3. Eine Anmerkung zur statistischen Signifikanz

Die vorliegende Bemerkung soll mögliche Wirkungen der unkritischen Anwendung von Signifikanztests bezüglich Annahmen zur stochastischen Unabhängigkeit und zur Identität zugrundeliegender Verteilungen vor Augen führen, auch um die vom Autor vertretenen alternativen Controlling-Techniken[1] weiter zu stützen.

In der statistischen Test-Theorie werden sogenannte Gegen- oder Alternativhypothesen H_1 aufgestellt um den aktuellen Wissensstand zu ergänzen oder in Frage zu stellen. An die binäre Logik (0 entspricht „falsch“ und 1 entspricht „wahr“) angelehnt ist die Alternativhypothese im allgemeinen die primär zu verifizierende Aussage, der eine komplementäre Aussage die sogenannte Nullhypothese H_0 gegenübergestellt wird.

Beispiel (In Anlehnung an Bortz[2])

Sind Lernerfolge, die bezüglich herkömmlicher Lehrmethoden erzielt wurden in Form von mittleren Testergebnissen aus einer Vielzahl von gleichartigen Tests bekannt, so dass wir nach dem zentralen Grenzwertsatz[3] den Lernerfolg nach der herkömmlichen Methode als normalverteilt mit bekanntem Erwartungswert μ und bekannter Varianz σ^2 (kurz: $N(\mu,\sigma^2)$-verteilt) ansehen können, so kann mittels einer Alternativhypothese die Behauptung geprüft werden inwieweit die Ergebnisse des Lernerfolges einer neuen Lehrmethode (beispielsweise) signifikant besser sind.

Nehmen wir also an, wir erzielen mit der neuen Lehrmethode Testergebnisse, die eine Verteilung der Ergebnisse, die sich durch eine $N(\mu_1,\sigma^2)$-Verteilung beschreiben lassen nahe legen, wobei $\mu_1>\mu_0$ gelte. Zur Prüfung des Wahrheitsgehalts der Aussage, die neue Lehrmethode sei besser als die herkömmliche Lehrmethode, lässt sich dann die

$$\text{Alternativhypothese } H_1\text{: } \mu_1>\mu_0$$

gegen die

$$\text{Nullhypothese } H_0\text{: } \mu_1\leq\mu_0$$

---

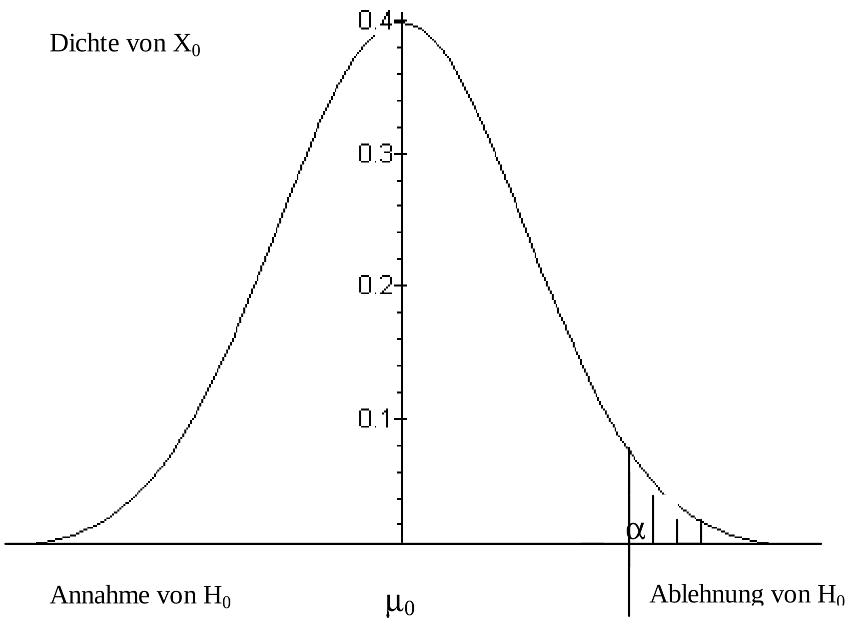
[1] Vgl. [Holz 01], den Anhang und die Discussion Paper Technik.doc, Ktarif.doc und Treiber.doc unter http://www.rankingweb.de/Paper.html.

[2] Vgl. [Bortz 99], Kapitel 4.

