Lösungen der Übungen zum Atelier "Statistik in Maturarbeiten"

Hansruedi Künsch

TMU Frauenfeld, 11. September 2019

1. Lichtgeschwindigkeit Fragestellungen: Wie lautet der Schätzwert für die Lichtgeschwindigkeit basierend auf diesen Daten? Wie genau ist die Lichtgeschwindigkeit damit bestimmt? Hatte Michelson's Messreihe einen systematischen Fehler?

Benutzter Test: Die Fragestellung betrifft die Lage eines metrischen Merkmals in einer Grundgesamtheit, der Menge aller Messungwerte, die Michelson mit seiner Versuchsanordnung hätte erhalten können. Messfehler werden häufig als normalverteilt angenommen, was für den t-Test spricht. Zum Vergleich führen wir aber auch den Wilcoxon-Test durch. Wir arbeiten nur mit den gemessenen Zeiten. Der Median und das Vertrauensintervall des Wilcoxon-Tests lassen sich direkt umrechnen auf die Geschwindigkeit, der P-Wert des Wilcoxon-Tests ändert sich dabei nicht. Beim arithmetischen Mittel und beim t-Test, spielt es jedoch eine Rolle, ob man die Werte für die Zeit oder für die Geschwindigkeit benutzt.

R-Befehle:

```
> licht <- read.csv("michelson.csv",header=TRUE,sep=";")
> licht <- licht$zeit
> mean(licht)
> median(licht)
> hist(licht)
> plot(licht,xlab="Nummer der Messung", ylab="Distanz")
> licht.korr <- licht[-c(6,10)] ## Weglassen der Beobachtungen 6 und 10
> t.test(licht,mu=33)
> t.test(licht.korr,mu=33)
> wilcox.test(licht,mu=33,conf.int=TRUE)
> wilcox.test(licht.korr,mu=33,conf.int=TRUE)
```

Bemerkung zum Einlesen der Daten: Das File michelson.csv muss im working directory von R abgespeichert sein. Mit dem Befehl `> getwd()` erhalten sie den Namen des working directory, mit dem Befehl `<- setwd("folder")` können Sie das directory mit dem namen folder zum working directory machen.

Interpretation: Aus dem Histogramm oder dem Plot der Messungen gegen die Reihenfolge sind zwei Werte ersichtlich, die weit weg vom Rest liegen und welche die Annahme einer Normalverteilung in Frage stellen. Es scheint plausibel, diese beiden Werte als Ausreisser zu bezeichnen und wegzulassen. Auf den Wilcoxon-Test und das zugehörige Vertrauensintervall hat das Weglassen der Ausreisser praktisch keine Auswirkungen. Hingegen wird beim t-Test der P-Wert dadurch deutlich kleiner und das Vertrauensintervall verschiebt sich nach rechts und wird kürzer. Dass der P-Wert durch das Weglassen der Ausreisser kleiner wird, erscheint zunächst paradox, denn die Ausreisser liegen ja weit weg vom hypothetischen Wert 33.02. Das Weglassen reduziert jedoch auch die Standardabweichung der Daten, und diese Standardabweichung beeinflusst ebenfalls den P-Wert. In jedem Fall ist es offensichtlich, dass Michelson's Messungen einen systematischen Fehler hatten. Das

Vertrauensintervall gibt an, in welchem Bereich das arithmetische Mittel etwa gelegen hätte, falls Michelson eine viel längere Messreihe gemacht hätte. Es berücksichtigt den statistischen, aber nicht den systematischen Fehler seiner Methode.

2. Geburten Fragestellungen: a) Wie wahrscheinlich ist eine Knabengeburt etwa? Wie genau ist diese Wahrscheinlichkeit bestimmt? Könnte der Knabenüberschuss auch zufällig zustande gekommen sein? b) Ist der Anteil der Zwillinge verschiedenen Geschlechts und damit auch der Anteil zweieiiger Zwillinge 2015 systematisch verschieden vom entsprechenden Anteil im Jahr 1975? Wie gross war etwa die Wahrscheinlichkeit für zweieiige Zwillinge in den Jahren 1975, bzw. 2015?

