W. Dörfler

Einführung in das statistische Testen

I. Testen von Hypothesen

I.1. Allgemeines

Das Grundproblem statistischer Tests läßt sich so fassen: Über die Verteilung einer Zufallsgröße (bzw. die Wahrscheinlichkeitsverteilung eines Merkmals in einer Grundgesamtheit) liegen gewisse Kenntnisse bzw. Annahmen vor, die man aus der Erfahrung oder durch theoretische Überlegungen gewonnen hat, oder die auch nur Angaben eines Geschäftspartners (Qualität einer Lieferung) sind. Durch Ziehen von Zufallsstichproben aus der jeweiligen realen oder gedachten Grundgesamtheit soll überprüft werden, wie weit die Kenntnisse bzw. Annahmen der "Realität" entsprechen, bzw. ob sie korrigiert werden sollen. Dabei sind grundsätzlich keine absoluten Aussagen zu erwarten, Fehler sind unvermeidbar, man möchte aber bei gegebener maximaler Fehlergröße die Wahrscheinlichkeit für einen derartigen Fehler wissen.

Beispiel 1. Eine Partei hatte bei den letzten Wahlen einen Stimmanteil von 43 % und möchte wissen, ob sie diesen Anteil gehalten hat. Die Vermutung, "Hypothese" ist, daß der Anteil $p \geq 43\ \%$ ist. Durch Befragung (zufällige Stichprobe) von n Personen soll die Hypothese überprüft, getestet werden.

Beispiel 2. Ein Großhändler behauptet, daß in einer umfangreichen Lieferung eines Massenartikels 3 % Ausschuß sind. Durch Ziehen einer Stichprobe (eines Loses) möchte ein Abnehmer mit großer Sicherheit (Wahrscheinlichkeit) überprüfen, ob der Ausschußanteil $\leq 3\ \%$ ist.

Beispiel 3. Eine Blechwalzstraße soll Blech der Dicke μ erzeugen, wobei eine Toleranz von maximal 0,01 mm zugelassen ist. Durch Überprüfung eines Loses soll festgestellt werden, ob die Maschine justiert werden muß oder nicht. Dabei soll die Wahrscheinlichkeit einer unnötigen Unterbrechung der Produktion nicht mehr als 0,1 % betragen.

I.2. Hypothesen über Anteile; Binomialverteilung

Zur Darstellung der grundlegenden Vorgangsweise bei derartigen Problemen werden wir den Fall von Beispiel 1 näher untersuchen. Die allgemeine Situation im Zusammenhang mit Anteilen eines Merkmals (etwa: Welche Partei wählt ein Bürger) in einer Grundgesamtheit läßt sich durch zufälliges Ziehen von Kugeln verschiedener Farbe aus einer Urne beschreiben. Da bei den üblicherweise in Frage kommenden Problemen die Grundgesamtheiten sehr groß sind, besteht zwischen dem Ziehen einer Stichprobe (Ziehen von n Kugeln ohne Zurücklegen) und dem Ziehen mit Zurücklegen kein wesentlicher Unterschied. Da die Wahrscheinlichkeitsverteilungen für letztere Ziehungsart einfacher sind, wollen wir annehmen, daß wir zufällig Kugeln aus der Urne mit Zurücklegen ziehen und aus den erhaltenen Farben Schlüsse auf die Anteile der Farben unter allen Kugeln zeihen wollen.

Anmerkung. Ist die Grundgesamtheit "klein" und nicht um Größenordnungen größer als der Umfang einer Stichprobe, so muß man zur Behandlung des gestellten Problems beim Ziehen von Stichproben als Teilmengen der Grundgesamtheit (Ziehen ohne Zurücklegen) die hypergeometrische Verteilung verwenden statt der hier betrachteten Binomialverteilung. Natürlich läßt sich die im Folgenden dargestellte Methode auch bei kleinen Grundgesamtheiten verwenden. Die prinzipielle Vorgangsweise ist jeweils dieselbe, der Test-Versuch ist jedoch verschieden.

Über die zu testende Urne treffen wir aufgrund bestimmter Vorinformationen die Annahme, daß der Anteil der weißen Kugeln 30 % ist. Diese Hypothese soll durch 100-maliges zufälliges Ziehen einer Kugel (mit Zurücklegen) getestet werden. Dabei zählen wir, wie oft eine weiße Kugel gezogen wird. Gefühlsmäßig sollten dabei erhaltene Anzahlen "nahe bei 30" die Hypothese stützen, solche "weit von 30" sie schwächen. Um dies zu präzisieren, stellen wir ein Gedankenexperiment an, indem wir alle denkbaren Ergebnisse solcher zufälliger Ziehungen von 100 Kugeln betrachten. Das n-malige Ziehen einer Kugel ist ein Zufallsexperiment, bei dem die Anzahl der weißen Kugeln die interessierende Zufallsgröße X ist: jedem Ausgang des Versuchs entspricht ein Wert k dieser Zufallsgröße. Für diese Werte gilt offenbar $0 \leq k \leq 100$. Das zufällige

Ziehen der 100 Kugeln mit Zurücklegen bedeutet insbesondere, daß die einzelnen Ziehungen voneinander unabhängig sind. Wenn daher der Anteil der weißen Kugeln in der Urne gleich p, $0 \leq p \leq 1$, ist, so ist die Verteilung der Zufallsgröße X eine Binomialverteilung u.zw. B(100,p):

$$P(X=k) = \binom{100}{k} p^k (1-p)^{100-k}, \quad 0 \leq k \leq 100$$

Mit P wird im folgenden stets die (aus dem Kontext festgelegte) Wahrscheinlichkeit bezeichnet.

Bei gegebener Verteilung der Grundgesamtheit (Kugeln in der Urne; bestimmt durch p) kann man also die Verteilung der "Testgröße X"= Anzahl der weißen Kugeln bei 100 zufälligen Ziehungen mit Zurücklegen ermitteln. Wir weisen darauf hin, daß dieser Zusammenhang ganz wesentlich die Zufälligkeit bzw. Unabhängigkeit der Ziehungen zur Voraussetzung hat.

Man kann X auch so interpretieren. Ist X_i, i=1,..., 100, die Zufallsgröße mit $X_i=1$, wenn die i-te Kugel weiß ist, und $X_i=0$, wenn die i=te Kugel nicht weiß ist, so ist $X=X_1+X_2+....+X_{100}$. Die X_i sind dabei unabhängige Zufallsgrößen mit B(1,p) als Verteilung. Die Zufallsgröße $\frac{1}{100}X=\frac{1}{100}(X_1+...+X_{100})$ ist dann der relative Anteil von "weiß" unter den 100 Ziehungen.

Unsere getroffene Annahme bedeutet nun: X ist nach B (100; 0,3) verteilt. Nun führen wir den Test-Versuch konkret durch und erhalten einen Wert x_o für X, $(0 \leq x_o \leq 100)$. Hier erscheint nun folgende Überlegung vernünftig: Ist der Wert x_o unter der Verteilung von X sehr unwahrscheinlich, so ist dies ein Argument gegen unsere Annahme. Noch stärker wird dieses Argument, wenn sogar die Wahrscheinlichkeit α dafür, daß X Werte annimmt, die so weit wie x_o oder noch weiter von 30 entfernt sind, klein ist. Dabei ist ja 30 der Erwartungswert von X. Was dabei "klein" heißen soll, hängt von der Fragestellung und der "Bedeutung" der Urne ab, d.h. welche Realsituation durch die Urne modelliert wird. Meist ist α =0,05 oder 0,01 oder auch 0,1. Als Formel läßt sich das Argument gegen die Hypothese "Anteil der weißen Kugeln ist 30 %" somit schreiben als

$$P(|X-30| \geq |x_o-30|) = \sum_{|k-30| \geq |x_o-30|} \binom{100}{k} 0{,}3^k \, 0{,}7^{100-k} \leq \alpha$$

Gleichwertig damit ist, daß der Wert x_o von X beim durchgeführten Versuch außerhalb des symmetrischen Intervalls um 30 liegt, in das X mit Wahrscheinlichkeit $1-\alpha$ fällt (wenn die Annahme stimmt):

$x_o \notin A = \{k \mid |k-30| \leq r\}$, wobei r kleinstmöglich gewählt ist,

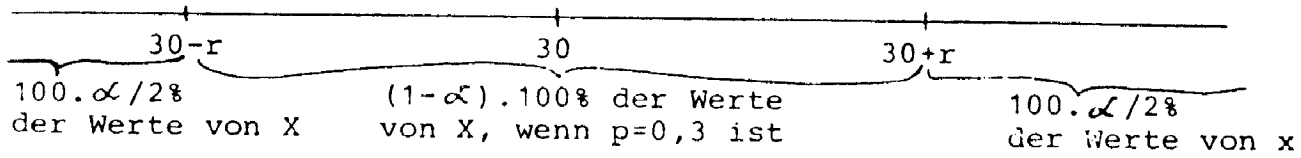
sodaß (P_X=Wahrscheinlichkeit unter der Verteilung von X):

$$P_X(A) = P_X(|X-30| \leq r) \geq 1-\alpha$$

Ist dagegen $x_o \in A$, so ist (bei der hier festgelegten Übereinkunft) das Ergebnis des Versuchs kein Argument gegen die Hypothese. Ein Argument gegen die Hypothese ergibt sich, wenn $|x_o-30| > r$ ist. Es gilt ferner:

$$P(|X-30| > r) \leq \alpha \quad [\text{bzw. } P(X < 30-r \vee X > 30+r) \leq \alpha]$$

Graphisch läßt sich das wie folgt veranschaulichen, wobei wir verwenden, daß B (100; 0,3) schon weitgehend eine symmetrische Verteilung ist:



Darauf beruht nun eine sogenannte Entscheidungsregel: Ergibt der Versuch ein derartiges Argument gegen die Annahme, so wird sie verworfen und man setzt Aktivitäten, als ob die Annahme falsch wäre (sie könnte ja auch noch immer stimmen!). Andernfalls verhält man sich so, als ob die Annahme zuträfe (sie könnte ja auch trotzdem falsch sein). Die im ersten Fall gezogenen Konsequenzen hängen ersichtlich von der jeweiligen konkreten Situation ab (Zurückweisung der Lieferung, Überprüfung der Politik usf.). Man nennt in diesem Zusammenhang das Intervall A auch den Annahmebereich und $\{k \mid |k-30| > r\}$ den Ablehnungsbereich.

