Theorie - Statistik und Wahrscheinlichkeitstheorie

27. Mai 2008

Inhaltsverzeichnis

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 heraus). Die Wahrscheinlichkeit 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 **kontinuierliche Zufallsvariablen**. Diskrete Zufallsvariablen haben höchstens abzählbar viele verschiedene Werte. Kontinuierlich stetige Zufallsvariablen können jeden beliebigen Wert in ihrem Definitionsbereich annehmen.

Der **Mittelwert** wird **Erwartungswert** μ genannt. Mit 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**

$$Z = \frac{(X - \mu)}{\sigma} \tag{0.1}$$

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 einer Schätzung der Varianz σ^2 haben wir immer eine χ^2_{n-1} - Verteilung, d.h. $n - 1$ Freiheitsgrade.

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

$$\overline{x} = \sum (\frac{x_i}{n})$$

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- und die Chi-Quadrat-Verteilung aus? (Skizze)

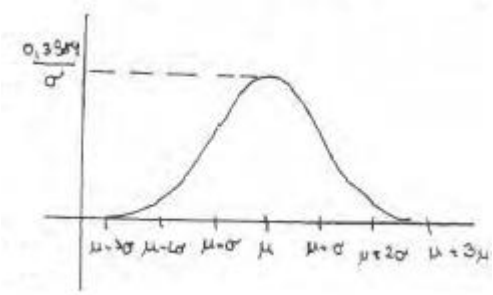
3.1. Normalverteilung N(μ,σ^2)

Die Konzentration (Dichte) der Normalverteilung kann man sich als Glockenkurve vorstellen. Viele quantitative Größen konzentrieren sich um einen bestimmten Wert. Traditionsgemäß dient die Normalverteilung als Approximation eines solchen Verhaltens.

- μ =Ortsparameter (Mittel)
- σ = Skalierungsparameter
- $\mu \pm \sigma$ = Wendepunkt
- $f(x) = \frac{1}{\sqrt{2\pi\sigma}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

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

- G = Verteilungsfunktion der Standardnormalverteilung
- g = Dichtefunktion der Standardnormalverteilung $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$



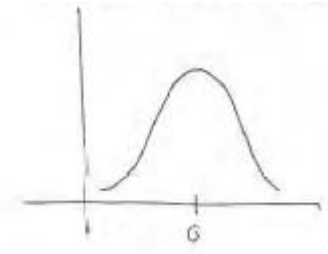
- Im Intervall $[-1, 1]$ liegen $68,25\%$ der Werte
- Im Intervall $[-2, 2]$ liegen $95,45\%$ der Werte
- Im Intervall $[-3, 3]$ liegen $99,73\%$ der Werte

3.2. 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 μ bei unbekannter Varianz σ^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.

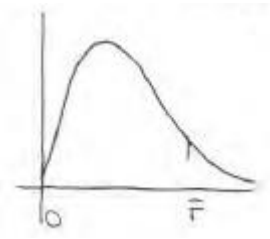


3.3. F - Verteilung

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

$$F = \frac{s_1^2}{s_2^2}(s_1^2 > s_2^2) \tag{0.2}$$

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.



3.4. χ^2 - Verteilung

Verteilung beginnt bei 0.

Gilt nur für $n \geqslant 3$, je mehr n, desto symetrischer wird die Verteilung

3.5. Binomialverteilung B(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 np und die Varianz $np(1-p)$.

Für große Werte von n lässt sich die Binomialverteilung durch die Normalverteilung mit dem Mittelwert $\mu = np$ 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 kleine Werte von n recht gut.

¡¡Graphik¿¿

3.6. Poissonverteilung

$$\sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda} \tag{0.3}$$

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

4. Was versteht man unter einem statistischen Test?

Ein statistischer Test liefert nach bestimmten Regeln einen Entscheidung darüber, ob eine vorgegebene Hypothese über die zu untersuchende Grundgeamtheit anhand von Daten aus einer Stichprobe verworfen werden muss oder nicht verworfen werden kann. In der Statistik vertehen wir unter Hypothese eine Annahme über die Verteilung einer Zufallvariablen.

