ablesen von $\chi^2_{n-1;\alpha/2}$ und $\chi^2_{n-1;1-\alpha/2}$ oder von $\chi^2_{n-1;\alpha}$ oder von $\chi^2_{n-1;1-\alpha}$

Auswertung:
- **H_0: $\delta^2 = \delta_0^2$**
  falls $T \leq \chi^2_{n-1;\alpha/2}$ oder $T \geq \chi^2_{n-1;1-\alpha/2}$
  ⇒ H_0 ablehnen, ansonsten beibehalten

- **H_0: $\delta^2 \geq \delta_0^2$**
  falls $T \leq \chi^2_{n-1;\alpha}$
  ⇒ H_0 ablehnen, ansonsten beibehalten

- **H_0: $\delta^2 \leq \delta_0^2$**
  falls $T \geq \chi^2_{n-1;1-\alpha}$
  ⇒ H_0 ablehnen, ansonsten beibehalten

2. ZWEISTICHPROBENFALL ($X_i \sim N(\mu_x, \delta_x^2)$-verteilt, n Versuchsdurchläufe und $Y_i \sim N(\mu_y, \delta_y^2)$-verteilt, m Versuchsdurchläufe)

1) Test zum Vergleich zweier Erwartungswerte μ_1 und μ_2

Vorgehensweise: **1. Errechnen von $\overline{X}$ und $\overline{Y}$**
2. Errechen von r = n + m - 2
3. Errechen von

$$T = \sqrt{\frac{nm*(n+m-2)}{n+m}} * \frac{\overline{X} - \overline{Y}}{\sqrt{(n-1)*S(x)^2 + (m-1)*S(y)^2}}$$

⇒ t-Verteilung mit r Freiheitsgraden
4. Ablesen von $t_{r;1-\alpha/2}$ oder von $t_{r;1-\alpha}$ oder von $t_{r;\alpha}$

Auswertung: • **H_0: $\mu_1 = \mu_2$**
falls $|T| \geq t_{r;1-\alpha/2}$
⇒ H_0 ablehnen, ansonsten beibehalten

• **H_0: $\mu_1 \geq \mu_2$**
falls $T < t_{r;\alpha}$
⇒ H_0 ablehnen, ansonsten beibehalten

• **H_0: $\mu_1 \leq \mu_2$**
falls $T > t_{r;1-\alpha}$
⇒ H_0 ablehnen, ansonsten beibehalten

2) Test zum Vergleich zweier Varianzen δ_1^2 und δ_2^2 ⇒ F-Test

Vorgehensweise: **1. Errechnen von $\overline{X}$ und $\overline{Y}$**
2. Errechen von $S(x)^2$ und $S(y)^2$
3. Errechen von $T = \frac{S(x)^2}{S(y)^2}$, wo $S(x)^2 \geq S(y)^2$
⇒ F-Verteilung mit (n-1,m-1) Freiheitsgraden
4. Ablesen von $F_{n-1;\, m-1;\, 1-\alpha/2}$

Auswertung: • **H_0: $\delta_1^2 = \delta_2^2$**
falls $T \geq F_{n-1;\, m-1;\, 1-\alpha/2}$
⇒ H_0 ablehnen, ansonsten beibehalten

3. WEITERE TESTS

1) Test auf eine bestimmte Wahrscheinlichkeit p_0 ⇒ Binomial-Test

1. Beidseitiger Test
Vorgehensweise: **Bestimmung des maximale k_u und minimalen k_o**

k_u: $\sum_{k=0}^{k_u-1} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \frac{\alpha}{2}$ **k_o:** $\sum_{k=k_o+1}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \frac{\alpha}{2}$

Auswertung: • **H_0: $p = p_0$**
falls $x < k_u$ oder $x > k_o$
⇒ H_0 ablehnen, ansonsten beibehalten

