Statistische Datenanalyse mit R, Teil 3 online, Statistische Tests

Dr. Andrea Denecke
Leibniz Universität IT-Services

Testdatensätze

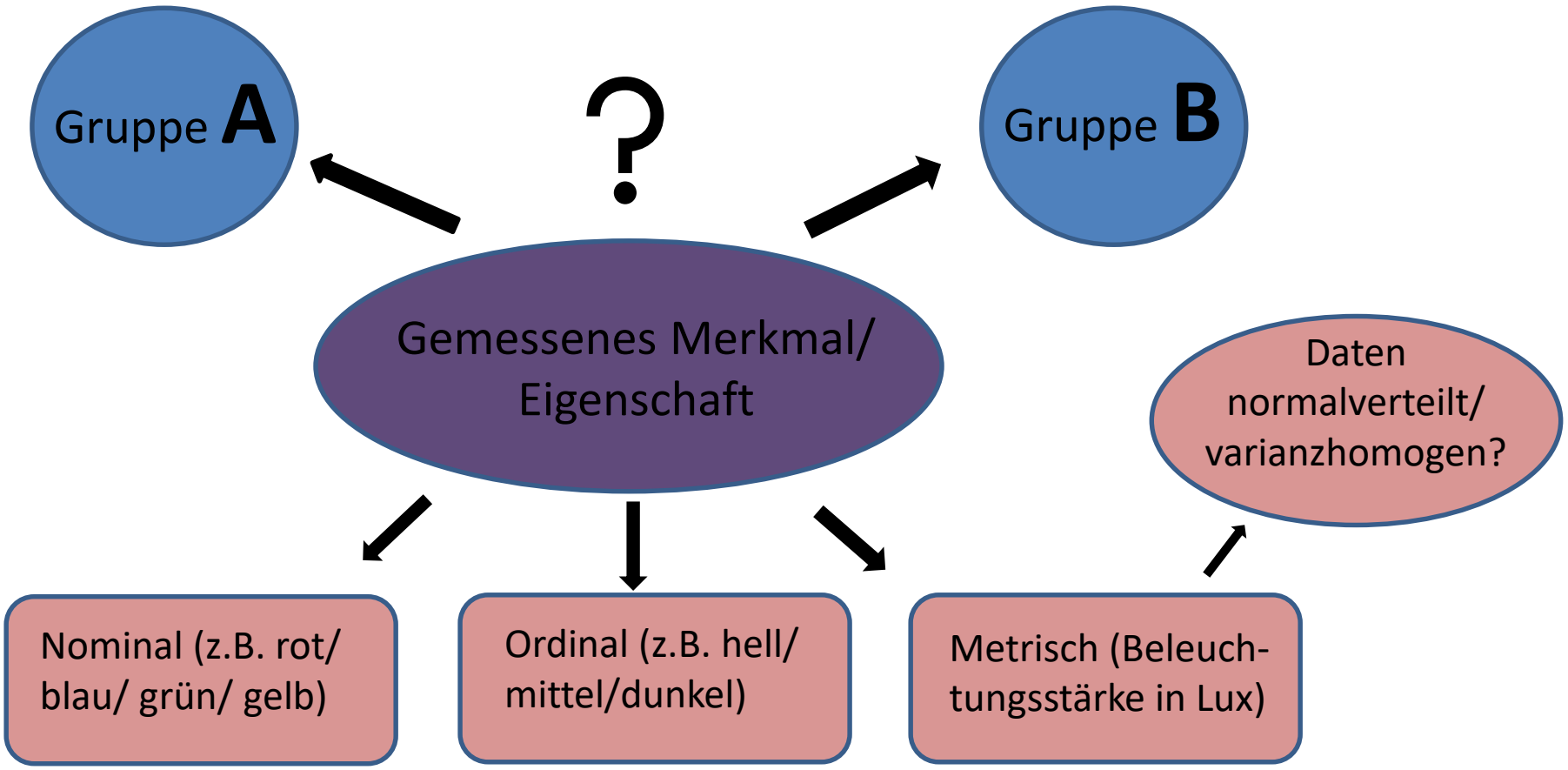
In R kann man auf viele Testdatensätze zurückgreifen:

`try(data(package = „datasets“))` zeigt die im Paket {datasets} verfügbaren Datensätze

`data(name)` öffnet den Datensatz, weitere Informationen über `help(name)`, Ansehen über `View(name)`.

