Fragenkatalog

Statistik

Univariate Verteilungen

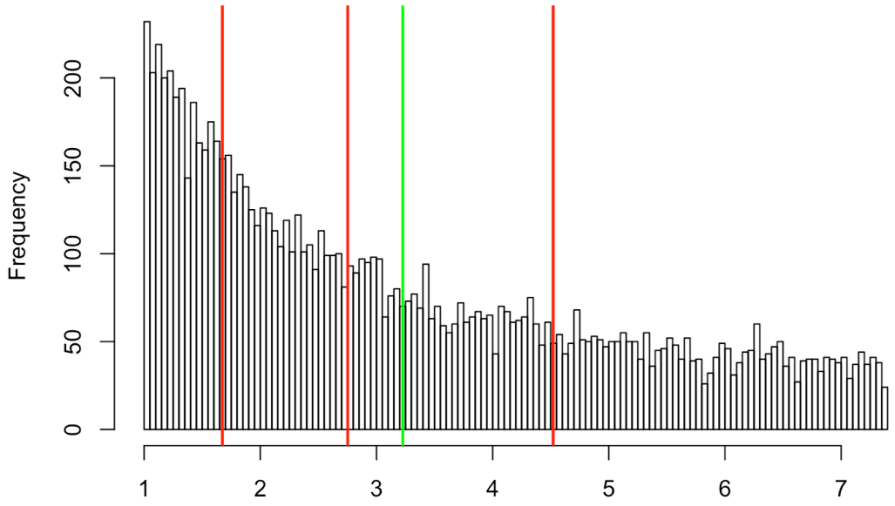
- Die Varianz einer beobachteten Variable kann nicht 0 werden.
  ➔ Falsch, die Varianz ist immer >= 0. Wenn alle Datenpunkte gleich sind ist sie 0, sie kann also 0 werden.

- Bei einer asymmetrischen Verteilung ist der Mittelwert immer größer als der Median.
  ➔ Falsch, größer oder kleiner, je nachdem ob die Verteilung rechtsschief oder linksschief ist.

- Wenn der Abstand der Quartile von links nach rechts abnimmt, ist die Verteilung a) linksschief b) rechtsschief c) gar nicht schief?
  ➔ A

- Ist es möglich, eine Varianz von 4 und eine Standardabweichung von 3 zu haben?
  ➔ Nein

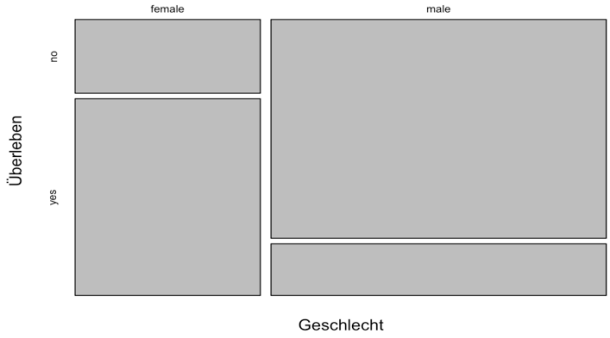
  Zeichnen Sie mit Hand die geschätzten Werte für a) 25%, 50%, 75% Quantil, und Mittelwert der folgenden Verteilung ein (es geht nur um die grobe relative Struktur):



Deskriptive Statistik mit 2 Variablen

- Welche Kombinationen von 2 Variablen unterschiedlichen Skalenniveaus (ungeordnet, metrisch) wurden in der Vorlesung durchgenommen?
  ➔ u-u, u-m, m-u, m-m

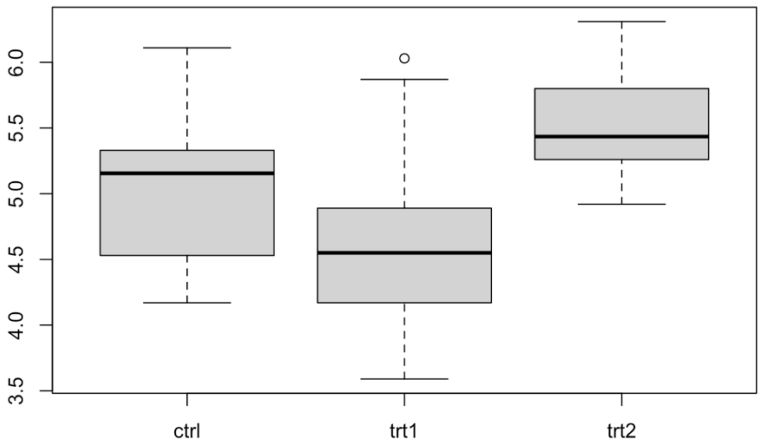
- Schauen Sie sich den folgenden Mosaikplot der Überlebenden der Titanic an:



- Haben relativ mehr Frauen überlebt? Woran sieht man das?
  ➔ ja, man sieht das am höheren vertikalen Übergang von ja / nein bei den Frauen.

- Haben absolut mehr Frauen überlebt? Woran sieht man das?
  ➔ sieht auch so aus - man muss den Flächeninhalt der beiden Felder vergleichen.

- Schauen Sie sich den folgenden Boxplot an



- Sind die Mediane der 3 Behandlungen gleich gleich?
  ➔ Nein

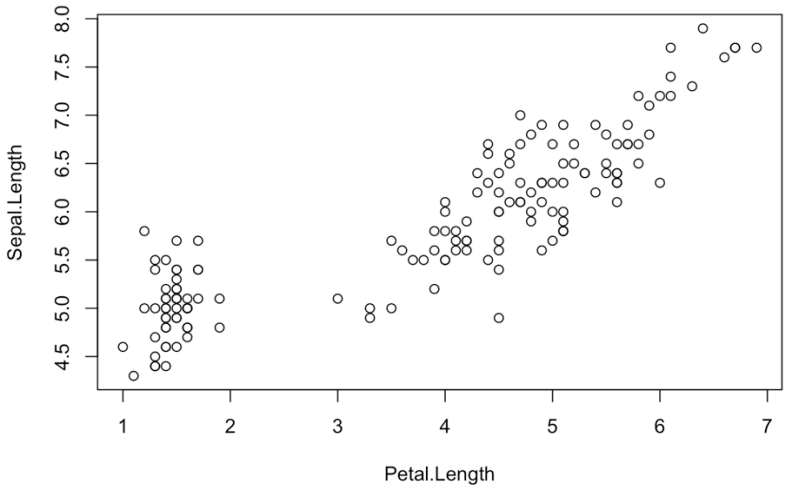
- Was können Sie über die Symmetrie / Schiefe der Verteilung der Kontrolle (ctrl) sagen?
  ➔ asymmetrisch, tendenziell linksschief (Median nach rechts verschoben).

- Gibt es Ausreißer / outlier?
  ➔ ja, Punkt in trt1.

- Welche Behandlung hat die kleinste Varianz?
  ➔ Trt2

- Ist der Unterschied zwischen den Behandlungen signifikant, d.h. gibt es einen Effekt der Behandlungen gegenüber der Kontrolle?
  ➔ das kann man aus so einem Plot nicht sagen!

Schauen Sie auf die folgende Korrelation zwischen Kelchblatt und Kronblatt von verschiedenen Blüten:



- Sind die Datenpunkte korreliert (deskriptiv), und wenn ja, positiv oder negativ?
  ➔ ja, positiv.

