Eindimensionale Deskriptive Statistik

Quantitative Daten: Gemessene Zahlen (Grösse und Gewicht)
Qualitative Daten: Nehmen Werte an (Geschlecht und Nationalität)

Arithmetisches Mittel (Mittelwert, Durchschnitt)

Wo ist die Mitte der Daten?

```
waageA <- 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)

mean(waageA)
## [1] 80.02077
```

Arithmetisches Mittel

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Empirische Varianz und Standardabweichung

Ist die empirische Varianz (und damit die Standardabweichung) gross, so ist die Streuung der Messwerte um das arithmetische Mittel gross.

```
var(waageA)
## [1] 0.000574359
sd(waageA)
## [1] 0.02396579
```

Empirische Varianz *var*(x) und Standardabweichung s_X

$$\mathrm{Var}(x) = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$s_x = \sqrt{\mathrm{Var}(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Median (mittlerer Wert)

Wert, bei dem die Hälfte der Messwerte unter oder gleich diesem Wert sind.
Die andere Hälfte ist gleich diesem Messwert oder darüber.

```
median(waageA)
## [1] 80.03
waageB <- c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95, 79.97)

median(waageB)
## [1] 79.97
```

Quartile

Unteres Quartil: Wert, wo 25% aller Beobachtungen kleiner oder gleich und 75% grösser oder gleich sind wie dieser Wert. Man wählt den nächstgrösseren Wert als unteres Quartil.
Oberes Quartil: Wert, wo 75% aller Beobachtungen kleiner oder gleich und 25% grösser oder gleich sind wie dieser Wert. Man wählt den nächstgrösseren Wert als oberes Quartil.

```
# Syntax für das untere Quartil: p=0.25

quantile(waageA, p = 0.25, type = 2)
##   25%
## 80.02
quantile(waageB, p = 0.25, type = 2)
##   25%
## 79.96
# Syntax für das obere Quartil: p=0.75

quantile(waageA, p = 0.75, type = 2)
##   75%
## 80.04
```

Quartilsdifferenz

Ist ein Streuungsmass für die Daten (oberes Quartil − unteres Quartil). Es misst die Länge des Intervalls, das etwa die Hälfte der mittleren Beobachtungen enthält. Je kleiner dieses Mass, umso näher liegt die Hälfte aller Werte um den Median und umso kleiner ist die Streuung

```
quantile(noten, p = c(0.25, 0.75), type = 2)
##  25%  75%
## 3.80 5.35
IQR(noten, type = 2)
## [1] 1.55
```

Hälfte der Lernenden liegen innerhalb von 1.55 Noten, nämlich zwischen 3.8 und 5.35
25 % der Klasse 3.8 oder weniger; rund 25 % der Klasse 5.35 und mehr

Quantile

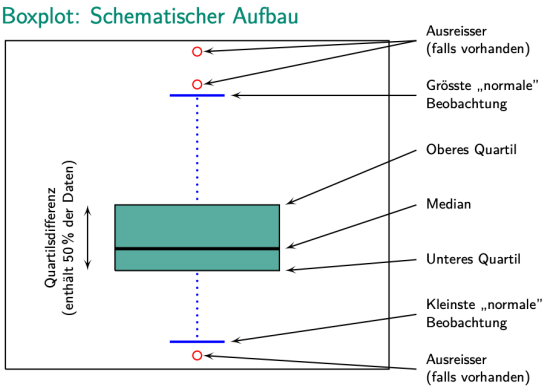
Quantile sind Quartile auf jede andere Prozentzahl verallgemeinert. Nachfolgend 10% und 70% -Quantil:

```
quantile(waageA, p = .1, type = 2)
##   10%
## 79.98
quantile(waageA, p = .7, type = 2)
##   70%
## 80.04
```

Knapp 10 % der Messwerte sind kleiner oder gleich 79.97
Entsprechend: Knapp 70 % der Messwerte kleiner oder gleich 80.04

Boxplot

In einem Boxplot werden numerische Daten in Quartile unterteilt, und zwischen den ersten und dritten Quartilen wird ein Feld gezeichnet, wobei eine zusätzliche Linie entlang des zweiten Quartils gezeichnet wird, um den Median zu kennzeichnen.




```
boxplot(waageA,
        col = "darkseagreen3"

)
```

```
boxplot(waageA, waageB,
        xlab = "Waage",
        col = c("orange", "lightblue")

)
axis(side = 1, at = c(1, 2), labels = c("A", "B"))
```

Boxplot: Darstellungen von verschiedenen Gruppen

tapply

Mittelwerte ganzer Datenreihen aus Spalten bestimmen. Erste Parameter gibt die Spalte für die Werte an, zweiter für die Namen und der dritte definiert, dass der Mittelwert berechnet werden soll.

```
tapply(diet$weight.loss, diet$Diet, mean)

##         1         2         3
## -3.300000 -3.025926 -5.148148
```

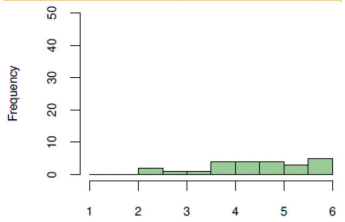
```
boxplot(weight.loss ~ Diet,
        data = diet,
        col = c("orange", "darkseagreen", "aquamarine")
```

boxplot(y ~ x), wobei y die Werte sind von denen R den Boxplot nehmen soll und x die Namen, nach denen die Werte geordnet werden sollen.

Histogramm

Gewöhnliches Histogramm (Frequency)

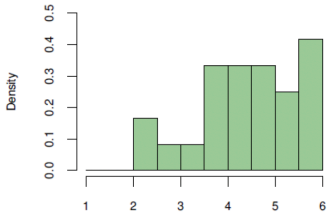
Höhe der Balken entspricht Anzahl der Beobachtungen in einer Klasse.




```
hist(waageA, col="darkseagreen")
```

Normiertes Histogramm (Density)

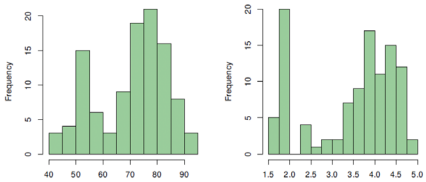
Balkenfläche entspricht dem prozentualen Anteil der jeweiligen Beobachtungen an der Gesamtanzahl Beobachtungen.




```
hist(waageA,freq = F)
```

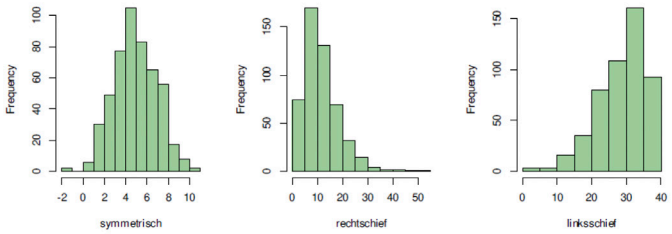
Bimodales Verhalten

Wenn es im Histogramm zwei Hügel gibt.



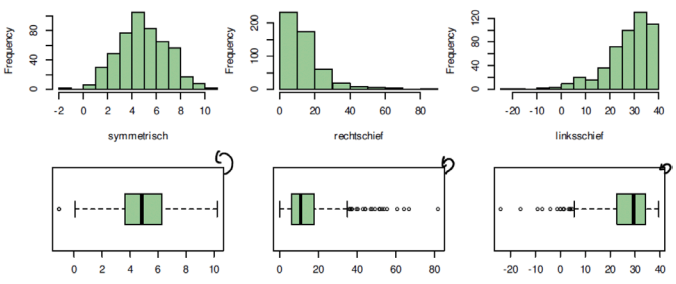
Schiefe von Histogrammen

Bezeichnung rechts und links bezieht sich immer auf die Seite, wo es weniger Daten hat.



Histogramm vs. Boxplott

Im Boxplot sind Lage, Streuung und Schiefe ersichtlich. Man sieht aber nicht, ob es bimodal ist.



Spick

Interquartilsdifferenz (IQR):
Die IQR ist kleiner, wenn die Daten weniger streuen (enger um den Median liegen).
Formel: IQR=Q3−Q1 (dritter Quartil minus erster Quartil).

Werte in der Box:
50 % der Werte liegen innerhalb der Box (zwischen Q1 und Q3).

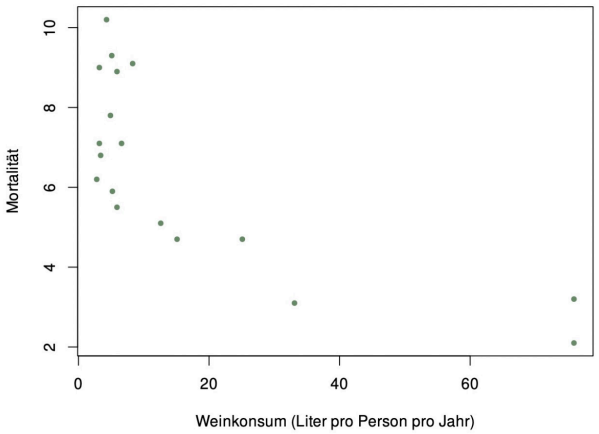
Zweidimensionale Deskriptive Statistik

Zweidimensionale Daten: An einem Versuchsobjekt werden jeweils zwei verschiedene Grössen gemessen.

Streudiagramm plot(x,y)

Plot deutet an, dass hoher Weinkonsum weniger Sterblichkeit wegen Herz-Kreislauferkrankung zu folge hat.

```
wein <- c(2.8, 3.2, 3.2, 3.4, 4.3, 4.9, 5.1, 5.2, 5.9, 5.9,
          6.6, 8.3, 12.6, 15.1, 25.1, 33.1, 75.9, 75.9)

mort <- c(6.2, 9.0, 7.1, 6.8, 10.2, 7.8, 9.3, 5.9, 8.9, 5.5,
          7.1, 9.1, 5.1, 4.7, 4.7, 3.1, 3.2, 2.1)

plot(wein, mort,
     xlab = "Weinkonsum (Liter pro Jahr)",
     ylab = "Mortalität",
     col = "blue",
     pch = 20

)
```

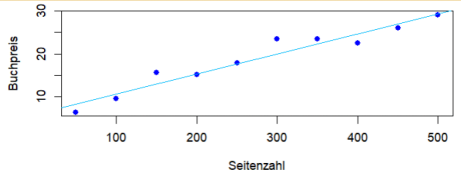


Regressionsgerade lm(y ~ x) + abline(lm(y ~ x))

Der Grundpreis ohne Seiten liegt bei 6.04. Pro zusätzliche Seite steigt der Preis um 0.04673.

```
seiten <- seq(50, 500, 50)

preis <- c(6.4, 9.5, 15.6, 15.1, 17.8, 23.4, 23.4, 22.5,
           26.1, 29.1)

lm(preis ~ seiten)

##
## Call:
## lm(formula = preis ~ seiten)
##
## Coefficients:
## (Intercept)       seiten
##     6.04000      0.04673
```

```
abline(lm(preis ~ seiten), col = "deepskyblue")
```



$$y = 6.04000 + 0.04673 * z$$

Korrelation

Dimensionslose Zahl zwischen −1 und +1. Korrelationskoeffizient erkennt nur lineare Zusammenhänge.

Wie kann man feststellen, ob ein linearer Zusammenhang der Daten besteht oder nicht?

Korrelationskoeffizient erkennt nur lineare Zusammenhänge

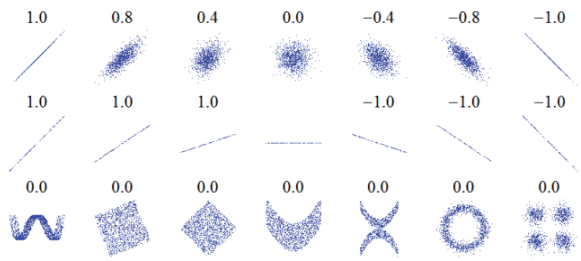
Wert sehr nahe bei 1: Starker linearer Zusammenhang (je mehr, desto mehr)

```
cor(seiten, preis)

## [1] 0.9681122
```

Empirische Korrelation

$$r = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2} \cdot \sqrt{(y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2) \cdot (\sum_{i=1}^{n}(y_i - \bar{y})^2)}}$$



Die Lösung dieses Optimierungsproblem ergibt:

$$b = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

wobei $\bar{x}$ und $\bar{y}$ die Mittelwerte der jeweiligen Daten

Residuum

Ein *Residuum* r_i ist die vertikale Differenz zwischen einem Datenpunkt (x_i, y_i) und dem Punkt $(x_i, a + bx_i)$ auf der gesuchten Geraden:

$$r_i = y_i - (a + bx_i) = y_i - a - bx_i$$

Spick

Steigung:

Die Steigung gibt an, wie stark die abhängige Variable (y) steigt oder fällt, wenn die unabhängige Variable (x) um eine Einheit zunimmt.

Beispiel: Steigung = 2 -> Für jede Erhöhung von x um 1 steigt y um 2.

Wahrscheinlichkeit

Wahrscheinlichkeitsmodell

Wahrscheinlichkeitsmodelle sind formale mathematische Beziehungen, die den möglichen Resultaten eines zufälligen Experiments ihre jeweiligen Wahrscheinlichkeiten zuordnen.

