Übungen mit dem Applet „Vergleich von zwei Mittelwerten“

1 Statistischer Hintergrund

1.1 Typische Fragestellungen

Häufig sollen zwei Lieferchargen, Fertigungsverfahren, Prozess- oder Produktvarianten verglichen werden. Man möchte wissen, ob sich die wahren aber unbekannten Mittelwerte der beiden Gruppen (=Grundgesamtheiten) unterscheiden oder nicht. Dazu wird aus jeder Gruppe eine Stichprobe vom Umfang n entnommen (der Einfachheit halber wird hier gleicher Stichprobenumfang unterstellt – das liefert auch die besten Ergebnisse).

Die Differenz der Stichprobenmittelwerte $\overline{\Delta x}$ ist ein Schätzwert für den wahren Unterschied (die wahre Differenz) der Mittelwerte der Grundgesamtheiten $\Delta\mu$ – sie unterliegt aber der Zufallsstreuung. Daher stellt sich die Frage: Unterscheiden sich die Mittelwerte der beiden Grundgesamtheiten wirklich oder ist der beobachtete Unterschied der Stichprobenmittelwerte nur auf die Zufallsstreuung zurückzuführen?

1.2 Fehler 1. und 2. Art

Wie beim Mittelwert einer Stichprobe (vgl. Applet "Vertrauensbereich für den Mittelwert") kann man einen Vertrauensbereich für die Differenz der Mittelwerte berechnen. Dieser enthält die wahre Differenz $\Delta\mu$ mit einem vom Anwender vorgebbaren Vertrauensniveau $1-\alpha$. Nur wenn der Vertrauensbereich für $\Delta\mu$ **den Wert 0 nicht enthält**, sagt man, die beobachtete Differenz $\overline{\Delta x}$ ist **signifikant** (ragt über die Zufallstreuung hinaus). Dadurch wird gewährleistet, dass ein Ergebnis nur mit einer Wahrscheinlichkeit α (dem Signifikanzniveau) als signifikant beurteilt wird, obwohl in Wirklichkeit kein Unterschied besteht (d.h. wenn $\Delta\mu=0$) – dies ist der Fehler 1. Art.

Der Preis für diese Absicherung gegen den Fehler 1. Art ist, dass auch kleine echte Unterschiede $\Delta\mu\neq0$ in der Zufallsstreuung untergehen. Manchmal wird ein in Wirklichkeit vorhandener Unterschied nicht bemerkt – dies ist der Fehler 2. Art. Gegen den Fehler 2. Art kann man sich nur mit einem ausreichend großen Stichprobenumfang n absichern.

1.3 Kurzbeschreibung des Applets

Im Applet wird unterstellt, dass Stichproben vom gleichen Umfang n (einstellbar im Bereich 2 bis 30) aus zwei Grundgesamtheiten gezogen werden, deren Mittelwerte sich um die wahre Differenz $\Delta\mu$ unterscheiden (einstellbar im Bereich von 0 bis 10). Die wahre Standardabweichung σ ist für beide Grundgesamtheiten gleich (einstellbar im Bereich 1 bis 3), ihr Wert ist dem Anwender aber normalerweise unbekannt.

Im Applet kann zwischen zwei Darstellungen umgeschaltet werden:

- **Verteilungen** enthält die Verteilung
  - der Einzelwerte (blau, abschaltbar, Normalverteilung wird angenommen),
  - die Verteilung der "Differenz der Mittelwerte", wenn $\Delta\mu=0$ (zentriert um 0) und
  - die Verteilung der "Differenz der Mittelwerte" für den vorgegebenen Wert für $\Delta\mu$.

- **Stichprobenergebnisse** enthält die aus den Ergebnissen von je 20 Stichproben berechnete Differenz der Mittelwerte und 95%- und 99%-Vertrauensbereiche für diese Differenz.