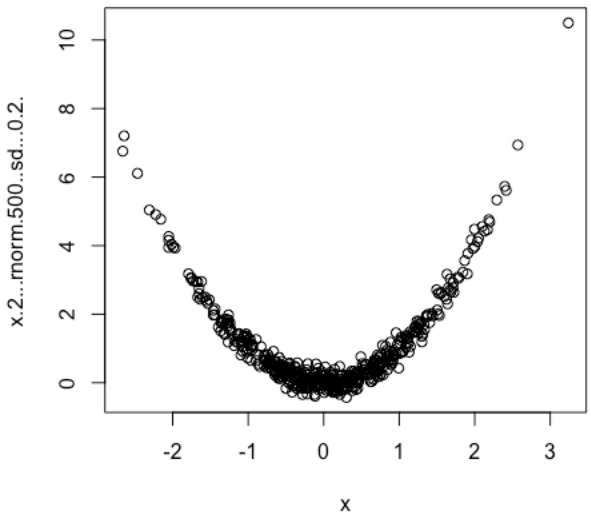
- Denken Sie, dass es einen deutlichen Unterschied zwischen den Korrelationskoeffizienten von Pearson und Spearman gibt?
  ➔ es gibt zwar diesen Cluster an Punkten links unten, aber da die Daten trotzdem mehr oder weniger linear sind, eher nicht (in der Tat, die Werte sind 0.87 und 0.88).

- Was ist der Unterschied zwischen deskriptiver und schließender Statistik?
  ➔ Deskriptive Statistik ist eine reine Zusammenfassung der Daten. Schließende Statistik benutzt Wahrscheinlichkeitsaussagen (wie wahrscheinlich wäre es diese Daten zu sehen, wenn kein Effekt da ist) um Schlüsse aus den Daten zu ziehen.

- Rechnen Sie Mittelwert, Standardabweichung und Schiefe für die folgenden Daten aus: - 1 , 0, 1
  ➔ Mittelwert: 0, sd = sqrt(1/3 * (-1^2 + 0^2 + 1^2)) = sqrt(2/3), Schiefe = 0

- Wenn der Mittewert wesentlich größer ist als der Median, was lässt sich dann über die wahrscheinliche Schiefe der Verteilung aussagen?
  ➔ Schiefe > 0 = rechtsschief. Tipp: lesen Sie noch mal die Artikel über Schiefe, Mittelwert usw. auf Wikipedia, siehe, z.B. https://de.wikipedia.org/wiki/Schiefe_(Statistik)

- Ist die Person Korrelation der folgenden Daten negativ, positiv, oder Null?



  ➔ Null

- Würde sich bei Anwendung eines Rangkorrelationskoeffizienten (z.B. Spearmann, siehe Vorlesung) eine andere Antwort ergeben?
  ➔ Nein, der wäre auch Null.

- Was ist der Unterschied zwischen Korrelation und Assoziation?
  ➔ Eine Beziehung zwischen 2 Variablen heißt Korrelation, wenn beide Variable metrisch sind, ansonsten Assoziation.

- Welche anderen Skalenniveaus gibt es denn noch, außer metrisch?
  ➔ nominal (ungeordnet), ordinal (geordnet)

- Geben Sie ein Beispiel für eine nominale Variable!
  ➔ Eine Variable mit den Werten: winzig, mittel, groß.

Maximum Likelihood

- Wie ist die Likelihood der beobachteten Daten D für ein gegebenes Modell M mit Parameter x definiert?
  ➔ p(D|M,x) - in Worten: die Wahrscheinlichkeits(dichte) für D, gegeben M und x
  Bemerkung: ich schreibe immer Dichte in Klammern, weil es sich mathematisch bei den Wahrscheinlichkeitsfunktionen von kontinuierlichen Daten um Dichtefunktionen handelt, d.h. die Normalverteilung ist eine Wahrscheinlichkeitsdichte.

- Beschreiben Sie die Idee des Maximum Likelihod Schätzers (MLE)!
  ➔ Man nimmt p(D|M,x), probiert alle möglichen Parameter aus, und nimmt den Parameter für den die Wahrscheinlichkeit der Daten am höchsten ist.

- Also ist der MLE der Parameter der an wahrscheinlichsten ist?
  ➔ Nein, er ist erst mal nur der Parameter für den die Wahrscheinlichkeit der Daten am höchsten ist. Nur wenn Sie die Zusatzannahme machen, dass auch alle Parameterwerte gleich wahrscheinlich sind dürften Sie sagen dass der MLE der wahrscheinlichste Parameter ist.

Hypothesentests

- Definieren Sie den p-Wert!
  ➔ p(d >= D | H0), in Worten: Wahrscheinlichkeit, die beobachteten oder extremere Daten zu bekommen, wenn H0 wahr ist.

- Wie ist in der vorherigen Definition “extremer” (also >=) definiert?
  ➔ Extremer bezieht sich auf die Teststatistik - diese ist ein Abstandsmaß, das der “Erfinder” des Testes wählt. Weil es hierfür unterschiedliche Möglichkeiten gibt, gibt durchaus verschiedene Tests mit gleichen H0, aber anderen Teststatistiken.

- Stellen Sie sich vor, sie leben in Venedig, und es gibt 3 Taxibootfirmen. Sie wollen wisse,n ob es Unterschiede in der Beförderungsgeschwindigkeit gibt und machen deshalb 300 Fahrten mit jeder Firma und stoppen die Zeit. Was wäre eine geeignete Nullhypothese H0 um auf einen Unterschied zu testen?
  ➔ H0: Die durchschnittliche Zeit aller 3 Firmen ist gleich.

- Bonusfrage: Nennen Sie eine der vielen möglichen sinnvollen Teststatistiken!
  ➔ Mögliche Antworten: Varianz der 3 Mittelwerte, Unterschied größter / kleinster Mittelwert, Durchschnitt der Differenzen der Mittelwerte, … [alles womit man die Firmen vergleichen könnte .. natürlich könnten Sie auch Mittewert durch Median ersetzen].

- In der Physik gibt es das sogenannte Standardmodell, das die Eigenschaften und Interaktionen der Material beschreibt. In den letzten 20 Jahren wurde das Standardmodell immer und immer wieder getestet, so “erfolgreich” dass die Physiker schon ein bisschen deprimiert sind, weil sie nichts neues entdecken. Was ist die

Nullhypothese, die bei diesen Tests angewandt wird?

➔ H0 = das Standardmodell.

Anmerkung: ich wollte Ihnen mit diesem Beispiel zeigen dass es 2 Möglichkeiten gibt, Nullhypothesen aufzustellen.

1 Wenn wir einen Effekt bestätigen wollen, stellen wir die umgekehrte Nullhypothese auf, also, dass es keinen Effekt gibt (das ist der Normalfall in der Analyse von biologischen Daten). Der Hintergrund hier ist, dass wir einen Effekt vermuten, aber nicht sicher sind.

2 Wenn der Effekt / das Modell aber schon gut bekannt ist, d.h. es eine feste Theorie gibt, wäre das normale Vorgehen diesen Effekt / Theorie als H0 aufzustellen, und zu schauen, ob es eine Abweichung von dem gibt was wir gerade als wahr ansehen.