- *Grundraum* Ω: Enthält alle möglichen Elementarereignisse ω
- *Ereignisse* A, B, C: Teilmengen des Grundraums
- *W'keiten* P, die zu den Ereignissen A, B, C gehören

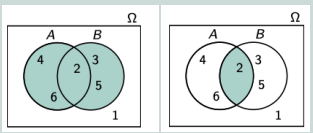
Beispiel: Würfelwurf

- Grundraum (die möglichen Ergebnisse)

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

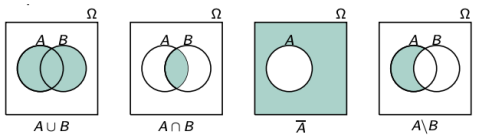
- Element $\omega = 2$ ist ein Elementarereignis
- Bedeutung: Beim Würfeln wurde die Zahl 2 geworfen
- Zahl 7: *Kein* Elementarereignis, da nicht im Grundraum Ω

- Ereignis A: Die geworfene Zahl ist gerade: $A = \{2, 4, 6\}$
- Ereignis B: Die geworfene Zahl ist Primzahl: $B = \{2, 3, 5\}$
- Ω wie gewohnt: $\Omega = \{1, 2, 3, 4, 5, 6\}$



Mengenlehre

Name	Symbol	Bedeutung
Vereinigung	$A \cup B$	A *oder* B, nicht-exklusives „oder"
Schnittmenge	$A \cap B$	A *und* B
Komplement	$\overline{A}$	*nicht* A
Differenz	$A \backslash B = A \cap \overline{B}$	A *ohne* B



Axiome der Wahrscheinlichkeit

Kolmogorov Axiome der Wahrscheinlichkeitsrechnung

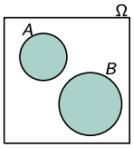
Jedem Ereignis A wird eine *W'keit* $P(A)$ zugeordnet, mit:

1. $P(A) \geq 0$
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ falls $A \cap B = \{\}$

A1: Wahrscheinlichkeit kann nicht negativ sein

A2: Mit $P(\Omega) = 1$: W'keiten eines Ereignisses zwischen 0 und 1

A3: Für zwei *disjunkte* Ereignisse:



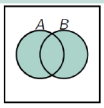
Stochastische Unabhängigkeit

Die stochastische Unabhängigkeit von Ereignissen impliziert, dass das Eintreten des einen keine Auswirkung auf die Wahrscheinlichkeit des Eintretens des anderen Ereignisses hat.

P(A U B) -> A *ODER* B:

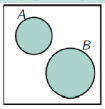
A und B nicht disjunkt und unabhängig:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



A und B disjunkt (gleichzeitiges Eintreten unmöglich) und abhängig:

$$P(A \cup B) = P(A) + P(B) \quad \text{falls } A \cap B = \{\}$$



P(A ∩ B) -> A *UND* B:

Sind die Ereignisse A und B *stochastisch unabhängig*, so gilt (↘ mit Schnittmenge)

$$P(A \cap B) = P(A) \cdot P(B)$$

- Unabhängig = keine gegenseitige Beeinflussung, Modell oft „Ziehen mit Zurücklegen".
- Abhängig = gegenseitige Beeinflussung, Modell oft „Ziehen ohne Zurücklegen".

Beispiel

- Ereignis A: Mit fairem Würfel eine eins oder zwei zu werfen
- Ereignis B: Kopf beim Werfen einer fairen Münze
- Werfen einer Münze keinen Einfluss auf das Resultat beim Würfelwurf
- Formel oben verwenden:

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

Gleichverteilung (Modell von Laplace)

- Annahme: Jedes Elementarereignis hat die gleiche W'keit
- Ereignis $E = \{\omega_1, \omega_2, \ldots, \omega_g\}$;
- Grundraum m Elemente
- W'keiten addieren sich zu 1 und deshalb:

$$P(\omega_k) = \frac{1}{|\Omega|} = \frac{1}{m}$$

Für ein Ereignis E im Laplace Modell gilt also

$$P(E) = \frac{g}{m} = \sum_{k:\ \omega_k \in E} P(\{\omega_k\})$$

- Man teilt die Anzahl der *„günstigen"* Elementarereignisse durch die Anzahl der *„möglichen"* Elementarereignisse

Beispiel: Laplace Modell

- Es werden zwei verschiedene (blau und rot) Würfel geworfen
- Wie gross ist die W'keit, dass die Augensumme 7 ergibt?
- Elementarereignis beschreibt die Augenzahlen auf beiden Würfeln
- Ergebnis in der Form 14 schreiben
- Ergebnis 14 ist *nicht* gleich 41
- Elementarereignisse:

$$\Omega = \{11, 12, \ldots, 16, 21 \ldots, 65, 66\}$$

- Anzahl Elementarereignisse:

$$|\Omega| = 36$$

- Ereignis E: Augensumme 7 wird gewürfelt
- Es gibt davon 6 Elementarereignisse:

$$E = \{16, 25, 34, 43, 52, 61\}$$

- Alle Elementarereignisse gleich wahrscheinlich: W'keit für Ereignis E:

$$P(E) = \frac{|E|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$$

Spick

Elementarereignis:

Ein Elementarereignis ist das kleinste mögliche Ergebnis eines Zufallsexperiments. Beispiel: Beim Würfeln ist „1" ein Elementarereignis.

Zufallsvariable

Zu jedem Elementarereignis ω wird mit der Funktion X(ω) = x eine Zahl zugeordnet. Die Werte, die die Zufallsvariable annehmen kann, wird als Wertemenge bezeichnet.

Zufallsvariable

Eine *Zufallsvariable* X ist eine *Funktion*:

$$\begin{aligned} X: \quad \Omega &\to \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$

- *Zufallsvariable* werden mit *Grossbuchstaben* X (oder Y, Z) bezeichnet
- Entsprechender *Kleinbuchstabe* x (oder y, z) stellt *konkreter Wert* dar, den die Zufallsvariable annehmen kann
- Ereignis, bei dem die Zufallsvariable X den Wert x annimmt:

$$X = x$$

Wahrscheinlichkeitsverteilung

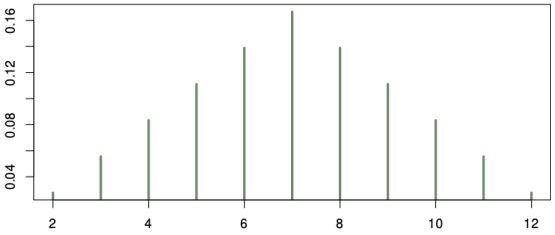
Eine Wahrscheinlichkeitsverteilung ist eine mathematische Funktion, bei der jedem möglichen Wert eines Zufallsexperiments eine bestimmte Wahrscheinlichkeit zugeordnet wird.

W'keitsverteilung für die Zufallsvariable X:

x	2	3	4	5	6	7	8	9	10	11	12
$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

<- muss sich auf 1 addieren !!!

Skizze:



Beispiel: Augensumme zweier Würfel

Wie gross ist die W'keit, die Augensumme 6 oder 8 zu würfeln?

- Gesucht: $P(X=6) + P(X=8)$:

$$P(X=6) + P(X=8) = \frac{5}{36} + \frac{5}{36} = \frac{10}{36} = \frac{5}{18}$$

Wahrscheinlichkeitsverteilung

Für *jede* Realisierung einer Zufallsvariable wird die zugehörige W'keit berechnet → *W'keitsverteilung* dieser Zufallsvariablen

Wahrscheinlichkeitsverteilung

„Liste" von $P(X=x)$ für alle *möglichen* Werte $x_1, x_2, \ldots, x_n$ heisst *diskrete W'keitsverteilung* der diskreten Zufallsvariablen X

Es gilt immer:

$$P(X=x_1) + P(X=x_2) + \ldots + P(X=x_n) = 1$$

Mit Summenzeichen:

$$\sum_{\text{alle möglichen } x} P(X=x) = 1$$

Alle W'keiten einer W'keitsverteilung ergeben 1

Kennzahlen einer Verteilung

- Standardabweichung mit R berechnen:

```
x <- 1 : 6
p <- 1 / 6

E_X <- sum(x * p)

var_X <- sum((x - E_X)^2 * p)
sd_X <- sqrt(var_X)

sd_X
## [1] 1.707825
```

- D.h.: Abweichung „durchschnittlich" 1.7 von 3.5

- Wurf eines fairen Würfels: Alle 6 möglichen Zahlen gleiche W'keit geworfen zu werden
- Zufallsvariable X sei die geworfene Zahl
- Erwartungswert $\mathrm{E}(X)$:

$$\begin{aligned} \mathrm{E}(X) &= x_1 \cdot P(X=x_1) + x_2 \cdot P(X=x_2) + \ldots + x_6 \cdot P(X=x_6) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{1}{6}(1+2+3+4+5+6) \\ &= 3.5 \end{aligned}$$

- Dieser Erwartungswert 3.5 nicht anderes als der Durchschnitt der Augenzahlen

Erwartungswert und Standardabweichung

- *Erwartungswert*:

$$\begin{aligned} \mathrm{E}(X) &= x_1 \cdot P(X=x_1) + x_2 \cdot P(X=x_2) + \ldots + x_n \cdot P(X=x_n) \\ &= \sum_{\text{alle möglichen } x} x P(X=x) \end{aligned}$$

- *Varianz* und *Standardabweichung*:

$$\begin{aligned} \mathrm{Var}(X) &= (x_1 - \mathrm{E}(X))^2 \cdot P(X=x_1) + \ldots + (x_n - \mathrm{E}(X))^2 \cdot P(X=x_n) \\ &= \sum_{\text{alle möglichen } x} (x - \mathrm{E}(X))^2 P(X=x) \end{aligned}$$

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}$$

Empirische und theoretische Kennzahlen

Das arithmetisches Mittel x nähert sich für immer mehr Versuche dem theoretischen Wert μX = E(x) an, sofern die Daten der Wahrscheinlichkeitsverteilung von X folgen

Die empirische Standardabweichung sX nähert sich für immer mehr Versuche dem theoretischen Wert σX an, falls die Daten der Wahrscheinlichkeitsverteilung von X folgen.

- Arithmetische Mittelwert $\bar{x}$: Aus *konkreten* Daten berechnet: Aus Messwerten $x_1, \ldots, x_n$ wird nach der Formel oben $\bar{x}_n$ berechnet
- Erwartungswert $\mathrm{E}(X)$: *Theoretischer Wert*, der sich aus dem Modell der W'keitsverteilung ergibt

$$P(M \mid R) = \frac{P(R \cap M)}{P(R)}$$

	M	F	
R	P(R ∩ M)	P(R ∩ F)	P(R)
$\bar{R}$	P($\bar{R}$ ∩ M)	P($\bar{R}$ ∩ F)	P($\bar{R}$)
	P(M)	P(F)	1

- Empirische Standardabweichung: s_X aus *konkreten* Daten berechnet: Aus Messwerte $x_1, \ldots, x_n$ wird nach Formel oben s_X berechnet
- Standardabweichung σ_X: Theoretischer Wert, der sich aus Modell der W'keitsverteilung ergibt

Bedingte Wahrscheinlichkeit

Wahrscheinlichkeit des Eintreten eines Ereignisses unter der Bedingung dass das Eintreten eines anderen Ereignisses bereits bekannt ist.

- Die *bedingte W'keit* ist die W'keit, dass das Ereignis A eintritt, wenn man schon weiss, dass B eingetreten ist
- Bezeichnung:

$$P(A \mid B)$$

- Längsstrich wird als „unter der Bedingung" gelesen
- Bedingte W'keit $P(A \mid B)$ wird definiert durch

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- Interpretation: $P(A \mid B)$ ist die W'keit für das Ereignis A, wenn man weiss, dass das Ereignis B schon eingetroffen ist

Satz von Bayes

Wenn A und B zwei Ereignisse sind und die Wahrscheinlichkeit P(A|B) gegeben ist, kannst Du P(B|A) berechnen.

Bayes' Theorem

Nützlicher Zusammenhang zwischen $P(A \mid B)$ und $P(B \mid A)$:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) P(A)}{P(B)}$$

Beispiel: Bayes Theorem liefert die gleiche Lösung wie vorher:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{0.9 \cdot (0.009 + 0.001)}{0.009 + 0.099} = \frac{0.009}{0.009 + 0.099} = 0.08$$

- 1000 Personen haben die Krankheit (1 %)
- 90 % dieser Personen werden positiv getestet: 900 Personen
- 99 000 haben die Krankheit nicht
- 10 % dieser Personen werden positiv getestet: 9900 Personen
- Anzahl positiv Getesteter:

$$900 + 9900 = 10\,800$$

- Unter diesen positiv getesteten sind aber bei weitem mehr Gesunde, die fälschlicherweise positiv getestet wurden
- W'keit, dass eine positiv getestete Person auch wirklich krank ist:

$$\frac{900}{10\,800} = 0.0833$$

Gesetz der totalen Wahrscheinlichkeit

Für Partionierung $A_1, \ldots, A_k$ und jedes beliebige Ereignis B gilt:

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_k)P(A_k) \\ &= \sum_{i=1}^{k} P(B|A_k)P(A_k) \end{aligned}$$

Normalverteilung

Punktwahrscheinlichkeit: Wahrscheinlichkeit, dass genau eine Körpergrösse gemessen wird.

