Analog kann man auch einseitige Vertrauensintervalle konstruieren. Sie enthalten alle Parameter, bei denen ein einseitiger Test nicht verwerfen würde. Beim t-Test sehen die einseitigen (1-$\alpha$)-Vertrauensintervalle so aus:

$$\text{Falls } H_A : \mu < \mu_0\text{: } (-\infty; \overline{x}_n + t_{n-1,1-\alpha} \cdot \frac{\hat{\sigma}_X}{\sqrt{n}}]$$
$$\text{Falls } H_A : \mu > \mu_0\text{: } [\overline{x}_n - t_{n-1,1-\alpha} \cdot \frac{\hat{\sigma}_X}{\sqrt{n}}; \infty)$$

---

**Beispiel A** Aggregation von Blutplättchen
Wir haben 10 Freiheitsgrade und $t_{10,0.975} = 2.23$. Das zweiseitige Konfidenzintervall für die Erhöhung der Blutplättchen-Aggregation nach dem Rauchen einer Zigarette ist somit (%-ige Zunahme)

$$I = 10.27 \pm 2.23 \cdot 7.9761/\sqrt{11} = [4.91, 15.63].$$

Insbesondere ist die Null nicht im Intervall $I$: das heisst, der Wert $\mu = 0$ ist nicht mit den Daten kompatibel (was wir bereits vom t-Test (siehe oben) wissen).

---

### 5.4.4 Tests für $\mu$ bei nicht-normalverteilten Daten

Der z- und t-Test sind optimal falls die Daten Realisierungen von normalverteilten Zufallsvariablen sind wie in (5.2). Optimalität bedeutet hier, dass dies die Tests sind, welche die beste Macht haben.

Wir betrachten hier die allgemeinere Situation, in der die Daten Realisierungen sind von

$$X_1, \ldots, X_n \text{ i.i.d. }, \tag{5.3}$$

wobei $X_i$ eine beliebige Verteilung hat. Wir bezeichen mit $\mu$ einen Lageparameter der Verteilung (z.B. $\mu$ = Median der Verteilung von $X_i$). Die Nullhypothese ist von der Form $H_0 : \mu = \mu_0$.

#### Der Vorzeichen-Test

Wir betrachten die Situation, wo die Daten Realisierungen von (5.3) sind, wobei die einzelnen $X_i$ nicht normalverteilt sein müssen. Der Vorzeichentest testet Hypothesen über den Median der Verteilung von $X_i$, den wir hier mit $\mu$ bezeichnen; im Falle einer symmetrischen Verteilung ist $\mu = \mathrm{E}(X_i)$. Wenn $\mu$ der Median der Verteilung von $X$ ist, dann ist die Wahrscheinlichkeit, dass eine Realisierung von $X$ grösser als $\mu$ ist genauso gross wie die Wahrscheinlichkeit, dass eine Realisierung von $X$ kleiner als $\mu$ ist. In anderen Worten: $P(X > \mu) = 0.5$. Der Vorzeichen-Test verwendet das folgendermassen:

1. **Modell:**

$$X_1, \ldots, X_n \text{ i.i.d. }, \tag{5.4}$$

   wobei $X_i$ eine beliebige Verteilung hat.

2. **Nullhypothese:** $H_0: \ \mu = \mu_0$, ($\mu$ ist der Median)
   **Alternative:** $H_A: \ \mu \neq \mu_0$ (oder einseitige Variante)

3. **Teststatistik:** $V$ : Anzahl $X_i$s mit $(X_i > \mu_0)$
   **Verteilung der Teststatistik unter $H_0$:** $V \sim \ \text{Bin}(n, \pi_0)$ mit $\pi_0 = 0.5$

4. **Signifikanzniveau:** $\alpha$

5. **Verwerfungsbereich für die Teststatistik:** $K = [0, c_u] \cup [c_o, n]$ falls $H_A: \ \mu \neq \mu_0$,
   Die Grenzen $c_u$ und $c_o$ müssen mit der Binomialverteilung oder der Normalapproximation berechnet werden.

6. **Testentscheid:** Entscheide, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich der Teststatistik liegt.

Vielleicht ist es Ihnen schon aufgefallen: Der Vorzeichen-Test ist nichts anderes als ein Binomialtest. Wenn wir $\mu_0 = 0$ wählen, entspricht die Teststatistik gerade der Anzahl "+" im Datensatz, daher der Name "Vorzeichentes". Wenn Sie den Binomialtest verstanden haben, müssen Sie für den Vorzeichen-Test also gar nichts neues lernen.