- Sie lesen einen Artikel über ein medizinisches Experiment. Getestet wurde ein Medikament gegen eine Kontrolle, und der p-Wert ist 0.03. Die Studie schreibt: "Die Wahrscheinlichkeit, dass das Medikament nicht wirkt ist 3%" - stimmen Sie zu?
  ➔ Nein, der p-Wert gibt die Wahrscheinlichkeit der Daten gegeben H0 an, nicht die Wahrscheinlichkeit von H0.

- Schreiben Sie eine korrekte Interpretation des obigen Ergebnises auf!
  ➔ "Die Wahrscheinlichkeit, dass die beobachteten Effekte oder stärker unter H0 (d.h. ohne Wirkung des Medikaments) auftreten ist 3%. Der Unterschied zwischen Kontrolle und Behandlung ist deshalb signifikant bei einem Signifikanzlevel von 5%."

- Definieren sie den Typ I Fehler (falschen positive)!
  ➔ Wahrscheinlichkeit, dass der p-Wert signifikant wird wenn H0 wahr ist.

- Wie viel Typ I Fehler erwarten Sie bei einem Signifikanzlevel von 7%?
  ➔ 7%

- Definieren Sie den Typ II Fehler (falsche negative)!
  ➔ Wahrscheinlichkeit, dass der p-Wert nicht signifikant wird, wenn H0 nicht wahr ist.

- Wie viel Typ II Fehler erwarten Sie bei einem Signifikanzlevel von 7%?
  ➔ Nur mit dieser Information kann man das nicht sagen, weil der Typ II von mehreren Faktoren abhängt.

- Nennen Sie 2 Faktoren, die Typ II Fehler beeinflussen, und die Richtung des Einflusses (negativ = Typ II nimmt ab wenn Faktor hoch geht).
  ➔ Mögliche Antworten: Signifikanzlevel - negativ; Varianz - positiv; Effektstärke: negativ; power / Stichprobengröße: negativ

- Definieren Sie die Teststärke / Power!
  ➔ Power = 1 - Typ I Fehler / in Worten: also Wahrscheinlichkeit, dass der p-Wert signifikant wird, wenn H0 nicht wahr ist.

- Sie testen 100 Gene auf eine Assoziation mit Krebs. Bei 5 Genen zeigt der Test Signifikanz an. Wie bewerten Sie dieses Ergebnis?
  ➔ 5 von 100 signifikante Tests sind zu erwarten, wenn keines der Gene eine Wirkung hat (wir machen hier multiples Testen). Das Ergebnis zeigt keine besondere Evidenz für einen Effekt an.

Bemerkung: Trotzdem würde man natürlich die 5 signifikanten mit einem weiteren Experiment noch mal nachtesten, man weiß ja nie. Aber statistisch ist so ein Ergebnis zu erwarten, auch wenn keines der Gene eine Wirkung hat.

- Sie bekommen von einer allwissenden Macht die Zusatzinformation, dass Sie in dem oben genannten Test eine Teststärke von 99% hatte. Sind Sie nun zuversichtlicher, dass Sie einen Effekt gefunden haben?
  ➔ Nein, denn wenn Ihre Teststärke fast 100% ist und Sie wirklich Assoziationen in den Datensatz hätten, dann müssten Sie ja trotzdem MEHR als 5% positive sehen.

- Definieren Sie die False Discovery Rate (FDR)!
  ➔ Die Rate an Experimenten, die Signifikanz anzeigt, obwohl kein Effekt da ist, wenn man viele Experimente macht.

- Wovon hängt die FDR ab?
  ➔ FDR hängt ab von Typ I und II Fehlerrate und der Wahrscheinlichkeit, dass H0 wahr ist (bzw. dass die in den Experimente getesteten Effekt da sind).

- In einer Reihe von Experimenten testen Sie Medikamente auf eine Wirkung. Ihre Power ist 100%. Sie schätzen, dass jedes 20. Medikament das Sie testen eine Wirkung haben sollte. Wie ist ihre FDR bei einem Signifikanzlevel von 5%? Es reicht, wenn Sie den Rechenweg aufschreiben, Sie müssen den Wert nicht ausrechnen.
  ➔ Antwort 0.95 * 0.05 / (0.95 * 0.05 + 0.05 * 1) - Rate an Typ I / Rate an signifikante Ergebnissen.

- Warum hat der p-Wert bei einem Signifikanzlevel von 0.05 genau 0.05 falsche Positive?
  ➔ Der p-Wert ist die Wahrscheinlichkeit einen gegebenen Wert (definiert durch die Teststatistik) oder größer zu erhalten, wenn H0 wahr ist - diese Definition impliziert, dass die Wahrscheinlichkeit einen p-Wert unter 5% zu bekommen 5% ist wenn H0 wahr ist.

- Was ist der Unterschied zwischen Typ I Fehler und der “False Discovery Rate”?
  ➔ Typ I Fehlerrate = wie häufig treten falsche Positive auf wenn H0 wahr ist. FDR = wie häufig treten falsche Positive auf wenn H0 mit einer Wahrscheinlichkeit p(H0) wahr ist.

- Geben Sie die false discovery rate an für Signifikanzlevel 0.05, Power = 0.4, p(H0) = 0.5!
  ➔ 0.05 * 0.5 / (0.05 * 0.5 + 0.4 * 0.5) -> Rate Typ I / (Rate Typ I + Rate echte Positive).

- Der p-Wert eines Experiments 0.04 - ist die folgende Aussage richtig: die Wahrscheinlichkeit, dass H0 wahr ist ist kleiner als 4%? Begründen Sie ihre Antwort.
  ➔ Nein. Begründung: p-Wert gibt Wahrscheinlichkeit der Daten gegeben H0 - wie wahrscheinlich es dann ist, dass H0 oder nicht H0 zutrifft kann man daraus nicht sehen. Wenn überhaupt müsste man für eine solche Aussage die FDR berechnen.

MLE & Regression

- Wie ist die Likelihood der beobachteten Daten D für ein gegebenes Modell M mit Parameter x definiert?
  ➔ $p(D|M,x)$ - in Worten: die Wahrscheinlichkeits(dichte) für D, gegeben M und x.

- In einem Experiment wurde der Effekt von Stickstoff auf das Wachstum von Pflanzen getestet - der Likelihood ist maximal für eine Verdoppelung des Wachstums. Die Autoren schreiben: "Die Wahrscheinlichkeit die beobachteten Daten zu erhalten ist maximal, wenn man annimmt, dass Stickstoff das Wachstum von den beobachteten Pflanzen verdoppelt" - ist diese Aussage Korrekt?
  ➔ Ja.

