Hilfsdokument für Statistik I - SoSe 17

Jakob Schwerter, M. Sc.

22 Juni, 2018

Contents

Einleitung

Sollten die Erklärungen hier nicht verständlich sein, bitte ich Sie, Fragen im Forum zu stellen. Gerne dürfen dann auch anderen Studierende antworten!

Das Dokument wird fortlaufend aktualisiert. Sollten im Forum Fragen bezüglich den Tipps gestellt werden, pflege ich die Antworten in das Dokument ein. Hinweise bezüglich Fehlern jeglicher Art bitte direkt per E-Mail an jakob.schwerter@uni-tuebingen.de.

Ab dem 28.08 wird immer das Datum angegeben, an dem es das letzte Update gab. Hierdurch soll eine neue Version verdeutlicht werden. Neu sind hier unteranderem die Themen

- (Es wurden nur ein paar Typos korrigiert)
- Die Grafik vom Polygonzug wurde aktualisiert

Das Hilfdokument wurde in *R Notebook* geschrieben. Hierbei können Rechnungen und Text zusammen in einem Dokument erstellt werden. Hierdurch sollte die Wahrscheinlichkeit von Rechenfehler stark reduziert sein, da die meisten Sachen nicht mehr per Hand berechnet werden. Die grau hinterlegten Felder zeigen den *R*-Code an, mit dem gearbeitet wurde. Für die Vorlesung oder die Klausur ist dies nicht weiter von Bedeutung. Interessierte Studierende können sich aber gerne das Programm *R* angucken und die Codes dabei verwenden. Es kann dabei helfen den Stoff besser nachvollziehen zu können.

Gründe für (und gegen) *R* können unter den folgenden Links gefunden werden:

http://r4stats.com/articles/popularity/

http://www.datacamp.com/community/tutorials/statistical-language-wars-the-infograph

http://sharpsightlabs.com/blog/r-recommend-data-science/

https://www.inwt-statistics.de/blog-artikel-lesen/Statistik-Software-R_SAS_SPSS_STATA_im_Vergleich.html

http://www.statistik-und-beratung.de/2015/02/spss-oder-r/

Allgemein

Tutoriumsanrechnung

Das Tutorium wird in Campur erst als bestanden angezeigt, sobald auch die Vorlesung durch die Klausur bestanden wurde. Grund hierfür ist, dass die Note der Klausur auch die Note für die *bung ist und dadurch nicht vor dem Schreiben der Klausur gegeben werden kann.

Vorbereitung für die Klausur

Die angeführten Tipps sind von Studierenden gegeben:

Klausur ernst nehmen und nicht nur nebenbei dafür lernen.

Übungsblatt vorbereiten, bei der Übung mitschreiben, Fehler deutlich markieren und hinterher nochmal durchrechnen. Übungsblätter vor der Klausur nochmal alle durchrechnen (und somit mindestens 4-mal durchgerechnet haben). Das bedeutet, man sollte die Aufgaben so oft durchrechnen, bis man sie auswendig kann. Man sollte also die Aufgaben in Gedanken klar lösen können, ohne die Lösungen aufschreiben zu müssen.

Übungsblätter von Anfang an mithilfe der Formelsammlung berechnen, damit man weiß wo die Formeln stehen und worauf man zurückgreifen kann.

Während des Lernens eine Zusammenfassung schreiben, in der auch Lösungen und Verbesserungen bezüglich eigener Fehler deutlich gemacht werden. Man sollte sich bei Fehlern klar aufschreiben, wieso ein Fehler entstanden ist und wie man ihn verbessert. Merksätze in die Zusammenfassung aufnehmen. Zusammenfassungen haben den größten positiven Effekt wenn sie selbstständig (oder in einer Gruppe) erstellt worden sind.

Für die Zusammenfassung sollten auch immer die jeweiligen Abschnitte des Hilfsdokuments gelesen werden, um möglich Unklarheiten zu lösen, eine unterschiedliche Erklärung zu kriegen oder Merksätze bezügliche den häufig vorkommenden Fehlern zu schreiben.