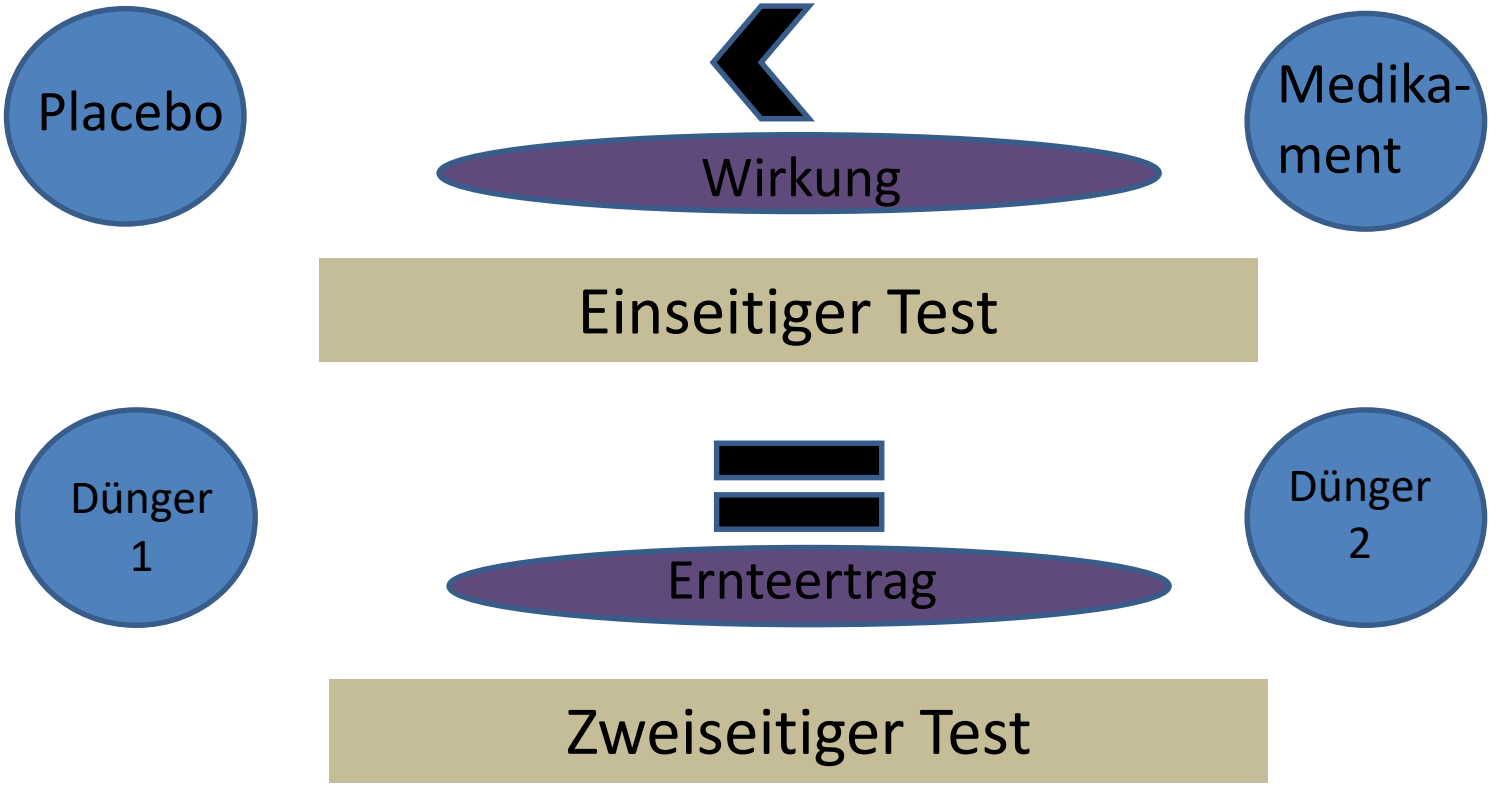
Auch in weiteren Paketen gibt es eine Reihe an Testdatensätzen, z. B. im Paket {car}. Zugriff erfolgt wie oben beschrieben.