Benutzte Tests: Bei a) geht es um den Anteil des kategorialen Merkmals "Geschlecht", also kommt der Binomialtest zum Zug. Bei b) geht es um den Anteil des kategorialen Merkmals "Verschiedenes Geschlecht" in den beiden Grundgesamtheiten, aus denen die Zwillingsgeburten 1975, bzw. 2015 gezogen wurden. Dazu verwendet man den Chi-Quadrat-Homogenitätstest. Wenn p die Wahrscheinlichkeit bezeichnet, dass Zwillinge zweieiig sind, dann ist die Wahrscheinlichkeit für "Verschiedenes Geschlecht" bei Zwillingen $\approx \frac{p}{2}$ (oder wegen a) genauer $= p \cdot 2 \cdot 0.51 \cdot 0.49 = p \cdot 0.4998$). Aus dem Vertrauensintervall für die Wahrscheinlichkeit für "Verschiedenes Geschlecht" bei Zwillingen erhält man also durch Multiplikation mit 2 ein Vertrauensintervall für die Wahrscheinlichkeit zweieiiger Zwillinge.

R-Befehle:

```
> binom.test(x=41111,n=80290)
> zwillinge <- matrix(c(754-205,205,1580-571,571),nrow=2)
> chisq.test(zwillinge)
> binom.test(205,754)
> binom.test(571,1580)
```

Interpretation: a) Der Zufall als Erklärung für den grösseren Anteil von Knaben kann aufgrund des P-Wertes praktisch ausgeschlossen werden. Die Wahrscheinlichkeit für eine Knabengeburt ist leicht höher als 50%, sie liegt mit 95% Sicherheit im Bereich $51.2\% \pm 0.35\%$. b) Die Nullhypothese, dass die Wahrscheinlichkeit für Zwillinge verschiedenen Geschlechts in beiden Jahren gleich war, wird klar verworfen. Ein Binomialtest berechnet je ein 95%-Vertrauensintervall für die Wahrscheinlichkeit für "Verschiedenes Geschlecht" in den Jahren 1975 und für 2015 (die Wahrscheinlichkeit, dass beide Intervalle den wahren Wert enthalten, ist dann $0.95^2 = 0.9025$). Daraus erhält man für die Wahrscheinlichkeit, dass ein Zwillingspaar zweieiig ist, im Jahr 1975 den Bereich $[48\%, 61\%]$ im Jahr 1975 und im Jahr 2015 den Bereich $[67.5\%, 77\%]$. Die Zunahme ist plausibel, denn sowohl das Alter der Mütter als auch die Anzahl von Hormonbehandlungen und künstlichen Befruchtungen haben zugenommen.

3. Nikotin und Blutgerinnung Fragestellung: Führt Nikotin dazu, dass sich Blutplättchen leichter aggregieren?

Benutzter Test: Es geht um den Vergleich von einem Lagemass in den zwei Grundgesamtheiten vor, bzw. nach dem Rauchen einer Zigarette. Es handelt sich um eine gepaarte Stichprobe. Zur Auswahl stehen der t-, der Vorzeichen- und der Wilcoxon-Test. Da der Anteil in Prozent gemessen wird, ist die Annahme der Normalverteilung fraglich. Aus medizinischen Gründen erwartet man eine Zunahme der Aggregation durch das Nikotin, daher kann einseitig getestet werden. Um zu sehen, dass das Ergebnis unterschiedlich ausfällt, wenn die gleichen Daten nicht aus einer gepaarten Stichprobe kommen, kann man noch die beiden 2-Stichproben-Tests benutzen.

R-Befehle

```
> 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(x=nachher,y=vorher,paired=TRUE,alternative="greater")
> wilcox.test(x=nachher,y=vorher,paired=TRUE,alternative="greater",conf.int=TRUE)
> nachher-vorher ## zum Bestimmen  der Vorzeichen von Hand
> binom.test(10,11,alternative="greater") ## P-Wert des Vorzeichentests
> library("DescTools")
> SignTest(x=nachher,y=vorher,alternative="greater")
> t.test(x=nachher,y=vorher,alternative="greater")
> wilcox.test(x=nachher,y=vorher,alternative="greater",conf.int=TRUE)
```