Hinweis

Für $\Delta\mu=0$ erhält man die "Verteilung der Differenz der Mittelwerte unter Berücksichtigung Zufallsstreuung der Standardabweichung" aus einer Umskalierung der t-Verteilung (vgl. Applet "Stetige Verteilungen") mit dem Faktor $\sqrt{\frac{2}{n}} \cdot \sigma$. Für $\Delta\mu\neq 0$ erhält man die nichtzentrale t-Verteilung, die hier aber zur Vereinfachung durch eine verschobene und umskalierte t-Verteilung angenähert wird. Diese Kurven werden hier nur zur besseren Veranschaulichung verwendet, eigentlich müsste man immer durch die Standard-abweichung teilen und die t-Verteilungen direkt vergleichen.

1.4 Ziel des Applets

Das Applet soll grafisch zeigen und plausibel machen, wie der Fehler 2. Art zustande kommt und dass und warum er umso kleiner ist,

- je größer der wahre Unterschied $\Delta\mu$
- je kleiner die Standardabweichung σ
- je größer der Stichprobenumfang n und
- je kleiner das Vertrauensniveau $1-\alpha$ bzw. je größer das Signifikanzniveau α ist.

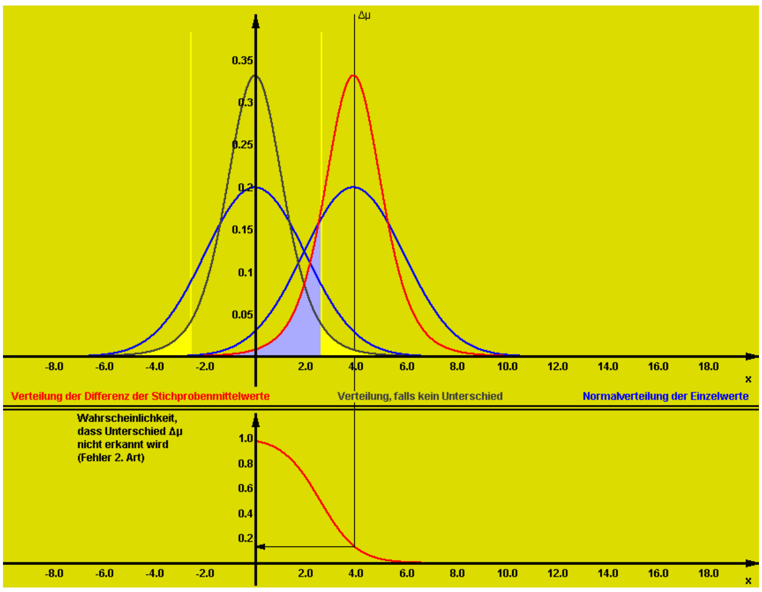
2 Visualisierungen mit dem Applet

2.1 Ausgangszustand nach dem Start des Applets

Nach dem Start des Applets stellt es im oberen Bild Verteilungen dar. Die Ausgangswerte sind

- Mittelwertunterschied $\Delta\mu = 5$
- Standardabweichung $\sigma = 2$ (daraus errechnet sich $\Delta\mu/\sigma = 5/2 = 2{,}5$)
- Stichprobenumfang $n = 16$ und
- Vertrauensniveau $1–\alpha = 0{,}9 = 90\%$.

Die beiden blauen Kurven zeigen zwei Normalverteilungen, deren Mittelwerte sich um $\Delta\mu = 5$ unterscheiden – ein Mittelwert wird willkürlich auf 0 gesetzt (fest), der andere auf $\Delta\mu = 5$ (variabel), die Mittelwerte bestimmen die Lage der Maxima. Für die weitere Betrachtung ist nur die Differenz der Mittelwerte wichtig. Die Standardabweichung (Breite bis zur steilsten Stelle) beträgt $\sigma = 2$ (variabel). Die Kurven stellen die (Dichten der) Verteilungen der Einzelwerte der beiden Grundgesamtheiten dar, die verglichen werden sollen. Die Fläche unter jeder Kurve ist 1.