Statistische Tests



Danach erfolgt die Auswahl eines geeigneten Tests, z.B. Chi²-Test `chisq.test()`, Wilcoxon-Test `wilcox.test()`, t-Test `t.test()`, …

Richtung des Tests

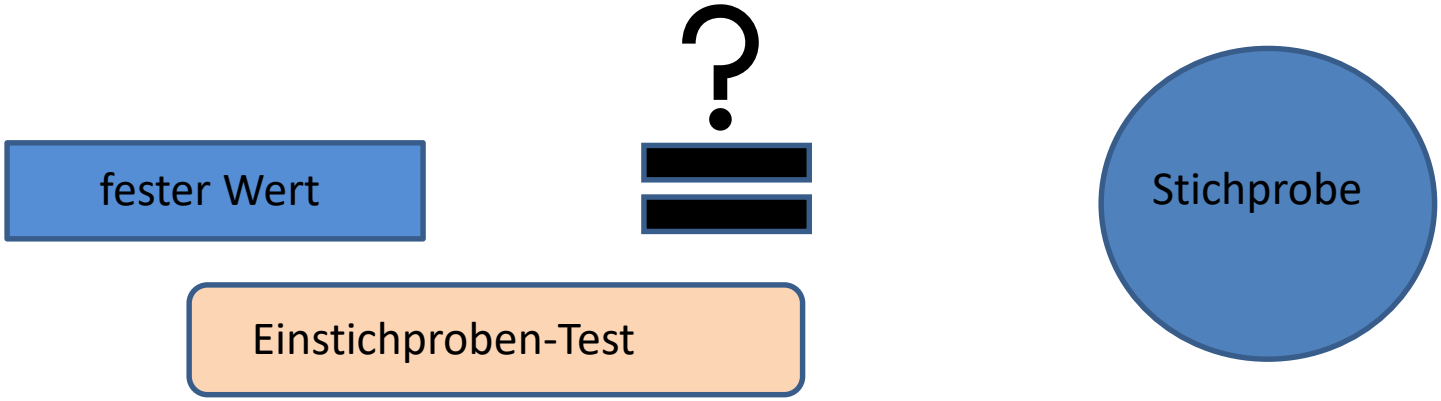


Wird definiert über das Argument

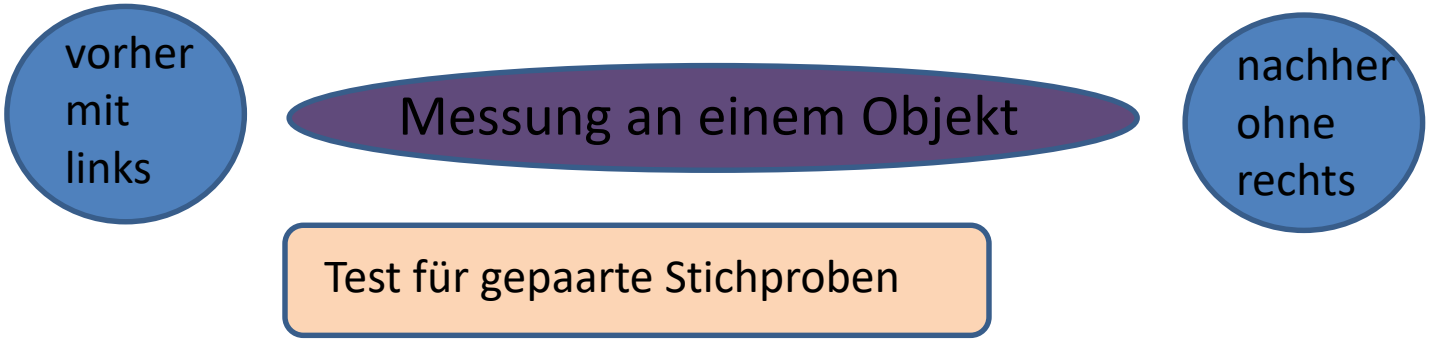
`alternative=` `„less“` `„greater“` `„two-sided“`