Eine Wochen vor der Klausur sollten alle Übungen mindestens 4-mal durchgerechnet worden sein und die Zusammenfassung fertig sein. In der letzten Woche wird das Wissen dann durch Wiederholung verfestigt und Verbindungen zwischen den Themen gefunden. Mögliche Unklarheiten mit Tutoren*innen oder dem Dozenten klären. Hierbei kann man die Zusammenfassung nochmal kürzen und das Hilfsdokument aufmerksam lesen. Man sollte aber die Option unterschätzen frühzeitig die Tutoren*innen oder den Dozenten (in der Sprechstunde oder während der Vorlesung) Fragen zu stellen.

Vorlesungsfolien einmal durchgehen um einen Überblick zu kriegen und dann nochmal um gezielt die Theorie zu erlernen. Altklausuren (auf Zeit) rechnen, damit man ein Gefühl für die Klausur kriegt. Die Aufgaben sind nicht allzu unterschiedlich und somit eine gute Übung.

Altklausuren, Übungsblätter UND Kontrollfragen sind eine gute Vorbereitung für die Klausur. Anwesenheit in der Vorlesung hilft vor allem bei den Kontrollfragen und dem Verständnis des Stoffes. Kontrollfragen vor der Vorlesung vorbereiten hilft schon von Anfang an aufmerksam zu lernen.

Klarmachen, was die Formeln bedeuten, wie sie anzuwenden sind und dass viele Rechnungen das Sture anwenden der Formeln sind.

Summenzeichen $\sum$

Das Summenzeichen $\sum$ wird verwendet, um Formeln abzukürzen.

Gegeben sei folgende Tabelle:

```
i <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16)
x_i <- c(1, 1, 2, 2, 2, 3, 3, 6, 6, 6, 6, 11, 11, 11, 16, 16)
dat <- data.frame(i, x_i)
knitr::kable(dat, col.names = c("$i$", "$x_{i}$"), bootstrap_options = c("striped",
    "hover", "condensed", "responsive"), full_width = F)
```

i	x_i
1	1
2	1
3	2
4	2
5	2
6	3
7	3
8	6
9	6
10	6
11	6
12	11
13	11
14	11
15	16
16	16

Die Summe des Merkmals X ist

$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} = 1 + 1 + 2 + 2 + 2 + 3 + 3 + 6 + 6 + 6 + 6 + 11 + 11 + 11 + 16 + 16 = 103$

Da dies Zeitaufwendig ist, nutzt man das SUmmenzeichen um die Schreibweise zu verkürzen.

$\sum_{i=1}^{16} x_i = 103$

Hier werden die einzelnen x_i's aufsummiert. Der untere Index zeigt an, wo man anfangen muss, bei $i = 1$. Der obere Index verdeutlicht, bei welcher Indexnummerierung man aufhören muss, bei $i = 16$. Generell wird eine Summer in der Vorlesung immer bei 1 anfangen, und bei n oder k aufhören. In dem Beispiel hier ist $n = 16$. Ein Beispiel mit k folgt.

Die Tabelle kann vereinfacht werden, da mehrere Zahlen mehrmals vorkommen:

```
i <- c(1, 2, 3, 4, 5, 6)
x_i <- c(1, 2, 3, 6, 11, 16)
n_i <- c(2, 3, 2, 4, 3, 2)
dat <- data.frame(i, x_i, n_i)
knitr::kable(dat, col.names = c("$i$", "$x_{i}$", "$n_{i}$"),
    bootstrap_options = c("striped", "hover", "condensed", "responsive"),
    full_width = F)
```

i	x_i	n_i
1	1	2
2	2	3
3	3	2
4	6	4
5	11	3

i	x_i	n_i
6	16	2

Somit kann man auch die oben angegebene Summe vereinfachen:

$\sum_{i=1}^{\mathbf{n}} x_i = \sum_{i=1}^{\mathbf{k}} n_i \cdot x_i = 3 \cdot 2 + 2 \cdot 3 + 4 \cdot 6 + 3 \cdot 11 + 2 \cdot 16 = 103$

Durch Vereinfachung der Tabelle wird hier k eingeführt. Die Gesamtmenge der Merkmalsausprägungen ist immer noch $n = 16$. Allerdings gibt es es, wenn man alle Merkmalsausprägungen die mehrmals vorkommen, gruppiert, nur noch $k = 6$ unterschiedliche Merkmalsausprägungen. Somit wird die Summe nur noch von 1 bis 6 durchgeführt. Hierfür muss aber das jeweilige x_i mit der jeweiligen absoluten Häufigkeit n_i multipliziert werden, damit das selbe Ergebnis herauskommt.

Zur Verdeutlichung der Summen noch folgende zufälligen Beispielreichnungen:

```
i <- c(1, 2, 3, 4, 5, 6)
x_i <- c(1, 12, 3, 6, 5, 16)
y_i <- c(12, 13, 2, 4, 30, 20)
z_i <- c(10, 20, 30, 40, 50, 60)
w_i <- c(11, 12, 13, 14, 15, 16)
v_i <- c(50, 45, 30, 22, 11, 5)
dat <- data.frame(i, x_i, y_i, z_i, w_i, v_i)
knitr::kable(dat, col.names = c("$i$", "$x_{i}$", "$y_{i}$",
    "$z_{i}$", "$w_{i}$", "$v_{i}$"), bootstrap_options = c("striped",
    "hover", "condensed", "responsive"), full_width = F)
```

i	x_i	y_i	z_i	w_i	v_i
1	1	12	10	11	50
2	12	13	20	12	45
3	3	2	30	13	30
4	6	4	40	14	22
5	5	30	50	15	11
6	16	20	60	16	5

$\sum_{i=1}^{6} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 1 + 12 + 3 + 6 + 5 + 16 = 43$

$\sum_{i=1}^{6} y_i = y_1 + y_2 + y_3 + y_4 + y_5 + y_6 = 12 + 13 + 2 + 4 + 30 + 20 = 81$

$\sum_{i=1}^{6} z_i = z_1 + z_2 + z_3 + z_4 + z_5 + z_6 = 10 + 20 + 30 + 40 + 50 + 60 = 210$

$\sum_{i=1}^{6} w_i = w_1 + w_2 + w_3 + w_4 + w_5 + w_6 = 11 + 12 + 13 + 14 + 15 + 16 = 81$

$\sum_{i=1}^{6} v_i = v_1 + v_2 + v_3 + v_4 + v_5 + v_6 = 50 + 45 + 30 + 22 + 11 + 5 = 163$

$\sum_{i=3}^{4} z_i = z_3 + z_4 = 30 + 40 = 70$

$\sum_{i=1}^{3} w_i = w_1 + w_2 + w_3 = 11 + 12 + 13 = 36$

$\sum_{i=5}^{6} v_i = v_5 + v_6 = 11 + 5 = 16$

$\sum_{i=1}^{6} x_i \cdot y_i = x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3 + x_4 \cdot y_4 + x_5 \cdot y_5 + x_6 \cdot y_6 = 12 + 156 + 6 + 24 + 150 + 320 = 668$

$\sum_{i=1}^{6} x_i - y_i = x_1 - y_1 + x_2 - y_2 + x_3 - y_3 + x_4 - y_4 + x_5 - y_5 + x_6 - y_6 = -11 + -1 + 1 + 2 + -25 + -4 = -38$

Zur Übung:

$\sum_{i=1}^{3} x_i \cdot w_i = 194$

$\sum_{i=1}^{3} x_i \cdot 5 = 80$

$\sum_{i=1}^{3} x_i - 2 = 10$

$\sum_{i=2}^{5} x_i + v_i = 342$

$\sum_{i=1}^{3} x_i + \sum_{i=1}^{3} v_i = 342$

$\sum_{i=2}^{5} x_i + v_{i-1} = 316$

Prozentzahlen

Tipp: Lieber keine Prozentzahlen, sondern Dezimalzahlen in die Tabellen aufnehmen und damit rechnen, damit es nicht zu Problem kommt. Prozentzahlen muss man nur am Ende haben, wenn explizit danach gefragt wird.