Interpretation: Der P-Wert ist am kleinsten beim t-Test (0.08%) und am grössten beim Vorzeichentest (0.25%). Es gibt in jedem Fall aber wenig Grund zu bezweifeln, dass die Aggregation erhöht wurde. Beim t- und beim Wilcoxon-Test stimmen die 95%-Vertrauensintervalle praktisch überein. Wenn man die Paarung ignoriert und die Daten fälschlicherweise mit der 2-Stichproben-Variante der Tests analysiert, erhält man einen P-Wert von 8.6%, bzw. 8.9% , d.h. die Nullhypothese "Gleicher Erwartungswert mit und ohne Nikotin" wird auf dem 5%-Signifikanzniveau beibehalten.

4. Körper- und Hirnmasse von Säugetieren Fragestellung: Gibt es einen Zusammenhang zwischen der Körper- und der Hirnmasse? Um was für eine Art von Zusammenhang handelt es sich? Kann man aus der Kenntnis der Körpermasse die Hirnmasse vorhersagen? Welche Tierarten unterscheiden sich deutlich vom Rest?

Benutzte Tests: Als Zusammenhangsmass kann man die Pearson- oder die die Rang-Korrelation verwenden, und man kann die Nullhypothese testen, dass diese Masse in der Grundgesamtheit aller Säugetierarten gleich null sind. Falls der Zusammenhang linear ist, kann die Hirnmasse mit Hilfe der Körpermasse vorhergesagt werden, und man kann ein Vertrauensintervall für die Steigung der Geraden in der Grundgesamtheit angeben.

R-Befehle: Zunächst muss das Excel-File als .csv File gespeichert werden.

```
> gewicht <- read.csv("gewicht.csv",header=TRUE,sep=";")
> gewicht
> plot(gewicht$Koerper,gewicht$Gehirn)
> plot(log(gewicht$Koerper),log(gewicht$Gehirn))
> plot(gewicht$Koerper,gewicht$Gehirn,log="xy")
  ## Logarithmische Skala, aber Werte auf den Achsen in urspruenglicher Skala
> cor.test(gewicht$Koerper,gewicht$Gehirn)
> cor.test(gewicht$Koerper,gewicht$Gehirn,method="spearman")
> cor.test(log(gewicht$Koerper),log(gewicht$Gehirn))
> cor.test(log(gewicht$Koerper),log(gewicht$Gehirn),method="spearman")
> regr <- lm(log10(gewicht$Gehirn) ~ log10(gewicht$Koerper))
> abline(regr)
> summary(regr)
> confint(regr)
> regr$residuals
```

Bemerkung zum Einlesen der Daten: gewicht.csv muss im working directory von *R* gespeichert sein, siehe Aufgabe 1. Das Argument `sep=";"` in `read.csv` ist nötig, wenn nicht das Komma, sondern der Stichpunkt als Separator benutzt wurde. Excel macht das offenbar so, meine open office Version jedoch nicht; deshalb war dieses Argument bei 1. nicht nötig.

Interpretation: Das Streudiagramm und damit auch die gewöhnliche Korrelation wird von den beiden Elefantenarten dominiert. Obwohl das Vertrauensintervall für die Korrelation weit weg von 0 und nahe bei 1 ist, bleibt ein linearer Zusammenhang fraglich. Mit

den logarithmierten Daten wird das Streudiagramm nicht mehr von den beiden Elefantenarten dominiert, und ein linearer Zusammenhang von Gehirn- und Körpermasse auf der logarithmischen Skala scheint plausibel. Die Resultate des Korrelationstests sind mit und ohne Logarithmieren fast gleich bei der gewöhnlichen Korrelation und exakt gleich bei der Rangkorrelation von Spearman. Dass sich die Rangkorrelation nicht ändert, war von vorneherein klar, da sich die Ränge bei monotonen Transformationen nicht ändern. Bei der gewöhnlichen Korrelation war das jedoch nicht abzusehen: Diese Daten sind ein Beispiel dafür, dass man damit nur schlecht zwischen ganz verschiedenen Streudiagrammen unterscheiden kann.