In der Beschreibung der Entscheidungsregel sind schon Fehler- bzw. Irrtumsmöglichkeiten angedeutet. Man könnte die Annahme verwerfen, obwohl sie zutrifft. Dies nennt man den Fehler 1. Art. Die Wahrscheinlichkeit dafür ist gerade die Wahrscheinlichkeit für ein "Gegenargument", wenn die Annahme zutrifft, d.h. X nach B(100; 0,3) verteilt ist, und diese ist wiederum gleich α . Die andere Irrtumsmöglichkeit ist, daß wir die Annahme nicht verwerfen, obwohl sie falsch ist: Fehler 2. Art. Die Wahrscheinlichkeit dafür hängt von dem tatsächlichen relativen Anteil der weißen Kugeln in der Urne ab. Ist dieser etwa 100p%, so ist die Testgröße X "in Wahrheit" nach B(100,p) verteilt. Die Wahrscheinlichkeit für einen Fehler 2. Art ist dann die Wahrscheinlichkeit unter B(100, p) dafür, daß X in den Annahmebereich fällt. Somit kann man zusammenfassend kurz sagen:

P(Fehler 1. Art)=P(Ablehnungsbereich) unter der getroffenen Hypothese

P(Fehler 2. Art)=P(Annahmebereich) unter der tatsächlichen Verteilung von X

Dabei wird der Annahmebereich so konstruiert, daß P(Fehler 1. Art) höchstens gleich α ist, wogegen P(Fehler 2. Art) offensichtlich von der unbekannten Verteilung in der Grundgesamtheit abhängt. Betrachtet man den Parameter p der Grundgesamtheit (den Anteil der weißen Kugeln) als variabel in $0 \leq p \leq 1$, so ergibt sich bei festgehaltenem Annahmebereich (d.h. bei festgelegter Entscheidungsregel) eine Funktion $p \mapsto P_p$(Fehler 2. Art), die man die Operationscharakteristik zur gegebenen Entscheidungsregel nennt. Wir werden diesen Zusammenhang an einem Beispiel illustrieren.

Wir betonen, daß mit den bisher durchgeführten Überlegungen die wesentlichen Schritte der meisten Testverfahren dem Prinzip nach dargelegt wurden. Unterschiede im Detail, in der technischen Durchführung entstehen durch verschiedene Testgrößen, unterschiedliche Verteilungen, Formulierung der Hypothese, Festlegung des Annahmebereiches u.a. Vor allem für die Festlegung der Entscheidungsregel (und damit der Wahrscheinlichkeiten für den Fehler 1. und 2. Art) ist die jeweilige Praxissituation aus-

schlaggebend. Es muß abgewogen werden, welches der beiden Risiken für einen Fehler 1. oder 2. Art eher vertretbar ist, und welches Ausmaß tolerierbar ist. Wir kommen in den Beispielen darauf zurück.

Beispiel. In einer Urne befinden sich 12 Kugeln und man weiß, daß mindestens zwei davon schwarz und die restlichen weiß sind. Die zu testende Hypothese H_o (man spricht auch von Nullhypothese) lautet: Es sind genau zwei schwarze Kugeln in der Urne. Der Test soll durch n-faches Ziehen mit Zurücklegen durchgeführt werden. Die Wahrscheinlichkeit p (der relative Anteil) für "schwarz" kann folgende Werte annehmen: $\frac{2}{12}, \frac{3}{12}, \ldots\ldots, \frac{11}{12}, 1$; dies sei die Menge W. Die Nullhypothese lautet somit:

$$H_o: p = p_o = \frac{1}{6}$$

und die sogenannte Gegenhypothese H_1

$$H_1: p \in W - \left\{\frac{1}{6}\right\} \quad \text{bzw.} \quad p > p_o.$$

Die Testgröße X="Anzahl der schwarzen Kugeln bei n Ziehungen" ist dann B(n,p) verteilt, wobei $p \in W$ unbekannt ist. Unter H_o ist X nach $B(n, \frac{1}{6})$ verteilt. Durch $H_1: p > p_o$ liegt hier schon ein Unterschied zur oben durchgeführten Überlegung vor, wo H_1 als $H_1: p \neq 0,3$ zu formulieren gewesen wäre. Das hat natürlich auch Einfluß auf die Entscheidungsregel: Der Annahmebereich wird solche Werte x von X umfassen, die nicht "allzu viel größer" als $\frac{n}{6}$ sind; $\frac{n}{6}$ ist dabei der Erwartungswert von X bei Gültigkeit von H_o. Man wird also ein $k \in \{0,1,\ldots,n\}$ festlegen und damit als Entscheidungsregel E_k formulieren (x ist der Wert von X beim Test):

$$E_k: \begin{cases} \text{für } x > k & H_o \text{ ablehnen} \\ \text{für } x \leq k & H_o \text{ nicht ablehnen.} \end{cases}$$

Bei der Entscheidungsregel E_k ist die Wahrscheinlichkeit α_k für den Fehler 1. Art gegeben durch

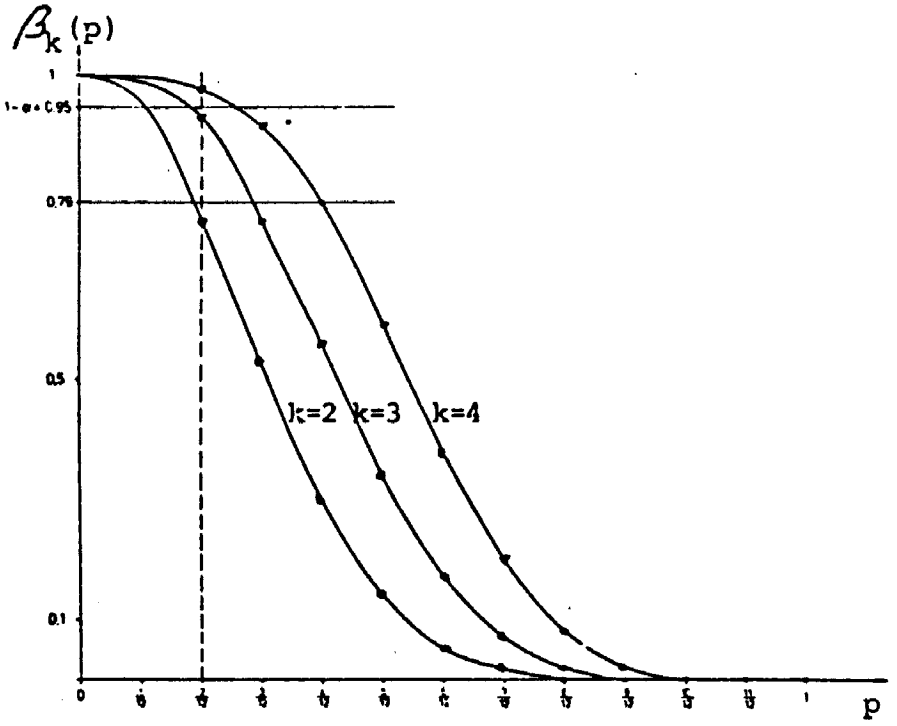
$$\alpha_k = P(X > k \mid H_o) = \sum_{i=k+1}^{n} \binom{n}{i} \left(\frac{1}{6}\right)^i \left(\frac{5}{6}\right)^{n-i}$$

Die Wahrscheinlichkeit β_k für den Fehler 2. Art hängt von der tatsächlichen Verteilung, d.h. vom Wert $p \in W$ ab:

$$\beta_k(p) = P(X \leq k \mid p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

Für jede Entscheidungsregel E_k erhalten wir damit eine Operationscharakteristik: $p \longmapsto \beta_k(p)$, die zunächst eine Funktion auf W ist, aber natürlich ohne weiteres als Funktion auf [0,1] aufgefaßt werden kann. Für k=2,3,4 sind in Figur 1 die Graphen von $\beta_k(p)$ eingetragen. Die Punkte auf den Graphen markieren die Funktionswerte für $p \in W$. Aus der Operationscharakteristik β_k ist auch α_k zu entnehmen:

$$\alpha_k = 1 - \beta_k(\tfrac{1}{6}).$$



Figur 1

Die Operationscharakteristik ist nun eine gute Hilfe bei der Festlegung der Entscheidungsregel, d.h. des k für E_k. In der Regel wird eine obere Schranke α für das Risiko 1. Art=P(Fehler 1. Art) vorgegeben, die man auch als Signifikanzniveau oder kurz Niveau des Tests bezeichnet. Es muß also $\alpha_k \leq \alpha$ sein, sodaß nur solche k in Frage kommen, wo auf dem Graphen von β_k Funktionswerte (Punkte) mit $\beta_k(\frac{1}{6}) \geq 1-\alpha$ existieren. Für $\alpha = 0{,}05$ ist dies nur für $k \geq 4$ der Fall, für $\alpha = 0{,}1$ für $k \geq 3$. Da mit zunehmendem k

der Annahmebereich größer wird, und damit auch $\beta_k(p)$ für jedes p wächst, nimmt man jedenfalls das kleinste k mit den erforderlichen Eigenschaften, also mit $\beta_k(\frac{1}{6}) \geq 1-\alpha$. So wählt man bei $\alpha = 0,1$ $k=3$ und aus den Operationscharakteristiken ist zu ersehen: $\beta_3(p) < \beta_4(p)$ für alle $p \in W$. Man entnimmt Figur 1 auch etwa $\beta_3(\frac{4}{12}) = 0,56$, d.h. sind tatsächlich 4 schwarze Kugeln in der Urne, so ist bei der Entscheidungsregel E_3 die Wahrscheinlichkeit, H_o nicht abzulehnen, gleich 0,56. Aus den Operationscharakteristiken ist auch noch ersichtlich, daß eine Verkleinerung der Wahrscheinlichkeit des Fehlers 1. Art eine Vergrößerung der Wahrscheinlichkeit für den Fehler 2. Art nach sich zieht und umgekehrt.

Den zuletzt erwähnten Zusammenhang zwischen den Fehlern 1. und 2. Art wollen wir in einem ganz einfachen Beispiel noch genauer untersuchen.

