Wahrscheinlichkeitsrechnung und
Statistik für Biologen
Versuchsplanung

Dirk Metzler

24. März 2026

Inhaltsverzeichnis

1 Warnung

Warnung

Für eine wissenschaftliche Publikation braucht man:

- Signifikanz ($\rightsquigarrow$ Stichprobenlänge groß genug?)
- Geeignete Auswahl der Stichprobe ($\rightsquigarrow$ Randomisierung)

Dies muss bei der Versuchsplanung beachtet werden!

Warnung

Erst denken, dann arbeiten!

Sonst kann wochen-/monatelange Laborarbeit vergebens sein.
Bei der Versuchsplanung (BEVOR man die Daten generiert) muss man u.a. folgende Fragen sinnvoll beantworten:

- „Wie groß muss die Stichprobe sein?“
- „An welchen Versuchsobjekten wird welche Methode angewendet?“ bzw „Wie wird die Stichprobe gesampelt?“

Um diese Fragen sinnvoll beantworten zu können, muss man sich die statistische Auswertung überlegen, BEVOR man die Daten generiert.

2 Stichprobenlänge

2.1 Allgemeines

Allgemeines
Je größer die Stichprobenlänge ist,

- desto wahrscheinlicher wird ein vorhandener Unterschied durch einen statistischen Test angezeigt
- desto kleinere Unterscheide können durch statistische Tests erkannt werden
- desto teurer wird der Versuch.

Es ist also wichtig, eine geeignete Stichprobenlänge zu wählen. Dazu muss man sich überlegen,

- welcher Unterschied durch die anzuwendenden Tests erkannt werden soll,
- wie groß die Variabilität in den Daten in etwa sein wird.

Allgemeines
Man benötigt:

- d = Unterschied, den man mindestens erkennen können möchte. (engl: detection level)
- einen ungefähren Wert s für die Standardabweichung, die man in den Daten erwartet (oft ein Wert aus Vorversuchen).
- $\alpha = \Pr_{H_0}(H_0$ wird (fälschlicherweise) abgelehnt). Meist 5%. α ist das Signifikanzniveau. Die Ws α heißt auch Fehler 1.Art.
- $\beta = \Pr_{\text{Alternative}}(H_0$ wird (fälschlicherweise) nicht verworfen). Die Wahl von β hängt stark vom Problem ab. $1-\beta$ ist die Testmacht. Die Ws β heißt auch Fehler 2.Art.

2.2 Einstichproben-Tests

Einstichproben-Tests
Frage: Ist der wahre Mittelwert gleich μ_0?
Beispiel: Kältesstress-Toleranz bei Fruchtfliegen.

Einstichproben-Tests
Die Chill-Coma Recovery Time (CCRT) ist die Zeit in Minuten, nach der die Fliege nach einem Kältekoma wieder aufwacht. In früheren Versuchen wurde bei *Drosophila ananassae* aus Bangkok eine mittlere CCRT von 46 gemessen.

Frage: Ist die CCRT bei *Drosophila ananassae* aus Kathmandu (Nepal) verschieden von 46?

Geplanter Test: (zweiseitiger) Einstichproben t-Test.

Ziel: Finde Unterschiede, die größer als $d = 4$ sind. Signifikanzniveau $\alpha = 5\%$. Testmacht $1 - \beta = 80\%$.

Vorwissen: Standardabweichung bei Vortest war $s = 11.9$

Frage: Bei wie vielen Fliegen muss ich die CCRT messen, um das Ziel zu erreichen?

Einstichproben-Tests
Frage: Stichprobenlänge für CCRT-Versuch?

Lösung: Es soll gelten:

$$n \geq \frac{s^2 \cdot (t_{1-\frac{\alpha}{2},n-1} + t_{1-\beta,n-1})^2}{d^2}$$

wobei $t_{1-\frac{\alpha}{2},n-1}$`<- qt(1-α/2,n-1)` das $(1 - \alpha/2)$-Quantil und $t_{1-\beta,n-1}$`<- qt(1-β,n-1)` das $(1 - \beta)$-Quantil der t-Verteilung ist.

Leider kannn man nicht einfach einsetzen, da die rechte Seite von n abhängt.

Entweder probiert man herum und sucht das kleinste n wofür die Ungleichung gilt.

Einstichproben-Tests
Oder man beginnt mit

$$n_0 = \frac{s^2 \cdot (z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{d^2}$$

wobei $z_{1-\frac{\alpha}{2}}$`<- qnorm(1-α/2)` das $(1 - \alpha/2)$-Quantil und $z_{1-\beta}$`<- qnorm(1-β)` das $(1 - \beta)$-Quantil der Normalverteilung ist. Die benötigte Stichprobenlänge findet man dann durch Iteration:

$$n_1 = \frac{s^2 \cdot (t_{1-\frac{\alpha}{2},n_0-1} + t_{1-\beta,n_0-1})^2}{d^2}$$
$$n_2 = \frac{s^2 \cdot (t_{1-\frac{\alpha}{2},n_1-1} + t_{1-\beta,n_1-1})^2}{d^2}$$

usw bis sich nichts mehr ändert.

Einstichproben-Tests
Zurück zum Beispiel:

$$n_0 = \frac{s^2 \cdot (z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{d^2} = \frac{11.9^2(z_{0.975} + z_{0.8})^2}{4^2} = 69.48 \approx 70$$
$$n_1 = \frac{s^2 \cdot (t_{1-\frac{\alpha}{2},n_0-1} + t_{1-\beta,n_0-1})^2}{d^2} = \frac{11.9^2(t_{0.975,69} + t_{0.8,69})^2}{4^2}$$
$$= 71.47 \approx 72$$
$$n_2 = \frac{s^2 \cdot (t_{1-\frac{\alpha}{2},n_1-1} + t_{1-\beta,n_1-1})^2}{d^2} = \frac{11.9^2(t_{0.975,71} + t_{0.8,71})^2}{4^2}$$
$$= 71.41 \approx 72$$

Antwort: Die Stichprobenlänge für den CCRT-Versuch sollte mindestens $n \geq 72$ sein.