Man formuliert eine Ausgangshypothese H_0 (=Nullhypothese) und stellt ihr als Gegenhypothese die 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 der Prüfgröße (Teststatistik) wird dann die Nullhypothese beibehalten oder zugunsten der Alternativhypothese verworfen.

5. Welche Arten von Tests gibt es?

Parametrische und nichtparametrische Tests:

Parametrische Tests heißen alle Tests, die an die Vorraussetzung einer bestimmten Verteilung mit entsprechenden Parametern gebunden sind. 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 Vorraussetzungen einer bestimmten 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 und wird sie trotzdem nicht verworfen, so nennt man das einen Fehler 2. Art. Das Auftreten eines Fehlers 2. Art wird mit β bezeichnet.

Entscheidung	**richtig**	**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 bezeichnet.

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.

Konfidenzintervall für

$$\mu : P(\overline{X} - z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \leqslant \mu \leqslant \overline{X} + z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}) = 1-\alpha \tag{0.4}$$

$$\mu : P(\overline{X} - t_{n-1,1-\frac{\alpha}{2}}S/\sqrt{n} \leqslant \mu \leqslant \overline{X} + t_{n-1,1-\frac{\alpha}{2}}S/\sqrt{n}) = 1-\alpha \tag{0.5}$$

falls σ durch S aus der Stichprobe geschätzt werden muss.

$$\sigma^2 : P((n-1)S^2/\chi^2_{n-1;1-\frac{\alpha}{2}} \leqslant \sigma^2 \leqslant (n-1)S^2/\chi^2_{n-1;\frac{\alpha}{2}}) = 1-\alpha \tag{0.6}$$

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

$$n = (2z_{1-\frac{\alpha}{2}}\sigma/d)^2 \tag{0.7}$$

8. Wozu wird eine Varianzanalyse durchgeführt? Vorraussetzungen?

8.1. Allgemeines

Varianzanalysen sind parametrische Mehrstichprobentests. Mehrstichprobentests sind statistische Tests für mehr als 2 Stichproben.

Bsp.: Die monatliche Gewichtszuname einer Anzahl von Tieren variiert von Tier zu Tier, auch wenn alle Bedingungen wie Futterart und -menge gleich groß sind. Diese Variation ist rein zufälliger Art. Werden die Tiere allerdings unterschiedlich gefüttert, so kommt die Variation bezüglich des Futters dazu $\rightarrow$ Die Varianzanalyse versucht, diese beiden Variationen zu erkennen.

Untersucht man 2 Einflüsse gleichzeitig, so spricht man von doppelter Varianzanalyse.

8.2. Voraussetzungen - einfache Varianzanalyse

n unabhängige Stichprobenwerte $x_{ij}; i = 1, \ldots, n; j = 1, \ldots, k$ von normalverteilten Zufallsvariablen mit gleicher Varianz, d.h.

$$X_{ij} \sim N(\mu_j, \sigma^2); n = \sum_{j=1}^{k} n_j \tag{0.8}$$

Wir wollen auf Gleichheit aller Mittelwerte testen

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
- $H_1 : \mu_r \neq \mu_s$ für mindestens ein $r \neq s$, r,$s = 1$,. . . ,k

Betrachtet wird die Quadratsumme der Abweichungen:

$$q = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \overline{x})^2 \tag{0.9}$$

und sie wird in 2 Quadratsummen zerlegt: $q = q_I + q_Z$

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

$q_Z = \sum_{j=1}^{k} n_j (\overline{x}_j - \overline{x})^2$ . . . Quadratsumme zwischen den Stichproben

Nun kann man zeigen, dass unter der Null-Hypothese, d.h. alle x_{ij} sind Realisationen unter der Verteilung $N(\mu_0, \sigma^2)$ mit $\mu_1 = \mu_2 = \cdots = \mu_k = \mu_0$ die Verteilungen der entsprechenden Zufallsvariablen von $\frac{q_I}{\sigma^2}$ und $\frac{q_Z}{\sigma^2}$ unabhängig und gleich χ^2_{n-4} bzw. χ^2_{n-1} sind. Das Verhältnis

$$F = \frac{q_I/(k-1)}{q_Z/n-k}$$

genügt einer $F_{k-1,n-k}$ Verteilung, kritischer Bereich: $F > F_{k-1,n-k,1-\alpha}$