Wahrscheinlichkeitsdichte: Wahrscheinlichkeit, dass ein Messwert in einem bestimmten Bereich liegt.
Wahrscheinlichkeit entspricht der Fläche zwischen a und b ur

Für eine W'keitsdichte $f(x)$ gelten folgende Eigenschaften:

- Es gilt:
  $$f(x) \geq 0$$
  Das heisst, die Kurve liegt überhalb der x-Achse
- W'keit
  $$P(a < X \leq b)$$
  entspricht der Fläche zwischen a und b unter $f(x)$
- Die gesamte Fläche unter der Kurve ist 1:
  - Dies ist die W'keit, dass *irgendein* Wert gemessen wird

Gaussverteilung

4.1 Gaussverteilung

Formel: $X \sim N(\mu, \sigma^2)$

- Durch Parameter μ (Erwartungswert/Mittelwert) Verschiebung der Kurve:
  - Nach rechts, falls μ positiv
  - Nach links, falls μ negativ
- Durch Parameter σ (Standardabweichung) wird die Kurve:
  - Schmal und hoch um μ, falls σ klein (nahe bei 0)
  - Weit und tief um μ, falls σ gross
- Die Wahrscheinlichkeit, dass eine Beobachtung höchstens
  - *eine* Standardabweichung vom Erwartungswert abweicht, ist etwa $^2/_3$
  - *zwei* Standardabweichung vom Erwartungswert abweicht, ist etwa *0.95*

Funktion	Was wird berechnet?	Wann verwenden?
$\text{pnorm}(x)$	Wahrscheinlichkeit $P(X \leq x)$	Wenn du wissen möchtest, wie viele Werte kleiner/gleich x sind.
$1 - \text{pnorm}(x)$	Wahrscheinlichkeit $P(X > x)$	Wenn du wissen möchtest, wie viele Werte größer x sind.
$\text{qnorm}(p)$	Wert x, sodass $P(X \leq x) = p$	Wenn du den Wert für eine bestimmte Wahrscheinlichkeit suchst.

Beispiel IQ-Test mit R

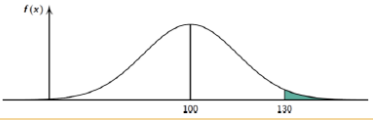
4.2 Beispiel IQ-Test mit R

IQ-Tests folgen einer Normalverteilung mit Mittelwert μ = 100 und Standardabweichung σ = 15

4.2.1 pnorm

Wie gross ist die Wahrscheinlichkeit, dass jemand einen IQ von mehr als 130 hat P(X > 130)?

Befehl: $1 - pnorm(\alpha, \mu, \sigma)$



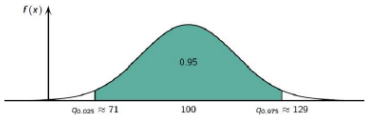

```
1 - pnorm(q = 130, mean = 100, sd = 15)

## [1] 0.02275013
```

⇨ Rund 2% der Bevölkerung sind hochbegabt.

4.2.2 qnorm

Welches Intervall enthält 95% der IQ's um den Mittelwert?

Befehl: $qnorm(q_\alpha, \mu, \sigma)$




```
qnorm(p = c(0.025, 0.975), mean = 100, sd = 15)

## [1]  70.60054 129.39946
```

⇨ Also haben 95% der Menschen einen IQ zwischen ca. 70 und 130. Diese Werte entsprechen einem Abstand von etwa 2 Standardabweichungen vom Mittelwert.

Gesetz der grossen Zahlen

Für $n \to \infty$ geht die Streuung gegen null.

Sind Zufallsvariablen i.i.d., so wird dasselbe unter den gleichen Bedingungen gemessen. independent, identically distributed

Kennzahlen von S_n → Summe

$$\text{E}(S_n) = n\mu$$

nimmt mit grösserem n zu:
$$\begin{cases} \text{Var}(S_n) = n\,\text{Var}(X_i) \\ \sigma(S_n) = \sqrt{n}\sigma_X \end{cases}$$

Kennzahlen von $\overline{X}_n$ → arithmetisches Mittel / Durchschnitt

$$\text{E}(\overline{X}_n) = \mu$$

nimmt mit grösserem n ab:
$$\begin{cases} \text{Var}(\overline{X}_n) = \dfrac{\sigma_X^2}{n} \\ \sigma(\overline{X}_n) = \dfrac{\sigma_X}{\sqrt{n}} \end{cases}$$ Standardfehler ($\frac{\sigma}{\sqrt{n}}$ mit u.m)

Zentraler Grenzwertsatz

Verteilung der Mittelwerte/Summen nähert sich mit wachsendem n einer Normalverteilung an.

Zentraler Grenzwertsatz

$X_1, \ldots, X_n$ i.i.d. mit irgendeiner Verteilung mit Erwartungswert μ und Varianz σ^2, dann gilt (ohne Beweis):

$$S_n \approx \mathcal{N}(n\mu, n\sigma_X^2)$$

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right)$$

- Approximation wird mit grösserem n i.A. besser
- Approximation besser, je näher die Verteilung von X_i bei der Normalverteilung $\mathcal{N}(\mu, \sigma_X^2)$ ist

Der zentrale Grenzwertsatz besagt, dass sich der Mittelwert und die Summe unabhängig und identisch verteilter Zufallsvariablen bei einer beliebigen Verteilung mit zunehmenden Stichprobenumfang der Normalverteilung annähern.

Summe → *Befehl:* $pnorm(\alpha, \mu * n, \sigma * \sqrt{n})$

Strassenverkehrsamt hat genug Streusalz gelagert, um mit einem Schneefall von insgesamt 80 cm pro Jahr fertigzuwerden. Täglich fallen im Mittel 1.5 cm mit einer Standardabweichung von 0.3 cm. Wie gross ist Wahrscheinlichkeit, dass das gelagerte Salz für die nächsten 50 Tage ausreicht?

```
pnorm(q = 80, mean = 50 * 1.5, sd = sqrt(50) * 0.3)

## [1] 0.9907889
```

Durchschnitt → *Befehl:* $pnorm(\alpha, \mu, \frac{\sigma}{\sqrt{n}})$

Die Lebensdauer eines bestimmten elektrischen Teils ist durchschnittlich 100 Stunden mit Standardabweichung von 20 Stunden. Wir testen 16 solcher Teile. Wie gross ist die Wahrscheinlichkeit, dass das Stichprobenmittel unter 104 Stunden liegt?

```
pnorm(q = 104, mean = 100, sd = 20/sqrt(16))

## [1] 0.7881446
```

Hypothesentest

Wichtiges statistisches Mittel, um zu entscheiden, ob der Mittelwert einer Messreihe zu einem bestimmten wahren Mittelwert passt oder nicht.

p-Wert ist ein Wert zwischen 0 und 1, der angibt, wie gut Nullhypothese und Daten zusammenpassen.

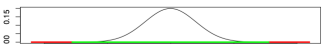
1. Verwerfe H_0 falls p-Wert $\leq \alpha$
2. Belasse H_0 falls p-Wert $> \alpha$

Wird die Nullhypothese verworfen, so deutet ein sehr kleiner p-Wert darauf hin, dass die Nullhypothese sicherer verworfen wird, als wenn er in der Nähe des Signifikanzniveaus ist.

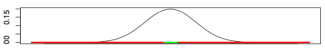
p-Wert ≈ 0.05 : schwach signifikant, "."
p-Wert ≈ 0.01 : signifikant, "$*$"
p-Wert ≈ 0.001 : stark signifikant, "$**$"
p-Wert $\leq 10^{-4}$: äusserst signifikant, "$***$"

Mit zunehmendem n wird die Wahrscheinlichkeit (p-Wert) immer kleiner, da die Standardabweichung mit grösser werdendem n kleiner wird. D.h. je mehr Messungen wir haben, umso gewichtiger ist eine Abweichung von wahrem Mittelwert

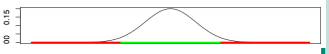
- Graphik: $\alpha = 0.0001$ (nahe bei 0)



- Graphik: $\alpha = 0.8$ (gross)



- Graphik: $\alpha = 0.05$



Berechne den p-Wert:
Der p-Wert gibt die Wahrscheinlichkeit an, dass das Ergebnis zufällig zustande kam, unter der Annahme, dass H0 wahr ist.
Signifikanzniveau (α):
Typischer Wert:α=0.05 (5 % Fehlerwahrscheinlichkeit).

Nullhypothese

$$H_0: \quad \mu = \mu_0 = 80$$

Alternativhypothese

$$H_A: \quad \mu \neq \mu_0 = 80 \quad \text{oder „<" oder „>"}$$

z-Test

Beim z-Test ist die Standardabweichung im Voraus bekannt.
1. Null- und Alternativhypothese aufstellen
2. Verwerfungsbereich bestimmen
3. p-Wert mit Messreihe berechnen
4. Testentscheid fällen

Das Bundesamt für Statistik behauptet, dass die durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz bei 180 cm mit einer Standardabweichung von 10 cm liegt. -> Einseitiger Test

Der p-Wert ist unter dem Signifikanzniveau von 0.05. Die Nullhypothese wird somit verworfen und die Alternativhypothese angenommen.

- *Nullhypothese* $H_0: \quad \mu_0 = 180$
- *Alternativhypothese* $H_A: \quad \mu < \mu_0 = 180$

```
qnorm(p = 0.05, mean = 180, sd = 10/sqrt(8))
## [1] 174.1846
```

Wählen zufällig 8 erwachsene Frauen aus, deren durchschnittliche Körpergrösse 171.54 cm beträgt.

```
pnorm(q = 171.54, mean = 180, sd = 10/sqrt(8))
## [1] 0.008359052
```

Wann? Große Stichprobe (n>30), Varianzen bekannt, normalverteilt.
Beispiel: Vergleich des Durchschnittsgewichts einer Region mit einem bekannten Mittelwert von 70 kg.

t-Test

Beim t-Test ist die Standardabweichung im Voraus unbekannt.
Es folgt eine zusätzliche Unsicherheit. Die t-Verteilung ist ähnlich der Normalverteilung, aber flacher, aufgrund der grösseren Unsicherheit.
Das Bundesamt für Statistik behauptet, dass die durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz
bei 180 cm liegt. -> Einseitiger Test
Wählen zufällig 10 Frauen aus und messen deren Körpergrösse.

Der p-Wert ist unter dem Signifikanzniveau von 0.05. Die Nullhypothese wird somit verworfen und die Alternativhypothese angenommen

```
groesse <- c(165.7, 156.7, 171.7, 180.3, 163.2, 166.7, 149.9,
    170.4, 163.4, 152.5)
t.test(groesse, mu = 180, alternative = "less")
##
##  One Sample t-test
##
## data:  groesse
## t = -5.4836, df = 9, p-value = 0.0001942
## alternative hypothesis: true mean is less than 180
## 95 percent confidence interval:
##      -Inf 169.382
## sample estimates:
## mean of x
##    164.05
```

Wann? Kleine Stichprobe (n<30), Varianzen unbekannt, normalverteilt.
Arten:
Gepaarter T-Test:
Beispiel: Gewichtsmessung vor und nach einer Diät.
Ungepaarter T-Test:
Beispiel: Vergleich der Durchschnittstemperatur in zwei Städten.

Wilcoxon-Test

Beim Wilcoxon-Test müssen die Daten nicht normalverteilt sein.
Er ist eine Alternative zum t-Test mit weniger Voraussetzungen.
Lediglich die Verteilung unter der Nullhypothese muss symmetrisch sein.

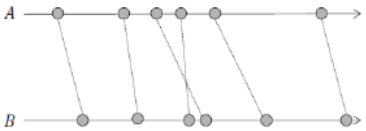
```
x <- c(79.98, 80.04, 80.0
    80.05, 80.03, 80.02,

wilcox.test(x, mu = 80, alternative = "two.sided")
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  x
## V = 69, p-value =
## alternative hypoth
```

Wann?: Daten nicht-normalverteilt.
Arten:
Wilcoxon-Vorzeichen-Rang-Test (gepaart):
Beispiel: Vorher-Nachher-Vergleich der Schlafqualität bei nicht-normalverteilten Daten.
Wilcoxon-Mann-Whitney-Test (ungepaart):
Beispiel: Vergleich von Gehältern in zwei Branchen bei nicht-normalverteilten Daten.