2. Einseitiger Test
Vorgehensweise: **Bestimmung des maximale k_u und minimalen k_o**

k_u: $\sum_{k=0}^{k_u-1} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \alpha$ **bzw. k_o:** $\sum_{k=k_o+1}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \alpha$

Auswertung: • **H_0: $p \geq p_0$**
falls $x < k_u$
⇒ H_0 ablehnen, ansonsten beibehalten

• **H_0: $p \leq p_0$**
falls $x > k_o$
⇒ H_0 ablehnen, ansonsten beibehalten

2) Test auf Häufigkeitwahrscheinlichkeiten von r verschiedenen Ausgängen
⇒ χ^2-Häufigkeitstest
→ Multinomialverteilung

$i \in [1,r]$ = möglicher Ausgang des Experiments (mit r verschiedenen Ausgängen**)**
h_i = beobachtete Häufigkeit des Ausgangs i
$p_i^{(0)}$ = geschätzte Wahrscheinlichkeit für den Ausgang i

Vorgehensweise: **1. Werte H_i und $p_i^{(0)}$ notieren**
2. Errechnen von T = $\sum_{i=1}^{r} \frac{\left(H_i - p_i^{(0)} n\right)^2}{p_i^{(0)} n}$
3. Ablesen von $\chi^2_{r-1;\,1-\alpha}$
Auswertung: • **H_0: $p_i = p_i^{(0)}$ (die geschätzten Einzelwkt. stimmen)**
falls $T \geq \chi^2_{r-1;\,1-\alpha}$
⇒ H_0 ablehnen, ansonsten beibehalten

3) Test auf die Unabhängigkeit zweier Ereignisse A und B
⇒ χ^2-Unabhängigkeitstest

a,b,c,d = beobachtete Häufigkeit der Ausgänge $(A \cap B)$,$(A^c \cap B)$,$(A \cap B^c)$ und $(A^c \cap B^c)$

Vorgehensweise: **1. Vierfeldertafel aufstellen und ausfüllen (nach folgendem Schema)**

	A	A^c	
B	**a**	**b**	**a+b**
B^c	**c**	**d**	**c + d**
	a+c	**b+d**	**n = a+b+c+d**

2. Errechnen von T = $\frac{n*(ad-bc)^2}{(a+b)*(c+d)*(a+c)*(b+d)}$
3. Ablesen von $\chi^2_{1;\,1-\alpha}$
Auswertung: • **H_0: A und B sind unabhängig**
falls $T \geq \chi^2_{1;\,1-\alpha}$
⇒ H_0 ablehnen, ansonsten beibehalten

d) Lineare Regression

0. Allgemeines zur Linearen Regression

Den (z.B. bei einem Versuch) erhobenen Werte haftet immer ein Meßfehler zufälliger Größe an. Daher wird sich keine eindeutige Gerade (hier werden nur lineare Beziehungen betrachtet) ergeben, wenn man die Punkte im Graphen einfach verbinden würde. Um nun die Gerade zu erhalten, die möglichst gut durch die erhaltenen Punktwolke führt, bestimmen wir die (empirische) Regressionsgerade „$Y = \hat{\beta}_0 + \hat{\beta}_1 X$ “ und dazu die (empirische) Regressionsgekoeffizienten $\hat{\beta}_0 + \hat{\beta}_1$.

1. Bestimmung der Regressionsgekoeffizienten $\hat{\beta}_0 + \hat{\beta}_1$

→ **X_k und Y_k** sind zwei zueinandergehörige, gemessene Werte (und identifizieren im Graphen zusammen einen Punkt)

$\Rightarrow P_1(X_1,X_2)$, $P_2(X_2,Y_2)$, etc.

Vorgehensweise: **1. Errechnen von $\overline{X}$ und $\overline{Y}$**

2. Errechnen von $\overline{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$

3. Errechnen von $\overline{XX} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

Auswertung: $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$ $\qquad \hat{\beta}_1 = \frac{\overline{XY} - \overline{X} * \overline{Y}}{\overline{XX} - \overline{X} * \overline{Y}}$