einseitig zweiseitig

Testmöglichkeiten 1



Wird über das Argument `mu=` *`Erwartungswert`* gesteuert

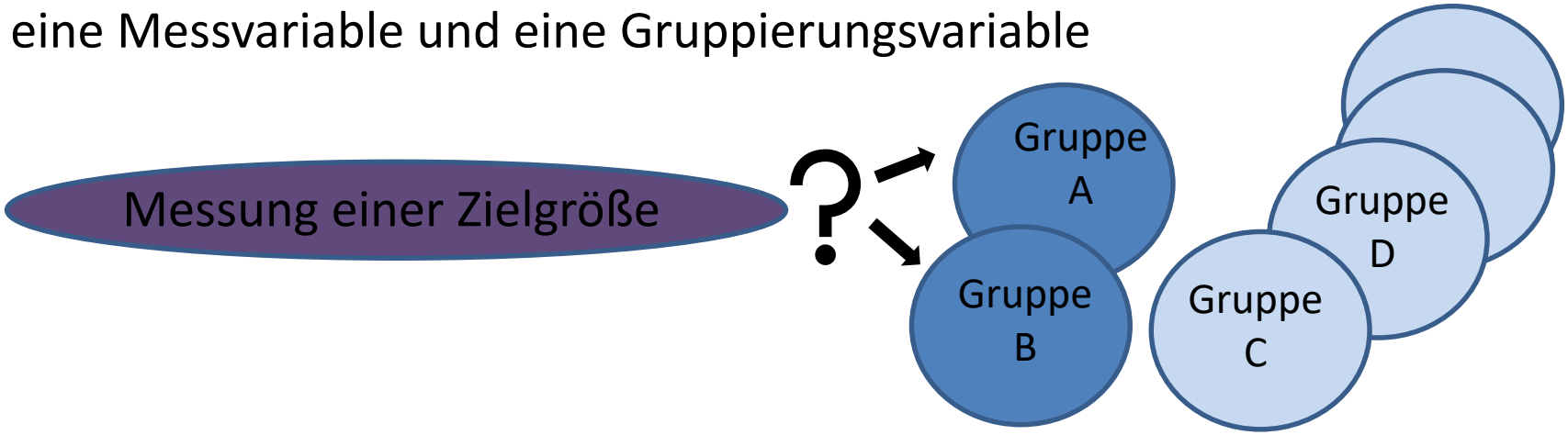


Wird über das Argument `paired=TRUE` gesteuert, hier habe ich zwei Variablen

Testmöglichkeiten 2

Test für unabhängige Stichproben

Man verwendet weder das Argument `mu= Erwartungswert` noch `paired=T` (oder man gibt `paired=F` an); hier hat man idR eine Messvariable und eine Gruppierungsvariable



ANOVA/ Kruskal-Wallis Test

Wenn man mehr als zwei Gruppen hat

`aov()`, `kruskal.test()`

Zusammenfassung in Folie 18

6

Metrische Variablen: Untersuchung auf Normalverteilung

Teil 3

Variablen auf Normalverteilung untersuchen:

- **visuelle Inspektion**, z.B. durch Q-Q-plots, Boxplots

`qqnorm()` mit `qqline()`, `boxplot()`

- **Shapiro-Wilk Test** `shapiro.test()`

wobei ein p-Wert < 0.05 Hinweis auf eine Abweichung von der Normalverteilung ist

- Unterschiede zwischen dem **Mittelwert** und dem **Median** der Variable weisen auf Abweichungen von der Normalverteilung hin ebenso wie Werte ≠ 0 (<-1 oder >1) der **Kurtosis** (Steilheit) und der **Schiefe**.

`skewness()`, `kurtosis()`, beide Paket {e1071}

- für Stichproben < 10 sollte generell keine Normalverteilung angenommen werden; für sehr große SP (>50) → zentraler Grenzwertsatz

Metrische Variablen: Untersuchung auf Varianzhomogenität

Variablen auf Varianzhomogenität untersuchen:

Sind die Verteilungen der Werte in den Gruppen gleich?

