

# **Bestimmung des Stichprobenumfangs bei biomedizinischen Experimenten**

**von**

**Berthold Schneider  
Institut für Biometrie  
Medizinische Hochschule Hannover**

## 1. Grundlagen der Statistik

**Statistik** befaßt sich mit Daten, die bei wiederholter Beobachtung oder Messung festgestellt werden.

Beispiele sind:

Geburten und Sterbefälle in verschiedenen Jahren und/oder Gebieten,  
Wetterbeobachtungen an den verschiedenen Tagen und/oder Stationen,  
Reaktion von Versuchstieren auf experimentelle Maßnahmen.

Kennzeichnend für diese Daten sind:

- a) **zufällige (regellose) Schwankungen** bei wiederholter Beobachtung oder Messung
- b) **stabile Häufigkeiten (Häufigkeitsverteilung)** der Meß- oder Beobachtungswerte (Merkmalwerte) in hinreichend großen Gesamtheiten.

Gegenstand der Statistik sind nicht die einzelnen Messungen oder Beobachtungen, sondern die **Gesamtheiten** der wiederholt beobachteten oder gemessenen Daten.

Man unterscheidet:

**Stichprobe** = eine endliche Anzahl ( $n$ ) von beobachteten oder gemessenen Daten.

**Grundgesamtheit (Population, Kollektiv)** = hypothetische Gesamtheit aller Daten, die bei unendlich häufiger Wiederholung der Messung oder Beobachtung zu erwarten wären.

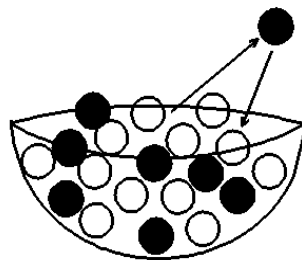
### Zusammenhang zwischen Stichprobe und Grundgesamtheit (Urnenschema):

Es wird angenommen, daß die in der Stichprobe beobachteten Daten 'zufällig' und 'unabhängig' aus der Grundgesamtheit entnommen sind.

#### Urnenschema

Alle Ergebnisse der Grundgesamtheit werden auf Kugeln geschrieben; die Kugeln werden in eine Urne geworfen und gut durchmischt.

Bei einem Ereignis wird eine Kugel zufällig entnommen und das Ergebnis notiert. Die Kugel wird dann wieder zurückgelegt (Unabhängigkeit).



**Ziel wissenschaftlicher Untersuchungen** ist es, aus den beobachteten oder gemessenen Daten einer Stichprobe, **Aussagen über die zugrunde liegende**

**Grundgesamtheit** zu machen (die Beobachtungen 'zu verallgemeinern'). Die Daten der Stichprobe werden dabei als **Realisationen** der Grundgesamtheit angesehen; sie 'realisieren' eine zufällige Auswahl der Grundgesamtheit und sind so repräsentativ für die Grundgesamtheit.

## 2. Charakterisierung von Grundgesamtheiten durch Wahrscheinlichkeiten

Die relative Häufigkeit, mit der einzelne Werte oder Mengen von Werten in der Grundgesamtheit vorkommen, ist die **Wahrscheinlichkeit**, mit der diese Werte bei Realisationen (Beobachtungen, Messungen) zu erwarten sind.

Im folgenden wird mit  $x$  ein möglicher Wert der Grundgesamtheit bezeichnet;  $X$  steht für 'alle Werte der Grundgesamtheit' (die u.U. bestimmte Bedingungen erfüllen). Man nennt  $X$  eine Zufallsgröße oder Zufallsvariable.  $\Pr(\dots)$  bedeutet 'Wahrscheinlichkeit für ...'.

Die **Wahrscheinlichkeitsverteilung  $F(x)$**  ist die Wahrscheinlichkeit, mit der Werte  $X \leq x$  in der Grundgesamtheit vorkommen:  $F(x) = \Pr(X \leq x)$ .  $F(x)$  steigt von 0 (für  $x_{\min} (-\infty)$ ) auf 1 (für  $x_{\max} (\infty)$ ) monoton an.

Die **Wahrscheinlichkeitsdichte  $f(x)$**  ist für ein kleines Intervall  $dx$  definiert als:  $f(x)dx =$  Wahrscheinlichkeit für Werte  $X$  zwischen  $x$  und  $x+dx = \Pr(x < X \leq x+dx)$ .  $f(x)$  gibt die Steilheit des Anstiegs von  $F(x)$  im Punkte  $x$  an:  $f(x) = dF(x)/dx$ .

**Parameter** von Grundgesamtheiten sind Kenngrößen, die bestimmte Eigenschaften der Grundgesamtheit (bzw. Verteilungsfunktion) charakterisieren. Beispiele für Parameter sind:

Der **Mittelwert (Erwartungswert)  $\mu$**  charakterisiert die mittlere Lage der Werte  $X$  in der Grundgesamtheit.

Die **Standardabweichung  $\sigma$**  charakterisiert die 'Streuung' der Werte in der Grundgesamtheit um den Mittelwert  $\mu$ .

### Normalverteilung

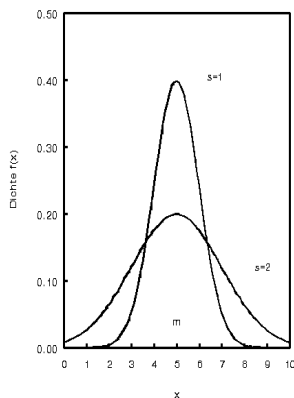
Realisationen von quantitativen Zufallsgrößen  $X$  können oft als Summen von vielen zufälligen Einflüssen angesehen werden, die den Meßwert positiv oder negativ beeinflussen. In der Grundgesamtheit kann die Verteilungsfunktion der Zufallsgröße  $X$  durch eine **Normalverteilung** mit Mittelwert  $\mu$  und Standardabweichung  $\sigma$  dargestellt werden:

$$\text{Verteilungsdichte: } f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{Verteilungsfunktion } F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{z^2}{2}} dz$$

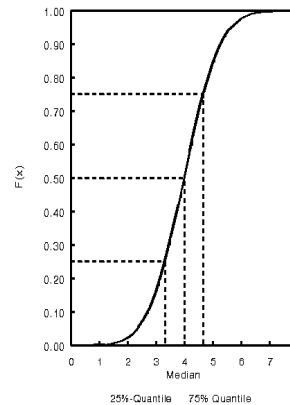
Die Wahrscheinlichkeit für Werte im Bereich

$\mu - \sigma$ bis $\mu + \sigma$	ist 68%	(einfacher Streuungsbereich)
$\mu - 2\sigma$ bis $\mu + 2\sigma$	ist 95%	(doppelter Streuungsbereich)
$\mu - 3\sigma$ bis $\mu + 3\sigma$	ist 99,4%	(dreifacher Streuungsbereich)

Dichten der Normalverteilung



Verteilungsfunktion F(x)  
Mittelwert=4, Standardabweichung=1



Die Normalverteilung mit Mittelwert 0 und Standardabweichung 1 nennt man die

**Standard-Normalverteilung**. Durch die Transformation:  $z = \frac{x - \mu}{\sigma}$  wird eine

Normalverteilung für Meßwerte  $x$  in die Standard-Normalverteilung überführt. Die Verteilungsdichte der Standard-Normalverteilung wird mit  $\varphi(z)$ , die Verteilungsfunktion mit  $\Phi(z)$  bezeichnet.  $\varphi(z)$  ist symmetrisch und  $\Phi(z)$  schief-symmetrisch um 0; d.h.  $\varphi(z) = \varphi(-z)$ ;  $\Phi(z) = 1 - \Phi(-z)$ .

Die **Quantile** einer Verteilung  $F(x)$  zur Wahrscheinlichkeit  $q$  ( $q$ -Quantile) ist der Merkmalwert  $x_q$ , für den gilt:  $F(x_q) = q$  (d.h. der Anteil  $q$  der Grundgesamtheit ist  $\leq x_q$ ). Die  $q$ -Quantile der Standard-Normalverteilung wird mit  $z_q$  bezeichnet. Wegen der Symmetrie:  $\Phi(z) = 1 - \Phi(-z)$  gilt:  $z_q = -z_{1-q}$ . Für  $q < 0.5$  ist  $z_q < 0$  und für  $q > 0.5$  ist  $z_q > 0$ .

Die Quantilen  $x_q$  der Standard-Normalverteilung

q	$z_q$	q	$z_q$
0.01	-2.326	0.5	0
0.025	-1.960	0.6	0.253
0.05	-1.645	0.7	0.524
0.1	-1.282	0.8	0.842
0.2	-0.842	0.9	1.282
0.3	-0.524	0.95	1.645
0.4	-0.253	0.975	1.960
0.5	0	0.99	2.326

### 3. Statistisches Schätzen

Ziel der statistischen Auswertung ist es, mit den Stichprobenergebnissen zuverlässige Aussagen über Parameter der Grundgesamtheit zu machen. Dies geschieht dadurch, daß aus den Stichprobenergebnissen **Schätzwerte (Statistiken)** für die interessierenden Parameter berechnet werden, die diese Parameter möglichst gut repräsentieren. Beispiele für Schätzwerte sind:

Schätzwert für die **Wahrscheinlichkeit**  $\pi$  eines Ereignisses ist die Häufigkeit  $h$ , mit der das Ereignis in der Stichprobe vorkommt:  $h = r/n$ , wenn bei  $n$  Beobachtungen das Ereignis  $r$  mal vorkommt.

Schätzwert für den **Mittelwert**  $\mu$  der Grundgesamtheit quantitativer Daten ist das arithmetische Mittel der Stichprobenwerte  $x_1, \dots, x_n$ :  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ . Schätzwert für die

**Varianz**  $\sigma^2$  der Grundgesamtheit ist die Stichprobenvarianz:  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ .

Schätzwert für die Standardabweichung  $\sigma$  ist die Wurzel  $s$ .

Da diese Schätzwerte aus den zufällig erhobenen Stichprobendaten berechnet werden, sind sie selbst Realisationen von Zufallsgrößen, die in der Gesamtheiten aller möglichen Wiederholungen der Stichprobe (Stichprobengesamtheit) zufällig variieren. Mit der Wahrscheinlichkeitsverteilung des Schätzwertes in dieser Gesamtheit wird die Genauigkeit des Schätzwertes charakterisiert. Von einem 'guten' Schätzwert ist zu fordern:

- a) **Unverzerrtheit (unbiasedness, Erwartungstreue)**: Der Mittelwert der Schätzwerte stimmt mit den zu schätzenden Parameter überein.
- b) **Effizienz**: Die Schätzwerte haben minimale Standardabweichung.