[3] Vgl. [Bortz 99], S.93: „Die Verteilung von Mittelwerten aus Stichproben des Umfangs n, die sämtlich derselben Grundgesamtheit entnommen wurden, geht mit wachsendem Stichprobenumfang in eine Normalverteilung über.“ Oder was die wesentlichen technischen Voraussetzungen deutlicher macht, [Bronstein 81], S. 715: „Kann eine Zufallsgröße als Summe einer großen Anzahl voneinander unabhängiger Summanden aufgefasst werden, von denen jeder zur Summe nur einen unbedeutenden Beitrag liefert, so ist diese Zufallsgröße annähernd normalverteilt.“

auf Signifikanz überprüfen, wozu die Hypothesen auf spezifische Verteilungen reduziert werden müssen, weshalb wir die Alternativhypothese mit der Vorstellung verbinden, die Zufallsvariable X_1 der Lehrerfolge der neuen Lehrmethode sei $N(\mu_1,\sigma^2)$-verteilt (kurz: $X_1 \sim N(\mu_1,\sigma^2)$) und die Nullhypothese mit der Vorstellung, die Zufallsvariable X_0 der Lehrerfolge der herkömmlichen Lehrmethode sei $N(\mu_0,\sigma^2)$-verteilt (kurz: $X_0 \sim N(\mu_0,\sigma^2)$).

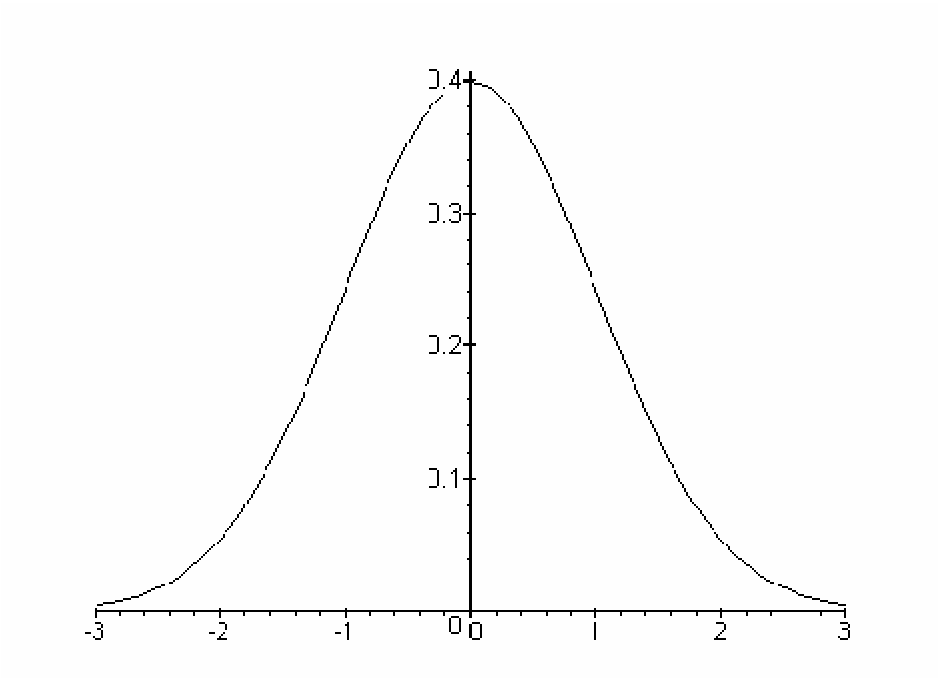
Mit der Festlegung eines Signifikanz-Niveaus α erhalten wir dann Signifikanz für H_1 sofern die Wahrscheinlichkeit $\mathbb{P}(X_0 > \mu_1) \leq \alpha$ ist. Sofern also bei Annahme der Richtigkeit der Nullhypothese nur mit einer Wahrscheinlichkeit kleiner oder gleich dem Signifikanzniveau α auch mit der herkömmlichen Lehrmethode Lernerfolge erzielt werden, die besser oder gleich dem Lernerfolg der neuen Lehrmethode sind. Sofern also der sogenannte Fehler 1. Art des Entscheidens für H_1 bei Richtigkeit von H_0 kleiner oder gleich dem Signifikanz-Niveau α ist. Oder einfacher sofern μ_1 im Ablehnungsbereich - auch kritischer Bereich genannt - von H_0 liegt.



Es sei nun nicht weiter auf mögliche andere beispielsweise zweiseitige Testkonstellationen oder auf die Güte von Tests im Sinne der Trennschärfe dieser eingegangen[4]. Wir wollen stattdessen den Gehalt der bis hier getroffenen Aussagen kritisch hinterfragen.

[4] Vgl. hierzu etwa [Bortz 99], Kapitel 4.

Die eindimensionale Normalverteilung[5]



$$X \sim N(\mu, \sigma^2)$$

empirisch:

Erwartungswert: $E(X) = \mu$ $\qquad \overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$

Varianz: $V(X) = \sigma^2$ $\qquad \overline{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n^2)$

und erwartungstreu $\qquad \hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n^2)$

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

Standardisierung zur N(0,1)-Verteilung: $Y = \frac{X - E(X)}{\sqrt{V(X)}}$

Transformationen: $X \sim N(\mu, \sigma^2), \quad Z = aX + b, \quad a \neq 0$

$$\Rightarrow \quad Z \sim N(a\mu + b, a^2\sigma^2)$$

---

[5] Vgl etwa [Heilmann 87], Kapitel 1.