---

**Beispiel A** Bluttplättchen-Aggregation
Die Nullhypothese ist $H_0: \ \mu = \mu_0 = 0$. Die realisierte Teststatistik ist dann $v = 10$ und der P-Wert bei einseitiger Alternative $H_A: \ \mu > \mu_0 = 0$ ist 0.005 (beim t-Test war der P-Wert = 0.00082).

---

Der Vorzeichentest stimmt immer, falls die Daten Realisierungen von (5.3) sind: das heisst, die Wahrscheinlichkeit für einen Fehler 1. Art ist kontrolliert durch $\alpha$ bei beliebiger Verteilung der $X_i$'s.

Vom Standpunkt der Macht gibt es keine eindeutige Antwort, ob der Vorzeichen- oder der t-Test besser ist. Wenn die Verteilung der $X_i$ langschwänzig ist, kann der Vorzeichentest grössere Macht haben. Weil der Vorzeichentest die Information nicht ausnützt, um wieviel die $X_i$ von dem Wert $\mu_0$ abweichen (siehe die Definition der Teststatistik $V$ oben), kann die Macht aber auch wesentlich schlechter sein als beim t-Test.

### Der Wilcoxon-Test

Der Wilcoxon-Test ist ein Kompromiss, der keine Normalverteilung voraussetzt wie der t-Test und die Information der Daten besser ausnützt als der Vorzeichen-Test.

Die Voraussetzung für den Wilcoxon-Test ist: Die Daten sind Realisierungen von (5.3) wobei die Verteilung der $X_i$'s stetig und symmetrisch ist bezüglich $\mu = \text{E}(X_i)$. Wir verzichten auf

die Formel für die Teststatistik und die Berechnung der Verteilung der Teststatistik unter der Nullhypothese $\mu = \mu_0$, da der P-Wert mit statistischer Software berechnet werden kann.

---

**Beispiel B** Blutplättchen-Aggregation
Die Nullhypothese ist $H_0 : \ \mu = \mu_0 = 0$. Der P-Wert bei einseitiger Alternative $H_A : \ \mu > \mu_0 = 0$ ist 0.002528.

---

Der Wilcoxon-Test ist in den allermeisten Fällen vorzuziehen: er hat in vielen Situationen oftmals wesentlich grössere Macht als der t- und als der Vorzeichen-Test, und selbst in den ungünstigsten Fällen ist er nie viel schlechter.

Wenn man trotzdem den t-Test verwendet, dann sollte man die Daten auch grafisch ansehen, damit wenigstens grobe Abweichungen von der Normalverteilung entdeckt werden. Insbesondere sollte der Normal-Plot (siehe Kap. 5.2.5) angeschaut werden.

## 5.5 Tests bei zwei Stichproben

Wir besprechen hier Methoden, um einen Vergleich zweier Methoden (Gruppen, Versuchsbedingungen, Behandlungen) hinsichtlich der Lage der Verteilung machen.

### 5.5.1 Gepaarte Stichprobe

#### Struktur der Daten

Wenn möglich sollte man eine Versuchseinheit beiden Versuchbedingungen unterwerfen: Es liegt eine **gepaarte Stichprobe** vor, wenn

- beide Versuchsbedingungen an derselben Versuchseinheit eingesetzt werden
- oder jeder Versuchseinheit aus der einen Gruppe genau eine Versuchseinheit aus der anderen Gruppe zugeordnet werden kann.

Die Daten sind dann von der folgenden Struktur:

$$x_1, \ldots, x_n \text{ unter Versuchsbedingung 1,}$$
$$y_1, \ldots, y_n \text{ unter Versuchsbedingung 2.}$$

Notwendigerweise ist dann die Stichprobengrösse $n$ für beide Versuchsbedingungen dieselbe. Zudem sind $x_i$ und $y_i$ abhängig, weil die Werte von der gleichen Versuchseinheit kommen.

---

**Beispiel A**
Wir testen den Muskelzuwachs duch ein Krafttraining. Dazu messen wir die Kraft von 10 Testpersonen zu Beginn des Trainings. Anschliessend durchlaufen alle Testpersonen ein 6-wöchiges Trainingsprogramm. Dann wird die Kraft erneut gemessen. Für jede Testperson gibt es also

zwei Messungen: Vorher und nachher, die Zuordnung ist eindeutig. Somit handelt es sich um gepaarte Stichproben.

---

**Beispiel B**
Die Wirksamkeit von Augentropfen zur Reduktion des Augeninnendrucks soll untersucht werden. Wir haben 12 Patienten. Bei jedem Patienten wählen wir zufällig ein Auge aus. In dieses Auge kommen die Augentropfen mit dem Wirkstoff. In das andere Auge kommen Tropfen ohne Wirkstoff (Placebo). Für jede Testperson haben wir also zwei Messungen: Eine für das rechte und eine für das linke Auge; die Zuordnung ist eindeutig. Somit handelt es sich um gepaarte Stichproben.