Die üblichen Schätzwerte (Maximum-Likelihood-Schätzer) sind (zumindest bei hinreichend großen Stichproben) unverzerrt und effizient. Die Genauigkeit, mit der sie den Parameter repräsentieren, hängt von der Variabilität der Daten **und vom Stichprobenumfang  $n$**  ab. Sie wird durch den **Standardfehler**, d.i. die Standardabweichung des Schätzwertes in der Stichprobengesamtheit, ausgedrückt. Der Standardfehler ist umgekehrt proportional zur Wurzel aus der Stichprobengröße  $n$ :

Schätzwert	Standardfehler
Häufigkeit $h=r/n$	$\sigma_h = \frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}$
Mittelwert $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Standardabweichung $s$	$\sigma_s \approx \frac{\sigma}{\sqrt{2n}}$ (bei Normalverteilung)

Bei Vergrößerung des Stichprobenumfangs  $n$  wird die Genauigkeit der Schätzung verbessert. Dieses **Gesetz der großen Zahl** bildet die Grundlage zur Bestimmung des Stichprobenumfangs: **Es ist für die beabsichtigte Aussage über die Grundgesamtheit die Genauigkeit vorzugeben. Der Stichprobenumfang  $n$  ist so groß zu wählen, daß die vorgegebene Genauigkeit eingehalten wird.**

### 3.1 Bestimmung des Stichprobenumfangs n bei Vorgabe des Standardfehlers

Es soll ein Parameter (z.B. Wahrscheinlichkeit, Mittelwert) geschätzt werden. Es wird die Größe des Standardfehlers vorgegeben und der Stichprobenumfang n so festgelegt, daß diese Vorgabe eingehalten wird:

**Schätzung des Mittelwertes  $\mu$  durch das arithmetische Mittel  $\bar{x}$  bei bekanntem  $\sigma$ :**

$$\text{Vorgabe: } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \Delta \quad n = \frac{1}{(\Delta / \sigma)^2} = \frac{1}{\delta^2} \quad \text{mit } \delta = \frac{\Delta}{\sigma}$$

$\Delta = \sigma/2$	$\delta = 0.50$	$n = 4$
$\Delta = \sigma/4$	$\delta = 0.25$	$n = 16$
$\Delta = \sigma/10$	$\delta = 0.10$	$n = 100$
$\Delta = \sigma/20$	$\delta = 0.05$	$n = 400$
$\Delta = \sigma/100$	$\delta = 0.01$	$n = 10000$

**Schätzung einer Wahrscheinlichkeit  $\pi$  durch die Häufigkeit h:**

$$\text{Vorgabe: } \sigma_h = \sqrt{\frac{\pi(1-\pi)}{n}} = \Delta \quad n = \frac{\pi(1-\pi)}{\Delta^2} = \frac{1}{\delta^2} \quad \text{mit } \delta = \frac{\Delta}{\sqrt{\pi(1-\pi)}}$$

$\pi = 0.5$	$\Delta = 0.1$	$\delta = 0.2$	$n = 25$
$\pi = 0.5$	$\Delta = 0.05$	$\delta = 0.1$	$n = 100$
$\pi = 0.5$	$\Delta = 0.01$	$\delta = 0.02$	$n = 2500$
$\pi = 0.9$	$\Delta = 0.1$	$\delta = 0.333$	$n = 9$
$\pi = 0.9$	$\Delta = 0.05$	$\delta = 0.167$	$n = 36$
$\pi = 0.9$	$\Delta = 0.01$	$\delta = 0.033$	$n = 900$

**Vorgabe des relativen Standardfehlers (Variationskoeffizient) der Häufigkeit h:**

$$\text{Vorgabe: } \frac{\sigma_h}{\pi} = \sqrt{\frac{1-\pi}{\pi \cdot n}} = v \quad n = \frac{1-\pi}{\pi \cdot v^2} = \frac{1}{\delta^2} \quad \text{mit } \delta = v \sqrt{\frac{\pi}{1-\pi}}$$

$\pi = 0.5$	$v = 0.1$	$\delta = 0.1$	$n = 100$
$\pi = 0.5$	$v = 0.05$	$\delta = 0.05$	$n = 400$
$\pi = 0.5$	$v = 0.01$	$\delta = 0.01$	$n = 10000$
$\pi = 0.9$	$v = 0.1$	$\delta = 0.3$	$n = 12$ (gerundet)
$\pi = 0.9$	$v = 0.05$	$\delta = 0.15$	$n = 45$ (gerundet)
$\pi = 0.9$	$v = 0.01$	$\delta = 0.03$	$n = 1112$ (gerundet)
$\pi = 0.1$	$v = 0.1$	$\delta = 0.033$	$n = 900$
$\pi = 0.1$	$v = 0.05$	$\delta = 0.017$	$n = 3600$
$\pi = 0.1$	$v = 0.01$	$\delta = 0.003$	$n = 90000$

### 3.2 Abschätzung der Wahrscheinlichkeit für seltene Ereignisse

Um die Wahrscheinlichkeit für seltene Ereignisse aus Experimenten oder Beobachtungen mit ausreichender relativer Genauigkeit abzuschätzen, sind sehr große Stichprobenumfänge erforderlich; z.B. ist zur Schätzung einer Wahrscheinlichkeit von 10% mit einer relativen Genauigkeit von 10% der Stichprobenumfang  $n=900$  erforderlich. Eine Möglichkeit, mit geringeren Stichprobenumfängen zu einer Aussage zu kommen, bietet das folgende Vorgehen: Es wird kein genauer Schätzwert für die Wahrscheinlichkeit  $\pi$  gesucht, sondern eine obere Grenze  $\pi_0$ , von der behauptet werden kann, daß sie mit einer vorgegebenen Zuverlässigkeit (Konfidenz; z.B. 95%) von der 'wahren' Wahrscheinlichkeit nicht überschritten wird. Da das Ereignis selten (d.h.  $\pi$  sehr klein) ist, ist bei kleinen Stichproben nicht zu erwarten, daß das Ereignis überhaupt beobachtet wird. Der Stichprobenumfang  $n$  wird nun so groß gewählt, daß für ein gegebenes  $\pi_0$  mit vorgegebener Zuverlässigkeit (Konfidenz) von z.B. 95% die Aussage, daß die Wahrscheinlichkeit  $\pi$  nicht größer als  $\pi_0$  ist, getroffen wird, wenn in der Stichprobe das Ereignis nicht beobachtet wird. Das erforderliche  $n$  kann durch folgende Überlegung ermittelt werden: Bei einer Ereigniswahrscheinlichkeit  $\pi_0$  ist die Wahrscheinlichkeit, in einer Stichprobe vom Umfang  $n$  das Ereignis **nicht** zu beobachten, gleich  $(1-\pi_0)^n$ . Diese Wahrscheinlichkeit soll höchstens 0.05 (5%) sein; d.h. die Behauptung, daß  $\pi \leq \pi_0$  ist, wenn das Ereignis in der Stichprobe nicht beobachtet wurde, soll höchstens in 5% der Fälle falsch und in 95% der Fälle richtig sein. Der erforderliche Stichprobenumfang  $n$  ergibt sich damit zu:

$$n = \frac{\log(0.05)}{\log(1 - \pi_0)}$$

Z.B. ist	für $\pi_0 = 0.1$	$n \approx 30$	zu wählen
	für $\pi_0 = 0.01$	$n \approx 300$	zu wählen
	für $\pi_0 = 0.001$	$n \approx 3000$	zu wählen.

### 3.3 Vorgabe der halben Breite eines Konfidenzintervalls

Der Standardfehler kennzeichnet die Variabilität des Schätzwertes, aber nicht die Lage des zu schätzenden Parameters. Den Stichprobenumfang durch Vorgabe des Standardfehlers (als Maß für die Genauigkeit des Schätzwertes) festzulegen, ist daher zwar eine sehr einfache, aber nicht ganz befriedigende Methode. Eine umfassendere Aussage über den Parameter liefert das **Konfidenzintervall** zu einer vorgegebenen Konfidenzwahrscheinlichkeit, die allgemein mit  $1-\alpha$  bezeichnet wird (für kleine Werte  $\alpha$ , z.B. 0.05 oder 0.01). Dies ist ein aus den Stichprobenwerten berechnetes Intervall, das den Wert des Parameters mit der vorgegebenen Konfidenz  $1-\alpha$  überdeckt. Dies bedeutet, daß bei wiederholter Stichprobennahme und Berechnung des Konfidenzintervalls der Anteil  $1-\alpha$  dieser Intervalle den 'wahren' Wert des Parameters enthält und nur der Anteil  $\alpha$  ihn nicht enthält.

Ein  $(1-\alpha)$ -Konfidenzintervall für den Mittelwert  $\mu$  der Daten  $x_1, \dots, x_n$ , deren Standardabweichung  $\sigma$  bekannt ist, ist das Intervall mit den beiden Grenzen:

$$\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{und} \quad \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

wobei  $\bar{x}$  der Mittelwert der Daten und  $z_{1-\alpha/2}$  die  $(1-\alpha/2)$ -Quantile der Standard-Normalverteilung sind.

Ein Maß für die Genauigkeit, mit der der Mittelwert  $\mu$  durch  $\bar{x}$  geschätzt wird, ist der **maximale Abstand der oberen oder unteren Grenze des Konfidenzintervalls** vom 'wahren' Wert  $\mu$ . Der Stichprobenumfang  $n$  soll so groß sein, daß dieser Abstand einen gegebenen Betrag  $\Delta$  einhält. Diese Forderung ist noch nicht eindeutig, da die Grenzen des Konfidenzintervalls von  $\bar{x}$  und damit von der Realisation einer Zufallsgröße abhängen, die vor der Stichprobennahme nicht bekannt ist. Man muß also noch angeben, mit welcher Wahrscheinlichkeit die aus den Stichprobenwerten zu berechnenden Grenzen den Abstand  $\Delta$  zu  $\mu$  einhalten sollen. Diese Wahrscheinlichkeit wird mit  $1-\beta$  bezeichnet. Die Forderung an  $n$  lautet somit: Es soll mit Wahrscheinlichkeit  $1-\beta$  gelten:

$$\mu - (\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}) < \Delta \quad \text{und} \quad (\bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}) - \mu < \Delta.$$

Durch Division beider Seiten der Ungleichungen mit  $\sigma/\sqrt{n}$  und Zusammenfassen beider Ungleichungen erhält man so die Forderung:

$$\Pr_{\mu} \left( \frac{|\bar{x} - \mu|}{\sigma} \sqrt{n} < \frac{\Delta}{\sigma} \sqrt{n} - z_{1-\alpha/2} \right) = 1 - \beta$$

Die Lösung dieser Gleichung ist:

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta/2})^2}{\delta^2}$$

mit  $\delta = \Delta/\sigma$ . Dieses Verhältnis ist zusammen mit  $\alpha$  und  $\beta$  vorzugeben und damit  $n$  zu berechnen.

In der Praxis ist  $\sigma$  meist nicht bekannt. Es sind dann in der Formel für die Grenzen des Konfidenzintervalls  $\sigma$  durch den aus den Daten berechneten Schätzwert  $s$  und die Quantile  $z_{1-\alpha/2}$  durch die Quantile  $t_{1-\alpha/2, n-1}$  der zentralen t-Verteilung mit  $n-1$  Freiheitsgrade zu ersetzen (vgl. Abschnitt 4.1). Der nach obiger Formel berechnete  $n$ -Wert ist in diesem Fall etwas zu klein. Einen genaueren Wert erhält man, wenn in der Formel die z-Quantilen durch die t-Quantilen mit  $n-1$  Freiheitsgrade ersetzt werden. Die Bestimmung von  $n$  ist dann allerdings iterativ durchzuführen. Die entsprechenden Werte von  $n$  können der Tabelle im Abschnitt 4.1 entnommen werden.

#### 4. Statistisches Testen

Mit statistischen Tests sollen **Hypothesen über Parameter von Grundgesamtheiten** anhand von Daten (Stichprobenergebnissen) überprüft werden.