Aus der Normalverteilung abgeleitete Prüfverteilungen[6]

<table>
<tr>
<td>

[1]: Die Summe unabhängiger normalverteilter Zufallsvariablen ist normalverteilt mit den addierten Erwartungswerten und Varianzen

$$\text{Ist } X_1 \sim N(\mu_1, \sigma_1{}^2),\ X_2 \sim N(\mu_2, \sigma_2{}^2) \quad \text{und} \quad X_1, X_2 \text{ unabhängig}$$

$$\Rightarrow \quad X_1 + X_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$$

</td>
<td>

[2]: Die Summe der Quadrate unabhängig N(0,1)-verteilter Zva'n ist Chi-Quadrat-verteilt mit n Freiheitsgraden

$$\text{Seien } X_1, \ldots, X_n \text{ unabhängig, identisch } N(0,1)-\text{verteilt}$$

$$\Rightarrow \quad X_1^2 + \ldots + X_n^2 \sim \chi_n^2 - \text{verteilt}$$

Beispiele:
Chi-Quadrat-Anpassungstest und Chi-Quadrat-Unabhängigkeitstest

</td>
</tr>
<tr>
<td>

[3]: t_n- oder studentverteilte Prüfgröße mit n Freiheitsgraden

$$\text{Seien } X_0, X_1, \ldots, X_n \text{ unabhängig, identisch } N(0,1)-\text{verteilt}$$

$$\Rightarrow \quad t = \frac{X_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}} \sim t_n - \text{verteilt}$$

Beispiel: t-Test, Vergleich von Mittelwerten bei unbekannter aber identischer Varianz zweier unabhängiger normalverteilter Größen

</td>
<td>

[4]: $F_{m,n}$- oder Fischer verteilte Prüfgröße mit m und n Freiheitsgraden

$$\text{Seien } X_1, \ldots, X_m \text{ und } Y_1, \ldots, Y_n \text{ unabhängig, identisch } N(0,1)-\text{verteilt}$$

$$\Rightarrow \quad F = \frac{\frac{1}{m}\sum_{i=1}^{m} X_i^2}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2} \sim F_{m,n} - \text{verteilt}$$

Beispiel: F-Test, Vergleich der Streuungen zweier unabhängiger normalverteilter Größen

</td>
</tr>
</table>

[6] Vgl etwa [Bortz 99], Abschnitt 2.5.

Kritisch an dem betrachteten vorgehen ist anzumerken, dass:

a) Die Parameter μ der den Hypothesen zugeordneten Verteilungen nur mittels der empirischen Werte $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ geschätzt festgelegt werden können, wobei n den Erhebungsumfang beschreibe. Genügt $\overline{x}_n$ den Voraussetzungen einer Version des Gesetzes der großen Zahlen[7], dann ist $\overline{x}_n$ ein „angemessener“ Schätzer für μ.

b) Nur für große n - allgemeinen[8] wird hier $n \geq 30$ gesetzt - kann für die Stichprobenmittel nach dem zentralen Grenzwertsatz die Normalverteilung als hinreichend gute Näherung der Verteilung herangezogen werden. Bei kleineren Stichprobenumfängen[9] ist die Student- oder auch t-Verteilung genannte Verteilung heranzuziehen, wobei dann aber vorausgesetzt werden muß, dass die Stichprobenwerte normalverteilt sind.

c) Sowohl die Gesetze der großen Zahlen, wie auch der zentrale Grenzwertsatz und die für das ableiten von Prüfverteilungen notwendigen Annahmen bei der Addition von Zufallsgrößen, beruhen ganz wesentlich auf der Annahme der (stochastischen) Unabhängigkeit beteiligter Zufallsgrößen.

Insbesondere der letztgenannte Tatbestand kann eine wesentliche Fehlerquelle für die mittels der statistischen Testtheorie erzielten Ergebnisse sein. Zwar lässt sich die Unabhängigkeit zweier Zufallsvariablen X, Y beispielsweise mittels des χ^2-Unabhängigkeits-Test prüfen, hierbei geht aber implizit die Annahme der Unabhängigkeit der erzielten Stichprobenresultate in die Festlegung der Prüfverteilung bereits ein. Wesentlicher ist aber, dass so gut wie keine Aussagen mehr möglich sind sofern die Annahme der stochastischen Unabhängigkeit nicht gegeben ist, da hierfür weder ein allgemeines Analogon zum zentralen Grenzwertsatz noch zu den Gesetzen der großen Zahlen existieren.

