Aufgabe Bootstrap p-Wert

- Erzeugen Sie sich zwei Stichproben X, Y vom Umfang 50 bzw. 75 aus der N(1.4, 2) für X und N(1,2) für Y.
- Simulieren Sie den p-Wert mittels Bootstrapping für t_p über die gepoolte Stichprobe für die Hypothese $\mu_x < \mu_y$.
- Nutzen Sie die Replikationszahlen 100, 1000, 10000!
- Simulieren Sie den p-Wert für die Hypothese $\mu_x < \mu_y$ mittels Bootstrapping für t_w über das stichprobenweise Mittelwertbereinigen.
- Führen Sie die Versuche mehrfach durch und vergleichen Sie die Größen. Ist das Ergebnis plausibel?

Lösung

- Zunächst t_p:

```
xd <- rnorm(50, mean=1.4, sd=sqrt(2))
yd <- rnorm(75, mean=1, sd=sqrt(2))

tp0 <- (mean(xd) - mean(yd)) /
        sqrt( (49 * var(xd) + 74 * var(yd))/123 *
                                        (1/50 + 1/75) )

tp <- function( xd, yd) {
  m <- length(xd)
  n <- length(yd)
  (mean(xd) - mean(yd)) /
        sqrt( ((m-1) * var(xd) + (n-1) * var(yd))/(n+m-2) *
                (1/m + 1/n) )
}
```

```
runs <- 1000
bootstraptp <- rep(NA, runs)
pooleddata <- c(xd, yd)

for ( i in 1:runs){
  xdstar <- sample(pooleddata,length(xd), replace=TRUE)
  ydstar <- sample(pooleddata,length(yd), replace=TRUE)
  bootstraptp[i] <- tp(xdstar, ydstar)
}

(boot.p.value <- length(which(bootstraptp >= tp0))/ runs)
```

- Dann t_w:

```
tw0 <- (mean(xd) - mean(yd)) /
        sqrt((var(xd)/length(xd) + var(yd)/length(yd)))
```

```
tw <- function(xd, yd) (mean(xd) - mean(yd)) /
        sqrt((var(xd)/length(xd) + var(yd)/length(yd)))

xdber <- xd - mean(xd)
ydber <- yd - mean(yd)

runs <- 1000
bootstraptp <- rep(NA, runs)

for ( i in 1:runs){
  xdstar <- sample(xdber,length(xd), replace=TRUE)
  ydstar <- sample(ydber,length(yd), replace=TRUE)
  bootstraptp[i] <- tw(xdstar, ydstar)
}

(boot.p.value <- length(which(bootstraptp >= tp0))/ runs)
```

- Vergleich der beiden Verfahren. Da das Poolen in einer Stichprobe die Varianz der resultierenden Stichprobe stark erhöht, erwartete ich zu einer gegebenen Ausgangsstichprobe (X, Y) die besseren Ergebnisse für den 2. Fall. Bessere Ergebnisse sollte in diesem Fall heißen, dass kleinere Bootstrap-p-Werte auftreten.

- Simulation:

```
pWertVergleich <- function(durchlaeufe=100, runs = 10000){
    result <- matrix(NA, ncol=2, nrow=durchlaeufe)
    for (j in 1:durchlaeufe){
    xd <- rnorm(50, mean=1.4, sd=sqrt(2))
    yd <- rnorm(75, mean=1, sd=sqrt(2))
    tp0 <- tp(xd, yd)
    bootstraptp <- rep(NA, runs)
    pooleddata <- c(xd, yd)
```

```
for ( i in 1:runs){
  xdstar <- sample(pooleddata,length(xd), replace=TRUE)
  ydstar <- sample(pooleddata,length(yd), replace=TRUE)
  bootstraptp[i] <- tp(xdstar, ydstar)
}
boot.p.value <- length(which(bootstraptp >= tp0))/ runs
result[j,1] <- boot.p.value
cat("p-Wert Bootstrap gepooled : ", boot.p.value, " --- ")
xdber <- xd - mean(xd)
ydber <- yd - mean(yd)
tw0 <- tw(xd, yd)
bootstraptw <- rep(NA, runs)
for ( i in 1:runs){
  xdstar <- sample(xdber,length(xd), replace=TRUE)
  ydstar <- sample(ydber,length(yd), replace=TRUE)
  bootstraptw[i] <- tw(xdstar, ydstar)
}
```

```
    boot.p.value <- length(which(bootstraptw >= tp0))/ runs
    result[j,2] <- boot.p.value
    cat("p-Wert Bootstrap mittelwertbereinigt : ",
        boot.p.value, "\n")
  }
  result
}

daten <- pWertVergleich()
 mean(daten[,1] -daten[,2])
[1] -0.00013
```

- Die Erwartung wurde nicht bestätigt. Warum?