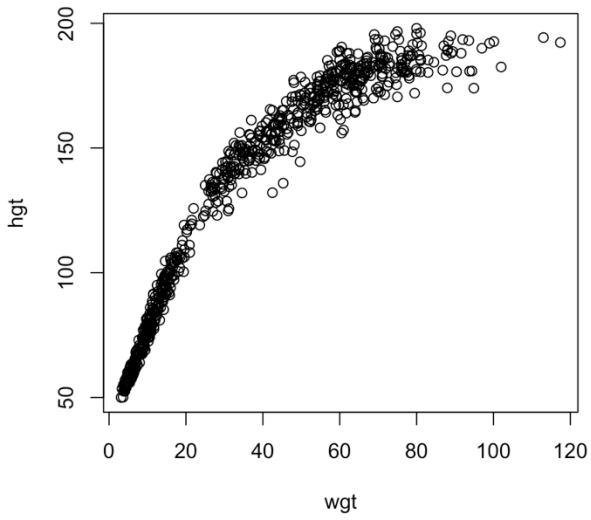
- Weiter unten schreiben die Autoren: "Der wahrscheinlichste Wert für den Effekt von Stickstoff ist 2 (Verdopplung)" - ist diese Aussage korrekt?
  ➔ Nein. Der MLE gibt den Wert für den die Wahrscheinlichkeit der Daten maximal ist. Um die Aussage umzudrehen, müssen wir Zusatzannahmen machen (alle Parameter gleich wahrscheinlich).

- Was sind die Annahmen der linearen Regression?
  ➔ Abh. Variable beschreibbar durch Polynom der unabhängigen Variable(n) + normalverteilte Streuung.

- Wie werden die Parameter in der linearen Regression bestimmt?
  ➔ Man sucht den MLE für die Annahmen der linearen Regression.

- Welche H0 steckt hinter den p-Werten der Parameter der linearen Regression?
  ➔ H0 = Parameter ist 0.

- Wir interessieren uns für den folgenden Zusammenhang zwischen Körpergröße (hgt) und Gewicht (wgt) aus einem Datensatz holländischer männlicher Kinder / Jugendlicher:

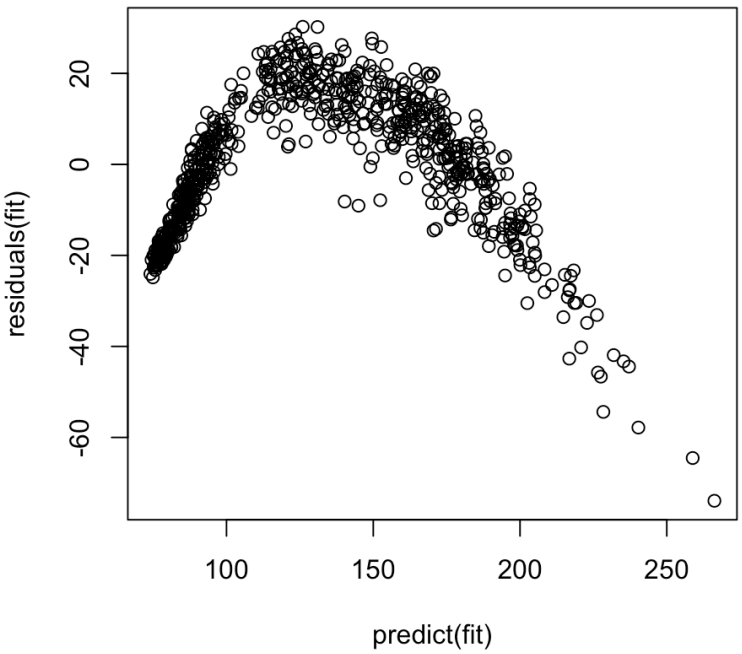


Interpretieren Sie die folgenden Regressiontabelle:

```
fit = lm(hgt ~ wgt, data = mice::boys, na.action =
"na.exclude")
summary(fit)
##
## Call:
## lm(formula = hgt ~ wgt, data = mice::boys, na.action =
"na.exclude")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.926 -11.639   1.329  12.532  30.231
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.79024    1.01156   68.00   <2e-16 ***
## wgt          1.68173    0.02207   76.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
##
## Residual standard error: 15.49 on 725 degrees of freedom
##   (21 observations deleted due to missingness)
## Multiple R-squared:  0.889,  Adjusted R-squared:  0.8889
## F-statistic:  5809 on 1 and 725 DF,  p-value: < 2.2e-16
```

➔ Gefittet wurde ein linearer Zusammenhang: hgt = a0 + a1 * wgt Der MLE für den Zusammenhang ist a0 = 68.79024, a1 = 1.68173. Körpergröße steigt also an mit dem Gewicht an. Für beide Werte kann die Nullhypothese (Parameter ist 0) mit hoher Signifikanz abgelehnt werden: p<0.0000000000000002

- Die Residuen des Models, geplottet gegen den vorhergesagten Wert (Körpergröße) sehen so aus:



Ist ein solcher Plot akzeptabel?

➔ Nein. Es gibt ein klares Pattern: für kleine und große Vorhersagen sind die Residuen systematisch negativ, d.h. das Modell überschätzt die Daten. Für mittlere Vorhersagen sind die Residuen systematisch positiv, d.h. das Modell unterschätzt die Daten.

- Können Sie sich denken was das Problem ist?
  ➔ Wie man an dem Plot sieht ist der Effekt nicht linear - man hätte die Variablen transformieren oder einen quadratische Term in das Modell aufnehmen müssen.

- Wie funktioniert die Methode des Maximum Likelihood Schätzers in Worten?
  ➔ Man stellt ein Modell für die Wahrscheinlichkeit der Daten gegeben die Parameter auf (Likelihood). Dann such man die Parameter für die diese Wahrscheinlichkeit der Daten maximal ist.

Die folgenden Fragen beziehen sich auf dieses Regressionsergebnis:

```
library(agridat)
fit2 <- lm(roots ~ leaves + I(leaves^2), data =
mercer.mangold.uniformity)
summary(fit2)
##
## Call:
## lm(formula = roots ~ leaves + I(leaves^2), data =
mercer.mangold.uniformity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.449  -8.508   0.672   9.817  36.662
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.54052   63.94141  -0.525 0.600485
## leaves       12.19976    2.59622   4.699  4.9e-06 ***
## I(leaves^2)  -0.09561    0.02623  -3.645 0.000342 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
##
## Residual standard error: 14.77 on 197 degrees of freedom
## Multiple R-squared:  0.4793, Adjusted R-squared:  0.474
## F-statistic: 90.68 on 2 and 197 DF,  p-value: < 2.2e-16
```

- Was für eine Kurve wird hier gefittet?
  ➔ Eine der folgenden Antworten wäre ausreichend: Polynom 2. Ordnung, quadratisches Polynom, y = a0 + a1x + a2 x2^2

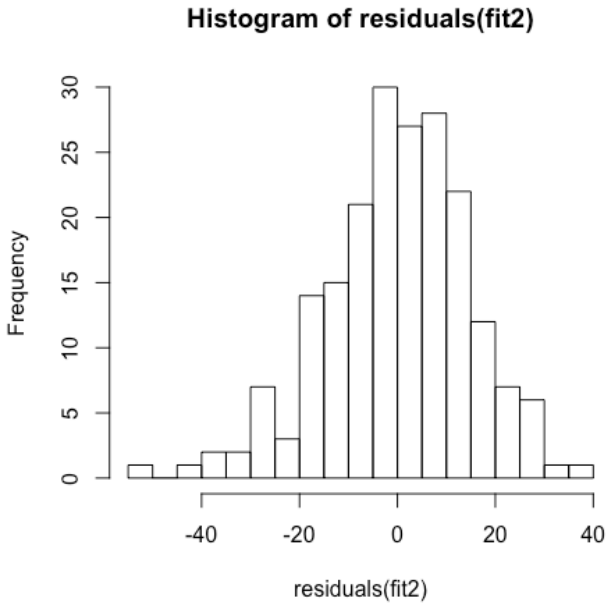
- Schreiben sie die Funktion auf, für die nach der oben stehenden Berechnung in R die Likelihood maximal wird (dem MLE)!
  ➔ roots = -33.54052 + 12.19976 * leaves + -0.09561 * leaves^2

- Der p-Wert für den Achsenabschnitt (Intercept) ist ja nicht signifikant. Aber was testet eigentlich dieser Wert, d.h. was ist die Nullhypothese?
  ➔ H0 ist: der Wert des Intercepts ist 0.