Beispiel. Wir gehen davon aus, daß die Wahrscheinlichkeit für "Schwarz" beim Ziehen von Kugeln aus einer Urne entweder $\frac{1}{6}$ oder $\frac{1}{2}$ beträgt und formulieren als Null- und Gegenhypothese:

$$H_o: \; p=\frac{1}{6}$$
$$H_1: \; p=\frac{1}{2}$$

Als Testgröße X wählen wir die Anzahl der schwarzen Kugeln bei n unabhängigen Ziehungen (mit Zurücklegen). Unter H_o (wenn H_o gilt) ist X somit nach $B(n, \frac{1}{6})$ und unter H_1 nach $B(n, \frac{1}{2})$ verteilt. Es sind wieder verschiedene Entscheidungsregeln E_k denkbar (x ist der Wert von X beim Test).

$$E_k = \begin{cases} \text{für } x > k & H_o \text{ ablehnen} \\ \text{für } x \leq k & \begin{cases} H_o \text{ nicht ablehnen} \\ (H_1 \text{ annehmen}) \end{cases} \end{cases}$$

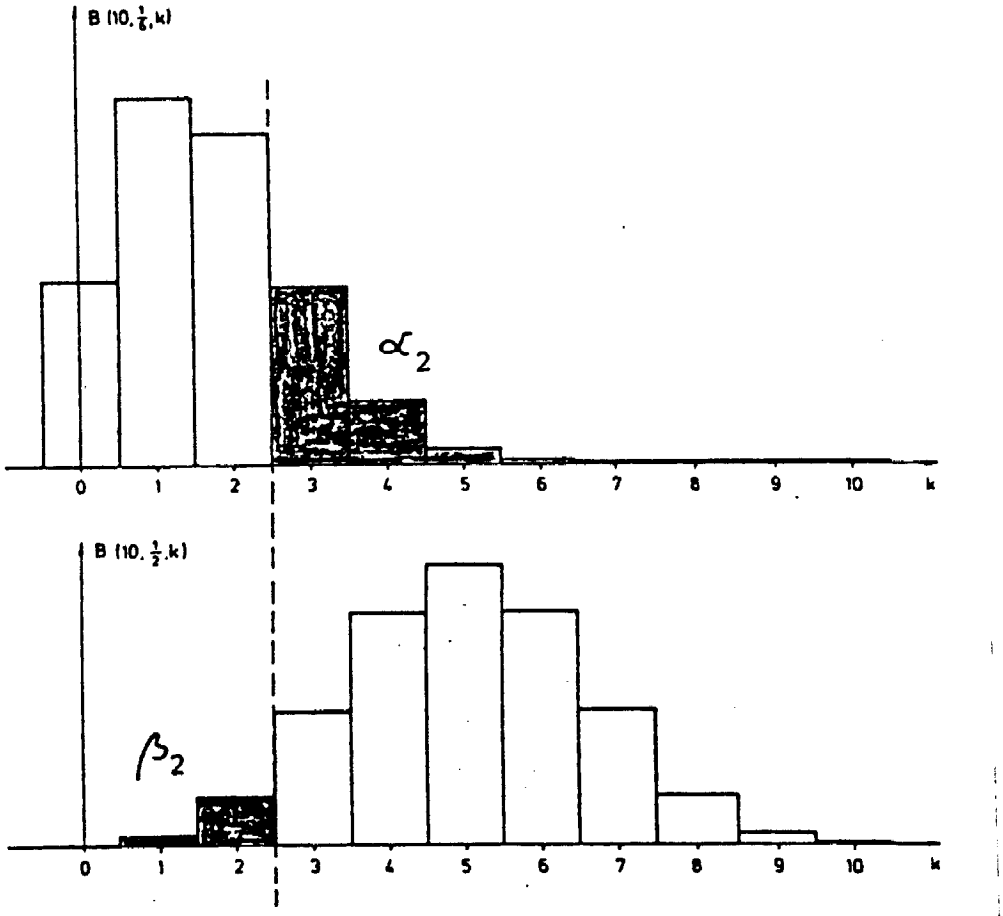
Man nennt dieses k auch den kritischen Wert. Die Wahrscheinlichkeiten α_k und β_k sind dann

$$\alpha_k = \sum_{i=k+1}^{n} \binom{n}{i} \left(\frac{1}{6}\right)^i \left(\frac{5}{6}\right)^{n-i} = P(X > k \mid H_o)$$

und

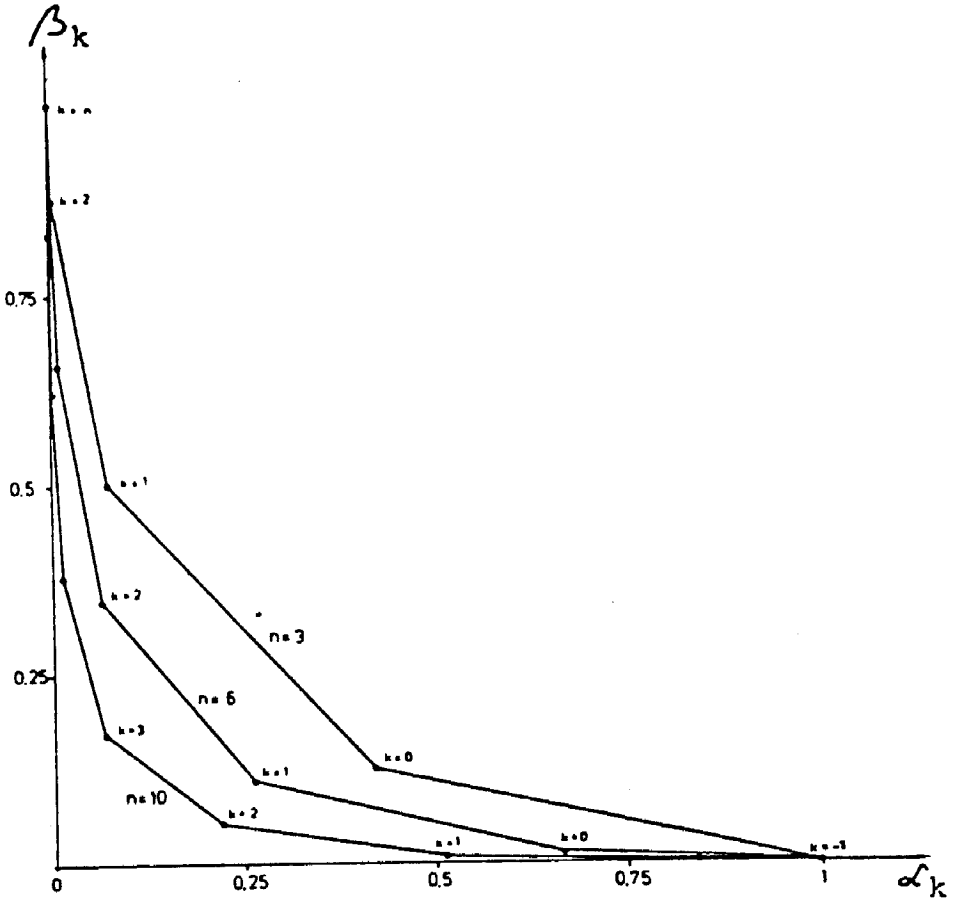
$$\beta_k = \sum_{i=0}^{k} \binom{n}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{n-i} = P(X \leq k \mid H_1)$$

Figur 2 zeigt diese Werte graphisch für n=10 und k=2.



Figur 2

Um den Zusammenhang zwischen α_k und β_k deutlicher zu machen, zeichnen wir die Paare (α_k, β_k), k=0,1,2,...,n, für verschiedene n (Anzahl der Ziehungen) in ein α,β-Koordinatensystem (Figur 3) und verbinden die entsprechenden Punkte zu einem Polygonzug. Der Wert k=-1 wird der Vollständigkeit halber hinzugenommen (H_o wird immer abgelehnt).



Figur 3

Man sieht, wie durch die Wahl von k (bei festem n) α_k und β_k variieren: was man bei dem einen gewinnt, verliert man beim anderen. Will man die Wahrscheinlichkeiten für beide Fehler verkleinern, so gelingt dies nur durch Erhöhung des Stichprobenumfanges n.

Anhand des letzten Beispiels sollen auch Möglichkeiten zur Festlegung einer Entscheidungsregel skizziert werden. Bei gegebener Schranke α für α_k sucht man einfach das kleinste k mit $\alpha_k \leq \alpha$, was gleichzeitig eine Minimierung der Wahrscheinlichkeit β_k eines Fehlers 2. Art ergibt. In vielen Fällen hat man irgendeine Art von Kosten, die durch einen Fehler 1. Art oder 2. Art verursacht werden, es seien dies die Beträge S_1 und S_2. Das können etwa die Kosten einer Neujustierung einer Maschine sein, der Verlust durch Ablehung einer Sendung, Verlust durch schlechte Ware u.a. Dann sind $S_1\alpha_k$ und $S_2\beta_k$ die Erwartungswerte des Verlusts, der

entstehenden Unkosten bei einem Fehler 1. bzw. 2. Art. Man wählt dann das k für die Entscheidungsregel E_k so, daß der größere der beiden Werte von $S_1\alpha_k$ und $S_2\beta_k$ (dieser hängt von k ab) minimal ist. In der im Beispiel modellhaft dargestellten Situation einer sogenannten einfachen Hypothese $H_o: p=p_o$ und einer einfachen Gegenhypothese $H_1: p=p_1$ hat man oft eine Wahrscheinlichkeitsverteilung für das Vorliegen von H_o bzw. H_1, die man etwa aus einer längeren Erfahrung mit diesem Test oder durch theoretische Überlegungen gewinnt. Es sei π_i die (subjektive) Wahrscheinlichkeit für H_i, i=1,2. Dann wird man k so wählen, daß das Gesamtrisiko $\pi_1 S_1\alpha_k + \pi_2 S_2\beta_k$ (das gewichtete Mittel der beiden Risiken $S_1\alpha_k$ und $S_2\beta_k$) minimal wird (Bayes-Prinzip).

I.3. Approximation durch Normalverteilung

Bei den bisher beschriebenen Tests war die Testgröße binomialverteilt nach B(n,p): Anzahl der "Erfolge" (etwa: weiße Kugel) bei n-facher unabhängiger Wiederholung des Versuchs. Für das praktische Arbeiten ist die Binomialverteilung schlecht geeignet. In vielen Fällen wird die Approximation durch die Normalverteilung ausreichend gut sein (in der Regel, sobald die Laplace-Bedingung $np(1-p) > 9$ erfüllt ist). Wir wiederholen dazu, daß der Erwartungswert von B(n,p) gleich np (für die Anzahl), gleich p (für die relativen Häufigkeiten) und die Varianz gleich np(1-p) bzw. $\frac{p(1-p)}{n}$ ist. Somit wird unter den entsprechenden Bedingungen B(n,p) durch die Normalverteilungen $N(np, np(1-p))$ bzw. $N(p, \frac{p(1-p)}{n})$ angenähert (Satz von Moivre-Laplace). Ist X die Anzahl der Erfolge, so ist demnach die standardisierte Zufallsgröße $\frac{X-np}{\sqrt{np(1-p)}}$ angenähert nach N(0,1) verteilt und daher gilt unter der Annahme, daß p der "wahre" Anteil ist:

$$P_p\left(a < \frac{X-np}{\sqrt{np(1-p)}} \leq b\right) \approx \Phi(b) - \Phi(a),$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist. Analog ist dies für $\frac{1}{n}X$, die relative Häufigkeit zu formulieren. Etwas anders ausgedrückt mit $\sigma = \sqrt{\frac{p(1-p)}{n}}$ und für $z > 0$ ergibt sich:

$$P_p\left(p - z\sigma < \frac{1}{n} X \leq p + z\sigma\right) \approx \Phi(z) - \Phi(-z).$$

Wir wollen diese Näherung für das Testproblem für eine Hypothese über einen unbekannten relativen Anteil (etwa: weiße Kugeln in einer Urne) verwenden. Ist etwa $H_o: p=p_o$ gegen $H_1: p\neq p_o$ zu testen und ist ein Signifikanzniveau α vorgegeben, so ermittelt man z (aus der entsprechenden Tabelle) derart, daß $\Phi(z)=1-\frac{\alpha}{2}$. Dann ist der Annahmebereich für die relative Häufigkeit $\frac{1}{n}X$ gegeben durch das Intervall

$$[p_o - z\sqrt{\frac{p_o(1-p_o)}{n}}\ ,\ p_o + z\sqrt{\frac{p_o(1-p_o)}{n}}]$$

in das $\frac{1}{n}X$ (n-fache Wiederholung des Versuchs) mit der Wahrscheinlichkeit $1-\alpha$ fällt, wenn H_o gilt; ist der Wert von $\frac{1}{n}X$ außerhalb dieses Intervalls, so wird H_o abgelehnt, sodaß $\alpha = P(\text{Fehler 1. Art})$ ist.