- Durch die Wahl beider Stichproben aus der Normalverteilungsfamilie erzeugen beide Verfahren identische Voraussetzungen unter H_0!

- Tatsächlich ist es bei Stichproben dieser Größenordung und Tests auf Mittelwerte gar nicht mehr so einfach deutlich von dieser Familie abzuweichen, da der ZGWS und das Gesetz der großen Zahlen bereits greifen. Auch ist fraglich, wenn ich diese Familie verlasse, ob die t-Statistiken überhaupt noch die richtigen Teststatistiken sind.

- In weiteren Experimenten ist es mir nur gelungen Fälle zu produzieren, in denen die erste Art die Nullhypothese zu erzwingen, also das Poolen, kleinere p-Werte erzeugt.

- Um zu einem begründeten Fazit im Vergleich dieser Verfahren zu kommen, müsste man noch bessere Experimente designen! Ist überhaupt immer ein kleinerer p-Wert besser?

Permutationstests

- Permutationstests sind eine Variante des Resampling, die auf Ziehen ohne Zurücklegen der kompletten Stichprobe $\mathbf{x}$ beruht.

- Diese Art Ziehung ist offensichtlich äquivalent zu einer reinen Umordnung der Beobachtungen.

- Den Name Permutationstest leitet sich von der Nullhypothese ab, dass jede Permutation der Originaldaten dieselbe Wahrscheinlichkeit habe!

- Es werden Eigenschaften von bekannten Untergruppen der Grundgesamtheit verglichen. Insbesondere sind Gruppenzugehörigkeiten bekannt und die Tests beziehen sich auf die Unterschiede in ausgewählten Teststatistiken der Untergruppen.

- In Permutationstests werden nicht nur einzelne Kenngrößen der Verteilungen verglichen, sondern implizit stets die Gleichheit der Verteilungen in Gänze auf den Prüfstand gestellt!

- Deshalb ist die Wahl der Teststatistik getrieben von der Alternative gegen die man die Nullhypothese abgrenzen will.

- Interessanterweise ist auch diese computerintensive Methode älter, als der Computer selbst.

- In die Literatur eingeführt von Sir R. Fisher bereits in den 1930er Jahren.

- Ein Speziallfall, nämlich Fischer's exakter Test in der Vierfeldertafel, ist bereits aus Statistik II bekannt.

Einführendes Beispiel für Permutationstests

- Aus: Resampling Methoden - Skript Dortmund 2005 - Jenö Reiczigel
- Zwei Behandlungen einer Anämie (Blutarmut) sollen verglichen werden. Gemessen wird das Hämoglobin in g/dl Blut.
- Folgende Daten liegen in den Gruppen vor:
  Gruppe B (Behandlung): 9.1, 10.3, 11.0 , 11.5, 11.9
  Gruppe K (Kontrolle): 8.1, 8.4, 9.2, 9.4
- Die Stichproben sind klein (kein ZGWS), unterschiedliche Stichprobenumfänge, es gibt keinen Grund für eine Normalverteilungsannahme.
- Die Frage ist nun, ob sich eine signifikante Steigerung des Hämoglobingehaltes des Blutes durch die Behandlung bereits mit diesen wenigen Daten belegen lässt.

- Wichtig! Die Nullhypothese H_0 in Permutationstests lautet: Es gibt keinen Unterschied in den Verteilungen der beiden Teilstichproben!

- Die Prüfgröße anhand derer man beispielsweise einen Unterschied entdecken möchte, sei die Differenz der Mittelwerte der Stichproben.

- Im Beispiel
$$T_0 = \bar{B} - \bar{K} = 10.76 - 8.78 = 1.98.$$
Ist diese Steigerung signifikant?

- Unter der Nullhypothese sind alle Beobachtungen aus einer Verteilung gezogen worden, also i.i.d.

- Damit wären die wahren Gruppenzugehörigkeiten belanglos, man dürfte also die "Schilder" B und K frei vertauschen ohne den Informationsgehalt der Stichprobe zu ändern.