Einstichproben-Tests
Bemerkung: Bei einer Testmacht von 80% erhält man in ca. 20% der Fälle, in denen sich die wahren Mittelwerte um d unterscheiden (also in 1 von 5 solcher Fälle), keine Signifikanz. Wenn man den Versuch 5 mal durchführt, so erhält man im Schnitt nur 4 mal Signifikanz selbst wenn der wahre Unterschied in etwa d ist.

Theoretischer Hintergrund
Angenommen, die Nullhypothese H_0 gilt nicht und der wahre Mittelwert μ_1 weicht um d vom μ_0 der H_0 ab.

Ziel: Wähle n so, dass die H_0 in diesem Fall mit Ws $\leq \beta$ nicht verworfen wird.

H_0 wird nicht verworfen falls

$$\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \leq t_{1-\frac{\alpha}{2},n-1}$$

Falls H_0 nicht wahr ist, sondern die wahre Verteilung einen Mittelwert $\mu_1 \geq \mu_0 + d$ hat, so wird H_0 mit Ws

$$\Pr_{\mu_1}\left(\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \leq t_{1-\frac{\alpha}{2},n-1}\right)$$

nicht verworfen. Diese Ws soll kleiner als β sein.

Theoretischer Hintergrund
Nun verwenden wir, dass $\frac{\overline{x}-\mu_1}{s/\sqrt{n}}$ unter $\Pr_{\mu_1}$ t-verteilt ist mit `df=n-1`:

$$\begin{aligned}
&\Pr_{\mu_1}\left(\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \leq t_{1-\frac{\alpha}{2},n-1}\right) \\
&\approx \Pr_{\mu_1}\left(\frac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq t_{1-\frac{\alpha}{2},n-1}\right) \\
&= \Pr_{\mu_1}\left(\frac{\overline{x} - \mu_1}{s/\sqrt{n}} \leq \frac{\mu_0 - \mu_1}{s/\sqrt{n}} + t_{1-\frac{\alpha}{2},n-1}\right)
\end{aligned}$$

Falls μ_1 der wahre Mittelwert ist, ist $\frac{\overline{x}-\mu_1}{s/\sqrt{n}}$ Student-t-verteilt mit $df = n-1$. Also ist obige Wahrscheinlichkeit genau dann $\leq \beta$, falls

$$\frac{\mu_0 - \mu_1}{s/\sqrt{n}} + t_{1-\frac{\alpha}{2},n-1} = t_{\beta,n-1} = -t_{1-\beta,n-1}.$$

Theoretischer Hintergrund
Dies ist $\leq \beta$, falls

$$\frac{\mu_0 - \mu_1}{s/\sqrt{n}} + t_{1-\frac{\alpha}{2},n-1} \leq t_{\beta,n-1} = -t_{1-\beta,n-1}.$$

Also muss gelten (bei Multiplikation mit $\mu_0 - \mu_1 < 0$ wird $\leq$ zu $\geq$)

$$\frac{\sqrt{n}}{s} \geq \frac{-t_{1-\beta,n-1} - t_{1-\frac{\alpha}{2},n-1}}{\mu_0 - \mu_1} = \frac{t_{1-\beta,n-1} + t_{1-\frac{\alpha}{2},n-1}}{\mu_1 - \mu_0}$$

Für $d = \mu_1 - \mu_0$ muss die Stichprobenlänge mindestens

$$n \geq \frac{s^2 \cdot (t_{1-\frac{\alpha}{2},n-1} + t_{1-\beta,n-1})^2}{d^2}$$

sein.

Einstichproben-Tests
Beispiel: Ist das Geschlechterverhältnis beim Kuhstärling bei der Geburt gleich 1 : 1?

Einstichproben-Tests
Frage: Ist die relative Häufigkeit von männlichen Kuhstärlingen bei der Geburt gleich $\frac{1}{2}$?

Geplanter Test: (zweiseitiger) Einstichproben z-Test.

Ziel: Finde Unterschiede, die größer als $d = 0.02$ sind. Signifikanzniveau $\alpha = 5\%$. Testmacht $1 - \beta = 80\%$.

Vorwissen: Nicht nötig.

Frage: Bei wie vielen neugeborenen Kuhstärlingen muss das Geschlecht bestimmt werden?

Einstichproben-Tests
Lösung: Das Geschlecht ist Bernoulli-verteilt (2 mögliche Werte) mit Standardabweichung $\sqrt{p(1-p)}$. Allerdings kennen wir p nicht. Vermutlich wird das Geschlechterverhältnis in etwa $1 : 1$ sein, also p nahe bei $\frac{1}{2}$. Als Näherung der Standardabweichung verwenden wir deshalb $s = \sqrt{\frac{1}{2}(1 - \frac{1}{2})} = \frac{1}{2}$.

Wähle n mindestens so groß, dass

$$n \geq \frac{s^2 \cdot (z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{d^2}$$

Einstichproben-Tests
Berechnung:

$$n_0 = \frac{s^2 \cdot (z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{d^2} = \frac{\frac{1}{2^2} \cdot (z_{0.975} + z_{0.8})^2}{(0.02)^2}$$
$$= 4905.55 \approx 4906$$

Die benötigte Stichprobenlänge wäre mindestens 4906! Diese Messreihe wird man vermutlich nicht durchführen wollen.

2.3 Zweistichproben-Test

Beispiel: Backenzähne von Hipparions

Beispiel: Backenzähne von Hipparions
Frage: Unterscheidet sich die mesiodistale Länge (mm) der Backenzähne von *Hipparion africanum* und *Hipparion libycum*

Geplanter Test: (zweiseitiger) ungepaarter Zweistichproben t-Test.