---

**Beispiel C**
Wir haben eine Gruppe von 15 eineiigen Zwillingen, die sich für eine Studie für ein Haarwuchsmittel gemeldet haben. Bei jedem Zwillingspaar wird eine Person zufällig ausgewählt und erhält das Medikament. Die andere Person des Zwillingspaares erhält ein Placebo. Nach drei Wochen misst man den Haarwuchs. Zu jeder Person aus der Gruppe mit Haarwuchsmittel kann man eindeutig eine Person aus der Gruppe ohne Haarwuchsmittel zuordnen. Somit handelt es sich um gepaarte Stichproben.

---

**Beispiel D**
Datensatz zu Blutplättchen-Aggregation, siehe Kapitel 2.2.1. Die Aggregation wurde für jede Person vor und nach dem Rauchen gemessen. Die Zuordnung aller Messungen in der Gruppe "vorhe" zu den Messungen in der Gruppe"nachhe" ist also eindeutig. Somit handelt es sich um gepaarte Stichproben.

---

**Test**

Bei der Analyse von gepaarten Vergleichen arbeitet man mit den Differenzen innerhalb der Paare,

$$u_i = x_i - y_i \quad (i = 1, \ldots, n),$$

welche wir als Realisierungen von i.i.d. Zufallsvariablen $U_1, \ldots, U_n$ auffassen. Kein Unterschied zwischen den beiden Versuchsbedingungen heisst dann einfach $E[U_i] = 0$ (oder auch Median$(U_i) = 0$, je nach Test). Tests dafür sind in Kapitel 5.4 beschrieben: Falls die Daten normalverteilt sind, eignet sich ein t-Test. Sonst kommt ein Vorzeichentest oder ein Wilcoxon-Test in Frage. Dabei ist zu beachten, dass die vorausgesetzte Symmetrie für die Verteilung von $U_i$ beim Wilcoxon-Test immer gilt unter der Nullhypothese, dass $X_i$ und $Y_i$ dieselbe Verteilung haben.

### 5.5.2 Ungepaarte Stichproben

Oft ist es nicht möglich, jeder Behandlungseinheit aus der einen Gruppe eine Behandlungseinheit aus der zweiten Gruppe eindeutig zuzuordnen. In diesem Fall ist eine Paarung nicht

möglich und man spricht von einer ungepaarten Stichprobe. Auch hier muss die Zuordnung zur Behandlungsgruppe durch das Los erfolgen um systematische Fehler zu vermeiden. (vgl. Abschnitt 5.6 unten).

### Struktur der Daten

Bei ungepaarten Stichproben hat man Daten $x_1, \ldots, x_n$ und $y_1, \ldots, y_m$ (siehe Kapitel 5.5.2), welche wir als Realisierungen der folgenden Zufallsvariablen auffassen:

$$\begin{aligned} &X_1, \ldots, X_n \text{ i.i.d. ,} \\ &Y_1, \ldots, Y_m \text{ i.i.d. ,} \end{aligned} \tag{5.5}$$

wobei auch alle $X_i$'s von allen $Y_j$'s unabhängig sind.

Bei einer solchen zufälligen Zuordnung von Versuchseinheiten zu einer von zwei verschiedenen Versuchsbedingungen spricht man von einer ungepaarten Stichprobe. Im Allgemeinen ist in einer ungepaarten Stichprobe $m \neq n$, aber nicht notwendigerweise. Entscheidend ist, dass $x_i$ und $y_i$ zu verschiedenen Versuchseinheiten gehören und als unabhängig angenommen werden können.

---

**Beispiel A**
Datensatz zu latenter Schmelzwärme von Eis in Kapitel 2.2.1. Wir haben die Schmelzwärme mit zwei verschiedenen Methoden hintereinander gemessen. Jede Messung ist entweder mit Methode A oder mit Methode B, aber nicht mit beiden gleichzeitig gemacht worden. Es gibt also keinen eindeutigen Zusammenhang zwischen den Messungen der Methode A und den Messungen der Methode B. Daher sind die beiden Stichproben ungepaart.

---

**Beispiel B**
Zufällige Zuordnung von 100 Testpatienten zu einer Gruppe der Grösse 50 mit Medikamenten-Behandlung und zu einer anderen Gruppe der Grösse 50 mit Placebo-Behandlung. Es gibt keine eindeutige Zuordnung von einem Patienten aus der Medikamenten-Gruppe zu einem Patienten in der Placebo-Gruppe. Daher handelt es sich um ungepaarte Stichproben, obwohl beide Gruppen gleich gross sind.

