Theorie

1.) Was ist eine Zufallsvariable?

Eine Zufallsvariable ist eine Variable, die ihre Werte in Abhängigkeit vom Zufall, d.h.: mit einer gewissen Wahrscheinlichkeit annimmt. (Wir ziehen aus der Grundgesamtheit eine Stichprobe daraus). Die Wahrscheinlichkeiten und damit die Zufallsvariable können oft durch eine Verteilung eindeutig charakterisiert werden. D.h.: Unter der Verteilung einer Zufallsvariablen versteht man die Gesetzmäßigkeit, nach der diese Zufallsvariable ihre Werte annimmt.
D.h.: eine Zufallsvariable ist die Zuordnung von Ereignissen eines "Zufallsexperiments" zu Zahlen
Man unterscheidet diskrete und kontinuierlich stetige Zufallsvariablen.
Diskrete Zufallsvariablen haben höchstens abzählbar viele verschiedene Werte. Kontinuierlich stetige Zufallsvariablen können jeden beliebigen Wert in ihrem Definitionsbereich annehmen.
Der Mittelwert einer Zufallsvariable wird Erwartungswert μ genannt. Bei der Varianz σ^2 gehört der Erwartungswert zu den Parametern, die eine Zufallsvariable charakterisieren.
Besitzt eine stetige Zufallsvariable X den Erwartungswert μ und die Varianz σ^2, dann kann man durch die Standardisierung $z = \frac{(X-\mu)}{\sigma}$ eine Zufallsvariable z erzeugen, deren Erwartungswert $\mu = 0$ und deren Varianz $\sigma^2 = 1$ ist. Diese Eigenschaft nutzt man, um Normalverteilungen mit beliebigen Parametern μ und σ in Standardnormalverteilungen zu transformieren

2.) Wie ist Freiheitsgrad definiert?

Freiheitsgrade = Anzahl der frei wählbaren Parameter:

z.B.: Bei der Schätzung der Varianz σ^2 haben wir immer eine χ^2_{n-1} - Verteilung, d.h.: n-1 Freiheitsgrade

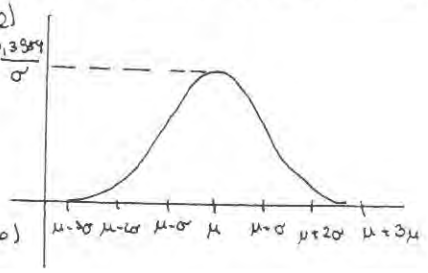
$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$ d.h.: ich kann (n-1) Parameter frei wählen, der n-te ist dann schon festgelegt.

3.) Wie sieht die Normal-, die t, die F-, die Chi-Quadrat Verteilung aus? (Skizze)

Normalverteilung $N(\mu, \sigma^2)$: Die Konzentration (Dichte) der Normalverteilung kann man sich als Glockenkurve vorstellen. Viele quantitative Größen konzentrieren sich oft um einen bestimmten Wert. Traditionsgemäß dient die Normalverteilung als Approximation eines solchen Verhaltens

μ = Ortsparameter (Mittel)
σ = Skalierungsparameter



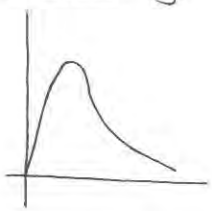
Bei $\mu \pm \sigma$ Wendepunkt

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\mu, x \in (-\infty, \infty)$
$\sigma > 0$

χ^2-Verteilung

Verteilung beginnt bei 0

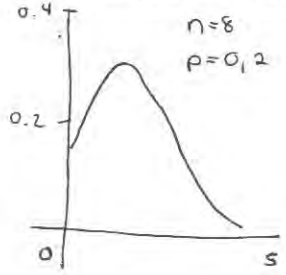


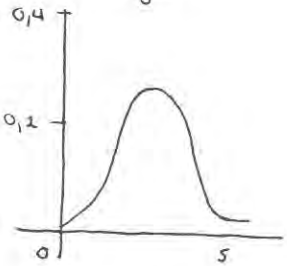
gilt nur für $n \geq 3$, je mehr n, desto symmetrischer wird die Verteilung

~~Hypothesis ein Beispiel~~

Binomialverteilung Bi(n;p)

Die Binomialverteilung dient als Modell, wenn bei einem Versuch bzw. einer Beobachtung zwei Möglichkeiten gegeben sind und diese mit den Wahrscheinlichkeiten p bzw (1-p) auftreten und n unabhängige Versuche mit der gleichen Einzelwahrscheinlichkeit p vorliegen. Die Binomialverteilung hat den Erwartungswert n·p und die Varianz np(1-p).
Für große Werte von n lässt sich die Binomialverteilung durch die Normalverteilung mit dem Mittelwert $\mu = n \cdot p$ und der Varianz $\sigma^2 = np(1-p)$ annähern. Dabei muss beachtet werden, dass die Binomialverteilung für Werte von p nahe 0 oder 1 sehr schief ist. Für p nahe 0.5 ist die Näherung allerdings auch für kleinere Werte von n recht gut





Poissonverteilung: $\sum_{0}^{\infty} p_i = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda}$

wird auch Verteilung der seltenen Ereignisse genannt. Für großes n und kleines p lässt sich die Binomialverteilung gut durch die Poissonverteilung annähern ($\lambda = n \cdot p$) (Faustregel $p < 0.1$, $n > 50$)

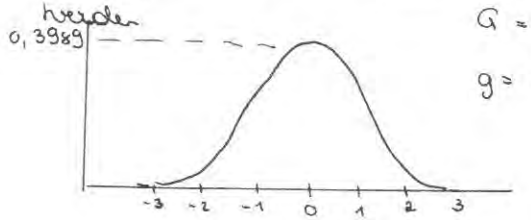
4. Was versteht man unter einem statistischen Test?