Ein linearer Zusammenhang der logarithmierten Werte entspricht einem Potenzzusammenhang der ursprünglichen Werte

$$\log_{10}(y) \approx m \log_{10}(x) + q \Leftrightarrow y \approx 10^q x^m$$

Das 95%-Vertrauensintervall für m ist gleich $[0.53, 0.81]$, enthält also den Wert 1 nicht. Daher bestehen berechtigte Zweifel an einem linearen Zusammenhang zwischen Körper- und Gehirnmasse in der ursprünglichen Skala. Die Tierarten mit der grössten positiven Abweichung von der Regressionsgeraden (den grössten Residuen) sind wie erwartet der Mensch und drei der vier Affenarten. Die grösste negative Abweichung findet man beim asiatischen Elefanten.

5. **Mehrfachbefragung** Fragestellung: Es geht um ein kategoriales Merkmal in zwei Grundgesamtheiten (wahlberechtigte Amerikaner zu zwei verschiedenen Zeitpunkten), und es handelt sich um eine gepaarte Stichprobe. Die Nullhypothese besagt, dass der Anteil der Zustimmung in der Grundgesamtheit beide Male gleich ist.

   Benutzter Test: McNemar-Test. Der Zustimmungsanteil ist bei der ersten Befragung $\frac{944}{1600} = 0.59$ und bei der zweiten $\frac{880}{1600} = 0.55$. Wenn man nicht beachtet, dass es sich um eine gepaarte Stichprobe handelt, benutzt man den Chi-Quadrat-Test. Um zu zeigen, dass der Unterschied wesentlich ist, berechnen wir auch den P-Wert mit dem Chi-Quadrat-Test.

```
> zustimmung <- matrix(c(794,150,86,570),nrow=2)
> mcnemar.test(zustimmung)
> zustimmung.2 <- matrix(c(944,656,880,720),nrow=2)
> chisq.test(zustimmung.2)
```

   Interpretation: Der P-Wert mit dem korrekten Test (McNemar) ist $4 \cdot 10^{-5}$, also besteht kein Zweifel, dass die Nullhypothese verworfen werden sollte. Verwendet man fälschlicherweise den Chi-Quadrat-Test, dann ist der P-Wert viel grösser nämlich 2.45%. Die Nullhypothese würde dann auf dem 1% Signifikanzniveau nicht verworfen. In der Stichprobe nimmt die Zustimmungsrate um 4% ab. Ein Vertrauensintervall für die Änderung der Zustimmungsrate in der Grundgesamtheit wäre auch von Interesse, wird aber von R nicht geliefert. Das liegt auch daran, dass es kein exaktes, aber viele leicht unterschiedliche genäherte Vertrauensintervalle gibt.

6. **Effekt von negativer Beeinflussung auf Intelligenztests** Fragestellung: Es geht um ein Lagemass, nämlich den Unterschied in den beiden Testresultaten in den zwei Grundgesamtheiten "Neutrale Bedingungen im zweiten Test", bzw. "Negative Beeinflussung im zweiten Test". Die Kontroll- und die Behandlungsgruppe bestehen aus verschiedenen Individuen, die Stichproben sind somit unabhängig.

   Verwendeter Test: Es stehen der 2-Stichproben-t-Test und der 2-Stichproben-Wilcoxon-Test zur Verfügung. Da man von der Vermutung ausgeht, dass sich die Resultate bei negativer Beeinflussung verschlechtern, kann ein einseitiger Test verwendet werden

   R-Befehle:

```
> kontrolle <- c(5,0,16,2,9)
> behandlung <- c(6,-5,-6,1,4)
> t.test(x=behandlung, y=kontrolle,alternative="less")
> wilcoxon.test(x=behandlung, y=kontrolle, alternative="less")
```

Interpretation: Es besteht nur wenig Evidenz, dass die negative Beeinflussung zu systematisch unterschiedlichen Testergebnissen führt (P-Wert von 6.2%, bzw. 11.1%). Entsprechend liegt 0 im Vertrauensintervall. Das ist jedoch kein Nachweis dafür, dass negative Beeinflussung keine Auswirkungen hat, es kann auch daran liegen, dass die Stichprobengrösse zu klein ist.

Bei Fragen können Sie mir ungeniert eine email schicken an `kuensch@stat.math.ethz.ch`.