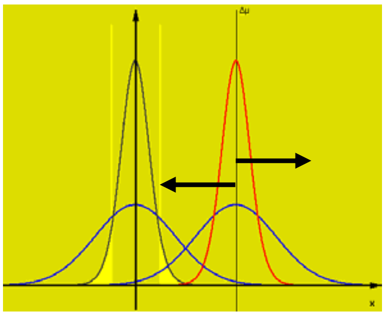
Die graue Kurve zeigt die Verteilung der "Differenz der Stichprobenmittelwerte" von zwei Stichproben des Umfangs n = 16 aus derselben Grundgesamtheit (d.h. $\Delta\mu = 0$). Aufgrund des großen Stichprobenumfangs ist sie deutlich schmäler und damit im Maximum höher als die Verteilung der Einzelwerte – die Fläche unter der Kurve ist ebenfalls 1. Das Maximum der Kurve liegt bei 0 und sie ist symmetrisch zu 0, da negative Differenzen genauso oft auftreten wie positive (der Unterschied ist ja rein zufällig).

Der innere, etwas dunklere Bereich enthält 90% der Fläche unter der Kurve, d.h. wenn zwei Stichproben des Umfangs n = 16 aus derselben Grundgesamtheit gezogen werden, so liegt die "Differenz der Mittelwerte" in 90% der Fälle in diesem Bereich – dies ergibt den 90%-Vertrauensbereich für die Differenz der Mittelwerte. Erhöhen Sie nun das Vertrauensniveau und beobachten Sie, wie der dunkle Bereich breiter wird – je höher das Vertrauensniveau, desto breiter ist der Vertrauensbereich (vgl. Applet "Vertrauensbereich für den Mittelwert").

Die rote Kurve zeigt die Verteilung der "Differenz der Stichprobenmittelwerte" für Stichproben aus Grundgesamtheiten, deren Mittelwerte sich um $\Delta\mu$ unterscheiden. Zwar erhält man aus jedem Stichprobenpaar eine andere Differenz, die Differenz ist aber praktisch immer größer als wenn $\Delta\mu = 0$ (kein Überlapp mit dem dunklen Bereich der grauen Kurve) – man erkennt fast immer, dass sich die Grundgesamtheiten unterscheiden.

2.2 Einfluss des wahren Mittelwertunterschieds $\Delta\mu$

Verändern Sie nun $\Delta\mu$ mit dem Schieber und beobachten Sie, wie sich die Verteilung der Einzelwerte (blau) und der "Differenz der Stichprobenmittelwerte" entsprechend verschieben.



Wird bei einem Vertrauensniveau von 90% $\Delta\mu$ kleiner als ca. 3,5 (alle anderen Parameter auf ihren Ausgangswerten), so schiebt sich die rote Kurve allmählich in den dunklen Bereich unter der grauen Kurve. Dies bedeutet, dass bei $\Delta\mu < 3{,}5$ auch Differenzen von Stichprobenmittelwerten auftreten können, die typisch für Stichproben sind, die sich nur zufällig unterscheiden. Aus diesen Stichprobenergebnissen ist nicht erkennbar, dass sich die Grundgesamtheiten unterscheiden – die Differenz wird als zufällig beurteilt, was natürlich falsch ist. Dies ist der Fehler 2. Art – ein tatsächlich vorhandener Unterschied wird nicht erkannt, er geht in der Zufallsstreuung unter.

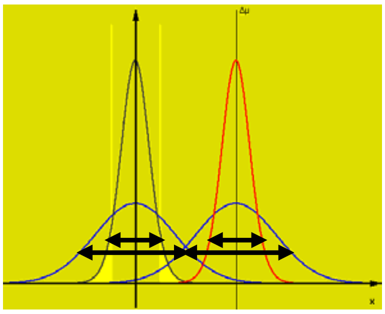
Die Wahrscheinlichkeit für den Fehler 2. Art wird durch die hellblaue Fläche gegeben – die Kurve im unteren Bild gibt ihre Größe an. Die schwarzen Linien helfen beim Ablesen. Bei $\Delta\mu = 2$ erhält man z.B. ca. 0,15 oder 15% – d.h. bei einem Vertrauensniveau von 90% (Signifikanzniveau von 10%) wird ein Unterschied von $\Delta\mu = 2 = \sigma$ nur bei ca. 85% der Stichproben mit $n = 16$ erkannt. Beachten Sie, dass die Verteilungen der Einzelwerte wesentlich stärker überlappen als die Verteilungen der Mittelwerte. Dies ist unerheblich für den Mittelwertvergleich.

