**Hypothesen: Fehler 1. und 2. Art, Power eines statistischen Tests**

Die äußerst wichtige Tabelle über die Zusammenhänge zwischen Fehler 1. und 2. Art bei der Aufstellung und dem Test von Hypothesen sei noch einmal in Erinnerung gerufen:

| | $H_0$ wahr ($H_1$ falsch) | $H_0$ falsch ($H_1$ wahr) |
|---|---|---|
| Akzeptiere $H_0$ (verwerfe $H_1$) | Richtige Entscheidung | Fehler 2. Art Falscher Nichtnachweis („Ignoranz“) |
| Verwerfe $H_0$ (Akzeptiere $H_1$) | Fehler 1. Art Falscher Nachweis („Aberglaube“) | Richtige Entscheidung |

Anhand des bereits behandelten Beispiels der Versuchsreihe zur Feststellung der Wirksamkeit eines Medikamentes werden der Zusammenhang zwischen Fehler 1. Art, Fehler 2. Art und der Power eines statistischen Tests erläutert:

**Fall 1: Einseitiger Test**

Es werden $H_o$ und $H_A$ formuliert

$H_o$: $p = p_o$; (bzw. $p_o \leq 0{,}4$ als zusammengesetzte Hypothese)
$H_A$: $p > p_o$

Zur Überprüfung der Hypothese wird ein Binomialexperiment durchgeführt; Dazu wird an n Personen die Wirksamkeit des Medikamentes überprüft und die Anzahl der mit Erfolg behandelten Patienten ermittelt. Ist die Anzahl *z* der erfolgreich behandelten Personen größer oder gleich einer kritischen Zahl *k*, dann wird die Hypothese $H_o$ verworfen, ansonsten wird $H_o$ akzeptiert.

X...Zufallsvariable, X = (0,1,2,.....,n)

Akzeptiere $H_o$, falls $(0 \leq X < k)$
Verwerfe $H_o$, falls $(k \leq X)$

Der kritische Wert k hängt nun ab von n, der Anzahl der Probanden, und von einer festgelegten Irrtumswahrscheinlichkeit $\alpha$ für den Fehler 1. Art. In den Naturwissenschaften ist eine Festlegung von 5% für den Fehler 1. Art allgemein üblich. Dabei wird meist von einer 2-seitigen Irrtumswahrscheinlichkeit von $\alpha$ = 5% ausgegangen, einseitig ist $\alpha/2$ bzw. sind $\alpha_r$ und $\alpha_l$ dann nur jeweils 2,5%. Bezeichnungsweise: $\alpha_r$ = rechts-seitig, $\alpha_l$ = links-seitige Irrtumswahrscheinlichkeit)

$$\alpha_r = \sum_{x=k}^{n} p(X = x) = \sum_{x=k}^{n} \binom{n}{x} p^x (1-p)^{n-x} \qquad (1)$$

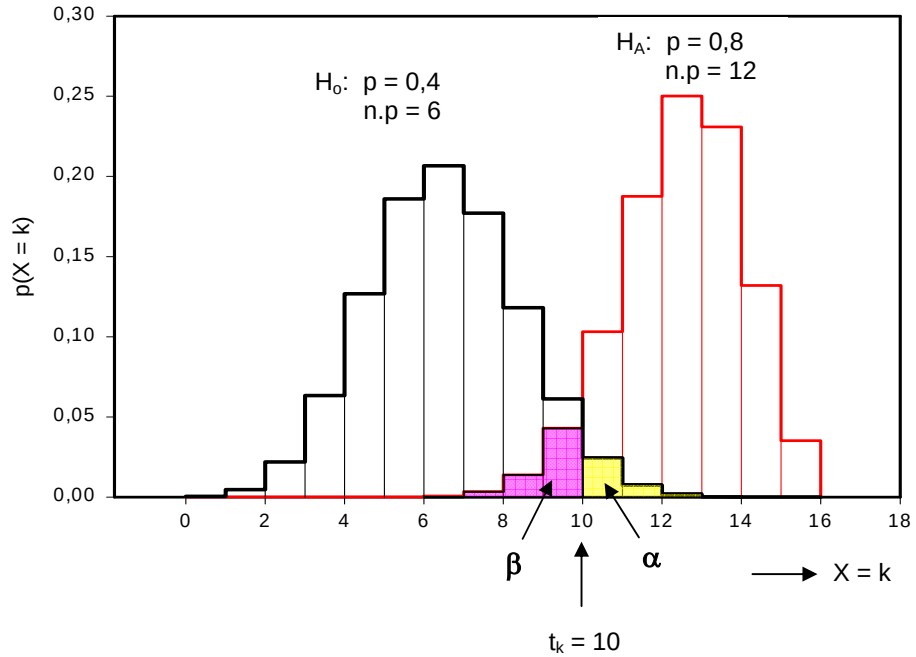
$\alpha_r$ wird ein Maximum für $p = p_o$, für kleinere Werte von p wird $\alpha_r$ ebenfalls immer kleiner.
Unter der Annahme, daß das Experiment 15 Personen umfaßt, ergeben sich als Funktion der „Wirksamkeitswahrscheinlichkeit“ des Mittels und einer festgelegten kritischen Region für $\alpha_r$, den Fehler 1. Art, folgende Werte:

**Einseitige $\alpha_r$ - Werte in Abhängigkeit von p und der kritischen Region**

| | Wirksamkeitswahrscheinlichkeit p | | | |
|---|---|---|---|---|
| kritische Region | 0,1 | 0,2 | 0,3 | 0,4 |
| $(7 \le X \le 15)$ | 0,0000 | 0,0181 | 0,1311 | 0,3902 |
| $(8 \le X \le 15)$ | 0,0000 | 0,0042 | 0,0500 | 0,2131 |
| $(9 \le X \le 15)$ | 0,0000 | 0,0008 | 0,0152 | 0,0950 |
| $(10 \le X \le 15)$ | 0,0000 | 0,0001 | 0,0037 | 0,0338 |
| $(11 \le X \le 15)$ | 0,0000 | 0,0000 | 0,0007 | 0,0093 |
| $(12 \le X \le 15)$ | 0,0000 | 0,0000 | 0,0001 | 0,0019 |
| $(13 \le X \le 15)$ | 0,0000 | 0,0000 | 0,0000 | 0,0003 |
| $(14 \le X \le 15)$ | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| $(15 \le X \le 15)$ | 0,0000 | 0,0000 | 0,0000 | 0,0000 |

Der kritische Bereich wird festgelegt als Intervall $(10 \le X \le 15)$, das einem $\alpha$ von 0,0338 entspricht. Damit bleibt $\alpha_r$ unterhalb 5%. Wäre $X = 9$ noch innerhalb dieses kritischen Bereiches, dann würde dadurch $\alpha_r$ auf 0,095 ansteigen. Damit würde $H_o$ zu häufig verworfen.

Fehler 1. Art ($\alpha$), Fehler 2. Art ($\beta$)




Der Fehler 2. Art, der als $\beta$-Fehler bezeichnet wird, tritt ein, wenn die Hypothese $H_A$ zutrifft, aber $H_o$ akzeptiert wird. Dieser Fehler 2. Art kann

genau angegeben werden, wenn $H_A$ ebenfalls genau festgelegt wird, z.B. in der Form $H_A$: p = 0,8.

$$\beta = \sum_{x=0}^{k-1} p(X = x) \;=\; \sum_{x=0}^{k-1} \binom{n}{x} p^x (1-p)^{n-x} \qquad (2)$$

β ist demnach der Anteil der Ereignisse, die unter der Gültigkeit von $H_A$ in den Akzeptanzbereich von $H_o$ fallen! Im angeführten Beispiel ist für $H_A$: p = 0,8; β = 0,0611.

**β - Werte in Abhängigkeit von p und der Akzeptanzregion**

| Akzeptanzregion | Wirksamkeitswahrscheinlichkeit | | | | |
|---|---|---|---|---|---|
| | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 |
| p(X<7) | 0,3036 | 0,0950 | 0,0152 | 0,0008 | 0,0000 |
| p(X<8) | 0,5000 | 0,2131 | 0,0500 | 0,0042 | 0,0000 |
| p(X<9) | 0,6964 | 0,3902 | 0,1311 | 0,0181 | 0,0003 |
| p(X<10) | 0,8491 | 0,5968 | 0,2784 | 0,0611 | 0,0022 |
| p(X<11) | 0,9408 | 0,7827 | 0,4845 | 0,1642 | 0,0127 |
| p(X<12) | 0,9824 | 0,9095 | 0,7031 | 0,3518 | 0,0556 |
| p(X<13) | 0,9963 | 0,9729 | 0,8732 | 0,6020 | 0,1841 |
| p(X<14) | 0,9995 | 0,9948 | 0,9647 | 0,8329 | 0,4510 |
| p(X<15) | 1,0000 | 0,9995 | 0,9953 | 0,9648 | 0,7941 |

Der Fehler β nimmt zu, je weiter die Akzeptanzregion zu größeren Werten hin ausgedehnt wird. Die Wahrscheinlichkeit sich zu irren, d.h. einen Fehler 2. Art zu begehen, steigt damit dramatisch an, weil die Hypothese $H_o$ noch immer akzeptiert wird, obwohl ihre Richtigkeit immer unwahrscheinlicher wird!