---

### Test: Zwei-Stichproben t-Test bei gleichen Varianzen

Die beiden Stichproben können gleiche oder unterschiedliche Varianz haben. Wir behandeln nur den Fall mit gleicher Varianz im Detail und erwähnen den Fall mit ungleicher Varianz nur kurz.

Im Detail sieht der Zwei-Stichproben t-Test folgendermassen aus:

1. **Modell:**

$$\begin{aligned} &X_1, \ldots, X_n \text{ i.i.d.} \quad \sim \mathcal{N}(\mu_X, \sigma^2), \\ &Y_1, \ldots, Y_m \text{ i.i.d.} \quad \sim \mathcal{N}(\mu_Y, \sigma^2). \end{aligned} \tag{5.6}$$

2. **Nullhypothese:**

$$H_0 : \ \mu_X = \mu_Y.$$

**Alternative:**

$$\begin{aligned} H_A : \ \mu_X \neq \mu_Y \ \text{(zweiseitig)} \\ \text{oder } H_A : \ \mu_X > \mu_Y \ \text{(einseitig)} \\ \text{oder } H_A : \ \mu_X < \mu_Y \ \text{(einseitig)} \end{aligned}$$

3. **Teststatistik:**

$$T = \frac{\overline{X}_n - \overline{Y}_m}{S_{pool}\sqrt{1/n + 1/m}}$$

wobei

$$\begin{aligned} S^2_{pool} &= \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2\right) = \\ &= \frac{1}{n+m-2}\left((n-1)\hat{\sigma}^2_x + (m-1)\hat{\sigma}^2_y\right). \end{aligned}$$

**Verteilung der Teststatistik unter $H_0$:** $T \ \sim \ t_{n+m-2}$.

4. **Signifikanzniveau:** $\alpha$

5. **Verwerfungsbereich für die Teststatistik:**

$$\begin{aligned} (-\infty, -t_{n+m-2,1-\alpha/2}] \cup [t_{n+m-2,1-\alpha/2}, \infty) \quad & \text{bei Alternative } H_A : \mu_X \neq \mu_Y, \\ [t_{n+m-2,1-\alpha}, \infty) \quad & \text{bei Alternative } H_A : \mu_X > \mu_Y, \\ (-\infty, -t_{n+m-2,1-\alpha}] \quad & \text{bei Alternative } H_A : \mu_X < \mu_Y. \end{aligned}$$

6. **Testentscheid:** Entscheide, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich der Teststatistik liegt.

Die Idee des Zwei-Stichproben t-Tests ist wie folgt. Man ersetzt die unbekannte Differenz $\mu_X - \mu_Y$ durch die Schätzung $\overline{X}_n - \overline{Y}_m$ und beurteilt, ob diese Schätung "nahe bei" 0 liegt ("weit weg von" 0 würde Evidenz für $H_A$ bedeuten). Dies wird so quantifiziert, dass man durch den geschätzten Standardfehler von $\overline{X}_n - \overline{Y}_m$ dividiert und dies als Teststatistik benutzt:

$$\begin{aligned} T &= \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{\mathrm{Var}}(\overline{X}_n - \overline{Y}_m)}} \\ &= \frac{\overline{X}_n - \overline{Y}_m}{S_{pool}\sqrt{1/n + 1/m}}. \end{aligned}$$

Unter der Annahme (5.6) und der Null-Hypothese $\mu_X = \mu_Y$ gilt dann:

$$T \ \sim \ t_{n+m-2}.$$

Die Wahl des Nenners in der Teststatistik $T$ ergibt sich aus

$$\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \sigma^2(\frac{1}{n} + \frac{1}{m}). \tag{5.7}$$

Beweis von (5.7):
1. $\overline{X}_n$ und $\overline{Y}_m$ sind unabhängig, weil alle $X_i$'s von allen $Y_j$'s unabhängig sind.
2. Wegen der Unabhängigkeit von $\overline{X}_n$ und $\overline{Y}_m$ gilt:
$\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \mathrm{Var}(\overline{X}_n) + \mathrm{Var}(-\overline{Y}_m) = \mathrm{Var}(\overline{X}_n) + \mathrm{Var}(\overline{Y}_m)$.
3. $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$ und $\mathrm{Var}(\overline{Y}_m) = \sigma^2/m$.
Somit ist mit Schritt 2: $\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \sigma^2(1/n + 1/m)$. □