Ein statistischer Test liefert nach bestimmten Regeln eine Entscheidung darüber, ob eine vorgegebene Hypothese über die zu untersuchende Grundgesamtheit anhand von Daten aus einer Stichprobe verworfen werden muss oder nicht verworfen werden kann. In der Statistik verstehen wir unter Hypothese eine Annahme über die Verteilung einer Zufallsvariablen. Man formuliert eine Ausgangshypothese H_0 (=Nullhypothese) und stellt ihr als Gegenhypothese der Alternativhypothese H_1 gegenüber. Dann gibt man ein Signifikanzniveau α vor und fordert, dass die Wahrscheinlichkeit des Verwerfens der Nullhypothese obwohl sie richtig ist, nicht größer als α ist. Aufgrund einer Prüfgröße (Teststatistik) wird dann die Nullhypothese beibehalten oder zugunsten der Alternativhypothese verworfen.

5. Welche Arten von Tests gibt es?

Parametrische und nichtparametrische Tests
Parametrisch heißen alle statistischen Tests, die an die Voraussetzung einer bestimmten Verteilung mit entsprechenden Parametern gebunden sind.

Eine Normalverteilung mit Erwartungswert $\mu = 0$ und der Varianz $\sigma^2 = 1$ heißt Standardnormalverteilung. Durch eine Standardisierung kann jede beliebige Normalverteilung in eine Standardnormalverteilung transformiert werden.



G = Verteilungsfkt. der Standardnormalverteilung

g = Dichtefkt. —"—

$$g(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

Im Intervall $[-1,1]$ liegen 68,26% der Werte

—"— $[-2,2]$ —"— 95,45% —"—

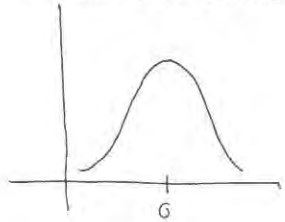
—"— $[-3,3]$ —"— 99,73% —"—

t-Verteilung

Die t-Verteilung ist eine Verteilung, die man aus einer Transformation von n unabhängigen normalverteilten Zufallsvariablen ableiten kann. Die t-Verteilung ist symmetrisch zu 0. Der Erwartungswert ist 0 und die Varianz ist $(n-1)(n-3)$. Mit wachsendem n nähert sich die Dichtefunktion der t-Verteilung immer mehr der Dichtefunktion der Standardnormalverteilung. Die Teststatistik eines t-Tests ist t-verteilt.

Die t-Verteilung verläuft umso flacher, je geringer der Stichprobenumfang bzw. die Anzahl der Freiheitsgrade m ist. Sie tritt daher zur Schätzung des Erwartungswertes m bei unbekannter Varianz s^2 an die Stelle der Normalverteilung.

Je größer allerdings der Stichprobenumfang ist, umso eher kann die t-Verteilung durch die einfacher zu handhabende Normalverteilung ersetzt werden.



schaut aus wie Normalverteilung, ist aber flacher

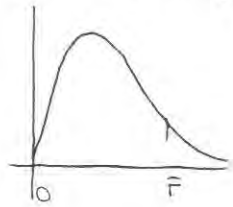
F-Verteilung:

Die F-Verteilung ist ein theoretisches Verteilungsmodell für eine positive, kontinuierliche Zufallsvariable.

Wenn 2 Varianzen unabhängiger zufälliger Stichproben der Umfänge n_1 und n_2 aus zwei normalverteilten Grundgesamtheiten mit gleicher Varianz sind, dann folgt die zufällige Variable

$$F = \frac{s_1^2}{s_2^2} \quad (s_1^2 > s_2^2)$$

einer F-Verteilung mit den Freiheitsgraden f_1 und f_2 als Parameter. Der Wert muss immer größer als 1 sein. Es handelt sich bei der F-Verteilung um eine stetige unsymmetrische Verteilung. Ihre Form hängt von den beiden Freiheitsgraden ab.



Die meisten parametrischen Tests sind unter Annahme der Normalverteilung entwickelt worden (z.B. t-Test und Varianzanalyse)
Nichtparametrisch heißen alle statistischen Tests, die nicht an die Voraussetzung einer bestimmter Verteilung mit entsprechenden Parametern gebunden sind

6.) Was ist ein Fehler 1. Art, was ein Fehler 2. Art?

