Zweite Hausarbeit in Statistik für Wirtschaftsinformatiker

HTW Berlin, Sommersemester 2017

Name, Matrikelnummer: Jenny Rothe, 544179
Name, Matrikelnummer: Laura Laugwitz, 544049

Formalitäten

Bitte bearbeiten Sie diese Hausarbeit in Zweiergruppen, in denen beide Studierende bei Herrn Spott oder beide bei Herrn Heimann eingeschrieben sind. Gruppen von drei oder mehr Studierenden sind nicht zugelassen. Setzen Sie bitte Ihre beiden Namen und Matrikelnummern oben ein.

Öffnen Sie das Dokument `titanic.Rmd` in RStudio, lösen Sie alle Aufgaben mit R und **fügen Sie alle Antworten zu diesem R-Markdown-Dokument hinzu, einschließlich des R-Codes**, wie Sie es bereits bei den Übungsblättern getan haben. Zusätzliche handgeschriebene Lösungen oder Erklärungen sind nicht zugelassen ebensowenig wie Lösungen, die mit anderer Software wie z.B. Microsoft Excel erstellt wurden.

Mehr Informationen über R-Markdown-Dokumente finden Sie im Internet unter http://rmarkdown.rstudio.com/lesson-1.html. Sie können die Musterlösungen in RMarkdown zu unseren Übungsblättern als Beispiele heranziehen.

Wenn Sie mit der Bearbeitung fertig sind, erzeugen Sie bitte in RStudio mit dem Knopf **Knit PDF** ein PDF-Dokument oder wählen alternativ über das Dreieck neben **Knit PDF** die Option **Knit HTML**, um ein HTML-Dokument zu erzeugen. Für **Knit PDF** ist die Installation einer LaTeX-Distribution wie MikTeX für Windows (miktex.org) oder MacTeX für Mac OS X (www.tug.org/mactex/) erforderlich. **Knit HTML** funktioniert auch ohne LaTeX. **Drucken Sie das so erzeugte Dokument aus und geben Sie es in Papierform ab**.

Abgabe

- **Elektronisch in Moodle bis spätestens Montag, 10.07.2017 um 16:00**
  - sowohl das RMarkdown-Quelldokument
  - als auch das daraus erzeugte PDF- oder HTML-Dokument

- **UND in Papierform spätestens am 10.07. bzw. 11.07.** bei Herrn Heimann bzw. Herrn Spott in den Übungen oder der Vorlesung. Spätere Abgaben werden nicht berücksichtigt und führen automatisch zu einer Bewertung mit 0 Punkten.

Die Ergebnisse aller Hausarbeiten werden zusammen mit 30% gewichtet, das Ergebnis der Klausur mit 70%.

Wichtig: Sobald Sie eine Hausarbeit abgeben, hat damit Ihre Prüfungsleistung für das Sommersemester 2017 begonnen, die mit der Klausur abgeschlossen wird. Wenn Sie eine Hausarbeit abgeben, aber die Klausur nicht im Sommersemester 2017 mitschreiben, sind Sie automatisch durchgefallen und die Punkte der Hausarbeiten verfallen. Diese Regelung ist in der Prüfungsordnung festgelegt.

Viel Erfolg,

Ihre Dozenten Martin Spott und Michael Heimann

Stand 10.07.2017

Aufgaben

Wie in der ersten Hausarbeit analysieren wir die Passagierdaten `titanic_data.csv` des Kreuzfahrtschiffes Titanic. Wir lesen die Daten ein und fügen die Spalte `Survived2` hinzu, die die Werte des Merkmals `Survived` anstatt 0/1 mit *no/yes* kodiert.

```
# setzen Sie hier den richtigen Pfad für titanic_data.csv ein
titanic_data <- read.csv("titanic_data.csv")
titanic_data$Survived2 <- factor(titanic_data$Survived,
                                 levels = c(0,1),
                                 labels = c("no", "yes"))
```

Aufgabe 1 (15 Punkte)

Wir untersuchen, ob sich die Überlebenschancen weiblicher und männlicher Passagiere unterscheiden.