Etwas anders ist das nächste

Beispiel. Ein Unternehmen bezieht einen Massenartikel bei einem bestimmten Produzenten und erfahrungsgemäß ist die Ausschußquote gleich 5%. Ein Konkurrenzangebot für denselben Artikel behauptet, eine Ausschußquote unter 5% zu haben. Es wird nun vom Unternehmer die Hypothese $H_o: p \geq 0{,}05$ für das Konkurrenzangebot gegen die Gegenhypothese $H_1: p < 0{,}05$ durch zufällige und unabhängige Auswahl von 400 Stück aus einer Lieferung aus dem Konkurrenzangebot getestet. Darunter waren 9 Ausschußstücke. Unter H_o ist die Wahrscheinlichkeit für Ausschuß mindestens 0,05 und daher sind "im für H_o schlechtesten Fall" X=Anzahl der Ausschußstücke bzw. ihre relative Häufigkeit $\frac{1}{400}X$ nach B(400; 0,05) verteilt. Ist die Gesamtlieferung groß gegenüber 400, so ist es ohne wesentlichen Einfluß, ob mit oder ohne Zurücklegen gezogen wird. Nun werden nur "sehr kleine" Werte von X bzw. $\frac{1}{400}X$ zur Ablehnung von H_o führen, was als Entscheidungsregel ergibt (x_o Wert von X beim Test):

Ist $\frac{1}{400}x_o \leq 0{,}05 - z\sqrt{\frac{0{,}05 \cdot 0{,}95}{400}} = c_o$ so wird H_o abgelehnt.

Die Bestimmung von z erfolgt entsprechend der Festlegung eines Signifikanzniveaus α als obere Schranke für P(Fehler 1. Art) über die Beziehung $\Phi(z)=1-\alpha$ bzw. $\Phi(-z)=\alpha$. Denn ein Fehler

1. Art entsteht genau dann, wenn $\frac{1}{400}x_o \leq 0{,}05-z\sqrt{\frac{0{,}05.0{,}95}{400}}$ gilt, obwohl $p \geq 0{,}05$ ist. Für $\alpha=0{,}01$ ist nach der Tabelle $z=2{,}33$ und somit $c_o=0{,}0246$. In unserer Stichprobe ist $x_o=9$, $\frac{x_o}{400}=0{,}0225$, sodaß H_o abgelehnt wird, was zur Konsequenz hat, das Konkurrenzangebot anzunehmen. Dabei ist die Irrtumswahrscheinlichkeit, also die Wahrscheinlichkeit H_o abzulehnen, obwohl H_o zutrifft, höchstens gleich $\alpha=0{,}01$.

I.4. Das allgemeine Schema

An dieser Stelle wollen wir versuchen, das allgemeine Schema der Vorgangsweise herauszuarbeiten, die beim Testen einer Hypothese H_o über einen Parameter der (unbekannten) Verteilung einer Zufallsgröße Y (eines Merkmals in einer Grundgesamtheit) gegen die Alternativ- oder Gegenhypothese H_1 eingesetzt wird. Dabei nimmt man von vornherein an oder weiß dies aus der Erfahrung oder der Theorie, daß die Verteilung von Y von einem bestimmten Typ ist (etwa B(1,p) oder eine Normalverteilung). Der Test besteht in der Durchführung eines Zufallsversuchs (Ziehen einer Stichprobe, Ziehen von Elementen mit Zurücklegen). Aus dem Zufallsversuch bzw. seinen n-fachen Wiederholungen gewinnt man eine neue Zufallsgröße X, die sogenannte Testgröße, deren Verteilung unter H_o bzw. unter anderen Annahmen über die Verteilung von Y aus dieser und der Art des Zufallsversuches bestimmbar ist. Insbesondere gilt für die Erwartungswerte $E(X)=E(Y)$. Bei jeder Durchführung des Tests ergibt sich ein bestimmter Wert x von X. Wenn der erhaltene Wert unter der Hypothese H_o "sehr unwahrscheinlich" ist, wird H_o abgelehnt, d.h. eine Aktion gesetzt, als ob H_1 zuträfe. Das "sehr unwahrscheinlich" dieser Entscheidungsregel wird dadurch konkretisiert, daß man Werte als sehr unwahrscheinlich auffaßt, die "weit genug" vom Erwartungswert von X (unter der Hypothese H_o) entfernt sind. Das "weit genug" schließlich wird dadurch festgelegt, daß die Menge der x, die zur Ablehnung von H_o führen (der Ablehnungsbereich) unter der Verteilung von X höchstens eine vorgegebene Wahrscheinlichkeit α hat (Signifikanzniveau). Dieser Ablehnungsbereich hat je nach der Formulierung von H_o und H_1 somit die Form $\{x \mid |x-E(X)| > r\}$ oder $\{x \mid x-E(X) > r\}$ oder $\{x \mid E(X)-x > r\}$. Man spricht im ersten Fall von einem zweiseitigen Test, andernfalls von einem einseitigen Test.

I.5. Test des Mittelwertes (Normalverteilung)

Bisher waren die Verteilungen (von Y und X) immer Binomialverteilungen (die zwar meist für X durch eine Normalverteilung approximiert werden können). Im nächsten Beispiel wird Y eine Normalverteilung besitzen, wie dies (angenähert) für viele Größen zutrifft (Länge, Stärke von Werkstücken, Füllmengen, physikalische Meßgrößen als Ergebnis von unabhängigen Messungen). Für die Konstruktion von Tests für Hypothesen über den Erwartungswert von Y benötigen wir den folgenden Satz:

Satz. Sind $X_1, X_2, \ldots, X_n$ n unabhängige nach $N(\mu,\sigma^2)$ verteilte Zufallsgrößen, so ist $\overline{X}=\frac{1}{n}(X_1+\ldots+X_n)$ nach $N(\mu, \frac{\sigma^2}{n})$ verteilt.

Anschaulich interpretiert ist das folgender Sachverhalt. Man hat eine in der Grundgesamtheit normalverteilte Größe Y mit Erwartungswert μ und Varianz σ^2. Man zieht eine zufällige Stichprobe vom Umfang n und erhält die Werte $x_1, x_2, \ldots, x_n$ und damit den (arithmetischen) Mittelwert $\overline{x}=\frac{1}{n}(x_1+x_2+\ldots+x_n)$. Denkt man sich alle möglichen zufälligen Stichproben gezogen, so ist $\overline{x}$ auffaßbar als der Wert einer Zufallsgröße $\overline{X}$ die man den Stichprobenmittelwert nennt. Offenbar ist $\overline{X}=\frac{1}{n}(X_1+\ldots+X_n)$, wo X_i wie Y verteilt ist und der i-ten Messung bei der Stichprobe entspricht. Dann sagt der Satz, daß der Erwartungswert des Stichprobenmittelwerts gleich $\mu=E(Y)$ und seine Varianz gleich $\frac{\sigma^2}{n}$ ist, $\sigma^2=V(Y)$. Damit ist der Stichprobenmittelwert bestens geeignet als Testgröße für Hypothesen über den Erwartungswert von Y.

Anmerkung. Die Aussage des Satzes über die Verteilung von X gilt (angenähert) auch dann, wenn die X_i nicht normalverteilt sind, aber wenigstens dieselbe Varianz haben und n groß ist (also bei oftmaliger Wiederholung des Versuchs, bei großen Stichproben). Der früher erwähnte Satz von Moivre-Laplace ist ein Spezialfall, wenn man bedenkt, daß die relative Häufigkeit als Stichprobenmittelwert aufgefaßt werden kann.

Beispiel. Die Füllmenge einer Flaschenabfüllanlage muß mindestens 1000 cm³ betragen. Anhand der Inhalte von 50 zufällig ausgewählten Flaschen soll entschieden werden, ob die Anlage neu eingestellt werden soll. Dabei weiß man aus der Erfahrung, daß

der Flascheninhalt jedenfalls eine Normalverteilung $N(\mu, 25)$ hat. Als Nullhypothese wird $H_o: \mu \leq 1000$ gewählt und als Gegenhypothese $H_1: \mu > 1000$. Wird H_o angenommen, so soll neu eingestellt werden. Als Signifikanzniveau wird $\alpha = 0{,}05$ vorgegeben, d.h. α ist die maximale Wahrscheinlichkeit dafür, daß eine nötige Neueinstellung nicht vorgenommen wird.

Die Testgröße ist nach den angestellten Überlegungen der Stichprobenmittelwert $\bar{X} = \frac{1}{50} \sum_{i=1}^{50} X_i$, wo X_i die Zufallsgröße "Inhalt der i-ten ausgewählten Flasche" ist; jedes X_i ist nach $N(\mu, 25)$ verteilt, wenn μ der "wahre" Erwartungswert der Füllmenge ist. Für $\mu = 1000$ ist somit $\bar{X}$ nach $N(1000, 1/2)$ verteilt. Zur Ablehnung von H_o werden "sehr große" Werte $\bar{x}$ von $\bar{X}$ führen. Zur Konstruktion des Ablehungsbereiches nimmt man $\mu = 1000$ an, weil derselbe Wert $\bar{x}$ von $\bar{X}$ für ein $\mu < 1000$ noch eher zur Ablehnung führt als für $\mu = 1000$. Nun ist aber α eine obere Schranke für die Wahrscheinlichkeit eines Fehlers 1. Art, sodaß der Ablehnungsbereich zu ermitteln ist aus:

$$P_{1000}(\bar{X} - E(\bar{X}) > \delta) = \alpha \quad \text{oder} \quad P_{1000}(\bar{X} > k) = \alpha$$

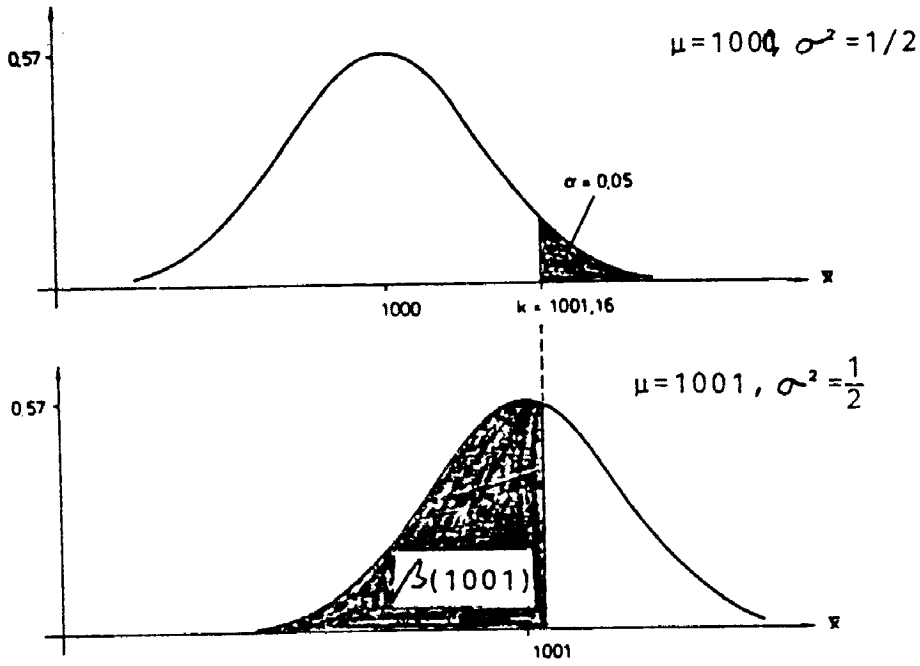
wo P_{1000} die Wahrscheinlichkeit unter der Annahme, daß $\mu = 1000$ gilt, bedeutet. Das k genügt somit der Beziehung

$$P_{1000}(\bar{X} \leq k) = P_{1000}\left(\frac{\bar{X} - 1000}{1/\sqrt{2}} \leq \frac{k - 1000}{1/\sqrt{2}}\right) = \Phi(\sqrt{2}(k - 1000)) = 1 - \alpha = 0{,}95.$$