Je kleiner der Unterschied $\Delta\mu$, desto größer wird der Fehler 2. Art. Bei sehr kleinen $\Delta\mu$ kann es sogar vorkommen, dass der Vergleich einen negativen Unterschied liefert, obwohl ein kleiner positiver Unterschied besteht.

(1 – Fehler 2. Art) wird auch als "Power" des Mittelwertvergleichs bezeichnet, es ist die Wahrscheinlichkeit, dass der Unterschied $\Delta\mu$ gefunden wird, falls er existiert.

2.3 Einfluss der Standardabweichung σ

Die Vergrößerung von σ verbreitert die Verteilung der Einzelwerte, damit auch die Verteilung der Mittelwerte. Damit erhöht sich die hellblaue Überlappfläche (die Wahrscheinlichkeit für den Fehler 2. Art). Die Powerkurve im unteren Bild wird breiter.

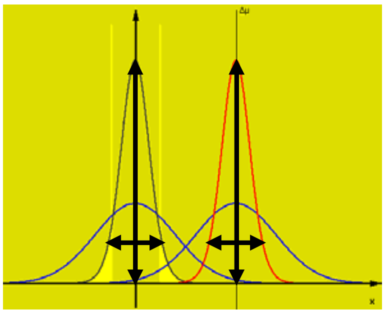


Der Fehler 2. Art hängt nur vom Verhältnis $\Delta\mu/\sigma$ ab.

Hinweis: Die y-Achse wird automatisch skaliert. Bitte achten Sie auf die Beschriftung – die Fläche unter den Kurven ist immer 1.

2.4 Einfluss des Stichprobenumfangs n

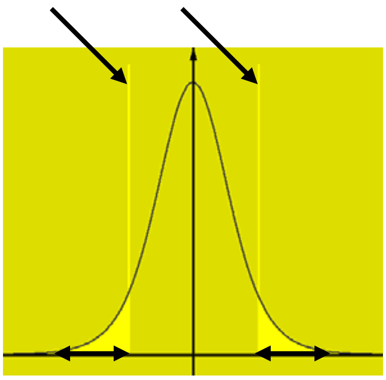
Die Vergrößerung von n hat keinen Einfluss auf die Verteilung der Einzelwerte, verschmälert aber die Verteilung der Mittelwerte (Umskalierung der y-Achse beachten). Damit verringert sich die hellblaue Überlappfläche (die Wahrscheinlichkeit für den Fehler 2. Art). Die Powerkurve im unteren Bild wird schmäler. Umgekehrtes gilt für eine Verkleinerung von n.



Je größer der Stichprobenumfang n, desto empfindlicher ist der Mittelwertvergleich. Dies kann umgekehrt verwendet werden, um aus der gewünschten oder geforderten Empfindlichkeit einen sinnvollen Stichprobenumfang zu berechnen.

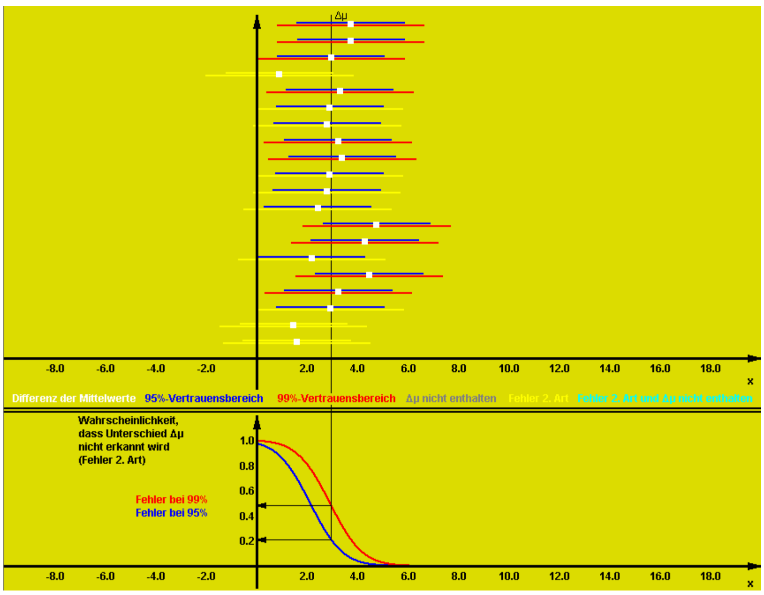
2.5 Einfluss des Vertrauensniveaus 1–α (Signifikanzniveau α)