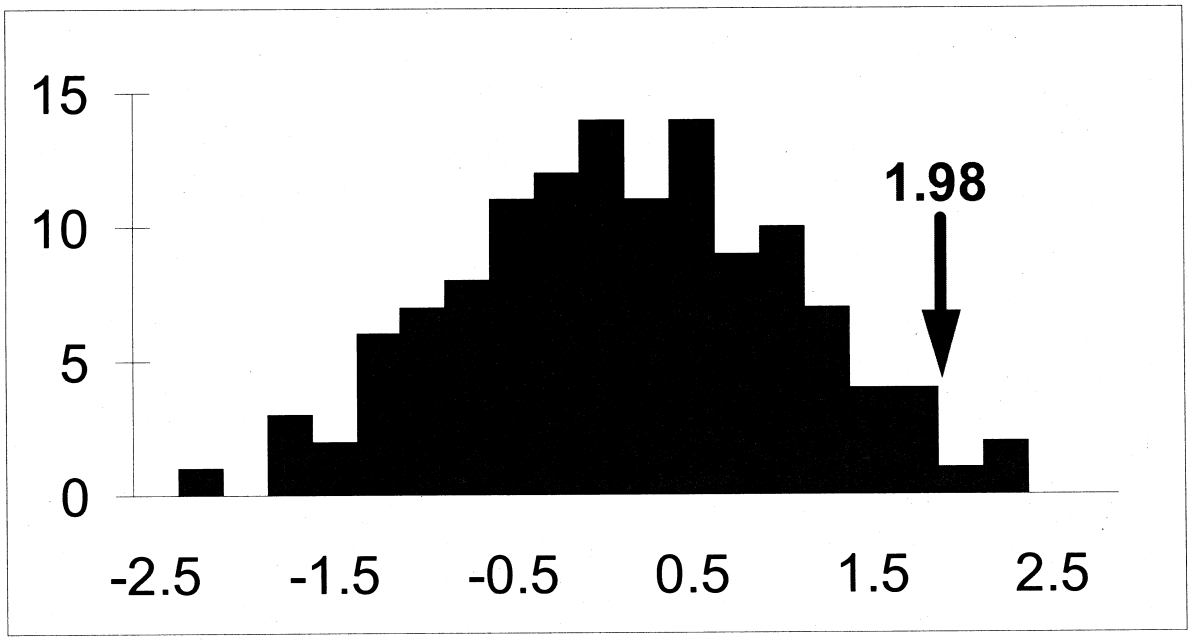
- Für einen (vollständigen) Permutationstest erhebt man für alle möglichen Permutationen P_i der Gruppenzugehörigkeiten die Teststatistik T_i und bestimmt aus diesen Daten den empirischen p-Wert von T_0.

- In unserem Beispiel gibt es $\frac{9!}{5!4!} = 126$ verschiedene Möglichkeiten die Gruppenzugehörigkeiten zu permutieren.

- Es werden also 126 Werte T_i berechnet. Diese befinden sich alle im folgenden Histogramm.

Histogramm aller Realisierungen des Permutationstests

Von den 126 Permutationen gibt es nur 3 mit Werten größer oder gleich 1.98.



Der Unterschied ist signifikant: $p = 3/126 = 0.0238$.

Formalisierung des (Zweistichproben-)Permutationstests

- Gegeben sei eine Stichprobe $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$ vom Umfang $N = m + n$, bestehend aus zwei Teilstichproben $\mathbf{x}_a$ vom Umfang m und $\mathbf{x}_b$ vom Umfang n aus bekannten Subpopulationen.
- Zum Mittelwertvergleich wird der Parameter $\hat{\theta} = \bar{x_a} - \bar{x_b}$ berechnet.
- Dann werden alle $\binom{N}{m}$ möglichen Zuordungen in m-elementige Teilstichproben generiert und diese jeweils x_a^* genannt.
- Die jeweils übrigen Elemente bilden x_b^*.
- Für jede solche Permutation wird $\hat{\theta}^* = \bar{x_a^*} - \bar{x_b^*}$ berechnet.

- Dann sei definiert

$$p_{perm} := \frac{\text{Anz. } \hat{\theta}^* \geq \hat{\theta}}{\binom{N}{n}}$$

als der p-Wert zur Teststatistik $\hat{\theta}$ aus dem Zweistichproben-Permutationstest für eine Hypothese, bei der große Werte zur Ablehnung führen ($\mu_{x_a} \leq \mu_{x_b}$).

- Werden tatsächlich alle Permutationen erhoben, spricht man von einem *exakten Test*.