Die Wirkung der Abhängigkeit von Zufallsgrößen bzw. deren Verteilungen wird im Folgenden am Beispiel der zweidimensionalen Normalverteilung verdeutlicht, um hiermit die mögliche Verfälschung von Tests bei Abhängigkeit vor Augen zu führen.

---

[7] Vgl. etwa [Bronstein 81], S. 712f, sinngemäß: Das arithmetische Mittel der Beobachtungen unabhängiger Wiederholungen des identischen Zufallsexperiments konvergiert mit wachsender Anzahl an wiederholungen (fast sicher) gegen den Erwartungswert der wahren Verteilung des Experiments.
[8] Vgl. etwa [Bortz 99], S. 94.
[9] Vgl. etwa [Bortz 99], S. 102.

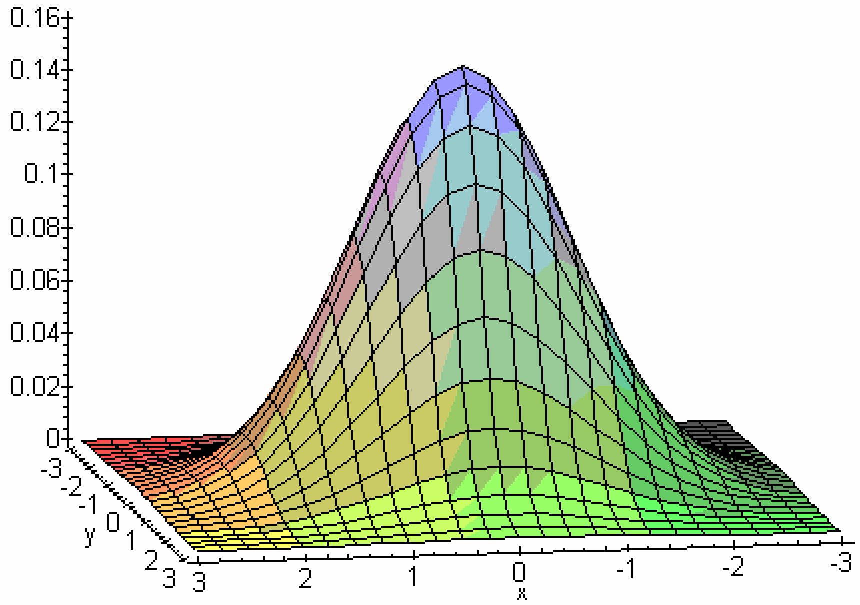
Die zweidimensionale Normalverteilung[10]

$X_1 \sim N(0,1)\,,\; X_2 \sim N(0,1)$

$$f(x_1, x_2) = \frac{e^{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)}}{\sqrt{(2\pi)^2(1-\rho^2)}}$$

ρ der Korrelationskoeffizient von X_1, X_2 .