## Power des statistischen Tests

Auf dem Fehler 2. Art bauen der Begriff und die Definition der ***„Power“*** eines Tests auf. Diese ist 1-β, bzw. gleich der Wahrscheinlichkeit eine falsche Nullhypothese richtig zu diagnostizieren. Die Bezeichnung "Power" wird auch in der deutschsprachigen Literatur verwendet. Mitunter steht für „Power“ auch „Sensitivität“ oder Trennschärfe.

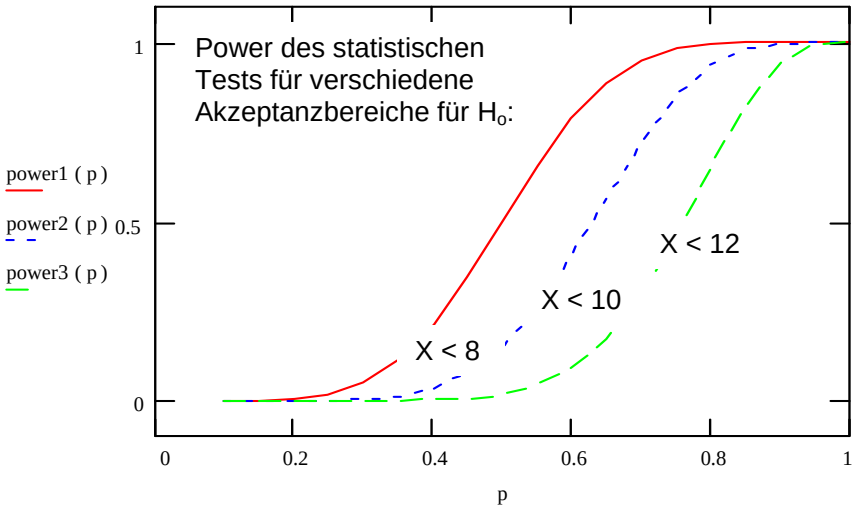
Die Power eines Tests hängt also vom kritischen Wert, bzw. der Akzeptanzregion, und von β ab, das sich aus der Alternativhypothese $H_A$ ergibt. Wird der kritische Wert zu niedrig angesetzt, dann erhöht sich zwar die Power eines Tests, gleichzeitig steigt aber auch α, der Fehler 1. Art. Die Hypothese $H_o$ wird also zu oft verworfen, obwohl sie richtig ist. Das wäre z.B. der Fall, wenn für Ho ein kritischer Wert von $t_k = 8$, bzw. für $H_o$ als Akzeptanzbereich X < 8 festgelegt würde. Umgekehrt verhält es sich, wenn der kritische Wert zu hoch angesetzt wird, z.B mit $t_k = 12$, bzw. einem Akzeptanzbereich von X < 12. In diesem Fall wird zwar α klein, oder sogar sehr klein werden, die Power des Tests nimmt allerdings ebenfalls ab, wenn die $H_A$ gleich bleibt. Um die Power des Tests wieder zu erhöhen, muß für die Alternativhypothese $H_A$ ein größerer Wert für p angenommen werden. Damit wächst aber die Differenz wischen $H_o$ und $H_A$, oder mit anderen Worten, der Test kann kleine Unterschiede nicht mehr erkennen. Das ist in vielen Fragestellungen natürlich nicht wünschenswert, deshalb ist es für die meisten

praktischen Fragestellungen auch irrelevant, unrealistisch kleine Werte für $\alpha$ und $\beta$ zu fordern, also eine gewisse Motivation für die Akzeptanz von 5% für $\alpha$. Der Fehler 2. Art ergibt sich dann automatisch aus der Alternativhypothese $H_A$. Ein guter statistischer Test soll sowohl einen kleinen Fehler 1.Art, als auch einen kleinen Fehler 2. Art aufweisen, d.h. die Power eines statistischen Tests soll möglichst nahe bei 1 liegen.