Antwortsätze

Tipp: Antwortsätze müssen nur gegeben werden, wenn explizit danach gefragt wird. Sollte in der Aufgabe *Brechnen Sie ...* stehen, muss nur die Rechnung aufgestellt werden. Steht ein *Begründen Sie kurz* oder ein *Antworten Sie kurz*, sollte ein kurzer Antwortsatz gegeben werden.

Berechnungen in der Klausur

Um die volle Punktzahl in den Aufgaben kriegen zu können, müssen sie neben dem Ergebniss auch Rechenschritte aufschreiben. Hier sollte mindestens immer die Formel in eingesetzer Form stehen. Also das man die jeweiligen Werte in die Formel schreibt, bevor etwas verrechnet wird. Ausnahme ist.

Sollte zum Beispiel eine zweidimensionale Verteilung gegeben sein und gefragt werden, bei wie vielen Personen $X = x_i$ und $Y = y_j$ zutrifft, sollte man n_{ij} mit den spezifischen i und j angegeben werden, bevor die Zahl für n_{ij} aufgeschrieben wird.

Runden / Brüche

Rechnen Sie immer mit Brüchen und nutzen Sie erst Dezimalzahlen wenn Sie ein Endergebnis haben. Runden Sie hierbei zur 4. Kommastelle. Sollten Sie früher runden entstehen Abweichungen und somit Fehler.

Rechnung nicht möglich

Gibt es z.B. bei einer statistischen Verteilung keinen Modus, sollte man dies schreiben. Ähnliches bei der Rechnung der Korrelation für nominale Merkmale. Auch hier sollten Sie klar angeben, dass die Rechnung nicht möglich ist, und den Grund dafür angeben.

Nur Berechnen, wonach gefragt ist

Sollte in der Aufgabenstellung nach der Varianz gefragt sein, muss nicht zusätzlich die Standardabweichung berechnet werden. Dies ist erst nötig, wenn explizit danach gefragt wird.

Weitere Referencen/Quellen?

http://www.crashkurs-statistik.de/ :
Gute Zusammenfassungen

https://de.khanacademy.org/
Gute Videos um die generellen Konzepte zuverstehen und weitere Zusatzaufgaben

http://www.mathods.com/mathe.php?con=Statistik

http://www.poissonverteilung.de/
Hilfreich für Statistik II

https://www.youtube.com/watch?v=kIZ9-mGbuN8&list=PLLTAHuUj-zHifw_3OhBTvQq2EGX5NedOy

https://www.youtube.com/watch?v=Iymj2-58Ilk

http://www.ted.com/talks/alan_smith_why_we_re_so_bad_at_statistics?utm_source=newsletter_daily&utm_campaign=daily&utm_medium=email&utm_content=button___2017-01-31#t-752755

Kapitel 2

Aufgaben bzgl. absoluter und relativer Häufigkeit

Am Anfang eine Tabelle mit allen Einträgen (x_i, n_i, h_i & H_i) mit den jeweiligen Summen aufstellen. Für jede Aufgabe eine neue Tabelle aufzustellen kostet unnötig Zeit. Die Berechnung der Summen der relativen Häufigkeit h_i ist hilfreich für die Überprüfung, ob die Rechnungen richtig sind.

```
i <- c(1, 2, 3, 4, 5, 6)
x_i <- c(1, 2, 3, 4, 5, 8)
n_i <- c(43, 42, 6, 2, 29, 6)
h_i <- round(n_i/sum(n_i), 4)
H_i <- cumsum(h_i)
dat <- data.frame(i, x_i, n_i, h_i, H_i)
knitr::kable(dat, col.names = c("$i$", "$x_{i}$", "$n_{i}$", "$h_{i}$", "$H(x_i)$"),
    bootstrap_options = c("striped", "hover", "condensed", "responsive"), full_width = F)
```