$$\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{V(X_1)V(X_2)}}$$

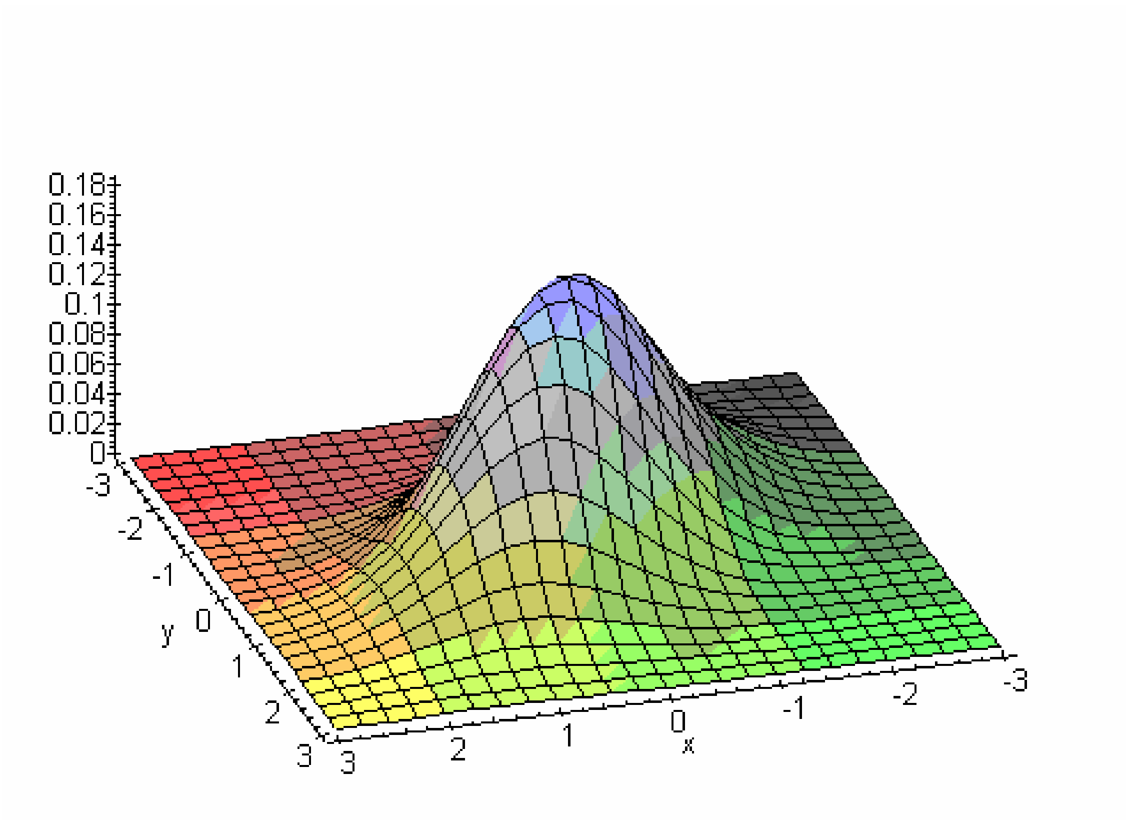


$X_1 \sim N(0,1)\,,\; X_2 \sim N(0,1)\,,$ X_1, X_2 unabhängig

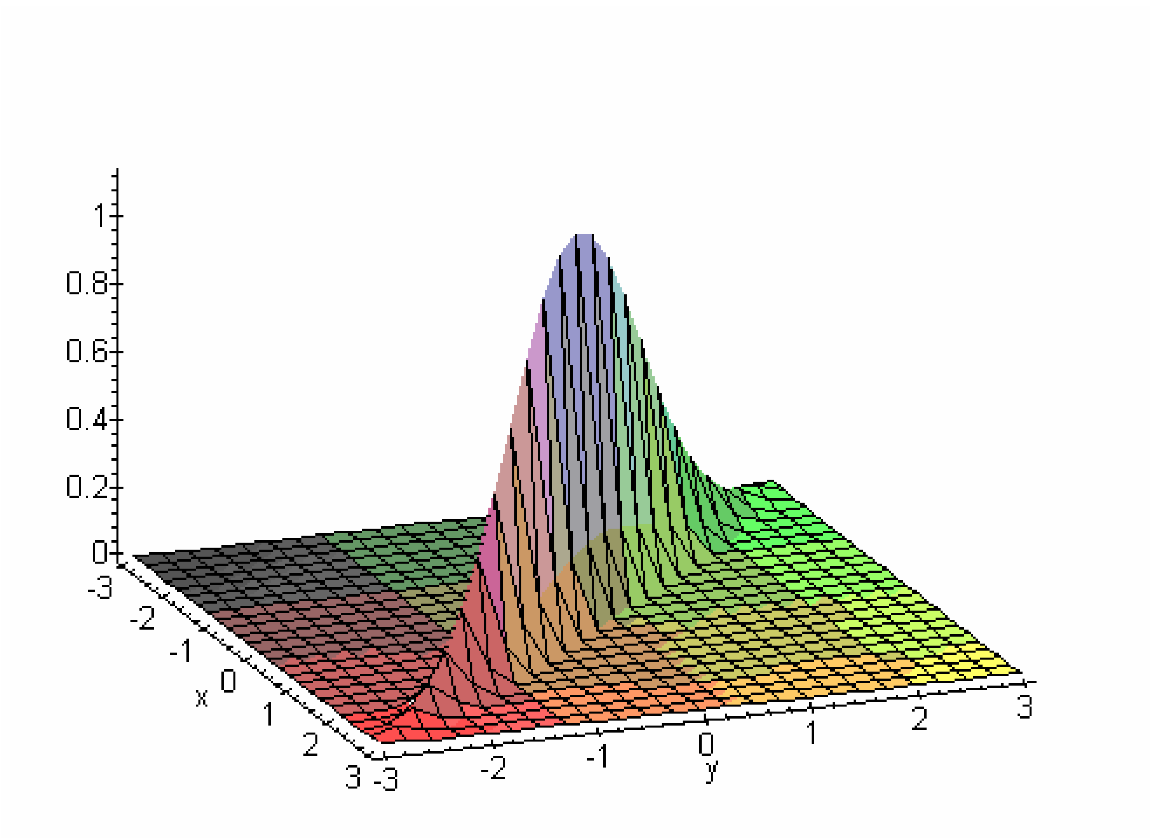
[10] Vgl. etwa [Elpelt 95], S. 64ff

Korrelierte zweidimensionale N(0,1)-Verteilungen

- ρ =0,5:



- ρ =-0,99:



Die stochastische Unabhängigkeit von gemeinsam gültigen Ereignissen E_x und E_y zweier Zufallsvariablen X und Y ist über die Multiplikation der Wahrscheinlichkeiten definiert:

$$\mathbb{P}(X = E_x, Y = E_y) \overset{\text{st. unabh.}}{:=} \mathbb{P}(X = E_x)\,\mathbb{P}(Y = E_y)\,.$$

Zufallsvariablen selbst sind dann als stochastisch unabhängig definiert, wenn die Multiplikationsregel für alle möglichen Kombinationen von Ereignissen von X, Y Gültigkeit besitzt. Sind die Zufallsvariablen X, Y stetig verteilt mit Dichten f(x) und g(y) und beschreibt h(x,y) die Dichte der gemeinsamen Verteilung von (X,Y), dann ist die stochastische Unabhängikeit von X, Y auch durch die Beziehung h(x,y) = f(x)g(y) gegeben.