a) Stellen Sie die Kontingenztabelle der absoluten Häufigkeiten von `Survived2` und `Sex` auf!

```
library(knitr)
h_surv_gender <- table(titanic_data$Survived2, titanic_data$Sex)
kable(addmargins(h_surv_gender))
```

	female	male	Sum
no	81	468	549
yes	233	109	342
Sum	314	577	891

b) Stellen Sie die Kontingenztabelle der relativen Häufigkeiten von `Survived2` und `Sex` in Prozent auf!

```
f_surv_gender <- prop.table(h_surv_gender)
kable(round(addmargins(f_surv_gender)*100, digits = 2))
```

	female	male	Sum
no	9.09	52.53	61.62
yes	26.15	12.23	38.38
Sum	35.24	64.76	100.00

c) Berechnen Sie folgende bedingte relative Häufigkeiten in Prozent:

i) f(Survived2 = yes | Sex = female)
ii) f(Survived2 = yes | Sex = male)
iii) f(Sex = female | Survived2 = yes)
iv) f(Sex = male | Survived2 = yes)

```
# f(Survived2 = yes | Sex = female)
i1 <- round((0.2615 / 0.3524)*100, 2)

# f(Survived2 = yes | Sex = male)
i2 <- round((0.1223 / 0.6476)*100, 2)

# f(Sex = female | Survived2 = yes)
i3 <- round((0.2615 / 0.3838)*100, 2)

# f(Sex = male | Survived2 = yes)
```

```
i4 <- round((0.1223 / 0.3838)*100, 2)
```

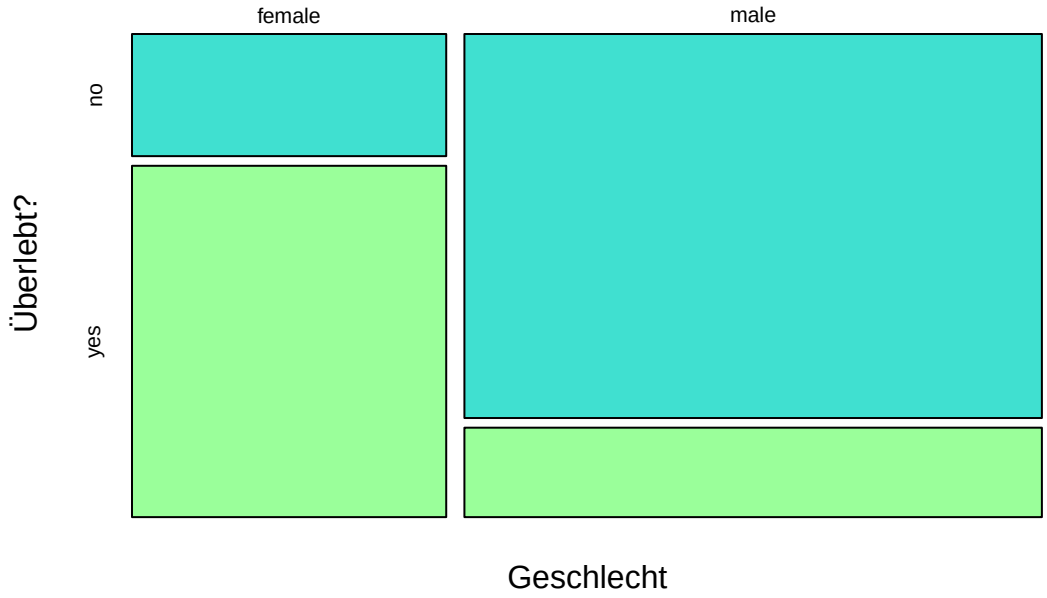
Unter den Frauen haben 74.21 % überlebt. Unter den Männern haben 18.89 % überlebt. Unter den Überlebenden waren 68.13 % Frauen und 31.87 % Männer.

d) Welche der relativen Häufigkeiten in c) beschreiben die Überlebenschance von Frauen bzw. von Männern?