---

**Beispiel A** Schmelzwärme von Eis, siehe Kapitel 2.2.1.

1. **Modell:** $X$: Mit Methode A gemessene Schmelzwärme in cal/g.
   $Y$: Mit Methode B gemessene Schmelzwärme in cal/g.

$$\begin{aligned} X_1, \ldots, X_n \text{ i.i.d.} &\sim \mathcal{N}(\mu_X, \sigma^2),\ n = 13 \\ Y_1, \ldots, Y_m \text{ i.i.d.} &\sim \mathcal{N}(\mu_Y, \sigma^2),\ m = 8. \end{aligned}$$

2. **Nullhypothese:**

$$H_0 :\ \mu_X = \mu_Y.$$

   **Alternative:**

$$H_A :\ \mu_X \neq \mu_Y \text{ (zweiseitig)}$$

3. **Teststatistik:**

$$T = \frac{\overline{X}_n - \overline{Y}_m}{S_{pool}\sqrt{1/n + 1/m}}$$

   wobei

$$S^2_{pool} = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2\right).$$

   **Verteilung der Teststatistik unter $H_0$:** $T \sim t_{n+m-2}$.

4. **Signifikanzniveau:** $\alpha = 0.05$

5. **Verwerfungsbereich für die Teststatistik:**

$$(-\infty, -t_{n+m-2,1-\alpha/2}] \cup [t_{n+m-2,1-\alpha/2}, \infty) \quad \text{bei Alternative } H_A : \mu_X \neq \mu_Y,$$

6. **Testentscheid:** Zunächst berechnen wir den beobachteten Wert der Teststatistik. Für die Mittelwerte ergibt sich $\overline{x} = 80.021$, $\overline{y} = 79.979$. Für die Schätzung der Varianz ergibt sich: $s^2_{pool} = 7.253 \cdot 10^{-4}$. Damit ist der beobachtete Wert der Teststatistik:

$$t = \frac{\overline{x} - \overline{y}}{s_{pool}\sqrt{1/n + 1/m}} = \frac{80.021 - 79.979}{\sqrt{7.253 \cdot 10^{-4}} \cdot \sqrt{1/13 + 1/8}} = 3.47.$$

Nun berechnen wir den konkreten Wert des Verwerfungsbereichs der Teststatistik. Aus der Tabelle entnehmen wir $t_{n+m-2,1-\alpha/2} = t_{19,0.975} = 2.093$. Daher ist der Verwerfungsbereich der Teststatistik:

$$(-\infty, -2.093] \cup [2.093, \infty)$$

Der beobachtete Wert der Teststatistik liegt also im Verwerfungsbereich der Teststatistik. Daher wird die Nullhypothese auf dem 5% Niveau verworfen.

---

### 5.5.3 Weitere Zwei-Stichproben-Tests bei ungepaarten Stichproben

#### Zwei-Stichproben t-Test bei ungleichen Varianzen

Anstelle der Annahme in (5.6) gelte:

$$\begin{aligned} X_1, \ldots, X_n \text{ i.i.d.} &\sim \mathcal{N}(\mu_X, \sigma_X^2), \\ Y_1, \ldots, Y_m \text{ i.i.d.} &\sim \mathcal{N}(\mu_Y, \sigma_Y^2). \end{aligned}$$

Die Verallgemeinerung des Zwei-Stichproben t-Tests für ungleiche Varianzen $\sigma_X^2 \neq \sigma_Y^2$ ist in der Literatur zu finden und in vielen statistischen Programmen implementiert. In den meisten Fällen erhält man ähnliche P-Werte wie unter der Annahme von gleichen Varianzen.

#### Zwei-Stichproben Wilcoxon-Test (Mann-Whitney Test)

Die Voraussetzungen für den Zwei-Stichproben Wilcoxon-Test, manchmal auch Mann-Whitney Test genannt, bezüglich (5.5) sind wie folgt:

$$\begin{aligned} &X_1, \ldots, X_n \text{ i.i.d.} \sim F_X, \\ &Y_1, \ldots, Y_m \text{ i.i.d.} \sim F_Y, \\ &F_X \text{ beliebige stetige Verteilungsfunktion, } F_Y(x) = F_X(x - \delta). \end{aligned}$$

Dies bedeutet, dass die Verteilung von $Y_j$ die um $\delta$ verschobene Verteilung von $X_i$ ist, denn: $P(Y_j \leq x + \delta) = F_Y(x + \delta) = F_X(x + \delta - \delta) = F_X(x) = P(X_i \leq x)$.