**Power = 1 – P(Fehler 2. Art, $\beta$)**
**= P($H_o$ zu verwerfen, wenn $H_o$ falsch ist)**
**= P($H_A$ akzeptieren, wenn $H_A$ richtig ist)**

**Power des statistischen Tests in Abhängigkeit von p und der Akzeptanzregion**

| Akzeptanzregion für $H_o$ | Wirksamkeitswahrscheinlichkeit | | | | |
|---|---|---|---|---|---|
| | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 |
| p(X<7) | 0,6964 | 0,9050 | 0,9848 | 0,9992 | 1,0000 |
| p(X<8) | 0,5000 | 0,7869 | 0,9500 | 0,9958 | 1,0000 |
| p(X<9) | 0,3036 | 0,6098 | 0,8689 | 0,9819 | 0,9997 |
| p(X<10) | 0,1509 | 0,4032 | 0,7216 | 0,9389 | 0,9978 |
| p(X<11) | 0,0592 | 0,2173 | 0,5155 | 0,8358 | 0,9873 |
| p(X<12) | 0,0176 | 0,0905 | 0,2969 | 0,6482 | 0,9444 |
| p(X<13) | 0,0037 | 0,0271 | 0,1268 | 0,3980 | 0,8159 |
| p(X<14) | 0,0005 | 0,0052 | 0,0353 | 0,1671 | 0,5490 |
| p(X<15) | 0,0000 | 0,0005 | 0,0047 | 0,0352 | 0,2059 |




**Fall 2: Zweiseitiger Test**

$H_o$ und $H_A$ werden formuliert, es werden der Akzeptanzbereich und die kritischen Werte festgelegt. In diesem Fall gibt es zwei kritische Werte, weil der Test zweiseitig vorgenommen wird.

$H_o$: $p = 0{,}4$
$H_A$: $p \neq 0{,}4$.

Nun wird ein Akzeptanzbereich für $H_o$ festgelegt. Dieser wird üblicherweise so gewählt, daß $\alpha$, der Fehler 1. Art, 0,05 beträgt. Bei einem zweiseitigen Test wird $\alpha$ zu gleichen Teilen auf die rechte und linke Seite der Verteilung aufgeteilt. Für n = 15 und p = 0,4 ist weder ein exakter Wert von $\alpha = 0{,}05$ noch eine symmetrische Aufteilung möglich, daher wird ein annähernd symmetrische Aufteilung mit einem Gesamt $\alpha$ in der Nähe von 0,05 gewählt.

| X = k | p(X = k) |
|---|---|
| 0 | 0,000470 |
| 1 | 0,004702 |
| 2 | 0,021942 |
| 3 | 0,063388 |
| 4 | 0,126776 |
| 5 | 0,185938 |
| 6 | 0,206598 |
| 7 | 0,177084 |
| 8 | 0,118056 |
| 9 | 0,061214 |
| 10 | 0,024486 |
| 11 | 0,007420 |
| 12 | 0,001649 |
| 13 | 0,000254 |
| 14 | 0,000024 |
| 15 | 0,000001 |

Passende $\alpha$-Werte ergeben sich für $X \leq 2$ und $X \geq 10$.

$p(X \leq 2) = 0{,}027$ (linksseitiges $\alpha$)

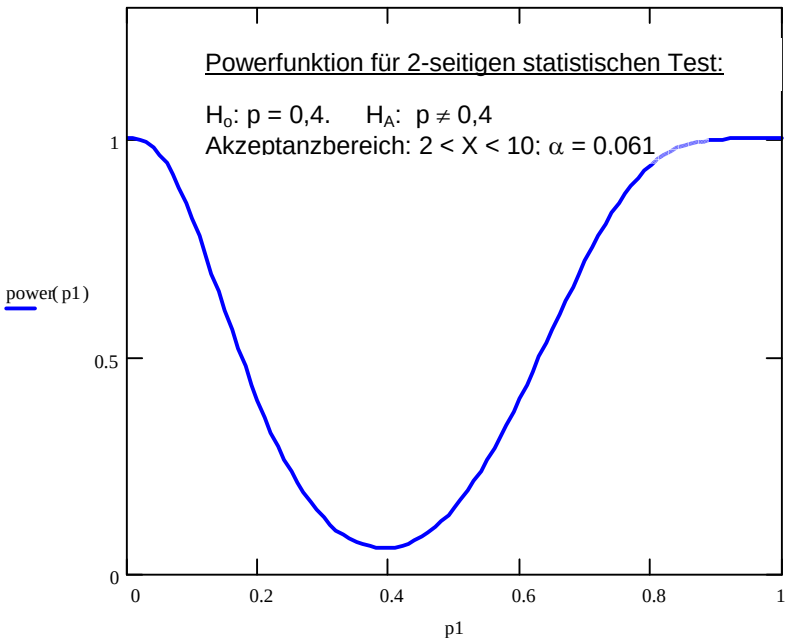
$p(X \geq 10) = 0{,}0338$ (rechtsseitiges $\alpha$)