Ziel: Finde Unterschiede, die größer als $d = 2.5$ mm sind. Signifikanzniveau $\alpha = 5\%$. Testmacht $1 - \beta = 80\%$.

Vorwissen: Standardabweichung bei *H. africanum* ist in etwa $s_A = 2.2$, bei *H. libycum* etwa $s_L = 4.3$.

Frage: Bei wie vielen Backenzähnen muss die mesiodistale Länge gemessen werden?
Lösung: In jeder Gruppe muss die Stichprobenlänge mindestens

$$n \geq \frac{(s_A^2 + s_L^2) \cdot (t_{1-\frac{\alpha}{2},2\cdot n-2} + t_{1-\beta,2\cdot n-2})^2}{d^2}$$

sein.

Beispiel: Backenzähne von Hipparions
Berechnung:

$$\begin{aligned}
n_0 &= \frac{(s_A^2 + s_L^2) \cdot (z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{d^2} \\
&= \frac{(2.2^2 + 4.3^2) \cdot (z_{0.975} + z_{0.8})^2}{2.5^2} = 29.3 \approx 30 \\
n_1 &= \frac{(s_A^2 + s_L^2) \cdot (t_{1-\frac{\alpha}{2},2\cdot n_0-2} + t_{1-\beta,2\cdot n_0-2})^2}{d^2} \\
&= \frac{(2.2^2 + 4.3^2) \cdot (t_{0.975,58} + t_{0.8,58})^2}{(2.5)^2} \\
&= 30.3 \approx 31 \\
n_2 &= \frac{(s_A^2 + s_L^2) \cdot (t_{1-\frac{\alpha}{2},2\cdot n_1-2} + t_{1-\beta,2\cdot n_1-2})^2}{d^2} \\
&= 30.28 \approx 31
\end{aligned}$$

Beispiel: Backenzähne von Hipparions
Antwort: Es müssen mindestens 31 Backenzähne von *H. africanum* und 31 Backenzähne von *H. libycum* vermessen werden.

2.4 Einseitige Tests

Wenn man einseitig testen will, so muss man in obigen Formeln $t_{1-\frac{\alpha}{2},n-1}$ durch $t_{1-\alpha,n-1}$ ersetzen.

Beispiel: Blutdruck senkendes Medikament

Frage: Senkt das Medikament den Blutdruck signifikant stärker als ein Placebo?

Geplanter Test: einseitiger ungepaarter Zweistichproben t-Test.

Ziel: Finde mit einer **Testmacht** $1 - \beta = 80\%$ Unterschiede, die signifikant größer als $d = 10$ sind, bei einem **Signifikanzniveau** $\alpha = 5\%$.

Vorwissen: Standardabweichung ist in jeder Gruppe in etwa $s = 20$.

Frage: Wie viele Testpersonen braucht man jeweils in der Kontrollgruppe und in der Versuchsgruppe?

Lösung: In jeder Gruppe muss die Stichprobenlänge mindestens

$$n \geq \frac{(s^2 + s^2) \cdot (t_{1-\alpha,2\cdot n-2} + t_{1-\beta,2\cdot n-2})^2}{d^2}$$ sein.

Ergebnis: $n = 51$.

2.5 Übersicht

Zweiseitiger Einstichproben t-Test
Geplanter Test: Zweiseitiger Einstichproben t-Test.
Ziel: Finde Unterschiede, die größer als d sind. Signifikanzniveau α. Testmacht $1-\beta$.
Vorwissen: Standardabweichung bei Vortest war s
Lösung: Es soll gelten:

$$n \geq \frac{s^2 \cdot (t_{1-\frac{\alpha}{2},n-1} + t_{1-\beta,n-1})^2}{d^2}$$

Zweiseitiger ungepaarter Zweistichproben t-Test
Geplanter Test: Zweiseitiger ungepaarter Zweistichproben t-Test.
Ziel: Finde Unterschiede, die größer als d sind. Signifikanzniveau α. Testmacht $1-\beta$.
Vorwissen: Die Standardabweichungen in den beiden Stichproben sind in etwa s_1 beziehungsweise s_2.
Lösung: In jeder Gruppe muss die Stichprobenlänge mindestens

$$n \geq \frac{(s_1^2 + s_2^2) \cdot (t_{1-\frac{\alpha}{2},2\cdot n-2} + t_{1-\beta,2\cdot n-2})^2}{d^2}$$

sein.

Zweiseitiger gepaarter Zweistichproben t-Test
Geplanter Test: Zweiseitiger gepaarter Zweistichproben t-Test.
Ziel: Finde Unterschiede, die größer als d sind. Signifikanzniveau α. Testmacht $1-\beta$.
Vorwissen: Standardabweichung der Differenz der beiden Stichproben ist in etwa s_d.
Lösung: In jeder Gruppe muss die Stichprobenlänge mindestens

$$n \geq \frac{s_d^2 \cdot (t_{1-\frac{\alpha}{2},n-1} + t_{1-\beta,n-1})^2}{d^2}$$

sein.

Einseitiger Einstichproben t-Test
Geplanter Test: Einseitiger Einstichproben t-Test.
Ziel: Finde Unterschiede, die größer als d sind. Signifikanzniveau α. Testmacht $1-\beta$.
Vorwissen: Standardabweichung bei Vortest war s
Lösung: Es soll gelten:

$$n \geq \frac{s^2 \cdot (t_{1-\alpha,n-1} + t_{1-\beta,n-1})^2}{d^2}$$

Einseitiger ungepaarter Zweistichproben t-Test
Geplanter Test: Einseitiger ungepaarter Zweistichproben t-Test.
Ziel: Finde Unterschiede, die größer als d sind. Signifikanzniveau α. Testmacht $1-\beta$.
Vorwissen: Die Standardabweichungen in den beiden Stichproben sind in etwa s_1 und s_2.
Lösung: In jeder Gruppe muss die Stichprobenlänge mindestens

$$n \geq \frac{(s_1^2 + s_2^2) \cdot (t_{1-\alpha,2\cdot n-2} + t_{1-\beta,2\cdot n-2})^2}{d^2}$$

sein.

