Statistik II

Paul Strimtu 3898312
Jakob Striegel 4351490

Abgabe 6: 24 Juni 2024

1 Sinusvenenthrombosen

In der folgenden Aufgabe soll mit Hilfe eines Binomialtests überprüft werden, ob die Anzahl an Sinusvenenthrombosen unter den Geimpften dafür spricht, dass das Risiko einer solchen Erkrankung durch die Astrazeneca-Impfung erhöht wird - dazu stehen die folgenden Daten zur Verfügung:

- Stichprobegröße $n = 5.000.000$
- W'keit krank ohne Impfung $p_0 = 0.000\,025\,\% = \frac{1}{4.000.000}$
- Beobachtete Anzahl an Sinusvenenthrombosen $k = 10$
- Signifikanziveau $\alpha = 0.01$

Im Folgenden sei auch für $i \in \{1, 2, 3, \cdots, 5.000.000\}$:

$$X_i = \begin{cases} 1 & \text{falls die } i\text{-te Person krank ist,} \\ 0 & \text{sonst.} \end{cases}$$

$H_0 : p_s \leq p_0$ besagt, dass die W'keit zur Erkrankung, nicht höher als p_0 ist.
$H_1 : p_s > p_0$ besagt, dass die W'keit zur Erkrankung mit der Impfung, höher als p_0 ist.

Seien dann

$$X \sim B(5.000.000, \frac{1}{4.000.000})$$
$$\mathbb{E}(X) = n \cdot p_0 = \frac{5.000.000}{4.000.000} = 1,25.$$

Ermittlung des Ablehnungsbereiches:

```
c <- 1-pbinom(1:10, size = 5000000, prob = 1/4000000);c

##  [1] 3.553642e-01 1.315323e-01 3.826903e-02 9.124269e-03 1.838082e-03
##  [6] 3.201275e-04 4.906446e-05 6.710912e-06 8.284818e-07 9.317918e-08

c2 <- subset(c, c < 0.01); c2[1]

## [1] 0.009124269
```

$1 - pbinom(4, \cdots) \approx 0,00912 < 0,01 = \alpha \Rightarrow C = \{5, 6, \cdots, 5.000.000\}$

Da $k = 10 \in C$, wird H_0 zum Niveau $\alpha = 0,01$ abgelehnt, d.h. dass die Impfung mit Astrazeneca das Risiko einer Sinusvenenthrombosen Erkrankung erhöht.

2 Buy-the-Dip

In dieser Aufgabe soll untersucht werden, ob die durchschnittliche prozentuale Veränderung des DAX an einem Tag systematisch von 0 abweicht, wenn der Index am Vortag um 5% oder mehr gefallen ist. Dazu werden zuerst die Tage mit einem Rückgang von mindestens 5% ($dayFallingBelow\_5$) ausgesucht und die %-Veränderungen des DAX an den jeweiligen Folgetagen ($dayNext$) gespeichert, um einen t-Test durchzuführen, der die systematische Abweichung von 0 überprüft.

$H_0 = \mu = \mu_0 = 0$.
$H_1 = \mu \neq \mu_0 = 0$.

```
DAX <- read.csv("DAX89-24.csv")$x*100

dayFallingBelow_5 <- which(DAX <= -5)

dayNext <- DAX[dayFallingBelow_5 + 1]
```

Sei dann die Teststatistik

$$T = \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

```
T <- (mean(dayNext)-0)/(sd(dayNext)/sqrt(length(dayNext)));T

## [1] 2.09492
```

$\Rightarrow T = 2.09492$

Der Ablehnungsbereich $C = (-\infty, -t_{1-\alpha/2,n-1}) \cup (t_{1-\alpha/2,n-1}, \infty)$ kann folgendermaßen in R berechnet werden:

```
qt(1-0.05/2, length(dayNext)-1)

## [1] 2.026192
```

$\Rightarrow C = (-\infty, -2.026192) \cup (2.026192, \infty)$
Somit gilt $T \in C$, sodass H_0 abgelehnt wird.