Sind Zufallsvariablen X, Y stochastisch unabhängig so sind Sie auch unkorreliert. Andererseits folgt im allgemeinen aus der Unkorreliertheit nicht die stochastische Unabhängigkeit[11]. Womit ein Maß zur Beurteilung der Unabhängigkeit allgemein nicht gegeben ist. Eine Ausnahme bildet hier der Fall, dass X, Y normalverteilt sind, wie man am Spezialfall oben leicht sieht. Wir wollen den Spezialfall zweier normalverteilter Zufallsvariablen nutzen, um die Wirkung von Abhängigkeit auf die Summe der Zufallsvariablen zu prüfen, da wie die Prüfverteilungen oben zeigen, die Addition von unabhängigen Zufallsvariablen zu wichtigen Prüfgrößen führt.

Sind X, Y Verteilungen über den ganzen Zahlen, dann gilt allgemein[12]:

$$\mathbb{P}(X + Y = z) = \sum_k \mathbb{P}(X = k)\,\mathbb{P}(X + Y = z \mid X = k) = \sum_k \mathbb{P}(X = k)\,\mathbb{P}(Y = z - k \mid X = k)$$

und im Fall der Unabhängigkeit von X,Y

$$\mathbb{P}(X + Y = z) \overset{\text{st.unabh.}}{=} \sum_k \mathbb{P}(X = k)\,\mathbb{P}(Y = z - k) = \sum_r \mathbb{P}(Y = r)\,\mathbb{P}(X = z - r)$$

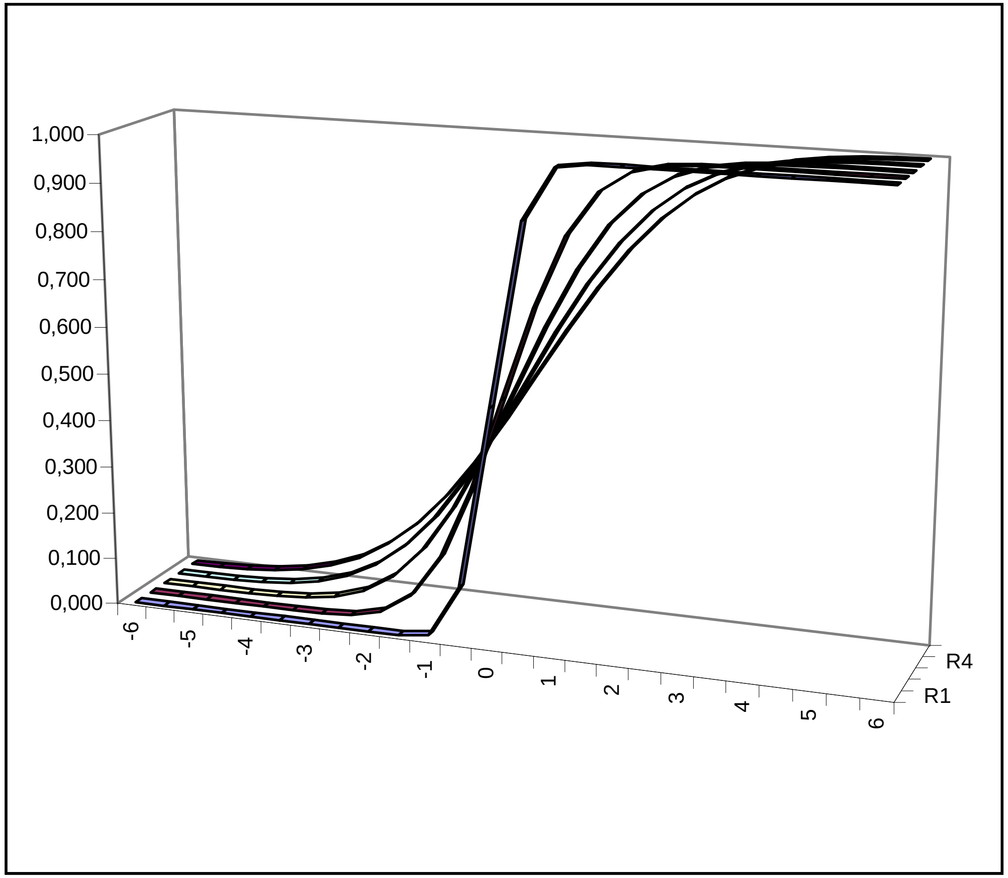
Im stetigen Fall ergibt sich die Verteilung von X+Y dann analog allgemein zu:

$$\mathbb{P}(X + Y \le z) = \int \mathbb{P}^X(dx)\,\mathbb{P}(X + Y \le z \mid X = x)\,.$$

[11] Vgl. etwa [Fisz 89], S.112
[12] Vgl. etwa [Heilmann 87], S. 61ff auch für die folgenden Überlegungen.