- **Visuelle Inspektion nach Gruppen** `boxplot(wertevar. ~ gruppvar.)` → sehen die Boxen ähnlich aus?
- **Levene Test** `leveneTest(wertevar. ~ gruppvar.)`, default ist `center=median`, das ist der Browne-Forsythe Test*, für originalen Levene-Test `center=mean` als Argument aufnehmen. Bei gepaarten SP `levene.var.test()`, Paket {PairedSamples}
- Bartlett Test für >2 Variablen `bartlett.test()`

Bei diesen Tests (Vergleich von Verteilungen) wird i.d.R. 0.2 als Signifikanzgrenze genommen (d.h. p-Werte kleiner als 0.2 weisen auf Varianzheterogenität hin).

* bei nicht normalverteilten Daten geeignet

Metrische Variablen: Anwendung von Funktionen

Genauere Untersuchung der Variablen

Viele Funktionen (mean, sd, shapiro.test,…) lassen keine Gruppierung einer Variable zu, also `mean(wertevar. ~ gruppvar.)` gibt eine Fehlermeldung. Um dies zu umgehen, gibt es einige Funktionen:

`tapply(wertevar., gruppvar., mean)` oder

`aggregate(wertevar.~ gruppvar., FUN=mean)`

Im Sinne von `tapply` gibt es auch die Funktionen `lapply()` (Ergebnis ist eine Liste) und `sapply()` (hilft bei nicht vektorisierten Funktionen)

Übung 3.1

Öffnen Sie den Testdatensatz „CO2“.

1. Untersuchen Sie die Variablen „conc“ und „uptake“ auf Normalverteilung (grafisch, Shapiro-Wilk-Test, Schiefe, Kurtosis,…)
2. Die Variable „Treatment“ dient nun als Gruppierungsvariable. Wie sieht es bei der Variable “uptake“ mit der Normalverteilung in den beiden Gruppen aus?
   Berechnen Sie den jeweiligen Mittelwert von „uptake“ in den beiden „Treatment“-Gruppen.
   Sind die beiden Gruppen varianzhomogen?

Unabhängige Stichproben

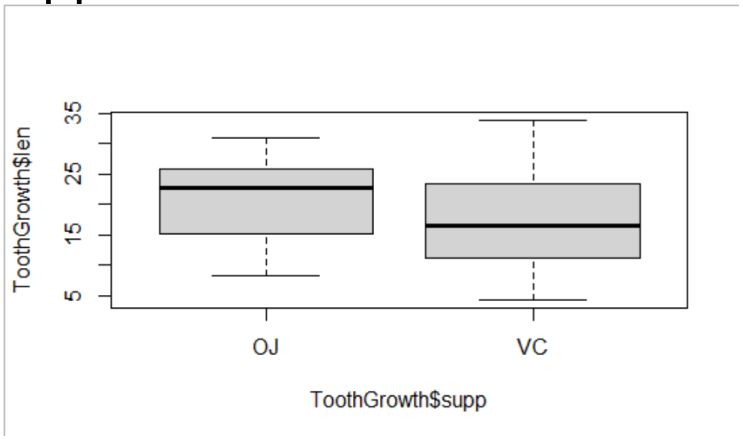
Der Testdatensatz **ToothGrowth** enthält Daten zum Zahnwachstum bei Meerschweinchen bei drei unterschiedlichen Vitamin C Dosen und zwei Darreichungsformen.

Zugriff: `data(ToothGrowth)`, Infos: `help(ToothGrowth)`

Zuerst untersuchen wir, ob die Darreichungsform (Orangensaft oder Ascorbinsäure) des Vitamin C einen Einfluss auf das Zahnwachstum hat.

Sind die „len“-Werte der beiden „supp“-Gruppen **normalverteilt** und **varianzhomogen**?

```
boxplot(ToothGrowth$len ~ ToothGrowth$supp)
```



Visuelle Bewertung: „OJ“ zeigt schon eine recht schiefe Verteilung, „VJ“ sieht relativ normalverteilt aus.

Die Varianzen sehen relativ homogen aus (Boxen sehen ähnlich aus).