Die Grundgesamtheit ist durch ihre **Verteilungsfunktion  $F(x; \Theta)$**  mit unbekanntem Parameter  $\Theta$  gekennzeichnet; d.h durch die Wahrscheinlichkeit für Realisationen von  $X$ , die kleiner oder gleich  $x$  sind:  **$F(x; \Theta) = \Pr(X \leq x; \Theta)$** .

Durch die Hypothesen werden die möglichen Parameterwerte in zwei Bereiche eingestuft, in den Bereich der **Nullhypothese** und in den der **Alternativhypothese**.

Man unterscheidet eine zweiseitige und einseitige Testung:

Bei **zweiseitiger** Testung gilt:

$$\begin{array}{ll} \text{Nullhypothese } H_0 & \Theta = \Theta_0 \\ \text{Alternativhypothese } H_1 & \Theta \neq \Theta_0 \end{array}$$

Bei **einseitiger** Testung gilt:

$$\begin{array}{ll} \text{Nullhypothese } H_0 & \Theta \leq \Theta_0 \quad (\text{oder } \Theta \geq \Theta_0) \\ \text{Alternativhypothese } H_1 & \Theta > \Theta_0 \quad (\text{oder } \Theta < \Theta_0) \end{array}$$

Der Unterschied zwischen der **Nullhypothese** und den Daten wird mit einer **Teststatistik  $T(\mathbf{x})$**  quantifiziert, die aus den Stichprobenergebnissen  $x_1, x_2, \dots, x_n$  berechnet wird. Der aus den Daten berechnete Wert von  $T(\mathbf{x})$  wird mit  $t_0$  bezeichnet. Bei zukünftigen Stichproben werden andere Daten und damit auch andere Werte  $t$  der Teststatistik  $T(\mathbf{x})$  vorkommen. In der Grundgesamtheit aller möglichen Wiederholungen der Beobachtungen (des Experiments) variiert der Wert von  $T(\mathbf{x})$  zufällig. In dieser Grundgesamtheit (Stichprobengesamtheit) ist  $T(\mathbf{x})$  eine Zufallsgröße  $T$ , deren Verteilung  $F_t(t)$  von der Verteilung  $F(x; \Theta)$  der Beobachtungen und dem Stichprobenumfang  $n$  abhängt.

Zur Bewertung von  $t_0$  wird die **Signifikanzwahrscheinlichkeit  $P$**  berechnet. Das ist die Wahrscheinlichkeit, bei zukünftigen Stichproben den gemessenen Unterschied  $t_0$  oder einen größeren Unterschied zu erhalten, wenn die Nullhypothese  $H_0$  gilt:

$$P = \Pr(T > t_0 | H_0) = 1 - F_t(t_0; \Theta_0)$$

Je größer der Unterschied der Daten zu  $H_0$  (und damit  $t_0$ ) ist, desto kleiner ist  $P$ .

Umfaßt  $H_0$  einen Bereich (z.B. bei einseitiger Testung den Bereich  $\Theta \leq \Theta_0$ ), dann ist  $P$  das Maximum der Wahrscheinlichkeit für  $T > t_0$  im Bereich von  $H_0$ . Bei einseitiger Testung wird dieses Maximum in der Regel für  $\Theta = \Theta_0$  angenommen.

### Test als Entscheidung:

Um zu entscheiden, ob die Nullhypothese oder Alternativhypothese angenommen werden soll, wird eine Schwelle  $\alpha$  (meist 0.05) vorgegeben und  $H_1$  angenommen ( $H_0$  abgelehnt), wenn  $P \leq \alpha$  (0.05) ist. Die Schwelle  $\alpha$  (0.05) ist die Wahrscheinlichkeit,  $H_0$  irrtümlich abzulehnen, wenn  $H_0$  gilt (**Fehler 1. Art**);  $\alpha$  ist die entsprechende Irrtumswahrscheinlichkeit 1. Art. Der  $t$ -Wert, bei dem  $H_0$  abgelehnt wird, ist die **Signifikanzschwelle**  $t_s$ . Es gilt:  $\Pr(T < t_s | H_0) = \alpha$ . Wenn  $H_0$  abgelehnt wird, nennt man den Unterschied zwischen Daten und Nullhypothese **signifikant**.

Die Annahme der Nullhypothese ( $P > \alpha$ ) bedeutet **nicht**, daß  $H_0$  mit großer Zuverlässigkeit gilt. Der Fehler,  $H_0$  anzunehmen ( $H_1$  abzulehnen), obwohl  $H_1$  gilt ( $\Theta \neq \Theta_0$ ), ist der **Fehler 2. Art**. Die Wahrscheinlichkeit hierfür bezeichnet man mit  $\beta$ . Sie hängt bei gegebenem  $\alpha$  vom Wert des Parameters  $\Theta$  und dem Stichprobenumfang  $n$  ab ( $\beta_{\alpha, n}(\Theta)$ ).

Die Wahrscheinlichkeit  $P_{\alpha, n}(\Theta)$ ,  $H_0$  bei gegebenem  $\alpha$  und  $n$  abzulehnen, wenn der Parameter den Wert  $\Theta$  hat, nennt man die **Powerfunktion** (Teststärke) zu gegebenem  $\alpha$  und  $n$ . Es gilt:  $P_{\alpha, n}(\Theta) = \Pr(T > t_s | \Theta) = 1 - F_t(t_s; \Theta) = 1 - \beta_{\alpha, n}(\Theta)$ .

### Tabelle der möglichen Testentscheidungen

Es gilt	Entscheidung für:	
	H <sub>0</sub>	H <sub>1</sub>
H <sub>0</sub>	richtig	<b>Fehler 1. Art</b>
H <sub>1</sub>	<b>Fehler 2. Art</b>	richtig

#### Festlegung des Stichprobenumfangs n:

Die Testschwelle  $\alpha$  (d.i. die Irrtumswahrscheinlichkeit 1. Art) wird (aus Aberglaube und Gewohnheit) fast immer auf 5% festgelegt. Damit ist die Genauigkeit fixiert, mit der eine gültige Nullhypothese zu Recht angenommen wird (nämlich 95%). Ziel eines Experiments ist es im allgemeinen aber nicht, eine Nullhypothese zu bestätigen, sondern sie abzulehnen, wenn eine relevante Abweichung  $\Theta_1 \neq \Theta_0$  besteht. Die Wahrscheinlichkeit dafür gibt die Powerfunktion  $P_{\alpha,n}(\Theta_1)$  an. Diese Funktion kennzeichnet somit die Genauigkeit des Tests hinsichtlich des eigentlichen Testzieles, nämlich des Auffindens relevanter Unterschiede. Der Stichprobenumfang ist daher so festzulegen, daß für einen relevanten Unterschied  $\Theta_1$  eine vorgegebene große Power  $P_{\alpha,n}(\Theta_1)$  erreicht wird. Hierfür sind folgende Schritte erforderlich:

- Vorgabe von  $\alpha$  (meist 0.05)
- Vorgabe eines von  $\Theta_0$  abweichenden relevanten Wertes  $\Theta_1$  (Referenzwert)
- Vorgabe von  $\beta$  bzw. der Power  $P_{\alpha,n}(\Theta_1)=1-\beta$  (meist  $\beta=0.2$  bzw.  $P_{\alpha,n}(\Theta_1)=0.8$ ).

Der Stichprobenumfang n (bzw. N bei mehr als einer Stichprobe) wird so berechnet, daß bei Gültigkeit des Referenzwertes  $\Theta_0$  und bei dem vorgegebenen  $\alpha$  die Irrtumswahrscheinlichkeit 2. Art höchstens den vorgegebenen Wert  $\beta$  besitzt bzw. die Power  $P_{\alpha,n}(\Theta_1)$  den vorgegebenen Wert  $1-\beta$  (z.B. 0.8) mindestens erreicht.

Dies wird im folgenden für einige wichtige Tests erläutert.

#### 4.1 Testen der mittleren Änderung (verbundener (paarweiser) t-Test)

Eine Meßgröße x wird zu Beginn und am Ende einer Periode bestimmt. Es wird danach gefragt, ob der Mittelwert der Meßgröße am Ende der Periode gegenüber dem Beginn nicht erhöht ist ( $H_0: \mu_{\text{Differenz}} \leq 0$ ) oder erhöht ist ( $H_1: \mu_{\text{Differenz}} > 0$ ). Es ist eine einseitige Testung durchzuführen.

Beispiel: Versuchsergebnisse bei n=8 Wiederholungen:

Beginn	$x_{1i}$ :	5	7	3	8	6	10	3	4
Ende	$x_{2i}$ :	8	12	9	6	5	11	9	8
Differenz	$d_i$ :	3	5	6	-2	-1	1	6	4

Mittelwert und Standardabweichung der Differenzen  $d_i$  sind:

$$\bar{d} = 2.75 \quad s^2_d = 9.64 \quad s_d = 3.10$$

Zu testen ist die Nullhypothese  $H_0: \mu_d \leq 0$  gegen die Alternativhypothese  $H_1: \mu_d > 0$ .

Als Teststatistik wird die t-Statistik genommen:

$$t = \frac{\bar{d}\sqrt{n}}{s_d}$$

Die Daten des Experiments ergeben den Wert  $t_0 = 2,5$  der Teststatistik.

Zur Berechnung der Signifikanzwahrscheinlichkeit  $P$  benötigt man die Kenntnis der Verteilung der Teststatistik, wenn  $H_0$  gilt. Unter  $H_0$  hat  $t$  (bei  $n > 6$ ) eine zentrale t-Verteilung  $F_{t,n-1}(t)$  mit  $n-1$  Freiheitsgraden, die 1908 von Gosset, einem leitenden Angestellten der Guinness-Brauerei, der unter dem Pseudonym 'Student' schrieb, erstmals publiziert wurde. Die Signifikanzschwelle  $t_s$  bei einem vorgegebenen  $\alpha$  ist die  $(1-\alpha)$ -Quantile  $t_{1-\alpha,n-1}$  der zentralen t-Verteilung mit  $n-1$  Freiheitsgraden. In der folgenden Tabelle sind diese Quantilen für  $\alpha=0.05$  und  $\alpha=0.025$  und verschiedene Freiheitsgrade angegeben. Bei  $\alpha=0.05$  und  $n-1=7$  Freiheitsgraden ist  $t_{1-0.05,7}=1.895$ . Der aus den Daten des Experiments berechnete t-Wert  $t_0$  ist 2.5 und somit größer als  $t_{1-\alpha,n-1}$ . Die Nullhypothese wird verworfen. Die mittlere Änderung ist **signifikant** größer als 0. Die Signifikanzwahrscheinlichkeit ist:  $P = \Pr(t > t_0 | \mu_d = 0) = (1 - F_{t,n-1}(2.5)) = 0.03$