Verteilungsfunktionen der Summen X+Y zweier korrelierter N(0,1)-Verteilungen

Korrel	-0,9	-0,5	0	0,5	0,9
-6	0,000	0,000	0,000	0,000	0,000
-5,5	0,000	0,000	0,000	0,000	0,002
-5	0,000	0,000	0,000	0,001	0,004
-4,5	0,000	0,000	0,001	0,004	0,010
-4	0,000	0,000	0,002	0,010	0,019
-3,5	0,000	0,000	0,006	0,021	0,035
-3	0,000	0,001	0,016	0,040	0,061
-2,5	0,000	0,006	0,037	0,073	0,099
-2	0,000	0,021	0,077	0,123	0,151
-1,5	0,000	0,064	0,142	0,192	0,220
-1	0,009	0,156	0,238	0,280	0,303
-0,5	0,118	0,306	0,360	0,385	0,398
0	0,499	0,499	0,499	0,499	0,499
0,5	0,881	0,693	0,638	0,613	0,601
1	0,990	0,843	0,761	0,718	0,696
1,5	0,999	0,934	0,856	0,807	0,779
2	0,999	0,977	0,922	0,876	0,848
2,5	0,999	0,993	0,961	0,926	0,900
3	0,999	0,997	0,983	0,958	0,938
3,5	0,999	0,998	0,993	0,978	0,964
4	0,999	0,999	0,997	0,989	0,980
4,5	0,999	0,999	0,998	0,995	0,989
5	0,999	0,999	0,998	0,997	0,994
5,5	0,999	0,999	0,999	0,998	0,997
6	0,999	0,999	0,999	0,999	0,998



Approximiert über die Werte der gemeinsamen Dichte der korrelierten N(0,1)-verteilten Zufallsvariablen (X,Y) an den

Stützstellen (-3,25; -2,75; -2,25; ...; 2,25; 2,75; 3,25) x (-3; -2,5; -2; ... ;2; 2,5; 3)

Mit der Tabelle und dem Schaubild oben ist die Verteilung der Summe X+Y zweier N(0,1)-verteilter Zufallsvariablen für die Korrelationen -0,9; -0,5; 0; 0,5; 0,9 dargestellt. Übliche Signifikanz-Niveaus sind α=0,05 und α=0,01. In den Verteilungen oben sind die dem Niveau α=0,01 am nächsten kommenden Ränder des Ablehnungsbereichs einer auf der Verteilung von X+Y basierenden Nullhypothese im obigen Sinne beispielshalber unterlegt kenntlich gemacht, wodurch sichtbar wird wie stark sich die Verletzung der Annahme der stochastischen Unabhängigkeit auf das wahre Signifikanz-Niveau auswirken kann.

So erhalten wir den Rand des Ablehnungsbereichs zum Signifikanz-Niveau von α=0,01 im unkorrelierten Fall (ρ=0) bei einem Wert von ca. 3,5, im Fall der Korrelation von 0,9 bei einem Wert von ca.4,5 und im Fall der negativen Korrelation von –0,9 schon bei einem Wert von 1, was bei der Korrelation von +0,9 einem Signifikanz-Niveau von ca. 30% entspräche.

Es bleibt anzumerken, dass hier symmetrische Verteilungen betrachtet sind, sogenannte links- oder rechtsschiefe Verteilungen sowie die fälschliche Annahme der Identität der den Stichprobenelementen zugrundeliegenden Verteilungen, können die aufgezeigte Wirkung der nicht gegebenen Unabhängigkeit auf die wahre Verteilung von Prüfgrößen noch wesentlich deutlicher werden lassen, womit lediglich verdeutlicht sei, dass die unkritische Anwendung von Tests, insbesondere in Bezug auf die Annahme der stochastischen Unabhängigkeit zu doch deutlichen Fehleinschätzungen führen kann. Im allgemeinen bleibt darüber hinaus die zeitliche Dimension der Erhebung und des Vergleichs von Stichproben unberücksichtigt was zu weiteren Fehlerquellen bzw. eher zufälligen Ergebnissen führen kann.

Hiermit soll nun keinesfalls das notwendige Testen von Aussagen in Frage gestellt werden. Instrumentarien des Prüfens bei Verletzung der notwendigen Annahmen für statistische Tests fehlen aber weitgehend.