Die Berechnung des P-Werts eines Zwei-Stichproben Wilcoxon-Tests kann mittels Computer erfolgen. Aus den gleichen Gründen wie im Fall einer Stichprobe (siehe Kapitel 5.4.4) ist der Wilcoxon-Test im Allgemeinen dem t-Test vorzuziehen.

## 5.6 Versuchsplanung

Genauso wichtig wie die Auswertung der Daten sind Überlegungen, wie man die Daten gewinnen soll. Bisher haben wir Vergleiche zwischen zwei "Behandlungen" besprochen (gepaart oder ungepaart). Allgemeiner geht es bei statistischen Studien meist darum, wie sich eine oder mehrere Einflussgrössen auf eine Zielgrösse auswirken. Die statistischen Methoden dafür werden wir im nächsten Kapitel noch kurz behandeln.

Zunächst muss man unterscheiden zwischen **Beobachtungsstudien** und **Experimenten**. Bei Beobachtungsstudien werden die Grössen von Interesse bei den **Beobachtungseinheiten** (Patienten, Tiere, Standorte, etc.) passiv erfasst oder gemessen. Bei einem Experiment hingegen werden die erklärenden Variablen vom Experimentator für jede **Versuchseinheit** aktiv festgelegt. Dazu zählen auch Studien, in denen die Versuchseinheiten (Patienten) verschiedenen Behandlungen ausgesetzt werden, sofern der Forscher die Reihenfolge der Behandlungen frei wählen kann. Experimente sind prinzipiell vorzuziehen, da sie Schlüsse auf Ursache-Wirkungs-Beziehungen zulassen.

Will man den Effekt einer Behandlung untersuchen, so braucht es eine **Kontrollgruppe** in der gleichen Studie, die sich zu Beginn der Studie möglichst wenig von der Gruppe mit der neuen Behandlung unterscheidet. Man darf also nicht Versuchseinheiten aus früheren Studien nehmen und sie als die eine Gruppe in einem Zwei-Stichproben-Vergleich verwenden. Es gibt auch oft effizientere Methoden als der Vergleich von zwei unabhängigen Stichproben für den Nachweis eines Behandlungseffekts. So ist meistens eine Versuchsanordnung, bei der sowohl die Kontrollbehandlung und die neue Behandlung auf die gleiche Versuchseinheit angewendet wird (gepaarter Vergleich) effizienter. Weitere solche Möglichkeiten sind in der Literatur unter Versuchsplanung (design of experiments) beschreiben.

Wie soll man im Zwei-Gruppen-Vergleich die Zuordnung der Versuchseinheiten zu den beiden Gruppen vornehmen ?, bzw. im gepaarten Fall: In welcher Reihenfolge soll man die beiden Behandlungen durchführen ? Um eine bewusste oder unbewusste systematische Bevorzugung der einen Gruppe zu vermeiden, soll die Zuordnung zufällig erfolgen (sogenannte **Randomisierung**). Zufällig heisst dabei nicht willkürlich, sondern mit Hilfe von Zufallszahlen oder einem physikalischen Zufallsmechanismus.

Bei human-medizinischen Studien ist es ausserdem wichtig, dass das Experiment wenn möglich **doppelblind** durchgeführt wird. Das heisst, dass weder die Person, welche die Behandlung durchführt oder deren Erfolg beurteilt, noch die Versuchsperson die Gruppenzugehörigkeit kennen. Dies ist wichtig, um den Effekt von Voreingenommenheit bei der Beurteilung auszuschalten. Weiter soll die empfangene Behandlung von allen Patienten gleich wahrgenommen werden – man gibt in der Kontrollgruppe also z.B. eine Placebo-Pille oder injiziert eine Kochsalzlösung. Damit garantiert man, dass ein allfälliger Unterschied zwischen den Gruppen wirklich auf die spezifische Behandlungsart zurückzuführen ist, und nicht etwa auf die erhöhte Aufmerksamkeit, die der Patient erfährt.

Nicht immer ist ein randomisiertes, doppelblindes Experiment möglich (aus ethischen oder praktischen Gründen), und man muss auf Beobachtungstudien zurückgreifen. Dies erschwert die Auswertung und Interpretation unter Umständen gewaltig, weil man Störeffekte praktisch nicht ausschliessen kann. Ein bekanntes Beispiel ist der Zusammenhang zwischen Rauchen und Lungenkrebs, der lange umstritten war, weil die genetische Veranlagung und der Effekt

des Lebensstils nicht auszuschliessen waren.