Das Signifikanzniveau α gibt die Wahrscheinlichkeit an, mit der die Hypothese verworfen wird, obwohl sie richtig ist. Diese Schlussfolgerung ist natürlich ein Fehler, der als Fehlerwahrscheinlichkeit oder Fehler 1. Art bezeichnet wird.
Ist eine Hypothese falsch wird sie trotzdem nicht verworfen, so nennt man das Fehler 2. Art. Das Auftreten eines Fehlers 2. Art wird mit β bezeichnet

Entscheidung	H_0 richtig	H_0 falsch
annehmen	$1-\alpha$	β
ablehnen	α	$1-\beta$
	1	1

Die Wahrscheinlichkeit, eine richtige Alternativhypothese im statistischen Test auch tatsächlich richtig zu erkennen, ist $(1-\beta) \Rightarrow$ Diese Wahrscheinlichkeit wird auch als Macht (Schärfe) des Tests.

7.) Was ist ein Konfidenzintervall?

Mittels einer "Zufallsstichprobe" kann man Aussagen über eine unbekannte Grundgesamtheit machen. Den Wertebereich, der den interessierenden Parameter mit Wahrscheinlichkeit $1-\alpha$ überdeckt, nennt man Konfidenzintervall
D.h.: Ein Konfidenzintervall ist ein geschätztes Intervall, welches den wahren Wert eines unbekannten Parameters (z.B.: Erwartungswert) mit vorgegebener Wahrscheinlichkeit $1-\alpha$ (= Überdeckungswahrscheinlichkeit) überdeckt
Konfidenzintervalle für

μ $P\left(\bar{x} - z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}\right) = 1-\alpha$

μ $P\left(\bar{x} - t_{n-1,1-\frac{\alpha}{2}}\, s/\sqrt{n} \leq \mu \leq \bar{x} + t_{n-1,1-\frac{\alpha}{2}}\, s/\sqrt{n}\right) = 1-\alpha$

falls σ durch s aus Stichprobe geschätzt werden muss

σ^2 $P\left((n-1)s^2/\chi^2_{n-1,1-\frac{\alpha}{2}} \leq \sigma^2 \leq (n-1)s^2/\chi^2_{n-1,\frac{\alpha}{2}}\right) = 1-\alpha$

Wie groß muss die Stichprobe bei gegebener maximaler Länge d des Konfidenzintervalls sein?

$$n = \left(2 z_{1-\frac{\alpha}{2}}\sigma/d\right)^2$$

8.) Wozu wird eine Varianzanalyse durchgeführt? Welche Voraussetzungen sind nötig?

Varianzanalysen sind parametrische Mehrstichprobentests. Mehrstichprobentests sind statistische Tests für mehr als 2 Stichproben.
Bsp.: Die monatliche Gerichtsannahme einer Anzahl von Tieren variiert von Tier zu Tier, auch wenn alle Be-

dingungen wie Futteral - und menge gleich groß ist. Diese Variation ist rein zufälliger Art. Werden die Tiere allerdings unterschiedlich gefüttert, so kommt die Variation bezüglich des Futters dazu => Bei Varianzanalyse versucht, diese beiden Variationen zu trennen. Untersucht man 2 Einflüsse gleichzeitig, so spricht man von doppelter Varianzanalyse.

Voraussetzungen: einfache Varianzanalyse

n unabhängige Stichprobenwerte

$$x_{ij}, i = 1, \dots, n_j \; ; \; j = 1, \dots, k$$

von normalverteilten Zufallsvariablen mit gleicher Varianz, d.h.: $X_{ij} \sim N(\mu_j, \sigma^2)$; $n = \sum_{j=1}^{k} n_j$.

Wir wollen auf Gleichheit aller Mittelwerte testen

H_0 : $\mu_1 = \mu_2 = \dots = \mu_k$

H_1 : $\mu_r \neq \mu_s$ für mind. ein $r \neq s$, $r,s = 1, \dots, k$

Betrachtet wird die Quadratsumme der Abweichungen:

$$q = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$ und sie wird in 2 Quadratsummen zerlegt:

$$q = q_I + q_2$$

$q_I = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$ = quadr. Abweichungen innerhalb jeder Stichprobe

$q_2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2$ = Quadratsumme zwischen den Stichproben

Nun kann man zeigen, dass unter der Null-Hypothese, d.h.: alle x_{ij} sind Realisationen der Verteilung $N(\mu_0, \sigma^2)$ mit $\mu_1 = \mu_2 = \dots = \mu_k = \mu_0$, die Verteilungen der entsprechenden Zufallsvariablen von q_I/σ^2 und q_2/σ^2 unabhängig und gleich χ^2_{n-k} bzw. χ^2_{k-1} sind.

Das Verhältnis $$F = \frac{q_2/(k-1)}{q_I/(n-k)}$$ genügt einer $F_{k-1,n-k}$ Verteilung.

Kritischer Bereich: $F > F_{k-1,n-k;\,1-\alpha}$

Voraussetzungen: doppelte Varianzanalyse

Wenn man Daten nach 2 Gesichtspunkten einteilt und sie nach diesen analysieren will, kann man die doppelte Varianzanalyse anwenden.