Das ergibt mittels der Tabelle für Φ:

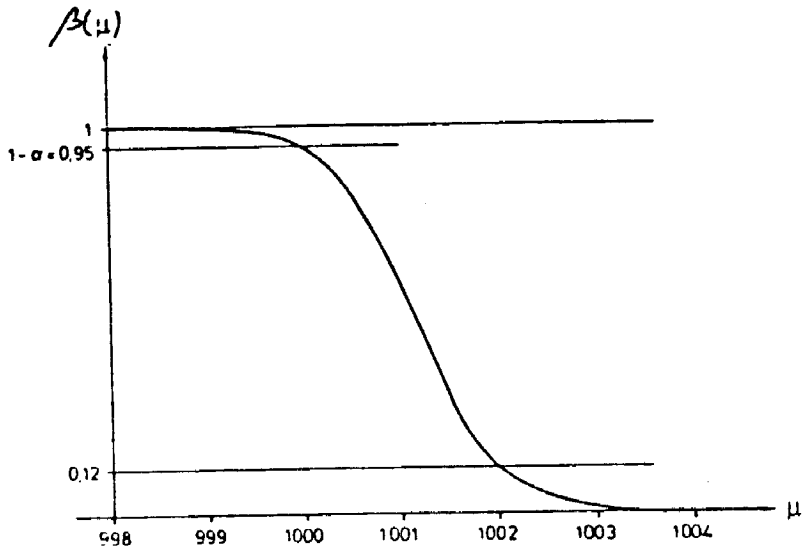
$$\sqrt{2}(k - 1000) = 1{,}645$$

und damit als kritischen Wert $k = 1001{,}16$. Der Ablehnungsbereich ist somit $\{\bar{x} \mid \bar{x} > 1001{,}16\}$ und ist aus Figur 4 zu entnehmen:



Figur 4

In dieser Figur wird auch die Wahrscheinlichkeit für einen Fehler 2. Art bei einem tatsächlichen Erwartungswert $\mu_1=1001$ gezeigt. Als Operationscharakteristik dieses Tests erhält man die in Figur 5 gezeigte Kurve. Dabei ist $\beta(\mu)=P_\mu(\bar{X} \leq 1001,16)=\Phi(\sqrt{2}(1001,16-\mu))$ die Wahrscheinlichkeit des Annahmebereichs, wenn μ der Erwartungswert der Füllmenge ist.



Figur 5

Beispiel. In der Situation des letzten Beispiels soll nun die Hypothese $H_o: \mu = 1000$ gegen die Alternativhypothese $H_1: \mu \neq 1000$ getestet werden (zweiseitiger Test). Beim Signifikanzniveau α ist dafür der Ablehnungsbereich bestimmt durch:

$$P_{1000}(|\bar{X}-1000| > k) \leq \alpha,$$

wo $\bar{X}$ wieder das Stichprobenmittel der Inhalte von 50 zufällig gewählten Flaschen ist. Daraus ergibt sich als Bedingung für k, weil $\bar{X}$ nach $N(1000, 1/2)$ und damit $\sqrt{2}(\bar{X}-1000)$ nach $N(0,1)$ verteilt ist:

$$P_{1000}\left(\frac{\bar{X}-1000}{1/\sqrt{2}} < -k\sqrt{2} \vee \frac{\bar{X}-1000}{1/\sqrt{2}} > k\sqrt{2}\right) =$$

$$\Phi(-k\sqrt{2}) + (1-\Phi(k\sqrt{2})) = \alpha$$

oder

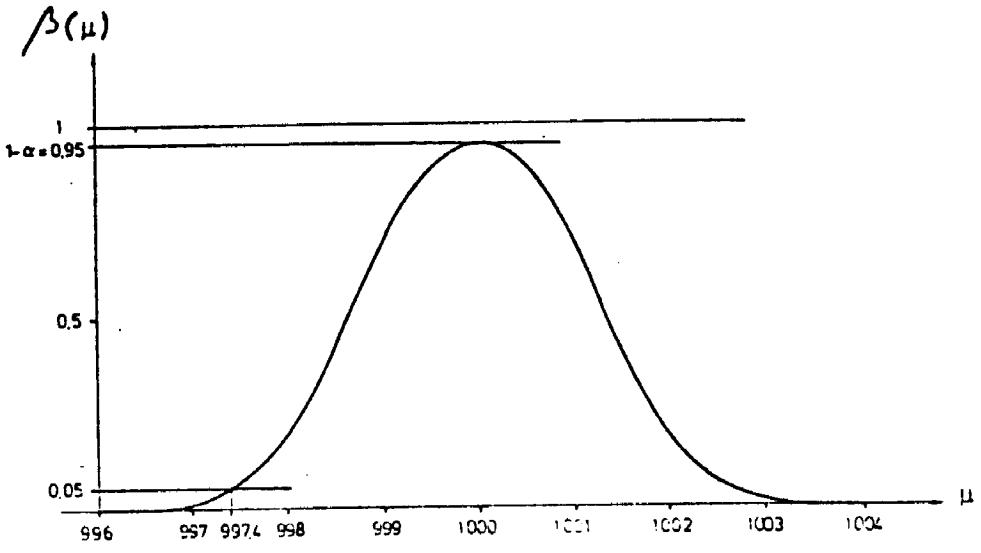
$$2\Phi(-k\sqrt{2}) = \alpha$$

$$\Phi(-k\sqrt{2}) = \alpha/2, \; -k\sqrt{2} = 1{,}96$$

Für $\alpha = 0{,}05$ ist dann $k = 1{,}386$ (nach Tabelle). Somit wird H_o verworfen, wenn $|\bar{x}-1000| > 1{,}386$ ist für den Mittelwert $\bar{x}$ der Stichprobe von 50 Flaschen. In Figur 6 ist wieder die Operationscharakteristik dieses Tests dargestellt, wobei

$$\beta(\mu) = P_\mu(|\bar{X}-1000| \leq 1{,}386) \text{ und speziell } \beta(1000) = 1-\alpha$$

die Wahrscheinlichkeit für einen Fehler 2. Art ist, wenn μ der "wahre" Erwartungswert der Füllmenge ist.



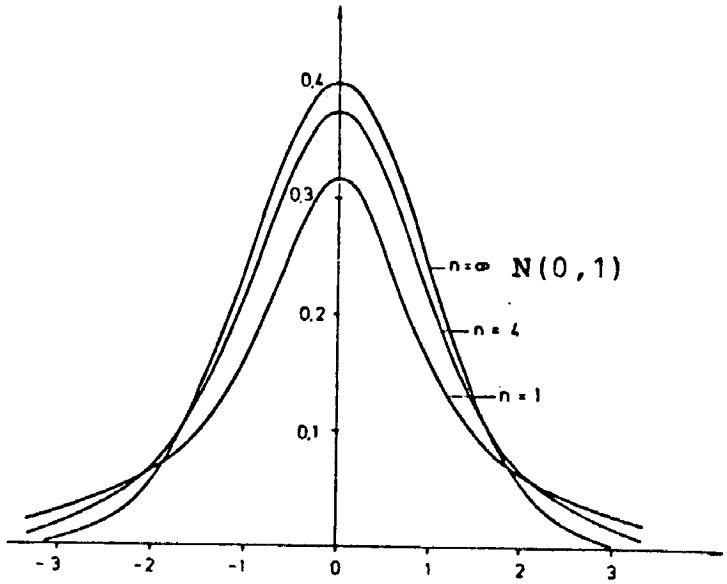
Figur 6

I.6. Test des Mittelwerts, t-Verteilung

Ergänzung. In den beiden letzten Beispielen hatten wir angenommen, daß die Varianz des untersuchten Merkmals "Füllmenge" bekannt ist. Ist dies nicht der Fall, so kann man entweder zunächst einen Schätzwert für die Varianz ermitteln (die empirische Varianz der Stichprobe) und dann so verfahren wie oben. Diese Vorgangsweise erhöht natürlich beträchtlich die Unsicherheit des Tests. Besser ist es, die folgende Testgröße W zu wählen, wo $\bar{X}$ der Stichprobenmittelwert und X_i die Zufallsgrößen der Werte der einzelnen Messungen bei einer zufälligen Stichprobe vom Umfang n sind (die X_i seien wieder identisch normalverteilt):

$$W = \frac{\sqrt{n}(\bar{X}-E(\bar{X}))}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-E(\bar{X}))^2\right]^{1/2}} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

Die Verteilung von W kann theoretisch ermittelt werden und ist eine t-Verteilung mit (n-1) Freiheitsgraden (auch: Student-Verteilung). Für einige Werte der Anzahl der Freiheitsgrade zeigt Figur 7 die Dichtefunktionen der t-Verteilungen und zum Vergleich auch die der N(0,1)-Verteilung, gegen die für $n \to \infty$ die t-Verteilungen konvergieren.