  ➔ Schauen Sie auf die Kennzahlen für die Verteilung der Residuen - ist die Verteilung stark asymmetrisch (Nein, nicht stark asymmetrisch, fast gleiche Abstände zwischen Q1, Median, Q3.

- Was sind denn eigentlich diese “Residuen”?
  ➔ Residuen = Abstände zwischen der Modellvorhersage (Polynom, siehe Frage oben) und den Daten, d.h. es gibt ein Residuum pro Datenpunkt.

- Wie sollten die Residuen bei der linearen Regression verteilt sein, und sieht die folgende Verteilung gut aus?



➔ Sie sollten normalverteilt sein. Die Verteilung sieht grob wie eine Normalverteilung aus (kleine Linksschiefe, aber ein paar Schwankungen sind natürlich immer da).

Multiple Regression, Störfaktoren, etc.

- Bzgl. der oben durchgeführten Analyse: welcher der drei Variablen ist der wichtigste Störfaktor: Alter, Sportlichkeit, Geburtsort?
  ➔ Alter, denn es ist zu erwarten, dass Alter einen großen Einfluss auf Gewicht (erklärende Variable) und Körpergröße (abhängige Variable) hat.

- Wie kann man den Effekt von Alter und Gewicht trennen?
  ➔ Multiple Regression.

- Wenn wir jetzt eine multiple Regression machen, erwarten Sie, dass der Effekt von Gewicht hoch oder runtergeht (in Bezug auf die Regression nur mit Gewicht)?
  ➔ Runter. Störfaktor Alter hängt positiv mit Gewicht (erklärende Variable) und Körpergröße (abhängige Variable) zusammen, also geht der Effekt von Gewicht hoch wenn der Störfaktor fehlt, und runter, wenn wir Alter reinnehmen. Schauen wir uns das mal an?

```
##
## Call:
## lm(formula = hgt ~ wgt + age, data = mice::boys, na.action
= "na.exclude")
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -30.367  -6.822   1.293   7.527  26.441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 69.48387    0.65230 106.521  < 2e-16 ***
## wgt          0.30576    0.04534   6.743 3.16e-11 ***
## age          5.48958    0.17177  31.960  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
##
## Residual standard error: 9.981 on 724 degrees of freedom
##   (21 observations deleted due to missingness)
## Multiple R-squared:  0.954,  Adjusted R-squared:  0.9538
## F-statistic:  7503 on 2 and 724 DF,  p-value: < 2.2e-16
```

- Wie stark hat sich der Effekt von Gewicht durch das Einbeziehen des Störfaktors geändert?
  ➔ Vorher 1.68, jetzt 0.3 -> d.h. der wahre Effekt ist nur ca. 20% des ursprünglichen Effekts, Körpergröße hängt vor allem am Alter und nicht am Gewicht.

- Sie erhalten einen großen Datensatz mit allen möglichen Eigenschaften von Personen und der Information ob eine Krebserkrankung vorliegt oder nicht. Die Frage ist, ob der Konsum von Fleisch zu einer höheren Wahrscheinlichkeit von Krebs führt. Nennen Sie 2 mögliche Störfaktoren, die man bei der Analyse dieses Datensatzes beachten sollte. Geben Sie eine Begründung für jeden Faktor!
  ➔ Alter - korreliert mit Krebs und potentiell mit Fleischkonsum (junge Leute mit größerer Wahrscheinlichkeit Vegetarier).
  ➔ Geschlecht - korreliert auf jeden Fall mit Fleischkonsum und potentiell mit Krebs (je nach Krebsart).

Bei einer solchen Frage ist wichtig, dass Sie glaubhaft darstellen, dass der Störfaktor mit beiden Variable (erkl. u. abh.) korreliert. Eine falsche Antwort wäre z.B. genetische Disposition - die hat zwar klar einen Einfluss auf den Krebs, aber es ist nicht klar warum diese mit Fleischkonsum korrelieren sollte!

Experimentelles Design

- Warum sollten Sie skeptisch sein, wenn Sie in einer Studie lesen, dass das Tragen von teurem Goldschmuck die Gesundheit fördert? (die Studie untersuchte 5000 zufällig ausgewählte Personen aus der deutschen Bevölkerung. In einer einfachen Regression zeigte sich, dass das Tragen von Goldschmuck ein signifikanter Prädiktor für die Lebenserwartung ist.
  ➔ Weil offensichtliche Störfaktoren nicht bedacht wurden (welche klären wir mit der nächsten Frage).

- Nennen Sie mindestens 2 offensichtliche Störfaktoren um einen Zusammenhang zwischen Goldschmuck und Lebenserwartung in einer Zufällig ausgewählten Stichprobe von Personen nachzuweisen, zusammen mit der Erklärung warum diese Faktoren Störfaktoren sind (Korrelation mit erkl. und abh. Variable)!
  ➔ Geschlecht (Frauen leben länge und Frauen tragen mehr Schmuck), Vermögen (vermögende Personen haben mehr Goldschmuck, und typischerweise auch bessere Gesundheit).

- Welcher Typ von Validität war in dieser Studie nicht gegeben?
  ➔ Interne Validität.

- Erklären sie am Beispiel der oben genannten Studie das Konzept der externen Validität.
  ➔ externen Validität = kann man die Studienergebnisse verallgemeinern. Externe Validität würde, z.B., die Frage bezeichnen ob man die Ergebnisse der gegebenen Stichprobe auf die deutsche Bevölkerung oder Menschen im Allgemeinen (ganze Welt) verallgemeinern kann.

- Wenn man wissen will wie viele Beobachtungen (Replikate) man für ein Expiriment braucht, macht man …
  ➔ eine Poweranalyse

Bioinformatik

Datenbanken

- Was wird in der NCBI-, der PDB- und der SCOP-Datenbank gesammelt?
  ➔ NCBI: National center for biotechnology information, Zugang zu DNA-, RNA, Protein-Datenbanken, Software und wissenschaftlicher Literatur.

  ➔ SCOP: Structural classification of proteins, Klassifikation von Proteinen aufgrund von Sequenz- und Strukturähnlichkeit.

  ➔ PDB: Protein data bank, Datenbank für 3D Strukturdaten von Proteinen und Nukleinsäuren.

- Wie ist die Pfam-Datenbank aufgebaut?
  ➔ Protein families, Kategorisieren von Proteindomänen, Mustererkennung mittels machine learning, Hinweis auf Struktur und Funktion, maschinelles Clustering und Mustererkennung durch HMMs.