Gesamtalpha $\alpha = 0{,}0271 + 0{,}0338 = 0{,}0609$

Nach der Festlegung des Akzeptanzbereiches kann für diesen Test mit der Hypothese Ho und mit diesem Akzeptanzbereich die Power des statistischen Tests berechnet werden. Für andere Akzeptanzbereiche erhält man eine andere Powerfunktion.

Die Powerfunktion $f(p) = 1 - \beta(p,\alpha)$;
Zur Erinnerung: $\beta$ ist die Wahrscheinlichkeit, daß unter der Gültigkeit der Alternativhypothese $H_A$ ein Stichprobenwert im Akzeptanzbereich von $H_o$ zu liegen kommt! Für die Powerberechnung sind also die Wahrscheinlichkeiten p1(X=k) über den Akzeptanzbereich zu summieren mit p1 als Variable.

$$\text{power}(p1) := 1 - \sum_{x=3}^{9} \frac{15!}{x! \cdot (15-x)!} \cdot p1^{x} \cdot (1-p1)^{15-x}$$



**Ermittlung des minimalen Stichprobenumfanges**

Die Größe der Stichprobe kann bestimmt werden, wenn eine maximale Abweichung von der Hypothese $H_o$ vorgegeben wird. Zur Erinnerung. Die Zufallsvariable

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}} \qquad (3)$$

ist t-verteilt mit FG ν: ν = n – 1. Dabei ist der Term $\overline{X} - \mu$ im Zähler die Abweichung vom Erwartungswert μ. Für das Aufstellen von Hypothesen ist diese Differenz eine Abweichung, die auf einem bestimmten α-Niveau getestet werden kann. Das heißt es lässt sich damit testen, ob bei einer vorgegebenen Irrtumswahrscheinlichkeit die festgelegte Abweichung noch innerhalb des Akzeptanzbereiches liegt oder eben nicht mehr. Damit lässt sich nun ein Konfidenzintervall für den Erwartungswert μ konstruieren. Wird eine bestimmte Signifikanzschwelle für α vorgegeben, dann wird aus dieser Gleichung eine Ungleichung, die nach n aufgelöst werden kann und ein minimales n liefert, dass die Bedingung noch erfüllt:

$$T = \frac{\overline{X} - \mu}{s / \sqrt{n}} \geq t_{\nu,\alpha}$$
$$n \geq \frac{t_{\nu,\alpha}{}^2 . s^2}{(\overline{X} - \mu)^2} \qquad (4)$$

Angenommen s=0,25, die Intervallbreite $\overline{X} - \mu$ legen wir fest mit 0,08 kg und α=10%. Aus der Berechnung von n ergibt sich eine prinzipielle Schwierigkeit, weil ja $t_{\alpha,\nu}$ von n abhängt. Wir müssen den Wert für n daher iterativ bestimmen. Für die Berechnung eines ersten Wertes kann statt $t_{\alpha,\nu}$ $z_\alpha$ aus der Standardardnormalverteilung genommen werden:

$z_{0,1} = 1{,}645$

die erste Abschätzung für n liefert $n \geq \frac{1{,}645^2 . 0{,}25^2}{0{,}08^2} = 26{,}43$

Mit n = 27 wird jetzt der entsprechende Wert aus der t Verteilung gesucht: $t_{\alpha=0,1,\nu=26} = 1{,}7056$. Damit wird n erneut berechnet:

$$n \geq \frac{1{,}7056^2 . 0{,}25^2}{0{,}08^2} = 28{,}41$$

und noch einmal für n = 29, bzw. ν=28; $t_{\alpha=0,1,\nu=28} = 1{,}7011$

$$n \geq \frac{1{,}7011^2 . 0{,}25^2}{0{,}08^2} = 28{,}26$$

Das Ergebnis ändert sich nur mehr unwesentlich. Die Mindestanzahl n ist damit n ≥ 29, um ein Konfidenzintervall der geforderten Breite zu erhalten.