8.3. Vorraussetzungen - doppelte Varianzanalyse

Wenn man Daten nach 2 Gesichtspunkten einteilt und 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. (k - Zeilen, p - Spalten)

$$\begin{Bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{k1} & x_{k2} & \ldots & x_{kp} \end{Bmatrix}$$

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

H_0: $-\ \mu_{01} = \mu_{02} = \cdots = \mu_{0p}$

$-\ \mu_{10} = \mu_{20} = \cdots = \mu_{k0}$

H_1: mindestens ein $\neq$

Die „totale„Quadratsumme

$$q = \sum_{i=1}^{k} \sum_{j=1}^{p} (x_{ij} - \bar{x})^2$$

wird in drei Teile aufgespalten:

$$q = q_Z + q_S + q_R$$

weitere Erklärung fehlt hier....

9. Wie funktioniert ein χ^2 - Test, wozu brauche ich ihn?

Der χ^2 - Test dient zum Überprüfen einer Hypothese über die Form einer Verteilung (z.B. Test auf Gleichverteilung, usw.). Dabei wird eine Klasseneinteilung aller Werte getroffen und die empirischen (gemessenen) Häufigkeiten werden mit den theoretischen (hypothetischen) Werten verglichen. Weichen sie stark voneinander ab, wird man die Hypothese verwerfen, sonst annehmen.

- k - Klassen
- h_i - absolute Häufigkeiten (Anzahl der Datenpunkte in der i-ten Klasse)
- φ_i - theoretische Wahrscheinlichkeit, dass ein Wert in die i-te Klasse fällt
- $e_i = n * p_i$ - die unter der Hypothese erwartete relative Häufigkeit
- n - Anzahl der Daten

$$T = \sum_{i=1}^{k} \frac{(h_i - e_i)^2}{e_i} \tag{0.10}$$

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 normalverteilt ist?

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

Hier muss angenommen werden, dass die Stichprobenvariablen $X_1, X_2, \ldots, X_n$ 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. (graphisch durch ein Wahrscheinlichkeitsnetz)

11. Wie kann ich feststellen ob μ und σ zweier Verteilungen gleich sind?

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

11.1. Vergleich der Mittel

Beim Vergleich 2er Gruppen von Beobachtungen vergleicht man im Allgemeinen ihre Mittelwerte und untersucht sie auf signifikante Unterschiede.

Voraussetzung: Beide Populationen, von denen die Beobachtungen kommen, sind normalverteilt und weisen die gleiche Varianz auf. Dabei können 2 wesentliche Fälle auftreten:

Jeder Wert der einen Stichprobe hängt mit einem Wert der anderen Stichprobe zusammen $\Rightarrow$ Bildung der Differenz und testen des Mittelwertes auf 0.

Stichproben sind voneinander unabhängig und eventuell nicht gleich groß $\Rightarrow$ Anwendung des 2-Stichproben-t-Tests

Stichprobe 1: $X_1, X_2, \ldots, X_n$

Stichprobe 2: $Y_1, Y_2, \ldots, Y_n$

Mittel μ_x , μ_y bzw. Schätzer $\overline{X}$, $\overline{Y}$ sowie σ_x^2 , σ_y^2. Als Teststatistik für den Test $\mu_x = \mu_y$

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n_1-1)s_x^2+(n_2-1)s_y^2}{n_1+n_2-2}(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{0.11}$$

verwendet. T ist t-verteilt mit $n_1 + n_2 - 2$ Freiheitsgraden, kritischer Bereich beim einseitigen Test $T > t_{n_1+n_2-2;1-\alpha}$

11.2. Vergleich der Varianzen

Manchmal ist es interessant zu wissen ob die Varianzen 2er Normalverteilungen, deren Mittel nicht bekannt sein müssen, als gleich angesehen werden können. Das Verhältnis der empirischen Varianzen zweier unabhängiger Normalverteilungen mit vorausgesetzer gleicher Varianz ist im wesentlichen F-verteilt.