- Da $\binom{N}{m}$ in der Praxis sehr schnell mit N und m riesig wird, ist eine Vollerhebung aller Permutationen in der Regel nicht möglich. In diesem Fall definiert man einen einen approximativen p-Wert für den Permutationstest über Monte-Carlo-Methoden.

- Anstatt alle Permutationen zu generieren beschränkt man sich auf B

zufällige Permutationen.

- Allerdings wirklich als Permutationen, nicht als Bootstrap, also Ziehen ohne Zurücklegen!

- Es werden B Stichproben ohne Zurücklegen des Umfangs n aus N gezogen und jeweils analog zum Vorgehen im vollständigen Fall die Statistiken $\hat{\theta}^* = \bar{x}_a^* - \bar{x}_b^*$ berechnet.

- Der approximative p-Wert für $\hat{\theta}$ ist dann definiert durch

$$p_{approx} := \frac{\text{Anz. } \hat{\theta}^* \geq \hat{\theta}}{B}.$$

- Die Ähnlichkeit im Vorgehen zu den Monte-Carlo bzw Bootstrap-Tests ist offensichtlich.

- Permutationstests sind nur für eine kleinere Anzahl von Hypothesen zu gebrauchen, da stets auf Gleichheit der Verteilungen überprüft wird!

- Die Teststatistiken werden nach den gewünschten Alternativen gewählt!

Konfidenzintervalle aus Permutationstests

- Zunächst einmal enthalten Permutationstests nicht unbedingt Parameter, da die vollständigen, exakten Verteilungen verglichen werden.
- In manchen Fällen lassen sich Hypothesen jedoch so umformulieren, dass sie sehr wohl bezüglich eines Parameters formuliert werden können.
- In diesen Fällen lassen sich Konfidenzintervalle für den Parameter der Hypothese durch die sog. Inversion des Testproblems angeben.
- Anhand eines Beispiels sei das prinzipielle Vorgehen erläutert.
- Im bekannten Test auf gleiche Mittelwerte, also auf gleiche Lage der Verteilung, kann $H_0 : F_X = F_Y$ so umformuliert werden, dass gegen eine

Verschiebung um einen Paramter θ getestet wird. Es wird angenommen $F_X(x) = F_Y(x + \theta)$ und die Hypothese wird zu $H_0 : \theta = 0$!

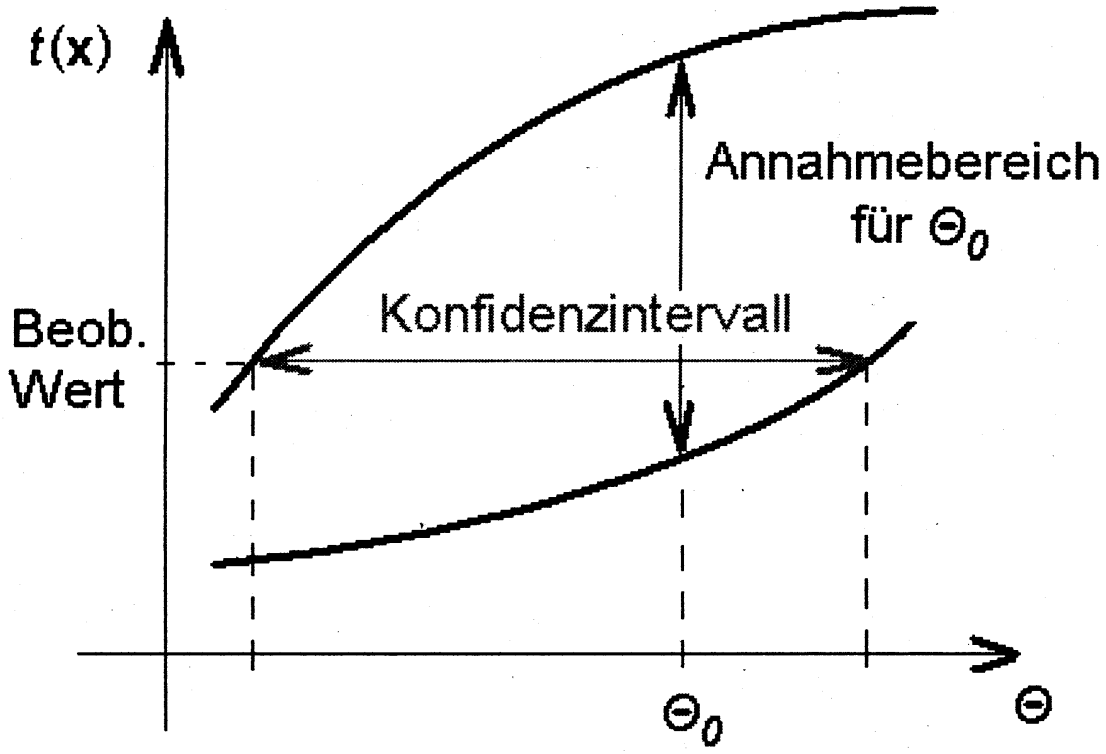
- Hat man nun einen exakten Test zum Niveau α für den Paramter θ über die Teststatisitk $T(\mathbf{x})$ für die Hypothese $H_0 : \theta = \theta_0$, dann definiert die Menge