Geg: Stichprobe von n Werten, die sich in k Gruppen und jede Gruppe in genau p Klassen einteilen lässt.

	p Spalten
k Zeilen	$x_{11}\ x_{12} \dots x_{1p}$
	$x_{k1} \dots \dots x_{kp}$

x_{ij} sind unabhängige Realisationen von normalverteilten Zufallsvariablen mit gleicher Varianz

H_0: $\bar{\mu}_{\cdot 1} = \bar{\mu}_{\cdot 2} = \dots = \bar{\mu}_{\cdot p}$

$\bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \dots = \bar{\mu}_{k\cdot}$.

H_1: mind. ein $\neq$

Die „totale" Quadratsumme $q = \sum_{j=1}^{k} \sum_{j=1}^{p} (x_{ij} - \bar{x})^2$ wird in 3 Teile aufgespalten: $q = q_2 + q_S + q_R$

$q_2 = p \sum_{i=1}^{k} (\bar{x}_{i\cdot} - \bar{x})^2$ = Quadratsumme der Mittelwerte zwischen den Zeilen

$q_S = k \sum_{j=1}^{p} (\bar{x}_{\cdot j} - \bar{x})^2$ = — " — Spalten

$q_R = \sum_{j} \sum_{j} (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2$ = quadr. Restsumme

Ist die Nullhypothese richtig, so kann man zeigen, dass die den Quadratsummen q_z, q_R, q_s entsprechenden Zufallsvariablen Q_z, Q_s, Q_R voneinander unabhängig sind und Q_z/σ^2, Q_s/σ^2 und Q_R/σ^2 χ^2-verteilt sind mit $k-1$, $p-1$ bzw. $(k-1)(p-1)$ Freiheitsgraden. Daraus ergibt sich, dass die Verhältnisse der mittleren Quadratsummen S^2_z/S^2_R und S^2_s/S^2_R mit

$S^2_z = Q_z/(k-1)$, $S^2_R = Q_R/[(k-1)(p-1)]$ und $S^2_s = Q_s/(p-1)$

F-Verteilungen mit $[(k-1), (k-1)(p-1)]$ bzw. $[(p-1), (k-1)(p-1)]$ Freiheitsgraden besitzen.

kritischer Bereich:

$$F = S^2_z/S^2_R > F_{k-1,(k-1)(p-1);1-\alpha}$$
$$F = S^2_s/S^2_R > F_{p-1,(k-1)(p-1);1-\alpha}$$

9. Wie funktioniert der Chi-Quadrattest? Wozu braucht ich ihn?

Der Chi-Quadrat-Test dient zum Überprüfen einer Hypothese über die Form der Verteilung (z.B.: Test auf Gleichverteilung, usw.)

Dabei wird eine Klasseneinteilung aller Werte getroffen und die empirischen (gemessenen) Häufigkeiten werden mit den theoretischen (hypothetischen) verglichen. Weichen Sie stark voneinander ab, wird man die Hypothese verwerfen, sonst annehmen.

k Klassen, h_i = obs. Häufigkeit = Anzahl von Datenpunkten in der i-ten Klasse

p_i = theoretische Wahrscheinlichkeit, dass ein Wert in die i-te Klasse fällt,

$e_i = n \cdot p_i$ = darunter die Hypothese erwartete absolute Häufigkeit, n = # der Daten

$$T = \sum_{i=1}^{k} \frac{(h_i - e_i)^2}{e_i}$$

die Verteilung von T strebt gegen eine χ^2_{k-1} Verteilung

kritischer Bereich $T > \chi^2_{k-1;1-\alpha}$

Die Verteilung von T ist nur asymptotisch bekannt, stimmt also nur für grosse n.

Enthält die hypothetische Verteilung noch unbekannte Parameter, die mit den gleichen Daten geschätzt werden müssen, so wirkt sich das auf die Verteilung von T aus. Werden r Parameter geschätzt, so besitzt T die asymptotische Verteilung χ^2_{k-r-1}.

10.) Wie kann man feststellen, ob eine Stichprobe normal verteilt ist?

- mittels Chi-Quadrat-Test (siehe Frage 9) } Anpassungstests
- mittels Kolmogorov-Smirnov-Test

Hier muss angenommen werden, dass die Stichprobenvariablen $X_1, X_2, \dots, X_n$ eine stetige Verteilungsfunktion F haben. Um die zugrunde liegende Verteilung F auf eine hypothetische F_0 zu testen

d.h.: $H_0: F(x) = F_0(x)$ $\forall x$ ist es naheliegend, die absolute Differenz $|F_n(x) - F_0(x)|$ bezüglich der empirischen Verteilungsfunktion F_n zu betrachten.

- grafisch mittels Wahrscheinlichkeitsnetz

11.) Wie kann ich feststellen, ob μ bzw. σ^2 zweier Verteilungen gleich sind?

In vielen Forschungsstudien liegt das Hauptinteresse im Vergleich zwei Gruppen statt im Vergleich einer Gruppe mit irgendwelchen bekannten Werten.

Vergleich der Mittel:

Beim Vergleich zwei Gruppen von Beobachtungen vergleicht man i. d. R. ihre Mittelwerte und untersucht sie auf signifikante Unterschiede.

Voraussetzung: Beide Populationen, von denen die Beobachtungen kommen, sind normalverteilt und weisen gleiche Varianz auf.

Dabei können 2 wesentliche Fälle auftreten

→ Jeder Wert der einen Stichprobe hängt mit einem Wert der anderen Stichprobe zusammen => Bildung der Differenz und Testen des Mittelwerts auf 0.