Wenn von X mit der Verteilung $N(\mu_x, \sigma)$ n_1 Stichprobenwerte und von Y mit der von X unabhängigen Verteilung $N(\mu_y, \sigma)$ n_2 Stichprobenwerte zur Verfügung stehen, so ist

$$F = \frac{S_X^2}{S_Y^2} \quad F_{n_1-1,n_2-1} \tag{0.12}$$

verteilt, kritischer Bereich ist $F > F_{n_1-1,n_2-1,1-\alpha}$

12. Regressionsproblem

12.1. Allgemein

Ein Regressionsproblem behandelt die Verteilung einer Variable, wenn mindestens eine andere Werte in nicht zufälliger Art annimmt. Bsp.: Verteilung des Gewichts von Männern mit ihrer Größe. Für jede Größe X bekommen wir eine gewisse Verteilung der Gewichte Y der Männer mit dieser gewissen 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 gegeben werden $\Rightarrow$ einfache lineare Regression

$$\mu_{y.x} = \hat{a} + \hat{b}(x - \overline{x})$$

$\hat{a}$ - Ordinatenabschnitt

$\hat{b}$ - Steigung der Regressionsgeraden

$\hat{a}$, $\hat{b}$ - 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.

12.2. Schätzung der Parameter

Die Parameter a und b müssen aus den Daten geschätzt werden.

$$\hat{a} = \overline{y}$$

$$\hat{b} = \frac{S_{xy}}{S_x^2}$$

wobei

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

die empirische Varianz der x - Werte bezeichnet,

$$S_{xy} = \frac{1}{n-1} \sum (x_i - \overline{x})^2 (y_i - \overline{y})^2$$

die empirische Kovarianz zwischen x und y. $\hat{y}_x$ geschätzter Mittlerer Wert von Y an der Stelle x, dann gilt:

$$\hat{y}_x = \hat{a} + \hat{b}(x - \overline{x}) \tag{0.13}$$

Eine erwartungsgetreue Schätzung für $\sigma^2 = \sigma_{y.x}$ ist

$$S^2 = \frac{n-1}{n-2}(s_y^2 - \hat{b}^2 s_x^2)$$

S = Standardfehler der Beobachtungen

„Methode der kleinesten Quadrate„: Die Regressionsgerade wird so gewählt, dass die Summe der quadrierten Residuen (Abweichungen zwischen gemessenen und geschätzten Werten) minimal wird.

12.3. Test auf Abhängigkeit

Eine häufig aufgestellte Hypothese ist die der Abhängigkeit der Variable 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 es genügend Grund zur Annahme, dass Y von x abhängt.

$H_1 : b \neq 0$

$$T = \frac{(\hat{b} - 0) s_x \sqrt{n-1}}{S} \tag{0.14}$$

Wenn die Verteilung von Y normal mit gleichem 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. Korrelationsproblem

13.1. Allgemein

Ein Korrelationsproblem betrachtet die gemeinsame Verteilung von 2 Variablen, von denen keine durch den Experimentator fixiert wird. D.h. es wird der Zusammenhang zwischen zwei zufälligen Größen betrachtet. In einer Stichprobe müssen immer paarweise Messungen vorliegen. Das Paar der betrachteten Zufallsvariablen (X, Y) sollten eine

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

Als Maß der Abhängigkeit zwischen X und Y zur Charakterisierung der bivariaten Verteilung dient die Kovarianz

$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)]$$

Die Korrelation zwischen X und Y ist definiert als

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

wobei ihr Wert zwischen -1 und 1 liegt. Unabhängigkeit der beiden Variablen bedeutet $\sigma_{xy} = 0$ und somit $\rho_{xy} = 0$.

Schätzung für

$$\rho = r_{xy} = \frac{1}{s_x s_y} \frac{1}{n-1} \sum (x_i - \overline{x})(y_i - \overline{y})$$

13.2. Test auf Unkorreliertheit

$H_0 : g = 0$

$H_1 : g \neq 0$

Sind die beiden Variablen X und Y voneinander unabhängig und 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. Wahrscheinlichkeit

Maß, das jedem Ereignis A aus ϑ eine nicht - negative 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 einem Wahrscheinlichkeitsmaß, das wir mit P bezeichnen.