Einseitiger gepaarter Zweistichproben t-Test
Geplanter Test: Einseitiger gepaarter Zweistichproben t-Test.
Ziel: Finde Unterschiede, die größer als d sind. Signifikanzniveau α. Testmacht $1-\beta$.
Vorwissen: Standardabweichung der Differenz der beiden Stichproben ist in etwa s_d.
Lösung: In jeder Gruppe muss die Stichprobenlänge mindestens

$$n \geq \frac{s_d^2 \cdot (t_{1-\alpha,n-1} + t_{1-\beta,n-1})^2}{d^2}$$

sein.

2.6 Stichprobenlänge ermitteln mit `R`

In `R` ermittelt man die benötigte Stichprobenlänge mit

```
power.t.test(n = , delta = , sd = , sig.level = ,
         power = ,
         type = c("two.sample","one.sample","paired"),
         alternative = c("two.sided", "one.sided") )
```

Die Argumente sind:

- `n =` Stichprobenlänge (pro Gruppe bzw pro Stichprobe)
- `delta =` d (minimale Differenz, detection level)
- `sd =` s (vermutete Standardabweichung pro Gruppe)
- `sig.level =` α (Signifikanzniveau)
- `power =` $1-\beta$ (Testmacht)

Genau eines der Argumente `n,delta,sd,sig.level,power` muss als `NULL` übergeben werden. Dieses wird dann berechnet.
Beispiele:

- CCRT bei *D. ananassae*: $d = 4$, $s = 11.9$, $\alpha = 5\%$, $\beta = 0.2$

```
> power.t.test(n=NULL, delta=4, sd=11.9,
+ sig.level=0.05, power=0.8,
+ type="one.sample", alternative="two.sided")


     One-sample t test power calculation

              n = 71.41203
          delta = 4
             sd = 11.9
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

- Relative Häufigkeit von männlichen Kuhstärlingen bei der Geburt: $d = 0.02$, $s = \sqrt{\frac{1}{2}(1-\frac{1}{2})}$, $\alpha = 5\%$, $\beta = 0.2$

```
> power.t.test(n=NULL, delta=0.02, sd=0.5,
+ sig.level=0.05, power=0.8,
+ type="one.sample", alternative="two.sided")

   One-sample t test power calculation

              n = 4907.471
          delta = 0.02
             sd = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

(wir verwenden `power.t.test` als Approximation, da `power.z.test` nicht existiert)

- Backenzähne von Hipparions: $d = 2.5$, $s = \sqrt{(2.2^2 + 4.3^2)/2}$, $\alpha = 5\%$, $\beta = 0.2$

```
> power.t.test(n=NULL, delta=2.5,
+ sd=sqrt( (2.2^2+4.3^2)/2 ),
+ sig.level=0.05, power=0.8,
+ type="two.sample", alternative="two.sided")

   Two-sample t test power calculation

              n = 30.28929
          delta = 2.5
             sd = 3.415406
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

- Blutdruck senkendes Medikament: (einseitiger Test) $d = 2$, $s = 4$, $\alpha = 5\%$, $\beta = 0.2$

```
> power.t.test(n=NULL, delta=2, sd=4,
+ sig.level=0.05, power=0.8,
+ type="two.sample", alternative="one.sided")

   Two-sample t test power calculation

              n = 50.1508
          delta = 2
             sd = 4
      sig.level = 0.05
          power = 0.8
    alternative = one.sided

NOTE: n is number in *each* group
```

Der Befehl `power.t.test()` kann auch dazu benutzt werden, die Testmacht zu berechnen, wenn man sich auf die Stichprobenlänge bereits festgelegt hat. Beispiel: CCRT bei *D. ananassae*: $n = 100$, $d = 4$, $s = 11.9$, $\alpha = 5\%$

```
> power.t.test(n=100, delta=4, sd=11.9,
+ sig.level=0.05, power=NULL,
+ type="one.sample", alternative="two.sided")

     One-sample t test power calculation

              n = 100
          delta = 4
             sd = 11.9
      sig.level = 0.05
          power = 0.9144375
    alternative = two.sided
```

2.7 F-Test

F-Test

Will man testen, ob die Mittelwerte bei 3 oder mehr Gruppen gleich sind, so verwendet man den F-Test. Um eine Aussage über die Stichprobenlänge treffen zu können, benötigt man die Variabilität innerhalb der Gruppen und die Variabilität zwischen den Gruppen (z.B. aus Vorversuchen).

Die Formel für die benötigte Stichprobe ist hier weniger übersichtlich. Deshalb konzentrieren wir uns auf die Berechnung mit `R`.

Wir zeigen an folgendem Beispiel, wie man den `R`-Befehl `power.anova.test()` einsetzt, um die benötigte Stichprobenlänge zu ermitteln.

Beispiel: Blutgerinnungszeit bei Ratten
Frage: Unterscheidet sich die Blutgerinnungszeit bei Ratten unter 4 verschiedenen Behandlungen?

Geplanter Test: F-Test.

Signifikanzniveau: $\alpha = 5\%$

Testmacht: $1 - \beta = 90\%$.

Vorwissen: Standardabweichung innerhalb jeder Gruppe ist in etwa $s_{\text{innerh}} = 2.4$. Beachte: $s^2_{\text{innerh}} = ss_{\text{innerh}}/\,\text{df}_{\text{innerh}}$. Standardabweichung zwischen den Gruppen ist in etwa $s_{\text{zw}} = 1.2$. Beachte: $s^2_{\text{zw}} = ss_{\text{zw}}/\,\text{df}_{\text{zw}}$.

Frage: Bei wie vielen Ratten muss die Blutgerinnungszeit gemessen werden?

Beispiel: Blutgerinnungszeit bei Ratten