Vergleich von zwei Stichproben

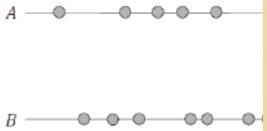
Gepaarte Stichproben

- Jede Beobachtung einer Gruppe kann eindeutig einer Beobachtung der anderen Gruppe zugeordnet werden
- Stichprobengrösse ist in beiden Gruppen zwangsläufig gleich
- abhängig voneinander



Ungepaarte Stichproben

- Keine Zuordnung von Beobachtungen möglich
- Stichprobengrössen können verschieden sein (müssen aber nicht!)
- Man kann die eine Gruppe vergrössern, ohne dass man die andere vergrössert



Gepaarte Stichproben
Gepaarte Stichproben:
Beispiel: Blutdruck eines Patienten vor und nach einer Behandlung. (Messungen an denselben Personen).
Ungepaarte Stichproben:
Beispiel: Vergleich der Körpergrößen von Männern und Frauen. (Unabhängige Gruppen).

p-Wert <= a H0 ablehnen -> Effekt signifikant
p-Wert > a H0 beibehalten -> kein signifikanter Effekt

Situation	Test
Vergleich von Mittelwerten, Varianz bekannt	Z-Test
Vergleich von Mittelwerten, kleine Stichprobe, normalverteilt	T-Test
Gepaarte Stichproben, normalverteilt	Gepaarter T-Test
Ungepaarte Stichproben, normalverteilt	Ungepaarter T-Test
Daten nicht-normalverteilt, gepaart	Wilcoxon-Vorzeichen-Rang-Test
Daten nicht-normalverteilt, ungepaart	Wilcoxon-Mann-Whitney-Test

```
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)

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

##
##  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
```

Die Nullhypothese ist, dass die Behandlung keine Wirkung hat. Sie wird in diesem Beispiel verworfen.

Lineare Regression

Auf Basis von erklärenden Variablen (Prädiktoren) eine passende Outputvariable (Zielgrösse) finden. Durch das hinzufügen eines Fehlerterms ε (Residuen), welcher unabhängig von den Prädikatoren ist, wird das Modell zu einer Gleichung.

Einfache lineare Regression

Einfaches Verfahren, um einen quantitativen Output Y auf der Basis einer einzigen Inputvariable X vorherzusagen.

$$Y \approx \beta_0 + \beta_1 X$$

- β_0 ist der y-Achsenabschnitt
- β_1 die Steigung der Geraden

Für zusätzliche CHF 1'000 Werbeausgaben werden 47.5 zusätzliche Einheiten des Produktes verkauft.
Die Nullhypothese wird mit p-Wert 2*10-16 verworfen.
Somit gibt es einen klaren Zusammenhang.
Die R2-Statistik erklärt 61.19% der Varianz durch das Modell.
Das Modell ist somit zu 2/3 akurat.

$$\text{Verkauf} \approx \beta_0 + \beta_1 \cdot \text{TV} \qquad Y \approx 7.03 + 0.0475X$$

```
confint(lm(Verkauf ~ TV), level = 0.95)
##                2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

Verkauf liegt ohne Werbung zwischen 6'130 und 7'935 Einheiten.
Für zusätzliche CHF 1'000 für TV-Werbung, werden durchschnittlich zwischen 42 und 53 Einheiten mehr verkauft.

```
summary(lm(Verkauf ~ TV))
##
## Call:
## lm(formula = Verkauf ~ TV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119,Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Multiple Lineare Regression

Das multiple lineare Modell verallgemeinert das einfache lineare Modell. Jeder erklärenden Variable wird ein eigener Steigungskoeffizient in einer Gleichung zugeordnet. Graphische Darstellung nur bis zwei erklärende Variablen möglich (Ebene).
βi: Durchschnittliche Änderung der Zielgrösse bei Änderung von Xi um eine Einheit, wenn alle anderen erklärenden Variablen gleichbleiben.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Steigung für Zeitung beschreibt die Änderung der Zielgrösse Verkauf, wenn man CHF 1'000 mehr für Zeitungswerbung ausgibt, wobei die anderen erklärenden Variablen TV und Radio gleichbleiben.
R2 erhöht sich, je mehr erklärende Variablen berücksichtigt werden.

$$\text{Verkauf} \approx \beta_0 + \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{Radio} + \beta_3 \cdot \text{Zeitung}$$
$$\text{Verkauf} \approx 2.94 + 0.046 \cdot \text{TV} + 0.189 \cdot \text{Radio} - 0.001 \cdot \text{Zeitung}$$

```
cor(data.frame(TV, Radio, Zeitung, Verkauf))
##                TV      Radio    Zeitung   Verkauf
## TV       1.00000000 0.05480866 0.05664787 0.7822244
## Radio    0.05480866 1.00000000 0.35410375 0.5762226
## Zeitung  0.05664787 0.35410375 1.00000000 0.2282990
## Verkauf  0.78222442 0.57622257 0.22829903 1.0000000
```

```
summary(lm(Verkauf ~ TV + Radio + Zeitung))
##
## Call:
## lm(formula = Verkauf ~ TV + Radio + Zeitung)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Zeitung     -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

In Märkten, wo mehr in die Werbung fürs Radio investiert wird, ist auch die Werbung für die Zeitung grösser, aufgrund des Korrelationskoeffizienten von 0.35. Aber Zeitungswerbung beeinflusst Verkäufe nicht. Zeitung schmückt sich hier mit fremden Lorbeeren, nämlich dem Erfolg von Radio auf Verkauf.
Zuerst entscheiden ob die erklärenden Variablen Einfluss auf die Zielgrösse haben und dann ein Modell aufstellen, welches nur diese Variablen enthält. Interaktionseffekt: lm(medv~lstat*age)

Qualitative Variablen

Quantitative Daten: Gemessene Zahlen (Grösse und Gewicht)
Qualitative Daten: Nehmen Werte an (Geschlecht und Nationalität)

Qualitative erklärende Variablen heissen auch Faktoren. Sie nehmen Stufen oder Levels an.
Ein Faktor kann durch eine Indikatorvariable ins Regressionsmodell aufgenommen werden. Es gibt immer eine Indikatorvariable weniger, als es Levels hat (hier zwei). Das Level ohne Indikatorvariable (hier Afroamerikaner) ist die Baseline.

β0: Durchschnittliche Kreditkartenrechnungen von Afroamerikanern
β1: Differenz der durchschnittlichen Rechnungen von Afroamerikanern und Asiaten
β2: Differenz der durchschnittlichen Rechnungen von Afroamerikanern und Kaukasiern

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{falls } i\text{-te Person asiatisch} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person kaukasisch} \\ \beta_0 + \varepsilon_i & \text{falls } i\text{-te Person afroamerikanisch} \end{cases}$$

```
summary(lm(balance ~ ethnicity))
##
## Call:
## lm(formula = balance ~ ethnicity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -531.00 -457.08  -63.25  339.25 1480.50
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         531.00      46.32  11.464   <2e-16 ***
## ethnicityAsian      -18.69      65.02  -0.287    0.774
## ethnicityCaucasian  -12.50      56.68  -0.221    0.826
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.9 on 397 degrees of freedom
## Multiple R-squared:  0.0002188,Adjusted R-squared:  -0.004818
## F-statistic: 0.0434 on 2 and 397 DF,  p-value: 0.9575
```

Asiaten haben um durchschnittlich $ 18.69 kleinere Rechnungen als die Afroamerikaner.
Kaukasier haben um durchschnittlich $ 12.50 kleinere Rechnungen als die Afroamerikaner.
Der p-Werte ist gross. Es gibt keinen signifikanten Unterschied bei den Kreditkartenrechnungen zwischen den Ethnien.

Serie 01

In **R** wird ein Vektor `temp` mit 10 Werten definiert.

Welche Behauptungen zu den **R** Befehlen `plot(...)` und `abline(...)` sind korrekt?

(Richtige Angaben geben 1 Punkt, flasche Angaben geben 0.5 Punkte Abzug.)

- [x] Mit `plot(temp, type="p")` werden die Werte von `temp` als runde Kreise im Plot eingezeichnet.
- [ ] Mit `abline(v=4)`wird in den durch `plot(temp)` erstellten Plot zusätzlich eine horizontale Linie mit Abstand 4 von der `x` Achse eingezeichnet.
- [x] Mit `abline(a=0,b=1)`wird in den durch `plot(temp)` erstellten Plot zusätzlich eine Gerade mit der Gleichung $y = x$ eingezeichnet.
- [ ] Mit dem einzigen Befehl `abline(a=1,b=3)` wird ein Plot mit einer Linie erstellt, auch wenn vorher der Befehl `plot(temp)` nicht aufgerufen wird.
- [ ] Es ist nicht möglich einzig mit dem Befehlt `plot(...)` die Werte von `temp` als runde Kreise zu zeichnen, so dass diese auch mit einer durchgezogenen Linie verbunden sind.
- [x] Mit `plot(temp,type="l",lty=2)` werden die Werte von `temp` durch eine gestrichelte Linie verbunden eingezeichnet.

Welche Behauptungen zum **R** Befehl `seq(...)` sind korrekt?

(Richtige Angaben geben 1 Punkt, falsche Angaben geben 0.5 Punkte Abzug.)

- [x] Der Befehl `seq(from=10,to=20,length.out=11)` gibt die Werte `10 11 12 13 14 15 16 17 18 19 20` zurück.
- [x] Der Befehl `seq(from=10,to=20,length.out=100)` gibt 100 Werte zurück.
- [ ] Der Befehl `seq(from=10,to=20,by=3)` gibt fünf Werte zurück.
- [x] Der Befehl `seq(from=10,to=20,by=2)` gibt die Werte `10 12 14 16 18 20` zurück.

Der folgende Vektor wird in **R** definiert:

```
Hunde <- c(10,11.5,9.5,12,15,9)
```

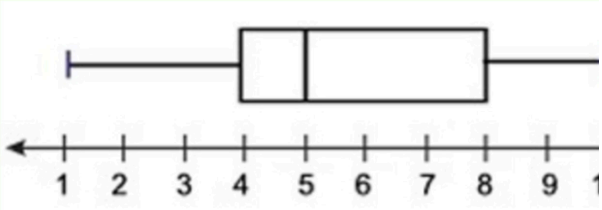
Welche Behauptungen sind korrekt?

(Richtige Angaben geben 1 Punkt, falsche Angaben geben 0.5 Punkte Abzug.)

- [ ] Mit `Hunde(c[1,3])` werden die Zahlen `10 9.5` zurückgegeben.
- [ ] Der Befehl `length(Hunde)` gibt den Wert `8` zurück.
- [x] Der Befehl `(Hunde >= 11)[c(1)]` gibt den Wert `FALSE` zurück.
- [x] Mit `Hunde[Hunde > 10]` werden die Zahlen `11.5 12.0 15.0` zurückgegeben.
- [ ] Mit `Hunde[0]` wird die Zahl `10` zurückgegeben.
- [ ] Der Befehl `order(Hunde)` gibt die Werte von `Hund` der Grösse nach aufsteigend zurück.
- [x] Der Befehl `(Hunde[c(1,4,6)] - 10)^2` gibt die Werte `0 4 1` zurück.
- [x] `Hunde - 10` ist ein Vektor gleicher Länge wie Hunde.
- [x] Der Befehl `sum(Hunde[Hunde < 10])` gibt den Wert `18.5` zurück.
- [ ] Mit `Hunde[1,5]` werden die Zahlen `10 15` zurückgegeben.

Serie 02

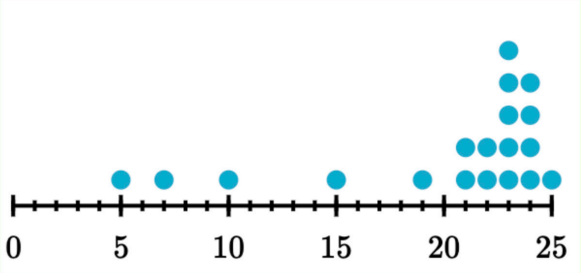
Markieren Sie alle richtigen Aussagen zum folgenden Boxplot:



- [ ] $Q_3 = 5$
- [ ] $IQR = 8$
- [ ] Der kleinste Ausreisser ist 1.
- [x] Der kleinste normale Wert ist 1.
- [x] $Median = 5$

Wir haben einen Datensatz mit 12 Zahlen, wobei die kleinste Zahl 20 ist, und die grösste Zahl 50 beträgt. Sei Q_1 das untere Quartile, und Q_3 das obere Quartile, sowie Q_1=Q_3. Markieren Sie alle richtigen Aussagen:

- [x] Es ist möglich, dass der Median 50 beträgt.
- [x] $Median = Q_3$
- [x] $Median = \frac{Q_1+Q_3}{2}$
- [ ] $Median \neq Q_2$



Wieviele Werte können wir als untere Ausreisser bezeichnen:

- ○ 4
- ○ 5
- ○ Keine
- ○ 2
- ◉ 3

Serie 03

In der `csv` Datei "`MadridWetterdaten1997-2015`" (download vor diesem Quiz in ILIAS) sind die täglich in Madrid gemessenen Werte für die **Temperatur** in Grad Celsius, für die **relative Luftfeuchtigkeit** in Prozent, für den **Luftdruck** in hPa und für die **Windgeschwindigkeit** in km/h von 1997 bis 2015 gespeichert.

Laden Sie diese Datei in **R** und entscheiden Sie durch eine entsprechende Datenanalyse in **R**, welche der folgenden Behauptungen wahr sind:

- [x] Das HIstogramm für die Windgeschwindigkeiten in Madrid ist **rechtsschief**.
- [x] Wenn ein Histogramm für einen Datensatz eine symmetrische Form hat, dann muss auch das entsprechende normierte Histogramm eine symmetrische Form haben.
- [ ] Im Streudiagramm mit der Temperatur auf der horizontalen Achse und der relativen Luftfeuchtigkeit auf der vertikalen Achse hat die Regressionsgerade die Gleichung $y = -1.928 + 86.236x$.
- [ ] Im Streudiagramm mit der Temperatur auf der horizontalen Achse und der relativen Luftfeuchtigkeit auf der vertikalen Achse hat die Regressionsgerade die Gleichung $y = 86.236 + 1.928x$.
- [x] Im Streudiagramm mit der Windgeschwindigkeit auf der horizontalen Achse und der relativen Luftfeuchtigkeit auf der vertikalen Achse sind die Punkte recht zerstreut, folgen aber einer leicht fallenden Regressionsgeraden.
- [x] Die Daten zeigen, dass bei **grösseren** Temperaturen die relative Luftfeuchtigkeit **kleiner** ist.
- [ ] Mit dem Befehl `abline(lm(windspeed ~ temperature),col="red")` wird in das Streudiagramm mit `windspeed` auf der horizontalen Achse und `temperature` auf der vertikalen Achse die passende Regressionsgerade in roter Farbe gezeichnet.
- [ ] Das Histogramm für die relative Luftfeuchtigkeit in Madrid ist **linksschief**.
- [x] Wenn die Windgeschwindigkeit in Madrid 0 km/h ist, beträgt der Luftdruck gemäss der Regressionsgerade schätzungsweise 1022.9076 hPa.
- [x] Am häufigsten wurden Luftdruckwerte **zwischen 1015 und 1020 hPa** gemessen.
- [ ] Für die Temperaturwerte erstellt R mit `breaks=25` ein Histogramm mit weniger Klassen als mit `breaks=30`.
- [x] Wenn in Madrid die Windgeschwindigkeit um 1 km/h steigt, dann nimmt der Luftdruck gemäss der Regressionsgerade schätzungsweise um 0.5381 hPa ab.

Serie 04

Bestmögliche Lösung:

Herr Borger organisiert eine Wohltätigkeits-Lotterie. Er verkauft 300 Lose für CHF 3.- pro Los. Die Wahrscheinlichkeit das jemand gewinnt ist 0.2. Jeder Preis kostet CHF 8.-. Der Profit wird einem Wohltätigkeits-Verein gespendet. Wieviel Geld spendet Herr Borger dem Verein?

Der Wert muss zwischen 420 und 420 liegen

Bestmögliche Lösung:

Im Quiz zur Serie 03 haben wir Wetterdaten aus Madrid analysiert. Laden Sie erneut die `csv` Datei "`MadridWetterdaten1997-2015`" und bestimmen Sie für alle sechs möglichen Paare der Daten `Temperatur`, `rel.Luftfeuchtigkeit`, `Luftdruck` und `Windgeschwindigkeit` den Korrelationskoeffizienten.

Hinweis: *Weil zum Teil Daten fehlen, müssen wir beim Berechnen des Korrelationskoeffizienten das Argument* `use = "complete.obs"` *setzen.*

Welche Behauptungen sind wahr?

- [ ] Alle sechs Korrelationskoeffizienten sind negativ.
- [x] Der stärkste lineare Zusammenhang besteht zwischen `Temperatur` und `rel.Luftfeuchtigkeit`.
- [x] Zwischen `Luftdruck` und `Windgeschwindigkeit` gibt es einen schwachen negativen linearen Zusammenhang.
- [ ] Der Korrelationskoeffizient zwischen `Luftfeuchtigkeit` und `Luftdruck` zeigt, dass es keinen Zusammenhang zwischen diesen Grössen gibt.

Bestmögliche Lösung:

Ein Sack enthält 10 Holzscheiben. Die Scheiben sind beschriftet mit den Zahlen 1 bis 10. Eine Scheibe wir zufällig aus dem Sack gezogen. Die Wahrscheinlichkeit, als Dezimalzahl, dass die gezogene Scheibe

- die Zahl 3 ist, beträgt 0.1 ,
- kleiner als 4 ist, beträgt 0.3 ,
- eine quadratische Zahl ist, beträgt 0.3 ,
- eine Primzahl ist, beträgt 0.4 .

Bestmögliche Lösung:

Eine Regressionsgerade $y = a + bx$ beschreibt eine Korrelation. Kreuzen Sie alle richtigen Aussagen an:

- [ ] Wenn $y = 1$ für alle x eines Datensatzes, dann besteht keine Korrelation.
- [x] Bei einem negativen Achsenabschnitt ist $y < 0$ wenn $x = 0$
- [ ] Gegeben ist ein Datenpunkt (x_4, y_4) und das Residium $r_4 < 0$. Es ist daher zwingend dass $y_4 > a + bx_4$.
- [ ] Bei einer negativen Korrelation ist $a < 0$

Serie 05

Welche der folgenden Behauptungen s

- [x] $P(X > 5) = 0.5$
- [x] $P(X = 8) = 0$
- [ ] $P(X = 5) = 0.5$
- [ ] $P(X \geq 10) = 0$

Swisslos verkauft jedes Jahr in der Weihnachtszeit ein **Adventslos**, bei dem unterschiedliche **Geldbeträge** k ausbezahlt werden. Aus den Angaben zur Losauflage kann man die **Wahrscheinlichkeiten** $P(X = k)$ für die Auszahlung k eines zufällig gekauften Loses ausrechnen. Die Daten zu den Auszahlungen k und ihre zugehörigen Wahrscheinlichkeiten P werden in R als Vektoren wie folgt eingelesen:

```
k <- c(0,100,150,200,250,300,350,400,450,500,550,600,650,1000,1100,5000,100000,1000000)
P <- c(0.662256,0.28934,0.015883,0.029239,0.00085,0.000132,
       0.000132,0.000199,0.00011,0.000806,0.00011,0.000066,
       0.000022,0.000773,0.000055,0.000006,0.000015,0.000006)
```

Kopieren Sie diese Zuweisungen in R und berechnen Sie ***mit einer Genauigkeit auf zwei Stellen nach dem Komma*** für die Zufallsvariable X der Auszahlungen

- den **Erwartungswert**: 46.47255
- und die **Standardabweichung**: 2480.656
- Ein Los kostet 100 Franken. Der **durchschnittlich zu erwartende Verlust pro Los** ist also: 53.52745

Bereits in altägyptischen Gräbern von 3500 v. Chr. wurden **manipulierte Würfel** gefunden. Aufgrund der Lage des Schwerpunkts hat jeder solche Würfel eine andere Wahrscheinlichkeitsverteilung für seine Auflageflächen. Die Wahrscheinlichkeiten für die Augenzahlen k bei einem Würfel aus dem antiken Rom sind:

k	1	2	3	4	5	6
P(X = k)	0.01	0.11	0.14	0.24	0.17	0.33

Berechnen Sie folgende Wahrscheinlichkeiten ***exakt als Dezimalzahl*** zwischen 0 und 1:

- Beim Werfen des Würfels kommt eine ungerade Zahl mit einer Wahrscheinlichkeit von 0.32 vor.
- Die Wahrscheinlichkeit $P(X \geq 5 \text{ und } X \leq 2)$ beträgt 0 .
- Die Wahrscheinlichkeit $P(3 \leq X < 5)$ beträgt 0.38 .
- Höchstens eine Zahl 4 zu werfen kommt bei diesem Würfel mit Wahrscheinlichkeit 0.5 vor.

Serie 06

Angenommen die beiden Ereignisse A und B seien stochastisch unabhängig. Kreuzen Sie an, welche Aussagen richtig sind.

- [x] $P(A|B) = P(A)$
- [x] $P(A \cap B) = P(A) \cdot P(B)$
- [ ] Generell gilt, wenn A eintritt, dann ist $P(B) = 0$.
- [ ] Da stochastisch unabhängig, gilt $A \cap B = \{\}$ und daher $P(A \cap B) = 0$.

Ein Unternehmen, das Handy-Batterien herstellt, hat zwei Fabriken: Fabrik A und Fabrik B. Fabrik A produziert 60% der Batterien, während Fabrik B die restlichen 40% produziert. Die Batterien von Fabrik A haben eine Ausfallrate von 2%, während die Batterien von Fabrik B eine Ausfallrate von 3% haben. Wenn zufällig eine Batterie aus dem Bestand des Unternehmens ausgewählt wird und sich als defekt erweist, wie hoch ist die Wahrscheinlichkeit, dass sie in Fabrik B hergestellt wurde?

Der Wert muss zwischen 0.4999 und 0.5001 liegen

Angenommen in einem totalitärem Regime werden alle Personen eingesperrt, die demonstrieren. Weiter nehme man an, die Wahrscheinlichkeiten, dass eine beliebige Person eingesperrt ist oder demonstriert seien beide 0.01. Wir bezeichnen das Ereignis dass jemand eingesperrt ist als E und das Ereignis dass jemand demonstriert als D. Kreuzen Sie alle richtigen Aussagen an. Tip: Um die richtige Antwort einer der Aussagen herzuleiten, verwenden Sie am besten das Gesetz der Totalen Wahrscheinlichkeit, und lösen daraus für die bedingte Wahrscheinlichkeit.

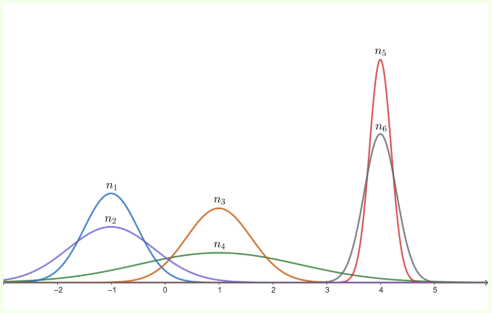
- [ ] E und D sind stochastisch unabhängig.
- [x] $P(E|\bar{D}) = 0$
- [x] $P(E) = 0.01$
- [ ] $P(E) = 0.99$

Serie 07

In einer gross angelegten Studie konnte 2015 gezeigt werden, dass oszillometrische Messgeräte den Blutdruck angenähert **normalverteilt** messen. Beim **systolischen Wert** haben diese Geräte eine Standardabweichung von **4.4 mmHg** und beim **diastolischen Wert** eine Standardabweichung von **3.4 mmHg**.

1. Mit welcher Wahrscheinlichkeit (Zahl zwischen 0 und 1; auf vier Stellen nach dem Komma gerundet) misst ein solches Messgerät bei einem Mann mit einem systolischen Blutdruck von 120 mmHg Werte grösser als 130 mmHg? 0.0115
2. Mit welcher Wahrscheinlichkeit (Zahl zwischen 0 und 1; auf vier Stellen nach dem Komma gerundet) misst ein solches Messgerät bei einer Frau mit einem diastolischen Blutdruck von 80 mmHg Werte zwischen 78 und 82 mmHg? 0.4436
3. In welchem symmetrischen Bereich (Werte gerundet auf zwei Stellen nach dem Komma) um 120 mmHg beim systolischen bzw. um 80 mmHg beim diastolischen Blutdruck liegen die Werte eines oszillometrischen Messgeräts zu 90% Wahrscheinlichkeit?
   - Beim systolischen Blutdruck im Bereich von 112.76 bis 127.24 .
   - Beim diastolischen Blutdruck im Bereich von 74.41 bis 85.59 .

Gegeben sind die Grafen von sechs verschiedenen Normalverteilungen nummeriert von n_1 bis n_6:



Ordnen Sie die folgenden **Dichtefunktionen** diesen Normalverteilungen zu:

Normalverteilung Nummer 1	passt zu	$\frac{1}{0.5\sqrt{2\pi}} e^{-0.5(\frac{x}{0.5})^2}$
Normalverteilung Nummer 2	passt zu	$\frac{1}{0.8\sqrt{2\pi}} e^{-0.5(\frac{x}{0.8})^2}$
Normalverteilung Nummer 3	passt zu	$\frac{1}{0.6\sqrt{2\pi}} e^{-0.5(\frac{x}{0.6})^2}$
Normalverteilung Nummer 4	passt zu	$\frac{1}{1.5\sqrt{2\pi}} e^{-0.5(\frac{x}{1.5})^2}$
Normalverteilung Nummer 5	passt zu	$\frac{1}{0.2\sqrt{2\pi}} e^{-0.5(\frac{x}{0.2})^2}$
Normalverteilung Nummer 6	passt zu	$\frac{1}{0.3\sqrt{2\pi}} e^{-0.5(\frac{x}{0.3})^2}$

Serie 08

Ungefähr 10 % der Menschen sind Linkshänder. Wenn wir der Linkshändigkeit den Wert 1 und der Rechtshändigkeit den Wert 0 zuweisen, dann folgt die Wahrscheinlichkeitsverteilung der Linkshändigkeit für die gesamte Bevölkerung einer Bernoulli-Verteilung. Die Standardabweichung der Population beträgt 0.3. Sie machen eine Studie über Linkshänder in einem bestimmten Land und wählen zufällig 100 Personen aus. Kreuzen Sie die richtigen Antworten an.

- [ ] Der Zentrale Grenzwertsatz kann nicht angewendet werden, da die Anzahl Linkshänder nicht normalverteilt sind.
- [x] Die Wahrscheinlichkeit dass nur 9.5% der Personen linkshändig sind beträgt 0.4338
- [ ] Die Wahrscheinlichkeit dass nur 9.5% der Personen linkshändig sind beträgt 0.0478

Wenn die Varianz von $\bar{X}$ gleich 25 ist und die Stichprobengrösse $n = 6$, beträgt die Standardabweichung der ursprünglichen Verteilung (Populationsstandardabweichung):

- (•) 150
- ( ) Kann nicht bestimmt werden.
- ( ) 12.25
- ( ) 2.04
- ( ) 4.17
- ( ) 25
- ( ) 5

Welche der folgenden Aussagen trifft auf den Standardfehler des Mittelwerts zu?

- ( ) Er ist kleiner als die Standardabweichung der Population.
- ( ) Er verringert sich, wenn die Stichprobengröße zunimmt.
- ( ) Er misst die Variabilität des Mittelwerts von Stichprobe zu Stichprobe.
- (•) Alle oben genannten
- ( ) Keine der oben genannten

Bei der Überprüfung der von einem Unternehmen ausgestellten Rechnungen stellt ein Prüfer fest, dass die Rechnungsbeträge einen Mittelwert von CHF 1'732 und eine Standardabweichung von CHF 298 aufweisen. Wie hoch ist die Wahrscheinlichkeit, dass der durchschnittliche Rechnungsbetrag für eine Stichprobe von 45 Rechnungen größer als CHF 1'800 ist?

- ( ) Kann ohne den Populationsmittelwert nicht bestimmt werden.
- ( ) 0.437
- ( ) 0.563
- (•) 0.063
- ( ) 0.937

Serie 09

Gemäss einer Studie von 2020 nutzen Kinder zwischen 11 und 13 Jahren ihr Smartphone **durchschnittlich 14 Stunden pro Woche**. Ein Forscherteam geht davon aus, dass dieser Wert inzwischen deutlich grösser ist. Um das zu klären, werden erneut **50'069 Kinder** zwischen 11 und 13 Jahren in Deutschland zu ihrer Smartphone-Nutzung befragt.

Die durchschnittliche Smartphone-Nutzung in dieser Stichprobe beträgt **14.22914 Stunden pro Woche** mit einer Standardabweichung von **10.43579 Stunden**.

Nun wird ein ***einseitiger*** Test auf dem 5%-Signifikanzniveau gemacht (auch wenn die Standardabweichung aus den Daten geschätzt wurde, kann hier mit der Normalverteilung gerechnet werden. Für ein so grosses n ist der Unterschied zwischen der t-Verteilung und der Normalverteilung sehr gering):

Die Nullhypothese lautet: H_0 gleich 14 .

Die Alternativhypothese lautet: H_A grösser 14 .

Der p-Wert beträgt 0.0000004481 . (Geben Sie das Resultat in der Form 0.xxxxxxxxxx an, also auf 10 Nachkommastellen genau!)

Die Nullhypothese H_0 wird verworfen .

Eine Kornmühle hat eine neue automatisierte Abfüllanlage für die Verpackung von Mehl in 500 Gramm Säcke in Betrieb genommen. Sie will nun testen, ob die Anlage korrekt abfüllt und macht zufällig 9 Stichproben:

501, 502, 500, 499, 501, 498, 500, 499, 501 (in Gramm)

Die Kornmühle geht davon aus, dass die Standardabweichung der Abfüllanlage dem Wert in der Spezifikation entspricht. Sie beträgt σ = 1 Gramm. Nun wird ein ***zweiseitiger*** Test auf dem 5%-Signifikanzniveau gemacht:

Die Nullhypothese lautet: H_0 gleich 500 .

Die Alternativhypothese lautet: H_A ungleich 500 .

Der Verwerfungsbericht ist $K = (-\infty,$ 499.3467 $) \cup ($ 500.6533 $, \infty)$. (Geben Sie das Resultat auf 6 Nachkommastellen an!)

Der Stichprobenmittelwert beträgt 500.111111 . (Geben Sie das Resultat auf 6 Nachkommastellen an!)

Die Nullhypothese H_0 wird nicht verworfen .

Serie 10

Ein Sportverband möchte testen, ob ein von Profisportlern verwendetes Nahrungsergänzungsmittel aufgrund der Erhöhung des Testosteronspiegels im Körper verboten werden sollte.

Die Testosteronwerte in Pikogramm/Milliliter von zehn Athleten wurden vor und nach der Einnahme des Nahrungsergänzungsmittels getestet, und die Ergebnisse sind in der folgenden Daten zusammengefasst:

Vorher (pg/ml) 65.83, 111.15, 106.18, 91.12, 97.43, 135.89, 69.45, 83.33, 157.88, 74.69

Nacher (pg/ml) 77.92, 129.27, 109.72, 97.68, 124.37, 147.12, 71.16, 81.27, 164.16, 79.51

Führen Sie einen Hypothesentest mit einem Signifikanzniveau von 1 % durch, um zu entscheiden, ob das Nahrungsergänzungsmittel verboten werden sollte oder nicht.

- ☐ Wir machen einen ungepaarten Test, da man beliebig viele Athleten wählen könnte.
- ☐ Wir machen einen zweiseitigen Test.
- ☑ Das Nahrungsergänzungsmittel sollte verboten werden.

Betrachten Sie einen einseitigen t-Test von $H_0 : \mu = 0$ gegen $H_A : \mu > 0$ zum Niveau $\alpha = 0.05$.

Die beobachteten *n* Datenpunkten haben einen empirischen Mittelwert grösser Null, so dass die **Nullhypothese verworfen wird**.

Entscheiden Sie, ob folgende Aussagen wahr oder falsch sind.

- ☑ Der *p*-Wert ist kleiner als 0.05
- ☑ Es gibt ein Niveau $\alpha < 1$, wo die Nullhypothese **nicht** verworfen wird.
- ☑ Man verwirft H_0 für ein Niveau $\alpha > 0.05$
- ☑ Wird ein zweiseitiger Test gemacht, könnte es je nach empirischen Mittelwert sein, dass die Nullhypothese **nicht** verworfen wird.

Serie 11

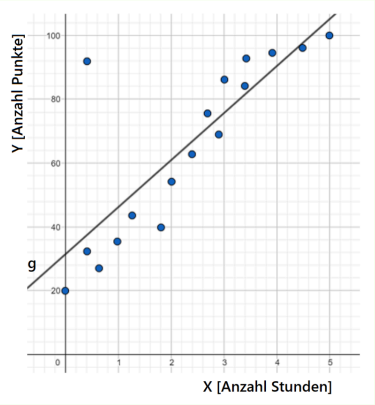
Eine Kellnerin möchte wissen, ob es einen Zusammenhang zwischen dem Rechnungsbetrag und dem gespendeten Trinkgelt gibt. Dazu sammelt sie Daten und macht eine lineare Regression. Die Ausgabe von R sieht wie folgt aus:

```
> summary(lm(Trinkgeld ~ Rechnungsbetrag))

Call:
lm(formula = Trinkgeld ~ Rechnungsbetrag)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8711 -1.1576 -0.0869  1.1578  5.7282

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.133455   0.363644   0.367    0.714
Rechnungsbetrag 0.174904   0.007034  24.865   <2e-16
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Residual standard error: 2.141 on 93 degrees of freed
Multiple R-squared:  0.8692,    Adjusted R-squared:
F-statistic: 618.3 on 1 and 93 DF,  p-value: < 2.2e-1
```

Wählen Sie alle wahren Behauptungen!

- ☑ Die Wahrscheinlichkeit, dass der Rechnungsbetrag keinen Einfluss auf das Trinkgeld hat, ist kleiner als 0.05.
- ☑ Im R Output steht, dass mit diesem Modell mehr als 86% von der Varianz im **Trinkgeldbetrag** erklärt wird.
- ☑ Die Nullhypothese, dass der Rechnungsbetrag keinen Einfluss auf das gespendete Trinkgelt hat, kann verworfen werden.
- ☑ Mit der Formel **Trinkgeld** $\approx 0.133455 + 0.174904 \cdot$ **Rechnungsbetrag** kann die Kellnerin aus dem Rechnungsbetrag ihr Trinkgeld schätzen.
- ☐ Pro 1 Franken mehr auf dem Rechnungsbetrag erhält die Kellnerin durchschnittlich 0.133455 Franken mehr Trinkgeld.
- ☐ Im R Output steht, dass mit diesem Modell mehr als 86% von der Varianz im **Rechnungsbetrag** erklärt wird.

Vor einer Prüfung wurden alle Teilnehmer gefragt, wie lange sie für die Prüfung geübt haben. Diese Angaben **X** wurden dann mit der erreichten Anzahl Punkte **Y** verglichen.



Beantworten Sie die folgenden Fragen zu den oben im Bild gezeigten Ergebnissen:

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch	
◉	○	Es hat mindestens eine Person mit 40 Punkten angegeben, dass sie 1.8 Stunden geübt hat.
○	◉	Die eingezeichnete Gerade **g** kann nicht die Regressionsgerade aus der Berechnung mit **lm(...)** in R sein, weil zu viele Datenpunkte unterhalb dieser Geraden **g** liegen.
○	◉	Der Koeffizient β_1 in der Regressionsgerade ist eine Zahl zwischen 0.5 und 5.
◉	○	Eine Person, die für so eine Prüfung nichts lernt, macht ***schätzungsweise*** 32 Punkte.

Serie 12

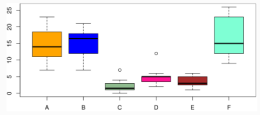
In einer multiplen Regressionsanalyse, wenn das Modell eine schlechte Anpassung zeigt, weist dies darauf hin, dass:

- ○ die Summe der Fehlerquadrate groß sein wird
- ○ der Standardfehler RSE der Schätzung groß sein wird. Der RSE is ähnlich der Standardabweichung, wobei hier die Definition $\mathrm{RSE} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-1}}$ gilt.
- ○ der multiple Bestimmtheitskoeffizient R^2 nahe null sein wird
- ◉ Alle oben genannten

Wenn die Varianz des Modells gleich 35 ist, und die Varianz Sample gleich 49 ist, dann ist die Varianz Differenz (RSS) gleich

- ○ 18
- ◉ 14
- ○ 12
- ○ 10

Wir testen Insektensprays. Dabei wurden 6 verschiedene Insektensprays verwendet, die auf verschiedenen Feldern versprüht wurden. Danach wurde die Anzahl Insekten gezählt, die sich auf dementsprechenden Feld nach dem Besprühen befanden. Je kleiner die Anzahl, umso wirksamer der Spray.



richtig	falsch	
◉	○	Für Spray F sind die Hälfte der Messwerte etwa 15 oder kleiner.
○	◉	Etwa 75% der Messwerte von Spray A sind ungefähr 7 oder höher.
○	◉	Spray A ist wirksamer als Spray C
◉	○	50% der Messwerte von Spray A liegen zwischen etwa 11 und 18

richtig	falsch	
◉	○	Für Spray B sind 25% der Messwerte ungefähr 18 oder grösser
○	◉	Ungefähr 25% der Messwerte von Spray F sind ungefähr 12 oder grösser
○	◉	Spray C hat die grössere Streuung bezüglich Interguartilsdifferenz als Spray F
◉	○	Ungefähr 50% der Messwerte von Spray F sind zwischen ungefähr 12 und 23

Bei einem Zufallsexperiment werden ein roter (r) und ein blauer (b) Würfel gleichzeitig geworfen. Wir nehmen an, dass sie „fair" sind, d. h. die Augenzahlen 1 bis 6 eines Würfels treten mit gleicher Wahrscheinlichkeit auf.

Beachten Sie: Falsche Antworten ergeben Punkteabzug.

richtig	falsch	
◉	○	r_5b_1 ist ein mögliches Elementarereignis
○	◉	Die Wahrscheinlichkeit, dass das Produkt der Augenzahlen 7 ist, ist 1/6
◉	○	Die Wahrscheinlichkeit, dass die Augensumme grösser 2 ist, ist 35/36
○	◉	Die Wahrscheinlichkeit, dass der rote Würfel 5 ist, ist 1/36

richtig	falsch	
○	◉	7 ist ein mögliches Elementarereignis
○	◉	Die Wahrscheinlichkeit, dass das Produkt der Augenzahlen 7 ist, ist 6/36
◉	○	Die Wahrscheinlichkeit, dass die Augensumme kleiner oder gleich 11 ist, ist 35/
◉	○	Die Wahrscheinlichkeit, dass der rote Würfel 5 ist, ist 1/6

Ein Multiple-Choice-Test besteht aus 15 Fragen, mit jeweils 5 Antwortmöglichkeiten, von denen genau eine richtig ist. Die Wahrscheinlichkeit dafür, eine Aufgabe richtig zu beantworten, ist also 0.2. Die Wahrscheinlichkeits- und Verteilungsfunktion sind gegeben durch:

k	8	9	10	11	12	13	14	15
$P(X \leq k)$	0.711	0.939	0.949	0.982	0.989	0.992	0.999	1

Beachten Sie: Es handelt sich hier um die kumulierten Wahrscheinlichkeiten $P(X \leq k)$ und nicht $P(X = k)$.

richtig	falsch	
○	◉	Die Wahrscheinlichkeit $P(X \leq 9)$ ist 0.228
○	◉	Die Wahrscheinlichkeit $P(X \geq 12)$ ist 0.989
◉	○	Die Wahrscheinlichkeit, dass genau 13 Fragen richtig beantwortet werden, ist 0.003
◉	○	Die Wahrscheinlichkeit, dass höchstens 10 Fragen richtig beantwortet werden, ist 0.969

richtig	falsch	
○	◉	Die Wahrscheinlichkeit $P(X \leq 10)$ ist 0.030
○	◉	Die Wahrscheinlichkeit $P(X > 11)$ ist 0.989
○	◉	Die Wahrscheinlichkeit, dass genau 14 Fragen richtig beantwortet werden, ist 0.999
◉	○	Die Wahrscheinlichkeit, dass mindestens 13 Fragen richtig beantwortet werden, ist 0.011

Ein U.S. Magazin, Consumer Reports, führte eine Untersuchung des Kalorien- und Salzgehaltes von verschiedenen Hotdog-Marken durch. Es gab drei verschiedene Typen von Hotdogs: Rind, „Fleisch" (Rind, Schwein, Geflügel gemischt) und Geflügel.

Die Resultate unten führen den Kaloriengehalt verschiedener Marken von Rind- und Geflügel-Hotdogs auf.

Rinds-Hotdog: 186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132
Geflügel-Hotdog: 129, 132, 102, 106, 94, 102, 87, 99, 170, 113, 135, 142, 86, 143, 152, 146, 144

Haben die beiden Hotdog-Arten verschiedenen Kaloriengehalt? Wir führen einen Hypothesentest auf 5% Signifikanzniveau durch und erhalten folgenden Output:

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 285.5, p-value = 0.0004549
## alternative hypothesis: true location shift is not equal to 0
```

x enthält die Rindsdaten und y die Geflügeldaten.