→ Stichproben sind voneinander unabhängig (und eventuell nicht gleich groß)

=> Anwenden des 2-Stichproben-t-Test

Stichprobe 1: $x_1, \dots, x_n$

Stichprobe 2: $y_1, \dots, y_n$

Mittel μ_x, μ_y bzw. Schätzer $\bar{x}, \bar{y}$ sowie s^2_x, s^2_y

Als Teststatistik wird für den Test $\mu_x = \mu_y$

$$T = \sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}} \cdot \frac{\bar{x}-\bar{y}}{\sqrt{(n_1-1)s^2_x + (n_2-1)s^2_y}}$$

verwendet. T ist t-verteilt mit n_1+n_2-2 Freiheitsgraden

kritischer Bereich beim einseitigen Test

$$T > t_{n_1+n_2-2;\,1-\alpha}$$

Vergleich der Varianzen

Manchmal ist es interessant zu wissen, ob die Varianzen zweier Normalverteilungen, deren Mittel nicht bekannt sein müssen, also gleich angesehen werden können.

Das Verhältnis der empirischen Varianzen zweier unabhängiger Normalverteilungen mit vorausgesetzter gleicher Varianz ist im wesentlichen F-verteilt. Wenn v. X mit der Verteilung $N(\mu_x, \sigma^2)$ n_1 Stichprobenwerte und von Y mit der von X unabhängigen Verteilung $N(\mu_y, \sigma^2)$ n_2 Stichprobenwerte zur Verfügung stehen; so ist

$$F = \frac{s^2_x}{s^2_y} \qquad F_{n_1-1, n_2-1} \text{ verteilt}$$

$H_0: \sigma^2_x = \sigma^2_y$

$H_1: \sigma^2_x > \sigma^2_y$

kritischer Bereich: $F > F_{n_1-1, n_2-1;\,1-\alpha}$

12. Was versteht man unter einem Regressionsproblem? Wie wird die Regressionsgerade angesetzt? Wo ist darin die Varianz enthalten? Wie berechnet man die Geradengleichung?

Ein Regressionsproblem behandelt die Verteilung einer Variablen, wenn mind. eine andere gewisse Werte in nicht zufälliger Art annimmt.

Bsp: z.B. Verteilung des Gewichts von Männer mit ihrer Größe.

Für jede gewählte Größe x schwanken wir eine gewisse Verteilung des Gewichts y der Männer mit dieser

Größe. Weil die Verteilung von Y von den Werten von x abhängt, wird Y auch als abhängige, x als unabhängige Variable bezeichnet.
Oft kann die Abhängigkeit der Mittelwerte von $Y (\mu_{Y.x})$ von x im Bereich der x-Werte durch eine gerade Linie angegeben werden => einfache, lineare Regression

$$\mu_{Y.x} = \underset{\sim}{a} + \underset{\sim}{b}(x - \bar{x})$$

a = Ordinatenabschnitt, b = Steigung der Regressionsgeraden

a, b = feste Parameter

„Methode der kleinsten Quadrate": Die Regressionsgerade soll so durch die Punktwolke gelegt werden, dass die Summe der Quadrate der Abweichungen möglichst klein, also minimiert wird.

Schätzung der Parameter

Die Parameter a und b müssen aus den Daten geschätzt werden

$$\hat{a} = \bar{Y}$$

$$\hat{b} = \frac{s_{xy}}{s^2_x}$$

wobei $s^2_x = \frac{1}{n-1} \sum (x_i - \bar{x})^2$ die empirische Varianz der x-Werte berechnet

$s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$ die empirische Kovarianz zwischen x und Y

$\hat{y}_x$ = geschätzter mittlerer Wert von Y an der Stelle x:
dann gilt: $\hat{y}_x = \hat{a} + \hat{b}(x - \bar{x})$

Eine erwartungstreue Schätzung für $\sigma^2 = \sigma^2_{y.x}$ ist

$$s^2 = \frac{n-1}{n-2} (s^2_y - \hat{b}^2 \cdot s^2_x)$$

s - Standardfehler der Beobachtungen

„Methode der kleinsten Quadrate": Die Regressionsgerade wird so gewählt, dass die Summe der quadrierten Residuen (= Abweichung zw. gemessenen und geschätzten Werten) minimal wird.

Test auf Abhängigkeit:

Eine häufig aufgestellte Hypothese ist die der Abhängigkeit der Variablen Y von x. Eine Methode, diese zu testen ist auf Gleichheit der Mittelwerte von Y bei allen Werten von x zu testen.

$H_0: b = 0$ (wird sie verworfen, so gibt dies genügend Grund zur Annahme, dass Y von x abhängt

$H_1: b \neq 0$

$$T = \frac{(\hat{b} - 0)\, s_x \sqrt{n-1}}{s}$$

Wenn die Verteilung von Y normal mit gleichen Mittel und Varianz für jedes x ist, so besitzt T eine t-Verteilung mit n-2 Freiheitsgraden

Kritischer Bereich: $|T| > t_{n-2,\,1-\frac{\alpha}{2}}$

13. Was ist ein Korrelationsproblem?

Ein Korrelationsproblem betrachtet die gemeinsame Verteilung von 2 Variablen, von denen keine durch den Experimentator fixiert wird

D.h.: es wird der Zusammenhang zwischen 2 zufällige Größen betrachtet.