**Tabelle der  $(1-\alpha)$ -Quantilen der zentralen t-Verteilung**

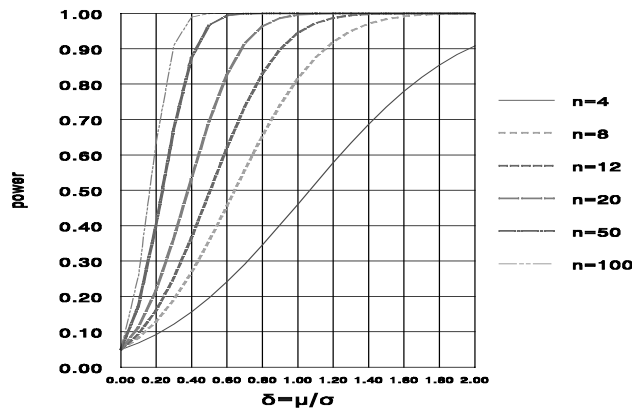
Freiheits- grade	$\alpha=0.05$	$\alpha=0.025$	Freiheits- grade	$\alpha=0.05$	$\alpha=0.025$
1	6.314	12.706	11	1.796	2.201
2	2.920	4.303	12	1.782	2.179
3	2.353	3.182	13	1.771	2.160
4	2.132	2.776	14	1.763	2.145
5	2.015	2.571	15	1.753	2.132
6	1.934	2.447	20	1.724	2.086
7	1.895	2.365	30	1.697	2.042
8	1.860	2.306	60	1.671	2.000
9	1.833	2.262	100	1.660	1.984
10	1.812	2.228	$\infty$	1.645	1.960

### Die Powerfunktion des Tests:

Für  $\mu_d$  größer als Null (Alternativhypothese) gibt die Power die Wahrscheinlichkeit an, mit der bei dem gegebenen Stichprobenumfang  $n$  ein signifikantes Ergebnis zu erwarten ist. Diese Power hängt aber nicht nur von  $\mu_d$  (sowie  $\alpha$  und  $n$ ) ab, sondern auch noch von der Standardabweichung  $\sigma_d$  der Differenzen in der Grundgesamtheit. Die t-Statistik hat unter der Alternative eine nichtzentrale t-Verteilung mit dem Nichtzentralitätsparameter  $nc = (\mu_d / \sigma_d) \sqrt{n}$ . Zur Berechnung der Power ist als Referenzwert nicht  $\mu_d$ , sondern der Quotient  $\delta = \mu_d / \sigma_d$  vorzugeben. In der folgenden Abbildung sind für verschiedene  $n$ -Werte die Powerfunktionen des t-Tests bei einseitiger Testung mit  $\alpha=0.05$  gezeigt.

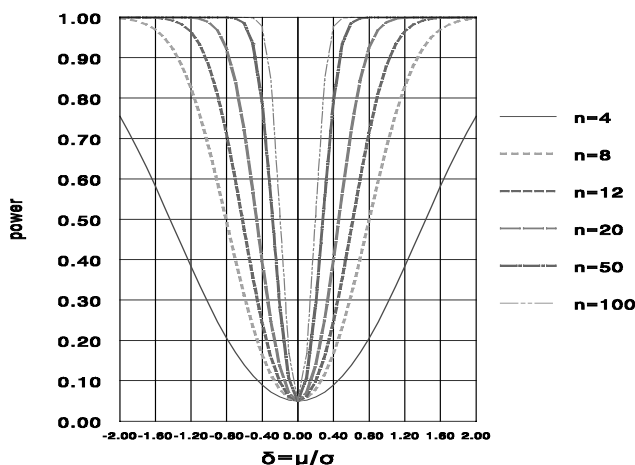
Man erkennt, daß mit zunehmendem  $n$  die Kurven immer steiler verlaufen. Für die Stichprobengröße  $n=8$  des Beispiels ist bei  $\delta=1.0$  mit einer Power von etwa 0.8 ein signifikantes Ergebnis zu erwarten.

**Power des verbundenen t-Tests  
zu  $\alpha=0.05$  (einseitig)**



Bei der Testung der Nullhypothese  $H_0: \mu_d=0$  gegen die zweiseitige Alternative  $\mu_d \neq 0$  ist als Teststatistik der Betrag von  $t$  ( $|t|$ ) zu nehmen.  $H_0$  wird abgelehnt, wenn dieser Betrag größer als die Signifikanzschwelle  $t_s$  ist; d.h. entweder  $t > t_s$  oder  $t < -t_s$  ist. Die Wahrscheinlichkeit, daß unter  $H_0$  entweder  $t > t_s$  oder  $t < -t_s$  eintritt, soll insgesamt  $\alpha$  sein. Dies ist gegeben, wenn  $t_s$  so gewählt wird, daß unter  $H_0$  die Wahrscheinlichkeit für  $t > t_s$  gleich  $\alpha/2$  und die Wahrscheinlichkeit für  $t < -t_s$  ebenfalls gleich  $\alpha/2$  ist. Wegen der Symmetrie der zentralen t-Verteilung um 0 ist als Signifikanzschwelle die  $(1-\alpha/2)$ -Quantile  $t_{1-\alpha/2, n-1}$  der zentralen t-Verteilung mit  $n-1$  Freiheitsgraden zu nehmen. Darin besteht der wesentliche Unterschied zwischen einseitiger und zweiseitiger Testung. Da  $t_{1-\alpha/2, n-1}$  stets größer als  $t_{1-\alpha, n-1}$  ist, hat für eine gegebene Alternative  $\delta = \mu_d / \sigma_d$  der zweiseitige Test eine geringere Power als der einseitige Test. In der folgenden Abbildung sind die Powerfunktionen des zweiseitigen Tests für  $\alpha=0.05$  und verschiedene  $n$ -Werte gezeigt.

**Power des verbundenen t-Tests  
zu  $\alpha=0.05$  (zweiseitig)**



Für  $n=8$  und  $c=1$  ist die Power bei zweiseitiger Testung etwa 0.7. Für den Referenzwert  $\delta=1$  wird die Power 0.8 mit dem Stichprobenumfang  $n=10$  erreicht.

### Bestimmung des Stichprobenumfangs

Der Stichprobenumfang  $n$  soll so groß sein, daß bei einem Referenzwert  $\delta = \mu_d / \sigma_d$  und der vorgegebenen Irrtumswahrscheinlichkeit 1. Art ( $\alpha$ ) die Nullhypothese  $H_0$  mit der vorgegebenen Power  $1 - \beta$  abgelehnt wird. Die Nullhypothese wird abgelehnt, wenn (bei einseitiger Testung)  $t > t_{1-\alpha, n-1}$  ist. Bei gegebenem  $\delta$  ist die Wahrscheinlichkeit dafür:

$P_{\alpha, n}(\delta) = 1 - F_{t, n-1}(t_{1-\alpha}, \delta \sqrt{n})$ , wobei  $F_{t, n-1}(\cdot, \cdot)$  die Verteilung der nichtzentralen t-Verteilung mit  $n-1$  Freiheitsgraden und dem Nichtzentralitätsparameter  $nc = \delta \sqrt{n}$  ist. Setzt man diesen Ausdruck gleich  $1 - \beta$ , dann kann daraus der erforderliche Stichprobenumfang berechnet werden. Die Berechnung muß iterativ erfolgen, da ja das gesuchte  $n$  auch in den Freiheitsgraden vorkommt. In der unten stehenden Tabelle sind für verschiedene  $\alpha$  und  $\beta$  die erforderlichen Stichprobenumfänge  $n$  angegeben. Bei einem zweiseitigen Test mit der (gesamten) Irrtumswahrscheinlichkeit  $\alpha = 0.05$  sind die Werte  $n$  in der Spalte mit  $\alpha = 0.025$  zu nehmen.

### Erforderlicher Stichprobenumfang $n$ beim verbundenen t-Test

$\delta$	$\alpha=0.025$	$\alpha=0.05$	$\alpha=0.025$	$\alpha=0.05$	$\alpha=0.025$	$\alpha=0.05$
	$\beta=0.2$	$\beta=0.2$	$\beta=0.1$	$\beta=0.1$	$\beta=0.05$	$\beta=0.05$
0.1	787	620	1053	858	1302	1084
0.2	199	156	256	216	327	272
0.3	90	71	119	97	147	122
0.4	52	41	68	55	84	70
0.5	34	27	44	36	54	45
0.6	24	19	32	26	39	32
0.7	19	15	24	19	29	24
0.8	15	12	19	15	23	19
0.9	12	10	16	13	19	15
1.0	10	8	13	11	16	13

Eine einfache Formel für den erforderlichen Stichprobenumfang  $n$  erhält man, wenn auch für den Test die Standardabweichung  $\sigma_d$  als bekannt angesehen wird und daher die Teststatistik  $z = (\bar{d} / \sigma_d) \sqrt{n}$  (bzw. bei zweiseitigem Test  $|z|$ ) genommen wird. Diese ist für  $\delta = \mu_d / \sigma_d$  normalverteilt mit dem Mittelwert  $\delta \sqrt{n}$  und der Standardabweichung 1. Die Nullhypothese wird abgelehnt, wenn  $z$  die Signifikanzschwelle  $z_s$  überschreitet, die bei einseitigem Test gleich der  $(1-\alpha)$ -Quantile  $z_{1-\alpha}$  und beim zweiseitigen Test gleich der  $(1-\alpha/2)$ -Quantile  $z_{1-\alpha/2}$  der Standardnormalverteilung ist. Die Power bei gegebenem  $\delta$  ist:  $P_{\alpha, n}(\delta) = 1 - \Phi(z_s - \delta \sqrt{n})$ . Der Stichprobenumfang  $n$  soll so groß sein, daß  $P_{\alpha, n}(\delta) = 1 - \beta$  ist. Die Lösung dieser Gleichung ist:

$$n = \frac{(z_s + z_{1-\beta})^2}{\delta^2}$$

bei einseitiger Alternative ist:  $z_s = z_{1-\alpha} = 1.64$  für  $\alpha = 0.05$

bei zweiseitiger Alternative ist:  $z_s = z_{1-\alpha/2} = 1.96$  für  $\alpha = 0.05$

$z_{1-\beta} = 0.84$  für  $\beta = 0.20$

Für  $\alpha = 0.05$  (zweiseitig) und  $\beta = 0.2$  ist  $n \approx 8/\delta^2$ . Ein Vergleich mit den Tabellenwerten (2. Spalte) zeigt, daß das nach dieser Formel bestimmte  $n$  meist nur um 2 zu klein ist.

## 4.2 Vergleich von zwei Mittelwerten (unverbundener t-Test)

Häufig werden in Experimenten zwei Versuchsanordnungen miteinander verglichen und es wird danach gefragt, ob sich die Versuchsergebnisse beider Anordnungen unterscheiden. Da die Versuchsergebnisse Realisationen von zwei Zufallsgrößen sind, muß die Frage präziser lauten: 'ob sich die Verteilungen der Versuchsergebnisse unterscheiden'. Wenn die Verteilungen in ihrer Form ähnlich sind, kann man sich darauf beschränken, nach Unterschieden in der Lage der Verteilungen zu fragen. Der am häufigsten verwendete Lage-Parameter einer Verteilung ist ihr Mittelwert  $\mu$ . Die Versuchsfrage läuft also auf einen Vergleich der Mittelwerte der beiden Verteilungen hinaus.

Beim Versuch 1 wurden die  $n_1$  Daten  $x_{1,1}, x_{1,2}, \dots, x_{1,n_1}$  erhalten, beim Versuch 2 die  $n_2$  Daten  $x_{2,1}, x_{2,2}, \dots, x_{2,n_2}$ . Die Mittelwerte und Standardabweichungen dieser Daten sind:  $\bar{x}_1$  und  $s_1$  sowie  $\bar{x}_2$  und  $s_2$ ; die Mittelwerte und Standardabweichungen der Verteilungen werden mit  $\mu_1$  und  $\sigma_1$  sowie  $\mu_2$  und  $\sigma_2$  bezeichnet. Zu testen sind:

bei einseitiger Testung:  $H_0: \mu_1 \leq \mu_2$  gegen  $H_1: \mu_1 > \mu_2$  (oder  $H_0: \mu_1 \geq \mu_2$  gegen  $H_1: \mu_1 < \mu_2$ )

bei zweiseitiger Testung:  $H_0: \mu_1 = \mu_2$  gegen  $H_1: \mu_1 \neq \mu_2$ .