Datenstrukturen

- Wie sind Graphen definiert?
  - ➔ Konzept um Anzahl und Relationen zwischen Objekten graphisch darzustellen. Tupel (V, E) mit e = V x V. Gerichtet oder ungerichtet.
  - ➔ V ist Menge an Knoten
  - ➔ E ist Menge an Kanten
  - ➔ Abfolge an Kanten heißt Pfad
  - ➔ Deg(u) beschreibt Anzahl Kanten, die u mit anderen Knoten verbindet

- Was zeichnet Bäume aus?
  - ➔ Graph G = (V, E), wenn es genau einen Pfad gibt, der Knoten verbindet.
  - ➔ Blätter: Knoten mit deg(1)
  - ➔ Wurzel: Ältester Vorfahre

Algorithmen

- Wozu dient die *O*-Notation?
  - ➔ Dient zur Abschätzung der Laufzeit eines Algorithmus bei unendlich großen Eingabemengen.

- Was besagt eine Laufzeit von $O(n^2)$? Wie ändert sich die Laufzeit, wenn *n* verdoppelt wird?
  - ➔ Polynom wächst mindestens exponentiell
  - ➔ Laufzeit vervierfacht sich bei Verdopplung von n

- Was wird unter Trainieren und maschinellem Lernen verstanden?
  - ➔ Generierung von Wissen aus Erfahrung, Ableiten von Wahrscheinlichkeiten aus bekannten Komplexen

Klassifikation

- Wie lautet die Bayessche Formel und wie wird sie bei Klassifikationsaufgaben benutzt?
  - ➔ $p(w_i/x) = p(x/w_i)\ p(w_i)\ /\ p(x)$
  - ➔ Falls $p(w_1/x) > p(w_2/x)$, entscheide zugunsten von w_1, ansonsten wähle w_2.

- Was ist ein Naiver Bayesscher Klassifikator?
  - ➔ Annahme: Das Vorkommen der Residuen ist unabhängig von deren Umgebung. Jedes Attribut hängt nur vom Klassenattribut ab.
  - ➔ Wahrscheinlichkeiten dürfen multipliziert werden!

Exakter Sequenzvergleich

- Welche Laufzeit hat der naive Ansatz zum Sequenzvergleich?
  ➔ Laufzeit von O(n m) bei n Länge Text und m Länge Suchmaske

- Was sind Suffix-Bäume? Was sind deren Vorteile?
  ➔ Jedes vorkommende Suffix wird durch einen Pfad von der Wurzel ausgehend zu einem Blatt repräsentiert. Blatt ist mit i markiert.
  ➔ Vorteile: Auswertung und Konstruktion in linearer Zeit, linearer Speicherverbrauch

Dotplots

- Berechnen Sie für zwei beliebige Sequenzen einen Dotplot!

- Wozu werden Suffixbäume benutzt und warum?
  ➔ Verwendung zur Genomsequenzierung und zur Assemblierung von vielen kleinen reads.

- Wozu und wie vergleicht man Genome mit Dotplots? Was kann aus den Plots abgeleitet werden?
  ➔ Vergleich von zwei Sequenzen/Genomen je Position, bei Homologität erscheint „dot"
  ➔ Laufzeit $O(n^2)$
  ➔ Ableiten von Insertionen, Deletionen, Inversionen, Pathogenität, repetitive Elemente, Intron/Exon Struktur, Terminatoren, Frameshifts, Regionen niedriger Komplexität.

- Wie interpretieren Sie Lücken, die in einem Dotplot auftreten?
  ➔ Insertionen, Deletionen, Introns bzw. keine Ähnlichkeit zwischen den Sequenzen.

Paarweiser Sequenzvergleich

- Auf welchem Paradigma beruht der Sequenzvergleich?
  ➔ Für alle Proteine gilt: In der Regel impliziert hohe Sequenzähnlichkeit auch ähnliche Funktion und/oder Struktur.

- Welche Distanzen zum Vergleich von Zeichenketten kennen Sie? Was sind ihre Vor- und Nachteile?
  ➔ Minkowski-Distanz, Errechnen des Abstandes mithilfe von Raumkoordinaten (Navigation), euklidischer und Manhattan-Abstand.
  ➔ Hamming-Distanz: Zählt Positionen, an denen Sequenzen unterschiedlich sind, darstellung im binären/ASCII Code möglich. Allerdings müssen die Sequenzen gleich lang sein.
  ➔ Levenshtein-Distanz: Minimale Anzahl von Editieroperationen, um Sequenz A in Sequenz B zu überführen. Erlaubt: Ersetzen,

Einfügen, Löschen. Finden des Minimums durch dynamisches Programmieren.

- Wie ist die Levenshtein-Distanz definiert? Wie wird sie berechnet?
  ➔ Minimale Anzahl von Editieroperationen, um Sequenz A in Sequenz B zu überführen. Berechnung:
  $D_{i,j} = \min(D_{i-1,j} + g(e), D_{i-1,j-1} + d(a_i, b_j), D_{i,j-1} + g(e))$

- Berechnen Sie für zwei beliebige Sequenzen ein Alignment!

- Was unterscheidet den NW- und den SW-Algorithmus?
  ➔ Needleman-Wunsch-Algorithmus: Berechnet optimalen globalen similarity score bzw. Alignments zwischen zwei Sequenzen.
  ➔ Smith-Waterman-Algorithmus: Berechnet optimalen lokalen similarity score bzw. Alignment zwischen 2 Sequenzen.

- Warum wird bem SW-Algorithmus die Null als Schwelle eingeführt?
  ➔ Score wird nicht negativ, jede Position hat die Chance, Beginn eines lokalen Alignments zu werden.
  ➔ Höchster Score steht irgendwo in der Matrix, nicht rechts unten. Backtracking vom höchsten Score bis hin zur 0.

- Wie und weshalb werden Lücken bewertet?
  ➔ Lücken durch mutagene Ereignisse wie Insertionen/Deletionen oder horizontalen Gentransfer. Eröffnen der Lücke wird teuer bewertet, Verlängern billiger. Wenige Lücken, dafür länger.

- Was unterscheidet Distanzen und Scores?
  ➔ Distanz sucht minimalen Abstand, Score sucht maximalen similarity score.

- Begründen Sie, weshalb der modulare Aufbau von Proteinen das Konzept der Vergleichsalgorithmen beeinflusst.
  ➔ Bereiche lokal hoher Ähnlichkeit sind durch größere Strecken ohne Übereinstimmung getrennt, was den Score stark herabsetzt, selbst wenn Proteine funktionell/strukturell sehr ähnlich sind. Außerdem können mutagene Ereignissewie Insertionen/Deletionen, sowie Translokationen und horizontaler Gentransfer zu Lücken im Alignment führen, obwohl sich die Proteine sehr ähnlich sind. Dies verfälscht das Ergebnis!

- Welcher Zusammenhang besteht bei Proteinen zwischen dem Anteil identischer Residuen und homologer Struktur?
  ➔ Ein hoher Anteil an identischen Residuen weist auf strukturelle und funktionelle Ähnlichkeit hin und somit ggf. auf die Zugehörigkeit zur selben Proteinfamilie und Homologie (gemeinsamer Ursprung).