**Power und Stichprobenumfang für Hypothesen der Art μ≠μo, μ>μo, μ<μo**

Soll eine Hypothese der Art μ≠μo, μ>μo, μ<μo mit einer Stichprobe überprüft werden, dann ist es wünschenswert zu wissen, wie groß die Stichprobe sein soll, um eine vorgegebene Abweichung bei gewünschter Signifikanz **und Power** des Tests erkennen zu können. Der erforderliche Umfang der Stichprobe kann angegeben werden, wenn $s^2$ als Schätzung der Varianz $\sigma^2$ der Stichprobe bekannt ist. Nun kann der Test näher spezifiziert werden: Irrtumswahrscheinlichkeiten für den Fehler 1.Art ($\alpha$) und für den Fehler 2. Art ($\beta$) werden festgelegt, sowie die maximal mögliche Differenz d zwischen $\mu$ und $\mu_o$, $d = |\mu-\mu_o|$, die unter diesen Forderungen noch erkannt werden soll. Um einen Test mit einem Signifikanzniveau von $\alpha$ mit der Power von $1-\beta$ durchzuführen, wird die minimale Stichprobengröße n:

$$n \geq \frac{s^2}{(\mu - \mu_o)^2}(t_{\alpha,\nu} + t_{\beta(1),\nu})^2$$

$\alpha$ kann dabei sowohl einseitige oder zweiseitige Irrtumswahrscheinlichkeit sein, je nachdem ob einseitig oder zweiseitig getestet wird. Wie beim Konfidenzintervall kann auch hier n nicht direkt berechnet werden, sondern nur über Iteration. Die Gleichung liefert eine bessere Abschätzung für n wenn $s^2$ eine gute Schätzung von $\sigma^2$ ist, d.h. $s^2$ sollte nicht aus einer kleinen Stichprobe geschätzt werden, sondern aus einer Stichprobe stammen vom ungefähren Umfang n.
Obenstehende Gleichung kann umgeformt werden, um bei vorgegebenem $\alpha$ und $\beta$, bzw. der Power $1-\beta$, die kleinste Abweichung d zu bestimmen, die für dieses akzeptierten Fehler noch ermittelt werden kann.

***Beispiel:*** 12 Ratten werden einem physischen Belastungstest unterzogen und es wird die Gewichtsdifferenz nach und vor dem Test ermittelt.
$s^2 = 1{,}5682g^2$;
$\bar{x} = -0{,}65g$;
n = 12

$H_o$: $\mu = 0$
$H_A$: $\mu \neq 0$

Wie groß müsste nun eine Stichprobe sein um die Nullhypothese verwerfen zu können? Der Test soll auf einem Signifikanzniveau von $\alpha$=5% mit einer 90%-igen Wahrscheinlichkeit einen Mittelwert zu erkennen, der sich von μo um nicht mehr als 1,0g unterscheidet.

Die Bestimmung von n erfolgt wieder iterativ. Man kann nun wieder z-Werten aus mit der Normalverteilung beginnen oder eine erste Schätzung für n vornehmen und mit dieser weiterrechnen.

Beginnen wir mit n = 25. $\nu$ = 24; $t_{0,05(2),24} = 2{,}064$; $\beta = 1-0{,}90 = 0{,}1$; $t_{0,10(1),24} = 1{,}325$. Nach obiger Gleichung wird n:

$$n \geq \frac{1{,}5682^2}{1{,}0^2}(2{,}064 + 1{,}325)^2 = 18{,}35$$

Im nächsten Schritt wird n = 19 als Näherung genommen und die Rechnung erneut durchgeführt.

n = 19. $\nu$ =18; $t_{0,05(2),18}$ = 2,101; $\beta$ = 1-0,90 = 0,1; $t_{0,10(1),18}$ = 1,330.

$$n \geq \frac{1{,}5682^2}{1{,}0^2}(2{,}101 + 1{,}330)^2 = 18{,}5$$

Aus diesem Ergebnis ist nun ersichtlich, dass $n \geq 18$ ist, d.h. die Stichprobe muß mindestens 19 Beobachtungen enthalten.