Wiederholung zu χ^2-Tests in Standardfällen und Fishers exaktem Test

Dirk Metzler

14. April 2020

1 χ^2-Test für eine feste Verteilung

Ein Experiment habe r mögliche Ausgänge (z.B. $r = 6$ beim Werfen eines Würfels), die Nullhypothese besage, dass Ausgang i mit Wahrscheinlichkeit p_i eintritt ($p_1 = p_2 = \cdots = p_6 = 1/6$ im Würfelbeispiel). Nehmen wir an, wir beobachten in n unabhängigen Wiederholungen des Experiments Ausgang i B_i mal. Unter der Nullhypothese erwarten wir $E_i := \mathbb{E}[B_i] = np_i$ mal Augang i zu beobachten. Geben die Beobachtungen Anlass, an der Nullhypothese zu zweifeln? Dazu berechnen wir die Statistik $X^2 = \sum_i \frac{(B_i - E_i)^2}{E_i}$ (sie ist unter der Nullhypothese approximativ χ^2-verteilt mit $r - 1$ Freiheitsgraden, kurz χ^2_{r-1}) und vergleichen den so berechneten Wert mit dem $(1 - \alpha)$-Quantil der χ^2_{r-1}-Verteilung.

Wir können das Werfen eines fairen Würfels mit R simulieren, beispielsweise mit dem Befehl `sample` (R gibt mit `?sample` oder `help(sample)` zusätzliche Hilfe aus):

```
> sample(1:6,size=1)
[1] 3
> sample(1:6,size=1)
[1] 6
> sample(1:6,size=1)
[1] 2
```

Um zu prüfen, ob ein Würfel fair ist, könnten wir ihn beispielsweise 12.000 mal werfen und die Abweichungen der empirischen Häufigkeiten der Augenzahlen vom theoretischen Wert (jeweils 2.000) mit der χ^2-Statistik bewerten. Der folgende R-Code tut dies (mit dem von R simulierten Würfel):

```
> ergebnis<-numeric(6)  # erzeuge einen Vektor der Länge 6
> ergebnis
[1] 0 0 0 0 0 0
> for (i in 1:12000) {
   wurf<-sample(1:6,size=1)
   ergebnis[wurf]<-ergebnis[wurf]+1
 }
```

Wir finden

```
> ergebnis
[1] 2017 1942 1922 1994 2037 2088
```

X^2 „von Hand“ ausrechnen:

```
> ergebnis-2000
[1]  17 -58 -78  -6  37  88
> (ergebnis-2000)^2
[1]  289 3364 6084   36 1369 7744
> (ergebnis-2000)^2/2000
[1] 0.1445 1.6820 3.0420 0.0180 0.6845 3.8720
> sum((ergebnis-2000)^2/2000)
[1] 9.443
```

Wie wahrscheinlich ist unter der Nullhypothese ein Wert ≥ 9.443 für X^2?

```
> pchisq(9.443,df=5,lower.tail=FALSE)
[1] 0.09264644
```

(Die Wahrscheinlichkeit für einen Wert ≤ 9.443 erhalten wir z.B. mit `pchisq(9.443,df=5)`, was $1 -$ `pchisq(9.443,df=5,lower.tail=FALSE)` ausrechnet.) Das alles kann auch R für uns erledigen:

```
> chisq.test(ergebnis,p=rep(1/6,times=6))

Chi-squared test for given probabilities

data:  ergebnis
X-squared = 9.443, df = 5, p-value = 0.09265
```

Demnach: Wenn wir dieses Experiment oft wiederholten (und der von R simulierte Würfel wirklich fair ist), würden wir in ca. 9% der Fälle einen so großen Wert von X^2 erwarten. Sollten wir Rs Würfel misstrauen (ein p-Wert von 0.09 ist zwar „nicht signifikant“ im üblichen Sinne, aber schon recht klein)? Wir „würfeln“ noch weitere 12.000 Mal:

```
for (i in 1:12000) {
   wurf<-sample(1:6,size=1)
   ergebnis[wurf]<-ergebnis[wurf]+1
 }
> ergebnis
[1] 4086 3978 3940 3943 3963 4090
```

und finden nun

```
> chisq.test(ergebnis)

Chi-squared test for given probabilities

data:  ergebnis
X-squared = 6.0495, df = 5, p-value = 0.3014
```

was unsere Zweifel an Rs (Pseudo-)Zufallsgenerator zerstreut. (Wenn Sie diese Befehle selbst mit R ausführen, werden Sie andere Werte finden, da R den Zufallszahlengenerator jedesmal beim Start „frisch“ initialisiert.)