Die relativen Häufigkeiten über die Spalten geben uns hier die korrekte Auskunft. Frauen haben zu 74.21% überlebt und Männer zu 18.89%.

e) Generieren Sie einen Mosaikplot der Merkmale `Survived2` und `Sex`. Spalten Sie dabei erst nach `Sex` auf, dann nach `Survived2`. Beschriften Sie die Achsen und vergeben Sie einen sinnvollen Titel!

```
mosaicplot(Sex ~ Survived2, data = titanic_data,
           main = "Überlebende auf der Titanic - Mosaikplot",
           ylab="Überlebt?", xlab="Geschlecht",
           col = c("turquoise", "palegreen1"))
```



f) Berechnen Sie den Phi-Koeffizienten von `Survived2` und `Sex`! Was misst dieser? Was sagt uns in diesem Fall der Wert von Phi?

```
library(DescTools)
p <- Phi(titanic_data$Survived2, titanic_data$Sex)
```

Phi gibt an, wie stark die statistische Abhängigkeit zwischen zwei Merkmalen ist. Phi kann zwischen 0 und 1 (starke Abhängigkeit) liegen. Mit 0.5433514 liegt Phi so gerade noch im Bereich des mittelstarken Zusammenhangs.

g) Sind die Merkmale `Survived2` und `Sex` statistisch unabhängig voneinander? Begründen Sie Ihre Antwort!

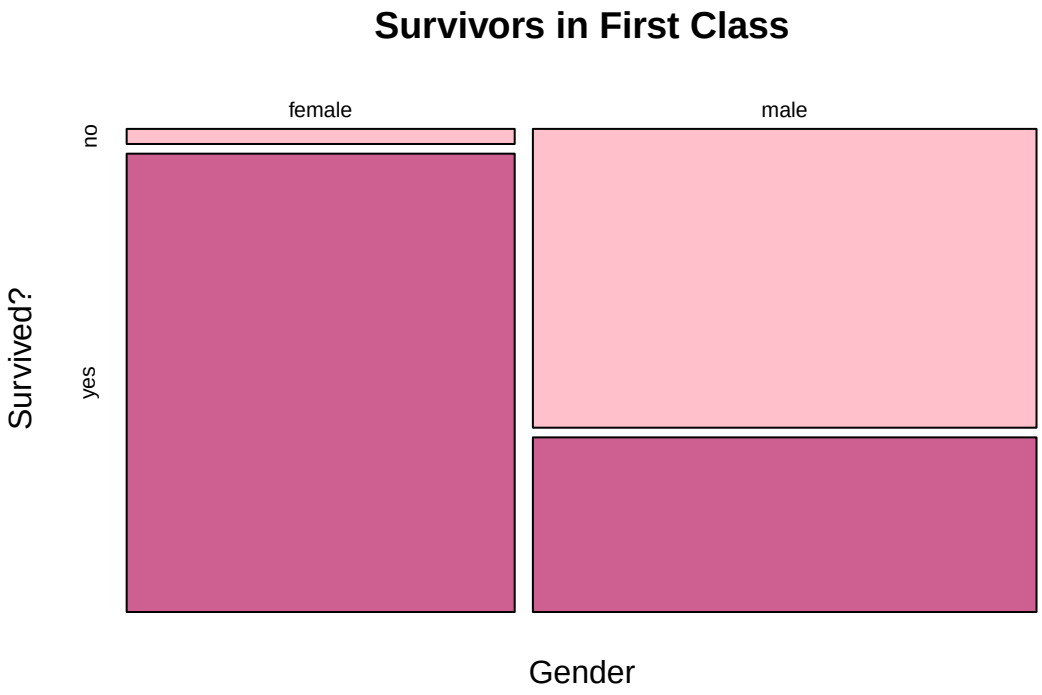
Die Merkmale sind nicht statistisch unabhängig voneinander. Wir können das sowohl an dem Mosaikplot erkennen (es befinden sich hier keine geraden Schnitte) als auch an dem Phi-Koeffizienten (er liegt im Bereich des mittelstarken Zusammenhangs).

Aufgabe 2 (9 Punkte)