Als Teststatistik wird die t-Statistik genommen; d.i. der Quotient aus der Differenz der beiden Stichproben-Mittelwerte  $\bar{x}_1 - \bar{x}_2$  zu einem Schätzwert  $s_{\bar{d}}$  für den Standardfehler dieser Differenz:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{d}}}$$

Falls man annehmen kann, daß sich die beiden Verteilungen zwar in ihren Mittelwerten, nicht aber in ihren Standardabweichungen unterscheiden ( $\sigma_1 = \sigma_2 = \sigma$ ), dann kann aus den beiden Standardabweichungen  $s_1$  und  $s_2$  ein Schätzwert  $s$  für die gemeinsame Standardabweichung  $\sigma$  berechnet werden:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Der Schätzwert für den Standardfehler ist damit:  $s_{\bar{d}} = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = s \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$  und die

Teststatistik  $t$  ist:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Für  $\mu_1 - \mu_2 = 0$  hat diese Statistik eine zentrale t-Verteilung mit  $n_1 + n_2 - 2$  Freiheitsgraden. Die Signifikanzschwelle  $t_s$  bei gegebenem  $\alpha$  ist somit:

einseitige Testung:  $t_s = t_{1-\alpha, n_1+n_2-2}$  zweiseitige Testung;  $t_s = t_{1-\alpha/2, n_1+n_2-2}$

Bei zweiseitiger Testung ist als Teststatistik der Betrag  $|t|$  zu nehmen.

Der Test soll an einem Beispiel demonstriert werden:

Versuch 1 ( $n_1=10$  Beobachtungen):  $x_{1i}$ : 10, 5, -2, 3, 8, 2, 0, 4, 2, 6.

Versuch 2 ( $n_2=6$  Beobachtungen):  $x_{2i}: 3, -4, 2, 0, 5, -2.$

Stichproben-Mittelwerte und Standardabweichungen sind:

Versuch 1:  $\bar{x}_1 = 3.8$   $s_1=3.6$

Versuch 2:  $\bar{x}_2 = 0.7$   $s_2=3.3$

Differenz: 3.1

Der Schätzwert für die gemeinsame Standardabweichung  $\sigma$  ist:  $s=3,5$ . Daraus folgt:

$$t = \frac{3.1}{3.5} \sqrt{\frac{10 \cdot 6}{10 + 6}} = 1.715$$

Die Zahl der Freiheitsgrade ist:  $10+6-2=14$ . Aus der Tabelle der  $(1-\alpha)$ -Quantilen der zentralen t-Verteilung (Abschnitt 4.1) entnimmt man für 14 Freiheitsgrade bei einseitiger Testung mit  $\alpha=0.05$  den Wert  $t_{0.95,14}=1.763$  und bei zweiseitiger Testung mit  $\alpha/2=0.025$  den Wert  $t_{0.975,14}=2.132$ . Die Nullhypothese kann in keinem Fall verworfen werden.

Die Powerfunktion hängt von der studentisierten Differenz  $\delta=(\mu_1-\mu_2)/\sigma$  ab und entspricht einer nichtzentralen t-Verteilung mit  $n_1+n_2-2$  Freiheitsgraden und dem Nichtzentralitäts-

parameter  $nc=\delta \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$ . Für  $n_1=n_2=n$  ist dieser Parameter:  $nc=\delta \sqrt{\frac{n}{2}}$ , wobei  $n$  der

Stichprobenumfang pro Versuch (Gruppe) ist. Um bei gegebenem  $\delta$  denselben Wert von  $nc$  wie beim verbundenen t-Test zu erhalten, muß pro Versuch der doppelte Stichprobenumfang genommen werden.

Der Stichprobenumfang  $n$  pro Versuch (gleiche Stichprobenumfänge für beide Versuche vorausgesetzt), der erforderlich ist, um bei einem Referenzwert  $\delta$  und gegebenem  $\alpha$  die Power  $1-\beta$  zu erreichen, kann analog zum verbundenen t-Test iterativ mit der nichtzentralen t-Verteilung berechnet werden. In der folgenden Tabelle sind für  $\alpha=0.025$  und  $\alpha=0.05$  sowie verschiedene  $\beta$ -Werte die erforderlichen Stichprobenumfänge  $n$  pro Versuch angegeben:

#### Erforderlicher Stichprobenumfang $n$ pro Versuch beim unverbundenen t-Test

$\delta$	$\alpha=0.025$	$\alpha=0.05$	$\alpha=0.025$	$\alpha=0.05$	$\alpha=0.025$	$\alpha=0.05$
	$\beta=0.2$	$\beta=0.2$	$\beta=0.1$	$\beta=0.1$	$\beta=0.05$	$\beta=0.05$
0.1	1571	1238	2103	1714	2600	2166
0.2	394	310	527	429	651	542
0.3	176	139	235	191	290	242
0.4	100	78	133	108	164	136
0.5	64	51	86	70	105	88
0.6	45	36	60	49	74	61
0.7	34	26	44	36	55	45
0.8	26	21	34	28	42	35
0.9	21	16	27	27	34	28
1.0	17	14	23	18	27	23

Die beim verbundenen t-Test hergeleitete Approximationsformel lautet für den unverbundenen t-Test:

$$n = \frac{2(z_s + z_{1-\beta})^2}{\delta^2}$$

Dies ergibt bei zweiseitiger Testung mit  $\alpha=0.05$  ( $z_s=1.96$ ) und  $\beta=0.2$  ( $z_{1-\beta}=0.842$ ):  $n=16/\delta^2$ .

## Anmerkungen

### a) Ungleiche Stichprobenumfänge $n_1$ und $n_2$

Sollen die Versuche mit ungleichen Stichprobenumfängen  $n_1$  und  $n_2$  durchgeführt werden, dann ist ein größerer Gesamtstichprobenumfang  $N=n_1+n_2$  erforderlich. Bezeichnet  $R=n_2/n_1$  das Verhältnis des geplanten Umfangs von Versuch 2 zu Versuch 1, dann lautet die Approximationsformel für den Gesamtstichprobenumfang  $N$ :

$$N = \frac{(z_s - z_{1-\beta})^2}{\delta^2} \cdot \frac{(1+R)^2}{R}$$

Der Gesamtumfang  $N$  vergrößert sich gegenüber  $N_{\text{gleich}}$  ( $R=1$ ;  $(1+R)^2/R=4$ ) um den Faktor  $(1+R)^2/(4R)$ . Im folgenden sind für einige Werte von  $R$  die entsprechenden Faktoren angegeben:

R	Faktor	R	Faktor	R	Faktor	R	Faktor
1.5	1.042	2.0	1.125	2.5	1.225	3.0	1.333

Für die reziproken Werte  $1/R$  erhält man denselben Faktor wie für  $R$ .

### b) Ungleiche Standardabweichungen $\sigma_1$ und $\sigma_2$

Bei ungleichen Standardabweichungen  $\sigma_1 \neq \sigma_2$  ist der Schätzwert für den Standardfehler

der Differenz der Mittelwerte:  $s_{\bar{d}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ . Dieser Standardfehler wird am kleinsten,

wenn die beiden Stichprobenumfänge so gewählt wurden, daß  $R=n_2/n_1=\sigma_2/\sigma_1$  gilt. In diesem Fall sind also gleiche Stichprobenumfänge nicht optimal. Die Verteilung der Teststatistik  $t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{d}}}$  kann durch eine t-Verteilung mit  $g=Rs_1^2/(Rs_1^2+s_2^2)$  approximiert

werden (Satterthwaite-Welch-Approximation). Der Nichtzentralitätsparameter bei  $\mu_1 \neq \mu_2$

ist:  $nc=(\mu_1-\mu_2)/\sigma_{\bar{d}}$  mit  $\sigma_{\bar{d}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ . Bei  $R=\sigma_2/\sigma_1$  gilt:  $\sigma_{\bar{d}} = (\sigma_1 + \sigma_2) / \sqrt{N}$  und

$nc=((\mu_1-\mu_2)/(\sigma_1+\sigma_2))\sqrt{N}$ . Damit läßt sich nach dem oben geschilderten Verfahren zu vorgegebenem  $\delta=(\mu_1-\mu_2)/(\sigma_1+\sigma_2)$  und Power  $P_{\omega,N}(c)$  der erforderliche Stichprobenumfang  $N$  bestimmen. Approximativ gilt:  $N=(z_s+z_{1-\beta})^2/\delta^2$ .

### c) Verteilungsannahmen

In vielen Lehrbüchern ist angegeben, daß der t-Test nur bei normal verteilten Daten angewandt werden darf. Dies ist so nicht richtig. Richtig ist, daß Gosset bei der Herleitung der (zentralen) t-Verteilung annahm, daß im Zähler die Differenz  $\bar{x}_1 - \bar{x}_2$  normalverteilt und der (vom Zähler unabhängige) Nenner  $s$  wie die Wurzel aus  $\sigma^2\chi^2/(n_1+n_2-2)$  verteilt sind, wobei  $\chi^2$  eine Summe aus  $n_1+n_2-2$  unabhängigen, standard-

normalverteilten Zufallsgrößen symbolisiert. Nun stehen aber im Zähler von  $t$  Mittelwerte und nach dem Hauptgrenzwertsatz konvergiert die Verteilung von Mittelwerten rasch gegen eine Normalverteilung, unabhängig von der Verteilung der Daten. Der Nenner beeinflusst mit zunehmendem  $n$  immer weniger die Verteilung. Die Verteilung der  $t$ -Statistik konvergiert also rasch gegen eine Normalverteilung. Der  $t$ -Test ist ein **asymptotisch verteilungsunabhängiger** Test. Überdies hat J. Lauter gezeigt, da fur eine sehr allgemeine Klasse von Verteilungen der Daten auch bei kleinem  $n$  die  $t$ -Statistik exakt eine  $t$ -Verteilung hat. Zahlreiche Simulationen haben dies bestatigt.

### 4.3 Vergleich von zwei Wahrscheinlichkeiten (Chi<sup>2</sup>-Test)

Haufig wird in Versuchen nur festgestellt, ob bei Versuchseinheiten (Versuchstieren) ein bestimmtes Ereignis (z.B. Tod) eintritt oder ausbleibt. Es interessiert die Wahrscheinlichkeit  $\pi$  fur das Eintreten des Ereignisses. Werden zwei Versuche mit unterschiedlichen Bedingungen durchgefuhrt, dann interessiert ein Vergleich der beiden Ereigniswahrscheinlichkeiten  $\pi_1$  und  $\pi_2$ . Die Ergebnisse beider Versuche werden in einer 4-Feldertafel zusammengefat:

	Ereignis	kein Ereignis	Gesamt
Versuch 1	a	b	$n_1$
Versuch 2	c	d	$n_2$
Gesamt	$m_1$	$m_2$	$N$

Dabei bedeuten:

- a = Zahl der Ereignisse im ersten Versuch
- b = Zahl ohne Ereignis im ersten Versuch
- $n_1$  = Stichprobenumfang des ersten Versuchs
- c = Zahl der Ereignisse im zweiten Versuch
- d = Zahl ohne Ereignis im zweiten Versuch
- $n_2$  = Stichprobenumfang des zweiten Versuchs
- $m_1 = a+c$  = Zahl der Ereignisse insgesamt
- $m_2 = b+d$  = Zahl ohne Ereignis insgesamt
- $N = n_1+n_2 = m_1+m_2$  = Gesamtzahl

Die Nullhypothese lautet:  $H_0: \pi_1=\pi_2$ , und die Alternative  $H_1: \pi_1\neq\pi_2$ .