Welche der folgenden Aussagen sind richtig?

richtig	falsch	
◉	○	Die Nullhypothese wird verworfen, da der p-Wert unter dem Signifikanzniveau liegt.
◉	○	Wir führen einen zweiseitigen Test durch.
○	◉	Die Alternativhypothese ist $\mu_x > \mu_y$
○	◉	Der Unterschied zwischen x und y ist nicht statistisch signifikant.

richtig	falsch	
◉	○	Wir nehmen an, dass die Daten nicht normalverteilt sind.
◉	○	Der Unterschied zwischen x und y ist statistisch signifikant.
◉	○	Die Alternativhypothese ist $\mu_x \neq \mu_y$
○	◉	Da die Stichprobengrössen unterschiedlich sind, führen wir einen gepaarten Test durch.

richtig	falsch	
◉	○	Die Nullhypothese wird verworfen, da der p-Wert unter dem Signifikanzniveau liegt.
◉	○	Wir führen einen zweiseitigen Test durch.
○	◉	Die Alternativhypothese ist $\mu_x > \mu_y$
○	◉	Der Unterschied zwischen x und y ist nicht statistisch signifikant.

Die MASS-Bibliothek enthält den Boston-Datensatz, der *medv* (median house value in $1000) für 506 Stadtviertel um Boston herum erfasst. Wir werden versuchen, medv mit 13 Prädiktoren wie *rm* (durchschnittliche Anzahl von Zimmern pro Haus), *age* (Durchschnittsalter der Häuser) und *crim* (Kriminalitätsrate) vorherzusagen.

Wir werden damit beginnen, die lm()-Funktion zu verwenden, um ein einfaches lineares Regressionsmodell mit *medv* als Zielvariable und *crim* als Prädiktor anzupassen. Wir erhalten folgenden Output:

```
> fit.fit = lm(medv ~ crim)
> summary(lm.fit)

##
## Call:
## lm(formula = medv ~ crim)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.957  -5.449  -2.007   2.512  29.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.03311    0.40914   58.74   <2e-16 ***
## crim        -0.41519    0.04389   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.484 on 504 degrees of freedom
## Multiple R-squared:  0.1508,Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

richtig	falsch	
○	◉	Das Modell lautet $\text{crim} = \beta_0 + \beta_1 \cdot \text{medv}$.
○	◉	Der Wert von $R^2 = 0.1508$ bedeutet, dass die Daten gut zum Modell passen.
○	◉	Der y-Achsenabschnitt ist -0.4152
◉	○	Die Steigung ist statistisch signifikant ungleich 0.

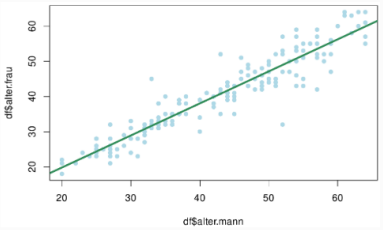
richtig	falsch	
◉	○	Das Modell lautet $\text{medv} = \beta_0 + \beta_1 \cdot \text{crim}$.
◉	○	Der Wert von $R^2 = 0.1508$ bedeutet, dass die Daten nicht gut zur Regressionsgeraden passen.
○	◉	Der y-Achsenabschnitt ist 24.03 bedeutet, dass pro zusätzlicher Prozentpunkt in Kriminalität, der Median des Hauspreises um etwa $24000 steigt.
○	◉	Die Steigung ist statistisch signifikant gleich 0.

richtig	falsch	
○	◉	Das Modell lautet $\text{medv} = \beta_0 + \beta_1 \cdot \text{latat}$.
○	◉	Der Wert von $R^2 = 0.1508$ bedeutet, dass die Daten nahe der Regressionsgeraden folgen.
◉	○	Der Steigung von -0.4142 bedeutet, dass pro zusätzlicher Prozentpunkt in Kriminalität, der Median des Hauspreises um etwa $400 fällt.
○	◉	Der y-Achsenabschnitt ist statistisch signifikant gleich 0.

Bitte lesen Sie die Aufgabe sorgfältig durch und überlegen Sie genau.

Wir haben aus eigener Erfahrung das Gefühl, dass bei Ehepaaren der Mann eher älter als die Frau ist. Nun wollen wir statistisch untersuchen, ob dem so ist.
In einer Untersuchung in England wurden das Alter (in Jahren) und die Körpergrösse (in cm) von 170 Ehepaaren untersucht.

Das Streudiagramm sieht wie folgt aus:



Die Regressionsgerade wird wie folgt bestimmt:

```
lm(df$alter.frau ~ df$alter.mann)

##
## Call:
## lm(formula = df$alter.frau ~ df$alter.mann)
##
## Coefficients:
##   (Intercept)  df$alter.mann
##         1.5740         0.9112
```

richtig	falsch	
○	◉	Aus dem Streudiagramm ist ein kein linearer Zusammenhang erkennbar.
◉	○	Die Regressionsgerade lautet $y = 1.574 + 0.9112x$
○	◉	Für jedes Jahr das der Ehemann älter ist, ist die Frau 1.574 Jahre älter
○	◉	Der Korrelationskoeffizient ist anähernd -1

richtig	falsch	
○	◉	Aus dem Streudiagramm ist ein quadratischer Zusammenhang erkennbar.
○	◉	Die Regressionsgerade lautet $y = 0.9112 + 1.574x$
○	◉	Für jedes Jahr das der Ehefrau älter ist, ist er Mann 0.911 Jahre älter
◉	○	Der Korrelationskoeffizient ist annähernd 1

Der Serumtest untersucht schwangere Frauen auf Babys mit Down-Syndrom. Der Serumtest ist ein sehr guter, aber nicht perfekter Test. Etwa 1 % der Babys haben das Down-Syndrom. Wenn das Baby das Down-Syndrom hat, besteht eine 90-prozentige Wahrscheinlichkeit, dass das Ergebnis positiv ausfällt. Wenn das Baby nicht betroffen ist, besteht immer noch eine 1-prozentige Wahrscheinlichkeit, dass das Ergebnis positiv sein wird. Eine schwangere Frau wurde getestet und das Ergebnis ist negativ. Wie gross ist die Wahrscheinlichkeit dass ihr Baby das Down-Syndrom nicht hat?

(Gerundet auf 5 Dezimalstellen)

- ○ 0.99859 ✗
- ○ 0.00141 ✗
- ◉ 0.99898 ✓
- ○ 0.00102 ✗

Für die Körpergrösse von 18-20jährigen Männern ergibt sich ein Mittelwert von 1.80 m bei einer Standardabweichung von 7.4 cm. Die Körpergrösse kann als normalverteilt angesehen werden.

Mit welcher Wahrscheinlichkeit ist ein zufällig ausgewählter Mann dieser Altersgruppe kleiner als 1.90cm?

X sei die Zufallsvariable für die Körpergrösse eines zufällig ausgewählten Mannes.

Welche der folgenden Aussagen beschreibt die gesuchte Wahrscheinlichkeit?

richtig	falsch	
◉	○	$P(X < 190)$
○	◉	1 - pnorm(q=190, mean=180, sd=7.4)
○	◉	qnorm(p=190, mean=180, sd=7.4)
◉	○	$1 - P(X \geq 190)$

Die Körpertemperatur von 10 Patienten wird zum Zeitpunkt der Verabreichung eines Medikaments (T 1) und 2 Stunden später (T 2) gemessen. Es soll geprüft werden, ob dieses Medikament eine fiebersenkende Wirkung hat.

Patient-Nr.	1	2	3	4	5	6	7	8	9	10
Temp. 1 in °C	39.1	39.3	38.9	40.6	39.5	38.4	38.6	39.0	38.6	39.2
Temp. 2 in °C	38.1	38.3	38.8	37.8	38.2	37.3	37.6	37.8	37.4	38.1

Wir führen einen Hypothesentest auf 5% durch um zu überprüfen, ob das Medikament fiebersenkend ist. Der R-Output zeigt:

```
##
##  Paired t-test
##
## data:  t.1 and t.2
## t = 5.6569, df = 9, p-value = 0.0001554
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7976252       Inf
## sample estimates:
## mean of the differences
##                    1.18
```

richtig	falsch	
○	◉	Wir führen einen zweiseitigen Test durch.
○	◉	Die Nullhypothese ist, dass das Medikament Wirkung hat.
◉	○	Da der p-Wert unter dem Signifikanzniveau liegt wird die Alternativhypothese angenommen.
◉	○	Der Wert 1.18 in der letzten Linie bedeutet, dass die durchschnittliche Temperatur um 1.18 ° C nach zwei Stunden tiefer war.

richtig	falsch	
○	◉	Wir führen einen ungepaarten Test durch.
◉	○	Die Nullhypothese ist, dass das Medikament keine Wirkung hat.
○	◉	Da 0 nicht im Vertrauensintervall ist, wird die Nullhypothese nicht verworfen.
○	◉	Der Wert 1.18 in der letzten Linie bedeutet, dass die durchschnittliche Temperatur um 1.18 ° C nach zwei Stunden grösser war

richtig	falsch	
○	◉	Wir führen einen gepaarten Test durch, da beide Stichproben gleich gross sind.
○	◉	Die Alternativhypothese ist, dass das Medikament keine Wirkung hat.
◉	○	Da 0 nicht im Vertrauensintervall ist, wird die Nullhypothese verworfen.
◉	○	Der Wert 1.18 in der letzten Linie bedeutet, dass die durchschnittliche Temperatur um 1.18 ° C nach zwei Stunden tiefer war.

Die MASS-Bibliothek enthält den Boston-Datensatz, der medv (median house value in $1000) für 506 Stadtviertel um Boston herum erfasst. Wir werden versuchen, medv mit 13 Prädiktoren wie rm (durchschnittliche Anzahl von Zimmern pro Haus) , di (Distanz vom Zentrum von Boston) und crim (Anteil der Kriminalität) vorherzusagen.

Wir passen ein multiples lineares Regressionsmodell mit der Zielvariable medv und den Prädiktoren crim, dis und rm. Das Signifikanzniveau ist 5%. Wir erha ten folgenden Output:

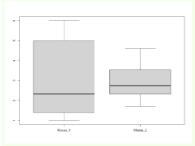
```
fit <- lm(medv ~ crim + rm + dis, data = Boston)

summary(fit)