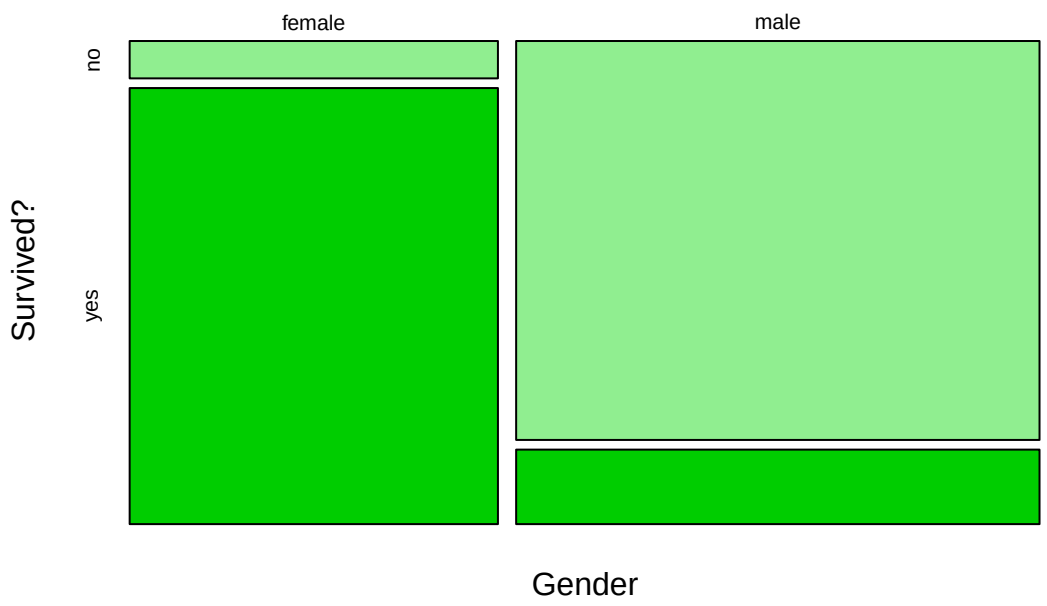
Wir untersuchen, ob die Überlebenschancen männlicher und weiblicher Passagiere in allen drei Passagierklassen `Pclass` in einem ähnlichen Verhältnis stehen.

a) Generieren Sie Mosaikplots der Merkmale `Survived2` und `Sex` wie in Aufgabe 1e), jeweils einen für die Passagiere der Klasse 1, 2 und 3! Beschriften Sie die Achsen und vergeben Sie einen sinnvollen Titel!

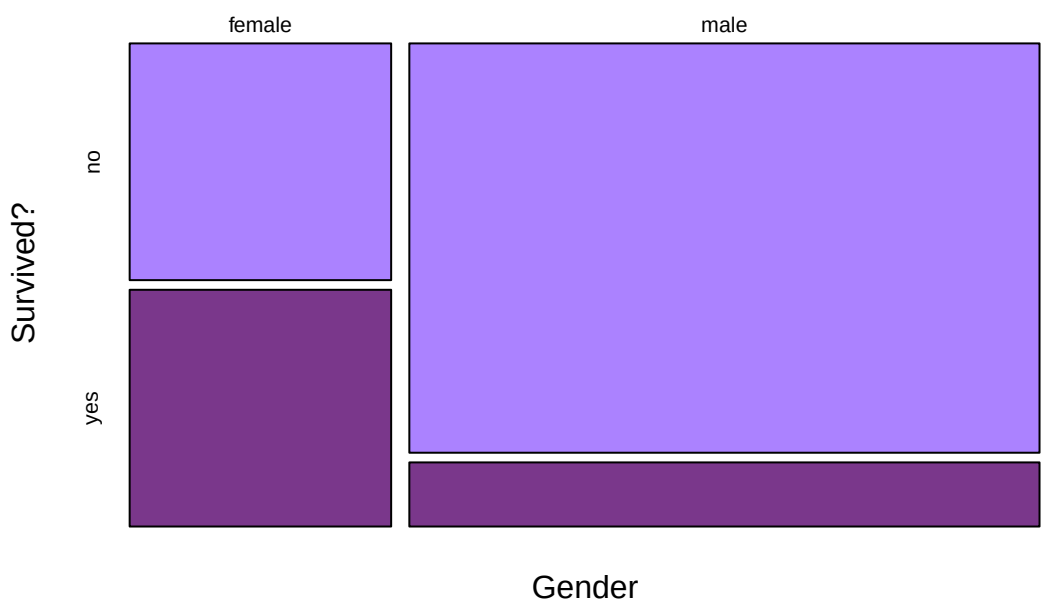
```
klasse1 <- subset(titanic_data, Pclass == "1")
klasse2 <- subset(titanic_data, Pclass == "2")
klasse3 <- subset(titanic_data, Pclass == "3")

tab_1 <- table(klasse1$Sex, klasse1$Survived2)
tab_2 <- table(klasse2$Sex, klasse2$Survived2)
tab_3 <- table(klasse3$Sex, klasse3$Survived2)

mosaicplot(tab_1, xlab = "Gender", ylab = "Survived?",
           main = "Survivors in First Class", col = c("pink", "hotpink3"))
```




```
mosaicplot(tab_2, xlab = "Gender", ylab = "Survived?",
           main = "Survivors in Second Class", col = c("lightgreen", "green3"))
```




```
mosaicplot(tab_3, xlab = "Gender", ylab = "Survived?",
           main = "Survivors in Third Class", col = c("mediumpurple1", "mediumorchid4"))
```