Die Vergrößerung des Vertrauensniveaus hat keinen Einfluss auf die Verteilung der Einzelwerte und der Mittelwerte. Allerdings verschiebt sich die Signifikanzgrenze. Je höher das Vertrauensniveau, desto größer muss der beobachtete Unterschied sein, dass er als signifikant beurteilt wird. Damit erhöht sich die hellblaue Überlappfläche (die Wahrscheinlichkeit für den Fehler 2. Art). Die Powerkurve im unteren Bild wird breiter.



2.6 Darstellung der Stichprobenergebnisse

Schalten Sie nun auf "Stichprobenergebnisse" um. Diese Darstellung soll helfen, die obigen Beobachtungen zu vertiefen. Dazu werden aus jeder Grundgesamtheit 20 Stichproben gezogen. Aus jedem Stichprobenpaar wird der 95%- und der 99%-Vertrauensbereich für die Differenz der Mittelwerte berechnet. Die 99%-Vertrauensbereiche (rot) sind immer breiter als die 95%-Vertrauensbereiche (blau). "Neue Simulation" liefert neue Stichprobenergebnisse.



Je nach Streuung innerhalb der Stichproben sind manche Vertrauensbereiche breiter als andere. Je nach Differenz der Stichprobenmittelwerte unterscheiden sich die Lagen der Vertrauensbereiche. Aber 95% der 95%-Vertrauensbereiche überdecken die wahre Differenz $\Delta\mu$ (analog für 99%) – wie im Applet "Vertrauensbereich für den Mittelwert". Vertrauensbereiche, die die wahre Differenz nicht überdecken, werden hier grau oder hellblau markiert. Für $\Delta\mu=0$ wäre dies ein Fehler 1. Art.

Ein neuer Aspekt ist jetzt, dass die beobachtete Differenz nur dann signifikant ist, wenn der Vertrauensbereich nicht auch die 0 überdeckt – dann könnte das beobachtete Ergebnis ja zufällig sein. Solche Vertrauensbereiche sind gelb oder hellblau markiert (bei den hellblau markierten Stichproben treten beide Fehler gleichzeitig auf). Dies sind Stichproben, bei denen der Fehler 2. Art auftritt. Ihr Anteil wird im unteren Bild dargestellt.

Da die 99%-Vertrauensbereiche immer breiter sind als die 95%-Vertrauensbereiche, überlappen auch mehr von ihnen mit 0, d.h. Fehler 2. Art sind wahrscheinlicher (vgl. 2.5 und die rote Kurve im unteren Bild).

Ein größerer Unterschied $\Delta\mu$ zieht alle Vertrauensbereiche weiter nach rechts, weniger von ihnen überlappen mit 0 und der Fehler 2. Art ist kleiner (vgl. 2.2).

Eine größere Standardabweichung σ führt zu breiteren Vertrauensbereichen. Bei festem Unterschied $\Delta\mu$ überlappen dann mehr Vertrauensbereiche mit 0 und der Fehler 2. Art ist größer (vgl. 2.3).

Ein größerer Stichprobenumfang n führt zu schmäleren Vertrauensbereichen. Bei festem Unterschied $\Delta\mu$ überlappen dann weniger Vertrauensbereiche mit 0 und der Fehler 2. Art ist kleiner (vgl. 2.4). Ist der Stichprobenumfang zu klein, gehen fast alle Unterschiede in der Zufallsstreuung unter – man hat kaum eine Chance, einen signifikanten Unterschied zu finden.

Schalten Sie immer wieder zwischen der Darstellung der Verteilungen und der Stichprobenergebnisse hin und her, um so aus dem Vergleich immer mehr Verständnis für die Zusammenhänge zu entwickeln.