Figur 7

Bei der Ermittlung des Ablehnungsbereiches geht man nun vollkommen analog vor wie oben, es wird nur die Normalverteilung durch die entsprechende t-Verteilung ersetzt, deren q-Quantile $t_{n,q}$ tabelliert sind. Ein symmetrischer Annahmebereich ist etwa bestimmbar aus:

$$P_{\mu_o}(|W| > k) = \alpha \Leftrightarrow k = t_{n-1;\ 1-\alpha/2}$$

wobei $\mu_o = E(\bar{X})$ ist unter der Hypothese $H_o: \mu = \mu_o$. Anders ausgedrückt ergeben sich für die Annahmebereiche bei ein- und zweiseitigen Tests die Bedingungen:

$$P_{\mu_o}(\mu_o - ks/\sqrt{n} \leq \bar{X} \leq \mu_o + ks/\sqrt{n}) = \alpha \Leftrightarrow k = t_{n-1;1-\alpha/2}$$

$$P_{\mu_o}(\bar{X} > \mu_o + ks/\sqrt{n}) = \alpha \Leftrightarrow k = t_{n-1;1-\alpha}$$

$$P_{\mu_o}(\bar{X} < \mu_o - ks/\sqrt{n}) = \alpha \Leftrightarrow k = t_{n-1;1-\alpha}$$

wobei s wie oben ist:

$$s = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu_o)^2\right]^{1/2}$$

II. Verteilungs- oder Anpassungstests

Im vorhergehenden Abschnitt wurden Testverfahren entwickelt, mit denen Hypothesen über die Parameter (Mittelwert, relativer Anteil, Standardabweichung) einer unbekannten Verteilung einer Zufallsgröße geprüft werden konnten. Dabei wird meist vorausgesetzt, daß man zumindest den Typ (Normalverteilung, Binomialverteilung usf.) der unbekannten Verteilung kennt. Daraus ergibt sich die für die Begründung und legitime Anwendung des Tests erforderliche Kenntnis der Stichprobenverteilung des jeweiligen Parameters (unter der Annahme einer zufälligen Stichprobenauswahl). In diesem Abschnitt werden Tests zur Überprüfung von Hypothesen über Wahrscheinlichkeitsverteilungen "im Ganzen" vorgestellt. Solche Tests heißen Verteilungs- oder Anpassungstests und werden etwa dann verwendet, wenn überprüft werden soll, ob zwei Zufallsgrößen (mit denselben Werten) signifikant verschiedene Verteilungen haben oder nicht. Wir beschränken uns hier auf das Wesentliche; wieder gibt es eine Vielzahl von Tests, die in den einschlägigen Handbüchern beschrieben werden, sh. Literaturliste.

Beispiel. Um zu überprüfen, ob ein Würfel fair ist, wird 60 mal gewürfelt und dabei ergeben sich folgende absolute Häufigkeiten H_i für die Augenzahl i (i=1,..., 6)

i	1	2	3	4	5	6
H_i	7	8	13	8	9	15

Ist dieses Ergebnis mit der Hypothese "fairer Würfel" verträglich, für den ja die Wahrscheinlichkeit für alle i gleich 1/6 ist? Die theoretisch zu erwartenden absoluten Häufigkeiten für den fairen Würfel wären dann jeweils 10.

Um eine Hypothese dieser Art testen zu können, muß man sich eine Testgröße als Funktion von Zufallsstichproben verschaffen, für die man durch theoretische Überlegungen die Verteilung in der Grundgesamtheit aller Zufallsstichproben (im Beispiel sind das die Würfelserien der Länge 10) bestimmen kann. Sind allgemein nun $p_1, p_2, \ldots, p_m$ die hypothetisch angenommenen

Wahrscheinlichkeiten für die Werte einer (diskreten) Zufallsgröße X und ist n der Stichprobenumfang (man führt n unabhängige Realisierungen von X durch), so sind np_i, i=1,..., m die zu erwartenden Häufigkeiten für das Auftreten der einzelnen Werte von X. In einer konkreten Testserie werden nun die Häufigkeiten n_i, i=1,..., m, für die Werte von X beobachtet. Man bildet dann folgende Testgröße χ^2 (chi-Quadrat).

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$$

In sie gehen die Quadrate der Abweichungen $n_i - np_i$ der tatsächlich beobachteten von den hypothetischen Werten ein, die noch bezogen werden auf die letzteren. Dann gilt (ohne Beweis):

Satz. Hat eine diskrete Zufallsgröße X eine Verteilung mit Wahrscheinlichkeiten p_i für die Werte i=1,...,m und werden in einer Stichprobe vom Umfang n die Häufigkeiten n_i beobachtet, so hat für große n die Zufallsvariable χ^2 (definiert auf den Zufallsstichproben des Umfangs n) mit

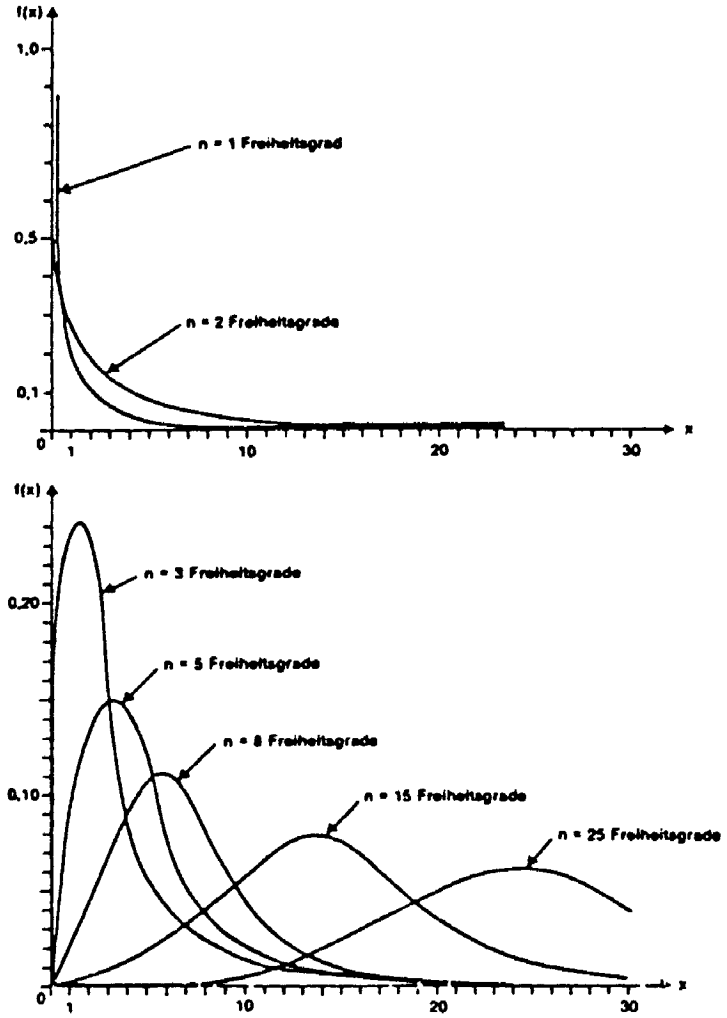
$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$$

angenähert eine χ^2-Verteilung mit m-1 Freiheitsgraden (m-Anzahl der verschiedenen Werte von X).

Ergänzung. Die χ^2-Verteilung mit r Freiheitsgraden ist definiert als die Verteilung der Zufallsgröße Y mit

$$Y = X_1^2 + X_2^2 + \ldots + X_r^2$$

wo die X_i unabhängige Zufallsgrößen sind, die alle standardnormalverteilt sind. Die Figur 8 zeigt die Dichtefunktion der χ^2-Verteilung für einige Werte von r.



Figur 8

Es gilt dann: Erwartungswert $\mu(Y) = r$ und Varianz $\sigma^2(Y) = 2r$. Die Werte der Verteilungsfunktion der χ^2-Verteilung sind für verschiedene $r \geq 1$ tabelliert; d.h. in den Tabellen werden für Werte p (der Wahrscheinlichkeit) die zugehörigen x angegeben, für die gilt $P(\chi^2 \leq x) = p$ (p-Quantile der χ^2-Verteilung).

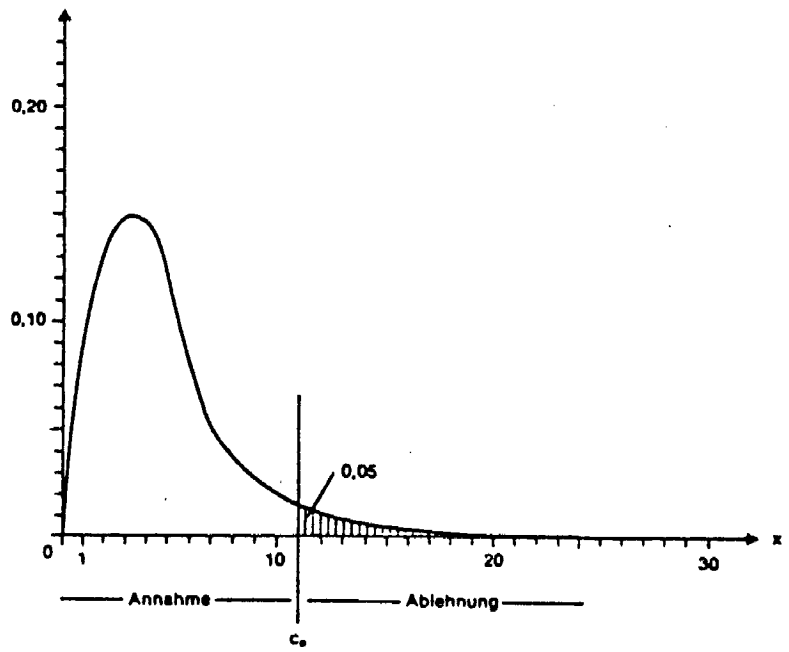
Unter der Hypothese H_o, daß X die Verteilung $\{p_1, p_2, \ldots, p_m\}$ hat, sind "große" Werte von χ^2 unwahrscheinlich. Ist daher c_o bei gegebenem α so bestimmt (aus der entsprechenden Tabelle), daß $P(\chi^2 > c_o) = \alpha$, $(P(\chi^2 \leq c_o) = 1 - \alpha)$ gilt, so wird unter der aufgestellten

Hypothese die Testgröße χ^2 mit Wahrscheinlichkeit α größer als c_o sein. Als Testregel legt man daher fest:

Ist $\chi^2 > c_o$, so wird die Hypothese H_o abgelehnt, andernfalls angenommen.

Es heißt auch hier α Signifikanzzahl, $\{x \mid 0 \leq x \leq c_o\}$ Annahmebereich und $\{x \mid x > c_o\}$ Verwerfungs- oder Ablehnungsbereich. Auch die Rolle von α ist dieselbe wie früher: es ist α die Wahrscheinlichkeit dafür, daß die Hypothese H_o abgelehnt wird, obwohl X die Verteilung $\{p_1, p_2, \ldots, p_m\}$ besitzt (Irrtumswahrscheinlichkeit).

Fortsetzung des Beispiels. Es werde $\alpha = 0{,}05$ gewählt. Es ist nach dem Satz die χ^2-Verteilung mit 6-1=5 Freiheitsgraden zu verwenden und aus der Tabelle ergibt sich $c_o = 11{,}07$, d.h.
$P(\chi^2 \leq 11{,}07) = 0{,}95$ oder $P(\chi^2 > 11{,}07) = 0{,}05$. Den Wert von χ^2 für die angeführte Stichprobe ermittelt man zu $\chi^2 = 5{,}20$, sodaß die Hypothese "fairer Würfel" angenommen werden kann. Die Figur 9 zeigt für dieses Beispiel den Annahme- und Ablehnungsbereich.



Figur 9

Anmerkungen. Für die Anwendung des oben beschriebenen χ^2-Verteilungstests sollen die Werte np_i nicht kleiner als 5 und n soll größer als 30 sein (damit der angeführte Satz gilt). Für m=2 (nur ein Freiheitsgrad) sollte man die sogenannte Yates-Korrektur anbringen: man ersetzt n_i-np_i durch $|n_i-np_i| -0,5$.