Profile, Scoring-Matrizen

- Was ist eine Konsensus-Sequenz? Was sind Profile? Was ein Sequenzlogo? Wie werden sie jeweils berechnet?
  - ➔ Konsensus-Sequenz: Sequenz, welche in der Summe am wenigsten von einer gegebenen Menge an Mustersequenzen abweicht. Berechnung heuristisch aus MSA.
  - ➔ Profil: Schema, das für jede Position in der Sequenz die Wahrscheinlichkeit für das Auftreten einer bestimmten Aminosäure/Base/Insertion/etc. durch eine positionsspezifische Häufigkeitstabelle berechnet. Sequenzprofile sind in verwandten Proteinen meist konserviert und nehmen eine Schlüsselrolle ein.
  - ➔ Sequenzlogo: Graphische Darstellung des Sequenzprofils. Darstellung der Konserviertheit von Aminosäuren/Basen in einer Proteinsequenz.

- Stellen Sie einen Bezug zwischen Chancen-Quotienten und dem Neyman-Pearson-Lemma her.
  - ➔ Voraussetzung: Beobachtungen H0 und H1 folgen bekannten, einfachen Wahrscheinlichkeitsdichten. Man erhält den besten Test durch eine Entscheidung, bei welcher die Nullhypothese verworfen wird, wenn der Likelihoodquotient f_0/f_1 einen bestimmten Wert unterschreitet, bzw. der Chancenquotient $p(x/H_1)$ / $p(x/H_0)$ einen bestimmten Wert überschreitet.

- Weshalb wird mit log-odds-Scores gerechnet?
  - ➔ Lambda ist Normalisierungsfaktor, Werte werden auf ganze Zahlen gerundet. Somit wird der Logarithmus größer null und der Score positiv, wenn die Aminosäure häufiger im Alignment gefunden wird, als durch Zufall erwartet.
    Ähnliche Aminosäuren, bzw. welche die leicht ineinander mutieren sowie seltene Aminosäuren erhalten somit einen höheren Score.

- Wie wurden die BLOSUM-Matrizen abgeleitet? Was unterscheidet die BLOSUM N-Matrizen voneinander?
  - ➔ Grundlage sind Aminosäureeigenschaften/Proteindomänen. Vergleichen von Sequenzblöcken homologer Proteine, N gibt Sequenzidentität an. Von jedem Sequenzpaar, das N% identische Residuen aufweist, wird eine Sequenz eliminiert. Somit enthalten die Blöcke nur noch Sequenzen, welche im paarweisen Vergleich untereinander max. N% identisch sind.
  - ➔ Somit werden verwandte Sequenzen stärker gewichtet.

- Wie unterscheiden sich verschiedene Scoring-Matrizen?
  - ➔ Bewertung verschiedener Gesichtspunkte: Einheitsmatrix, empirische Matrizen (BLOSUM, PAM), basierend auf chemischen Eigenschaften oder genetischem Code.

- Bei welcher Fragestellung empfiehlt es sich, von der Standardeinstellung abzuweichen?
  ➔ Identifikation von Proteindomänen
  ➔ Betrachten evolutionärer Vorgänge

- Wozu dient die affine Kostenfunktion?
  ➔ Basieren auf Ähnlichkeitsmatrizen, die für jedes Paar angeben, wie wahrscheinlich es ist, dass dieses durch Evolution entstanden ist. Identische Paare erzielen sehr hohen Score, ähnliche hohen Score, stark unterschiedliche setzen Score herab. Rechnerisch bestes Alignment mit höchstem Score ist dann gleichzeitig das, welches bei Homologie zu erwarten wäre.

Markov-Ketten

- Was sind homogene und nicht-homogene Markov-Ketten?
  ➔ Angeben von Wahrscheinlichkeiten für das Eintreten bestimmter Ereignisse.
  ➔ Homogen: Übergangswahrscheinlichkeiten unabhängig von aktuellen Zeitpunkten.
  ➔ Nicht-homogen: Übergangswahrscheinlichkeiten abhängig vom Zeitpunkt, „Jeder dritte Zug bevorzugt nach Norden".

- Welche Anwendungen gibt es in der Bioinformatik?
  ➔ Modellierung und Identifikation von CpG-Inseln, methyliertes C mutiert gern zu T, Mutantion nahe Promotor jedoch unterdrückt, CpG-Inseln weisen also auf Promotoren hin.

- Wie unterscheiden sich Markov-Ketten von HMMs?
  ➔ Bei HMMs sind die Zustände verborgen, bei Markov-Ketten nicht. Außerdem kommen bei HMMs die Emissionswahrscheinlichkeiten hinzu.

Hidden-Markov-Modelle

- Wie ist ein HMM definiert?
  ➔ Ein HMM ist ein diskreter stochastischer Prozess, welcher die Beobachtung $x_1...x_n$ und die mit ihr verschränkte Zustandsfolge $Pi_1...Pi_n$ in n Schritten erzeugt.
  ➔ System wird durch Markov-Kette mit unbeobachteten Zuständen modelliert. Zustände verborgen, Zustandswahrscheinlichkeiten nur vom aktuellen Zustand abhängig, nicht vom vorherigen. Es werden nur Emissionen beobachtet, die je nach Zustand mit gewissen Wahrscheinlichkeiten auftreten.

- Was ist der Viterbi-Pfad und wie wird er berechnet?
  ➔ Viterbi-Algorithmus berechnet wahrscheinlichste Abfolge von verborgenen Zuständen bei einem gegebenen HMM und einer beobachteten Sequenz von Emissionen. Diese Abfolge nennt man Viterbi-Pfad.

- Wo werden HMMs in der Bioinformatik eingesetzt?
  ➔ Pfam-Datenbank, Genvorhersage, CpG-Inseln, Hmmer, 3D-Homologie-Modellierung

- Wie werden HMMs aus MSAs abgeleitet?
  ➔ Ableiten der Übergangswahrscheinlichkeiten aus MSAs, Emissionswahrscheinlichkeit definiert sich auch durch Spaltenzusammensetzung für eine Position.

- Warum sind HMMs zum Charakterisieren von Proteinfamilien besser geeignet als z.B. Profile?
  ➔ Nicht homogenes HMM, Wahrscheinlichkeiten für Insertionen/Deletionen hängen vom Zeitpunkt/Position ab. So werden Lücken positionsabhängig bewertet!

BLAST

- Was unterscheidet BLAST von den exakten Vergleichsverfahren?
  ➔ Keine perfekte Übereinstimmung von Nöten, hinreichende Ähnlichkeit zwischen Eingabe und DB-Sequenz ausreichend (BLOSUM62).

- Wie arbeitet PSI-BLAST und wozu dient es?
  ➔ Iteratives Verfahren um aus Treffern einer BLASt-Suche Profile abzuleiten. Für jeden Iterationsschritt wird eine positionsspezifische Scoring-Matrix generiert.
  Bestimmen der Sequenzen für MSA, Konstruieren MSA, Berechnen des Profils aus Verbundwahrscheinlichkeiten für das anschließende BLASTen.