```
> power.anova.test(groups=4, n=NULL, between.var=1.2^2,
           within.var=2.4^2, sig.level=0.05, power=0.9)

 Balanced one-way analysis of variance power calculation

      groups = 4
           n = 19.90248
```

```
 between.var = 1.44
  within.var = 5.76
   sig.level = 0.05
       power = 0.9
NOTE: n is number in each group
```

Antwort: Für jede der 4 Behandlungen braucht man mindestens 20 Ratten.

3 Stichprobenwahl

3.1 Überspitzte Beispiele

Um die Problematik der Stichprobenwahl zu verdeutlichen, beginnen wir überspitzten Beispielen.

- Um die Parteienpräferenz in Deutschland zu messen, stellt ein Wahlforschungsunternehmen die Sonntagsfrage („Was würden Sie wählen, wenn kommenden Sonntag Bundestagswahl wäre“) an 1000 zufällig ausgewählte Bürger aus Garmisch-Partenkirchen.[5mm] Keine repräsentative Stichprobe![5mm] Die Einwohner- und Meinungsstruktur von Garmisch-Partenkirchen ist möglicherweise nicht typisch für Deutschland.

- Um die Chill-Coma Recovery Time (CCRT) der europäischen *Drosophila melanogaster* mit der taiwanesischen Population zu vergleichen, werden *Drosophila melanogaster* an jeweils 10 verschiedenen Orten in Frankreich, Spanien und Italien gesampelt.[4mm] Keine repräsentative Stichprobe![4mm] Die CCRT von Fruchtfliegen in Südeuropa ist nicht typisch für die CCRT europäischer Fruchtfliegen.

- Um die Blätterdichte in oberbayerischen Wäldern zu messen, wird in 10 zufällig ausgewählten oberbayerischen Wäldern die Blätterdichte entlang des Waldrandes und entlang von Waldwegen gemessen.[3mm] Keine repräsentative Stichprobe![3mm] Am Waldrand und auch entlang von Waldwegen ist die Blätterdichte überdurchschnittlich hoch.

- Waldameisen

Es sollen 100 französische Waldameisen gesampelt werden. Dazu wird ein Ameisennest zufällig in Frankreich ausgewählt und hiervon 100 Ameisen genommen.[5mm] Keine repräsentative Stichprobe der Länge 100![5mm] Die erste gesampelte Ameise ist wohl eine typische französische Waldameise. Die weiteren sind aber vermutlich mit der ersten Ameise nahe verwandt. Für eine Stichprobe der Länge 100 braucht man 100 'unabhängige' Ameisen. Kommen die 100 Ameisen aus demselben Ameisennest, so können sie Geschwister sein und sind dann sicherlich nicht unabhängig voneinander.

- 20 zufällig ausgewählte Studierenden werden eingeladen, an einem Versuch teilzunehmen. Die ersten 10 Studierenden, die am Versuchsort ankommen, bilden die Kontrollgruppe. Die weiteren 10 Studierenden bilden die Versuchsgruppe.[5mm] Die beiden Versuchsgruppen sind nicht identisch verteilt![5mm] Die Kontrollgruppe besteht aus pünktlicheren Studierenden. Diese Gruppe könnte engagierter am Versuch teilnehmen. Dadurch wird das Ergebnis verfälscht.

Dr. X könnte argumentieren: *Wir haben mit einem zusätzlichen Test gezeigt, dass es die Reihenfolge der studierenden keinen Einfluss auf die Versuchsergebnisse hatte.*

Was halten Sie von diesem Argument?

Eine solche Argumentation ist aus statistischer Sicht Blödsinn!

- Ein statistischer Test kann niemals zeigen, dass ein Effekt nicht existiert.
- Vermutlich meint Dr. X, dass er einen statistischen Test durchgeführt hat, bei dem es keinen statistisch signifikanten Zusammenhang zwischen Pünktlichkeit und Versuchsergebnis gab.
- Man darf aber aus nicht-Signifikanz niemals schließen, dass es den Effekt nicht gibt.
- Vielleicht ist der Effekt so schwach, dass der Vortest geringe Macht hatte, aber immer noch stark genug um die spätere statistische Analyse zu verfälschen.

3.2 Zufallsstichprobe

Zufallsstichprobe

Eine Zufallsstichprobe der Länge n aus einer Gesamtpopulation der Größe N erhält man wie folgt:

- Nummeriere N identische Kugeln von 1 bis N.
- Durchmische die N Kugeln in einem Beutel oder ähnlichem.
- Ziehe (ohne Zurücklegen) n Kugeln.

Die zu den Nummern auf den Kugeln gehörigen Individuen in der Gesamtpopulation biden dann eine Zufallsstichprobe.

Beispiel

Ziel: Man möchte eine Umfrage unter allen Bachelor-Studierenden der Biologie an der LMU München durchführen. Es zu aufwändig ist, alle Studierenden zu befragen. Also möchte man 50 Studierenden zufällig auswählen, um diese dann zu befragen.

Vorgehen: Die Anzahl N an Studierenden ist bekannt. Nun nummerieren wir die Studierenden durch und ziehen 50 Nummern rein zufällig. Dies könnte man in `R` durchführen:

```
sample(1:N, size=50, replace=FALSE)
```

Dieses Vorgehen wird oft als Lotterieverfahren bezeichnet.
In Anwendungen ist dies meist nicht möglich, da

- die Größe der Gesamtpopulation meist unbekannt ist (zB: Anzahl an Ameisen, Anzahl an Drosophila melanogaster)
- beziehungsweise es bei großen Populationen schwierig ist, den Individuen Nummern zuzuweisen.

Eine Zufallsstichprobe ist Teil einer Gesamtpopulation, die durch einen Auswahlprozess mit Zufallsprinzip aus der Gesamtpopulation entnommen wird und stellvertretend, repräsentativ für die Gesamtpopulation ist.

Ein Teil einer Gesamtpopulation kann auch dann als repräsentative Stichprobe angesehen werden, wenn das Auswahlverfahren zwar nicht zufällig, aber von den auszuwertenden Merkmalen stochastisch unabhängig ist.

Anders formuliert: Die Stichprobe muss bezüglich den auszuwertenden Merkmalen typisch für die Gesamtpopulation sein.

Betrachtet man eine „Stichprobe, die gerade zur Hand ist“ und die keine Zufallsstichprobe ist, so darf man Aussagen über die Stichprobe nicht auf die Gesamtpopulation verallgemeinern.

Beispiel
Ziel: Stichprobe von 100 Mäusen.
Beachte: Für die statistische Analyse wird Unabhängigkeit vorausgesetzt. Insbesondere dürfen die Mäuse nicht verwandt sein.

Falsch: 100 Mäuse von demselben Bauernhof. Denn: Von demselben Bauernhof sammelt man mit gewisser Ws verwandte Mäuse. Extremfall: Nimmt man 100 Klone derselben Maus, so ist die tatsächliche Stichprobenlänge gleich 1 (= Anzahl voneinander unabhängiger Mäuse).

Richtig: (Wird jedenfalls in der Literatur akzeptiert)

- Je eine Maus pro Bauernhof.
- Bauernhöfe müssen mindestens 1km voneinander entfernt sein.

Beispiel
Beachte: Sampelt man Mäuse von verschiedenen Bauernhöfen in der Gegend von Memmingen, so ist die Stichprobe nur repräsentativ für die Region Memmingen.

Es darf bezweifelt werden, ob diese Stichprobe repräsentativ für Deutschland oder gar Europa ist.

3.3 Elimination von nicht-interessierenden Einflussgrößen

Nun geht es nicht mehr um Zufallsstichproben, sondern um die Einteilung von Versuchsobjekten in verschiedene Behandlungsgruppen.

Prinzipien der Versuchsplanung
Wir sprechen nun von Einflussgrößen bzw von Einflussfaktoren und von Zielgrößen.
Einflussgröße kann so ziemlich alles sein:

- Wurde die Behandlung angewendet: Ja oder Nein?
- Wer hat die Messung durchgeführt?
- Wurde ein großes oder kleines Reagenzglass verwendet?
- Wir waren die Lichtverhältnisse im Labor während des Versuchs?

Prinzip

Nicht interessierende Einflussgrößen sind im Versuch möglichst konstant zu halten.

Prinzipien der Versuchsplanung

Beispiele für die Einhaltung dieses Prinzips:

- Derselbe Experimentator für alle Versuche.
- Doppelblind: Weder Experimentator, der den Effekt misst (z.B. diagnostizierender Arzt), noch Versuchsperson wissen, zu welcher Behandlungsgruppe die Versuchsperson gehört. (Ausschluss von subjektiven Einflussfaktoren).
- Dieselben oder zumindest baugleiche Materialien und Laborbedingungen bei allen Versuchen.
- Reihenfolge der Behandlungsgruppen ist zufällig. (Also nicht: Versuchsgruppe, Kontrollgruppe, Versuchsgruppe, Kontrollgruppe, ...

3.4 Blockbildung

Sind die Versuchsobjekte sehr unterschiedlich, so empfiehlt sich eine Zusammenfassung von sehr ähnlichen Versuchsobjekten zu Untergruppen. Die für das Versuchsziel wichtigen Vergleiche werden dann möglichst innerhalb der Blöcke vorgenommen.

Beachte: Die Bildung von Blöcken ist nur dann sinnvoll, wenn die Streuung zwischen den Versuchsobjekten deutlich größer ist als die Streuung zwischen den verschiedenen Behandlungen.

Zweck der Blockbildung ist es, die Genauigkeit blockinterner Vergleiche zu erhöhen.

Beispiel

Frage: Wirkt eine gewisse Diät besser als Placebo?

Problem: Nehmen wir, die Diät verringert das Gewicht tatsächlich im Mittel um 3 kg. Da das Gewicht bei den Versuchspersonen aber sehr stark zwischen 50 kg und 130 kg schwankt, braucht man sehr viele Versuchspersonen, um den kleinen Unterschied festzustellen.

Lösung: Unterteile die Versuchspersonen in Untergruppen gleicher Gruppengröße, so dass die Personen in jeder Untergruppe ähnliches Gewicht haben. Jede Untergruppe wird dann in Diätgruppe und Kontrollgruppe aufgeteilt. Die Gewichtsvergleiche finden dann in jeder Untergruppe statt.

Beispiel: Experimentatoreneffekt

Die vier Bio-Studierenden Lukas, Leon, Laura und Lisa sollen untersuchen, wie unterschiedlich sich vier verschiedene Nährmedien A, B, C, D auf das Wachstum von je 20 Zellkulturen auswirken.

Problem: Vielleicht gibt es Unterschiede zwischen den vier Studierenden z.B. bei der Geschicklichkeit beim Pipettieren.

Ganz falsch: Lukas behandelt die 20 Zellkulturen mit A, Leon 20 mit B, Laura 20 mit C und Lisa 20 mit D. Wenn es signifikante Unterschiede zwischen A und B gibt, kann man nicht ausschließen, dass es nur daran lag, dass Lukas und Leon unterschiedlich gearbeitet haben.

Lösung: Jede(r) behandelt für jedes Nährmedium jeweils 5 Zellkulturen. Die das Experiment durchführende Person wird jeweils vermerkt und mögliche Experimentatoreneffekte werden in der Analyse berücksichtigt, z.B. als Faktor bei einer Varianzanalyse oder einem linearen Modell, oder durch Blockbildung herausgemittelt.

Beispiel: Experimentatoreneffekt, Variante

Die vier Bio-Studierenden Lukas, Leon, Laura und Lisa sollen untersuchen, wie unterschiedlich sich vier verschiedene Behandlungen A, B, C, D auf das Wachstum von je 20 Zellkulturen auswirken. Da sich die vier Behandlungen sehr unterschiedlich und kompliziert sind, können Studierende nur jeweils zwei Arten der Behandlung erlernen und durchführen.

Falsch: Lukas und Laura behandelen jeweils 10 Zellkulturen mit A und jeweils 10 mit B und Leon und Lisa behandeln jeweils 10 mit C und 10 mit D. Wenn die mit A und B behandelten signifikant anders sind als die mit C und D behandelten, kann man nicht ausschließen, dass es nur daran lag, dass Lukas und Laura anders gearbeitet haben als Lisa und Leon.

Besser: Lukas behandlt 10 mit A und 10 mit B, Laura behandelt 10 mit C und 10 mit D, Leon 10 mit A und 10 mit C und Lisa 10 mit B und 10 mit D. Experimentatoreneffekte können in der Analyse berücksichtigt und besser von Behandlungseffkten unterschieden werden.

Wir simulieren, dass Behandlung A und B einen Effekt hatten, aber nicht die Experimentatoren, die die Effekte durchgeführt haben:

```
treatment <- rep(c("A","B","C","D"),each=20)
scientist <- rep(c(rep(c("Lukas","Laura"),2),
                   rep(c("Leon","Lisa"),2)),
                 each=10)
obs <- round(rnorm(80,mean=40,sd=3) +
              10 * (treatment == "A" | treatment =="B"),2)
```

```
> data.frame(obs, treatment, scientist)
    obs treatment scientist
1  50.02         A     Lukas
2  51.16         A     Lukas
.                .
.                .
10 53.45         A     Lukas
11 52.98         A     Laura
.                .     .
.                .     .
20 46.92         A     Laura
21 47.87         B     Lukas
.                .     .
.                .     .
30 52.76         B     Lukas
31 48.27         B     Laura
.                .     .
.                .     .
40 46.93         B     Laura
41 40.35         C      Leon
.                .     .
.                .     .
50 39.79         C      Leon
51 44.33         C      Lisa
.                .     .
.                .     .
60 35.67         C      Lisa
61 42.11         D      Leon
.                .     .
.                .     .
70 36.14         D      Leon
71 42.36         D      Lisa
.                .     .
.                .     .
80 38.88         D      Lisa
```

Die Varianzanalyse kann keinen signifikanten Effekt der Behandlung erkennen, da es sich auch um einen Experimentatoreneffekt handeln könnte:

```
> drop1(lm(obs~treatment+scientist),test="F")
Single term deletions

Model:
obs ~ treatment + scientist
          Df Sum of Sq    RSS    AIC F value Pr(>F)
```

```
<none>                  579.58 170.42
treatment  2     7.957 587.53 167.51  0.5080 0.6038
scientist  2    18.878 598.45 168.99  1.2051 0.3055
```

Jetzt gehen wir davon aus, dass sich die Experimentatoren klüger auf die Versuche verteilt haben:

```
> scientist <- rep(c("Lukas", "Leon",
+                    "Lukas", "Lisa",
+                    "Laura", "Leon",
+                    "Laura", "Lisa"),
+                  each=10)
```

(Wir können die selben Daten verwenden, da wir ohnehin keinen Experimentatoreneffekt simuliert haben.)

```
> data.frame(obs, treatment, scientist)
     obs treatment scientist
1  50.02         A     Lukas
2  51.16         A     Lukas
.                .         .
.                .         .
10 53.45         A     Lukas
11 52.98         A      Leon
.                .         .
.                .         .
20 46.92         A      Leon
21 47.87         B     Lukas
.                .         .
.                .         .
30 52.76         B     Lukas
31 48.27         B      Lisa
.                .         .
.                .         .
40 46.93         B      Lisa
41 40.35         C     Laura
.                .         .
.                .         .
50 39.79         C     Laura
51 44.33         C      Leon
.                .         .
.                .         .
60 35.67         C      Leon
61 42.11         D     Laura
.                .         .
.                .         .
70 36.14         D     Laura
71 42.36         D      Lisa
.                .         .
.                .         .
80 38.88         D      Lisa
```

Jetzt kann die Varianzanalyse den Behandlungseffekt von einem möglichen Experimentatoreneffekt unterscheiden:

```
> drop1(lm(obs~treatment+scientist),test="F")
Single term deletions

Model:
obs ~ treatment + scientist
          Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>                  569.69 171.04
treatment  3   1028.22 1597.91 247.55  43.919 2.492e-16 ***
scientist  3     15.22  584.91 167.15   0.650    0.5855
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3.5 Balanced Design vs Non-Balanced Design

Balanciertes Design bedeutet, dass jede Gruppe aus gleich vielen Versuchsobjekten besteht. In jeder Behandlungsgruppe hat man also dieselbe Stichprobenlänge.

Im Normalfall bevorzugt man ein balanciertes Versuchs-Design

Vorteil des balancierten Versuchs-Designs:

- Die Effekte korrelierter Einflussfaktoren, z.B. von Geschlecht und Körpergröße, lassen sich trennen.
- Manche statistische Verfahren setzen balanciertes Design voraus (z.B Tukey's simultane Konfidenzintervalle).

Nachteil des balancierten Versuchs-Designs: Eine balanciertes Design ist in der Regel nicht repräsentativ.

Beispiel: Die untypische Gewichtsklasse $140 - 150$ kg wird im balancierten Design überrepräsentiert.

3.6 Randomisierung

Randomisierung

Randomisierung ist die zufällige Zuordnung der Behandlungen zu den gegebenen Versuchsobjekten.