2 χ^2-Test auf Unabhängigkeit (oder Homogenität)

Rosen und Jerdee (Influence of sex role stereotypes on personnel decisions, *J. Appl. Psych.* **59**, 9–14, 1974) berichten folgendes Experiment: 48 Teilnehmern eines Management-Kurses wurde je eine (fingierte) Personalakte vorgelegt, und sie sollten anhand der Aktenlage entscheiden, ob sie die betreffende Person befördern oder die Akte zunächst ablegen und weitere Kandidaten begutachten würden. Die Akten waren identisch bis auf die Geschlechtsangabe — 24 waren als „weiblich“ und 24 als „männlich“ gekennzeichnet — und wurden rein zufällig an die Teilnehmer verteilt. Es kam zu folgendem Ergebnis:

	Weiblich	Männlich
Befördern	14	21
Ablegen	10	3

Kann das Zufall sein? In 35 von 48 Fällen wurde „Befördern“ entschieden, unter der Nullhypothese, dass Geschlechtsmarkierung und Beförderungsentscheidung unabhängig sind, würden wir also

```
> 24*35/48
[1] 17.5
```

beförderte männliche und ebensoviele beförderte weibliche Akten erwarten (und entsprechend jeweils 6.5 abgelegte). Die X^2-Statistik ist

```
> (17.5-14)^2/17.5+(21-17.5)^2/17.5+(10-6.5)^2/6.5+(3-6.5)^2/6.5
[1] 5.169231
```

sie ist unter der Nullhypothese „Geschlechtsmarkierung und Beförderungsentscheidung sind unabhängig“ approximativ χ^2-verteilt mit einem Freiheitsgrad ($1 = 4 - 1 - 1 - 1 = (2-1)\cdot(2-1)$: 4 Zellen, ein Freiheitsgrad geht für die feste Gesamtsumme, einer für das (prinzipiell) unbekannte Geschlechterverhältnis und einer für die (prinzipiell) unbekannte Beförderungswahrscheinlichkeit „verloren“; allgemein für eine $r \times s$-Häufigkeitstabelle $r \cdot s - 1 - (r-1) - (s-1) = (r-1)(s-1)$ Freiheitsgrade). Die Wahrscheinlichkeit, das eine χ^2_1-verteilte Zufallsgröße einen Wert ≥ 5.169231 annimmt, ist

```
> pchisq(5.169231, df=1,lower.tail=FALSE)
[1] 0.02299039
```

(Wir haben also berechtigte Zweifel an der Nullhypothese.)

Dasselbe mit R:

```
> pers<-matrix(c(14,10,21,3),ncol=2)
> pers
     [,1] [,2]
[1,]   14   21
[2,]   10    3

> chisq.test(pers,correct=FALSE)

Pearson's Chi-squared test

data:  pers
X-squared = 5.1692, df = 1, p-value = 0.02299
```

Bemerkung: Voreingestellt benutzt R in diesem Fall die sogenannte Yatessche Stetigkeitskorrektur (siehe `?chisq.test`), d.h. es berechnet $\widetilde{X}^2 = \sum_i \frac{(B_i - E_i - 0.5)^2}{E_i}$:

```
> chisq.test(pers)

Pearson's Chi-squared test with Yates' continuity correction

data:  pers
X-squared = 3.7978, df = 1, p-value = 0.05132
```

3 Fishers exakter Test

Der χ^2-Test auf Unabhängigkeit beruht auf einer Approximation: Für eine große Anzahl Beobachtungen n ist X^2 approximativ χ^2-Verteilt. Im Fall von 2×2-Häufigkeitstabellen können wir alternativ eine auf R.A. Fisher zurückgehende Permutationsidee verwenden:

Unter der Nullhypothese „Geschlechtsmarkierung und Beförderungsentscheidung sind unabhängig“ können wir die Beobachtungen alternativ folgendermaßen erklären: Es gab 35 wohlgesonnene und 13 strenge Gutachter, und die Akten samt ihren Geschlechtsmarkierungen wurden rein zufällig (ohne zurückzulegen) aus einer Urne gezogen und auf die Gutachter verteilt. Sei H die Anzahl männlich markierter Akten unter den 35 auf die wohlgesonnenen Gutachter entfallenen Akten. Es gibt $\binom{48}{35}$, mit R:

```
> choose(48,35)
[1] 192928249296
```

Möglichkeiten, 35 Akten aus den insgesamt 48 zu wählen. Unter der Nullhypothese ist

$$\mathbb{P}(H = 21) = \frac{\binom{24}{21}\binom{24}{14}}{\binom{48}{35}}$$

mit R:

```
> choose(24,21)*choose(24,14)/choose(48,35)
[1] 0.02057543
```

H ist unter der Nullhypothese *hypergeometrisch* verteilt mit Parametern $24, 24, 35$ (wir schreiben auch $\text{hypergeom}_{24,24,35}$; allgemein: Eine Urne enthalte m weiße und n schwarze Kugeln, wir ziehen k Kugeln ohne Zurücklegen. $\text{hypergeom}_{m,n,k}$ ist die Verteilung der Anzahl weißer Kugeln unter den gezogenen.) R kennt die hypergeometrische Verteilung: `dhyper` (Verteilungsgewichte), `phyper` (Verteilungsfunktion), `qhyper` (Quantilfunktion), `rhyper` (Simulation), siehe z.B. `?dhyper`. Die Wahrscheinlichkeit, im Beispiel (unter der Nullhypothese) eine derart extreme Bevorzugung der „männlichen“ Akten zu sehen, d.h. 21 oder mehr davon unter den „Beförderten“, ist

```
> dhyper(21,24,24,35)+dhyper(22,24,24,35)+dhyper(23,24,24,35)+dhyper(24,24,24,35)
[1] 0.02449571
```

(ein äquivalenter R-Befehl ist `phyper(20,24,24,35,lower.tail=FALSE)`). Wir haben also berechtigte Zweifel an der Nullhypothese (nebenbei bemerkt: der approximative p-Wert des χ^2-Tests und der „exakte“ p-Wert von Fishers Test sind hier fast identisch).