- Welche "algorithmischen Tricks" werden in BLAST umgesetzt, um die Geschwindigkeit zur Berechnung eines Alignments zu steigern?
  ➔ Preprocessing: Induzierung, Heraussuchen der Sequenzen der Datenbank, in welcher n-mere der Suchsequenz vorkommen
  ➔ Frühzeitiger Abbruch: Abbrechen des Vergleichs, wenn lokaler Score hinreichend niedrig

- Welche Art von Alignments bestimmt BLAST überhaupt?
  ➔ Lokale Alignments.

- Welche Trefferquote erwarten Sie bei Homologen, die ca. 40 % Sequenzidentität besitzen?
  ➔ 11% (FASTA: 15%).

- Wie ist der E-value definiert? Welche E-Werte sind statistisch signifikant?
  ➔ Bezeichnet Wahrscheinlichkeit, den betrachteten Treffer durch Zufall zu generieren bei gegebener Datenbank. Signifikanz bzw. cut-off ist davon abhängig, was man sucht. Ab 1e-10 meist?

- Welches Programm wählen Sie, wenn Sie globale alignments berechnen wollen?
  ➔ Needleman-Wunsch-BLAST, bzw. ClustalW bei MSAs.

- Erläutern Sie die Vorgehensweise von PSI-BLAST. Um welchen Faktor und weshalb ist PSI-BLAST empfindlicher als BLAST?
  ➔ PSI-BLASt ist viel empfindlicher was die Ermittlung weit verwandter Proteine betrifft. Für jeden Schritt wird positionsspezifische Scoring-Matrix generiert, Berechnen eines Profils aus Verbundwahrscheinlichkeiten und Einbeziehen der nächsten Verwandten.
  ➔ Zuerst wird eine Liste aller sehr ähnlichen Proteine erstellt. Über diesen Proteinen wird ein Profil erstellt, eine Art gemittelte Sequenz. Daraufhin sendet man mit diesem Profil erneut eine Suchanfrage an die Proteindatenbank und erhält eine größere Gruppe ähnlicher Sequenzen. Mit dieser Gruppe kann man wieder ein neues Profil erstellen und den Prozess beliebig oft wiederholen. Dadurch, dass verwandte Proteine in die Suche miteinbezogen werden, ist PSI-BLAST viel empfindlicher bei der Ermittlung weit entfernter Verwandtschaften als das gewöhnliche Protein-Protein BLAST.

Multiple Sequenzalignments

- Was unterscheidet T-Coffee von ClustalW?
  ➔ ClustalW: Iteratives, progressives MSA unter Verwendung eines guideTree (Probleme: Falsche anfängliche Paarung kann nicht korrigiert werden, Alignments stark von Scoring-System abhängig, basiert auf globalem Alignment, ungeeignet für Sequenzen mit wenig gemeinsamen Domänen)
  ➔ T-Coffee: Kombination von lokalem und globalem Alignment, zunächst paarweiser Vergleich zur Berechnung von PW-Scores (unterschiedlichste Methoden, Berechnung lokaler Ähnlichkeiten, strukturelle Korrespondenzen). Diese Scorres steuern iteratives Alignment, restlicher Algorithmus wie bei ClustalW. Bewertung der Alignments durch Anteile der identischen Residuen.

- Für welche Fragestellungen werden MSAs berechnet?
  ➔ Charakterisierung wichtiger Residuen, Berechnen von Profilen, Phylogenetische Analyse, Strukturvorhersage.

- In welchen Algorithmen werden sie weiterverarbeitet?
  ➔ Z.B. Viterbi-Algorithmus

- Was unterscheidet einen paarweisen Vergleich von einem MSA?
  ➔ Alignment mehrerer Sequenzen lässt konservierte Residuen besser hervortreten (Funktionalität, Sekundär- sowie Tertiärstruktur). In paarweisen Alignments kann die Relevanz einzelner Residuen nicht bestimmt werden.

Phylogenetische Verfahren

- Was unterscheidet die Begriffe Homologie und Ähnlichkeit?
  ➔ Homologie: Ähnlichkeit von Prozessen/Strukturen/Verhalten durch gleichen evolutionären Ursprung/Vorfahren.
  ➔ Ähnlichkeit: Neutraler Begriff, in Prozent anzugeben.

- Welches Signal wird bei der Berechnung von Bäumen üblicherweise ausgewertet?
  ➔ Moleküle werden als Uhren verwendet, Mutationsrate muss genau zum Zeitraum passen, der untersucht werden soll, langsam laufende Uhren von Nöten (sonst Rückmutation möglich)
  ➔ 16S rRNA, nahe verwandte Bakterien

- Welche Verfahren werden bei welcher Vorgehensweise verwendet?
  ➔ Phaenetische Verfahren:Bewertung des Phänotyps ohne Evolutionsmodell, distanzbasierte Methoden, Clusterverfahren.
  ➔ Kladistische Verfahren: Bewerten evolutionärer Vorgänge unter Verwendung eines Modells, Maximum Parsimony und Maximum Likelihood.

- Wozu dient das Neighbour-Joining Verfahren? Welche Art von Baum wird hierbei berechnet?
  ➔ Generiert aus Matrix D einen Baum mit ungerichteten und gewichteten Kanten.
  ➔ Additiver Baum, wenn D additiv.

- Welches Ziel verfolgen Maximum-Parsimony-Verfahren, welches ML-Ansätze?
  ➔ Maximum-Parsimony: Finde minimale Anzahl von Mutationen, um vorliegende Sequenzmenge zu beschreiben.
  ➔ ML-Ansatz: Bestimme den Baum mit höchster Wahrscheinlichkeit für die vorliegenden Daten.

- Wie wird die Likelihood eines Baumes berechnet?
  ➔ Mithilfe von Markov-Prozess und Modell für evolutionäre Entwicklung. Jede Base i mutiert mit konstanter Mutationsrate, also Übergang von Base i zu Base j mit konstanter Rate. Berechnen der Wahrscheinlichkeit, dass nach t Generationen keine Mutation auftritt. Daraus Berechnen der Übergangswahrscheinlichkeiten. Voraussetzung: Alle Positionen mutieren unabhängig voneinander.

- Wie funktioniert Quartett-Puzzle?
  ➔ Rekonstruktion eines Baumes aus allen bekannten Quartetten, Einfügen aller Bäume und Summieren der Scores an den Kanten, an denen unbekanntes Taxon nicht eingefügt werden soll. Mehrfaches Ausführen des Algorithmus, dann Ableiten des Konsensusbaum und Berechnen der Likelihood. Optimierung der Kantenlängen.

- Wozu dient das Bootstrapping, wie wird es ausgeführt?
  ➔ Resampling Methode mit konstantem Stichprobenumfang. Spalten werden mehrfach verwendet und Prozentzahlen generiert, wie oft angezeigte Kante in allen Bäumen vorkommt. Dies gibt Zuverlässigkeit/Reproduzierbarkeit für diese Kante an.

- Wie werden Outgroups gewählt?
  ➔ Zusätzliche Spezies(gruppe), welche mit allen anderen Kandidaten nur wenig verwandt ist. Dient als Plausiblitätskontrolle, wenn Sequenzen der outgroup breit über den Baum verstreut, ist dieser nicht interpretierbar.

- Wie kann in wurzellosen Bäumen die Lage der Root eingegrenzt werden?
  ➔ GC-Gehalt der rRNA, Gehalt an Aminosäuren, nhPhyloBayes.