Vorgehen: Nummeriere die Versuchsobjekte und wende das Lotterieverfahren an.
Beispiel: Ein Medikament zur Steigerung der Konzentration soll getestet werden an 20 Studierenden.

Falsch: Die 10 Studierenden, die zuerst im Labor eintreffen, bekommen das Medikament. Die nächsten 10 Studierenden bekommen das Placebo. Problem hier: Pünktlichere Studierenden können sich vielleicht von vornherein besser konzentrieren.

Richtig: Die Studierenden werden von 1 bis 20 durchnummeriert. Die Kontrollgruppe besteht dann aus den Studierenden mit Nummern

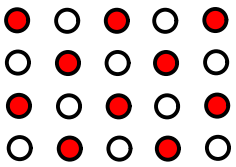
```
sample(1:20,size=10,replace=FALSE)
19 16  1 13 18 10  2  5  9 14
```

(Natürlich gibt es viele weitere Verfahren, eine Zufallszuordnung zu erreichen.)
Braucht man wirklich einen Zufallsgenerator oder kann man auch einfach eine beliebige Reihenfolge wählen?

Problem: Ein von Menschen erdachter Pseudo-Zufall ist oft nicht zufällig genug, siehe z.B.

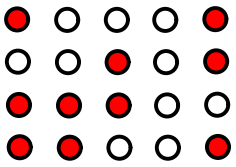
David F. Marks and John Colwell (2000) The Psychic Staring Effect: An Artifact of Pseudo Randomization. *Skeptical Inquirer*

Selbst Computer können in der Regel nur Pseudo-Zufall generieren, aber je nach Anwendungsgebiet stehen unterschiedlich sorgfältige Verfahren zur Verfügung.
Beispiel: Räumliche Anordnung von Behandlungsgruppen, etwa Pflanzen auf einem Feld (oder Reaktionsgefäße in einem Rack).
Problem: es könnte räumliche Effekte geben, z.B. unterschiede in der Bodenqualität zwischen verschiedenen Bereichen eines Feldes.
Mögliche Anordnung bei zwei Behandlungsgruppen mit je 10 Pflanzen (oder Reaktionsgefäßen):



Vorteil: großflächige Schwankung in der Bodenqualität sollten zwischen den Gruppen ausgeglichen sein

Problem: schachbrettartige Schwankungen theoretisch möglich, z.B. durch Art wie das Feld gepflügt oder bewässert wurde (vielleicht unplausibel, aber als Einwand schwer zu entkräften).
Besser? Kommt darauf an...



Falls von Experimentator so nach Gutdünken gesetzt:
sehr schlecht! Einwand: Könnte bei Auswahl bewusst oder unbewusst von Bodenqualität beeinflusst gewesen sein.
Besser? Kommt darauf an...

```
> x <- rep(c(0,1),10)
> matrix(sample(x),nrow=4)
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    0    0    0    1
[2,]    0    0    1    0    1
[3,]    1    1    1    0    0
[4,]    1    1    0    0    1
```

Falls randomisiert erzeugt (siehe R-Code):
sehr gut! Einwände können mit stochastischer Argumentation zurückgewiesen werden.

Stoch. Argumentation bei Randomisierung
Seien

$v_1, \ldots, v_{20}$ die Effekte der Positionen auf die Zielvariable.

$J(1), \ldots, J(20)$ die zufälligen Positionen der 20 Pflanzen.

$Z_i = v_{J(i)}$ Der Effekt and der Position von Pflanze i.

μ_0, μ_1 die Effekte der beiden Behandlungen auf die Pflanzen,d.h.

$\mathcal{N}(\mu_j, \sigma^2)$ wäre bei typischem t-Test-Szenario die Verteilung der Zielvariablen, falls es keinen Effekt der Position gibt.

Da $J(i)$ zufällig ist, ist auch Z_i zufällig, und sei σ_Z^2 die Varianz von Z_i.
Beobachtete Werte Y_i aus Gruppe j: $\mathbb{E}Y_i = \mu_j + \overline{v}$, $\mathrm{Var}(Y_i) = \sigma^2 + \sigma_Z^2$
Führt man also einen t-Test durch, testet man, ob

$$\mu_0 + \overline{v} = \mu_1 + \overline{v},$$

und das ist äquivalent zu unserer eigentlichen Frage, ob $\mu_0 = \mu_1$.

Sind die Voraussetzungen des t-Tests wirklich erfüllt?

- Schon mal gut: Varianzen in beiden Gruppen sind gleich $\sigma^2 + \sigma_Z^2$.
- ungefähr normalverteilt? Überprüfen wie sonst auch.
- Sind die $Y_1, \ldots, Y_{20}$ bzw. die $Z_1, \ldots, Z_{20}$ die unabhängig? Nicht ganz, aber fast: letztere wären unabhängig, wenn sie nicht ohne sondern mit Zurücklegen aus der empirischen Verteilung der $(v_1, \ldots, v_{20})$ gezogen worden wären.

Wir vernachlässigen hier die leichten Abhängigkeiten zwischen den $Z_1, \ldots, Z_{20}$ und damit den $Y_1, \ldots, Y_{20}$. Ein alternativer Ansatz wäre, davon auszugehen, dass die $(v_1, \ldots, v_{20})$ bereits selbst Zufallsvariablen sind, zwischen denen es Abhängigkeiten gibt, die durch das Randomisieren (weitgehend?) verloren gehen.

Was Sie u.a. erklären können sollten

- Berechnung von nötigen Stichprobenlängen
  - Theoretische Herleitung
  - Spezialfälle für Varianten des t-Test und des F-Tests
  - Verwendung der R-Befehle `power.t.test` und `power.anova.test`

- Was macht eine Stichprobe repräsentativ?
- Randomisierung: Wie und warum?
- Blockbildung
- Balanciertes Design