Das folgende Beispiel zeigt, wie der χ^2-Verteilungstest auch für stetige Zufallsgrößen unter Verwendung einer geeigneten Klasseneinteilung verwendet werden kann.

Beispiel. Es soll getestet werden, ob der Mietzins X (angenähert) normalverteilt ist. Dazu wurden n=393 Vierpersonenhaushalte mit mittlerem Einkommen befragt. Es ergab sich folgende Häufigkeitsverteilung:

Miete (DM) von ... bis unter	abs. Häufigkeiten n_i in der Stichprobe	hypothetische Häufigkeiten p_i	np_i
<200	59	0,1230	48,34
200-250	83	0,1716	67,44
250-300	67	0,2413	94,83
300-350	81	0,2252	88,50
350-400	58	0,1488	58,48
400	45	0,00901	35,41

Als Mittelwert der Stichprobe ergab sich $\bar{x}$=293 DM, als Standardabweichung sei 80 DM angenommen. Zunächst lautet dann die "vernünftigste Hypothese", daß X nach N(293; 80) normalverteilt ist. Aus dieser Annahme lassen sich für jede Klasse obiger Klasseneinteilung die Wahrscheinlichkeiten p_i ermitteln, mit denen X in diese Klasse fällt; diese sind in der Tabelle angeführt. Aus diesen ergeben sich wieder die erwarteten Häufigkeiten np_i. Nun kann man die Klassen als die 6 Werte einer diskreten Zufallsgröße ansehen und den obigen χ^2-Verteilungstest anwenden. Allerdings ist dabei die Regel zu beachten, daß für jeden geschätzten Parameterwert der hypothetischen Verteilung (hier μ und σ) die Anzahl der Freiheitsgrade der zu verwendenden χ^2-Verteilung um 1 zu vermindern ist. Daher ergeben sich hier nicht 6-1=5 sondern nur 3 Freiheitsgrade. Aus der entsprechenden Tabelle entnimmt man für α=0,05 den kritischen Wert c_o=7,81.

Der Wert von χ^2 ergibt sich als:

$$\chi^2 = \sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i} = 17{,}35 > c_o$$

Daher ist die Hypothese "X ist normalverteilt" abzulehnen.

Ergänzung. Auch auf der χ^2-Verteilung baut ein Test für die Unabhängigkeit zweier Merkmale X und Y auf. Er kann hier nur kurz skizziert werden. Die Häufigkeitsverteilung einer Stichprobe vom Umfang n des zweidimensionalen Merkmals (X,Y) sei in einer Kontingenztafel mit Werten n_{ij} für die Häufigkeit der Ausprägung (x_i, y_j), $i=1,\ldots,r$, $j=1,\ldots,s$, vorgelegt. Im Falle der idealen Unabhängigkeit von X und Y würde gelten $n_{ij}=n_{i.}\,n_{.j}/n$, wobei $n_{i.}$, $n_{.j}$ die Randhäufigkeiten sind:

$$n_{i.} = \sum_{j=1}^{s} n_{ij} \quad \text{und} \quad n_{.j} = \sum_{i=1}^{r} n_{ij}.$$

Es ist also $n_{i.}$ bzw. $n_{.j}$ die Häufigkeit von x_i bzw. y_j. Es gilt nun für die Testgröße

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_{i.}\,n_{.j}/n)^2}{n_{i.}\,n_{.j}/n}$$

daß sie für große n angenähert eine χ^2-Verteilung mit $(r-1)(s-1)$ Freiheitsgraden besitzt. Man wird daher die Hypothese der Unabhängigkeit von X und Y ablehnen, wenn $\chi^2 > c_o$ gilt, wo c_o bestimmt wird aus $P(\chi^2 > c_o) = \alpha$ mit vorgegebener Irrtumswahrscheinlichkeit α (Wahrscheinlichkeit P zur entsprechenden χ^2-Verteilung).

III. Zeichentest und Rangtest

In den beiden vorangehenden Abschnitten wurde zur Annahme oder Ablehnung einer Hypothese immer die Kenntnis der (theoretischen) Verteilung einer entsprechend konstruierten Testgröße verwendet, wobei oft noch ein entsprechend großer Stichprobenumfang vorausgesetzt werden mußte. Es gibt aber Situationen,wo eine solche Vorgangsweise nicht möglich ist, weil die Verwendung bekannter Stichprobenverteilungen unzulässig ist (etwa: zu kleine Stichproben). Man muß dann auf sogenannte verteilungsunabhängige oder nichtparametrische Tests ausweichen, von denen wir hier zwei einfache Spezialfälle besprechen: Vorzeichentest und Rangtest. Diese Tests sind vor allem anwendbar, wenn es sich nicht um quantitative Merkmale, sondern um Rangmerkmale handelt: Qualitätsstufen, Noten, Intelligenzquotienten u.ä. Bei diesen Merkmalen kommt es im wesentlichen nur auf die Reihenfolge der Werte an, sodaß auch in den Tests nur die "kleiner-Beziehung" zwischen den Werten verwendet wird. Die Tests sind aber auch für quantitative Merkmale einsetzbar, wenn eben andere Tests versagen.

Beispiel zum Vorzeichentest. Es werden zwei Reifenprofile A und B auf die Länge des Bremsweges getestet und zwar an fünf PKW's von je vier Testfahrern, sodaß je 20 Werte des Bremsweges vorliegen: $x_1, x_2, \ldots, x_{20}$ für A und $y_1, y_2, \ldots, y_{20}$ für B. Man kann diese Werte als Realisierungen von Zufallsgrößen $X_1, X_2, \ldots, X_{20}$ und $Y_1, Y_2, \ldots, Y_{20}$ ansehen, wobei die zweidimensionalen Zufallsgrößen (X_i, Y_i), $i=1,\ldots,20$, als unabhängig angesehen werden können. Dagegen sind X_i und Y_i für jedes i sicher nicht unabhängig (gleicher Fahrer, gleiches Auto!). Die Differenzen $D_i=X_i-Y_i$ kann man dagegen wieder als unabhängige Zufallsgrößen ansehen, die nur vom Profil bestimmt sind (Einfluß von Auto und Fahrer hebt sich in der Differenz weg!). Daher kann man für alle D_i auch dieselbe Verteilung annehmen mit Verteilungsfunktion F. Nun machen wir die Hypothese, daß die Profile gleichwertig sind in ihrer Bremswirkung. Dann muß $-D_i=Y_i-X_i$ dieselbe Verteilung haben wie D_i, also auch dieselbe

Verteilungsfunktion F. Daraus folgt aber $F(0)=1-F(0)$ und daher $F(0)=1/2$, d.h. die Wahrscheinlichkeit dafür, daß $D_i > 0$ bzw. $D_i < 0$ ist, ist unter unserer Hypothese gleich 1/2. M.a.W.: Unter der Hypothese der Gleichwertigkeit ist die Wahrscheinlichkeit für eine positive (bzw. negative) Differenz X_i-Y_i gleich 1/2. Da die Differenzen unabhängig sind, ist daher die Anzahl V der positiven (negativen) Differenzen binomialverteilt nach B(20, 1/2) und man kann diese Anzahl V als Testgröße verwenden. Für unter der Hypothese sehr unwahrscheinliche Werte dieser Testgröße wird man dann die Hypothese ablehnen. Dazu gibt man wieder ein Signifikanzniveau α vor und bestimmt aus den Tabellen der Binomialverteilung den größten Wert k mit der Eigenschaft (hier ist n=20 die Anzahl der Meßdatenpaare):

$$P(V \leq n-k) > 1-\alpha/2 \qquad P(V) > n-k) \leq \alpha/2$$

Dann ist $P(V<k \text{ oder } V>n-k) \leq \alpha$, weil B(n, 1/2) symmetrisch ist; der Erwartungswert von V ist unter B(n, 1/2) gleich n/2 (gleichbedeutend mit: gleich viele positive wie negative Vorzeichen), sodaß man sehr kleine und sehr große Werte von V als unwahrscheinlich ansehen muß, wenn die Hypothese gilt. Damit ergibt sich als Entscheidungsregel : Ist $V<k$ oder $V>n-k$ für dieses k, so wird man die Hypothese der Gleichwertigkeit ablehnen. Ist $k \leq V \leq n-k$, so ist die Hypothese nicht abzulehnen.

In unserem Beispiel mögen nun folgende Daten vorliegen:

i	x_i	y_i	x_i-y_i	Vorzeichen
1	44,5	44,7	-0,2	-
2	55,0	54,8	0,2	+
3	52,5	55,6	-3,1	-
4	50,2	55,2	-5,0	-
5	45,3	45,6	-0,3	-
6	46,1	47,7	-1,6	-
7	52,1	53,0	-0,9	-
8	50,1	49,9	0,2	+
9	50,7	52,3	-1,6	-
10	49,2	50,7	-1,5	-
11	47,3	46,1	1,2	+
12	50,1	52,3	-2,2	-
13	51,6	53,9	-2,3	-
14	48,7	47,1	1,6	+
15	54,2	57,2	-3,0	-
16	46,1	52,7	-6,6	-
17	49,9	48,0	1,9	+
18	52,3	54,9	-2,6	-
19	48,7	51,4	-2,7	-
20	56,1	56,9	-0,8	-

Als Irrtumswahrscheinlichkeit wählen wir $\alpha = 0{,}05$, sodaß sich aus der Tabelle für die Binomialverteilung $k=6$ ergibt, weil $P(V \leq 13) = 0{,}9423$, $P(V \leq 14) = 0{,}9793$ und $P(V \leq 15) = 0{,}9941$ gilt. Damit ist die Hypothese abzulehnen, wenn $V > 14$ oder $V < 6$ ist. Bei uns ist $V=5$ (V=Anzahl der positiven Differenzen bzw. Vorzeichen), sodaß die Hypothese der Gleichwertigkeit der Profile abgelehnt werden muß. Da mehr negative Differenzen vorliegen, d.h. der Bremsweg bei Profil B öfter länger war als der bei A, wird man annehmen, daß das Profil A bessere Bremseigenschaften hat.

Nachdem im Beispiel die Grundidee des Vorzeichentests erläutert wurde, soll noch erwähnt werden, daß man für größere Stichprobenumfänge wieder die Binomialverteilung $B(n,p)$ durch die entsprechende Normalverteilung $N(np,\ np(1-p))$ ersetzen wird. Den kritischen Wert k erhält man dann aus der tabellierten Verteilungsfunktion Φ von $N(0,1)$ über die standardisierte Größe $(V-n/2)/\sqrt{n/4}$ als den kleinsten Wert k mit:

$$\Phi\left(\frac{k-n/2}{\sqrt{n/4}}\right) \leq 1-\alpha/2$$

oder

$$\Phi\left(\frac{2k-n}{\sqrt{n}}\right) \leq 1-\alpha/2.$$

M.a.W.: Es ist $\frac{2k-n}{\sqrt{n}}$ das $(1-\alpha/2)$-Quantil $z_{1-\alpha/2}$ der $N(0,1)$-Verteilung und somit $k=(\sqrt{n}\, z_{1-\alpha/2}+n)/2$. Bekanntlich ist für $\alpha = 0{,}05$ $z_{0,975} \approx 1{,}96 \approx 2$, sodaß für dieses α der kritische Wert k sich ergibt zu $k=(2\sqrt{n}+n)/2 = \sqrt{n}+n/2$. Somit haben wir die einfache Regel:

Auf dem Signifikanzniveau von $\alpha=0{,}05$ ist die Hypothese der Gleichwertigkeit bei n Paaren (x_i, y_i), n groß, abzulehnen, wenn $V > n/2+\sqrt{n}$ oder $V < n/2-\sqrt{n}$ gilt; V=Anzahl der positiven Vorzeichen.