Da es sich um eine zweiseitige Testung handelt, kann man als Teststatistik das Quadrat der Differenz der beiden beobachteten Ereignishaufigkeiten  $h_1=a/n_1$  und  $h_2=c/n_2$  nehmen, das auf einen Schatzwert fur die Varianz der Differenz  $h_1-h_2$  bezogen (d.h. dadurch dividiert) wird. Nach einigen Umformungen ergibt dies die

$$\text{Teststatistik: } \chi^2 = \frac{(a \cdot d - b \cdot c)^2 N}{n_1 \cdot n_2 \cdot m_1 \cdot m_2}.$$

Unter der Nullhypothese hat diese Teststatistik (wenn  $h_1$  und  $h_2$  als Realisationen normal verteilter Zufallsgrößen angesehen werden können) eine  $\chi^2$ -Verteilung mit einem Freiheitsgrad (d.i. die Verteilung des Quadrats einer standard-normalverteilten Zufallsgröe  $X$ ). Die Nullhypothese, da die Wahrscheinlichkeiten  $\pi_1$  und  $\pi_2$  gleich sind, wird mit Irrtumswahrscheinlichkeit  $\alpha=0.05$  verworfen, wenn  $X^2>3.8$  ist.

Zur Bestimmung der Stichprobenumfänge  $n_1$  und  $n_2$ , muß die Power des Testes bei unterschiedlichen  $\pi_1$  und  $\pi_2$  bekannt sein. Diese Power hängt von diesen beiden Wahrscheinlichkeiten und nicht nur von der Differenz  $\pi_1 - \pi_2$  ab. Eine Formel für den insgesamt erforderlichen Stichprobenumfang  $N = n_1 + n_2$  bei gegebenen  $\pi_1$ ,  $\pi_2$ ,  $R = n_2/n_1$ ,  $\alpha$  und  $\beta$  lautet:

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\pi_1 - \pi_2)^2} \cdot \frac{1+R}{R} \cdot (\pi_1(1-\pi_1)R + \pi_2(1-\pi_2))$$

Bei der Herleitung dieser Formel wurde angenommen, daß die Ereignishäufigkeiten  $h_1$  und  $h_2$  Realisationen von unabhängigen, normal verteilten Zufallsgrößen mit den Mittelwerten  $\pi_1$  bzw.  $\pi_2$  und den Varianzen  $\pi_1(1-\pi_1)/n_1$  bzw.  $\pi_2(1-\pi_2)/n_2$  sind.  $z_{1-\alpha/2}$  und  $z_{1-\beta}$  sind die entsprechenden Quantile der Standard-Normalverteilung.

In der folgenden Tabelle sind die Stichprobenumfänge  $N$  für  $R=1$  angegeben, die erforderlich sind, um bei dem in der ersten Zeile angegebenen  $\pi_1$  und dem in der ersten Spalte angegebenen  $\pi_2$  ( $\pi_1 > \pi_2$ ) mit der Power= 0.8 und  $\alpha=0.05$  ein signifikantes Ergebnis erwarten zu können.

**Tabelle für gesamten Stichprobenumfang N  
bei gleichen Stichprobengrößen  $n_1 = n_2 = N/2$  ( $R=1$ )  
für  $\alpha=0.05$  (zweiseitig) und  $\beta=0.2$  (Power=0.8)**

$\pi_2$	$\pi_1$							
	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	394	118	58	34	22	14	10	6
0.2		582	158	72	40	24	14	10
0.3			708	182	80	42	24	14
0.4				770	190	80	40	22
0.5					770	182	72	34
0.6						708	158	58
0.7							582	118
0.8								394

#### 4.4 Vergleich vom mehr als 2 Mittelwerten (Varianzanalyse)

In Versuchen werden oft mehrere Versuchsbedingungen systematisch variiert und es wird gefragt, ob sich die Mittelwerte der Ergebnisse dabei ändern. Als Beispiel sei ein Versuch genannt, bei dem verschiedene Stoffe und eine Negativkontrolle jeweils verschiedenen Stichproben von Tieren appliziert werden und gefragt wird, ob sich die mittleren Konzentrationen eines Toxins in der Leber der Tiere bei den verschiedenen Stoffen unterscheiden. Die Stoffe bilden einen **Faktor**, der systematisch in **Stufen**

gegeben wird. Werden insgesamt  $k$  Stufen gegeben, die mit dem Index  $i = 1, \dots, k$  gekennzeichnet sind, und wird jede Stufe bei  $n$  Versuchseinheiten (Tiere, Wiederholungen) geprüft, dann wird für den Meßwert  $x_{ij}$  bei der Wiederholung  $j$  der Stufe  $i$  ein lineares Modell angesetzt:

$$x_{ij} = \mu + \alpha_i + e_{ij} \quad (i=1, \dots, k, j=1, \dots, n)$$

Dabei ist  $\mu$  das Gesamtmittel der Meßwerte (über alle Stufen und Wiederholungen) und  $\alpha_i$  der Effekt der  $i$ -ten Stufe. Dieser ist definiert als die Abweichung des Mittelwertes  $\mu_i$  der  $i$ -ten Stufe vom Gesamtmittel:  $\alpha_i = \mu_i - \mu$ . Da das Gesamtmittel  $\mu$  der Mittelwert aller  $\mu_i$  ist, muß gelten:  $\sum \alpha_i = 0$ . Die Größen  $e_{ij}$  sind die Residualterme, die die Zufallsvariation innerhalb der Versuchsbedingungen repräsentieren. Sie werden als Realisationen unabhängiger Zufallsgrößen mit dem Mittelwert 0 und einer Varianz  $\sigma^2$ , die bei allen Versuchsbedingungen (Stufen) gleich ist (Residualvarianz), angesehen.

Die globale Nullhypothese, daß kein Einfluß der Versuchsbedingungen auf die Meßwerte besteht, lautet mit diesen Bezeichnungen:  $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$ . Die Alternative  $H_1$  besagt, daß diese Aussage nicht gilt. Da die Summe der Effekte  $\alpha_i$  Null ist, muß unter  $H_1$  für mindestens zwei Indizes  $m$  und  $n$  gelten:  $\alpha_m \neq 0$  und  $\alpha_n \neq 0$  (d.h.  $\mu_m - \mu_n \neq 0$ ).

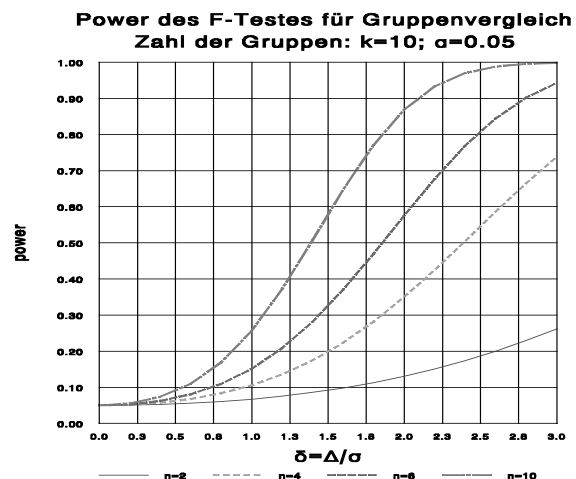
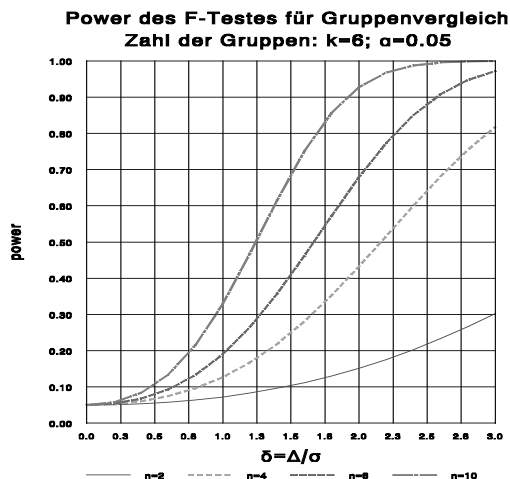
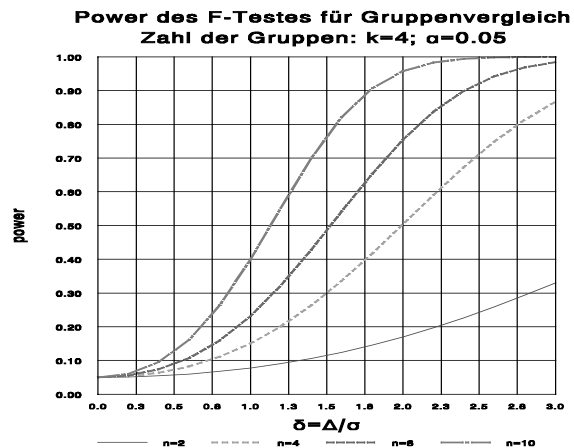
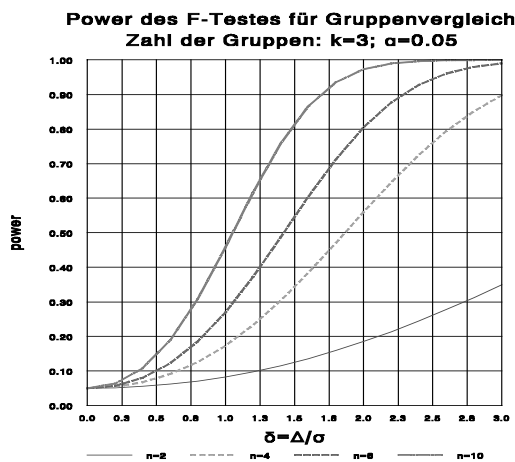
Zum Test der Nullhypothese wird eine Varianzanalyse der Daten  $x_{ij}$  durchgeführt; d.h. die gesamte Summe der Abweichungsquadrate  $SQ = \sum_i \sum_j (x_{ij} - x_{..})^2$  wird in den Anteil 'zwischen den Gruppen':  $SQ_{\text{zwischen}} = n \sum_i (x_i - x_{..})^2$  und den 'innerhalb der Gruppen':  $SQ_{\text{innerhalb}} = \sum_i \sum_j (x_{ij} - x_i)^2$  zerlegt. Dabei bedeuten  $x_i = \sum_j x_{ij} / n$  den mittleren Meßwert der Gruppe  $i$  und  $x_{..} = \sum_i \sum_j x_{ij} / (nk)$  das Gesamtmittel.  $SQ_{\text{zwischen}}$  hat  $k-1$  Freiheitsgrade (d.h. der Ausdruck kann als Summe von  $k-1$  Quadraten unabhängiger Realisationen von Zufallsgrößen dargestellt werden),  $SQ_{\text{innerhalb}}$   $k(n-1)$  Freiheitsgrade. Beide Größen sind stochastisch unabhängig, wenn die Meßwerte  $x_{ij}$  Realisationen unabhängiger, normal verteilter Zufallsgrößen sind. Teststatistik für die globale Nullhypothese  $H_0: \alpha_i = 0$  ( $i=1 \dots k$ ) ist:

$$F = \frac{SQ_{\text{zwischen}} / (k-1)}{SQ_{\text{innerhalb}} / k(n-1)}$$

Die Verteilung dieser Statistik wurde unter der Annahme, daß die Meßwerte  $x_{ij}$  Realisationen normal verteilter Zufallsgrößen sind, von Sir R.A. Fisher hergeleitet und wird deshalb ihm zu Ehren F-Verteilung genannt. Sie hängt bei Gültigkeit von  $H_0$  (zentrale F-Verteilung) nur von den beiden Freiheitsgraden  $k-1$  und  $k(n-1)$  ab.  $H_0$  wird bei gegebenem  $\alpha$  (z.B. 0.05) abgelehnt, wenn der berechnete F-Wert größer als die  $(1-\alpha)$ -Quantile  $F_{1-\alpha}$  der entsprechenden zentralen F-Verteilung ist.

