CAS Datenanalyse HS16 - DeskStat

Anpassungs- und Unabhängigkeitstests

Anpassungstests

- Ein Merkmal heisst multinominal wenn es kategorisch ist und in diskrete Klassen unterteilt wurde.

Anpassungstests

- Ein Merkmal heisst multinominal wenn es kategorisch ist und in diskrete Klassen unterteilt wurde.
- Wir vergleichen die beobachteten Häufigkeiten dieser Klassen mit erwarteten Häufigkeiten.

Anpassungstests

- Ein Merkmal heisst multinominal wenn es kategorisch ist und in diskrete Klassen unterteilt wurde.
- Wir vergleichen die beobachteten Häufigkeiten dieser Klassen mit erwarteten Häufigkeiten.
- H_0: die beobachteten und die erwarteten Häufigkeiten sind gleich.

Anpassungstests

- Ein Merkmal heisst multinominal wenn es kategorisch ist und in diskrete Klassen unterteilt wurde.
- Wir vergleichen die beobachteten Häufigkeiten dieser Klassen mit erwarteten Häufigkeiten.
- H_0: die beobachteten und die erwarteten Häufigkeiten sind gleich.
- H_a: die beobachteten und die erwarteten Häufigkeiten sind verschieden.

Anpassungstests

- Die Abweichung zwischen den beiden Häufigkeiten wird die dem χ^2-Wert gemessen:

$$\chi^2 = \sum_i \frac{(f_i - e_i)^2}{e_i}$$

Anpassungstests

- Die Abweichung zwischen den beiden Häufigkeiten wird die dem χ^2-Wert gemessen:

$$\chi^2 = \sum_i \frac{(f_i - e_i)^2}{e_i}$$

- Mit dem zugehörigen p-Wert der χ^2-Verteilung finden wir die Testentscheidung.

Anpassungstests

Problem: Die Datenmenge `survey` enthält auch Informationen zum Rauchverhalten der australischen Studierenden aus Adelaide.

```
library(MASS)
levels(survey$Smoke)

## [1] "Heavy" "Never" "Occas" "Regul"

smoke.freq <- table(survey$Smoke)
smoke.freq

##
## Heavy Never Occas Regul
##    11   189    19    17
```

Anpassungstests

Problem: Aufgrund einer früheren Vollumfrage kennt die Unileitung die Rauchstatistiken.

Heavy	Never	Occassionaly	Regular
4.5%	79.5%	8.5%	7.5%

Entscheiden Sie, ob die Stichprobe aus **survey** die Behauptung der Unileitung stützt. Arbeiten Sie mit einem Signifikanzniveau von 5%.

Anpassungstests

Antwort:

```
smoke.prob = c(.045, .795, .085, .075)
chisq.test(smoke.freq, p=smoke.prob)


##
##  Chi-squared test for given probabilities
##
## data:  smoke.freq
## X-squared = 0.10744, df = 3, p-value = 0.9909
```

Anpassungstests

Antwort: Der p-Wert ist deutlich grösser als 5%. Die Nullhypothese H_0 wird daher nicht verworfen. Die Stichprobe verträgt sich mit der Behauptung der Unileitung.

Aufgabe: Anpassungstests

Problem: Die Unileitung vermutet folgendes Rauchverhalten ihrer Studierenden.

Heavy	Never	Occassionaly	Regular
4.5%	79.5%	8.5%	7.5%

Prüfen Sie, ob die Stichprobe aus **survey** sich mit dieser Behauptung verträgt. Bestimmen Sie den p-Wert, ohne auf die Funktion `chisq.test` zurückzugreifen.

Lösung: Anpassungstests

```
f = table(survey$Smoke)
e = smoke.prob*length(survey$Smoke)
e
```

```
## [1]  10.665 188.415  20.145  17.775
```

```
d = f-e
chi = sum(d*d/e)
chi
```

```
## [1] 0.1112089
```

```
df = length(f)-1
pchisq(chi, df=df, lower=FALSE)
```

```
## [1] 0.9904592
```

Unabhängigkeitstests

- Die Zufallsvariablen X und Y sind unabhängig, wenn die eine Wahrscheinlichkeitsverteilung die andere nicht beeinflusst.

Unabhängigkeitstests

- Die Zufallsvariablen X und Y sind unabhängig, wenn die eine Wahrscheinlichkeitsverteilung die andere nicht beeinflusst.
- Wir vergleichen die beobachteten Schnitthäufigkeiten mit den erwarteten Häufigkeiten bei Unabhängigkeit.

Unabhängigkeitstests

- Die Zufallsvariablen X und Y sind unabhängig, wenn die eine Wahrscheinlichkeitsverteilung die andere nicht beeinflusst.
- Wir vergleichen die beobachteten Schnitthäufigkeiten mit den erwarteten Häufigkeiten bei Unabhängigkeit.
- H_0: die beobachteten und die erwarteten Häufigkeiten sind gleich. Die beiden Zufallsvariablen sind unabhängig.

Unabhängigkeitstests

- Die Zufallsvariablen X und Y sind unabhängig, wenn die eine Wahrscheinlichkeitsverteilung die andere nicht beeinflusst.
- Wir vergleichen die beobachteten Schnitthäufigkeiten mit den erwarteten Häufigkeiten bei Unabhängigkeit.
- H_0: die beobachteten und die erwarteten Häufigkeiten sind gleich. Die beiden Zufallsvariablen sind unabhängig.
- H_a: die beobachteten und die erwarteten Häufigkeiten sind verschieden. Die beiden Zufallsvariaben sind abhängig.

Unabhängigkeitstests

Problem: Untersuchen Sie, ob das Rauch- und Sportverhalten der Studierenden aus **survey** unabhängig sind. Die entsprechenden Variablen sind `smoke` und `Exer`. Arbeiten Sie mit einem Signifikanzniveau von 5%.

Unabhängigkeitstests

Antwort:

```
library(MASS)
tbl = table(survey$Smoke, survey$Exer)
tbl
```

```
##
##          Freq None Some
##    Heavy    7    1    3
##    Never   87   18   84
##    Occas   12    3    4
##    Regul    9    1    7
```

Unabhängigkeitstests

Antwort:

```
chisq.test(tbl)

## Warning in chisq.test(tbl): Chi-squared approximation may be
incorrect

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

H_0 wird nicht verworfen.

Unabhängigkeitstests

Erweiterte Antwort: Die Warnung beim `chisq.test` erscheint, weil gewisse Zelleneinträge der Tabelle `tbl` zu gering sind (<5). Wir fassen daher die zweite und dritte Spalte zu einer neuen Spalte zusammen.

```
ctbl = cbind(tbl[,"Freq"],tbl[,"None"] + tbl[,"Some"])
ctbl
```

```
##       [,1] [,2]
## Heavy    7    4
## Never   87  102
## Occas   12    7
## Regul    9    8
```

Unabhängigkeitstests

Erweiterte Antwort:

```
chisq.test(ctbl)


##
##  Pearson's Chi-squared test
##
## data:  ctbl
## X-squared = 3.2328, df = 3, p-value = 0.3571
```