15. Allgemeiner Additionssatz

$$P(A \cup B) = P(A) + P(B) + P(A \cap B) \tag{0.15}$$

16. Bedingte Wahrscheinlichkeit - 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 zwei Ereignisse A und H aus ϑ mit $P(H) > 0$ die bedingte Wahrscheinlichkeit von A unter H als

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

$P(.|H)$ ist wieder ein Wahrscheinlichkeitsmaß.

Multiplikationsregel für Wahrscheinlichkeiten:

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

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

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

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

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

17. Definition der Binomialverteilung

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

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

Es gibt aber $\begin{pmatrix} n \\ i \end{pmatrix}$ Möglichkeiten der Reihenfolge der Versuchsausgänge $\Rightarrow$

$$P(E_i) = \begin{pmatrix} n \\ i \end{pmatrix} p^i(1-p)^{n-i}$$

18. Formale Definition der Zufallsvariable

Eine Zufallsvariable X ist eine Abbildung von (Ω, ϑ) in $(\mathbb{R}, \zeta)$, wobei für jedes $B \in \zeta$ gilt:

$$X^{-1}(B) = \{w | X(w) \in B\} = A \in \vartheta$$

19. Eigenschaften der Verteilungsfunktion

1. F ist monoton wachsend
2. F ist rechtsseitig stetig
3. $\lim_{x\to\infty} F(x) = F(\infty) = 1$
   $\lim_{x\to-\infty} F(x) = F(-\infty) = 0$

20. Poissonverteilung

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

$$p_i = \frac{\lambda^i}{i!} e^{-\lambda}, i = 0, 1, 2, \ldots$$

für ein gewisses $\lambda > 0$ definiert.

$$\sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda} \tag{0.16}$$

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

21. 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, weißt 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 verzerren. Im Wahrscheinlichkeitsnetz wird die Ordinate zwischen 0 und 1 nicht in gleich große Teile geteilt, sondern die Abstände werden proportional zu $G^{-1}(F_n(x))$ über x dargestellt. Wenn nun die Daten ungefähr normalverteilt sind, so wird die Treppenfunktion etwa auf einer Gerade zu liegen kommen.

22. Erwartung und Varianz

$h =$ reele Funktion der Zufallsvariablen X, dann ist der Mittelwert oder die mathematische Erwartung von $h(x)$ im Falle eine stetigen Zufallsvariablen

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

im diskreten Fall

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

Ist $h(X) = x \Rightarrow$

$$\mu = \mathbb{E}(X) = \int xf(x)dx$$

bzw.

$$\mu = \mathbb{E}(X) = \sum x_i p_i$$

$$\sigma^2 = VAR(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}[X - \mathbb{E}X]^2$$
$$VAR(X) = \mathbb{E}(X) - (\mathbb{E}X)^2$$

23. Additionssatz für Mittelwerte

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

$$\mathbb{E}(X_1 + X_2 + \cdots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n) \tag{0.17}$$

24. Multiplikationssatz für Mittelwerte

Für n **unabhängige** Zufallsvariablen $X_1, X_2, \ldots, X_n$, deren Mittelwerte existieren gilt:

$$\mathbb{E}(X_1 X_2 \ldots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2)\ldots\mathbb{E}(X_n) \tag{0.18}$$

25. Varianz der Summe zweier Zufallsvariablen

$$Z = X + Y$$
$$\sigma_X^2 = VAR(X), \qquad \sigma_Y^2 = VAR(Y), \qquad \sigma_Z^2 = VAR(Z)$$

$$\sigma_Z^2 = \mathbb{E}(Z - \mathbb{E}Z)^2 = \tag{0.19}$$
$$\mathbb{E}(X + Y - \mathbb{E}(X + Y))^2 = \tag{0.20}$$
$$\mathbb{E}(X - \mathbb{E}X)^2 + \mathbb{E}(Y - \mathbb{E}Y)^2 + 2\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \tag{0.21}$$
$$\sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \tag{0.22}$$

$$\sigma_{XY} = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)$$

Kovarianz der Zufallsvariablen X und Y.