Gilt die Nullhypothese nicht, d.h. sind mindestens zwei  $\alpha_i$ -Werte verschieden von Null (und damit  $\sum \alpha_i^2 > 0$ ), dann hängt die Verteilung der Statistik  $F$  von den beiden Freiheitsgraden und dem Nichtzentralitätsparameter  $nc = n \sum_i \alpha_i^2 / \sigma^2 = n \delta^2 / 2$  ab, wobei  $\sigma^2$  die Residualvarianz ist und  $\delta^2 = 2(\sum_i \alpha_i^2 / \sigma^2)$  gesetzt wurde. Für einen vorgegebenen Wert  $\delta$  ist die Power die Wahrscheinlichkeit, mit der bei diesem  $\delta$ -Wert  $F > F_{1-\alpha}$  zu

erwarten ist. In den folgenden Abbildungen sind die Powerfunktionen für verschiedene Anzahlen  $k$  von Gruppen und Wiederholungen  $n$  gezeigt:



Auf der Abszisse sind die Werte  $\delta=\Delta/\sigma$  aufgetragen. Zur Interpretation des Wertes  $\Delta$  ist anzunehmen, daß sich nur zwei Mittelwerte  $\mu_m$  und  $\mu_n$  um den Betrag  $\Delta$  unterscheiden und alle anderen Mittelwerte gleich  $(\mu_m+\mu_n)/2$  sind. Für die Effekte  $\alpha_i$  bedeutet dies, daß die beiden Effekte  $\alpha_m$  und  $\alpha_n$  verschiedene Vorzeichen und jeweils den Betrag  $\Delta/2$  haben und alle anderen Effekte  $\alpha_i=0$  sind. Diese Annahme entspricht einem Extremfall. Meistens dürften alle Gruppenmittelwerte kleinere oder größere Unterschiede haben und damit alle  $\alpha_i^2$  größer als 0 sein. Der Wert  $\delta^2$  ist der doppelte Quotient aus der Summe der  $\alpha_i^2$  zur Residualvarianz  $\sigma^2$ .

Der Stichprobenumfang  $n$  pro Gruppe, der erforderlich ist, um bei gegebenem  $\alpha$  und einem Referenzwert  $\delta$  mit der Power  $1-\beta$  ein signifikantes Ergebnis erwarten zu können, kann aus den Powerfunktionen hergeleitet werden. In der folgenden Tabelle sind für  $\alpha=0.05$  und  $1-\beta=0.8$  sowie für verschiedene Anzahlen  $k$  von Gruppen und  $\delta$ -Werte die erforderlichen Stichprobenumfänge  $n$  angegeben.

**Stichprobenumfang n pro Gruppe  
beim Vergleich der Mittelwerte von k Gruppen mit dem F-Test**  
 $\alpha=0.05, \beta=0.2, \text{Power}=0.8, \delta=\Delta/\sigma, \delta^2=2\sum\alpha_i^2/\sigma^2$

$\delta$	k=3	k=4	k=5	k=6	k=8	k=10
1.0	21	23	25	27	30	33
1.2	15	17	18	19	21	23
1.4	11	13	14	15	16	17
1.6	9	10	11	11	13	14
1.8	8	8	9	9	10	11
2.0	6	7	7	8	9	9
2.2	6	6	6	7	7	8
2.4	5	5	6	6	6	7
2.6	5	5	5	5	6	6
2.8	4	4	5	5	5	5
3.0	4	4	4	4	5	5

Wird die globale Nullhypothese, daß zwischen den Mittelwerten der Gruppen (Versuchsbedingungen) keine Unterschiede bestehen, abgelehnt, dann möchte man die Gruppen herausfinden, zwischen denen signifikante Unterschiede bestehen. Ein allgemeines und auch teststarkes Verfahren basiert auf dem Abschluß-Testprinzip: Es wird zunächst mit dem F-Test die Globalhypothese, daß zwischen allen k Gruppen keine Unterschiede bestehen, mit vorgegebenem  $\alpha$  geprüft. Wird diese Hypothese nicht abgelehnt, dann ist das Verfahren beendet. Wird die Hypothese abgelehnt, dann werden jeweils k-1 Gruppen auf dem Niveau  $\alpha$  getestet. Das Verfahren wird nur mit den Gruppen fortgesetzt, für die sich signifikante Unterschiede ergeben haben. Bei diesen Gruppen werden die Unterschiede zwischen allen k-2 Gruppen zum Niveau  $\alpha$  getestet usw. Am Ende erhält man die Paare von Gruppen, deren Mittelwerte sich signifikant unterscheiden. Der Vorteil dieses Verfahrens besteht darin, daß jeder der Tests zum vorgegebenen Niveau  $\alpha$  durchgeführt wird und die multiple Irrtumswahrscheinlichkeit, bei mindestens einem dieser Tests eine Nullhypothese fälschlich zu verwerfen, gleichgültig welche und wieviele Nullhypothesen gelten, nicht größer als  $\alpha$  ist.

Das oft angewandte Verfahren, alle Paare von Gruppen (oder die interessierenden Paare) mit jeweils einem t-Test zu testen, hat nicht diese Eigenschaft. Werden m t-Tests auf dem Niveau  $\alpha$  durchgeführt, dann ist die multiple Irrtumswahrscheinlichkeit größer als  $\alpha$ . Sie kann (nach einer Ungleichung von Bonferroni) maximal  $m \cdot \alpha$  (bzw. 1) werden. Um daher bei den m Tests das multiple Niveau  $\alpha$  einzuhalten, müssen die einzelnen Tests zum Niveau  $\alpha^* = \alpha/m$  durchgeführt werden. Bei größeren m-Werten ist  $\alpha^*$  sehr klein und die einzelnen Tests sind sehr testschwach. Das Abschluß-Testverfahren ist daher vorzuziehen. Wurde n so gewählt, daß der F-Test bei  $\delta = \Delta/\sigma$  mit der Power  $1-\beta$  ein signifikantes Ergebnis erwarten läßt, dann gilt mindestens diese Power auch für den Vergleich von zwei Gruppen, deren Mittelwerte sich um  $\Delta$  unterscheiden.

#### 4.5 Stichprobengröße bei Pilotversuchen (Einstichprobenprobleme)

Oft sollen in einem Pilotversuch orientierende Informationen über Mittelwerte, Standardabweichungen oder Wahrscheinlichkeiten erhalten werden, die dann zur genaueren Planung des größeren Versuchs verwendet werden. Soweit bei dem Pilotversuch nur Parameter der Meßgrößen unter einer Versuchsbedingung geschätzt werden, können zur Stichprobenplanung des Pilotversuchs die im Abschnitt 3 besprochenen Verfahren angewandt werden.

Soll im Pilotversuchen festgestellt werden, ob mit dem geplanten Verfahren überhaupt ein relevanter Effekt erreicht wird, kann ein statistischer Test durchgeführt werden (Einstichprobentest). Wird z.B. danach gefragt, ob der Mittelwert der Meßwerte einen relevanten Wert  $\mu_{\text{relevant}}$  besitzt, dann sind zwei Werte  $\mu_0 < \mu_{\text{relevant}}$  und  $\mu_1 > \mu_{\text{relevant}}$  vorzugeben und die Nullhypothese  $H_0: \mu \leq \mu_0$  gegen die Alternative  $\mu > \mu_0$  auf dem Niveau  $\alpha$  (z.B. 0.05) zu testen. Der Stichprobenumfang  $n$  soll so groß sein, daß für  $\mu = \mu_1$  die Nullhypothese mit der Power  $1 - \beta$  (z.B. 0.8) abgelehnt wird. Zum Test wird die t-Statistik:  $t = \frac{\bar{X} - \mu_0}{s} \sqrt{n}$  verwendet. Für die Stichprobengröße  $n$  gilt:

$$n \approx \frac{(z_{1-\alpha} + z_{1-\beta})^2}{(\mu_1 - \mu_0)^2 / \sigma^2} \cdot \text{Genauere Werte können der Tabelle zum verbundenen t-Test (Abschnitt 4.1) entnommen werden, wobei } \delta = (\mu_1 - \mu_0) / \sigma \text{ zu setzen ist.}$$

Werden im Pilotversuch Ereignisse beobachtet und soll festgestellt werden, ob die Wahrscheinlichkeit  $\pi$  für das Ereignis einen relevante Wert  $\pi_{\text{relevant}}$  erreicht, kann ähnlich vorgegangen werden. Es sind ein  $\pi_0 < \pi_{\text{relevant}}$  und ein  $\pi_1 > \pi_{\text{relevant}}$  sowie  $\alpha$  (z.B. 0.05) und  $1 - \beta$  (z.B. 0.8) festzulegen. Die Wahrscheinlichkeit wird als ungenügend angesehen, wenn die Nullhypothese  $H_0: \pi \leq \pi_0$  zum Niveau  $\alpha$  nicht verworfen werden kann. Die Stichprobengröße  $n$  soll so groß sein, daß für  $\pi = \pi_1$  die Power mindestens  $1 - \beta$  ist. Die erforderliche Stichprobengröße ergibt sich aus der Formel:

$$n = \frac{(z_{1-\alpha} \sqrt{\pi_0 \cdot (1 - \pi_0)} + z_{1-\beta} \sqrt{\pi_1 \cdot (1 - \pi_1)})^2}{(\pi_1 - \pi_0)^2}$$

Die Nullhypothese wird verworfen, wenn die Zahl der Ereignisse gleich oder größer als

$$r = n \cdot \pi_0 + z_{1-\alpha} \sqrt{n \cdot \pi_0 (1 - \pi_0)}$$

ist. In der folgenden Tabelle sind für  $\alpha = 0.05$  und  $1 - \beta = 0.8$  sowie verschiedenen Werte  $\pi_0$  und  $\pi_1$  die Werte  $r$  und  $n$  (als  $r/n$ ) angegeben:

$\pi_0$	$\pi_1$							
	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	11/69	5/20	3/10	2/6	2/4	2/3	1/2	1/1
0.2		29/109	10/29	5/13	4/8	3/5	2/3	2/2
0.3			50/136	15/35	8/16	5/9	4/5	3/3
0.4				71/151	21/38	10/16	7/9	4/5
0.5					87/153	24/37	11/15	7/8
0.6						95/142	25/33	11/13
0.7							92/119	23/26
0.8								73/83

#### 4.6 Reduktion der Stichprobe durch sequentielles Vorgehen

Bei der Bestimmung der Stichprobengröße sind ein Referenzwert  $\delta$  und die dazugehörige Power vorzugeben. Sind die tatsächlichen Unterschiede größer als  $\delta$ , dann könnte man bereits mit einer kleineren Stichprobe zu einer Entscheidung kommen. Dieser Nachteil der Vorgabe eines festen Stichprobenumfangs kann durch ein sequentielles Vorgehen vermieden werden. Dabei werden ebenfalls  $\alpha$  und die für einen Referenzwert  $\delta$  verlangte Power  $1-\beta$  vorgegeben. Der Stichprobenumfang  $n$  wird aber nicht festgelegt. Es wird vielmehr im Verlauf der Datenerhebung zu vorgegebenen oder frei wählbaren Zeitpunkten überprüft, ob mit den bisher erfaßten Ergebnissen eine Entscheidung getroffen werden kann, oder die Datenerhebung fortgesetzt werden muß. Der Stichprobenumfang ist bei diesem Vorgehen keine feste Zahl, sondern eine Zufallsgröße. Um die mit  $\alpha$  und der Power  $P(c)=1-\beta$  festgelegte Genauigkeit einzuhalten, kann der Stichprobenumfang bis zu einer Entscheidung größer als bei fester Vorgabe sein. Dies ist aber weniger wahrscheinlich. Häufiger wird das Verfahren mit einem geringeren Stichprobenumfang beendet werden können. Der zu erwartende Stichprobenumfang kann bis zur Hälfte des fest vorgegebenen Stichprobenumfangs reduziert sein.