b) Berechnen Sie den Phi-Koeffizienten von `Survived2` und `Sex` jeweils für Klasse 1, 2 und 3!

```
p1 <- Phi(klasse1$Sex, klasse1$Survived2)
p2 <- Phi(klasse2$Sex, klasse2$Survived2)
p3 <- Phi(klasse3$Sex, klasse3$Survived2)

p1
```

```
## [1] 0.6152121
```

```
p2
```

```
## [1] 0.7531211
```

```
p3
```

```
## [1] 0.387313
```

c) Was bedeutet es, dass der Phi-Koeffizient für Klasse 2 größer ist als für die anderen beiden Klassen?

In der zweiten Klasse war das Geschlecht der Reisenden am Relevantesten für deren Überleben (im Vergleich zu den anderen Klassen). Es besteht hier nämlich ein starker Zusammenhang (0.7531211) zwischen Geschlecht und Überleben. Bei der ersten Klasse ist der Zusammenhang zwischen Geschlecht und Überleben noch mäßig relevant (0.6152121). In der dritten Klasse besteht zwischen den beiden Merkmalen ein eher schwacher Zusammenhang (0.387313).

Aufgabe 3 (12 Punkte)

Wir untersuchen die Überlebenschancen in verschiedenen Altergruppen.

a) Gruppieren Sie `Age` wie in Aufgabe 3 der ersten Hausarbeit in die Intervalle [0,5), [5,10), [10, 15) usw! Beachten Sie wieder, zu welcher Gruppe die Grenzwerte der Intervalle gehören. Hinweis: Sehen Sie sich die Funktion `cut()` an, deren Syntax der von `hist()` ähnelt.

Fügen Sie die Altersgruppe als Spalte `AgeGroup` dem Data-Frame `titanic_data` zu.

```
# Gruppierung von Age
bins <- c(seq(0,85,5))
age_group <- table(cut(titanic_data$Age, breaks = bins,
                       include.lowest = TRUE, right = FALSE))
age_group
```

```
##
##   [0,5)  [5,10) [10,15) [15,20) [20,25) [25,30) [30,35) [35,40) [40,45)
##      40      22      16      86     114     106      95      72      48
## [45,50) [50,55) [55,60) [60,65) [65,70) [70,75) [75,80) [80,85]
##      41      32      16      15       4       6       0       1
```

```
# Anlegen und Befüllen der Spalte 'AgeGroup'
# Das wäre auch einfacher gegangen, ist mir aber erst später aufgefallen - egal :)

library(TeachingDemos)
titanic_data[, "AgeGroup"] <- 0
for(i in 1:891){
  if(is.na(titanic_data[i, 6])){
    titanic_data[i, 14] <- NA
  }
  else if(0.0 %<=% titanic_data[i, 6] %<%5.0){
    titanic_data[i, 14] <- "[0,5)"
  }
  else if (5.0 %<=% titanic_data[i, 6] %<%10.0){
    titanic_data[i, 14] <- "[05,10)"
  }
  else if (10.0 %<=% titanic_data[i, 6] %<%15.0){
    titanic_data[i, 14] <- "[10,15)"
  }
```

```
  else if (15.0 %<=% titanic_data[i, 6] %<%20.0){
    titanic_data[i, 14] <- "[15,20)"
  }
  else if (20.0 %<=% titanic_data[i, 6] %<%25.0){
    titanic_data[i, 14] <- "[20,25)"
  }
  else if (25.0 %<=% titanic_data[i, 6] %<%30.0){
    titanic_data[i, 14] <- "[25,30)"
  }
  else if (30.0 %<=% titanic_data[i, 6] %<%35.0){
    titanic_data[i, 14] <- "[30,35)"
  }
  else if (35.0 %<=% titanic_data[i, 6] %<%40.0){
    titanic_data[i, 14] <- "[35,40)"
  }
  else if (40.0 %<=% titanic_data[i, 6] %<%45.0){
    titanic_data[i, 14] <- "[40,45)"
  }
  else if (45.0 %<=% titanic_data[i, 6] %<%50.0){
    titanic_data[i, 14] <- "[45,50)"
  }
  else if (50.0 %<=% titanic_data[i, 6] %<%55.0){
    titanic_data[i, 14] <- "[50,55)"
  }
  else if (55.0 %<=% titanic_data[i, 6] %<%60.0){
    titanic_data[i, 14] <- "[55,60)"
  }
  else if (60.0 %<=% titanic_data[i, 6] %<%65.0){
    titanic_data[i, 14] <- "[60,65)"
  }
  else if (65.0 %<=% titanic_data[i, 6] %<%70.0){
    titanic_data[i, 14] <- "[65,70)"
  }
  else if (70.0 %<=% titanic_data[i, 6] %<%75.0){
    titanic_data[i, 14] <- "[70,75)"
  }
  else if (75.0 %<=% titanic_data[i, 6] %<%80.0){
    titanic_data[i, 14] <- "[75,80)"
  }
  else if (80.0 %<=% titanic_data[i, 6] %<%85.0){
    titanic_data[i, 14] <- "[80,85)"
  }
}
```