Die Schlussfolgerung kann daher nur sein, dass es notwendig wird real gegebene Datenlagen besser vor Augen zu führen um ein verantwortungsvolleres Entscheiden zu ermöglichen, wozu die alternativen Controlling-Techniken des Autors entwickelt sind[13].

Eine Testsituation die auch explizit den begrenzten Gehalt der Aussagen statistischer Signifikanz vor Augen führt ist mittels der folgend beschriebenen Diskriminanzanalyse ab der Version Regio2004.xls den Analyse-Dateien beigefügt.

---

[13]Vgl. [Holz 01], den Anhang oder aktueller das Dokument http://www.rankingweb.de/MANUAL.pdf und die Datei Regio2003.xls unter http://www.rankingweb.de/Downloads.html

Diskriminanzanalyse über Mahalanobis-Distanzen[14]

Seien X_1, .., X_J Zufallsvariablen, die die Ausprägungen der Merkmale einer Objektmenge X beschreiben und i ein X in zwei Gruppen differenzierender Index mit n_i, i=1,2 Objekten aus X. Unter der Annahme, dass die X_{i1}, .., X_{iJ} multivariat normal verteilt sind mit Erwartungswerten $\mu_i=(\mu_{i1},...,\mu_{iJ})$ und Covarianzmatrix Σ,

$$\bar{x}_i = (\bar{x}_{i1},..,\bar{x}_{iJ}) \quad \text{mit} \quad \bar{x}_{ij} = \frac{1}{n_i}\sum_{t=1}^{n_i} x_{ijt}\ , \ \text{i=1,2 und}$$

$$s_{pq} = \frac{1}{n_1+n_2-2}\sum_{i=1}^{2}\ \sum_{t=1}^{n_i}(x_{ipt}-\bar{x}_{ip})(x_{iqt}-\bar{x}_{iq}), \quad S_J = \{s_{p,q} : p,q = 1,..,J\}$$

Schätzer für $\mu_i=(\mu_{i1},...,\mu_{iJ})$ und die Komponenten von Σ, ist die Distanz einer Auswahl von $k \leq J$ Merkmalen gegeben durch

$$D_k^2 = \sum_{p=1}^{k}\ \sum_{q=1}^{k}(\bar{x}_{1p}-\bar{x}_{2p})s'_{pq}(\bar{x}_{1q}-\bar{x}_{2q}), \quad p,q = 1,..,J$$

s'_{pq} die Komponenten der Inversen von S_k und die Nullhypothese der Identität der Erwartungen gegeben k Merkmale unter Hinzunahme eines weiteren Merkmales k+1 aus 1,..,J

$$H_0 : E\left[X_{1,k+1} \middle| X_{1,1},..,X_{1.k}\right] = E\left[X_{2,k+1} \middle| X_{2,1},..,X_{2.k}\right]$$

kann aus der Kenntnis der Verteilung von

$$F = \frac{(n_1 n_2)(n_1+n_2-k-2)}{(n_1+n_2)(n_1+n_2-2)} \frac{\left[D_{k+1}^2 - D_k^2\right]}{(1+\dfrac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)}D_k^2)}$$

als Fischer-Verteilung mit (1, n_1+n_2-k-2) Freiheitsgraden zum Niveau α mittels

$$F > F_\alpha(1, n_1+n_2-k-2)$$

geprüft werden. Intuitiv wird ersichtlich, dass je größer D_k^2 ist, desto größer auch der diskriminierende Charakter der gewählten Gruppen im Sinne des Abstandes der Merkmale ist.

---

[14] Vgl. J. Van Eeghen, E.K. Greup, J.A. Nijssen: „Rate Making“, Surveys of Actuarial Studies No. 2, 1983, Nationale-Nederlanden N.V., Rotterdam, Niederlande.

Literatur:

J. Bortz: „Statistik für Sozialwissenschaftler“, Springer 5. Aufl. 1999.

I.N. Bronstein, K.A. Semendjajew: „Taschenbuch der Mathematik“, Teubner 1981.

J. Van Eeghen, E.K. greup, J.A. Nijssen: „Rate Making“, Surveys of Actuarial Studies No. 2, 1983, Nationale-Nederlanden N.V., Rotterdam, Niederlande.

B. Elpelt, J. Hartung: „Multivariate Statistik“, Oldenburg 1995.

M. Fisz: „Wahrscheinlichkeitsrechnung und mathematische Statistik“, VEB Berlin 1989.

W.-R. Heilmann: „Grundbegriffe der Risikotheorie“, VVW Karlsruhe 1987.

R. Holz: „Großstädte-Ranking 2001“, Shaker Aachen 2002.

Und das Discussion Paper http://www.rankingweb.de/MANUAL.pdf