Unabhängige Stichproben

Bestätigung über Tests:

Um den Shapiro-Wilk-Test nach der Darreichungsform zu gruppieren, muss man den Befehl `tapply(ToothGrowth$len,ToothGrowth$supp, shapiro.test)` verwenden.

Das Ergebnis des Shapiro-Wilk-Tests bestätigt unsere Beobachtung: p-Wert „OJ“ 0.02359 (sign. Abweichung von NV da <0.05), „VJ“ 0.4284 (NV annehmbar)

Beim Levene-Test `leveneTest(len ~ supp, data=ToothGrowth)` bestätigt sich mit einem p-Wert von 0.275 auch unsere Beobachtung, dass die Varianzen als homogen angesehen werden können (Wert >0.2).

Unabhängige Stichproben

Wir würden also auf einen **nicht-parametrischen Test** zurück greifen (da „OJ“ nicht normalverteilt ist), in diesem Fall den Wilcoxon-Test für unabhängige Stichproben. Wir testen zweiseitig, da wir keine Richtung der Ergebnisse vorgeben/ wissen.

```
wilcox.test(ToothGrowth$len ~ ToothGrowth$supp,
alternative="two.sided")
```

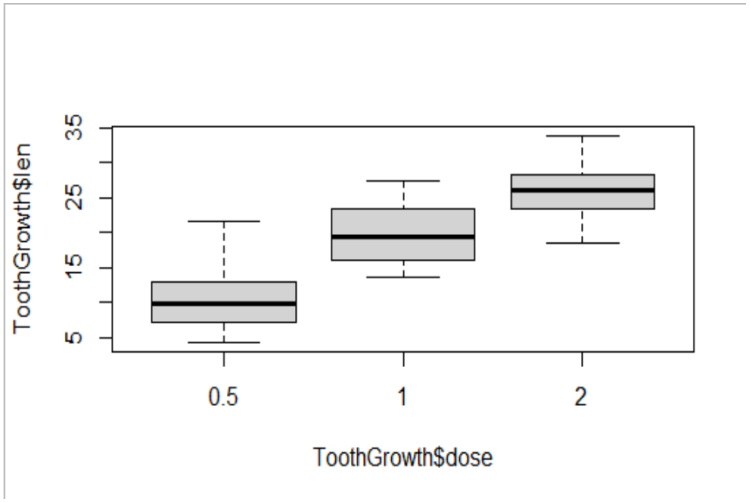
Der p-Wert von 0.0645 zeigt, dass keine signifikanten Unterschiede in der Zahnlänge bei den unterschiedlichen Vitamin C-Darreichungs-formen festgestellt werden können (da >0.05, wenn auch nur knapp).

Unabhängige Stichproben

Als nächstes würden wir gerne bei diesem Datensatz den Einfluss der **Vitamin C-Dosis (dose) auf die Zahnlänge** untersuchen. Auch hier untersuchen wir, ob die einzelnen Werte normalverteilt sind:

```
boxplot(ToothGrowth$len ~ ToothGrowth$dose)
```

Die Boxen sehen halbwegs normalverteilt und varianzhomogen aus, der Shapiro-Wilk Test (`tapply(ToothGrowth$len, ToothGrowth$dose, shapiro.test))` zeigt auch keine Einwände gegen die Annahme der Normalverteilung (p-Werte 0,25, 0,16 und 0,90).



Beim Levene-Test (Überprüfung Varianzhomogenität)

```
leveneTest(ToothGrowth$len ~ ToothGrowth$dose)
```

erhalten wir allerdings eine Fehlermeldung!

Unabhängige Stichproben

`str(ToothGrowth)` zeigt, dass die Variable „dose“ vom Typ „numeric“ ist, „supp“ hingegen „Factor“. Der Levene-Test verlangt als Gruppierungsvariable einen Factor, also müssen wir die Variable dose in eine Factor- Variable abändern. Dieses kann als Unterbefehl innerhalb der leveneTest-Funktion geschehen:

```
leveneTest(ToothGrowth$len ~
factor(ToothGrowth$dose))
```

Hier erhalten wir einen p-Wert von 0,53, Varianzhomogenität kann also angenommen werden.