Für den Fall der Unabhängigkeit von X und Y gilt:

$$\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$$

‚sodass die Kovarianz verschwindet. Der umgekehrte Schluss ist nicht zulässig.

26. Additionssatz für Varianzen

Die Varianz einer Summe **unabhängiger** Zufallsvariablen, deren Varianzen existieren, ist gleich der Summe der Varianzen:

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$$

27. Kovarianz & Korrelation

Die Kovarianz σ_{XY} von zwei Zufallsvariablen X und Y stellt ein Maß für die Abhängigkeit beider dar. Eine Standardisierung dieses Maßes, erhält man, indem man es durch die Streuung von X und Y dividiert

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

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

Schätzungen für die Kovarianz σ_{XY}

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

Korrelation ρ_{XY}

$$R_{XY} = \frac{S_{XY}}{S_X S_Y}$$

28. Zentraler Grenzwertsatz - Aussage

Bei steigendem Stichprobenumfang (aus Gleichverteilung) ist:

- das empirische Mittel $\bar{X}$ für verschiedene n ungefähr gleich (bis auf kleine zufällige Fehler)
- die Varianzen S_X^2 von $\bar{X}$ werden bei steigendem Stichprobenumfang kleiner
- multipliziert man die Varianz S_X^2 allerdings mit n, so erhält man allerdings ungefähr denselben 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.

29. Was ist das Ziel der analytischen Statistik?

Die analytische Statistik soll eine Verbindung zwischen der Theorie und der Wirklichkeit herstellen. D.h. inwieweit und wie können Schlüsse von der Stichprobe auf die Grundgesamtheit gezogegen werden.

30. 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 $(x_1, x_2, \ldots, x_n)$ sind dann wiederholte Realisationen der Zufallsvariablen X, die normalerweise als unabhängig voneinander betrachtet werden.

31. Punktschätzungen

Angenommen: Die Verteilung der Stichprobenelemente x_i enthält einen unbekannten Parameter θ und es gibt eine Funktion t, die aus den Stichprobenwerten den Wert von θ näherungsweise berechnen.

$$\hat{\theta} = t(x_1, x_2, \ldots, x_n)$$

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 $\mathbb{E}(T) = \theta \quad \Rightarrow$

erwartungsgetreu oder unverzerrter Schätzer ⇒ Das arithmetische Mittel $\bar{x}$ der Stichprobe ist ein erwartungsgetreuer 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 erwartungsgetreuer 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 wichtigste. Sie wählt im Wesentlichen jenen Wert des Parameters, der die Stichprobe als „wahrscheinlichstes„Resultat erscheinen lässt.

32. Schätzungen des Mittels einer Population

Tests bezüglich des Mittels μ einer Population stützen sich auf stützen sich auf das Stichprobenmittel $\bar{X}$ und dessen Verteilung. Die Hypothese soll lauten $\mu = \mu_0$, wobei μ_0 ein speziell gewählter Wert ist. Ist sie richtig, so werden sich die Werte von $\bar{X}$ zufällig um μ_0 streuen. Die Wahrscheinlichkeit, dass $\bar{X}$ in den kritischen Bereich (führt zum Verwerfen der Hypothese) fällt, wird Signifikanzzahl/ niveau genannt.

Voraussetzung:

$$X \sim N(\mu_0, \sigma^2) \Rightarrow \bar{X} \sim N(\mu_0, \sigma^2/n)$$

und es isst besser mit der standardisierten Größe

$$Z = \frac{\bar{X} - \mu_0}{\sigma^2/n}$$

zu arbeiten, die $N(0,1)$ verteilt ist. Wenn ein konkreter 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 ⇒ Wenn also μ_0 in das Konfidenzintervall fällt, wird auch der Test die Hypothese $H_0 : \mu = \mu_0$ nicht verwerfen und umgekehrt.

33. Hypothesentest für Varianz

Die Standardabweichung dient als Maß für die Variabilität einer Meßgröße und ist ebenso wichtig wie das Mittel. χ^2 - Verteilung mit n bzw. $n-1$ Freiheitsgraden, je nachdem ob μ bekannt ist oder durch $\bar{x}$ aus der Stichprobe geschätzt werden muss.

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

bzw.

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