b) Berechnen Sie für jede Altersgruppe [0,5), [5,10) usw. die Überlebenschance (relative Häufigkeit zu überleben) in Prozent!

```
# Erstelle Kontingenztabelle / absolute Häufigkeiten
h_age_survival <- table(titanic_data$AgeGroup, titanic_data$Survived2)
# leider geht hier [75, 80) verloren (weil 0? oder wegen den Levels?)
kable(addmargins(h_age_survival))
```

	no	yes	Sum
[0,5)	13	27	40

	no	yes	Sum
[05,10)	11	11	22
[10,15)	9	7	16
[15,20)	52	34	86
[20,25)	75	39	114
[25,30)	68	38	106
[30,35)	55	40	95
[35,40)	39	33	72
[40,45)	30	18	48
[45,50)	25	16	41
[50,55)	18	14	32
[55,60)	10	6	16
[60,65)	9	6	15
[65,70)	4	0	4
[70,75)	6	0	6
[80,85)	0	1	1
Sum	424	290	714

```
# Ermittle gemeinsame relative Häufigkeiten
f_age_survival <- prop.table(h_age_survival)
f_age_survival_margins <- addmargins(f_age_survival)
kable(round(f_age_survival_margins*100, 2))
```

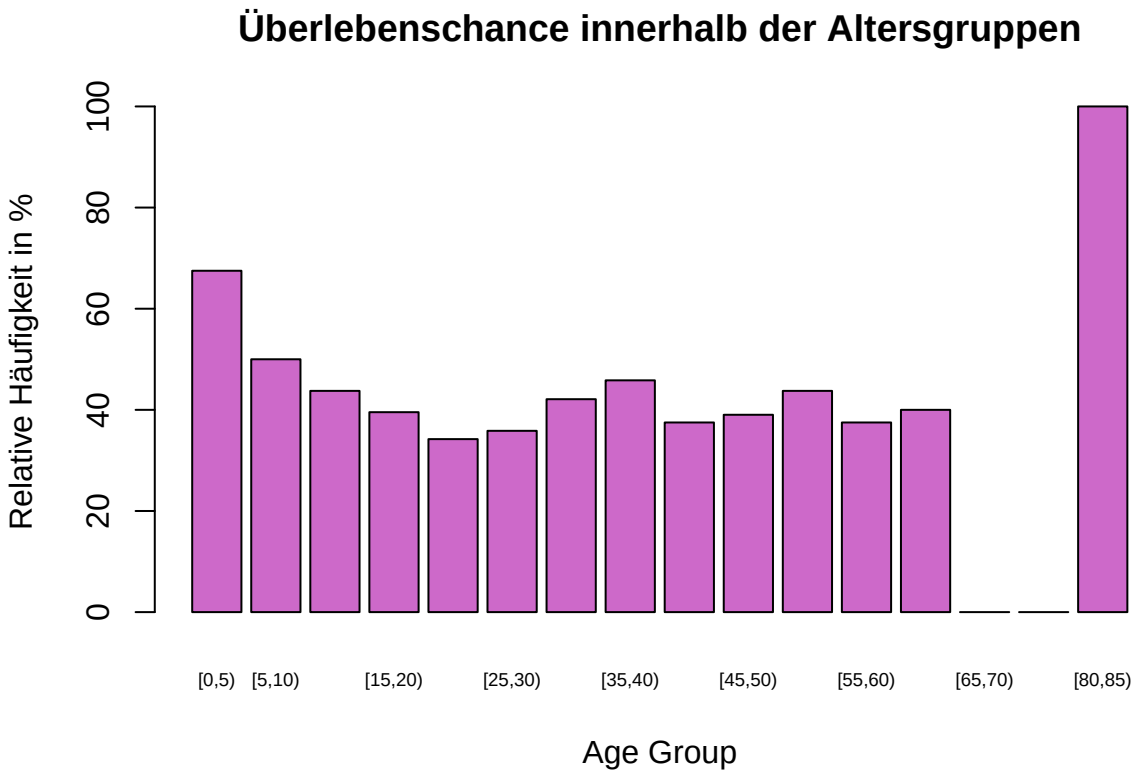
	no	yes	Sum
[0,5)	1.82	3.78	5.60
[05,10)	1.54	1.54	3.08
[10,15)	1.26	0.98	2.24
[15,20)	7.28	4.76	12.04
[20,25)	10.50	5.46	15.97
[25,30)	9.52	5.32	14.85
[30,35)	7.70	5.60	13.31
[35,40)	5.46	4.62	10.08
[40,45)	4.20	2.52	6.72
[45,50)	3.50	2.24	5.74
[50,55)	2.52	1.96	4.48
[55,60)	1.40	0.84	2.24
[60,65)	1.26	0.84	2.10
[65,70)	0.56	0.00	0.56
[70,75)	0.84	0.00	0.84
[80,85)	0.00	0.14	0.14
Sum	59.38	40.62	100.00

```
# Ermittle Überlebenschance (relative Häufigkeit über Zeile)
foo <- f_age_survival_margins
for(i in 1:17){
  foo[i,2] <- (f_age_survival_margins[i,2] / f_age_survival_margins[i,3]) * 100
}
foo <- foo[,2]
ueberlebenschancen <- data.frame(foo)
kable(ueberlebenschancen)
```

	foo
[0,5)	67.50000