Also: Werte normalverteilt und varianzhomogen → parametrischer Test

Unabhängige Stichproben

Da es mehr als zwei Gruppen sind, wählt man die ANOVA (nicht-parametrisch: Kruskal-Wallis):

```
summary(aov(ToothGrowth$len ~ factor(ToothGrowth$dose)))
```

Dieser Test ergibt einen p-Wert von 9.53e-16, es liegen also signifikante Unterschiede zwischen den Gruppen vor.

Wo genau diese Unterschiede liegen, kann man mit einem Post-Hoc Test nach Tukey untersuchen (auch hier muss wieder beachtet werden, dass die Variable dose als Factor in die Funktion genommen wird):

```
TukeyHSD(aov(ToothGrowth$len ~ factor(ToothGrowth$dose)))
```

Für alle drei Einzelvergleiche liegt der adjustierte p-Wert deutlich unter 0.05, es liegen also signifikante Unterschiede zwischen allen drei Gruppen vor.

Zusammenfassung

- Falls erforderlich, auf **Normalverteilung** untersuchen:
  visuell: `boxplot(wertevar. ~ gruppvar.)`
  Test: `shapiro.test(wertevar.)` bzw. `tapply(wertevar., gruppvar., shapiro.test)`
- F.e., auf Varianzhomogenität untersuchen:
  visuell wie oben über boxplot
  Test: `leveneTest(wertevar.~gruppvar.)`
- **Bei NV:** Unabhängige SP: `t.test(wertevar. ~gruppvar., …)`,
  EinSP: `t.test(wertevar., mu=`*`Erwartungswert`*`, …)`
  gepaarte SP: `t.test(wertevar1, wertevar2, paired=T,…)`
  Argumente innerhalb der Klammern steuern
  Fragestellung: `alternative=„t“` (2-seitig),`„l“` oder `„g“` (1-seitig)
  Varianzhomogenität ja/nein: `var.equal=T / var.equal=F`
- **Bei nicht NV**: alles (fast) wie oben, nur `wilcox.test()` verwenden
- *Bei > 2 Gruppen*, NV: `aov()`, nicht NV: `kruskal.test()`

Übung 4

Öffnen Sie den Datensatz „nitrate“ in R

1. Erstellen Sie eine Variable, die den Viehbestand (livestock) in die Gruppen 1 (niedrig, <=59000) und 2 (hoch, >59000) einteilt.
2. Untersuchen Sie die Hypothese, dass Regionen mit hohen Viehbeständen (Gruppe 2) signifikant höhere Nitratgehalte (nitrate) haben als solche mit geringeren Viehbeständen.
3. Versuchen Sie, grafisch den Nitratgehalt (nitrate) der verschiedenen Bodenarten (soil) darzustellen.
4. Bestehen Unterschiede hinsichtlich des Nitratgehalts zwischen den vier Bodenarten? (Verzichten Sie auf die paarweisen Vergleiche)
5. *Für Schnelle*: Lösen Sie Aufgabe 1 mithilfe der Funktion cut.

Zusatz

Um bei einem signifikanten Ergebnis des Kruskal-Wallis-Tests noch paarweise Vergleiche anschließen zu können, kann man z.B. die Funktion `pairw.kw()`, Paket {asbio}, verwenden:

```
pairw.kw(nitrate$nitrate, factor(nitrate$soil))
```

95% Confidence intervals for Kruskal-Wallis comparisons

	Diff	Lower	Upper	Decision	Adj. P-value
Avg.rankl-Avg.ranks	-15.38889	-29.34422	-1.43356	Reject H0	0.021736
Avg.rankl-Avg.rankt	-1.53175	-13.23507	10.17158	FTR H0	1
Avg.ranks-Avg.rankt	13.85714	-0.69869	28.41297	FTR H0	0.072107
Avg.rankl-Avg.ranku	-8.93889	-19.60915	1.73138	FTR H0	0.162561
Avg.ranks-Avg.ranku	6.45	-7.28896	20.18896	FTR H0	1
Avg.rankt-Avg.ranku	-7.40714	-18.8516	4.03731	FTR H0	0.526322

Hier zeigen sich also nur signifikante Unterschiede zwischen „l“ und „s“.