Mehr Wiederholungen reduzieren die Unsicherheit. Aus Kosten- und Zeitgründen will man jedoch möglichst wenige Wiederholungen machen. Die Statistik kann berechnen, wie viele Wiederholungen nötig sind, um einen Behandlungseffekt von vorgegebener Grösse mit einer vorgegebenen Wahrscheinlichkeit zu entdecken, sofern auch die Streuung bei gleicher Behandlung bekannt ist oder wenigstens abgeschätzt werden kann.

Eine Warnung vor sogenannten **Scheinwiederholungen** ist an dieser Stelle angebracht. Eine echte Wiederholung ist die kleinste Einheit, bei welcher eine Behandlung unabhängig von anderen Einheiten angewendet werden kann. Untersucht man z.B. den Effekt von einem Schadstoff im Wasser auf Fische, dann sind mehrere Fische im gleichen Aquarium Scheinwiederholungen, denn man kann bei Fischen im gleichen Aquarium die Exposition nicht individuell verändern. Ein allfälliger signifikanter Unterschied zwischen den Fischen in zwei verschieden behandelten Aquarien könnte auch durch irgendwelche unbeabsichtigte andere Unterschiede zwischen ihnen oder durch eine gegenseitige Beeinflussung der Fische im gleichen Aquarium verursacht sein. Man muss also pro Behandlung einige Aquarien haben, oder den Versuch auf geeignete Weise mehrmals wiederholen, damit man statistisch korrekt Unterschiede zwischen Behandlungen nachweisen kann. Scheinwiederholungen gibt es natürlich auch bei Beobachtungsstudien. Auch dort muss man sich fragen, was die Beobachtungseinheit ist, z.B. ein Individuum oder eine Gruppe.

## 5.7 Software

Praktisch alle Methoden, die in diesem Kapitel vorgestellt wurden, stehen in der Statistik-Software `R` (und auch in den meisten anderen üblichen Softwarepaketen) zur Verfügung. Sehen Sie sich mal die Hilfefiles der folgenden Funktionen an (wenn die Funktion z.B. `mean` heisst, dann können Sie mit `?mean` das zugehörige Hilfefile aufrufen).

### 5.7.1 Verschiedenes

Empirischer Mittelwert, empirische Varianz und empirische Standardabweichung lassen sich mit den Befehlen `mean`, `var` und `sd` berechnen. Quantile lassen sich mit der Funktion `quantile` berechnen. Für Histogramme, Boxplots, die empirische kumulative Verteilungsfunktion und Normal-Plots verwenden Sie die Funktionen `hist`, `boxplot`, `ecdf` und `qqnorm`.

Die Uniforme Verteilung, Exponentialverteilung und die Normalverteilung stehen unter den Kürzeln (siehe Kapitel 3.8) `unif`, `exp` und `norm` zur Verfügung.

### 5.7.2 Zwei-Stichproben t-Test für ungepaarte Stichproben

Sowohl der Ein-Stichproben t-Test als auch der Zwei-Stichproben t-Test (sowohl mit gleicher als auch mit ungleicher Varianz) sind unter `t.test` implementiert. Der Ein-Stichproben und Zwei-Stichproben Wilcoxon-Test ist unter `wilcox.test` verfügbar.

---

**Beispiel A** Wir berechnen den Zwei-Stichproben t-Test für ungepaarte Stichproben. Zunächst lesen wir die Daten ein:

```
> x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05,
         80.03, 80.02, 80.00, 80.02)
> y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)
```

Nun führen wir den t-Test mit durch. Wir nehmen zunächst an, dass die Streuung bei beiden Messmethoden gleich ist:

```
t.test(x, y, altervative = "two.sided", mu = 0, paired = FALSE,
       var.equal = TRUE, conf.level = 0.95)
```

Die ersten beiden Argumente enthalten die Daten der beiden Stichproben. Das Argument `alternative` gibt an, ob die Alternative einseitig (und wenn ja in welche Richtung mit `alternative =‘‘greater‘‘` und `alternative =‘‘les‘‘`) oder zweiseitig (mit `alternative = ‘‘two.sided‘‘`) ist. Das Argument `mu` gibt an, welcher Unterschied in den Mittelwerten der beiden Gruppen in der Nullhypothese getestet werden soll. Wenn man testen will, ob die beiden Gruppenmittelwerte gleich sind, ist `mu=0` die richtige Wahl. `paired = FALSE` gibt an, dass es sich um zwei ungepaarte Stichproben handelt. `var.equal = TRUE` gibt an, dass die Streuungen in den beiden Stichproben gleich gross sind. Mit `conf.level = 0.95` wird ein 95%-Vertrauensintervall des Unterschieds zwischen den beiden Gruppenmittelwerten ausgegeben.