[05,10)	50.00000
[10,15)	43.75000
[15,20)	39.53488
[20,25)	34.21053
[25,30)	35.84906
[30,35)	42.10526
[35,40)	45.83333
[40,45)	37.50000
[45,50)	39.02439
[50,55)	43.75000
[55,60)	37.50000
[60,65)	40.00000
[65,70)	0.00000
[70,75)	0.00000
[80,85)	100.00000
Sum	40.61625

c) Stellen Sie das Ergebnis aus b) als Balkendiagramm da – für jede Altersgruppe einen Balken! Beschriften Sie alle Axen und geben dem Diagramm einen sinnvollen Titel!

```r
dims <- c("[0,5)", "[5,10)", "[10,15)", "[15,20)", "[20,25)",
          "[25,30)", "[30,35)", "[35,40)", "[40,45)", "[45,50)",
          "[50,55)", "[55,60)", "[60,65)", "[65,70)", "[70,75)", "[80,85)")

# f( survived | $age_group )
barplot(ueberlebenschancen$foo[1:16], names.arg = dims,
        main = "Überlebenschance innerhalb der Altersgruppen",
        xlab = "Age Group", ylab = "Relative Häufigkeit in %", cex.names = 0.6,
        col = "orchid3")
```



d) Welche zwei Altersgruppen hatten die größte Überlebenschance? Sind beide Altersgruppen statistisch gesehen gleich bedeutend?

Innerhalb der Altergruppen [0,5) und [80, 85) war die Chance, zu überleben, am Höchsten. Bei letzterer lag diese sogar bei 100%. Allerdings besteht hier ein deutlicher Unterschied, weil in der Gruppe [80, 85) nur eine Person existierte - und die hat dann überlebt. In der Gruppe [0,5) haben über 67% überlebt. Allerdings bestand diese Gruppe aus 40 Personen. Dieser Unterschied zeigt sich auch in der gemeinsamen relativen Häufigkeit: Während im Bezug auf die Gesamtzahl aller Personen die Überlebenschance der Gruppe [0,5) bei 3.78% lag, ist sie bei der Gruppe [80, 85) bei 0.14%.