$$KI_{1-\alpha,perm} := \{\theta_0 : T_0 \in A(\theta_0)\}$$

ein exaktes Konfidenzintervall für θ zum Niveau $(1-\alpha)$. $A(\theta_0)$ bezeichnet den Bereich, in dem die Hypothese H_0 nicht verworfen wird.

- Das folgende Bild macht diese Größen anschaulich.

Konfidenzintervall durch Inversion eines Tests



Mathematische Grundlagen der Zulässigkeit von Permutationstests

- Der theoretische Begriff hinter diesen Tests ist die sog. Austauschbarkeit von Zufallsvariablen.

- **Definition:** Seien $X_1, \ldots, X_n$ ZVen mit der gemeinsamen Verteilungsfunktion F. Diese ZVen werden *austauschbar* genannt, wenn

$$F(x_1, x_2, \ldots, x_n) = F(x_{i1}, x_{i2}, \ldots, x_{in})$$

  für alle Werte $x_1, x_2, \ldots, x_n$ und alle Permutationen $(x_{i1}, x_{i2}, \ldots, x_{in})$ dieser Wert.

- Bei vorliegender Austauschbarkeit ist ein Permutationstest stets exakt und unverzerrt! (Lehmann, 1986)

- Wann liegt Austauschbarkeit vor?

- Hinreichend ist $X_1, X_2, \ldots, X_n$ sind i.i.d.

- Auch Stichprobenziehen mit Zurücklegen erzeugt austauschbare ZVen.

- Unter Austauschbarkeit besitzt jede Permutation der n Beobachtungen dieselbe Wahrscheinlichkeit $\frac{1}{n!}$!

Bemerkungen zu Permutationstests

- Bei Permutationstests handelt es sich immer um bedingte Tests. Alle Größen werden stets bedingt unter der beobachteten Stichprobe berechnet.

- Alle abgeleiteten Größen gelten also nur für die beobachtete Stichprobe!

- Darf man von dieser Stichprobe auf die Gesamtheit schliessen?

- Die Kehrseite dieser Betrachtungen ist die Randomisation in klinischen Studien.

- Die Patienten sind keine zufällige Stichprobe der Bevölkerung, trotzdem werden die Behandlungen zufällig zugeordnet. Als Analysemethode wird

trotzdem fast immer der t- oder der F-Test gewählt, obwohl diese Tests eher für zufällige Stichproben geeignet wären!

- Der größte Vorteil der Permutationstests ist die freie Wahl der Teststatistik, je nach gewünschter Alternative!

Aufgabe zu Permutationen

- Implementieren Sie einen approximativen Permutationstest für die Stichprobenkonstellation $X \sim N(1.4, 2)$ vom Umfang 50, $Y \sim N(1, 2)$ vom Umfang 75 und die Hypothese, dass $\mu_X = \mu_Y$ über die Teststatistik $\bar{X} - \bar{Y}$. (Dies ist äquivalent zur Bestimmung des p_{approx}.)

- Überlegen Sie hierzu, welche Parameter festgelegt und aus welchen Grundgesamtheiten welche Stichproben gezogen werden müssen.

- Inwiefern ist Fischer's exakter Test ein Permutationstest?

Zusatzaufgaben

- (Zusatzaufgabe) Schreiben Sie ein Programm, dass alle Permutationen eines gegebenen Vektors ausgibt.

- (Zusatzaufgabe) Schreiben Sie ein Programm, dass alle Permutationen für den obigen Zweistichprobenfall erzeugt. Gegeben seien zwei Vektoren, einmal die Messwerte und einmal die Gruppenzugehörigkeiten.