---

Obiges Beispiel haben wir in Kapitel 5.5.2 besprochen. Am besten überzeugen Sie sich selbst davon, dass die Software tatsächlich das gleiche Resultat wie die Rechnung von Hand liefert. Die Ausgabe des Computers sieht folgendermassen aus:

```
Two Sample t-test

data:  x and y
t = 3.4722, df = 19, p-value = 0.002551
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01669058 0.06734788
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

In der Zeile `t=...` steht zunächst der beobachtete Wert der Teststatistik: $t = 3.47$. Unter der Nullhypothese folgt die Teststatistik einer t-Verteilung mit $df = 19$ Freiheitsgraden. Das ergibt bei einer zweiseitigen Alternative (siehe Zeile `alternative hypothesis: ...`) einen P-Wert von 0.002551. Der Unterschied ist also auf dem auf dem 5% Signifikanzniveau signifikant, weil der P-Wert kleiner als 5% ist. Der Computer berechnet auch das 95%-Vertrauensintervall des Unterschieds in den Gruppenmittelwerten: Mit 95% Wahrscheinlichkeit ist der Gruppenmittelwert von `x` um eine Zahl im Bereich $[0.0167, 0.0673]$ grösser als der Gruppenmittelwert von `y`

[2]. In der letzten Zeile werden schliesslich noch die Mittelwerte der beiden Gruppen angegeben. Beachten Sie, dass kein Verwerfungsbereich ausgegeben wird.

### 5.7.3 Zwei-Stichproben t-Test für gepaarte Stichproben

Einen t-Test für gepaarte Stichproben kann man leicht durchführen, indem man das Argument `paired = TRUE` verwendet.

---

**Beispiel A** Vergleichen Sie die Ergebnisse des Computers mit den Berechnungen, die wir in 5.4.2 von Hand durchgeführt haben. Sie sollten identisch sein. Zunächst lesen wir wieder die Daten ein:

```
> vorher <- c(25,25,27,44,30,67,53,53,52,60,28)
> nachher <- c(27,29,37,56,46,82,57,80,61,59,43)
```

Dann führen wir den t-Test für gepaarte Stichproben durch:

```
t.test(nachher, vorher, altervative = "two.sided", mu = 0, paired = TRUE,
       conf.level = 0.95)
```

Die Interpretation der Argumente ist wie im vorhergehenden Beispiel. Der Output ist (Interpretation ist ähnlich wie im vorhergehenden Beispiel):

```
Paired t-test

data:  nachher and vorher
t = 4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.91431 15.63114
sample estimates:
mean of the differences
               10.27273
```

Der Unterschied der Gruppenmittelwerte hat bei einer zweiseitigen Alternative (siehe Zeile `alternative hypothesis: ...`) einen P-Wert von `0.0016` und ist somit auf dem 5% Signifikanzniveau signifikant. Der Wert der Teststatistik ist `4.27` und folgt unter der Nullhypothese einer t-Verteilung mit `df = 10` Freiheitsgraden. Der Unterschied `nachher-vorher` [3] ist `10.27`. Ein 95%-Vertrauensintervall für diese Differenz ist: `[4.91, 15.63]`.

---

### 5.7.4 t-Test für eine Stichprobe

Der t-Test für nur eine Stichprobe lässt sich leicht berechnen, indem man das zweite Argument im Funktionsaufruf einfach weglässt.

[2] Die null ist nicht enthalten, also ist der Unterschied der Mittelwerte signifikant

[3] Allgemein: Erstes Argument minus zweites Argument im Funktionsaufruf

**Beispiel B** Wir testen wie in Kapitel 5.4.2, ob die Beobachtungen in Gruppe A mit der Nullhypothese $H_0 : \mu = 80.00$ verträglich ist. Zunächst wieder die Dateneingabe:

```
> x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05,
         80.03, 80.02, 80.00, 80.02)
```

Und nun der t-Test für eine Stichprobe:

```
t.test(x, altervative = "two.sided", mu = 80.00, conf.level = 0.95)
```

Der Computer liefert:

```
One Sample t-test

data:  x
t = 3.1246, df = 12, p-value = 0.008779
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
 80.00629 80.03525
sample estimates:
mean of x
 80.02077
```

Der beobachtete Wert der Teststatistik ist `3.12` und folgt unter der Nullhypothese einer t-Verteilung mit `df = 12` Freiheitsgraden. Der P-Wert mit einer zweiseitigen Alternative ist `0.008779` und ist somit auf dem 5% Signifikanzniveau signifikant. Der beobachtete Mittelwert der Daten ist `80.02`. Ein 95%-Vertrauensintervall für den wahren Mittelwert der Messungen ist `[80.006, 80.035]`.