In einer Stichprobe müssen immer paarweise Messungen vorliegen. Das Paar der betrachteten Zufallsvariablen (x, y)

sollen eine bivariate Normalverteilung aufweisen. D.h.: Bei einem fixen Wert von X besitzt Y eine Normalverteilung und umgekehrt.

Das Maß der Abhängigkeit zwischen X und Y zur Charakterisierung dieser bivariaten Verteilung dient die Kovarianz. $\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)]$

Die Korrelation zwischen X und Y ist definiert als:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \qquad \text{(Wertebereich zw. -1 und 1)}$$

Unabhängigkeit der beiden Variablen bedeutet $\sigma_{XY} = 0$ und damit $\rho_{XY} = 0$

Schätzung für $\rho = r_{XY} = \frac{1}{s_X \cdot s_Y} \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$

Test auf Unkorreliertheit

$H_0: \rho = 0$

$H_1: \rho \neq 0$

Sind die beiden Variablen X und Y voneinander unabhängig & normalverteilt, so besitzt die Statistik

$$T = R\sqrt{\frac{n-2}{1-R^2}}$$ eine t_{n-2}-Verteilung, wobei R die Zufallsvariable bezeichnet, die die Werte des empirischen Korrelationskoeffizienten r_{XY} annimmt. kritischer Bereich: $|T| > t_{n-2;\,1-\frac{\alpha}{2}}$

14.) Was versteht man unter Wahrscheinlichkeit?

- Das, das jedem Ereignis A aus $\mathcal{A}$ eine nicht-negative, reelle Zahl $\mu(A)$ zuordnet und das bestimmte Eigenschaften aufweist.

Ein Maß μ ist eine Funktion vom Ereignisraum in $[0, \infty]$, wobei, wenn $\{A_1, A_2, \ldots\}$ eine Zerlegung von A darstellt, gilt (σ-Additivität): $\mu(A) = \sum_{i=1}^{\infty} \mu(A_i)$. Gilt außerdem $\mu(\Omega) = 1$, dann spricht man von einer Wahrscheinlichkeitsmaß, das wir mit P bezeichnen.

15.) Allgemeiner Additionssatz?

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

16.) Was bedeutet „Bedingte Wahrscheinlichkeit"? Was versteht man in diesem Zusammenhang unter Unabhängigkeit?

Manchmal ist es leichter, die Wahrscheinlichkeit des Eintretens eines Ereignisses zu definieren, wenn man weiß, dass ein anderes bereits eingetreten ist.

Allgemein definiert man für 2 Ereignisse A und H aus $\mathcal{A}$ mit $P(H) > 0$ die bedingte Wahrscheinlichkeit von A unter H als:

$$P(A|H) = \frac{P(A \cap H)}{P(H)}$$ $P(\cdot|H)$ ist wieder ein Wahrscheinlichkeitsmaß

Multiplikationsregel für Wahrscheinlichkeiten

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

Wenn für 2 Ereignisse A und B $(P(B) > 0)$ gilt

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

also, dass das bedingte Ereignis die gleiche Wahrscheinlichkeit aufweist wie das nichtbedingte, dann sind A und B unabhängig.

Bei A und B unabhängig gilt:

$$P(A \cap B) = P(A) \cdot P(B).$$

17.) Definition der Binomialverteilung:

Binomialverteilung $Bi(n,p)$: Es werden n unabhängige Versuche durchgeführt, von denen jeder Versuchsausgang mit Wahrscheinlichkeit p gut und $(1-p)$ schlecht ausgeht. Nachdem die n Versuche unabhängig voneinander durchgeführt werden, ist die Wahrscheinlichkeit, dass i Versuche gut und die restlichen $(n-i)$ Versuche schlecht ausgehen:

$$\left(\prod_{j=1}^{i} P(x_j=1)\right)\left(\prod_{j=i+1}^{n} P(x_j=0)\right) = p^i (1-p)^{n-i}$$

Es gibt aber $\binom{n}{i}$ Möglichkeiten der Reihenfolge der Versuchsausgänge $\Rightarrow$

$$P(E_i) = \binom{n}{i} \cdot p^i \cdot (1-p)^{(n-i)}$$

18.) Formale Definition für Zufallsvariable

Eine Zufallsvariable X ist eine Abbildung von $(\Omega, \mathcal{A})$ in $(\mathbb{R}, \mathcal{L})$, wobei für jedes $B \in \mathcal{L}$ gilt $X^{-1}(B) = \{\omega \mid X(\omega) \in B\} = A \in \mathcal{A}$

19.) Eigenschaften der Verteilungsfunktion?

a) F ist monoton wachsend

b) F ist rechtsseitig stetig

c) $\lim_{x \to \infty} F(x) = F(\infty) = 1$

$\lim_{x \to -\infty} F(x) = F(-\infty) = 0$

20.) Poissonverteilung?