Zuletzt kann wird noch der p-Wert bestimmt:

```
pt(-T,length(dayNext)-1)+1-pt(T,length(dayNext)-1)

## [1] 0.04308197
```

$\Rightarrow p = 0.04308197$

3 Roulette

Um die Gleichverteilung der 37 möglichen Ausgänge beim Rouletterad zu überprüfen, kann ein Chi-Quadrat-Anpassungstest durchgeführt werden.

$H_0 : P(X_i = k) = p_k$ für $k = 1, \cdots, 37$
$H_1 : P(X_i = k) \neq p_k$ für min. ein k

Sei dann die Verteilung der Teststatistik mit $a, b \geq 30$

$$\chi^2 = \sum_{k=1}^{K} \frac{(h_k - n \cdot p_k)^2}{n \cdot p_k}$$

```
Roulette <- read.csv("RouletteDaten.csv")$x

absFrequency <- table(Roulette)

n <- length(Roulette)
k <- length(absFrequency)

chisq <- sum((table(Roulette)-(n/37))^2 / (n/37));chisq

## [1] 48.506
```

Der Ablehnungsbereich $C = (\chi^2_{1-\alpha,K-1}, \infty)$ kann folgendermaßen in R berechnet werden:

```
qchisq(1-0.05,k-1)

## [1] 50.99846
```

$\Rightarrow C = (50.99846, \infty)$
Somit ist $\chi^2 = 48.506 \notin C$, sodass H_0 beibehalten wird. Es gibt also keine statistische Evidenz dafür, dass es bei einem Signifikanzniveau von $\alpha = 0.05$ eine Ungleichverteilung der 37 verschiedenen Ausgänge des Rouletterades gibt.

Zuletzt kann wird noch der p-Wert bestimmt:

```
1-pchisq(chisq,k-1) #p wert

## [1] 0.07960126
```

$\Rightarrow p = 0.07960126$

4 Einstiegsgehälter

Um die Erwartungen hinsichtlich des Einstiegsgehalts zwischen weiblichen und männlichen Studierenden zu vergleichen, kann ein approximativenr Zwei-Stichproben-Gaußtest (Z-Test) durchgeführt werden.

$H_0 : \mu_W \leq \mu_M$ besagt, dass das Einstiegsgehalt der Männer höher ist.
$H_1 : \mu_W > \mu_M$ besagt, dass das Einstiegsgehalt der Frauen höher ist.

Sei dann die Verteilung der Teststatistik mit $a, b \geq 30$

$$Z = \frac{\bar{M}_a - \bar{W}_b}{\sqrt{\sigma_M^2/a + \sigma_W^2/b}}$$

```
dat <- read.csv("Gehaelter.csv")

mG <- dat$Gehalt[dat$Geschlecht == "maennlich"]
wG <- dat$Gehalt[dat$Geschlecht == "weiblich"]

n_mG <- length(mG)
n_wG <- length(wG)

mean_mG <- mean(mG)
mean_wG <- mean(wG)

var_mG <- sd(mG)
var_wG <- sd(wG)

z <- (mean_mG-mean_wG) / sqrt((var_mG^2/n_mG)+(var_wG^2/n_wG));z

## [1] 1.158768
```

$\Rightarrow Z = 1.158768$

Der Ablehnungsbereich $C = (z_{1-\alpha}, \infty)$ kann folgendermaßen in R berechnet werden:

```
qnorm(1-0.01)

## [1] 2.326348
```

$\Rightarrow C = (2.326348, \infty)$
Somit ist $Z = 1.158968 \notin C$, sodass H_0 beibehalten wird. Es gibt also eine statistische Signifikanz dafür, dass Männer ein höheres Einstiegsgehalt als Frauen bekommen.

Zuletzt kann wird noch der p-Wert bestimmt:

```
1-pnorm(z)

## [1] 0.1232754
```

$\Rightarrow p = 0.1232754$