Bei dem im vorigen Abschnitt besprochenen Einstichprobentest der Hypothese  $H_0: \pi \leq \pi_0$  gegen  $\pi > \pi_0$  mit der Vorgabe der Power  $1-\beta$  für  $\pi = \pi_1$ , kann der zu erwartende Stichprobenumfang bei einem zweistufigen Verfahren reduziert werden. In der ersten Stufe wird mit einer Stichprobe vom Umfang  $n_1$  entschieden, ob bereits  $H_0$  angenommen werden kann oder der Versuch fortzusetzen ist.  $H_0$  wird angenommen, wenn die Zahl der Ereignisse kleiner oder gleich einer Zahl  $r_1$  ist. Bei Fortsetzung wird eine weitere Stichprobe vom Umfang  $n_2$  genommen und  $H_0$  abgelehnt, wenn in beiden Stichproben vom Umfang  $n = n_1 + n_2$  höchstens  $r$  Ereignisse beobachtet wurden. Wurden mehr als  $r$  Ereignisse beobachtet, dann wird  $H_0$  abgelehnt und  $H_1$  angenommen. Für den Fall  $\pi_1 - \pi_0 = 0.2$ ,  $\alpha = 0.05$  und  $1 - \beta = 0.8$  sind in der folgenden Tabelle die erforderlichen Zahlen  $r_1/n_1$ ,  $r/n$  sowie der bei  $\pi = \pi_0$  zu erwartende Stichprobenumfang  $En(\pi_0)$  angegeben (aus: Simon R.: Optimal Two-Stage Designs for Phase II Clinical Trials, Controlled Clinical Trials 10, 1-10 (1989)).

$\pi_0$	$\pi_1$	$r_1/n_1$	$r/n$	$En(\pi_0)$	$\pi_0$	$\pi_1$	$r_1/n_1$	$r/n$	$En(\pi_0)$
0.05	0.25	0/9	2/17	12	0.40	0.60	7/16	23/46	25
0.10	0.30	1/10	5/29	15	0.50	0.70	8/15	26/43	24
0.20	0.40	3/13	12/43	21	0.60	0.80	7/11	30/43	21
0.30	0.50	5/15	18/46	24	0.70	0.90	4/6	22/27	15

Um z.B. den Test für  $\pi_0 = 0.20$  und  $\pi_1 = 0.40$  durchzuführen, ist zunächst eine Stichprobe mit  $n = 13$  zu nehmen.  $H_0$  wird angenommen, wenn dabei 3 oder weniger Ereignisse beobachtet werden. Werden mehr als 3 Ereignisse beobachtet, dann wird eine zweite Stichprobe mit  $n = 30$  genommen und  $H_0$  angenommen, wenn in der gesamten Stichprobe mit  $n = 43$  höchstens 12 Ereignisse beobachtet werden, sonst abgelehnt. Die mittlere Stichprobengröße ist 21. Aus der Tabelle des vorherigen Abschnitts geht hervor, daß bei fest vorgegebem Umfang eine Stichprobe mit  $n = 29$

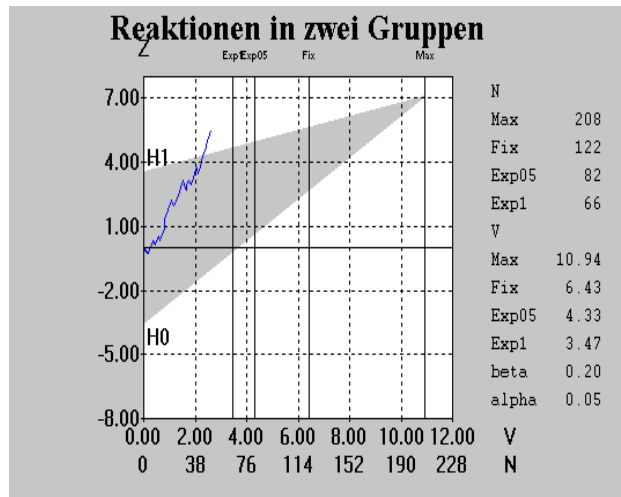
zu nehmen wäre und die Nullhypothese angenommen wird, wenn dabei weniger als 10 Ereignisse beobachtet werden.

Ein zweistufiges Vorgehen ist auch dann angebracht, wenn bei der Schätzung oder Testung von Mittelwerten keine Kenntnisse über die Standardabweichung  $\sigma$  vorliegen. In der ersten Stufe mit Stichprobenumfang  $n_1$  wird ein Schätzwert für  $\sigma$  ermittelt. Damit wird dann der Stichprobenumfang für die zweite Stufe so festgelegt, daß die Genauigkeitsforderungen erfüllt sind.

Zur Bestimmung eines Konfidenzintervalls für den Mittelwert mit vorgegebener Breite  $2\Delta$  wurde dieses Verfahren bereits 1945 von Stein vorgeschlagen. In der ersten Stufe wird eine Stichprobe mit (nicht zu großem)  $n_1$  genommen und mit den Daten  $\bar{x}_1$  und  $s_1^2$  berechnet. Hat das damit bestimmte  $(1-2\alpha)$ -Konfidenzintervall bereits die vorgegebene Breite  $2\Delta$ , dann wird das Verfahren beendet. Andernfalls wird eine zweite Stichprobe vom Umfang  $n_2$  genommen.  $n_2$  ist die kleinste ganze Zahl, die größer oder gleich  $(s_1^2/q) \cdot n_1 + 1$  ist, wobei  $q = (\Delta/t_{1-\alpha, n_1-1})^2$  ist. Aus den Werten beider Stichproben wird der Mittelwert  $\bar{x}$  berechnet. Das gesuchte Konfidenzintervall ist  $\bar{x} \pm \Delta$ . Dieses Verfahren wurde später von Bauer et al. auf mehr als 2 Stufen erweitert (Bauer P., Schreiber V. und Wohlzogen F.X.: Sequentielle statistische Verfahren. Gustav Fischer Verlag Stuttgart, New York, 1986).

Beim unverbundenen t-Test kann man ähnlich vorgehen. In einer Pilotstudie mit 2 Stichproben vom Umfang  $n_1$ /Gruppe wird ein vorläufiges  $t'$  berechnet und eine Schätzung  $s_1^2$  für die unbekannte Varianz  $\sigma^2$  ermittelt. Mit dieser Schätzung wird eine zweite Studie so geplant, daß bei vorgegebenem Unterschied  $\mu_1 - \mu_2$  mit der Power  $1 - \beta$  zum Niveau  $\alpha$  ein signifikantes Ergebnis zu erwarten ist. Das ausführliche Verfahren wurde von Friede T. und Kieser M. auf der 43. Jahrestagung der GMDS vom 14.-16. September 1998 in Bremen vorgetragen.

Allgemeine sequentielle Testverfahren für Vergleiche von 2 Stichproben sind die **Sequentiellen Dreieckspläne**. Dabei werden aus den Daten  $x_1, x_2, \dots$  fortlaufend Testgrößen  $Z_m$  berechnet, die die Information aller bis zum  $m$ -ten Wert  $x_m$  erfaßten Daten zusammenfassen. Der Verlauf von  $Z_m$  über den Stichprobenumfang  $m$  wird in ein Koordinatensystem eingezeichnet (Sequenzpfad). In diesem System wird ein Dreieck eingezeichnet. Verläuft der Sequenzpfad innerhalb des Dreiecks, dann wird die Datenerhebung fortgesetzt, erreicht er eine der Grenzen, dann wird abgeschlossen und die entsprechende Entscheidung getroffen.



Diese Pläne und Auswertungen können mit dem Programm TRIQ (BioMath GmbH Joachim-Jungius-Str. 9, 18059 Rostock) erstellt werden.

### Abschließende Bemerkungen

Zum Abschluß sei noch auf zwei Bücher verwiesen, in denen die Probleme der Stichprobenplanung ausführlich behandelt werden. Im zweiten Buch werden auch allgemeine Probleme der Versuchsplanung bei komplizierteren Fragestellungen, wie z.B. allgemeine Regressionsprobleme, mehrfaktorielle Versuche, Dosis-Wirkungsanalysen, behandelt:

**Bock J.;** Bestimmung des Stichprobenumfangs für biologische Experimente und kontrollierte klinische Studien. R.Oldenbourg Verlag, München Wien 1998

**Rasch D., Guiard V. und Nürnberg G.:** Statistische Versuchsplanung. Gustav Fischer Verlag, Stuttgart Jena New York 1992

Aus dem erstgenannten Buch sollen folgende Schlußfolgerungen und Empfehlungen zitiert werden:

- "Der erste und wichtigste Schritt bei der Planung des Umfangs einer Studie oder Experiments ist die **Präzisierung der Aufgabenstellung**. Dazu gehören die Festlegung des Aussagebereichs, der statistischen Fragestellung, der primären Variablen, des Designs (Versuchsanlage), der statistischen Methode und der Anforderungen an die Güte und Genauigkeit.
- Obgleich letztendlich ein konkreter Umfang festgelegt werden muß, sollte seine Bestimmung nicht auf die formale Berechnung mit Hilfe einer passenden Formel oder Software reduziert werden. Vielmehr sollten durch die Berechnung verschiedene Varianten abgeklärt werden, welches das geeignete Design der Studie ist, wie der Umfang von den Planungsparametern und der Analysenmethode abhängt bzw. sich die Aussagekraft der Studie in Abhängigkeit vom Umfang verändert.....
- So gesehen ist die Berechnung des Stichprobenumfangs nicht Selbstzweck, sondern ein Hilfsmittel zur Planung einer Studie mit ausreichendem "**statistischem Auflösungsvermögen**" - vergleichbar mit der Bereitstellung eines optischen Geräts mit ausreichendem optischen Auflösungsvermögen."