Poissonverteilung $P(\lambda)$: Die möglichen Werte der Zufallsvariablen X sind $x_i = 0, 1, 2, \ldots$ Die Wahrscheinlichkeitsfunktion ist durch

$$p_i = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0,1,2,\ldots$$ für ein gewisses $\lambda > 0$ definiert

$$\sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda} = e^{-\lambda} \cdot e^{\lambda} = 1$$

Die Poissonverteilung wird auch Verteilung der seltenen Ereignisse genannt. Für grosses n und kleines p lässt sich die Binomialverteilung gut durch die Poissonverteilung annähern

21.) Was ist das Wahrscheinlichkeitsnetz?

Manchmal ist es günstiger, die Verteilungsfunktion statt der Dichte zu betrachten. Wenn $x_1, \ldots, x_n$ n Datenpunkte bezeichnen, so heißt die Funktion

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(x_i)$$

empirische Verteilungsfunktion, wobei I die Indikatorfunktion ist. Diese Treppenfunktion F_n gibt die relative Summenhäufigkeit an, weist n Sprünge der Größe $\frac{1}{n}$ auf und hat alle Eigenschaften der Verteilungsfunktion. Für den Zweck der Prüfung auf Normalverteilung ist es von Vorteil, die Skalierung der Ordinate zu verändern. Im Wahrscheinlichkeitsnetz wird die Ordinate so, dass 0 und 1 nicht in gleich große Intervalle geteilt sondern die Abstände werden proportional zu G^{-1} aufgetragen. (G = Verteilungsfkt d. Standardnormalverteilung) Somit wird eigentlich $G^{-1}(F_n(x))$ über x dargestellt. Wenn nun die Daten ungefähr normalverteilt sind, so wird die Treppenfunktion etwa auf einer Geraden zu liegen kommen.

22.) Erwartung und Varianz

h = reelle FKt der Zufallsvariablen X, dann ist der Mittelwert oder die math. Erwartung von h(X) im Falle einer stetigen Zufallsvariable

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)\, f(x)\, dx$$

im diskreten Fall

$$E[h(X)] = \sum_{i=1}^{\infty} h(x_i)\, p_i$$

Ssl $h(x) = x \Rightarrow \mu = E(X) = \int x\, f(x)\, dx$

bzw $\Rightarrow \mu = E(X) = \sum x_i p_i$

$$\sigma^2 = VAR(X) = E(X-\mu)^2 = E(X - E(X))^2$$

$$VAR(X) = E(X)^2 - (EX)^2$$

Additionssatz für Mittelwerte

Der Mittelwert einer Summe von Zufallsvariablen, deren Mittelwerte existieren, ist gleich der Summe dieser Mittelwerte

$$E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$$

Multiplikationssatz für Mittelwerte

Für n unabhängige Zufallsvariable $X_1, \dots, X_n$, deren Mittelwerte existieren, gilt

$$E(X_1 X_2 \dots X_n) = (EX_1)(EX_2) \dots (EX_n).$$

Varianz der Summe zweier Zufallsvariablen $Z = X + Y$

$\sigma_x^2 = VAR(X)$, $\sigma_y^2 = VAR(Y)$, $\sigma_z^2 = VAR(Z)$

$$\sigma_z^2 = E(Z - EZ)^2 = E(X + Y - E(X+Y))^2 =$$

$$E(X - EX)^2 + E(Y - EY)^2 + 2E[(X - EX)(Y - EY)]$$

$$= \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$$

$$\sigma_{xy} = E[(X - EX)(Y - EY)] = E(XY) - (EX)(EY)$$

= Kovarianz der Zufallsvariablen X und Y

Für den Fall der Unabhängigkeit von X und Y gilt: $E(XY) = (EX)(EY)$, sodass folgt $\sigma_{xy} = 0 \Rightarrow$ die Kovarianz verschwindet. Der umgekehrte Schluss ist nicht zulässig.

Additionssatz für Varianzen:

Die Varianz einer Summe unabhängiger Zufallsvariablen, deren Varianzen existieren, ist gleich der Summe der Varianzen: $\sigma_z^2 = \sigma_y^2 + \sigma_x^2$

23.) Was bedeutet Kovarianz? Was Korrelation?

Die Kovarianz σ_{xy} von zwei Zufallsvariablen X und Y stellt ein Maß für die Abhängigkeit der Beiden dar.

Eine Standardisierung dieses Maßes erhält man, indem man es durch die Streuung von X und Y dividiert

$$\varrho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Diese Größe kann nur Werte zwischen -1 und 1 annehmen.

Schätzung für die Kovarianz σ_{xy}

$$s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Korrelation ϱ_{xy}

$$R_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

24.) Zentraler Grenzwertsatz - Aussage

Bei steigendem Stichprobenumfang (aus Gleichverteilung) ist:

→ das empirische Mittel von $\bar{x}$ für verschiedene n ungefähr gleich (bis auf kl. zufälliger Fehler)

→ die Varianzen $s^2_{\bar{x}}$ von $\bar{x}$ werden bei steigendem Stichprobenumfang kleiner

→ Multipliziert man die Varianz s^2_x allerdings mit n, so erhält man ungefähr den selben Wert

SATZ: Besitzt die Verteilung der Grundgesamtheit eine endliche Varianz, so ist die Verteilung der arithmetischen Mittel von Zufallsstichproben approximativ normal, sofern der Stichprobenumfang genügend groß ist.