##
## Call:
## lm(formula = medv ~ crim + rm + dis, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.247  -2.930  -0.570   2.390  39.072
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -29.45838    2.60010 -11.330  < 2e-16 ***
## crim         -0.25405    0.03532  -7.193 2.32e-12 ***
## rm            8.34257    0.40870  20.413  < 2e-16 ***
## dis           0.14627    0.14392   0.870    0.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.238 on 502 degrees of freedom
## Multiple R-squared:  0.5427,Adjusted R-squared:  0.5399
## F-statistic: 198.6 on 3 and 502 DF,  p-value: < 2.2e-16
```

◉	○	Der p-Wert für rm ist kleiner als das Signifikanzniveau und somit wird die Nullhypothese $\beta_2 = 0$ verworfen.
◉	○	Der p-Wert 0.38 für dis bedeutet, dass es keinen statistisch signifkanten Unterschied im Medianwert gibt, ob man nahe oder weit vom Zentrum entfernt lebt.
○	◉	Der Koeffizient für dis bedeutet, dass pro Mile Abstand mehr vom Zentrum von Boston, der Medianpreis um etwa $126 fällt.
○	◉	Der Wert $R^2 = 0.54$ ist statistisch signifikant.

richtig	falsch	
○	◉	Der p-Wert für dis ist grösser als das Signifikanzniveau und somit wird die Nullhypothese $\beta_3 = 0$ verworfen.
○	◉	Der p-Wert um F-Wert ist signifikant und somit hängt kein Prädiktor mit der Zielvariable zusammen.
◉	○	Der Koeffizient für dis bedeutet, dass pro Mile Abstand mehr vom Zentrum von Boston, der Medianpreis um etwa $126 steigt.
○	◉	Der Wert $R^2 = 0.54$ ist statistisch nicht signifikant.

richtig	falsch	
○	◉	Der p-Wert für dis ist grösser als das Signifikanzniveau und somit wird die Alternativhypothese $\beta_3 = 0$ angenommen.
◉	○	Der p-Wert um F-Wert ist signifikant und somit hängt mindestens ein Prädiktor mit der Zielvariable zusammen.
◉	○	Der Koeffizient für rm bedeutet, dass pro Zimmer mehr, der Medianpreis um etwa $8300 steigt.
○	◉	Der Wert 0.54 bedeutet, dass etwas 50% der Variablen einen Einfluss auf medv haben.

Bestmögliche Lösung:

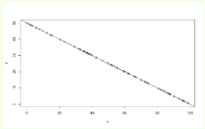


Betrachten Sie den Boxplot oben für die Noten zweier Schulklassen. Beantworten Sie, ob die unteren Aussagen wahr oder falsch sind.

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch	
◉	○	Die Median-Note von Klasse 2 ist höher als die von Klasse 1.
○	◉	Die Quartilsdifferenz von Klasse 1 ist kleiner 2.
○	◉	Beide Datensätze sind normalverteilt.
◉	○	Beide Datensätze enthalten keine Ausreisser.

Bestmögliche Lösung:



In einem Scatterplot werden die Grössen x und y gegenübergestellt. Alle Punkte liegen auf einer absteigenden Geraden, wobei diese Gerade auch den durch die lineare Regression ermittelten Schätzwerten für y entspricht. Für diese Schätzung wurden R^2 und RSS berechnet.

Welche Aussagen sind hier wahr?

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch	
○	◉	der RSS beträgt 1.0.
○	◉	der R^2 beträgt -1.0.
◉	○	Die Korrelation zwischen X und Y beträgt -1.
◉	○	Gemäss dem Regressionsmodell hat die Grösse x Einfluss auf y.

Bestmögliche Lösung:

Angenommen die Wahrscheinlichkeit, dass es an einem beliebigen Tag regnet (Ereignis R) sei 0.05 .
Die Wahrscheinlichkeit, dass der Rasen an einem beliebigen Tag gesprengt wird (Ereignis G) sei 0.1, wobei der Rasen nie gesprengt wird, wenn es an diesem Tag auch regnet.

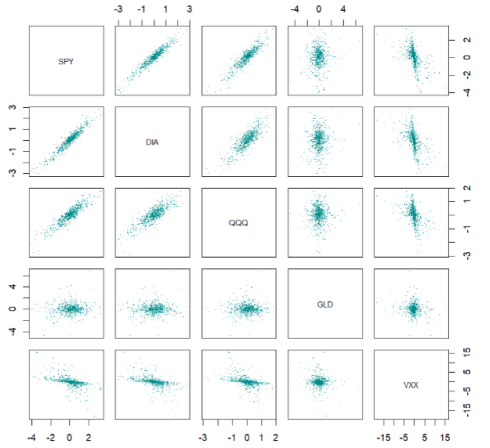
Welche Aussagen sind wahr?

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch	
○	◉	R und G sind stochastisch unabhängig.
◉	○	$P(R \cup G) = 0.15$
◉	○	$P(R\|G) = 0$
◉	○	$P(G\|R) = 0$

Bestmögliche Lösung:

Im Datensatz EFTs sind Tagesrenditen von den fünf börsengehandelten Indexfonds SPY, DIA, QQQ, GLD und VXX gespeichert. Mit der Funktion pairs() wurden Streudiagramme für die Variablen des Datensatzes EFTs erstellt.



Ordnen Sie die folgenden Werte korrekt zu:

round(cor(EFTs$SPY, EFTs$DIA), 2)	passt zu	0.95
round(cor(EFTs$QQQ, EFTs$GLD), 2)	passt zu	0.04
round(cor(EFTs$SPY, EFTs$VXX), 2)	passt zu	-0.55
round(cor(EFTs$QQQ, EFTs$DIA), 2)	passt zu	0.83

Bestmögliche Lösung:

Die National Collegiate Athletic Association hat ein neues Trainingsprogram entwickelt, das die Sprunghöhe von Basketballspielern erhöhen soll. Um die Wirksamkeit des Programs zu testing, wurden zufällig 12 College Basketballspieler rekrutiert, und deren Sprunghöhe vor und nach dem einmonatigen Training gemessen.

Die folgenden Daten zeigen für jeden Spieler die maximale Sprunghöhe (in Zoll) vor und nach dem Training:

Vorher: 22, 24, 20, 19, 19, 20, 22, 25, 24, 23, 22, 21

Nachher: 23, 25, 20, 24, 18, 22, 23, 28, 24, 25, 24, 20

Der folgende Code zeigt wie der Hypothesentest in R ausgeführt wurde:

```
#define before and after max jump heights
before <- c(22, 24, 20, 19, 19, 20, 22, 25, 24, 23, 22, 21)
after <- c(23, 25, 20, 24, 18, 22, 23, 28, 24, 25, 24, 20)

#perform paired samples t-test
t.test(x = before, y = after, paired = TRUE)

        Paired t-test

data:  before and after
t = -2.5289, df = 11, p-value = 0.02803
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.3379151 -0.1620849
sample estimates:
mean of the differences
                  -1.25
```

Beantworten Sie die folgenden Fragen als Richtig oder Falsch:

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch	
○	◉	Die Nullhypothese sagt aus, dass das Program wirksam ist.
○	◉	Da der Vertrauensintervall als 95% angegeben ist, wird alpha = 0.025 verwendet.
○	◉	Der p-Wert liegt über alpha und die Nullhypothese wird verworfen.
◉	○	Bei einem Vertrauensintervall von 99% ist die Obergrenze des Intervalls positiv.

Bestmögliche Lösung:

Aus einer Zusammenstellung des Erfolgs der Wetterprognosen ist bekannt, dass der Wetterbericht

- mit Wahrscheinlichkeit $\frac{1}{4}$ schönes Wetter fälschlicherweise als schlecht voraussagt und
- mit Wahrscheinlichkeit $\frac{4}{5}$ schlechtes Wetter korrekt als schlecht voraussagt.

Eine Person glaubt mit Wahrscheinlichkeit $\frac{2}{3}$, dass das Wetter Morgen schön wird. Dann hört sie den Wetterbericht sagen, dass es schlecht wird. Wie sollte die Person dadurch ihre subjektive Wahrscheinlichkeit für schönes Wetter ändern?

Wir bezeichnen mit A das Ereignis, dass es Morgen schön sein wird und mit B das Ereignis, dass der Wetterbericht schlechtes Wetter voraussagt.

Die Person erhält fünf Vorschläge für diese neue Wahrscheinlichkeit:

- Vorschlag (a): $P(A|B) = \frac{\frac{2}{3} \cdot \frac{1}{4}}{\frac{2}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{4}{5}}$
- Vorschlag (b): $P(A|B) = \frac{2}{3} \cdot \frac{1}{4}$
- Vorschlag (c): $P(A|B) = \frac{\frac{2}{3} \cdot \frac{1}{4}}{\frac{2}{3} \cdot \frac{1}{4} + \frac{2}{4} \cdot \frac{1}{5}}$
- Vorschlag (d): $P(A|B) = \frac{1}{3} \cdot \frac{4}{5}$
- Vorschlag (e): $P(A|B) = \frac{\frac{2}{3}}{\frac{1}{4} + \frac{4}{5}}$

Welche dieser Rechnungen ist für die neue Wahrscheinlichkeit korrekt?

○ Vorschlag (e)
○ Vorschlag (d)
○ Vorschlag (c)
○ Vorschlag (b)
◉ Vorschlag (a)

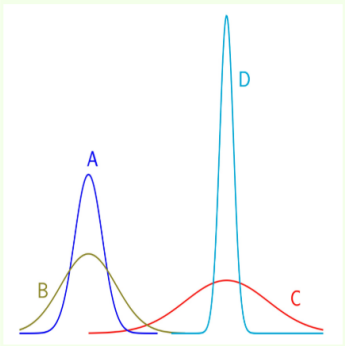
Bestmögliche Lösung:

Eine einfache lineare Regression ergibt die Gerade $\hat{y} = 32 + 0.4x$. Eines der Subjekte, Elisabet, hat $x = 60$ und $y = 52$. Berechne das Residuum von Elisabeth.

Der Wert muss zwischen -4.01 und -3.99 liegen

Bestmögliche Lösung:

Ordnen Sie jedem Parameterpaar die richtige Normalverteilungskurve zu. Dazu ziehen Sie mit der Maus das passende Parameterpaar auf den Buchstaben auf der linken Seite.



A	passt zu	mean = 5, sd = 1
B	passt zu	mean = 5, sd = 2
C	passt zu	mean = 15, sd = 3
D	passt zu	mean = 15, sd = 0.5

Bestmögliche Lösung:

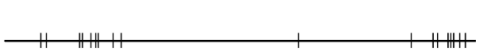
Wir führen eine einfache, lineare Regression mit R aus, wobei wir $H_0 : \beta_1 = 0$ gegen $H_A : \beta_1 \neq 0$ testen. Die R Ausgabe gibt einen p– Wert von 0.015 an. Wir schliessen daraus dass

A. $H_0 : \beta_1 = 0$ ist wahr mit Wahrscheinlichkeit 0.015
B. $H_0 : \beta_1 = 0$ wird verworfen mit $\alpha = 0.05$
C. $H_A : \beta_1 \neq 0$ ist wahr mit Wahrscheinlichkeit 0.015
D. $H_A : \beta_1 \neq 0$ wird verworfen mit $\alpha = 0.05$
E. Beides, A und B.

○ A
◉ B
○ C
○ D
○ E

Bestmögliche Lösung:

In der Abbildung sehen wir die numerischen Werte eines eindimensionalen Datensatzes als dünne Striche auf der Zahlenachse dargestellt. Die Daten wurden mit einem Gerät gemessen, das Werte auf Millimeter rundet.



Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch	
○	◉	Aus der graphischen Darstellung ist immer klar, wie viele Messungen der Datensatz enthält.
○	◉	Die Quartilsdifferenz der Daten ändert sich nicht, wenn wir einen Wert im Datensatz (egal welchen) abändern.
◉	○	Der Mittelwert der Daten ändert sich, wenn wir einen Wert im Datensatz (egal welchen) abändern.
◉	○	Die Daten sind nicht normalverteilt.

Bestmögliche Lösung:

Ein Datensatz zu Prüfungsergebnissen der Berufsmaturität wird untersucht. Es wurde ein multiples lineares Modell mit der Zielvariablen BMSchnitt (Abschlussnote von 1 bis 6) und den Prädiktoren IQ, Motivation, Grösse und Geschlecht angepasst. Der R Output sieht wie folgt aus:

```
Call:
lm(formula = BMSchnitt ~ IQ + Motivation + Größe + Geschlecht, data = BM)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70065 -0.13915  0.01281  0.16483  0.49899

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.305073   0.714277   0.427    0.671
IQ             0.037845   0.004435   8.533 4.91e-11 ***
Motivation     0.150587   0.024609   6.119 1.92e-07 ***
Größe         -0.434140   0.336441  -1.290    0.203
GeschlechtMann -0.068264   0.090344  -0.756    0.454
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2743 on 46 degrees of freedom
Multiple R-squared:  0.9056,    Adjusted R-squared:  0.8974
F-statistic: 110.4 on 4 and 46 DF,  p-value: < 2.2e-16
```

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch	
○	◉	Gemäss den Angaben (Signifikanz ausser acht gelassen) schliessen Frauen schlechter ab als Männer.
○	◉	Wenn der IQ um eine Einheit zunimmt, steigt der BMSchnitt um 0.037845 Prozent.
◉	○	Die Prädiktorvariable Motivation liegt ausserhalb des Bereichs, bei dem von einer zufälligen Verteilung zwischen Prädiktorvariable und Zielvariable ausgegangen werden kann.
◉	○	Der Anteil der Varianz in den Daten, der durch dieses Modell erklärt wird, beträgt ca. 90 Prozent.

Bestmögliche Lösung:

Die Einkommensverteilung in einigen Ländern der Dritten Welt gilt als keilförmig (viele sehr arme Menschen, sehr wenige Menschen mit mittlerem Einkommen und noch weniger wohlhabende Menschen). Angenommen, wir wählen ein Land mit einer keilförmigen Verteilung aus. Lassen Sie das durchschnittliche Gehalt 2000 $ pro Jahr mit einer Standardabweichung von 8000 $ betragen. Wir befragen zufällig 1000 Einwohner dieses Landes und erhalten einen Stichprobenmittelwert $\bar{X}$. Berechnen Sie die Wahrscheinlichkeit

$P(2100 < \bar{X} < 2200)$

auf mindestens vier Stellen nach dem Komma genau (maximal 15 Zeichen erlaubt).

Der Wert muss zwischen 0.13155 und 0.13185 liegen