Beispiel. Eine andere Anwendungssituation wäre die folgende: Es werden n Personen mit Schlafstörungen zwei Medikamente A und B zum Ausprobieren gegeben und jede Person soll angeben, welches Medikament wirksamer war. Als Vorzeichen vereinbart man +, wenn A wirksamer als B angegeben wird, andernfalls -. Mit dem Vorzeichentest

wird dann die Hypothese "A und B sind gleichwertig" getestet. Dabei ist unter dieser Hypothese wieder $P(+)=P(-)=1/2$ und somit sind die obigen Überlegungen unverändert anwendbar.

Es ist klar, daß im Falle quantitativer Größen durch den Vorzeichentest nur ein sehr geringer Teil der verfügbaren Information ausgenutzt wird und daher der Test auch nicht sonderlich scharf ist.

Anmerkung. Für den Vorzeichentest in der vorgestellten Form ist es wichtig, daß die Differenzen $D_i=X_i-Y_i$ bzw. die jeweiligen "Vorzeichen", die angeben, ob X_i "größer" als Y_i oder "kleiner" als Y_i ist (im Beispiel war "größer"="wirksamer") als unabhängige Zufallsgrößen angesehen werden können. Dafür ist erforderlich, daß darauf, ob "$X_i < Y_i$" oder "$X_i > Y_i$" ist, (in Abhängigkeit von der Bedeutung von "$<$") nur die jeweils zu untersuchende Größe (Profil, Medikament) Einfluß hat. Das kann man durch Untersuchung an denselben Beobachtungseinheiten (selbes Auto und selber Fahrer, dieselbe Person) oder durch Eliminierung anderer Einflüsse erreichen. Will man etwa zwei Lehrmethoden A und B an Schülern testen, so wird man zwei Versuchsgruppen von Schülern $x_1,\ldots,x_n$ (für Methode A) und $y_1,\ldots,y_n$ (für Methode B) so zusammenstellen, daß x_i und y_i weitestgehend im entsprechenden Fach bisher identische Leistungen gezeigt haben. Dazu kann man auch vorher einen adäquaten Leistungstest verwenden. Nach dem Lehrversuch wird festgestellt, wer von x_i und y_i "besser" ist - ergibt wieder die "Vorzeichen" + bzw. -. Man spricht von verbundenen oder abhängigen Stichproben.

Vorzeichen- Rangtest (Test von Wilcoxon für abhängige Stichproben). Dieser Test geht wieder von zwei abhängigen Stichproben $x_1,\ldots,x_n$ und $y_1,\ldots,y_n$ eines bestimmten Merkmals aus, bei dem Differenzen zwischen Werten sinnvoll sind und der Größe nach geordnet werden können. Ein typisches derartiges Merkmal ist die Temperatur. Man ordnet die Differenzen $d_i=x_i-y_i$ dem Betrag nach nach der Größe und verbindet so mit jedem d_i einen Rangplatz $1,2,\ldots,n$ in dieser Anordnung. Dann bildet man die Summen R_+ und R_- der Ränge der positiven bzw. negativen Differenzen d_i, sowie die Summe R aller Ränge, $R=\frac{n(n+1)}{2}$. Unter der Hypothese der "Gleichartigkeit" der x-Werte und der y-Werte ist dann R/2 der Erwartungswert für R_+ und

R_-, sodaß absolut "große" Differenzen zwischen R/2 und R_+ bzw. R_- zur Ablehnung dieser Hypothese führen werden. Den kritischen Wert c_o, d.h. die Grenze zwischen Annahmebereich und Ablehnungsbereich, entnimmt man einer Tabelle, die ihn für verschiedene Signifikanzniveaus α und Stichprobenumfänge n enthält:

n	0,01	0,05
6	-	10,5
7	-	12,0
8	18,0	14,0
9	20,5	16,5
10	24,5	19,5
11	28,0	22,0
12	32,0	25,0
13	35,5	28,5
14	39,5	31,5
15	44,0	35,0
16	48,0	38,0
17	53,5	41,5
18	57,5	45,5
19	63,0	49,0
20	67,0	53,0
21	72,5	56,5
22	77,5	60,5
23	83,0	65,0
24	89,0	69,0
25	94,5	73,5

Ist bei gegebenem α daher der Betrag der Differenz R_+-R/2 (bzw. R_--R/2) größer als das entsprechende c_o, so ist die Hypothese der Gleichwertigkeit abzulehnen.

Numerisches Beispiel. Eine Überprüfung der Thermometer in 10 Tiefkühltruhen ergab die folgenden Differenzen $d_i=x_i-y_i$ zwischen vom Thermometer in der Truhe angezeigter Temperatur x_i und wahrer Temperatur y_i. Diese sind in der folgenden Tabelle schon dem Absolutwert nach geordnet:

d_i	-0,5°	-0,5°	2,0°	-2,5°	3,5°	4,0°	5,5°	5,5°	7,5°	12,0°
Rang	1	2	3	4	5	6	7	8	9	10
Vorzeichen	-	-	+	-	+	+	+	+	+	+

Treten dabei gleiche Differenzen auf, so können diese beliebig angeordnet werden. Hier ist $R_- = 7$, $R_+ = 48$, $R = 55$, $R/2 = 27{,}5$ und $|R_+ - R/2| = 20{,}5$. Für $\alpha = 0{,}05$ und $n = 10$ ist nach der Tabelle $c_o = 19{,}5$, also ist die Hypothese "Thermometer zeigen im Mittel die richtige Temperatur" abzulehnen. Man sieht übrigens sofort, daß hier der zwar einfachere Zeichentest nicht zur Ablehnung der Hypothese führt. Der Vorzeichen-Rangtest ist also trennschärfer u.zw. deswegen, weil er mehr Information benutzt (auch die Größenordnung und nicht nur das Vorzeichen der Differenzen).

Schlußbemerkung. Auch in diesem Kapitel wurden nur die Grundzüge einiger einfacher Testverfahren herausgearbeitet. Für die schulische Ausbildung ist es empfehlenswert, einige der wichtigsten Tests gründlich und mit vollständigen Überlegungen zu ihrem Einsatzbereich und ihrer Begründung durchzuführen. Ein übertriebenes Hineinstopfen bloß rezeptartiger Verfahren bringt kein Verständnis der grundlegenden Methoden und die Rezepte werden schnell wieder vergessen. Hat der Schüler jedoch das Grundsätzliche verstanden, so kann er sich in der beruflichen Praxis schnell die notwendigen Techniken erarbeiten und sie auch verständig und adäquat einsetzen.

LITERATURHINWEISE

S - Schulbücher

ANDERSON, O./POPP, W./SCHAFFRANEK, M./STEINMETZ, D./STENGER, H.: Schätzen und Testen. Eine Einführung in die Wahrscheinlichkeitsrechnung und schließende Statistik. Berlin-Heidelberg-New York 1976.

BARTH, F./BERGOLD, H./HALLER, R.: Tabellen zur Stochastik. München 1976.

S BARTH, F./BERGOLD, H./HALLER, R.: Stochastik 1 und 2. München 1976 und 1974.

S BERG, D./HORN, R./SCHMIDT, G.: Statistik und Wahrscheinlichkeitsrechnung. G. Mathematik S-2 (Hrsg.: B. Andelfinger). Freiburg 1980.

S BORGES, R.: Statistik und Wahrscheinlichkeitsrechnung. Mathematik S-2 (Hrsg.: B. Andelfinger), Freiburg 1975.

BORTZ, J.: Statistik für Sozialwissenschaftler. Berlin-Heidelberg-New York 1977.

S ENGEL, A.: Wahrscheinlichkeitsrechnung und Statistik. Bd. 1 und Bd. 2. Stuttgart 1973 und 1976

S FEUERPFEIL, J./HEIGEL, F./VOLPERT, H.: Stochastik (Grundkurs). München 1975.

S FEUERPFEIL, J./HEIGEL, F./VOLPERT, H.: Stochastik (Grundkurs) - Lösungen. München 1977.

GOLDBERG, S.: Die Wahrscheinlichkeit. Eine Einführung in die Wahrscheinlichkeitsrechnung und Statistik, 3. Auflage. Braunschweig 1972.

S HEIGL, F./FEUERPFEIL, J.: Stochastik. München 1975.

S HEIGL, F./FEUERPFEIL, J.: Stochastik (Leistungskurs). 2. Auflage. München 1976.

S HEIGL, F./FEUERPFEIL, J.: Stochastik (Leistungskurs) - Lösungen. München 1976.

HELLER, W.-D./LINDENBERG, H./NUSKE, M./SCHRIEVER, K.H.: Studien- und Unterrichtsmaterial zur Lehrerfortbildung.
Bd. 1,2: Wahrscheinlichkeitsrechnung
Bd. 3: Beschreibende Statistik
Bd. 4: Schließende Statistik
Alles mit vollständig gelösten Aufgaben. Birkhäuser Verlag, Basel 1979

S INEICHEN, R.: Einführung in die elementare Statistik und Wahrscheinlichkeitsrechnung. 5. Auflage. Luzern-Stuttgart 1977.

S INEICHEN, R.: Elementare Beispiele zum Testen statistischer Hypothesen. 2. Auflage. Zürich 1978.

KREYSZIG, E.: Statistische Methoden und ihre Anwendungen. 6. Auflage. Göttingen 1977.

S LAUTER, J./RÜDIGER, K.: Wahrscheinlichkeitsrechung und Statistik. Mathematik Sekundarstufe II. (Hrsg,: Kuypers, W./Lauter. J.). Düsseldorf 1973.

MAIBAUM, G.: Wahrscheinlichkeitstheorie und mathematische Statistik. Berlin 1976.

PFANZAGL, J.: Allgemeine Methodenlehre der Statistik I, 5. Auflage. Berlin 1972

PFANZAGL, J.: Allgemeine Methodenlehre der Statistik II, 5. Auflage. Berlin 1978.

SCHAICH, E.: Schätz- und Testmethoden für Sozialwissenschaftler. München 1977.

STRICK, H.K.: Einführung in die beurteilende Statistik. (Materialien für die Sekundarstufe II). Hannover 1980.

WALSER, W.: Wahrscheinlichkeitsrechnung. Stuttgart 1975.

WETZEL, W.: Statistische Grundausbildung für Wirtschaftswissenschaftler, Band 2: Schließende Statistik. Berlin 1973.

WIEDLING, H.: Statistische Verfahren, Band 2: Schließende Statistik. Gernsbach 1979.