25.) Was ist das Ziel der analytischen Statistik?

Die analytische Statistik soll eine Verbindung zwischen der Theorie & der Wirklichkeit herstellen D.h.: sie versucht Schlüsse von einer Stichprobe auf die Grundgesamtheit gezogen werden.

26.) Was sind Stichproben?

Untermenge einer Population

Nach der Einführung von Zufallsvariablen X kann ein Stichprobenwert x_i auch als Realisation, als konkret angenommener Wert von X, aufgefasst werden.

Die verschiedenen Stichprobenwerte $(x_1, \dots, x_n)$ sind dann wiederholte Realisationen der Zufallsvariablen X, die normalerweise als unabhängig voneinander betrachtet

27.) Punktschätzungen?

Angenommen: Die Verteilung der Stichprobenelemente X_i enthält einen unbekannten Parameter Θ und es gibt eine Fkt t, die aus den Stichprobenwerten den Wert von Θ näherungsweise berechnen $\hat{\Theta} = t(x_1, \dots, x_n)$

t = Schätzfunktion oder Schätzer.

Eine Realisation des Schätzers heißt Schätzwert oder Schätzung.

Ein Schätzer ist eine Zufallsvariable, daher kann seine mathematische Erwartung untersucht werden. Wenn also t den Parameter Θ schätzt, so soll gelten $E(T) = \Theta$ => erwartungstreuer oder unverzerrter Schätzer => Das arithmetische Mittel $\bar{x}$ der Stichprobe ist ein erwartungstreuer Schätzer des Mittels der Verteilung oder des Populationsmittels.

$\bar{x}$ stellt auch einen konsistenten Schätzer dar.

Die Güte eines Schätzers hängt von seiner Variabilität ab => d.h.: je kleiner seine Varianz, desto besser. Man sagt, ein erwartungstreuer Schätzer ist wirksam oder effizient, wenn er die kleinstmögliche Varianz aufweist.

Es gibt verschiedene Verfahren, um brauchbare Schätzer für Parameter einer Verteilung zu finden

Die Maximum-Likelihood-Methode ist die richtige Sie wählt im Wesentlichen jenen Wert des Parameters der die Stichprobe als „wahrscheinlichstes" Resultat erscheinen lässt

28.) Schätzungen ~~eines~~ des Mittels einer Population

Tests bezüglich des Mittels μ der Population stützen sich auf das Stichprobenmittel $\bar{x}$ und dessen Verteilung.

Die Hypothese soll lauten $\mu = \mu_0$, wobei μ_0 ein speziell Wert ist. Ist sie richtig, so werden sich die Werte von $\bar{x}$ zufällig um μ_0 streuen. Die Wahrscheinlichkeit, dass

$\bar{x}$ in kritischen Bereich (fällt zum Verwerfen der Hypothese führt, wird Signifikanzzahl/niveau genannt.
Voraussetzung $X \sim N(\mu_0, \sigma^2) \Rightarrow \bar{x} \sim N(\mu_0, \sigma^2/n)$
und es ist besser mit der standardisierten Größe $z = (\bar{x} - \mu_0) / (\sigma/\sqrt{n})$ zu arbeiten, die $N(0,1)$ verteilt ist.
Wenn ein bestimmter Wert von z absolut größer ist als eine bestimmte Schranke, wird die Hypothese verworfen. $|z| > z_{1-\frac{\alpha}{2}}$
Einstichproben-t-Test: Im Falle einer unbekannten Varianz σ^2 wird σ^2 aus der Stichprobe geschätzt und z wird durch t_{n-1} ersetzt
Die Grenzen des Konfidenzintervalls und des kritischen Bereichs fallen zusammen $\Rightarrow$ Wenn also μ_0 in das Konfidenzintervall fällt, wird auch der Test die Hypothese $H_0: \mu = \mu_0$ nicht verwerfen und umgekehrt.

29. Hypothesentest für Varianz

Standardabweichung dient als Maß für die Variabilität einer Messgröße und ist ebenso wichtig wie das Mittel

$H_0: \sigma^2 = \sigma_0^2$

$\frac{n}{\sigma_0^2} \sum_{i=1} (x_i - \bar{x})^2$ χ^2-Verteilung mit n bzw. n-1 Freiheitsgraden, jenachdem, ob μ bekannt ist oder $\bar{x}$ aus Stichprobe geschätzt werden muss

$$T = \sum_{i=1}^{n} (x_i - \mu)^2$$

bzw $$T = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Wartung	Funktionsdauer (in Tagen)									
A	310	1073	1361	1843	965	312	680	231	968	1056
B	379	1238	1114	1278	2091	526	1021	378	758	1296

a) Sind die Funktionsdauern unabhängig? (Sicherheit $1 - \alpha = 95\ \%$) (3)

b) Führt der Wartungsplan B zu kürzeren Funktionszeiten? (Sicherheit $1 - \alpha = 95\ \%$) (2)

Mündliche Prüfung: Freitag, 28 05 1993, im Dienstzimmer von Herrn Prof. Dutter (bitte tragen Sie sich dazu in die im Sekretariat aufliegende Liste ein).