i	x_i	n_i	h_i	$H(x_i)$
1	1	43	0.3359	0.3359
2	2	42	0.3281	0.6640
3	3	6	0.0469	0.7109
4	4	2	0.0156	0.7265
5	5	29	0.2266	0.9531
6	8	6	0.0469	1.0000

i steht für *Arbeiter* (1), *Angestellter* (2), *Beamter* (3), *Zeit/Berufssoldat* (4), *Selbstständig* (5) und *Weiß ich nicht* (8). (Keine Merkmalsausprägungen für mithelfende Familienzugehörigkeit und freier Mitarbeiter hier). k ist hier gleich 6 (immer die letzte Zahl in der i-Spalte. Weiter gilt $n = \sum_{i=1}^{k} n_i$ 128, $h_i = n_i/n$ und $H(x_i) = \sum_{i|xi \leq x} h_i$. Die meisten Väter der Studierenden sind Arbeiter oder Angestellte und als nächstes Selbstständig.

Aus dieser Tabelle können dann alle weiteren Aufgaben beantwortet werden. Es sollte klar sein, dass wenn n_i (relativ) groß ist, dass dann auch h_i größer sein sollte und weiter auch der Zuwachs in $H(x_i)$. Ich empfehle in den Tabellen keine Prozentzahlen zu schreiben, damit man bei den Rechnung nicht versehentlich einen Fehler macht.

Berechnung von relativen Häufigkeiten:

Wenn die kumulierte relative Häufigkeit H_i für einen Wert zwischen zwei Grenzen berechnet werden soll, rechnet man mit den jeweiligen Grenzen:

$$\begin{aligned} H_{rel}(2 < x \leq 4) &= H_{rel}(x \leq 4) - H_{rel}(x < 2) = H_{rel}(x \leq 4) - H_{rel}(x \leq 1) \\ &= H(4) - H(1) = 0.7266 - 0.3359 = 0.3907 \end{aligned}$$

Der Vorteil hieran besteht dabei, dass man die Werte $H(4) = 0.7265$ und $H(1) = 0.3359$ aus der Tabelle ablesen kann und keine neue Rechnung aufstellen muss.

Ähnlich verhält es sich, wenn man eine relative Häufigkeit für einen Wert größer einer Grenze berechnen muss. Auch hier ist die Rechnung mit dem Gegensatz meistens schneller:

$$H_{rel}(x > 3) = 1 - H_{rel}(x \leq 3) = 1 - H(3) = 1 - 0.7109 = 0.2891$$
$$H_{rel}(x \geq 3) = 1 - H_{rel}(x < 3) = 1 - H_{rel}(x \leq 2) = 1 - H(2) = 1 - -0.664 = \text{`}1 - H_i[2]\text{`}$$

Beim letzten Beispiel sollte man sich genau die Symbole bezüglich der Ungleicheiten, also größer ($>$), größer gleich ($\geq$), kleiner ($<$), und kleiner gleich ($\leq$) angucken!

Bei dem ersten ist die 3 nicht mit inbegriffen, also muss bei dem Gegensatz der kumulierten relativen Häufigkeit die 3 mit aufgenommen werden. Sonst wäre die 3 in keinem der beiden Gruppen und würde fehlen!

Bei der zweiten Berechnung ist die 3 am Anfang mit inbegriffen. Somit muss beim Gegensatz die 2 ausgelassen werden.

Fernab des Beispiel oben, generell noch mal bezüglich den Gegensätzen:

Es gibt bei solchen Berechnung immer genau zwei Gruppen, in dem die Zahlen größer oder kleiner x_i fallen. Wenn die Gegenhäufigkeit berechnet wird, muss man sicher gehen, dass man genau die relative Häufigkeit für die komplette Gegenmenge berechnet.

Sei $x_i = 1, 2, 3, 4, 5$ (hier sind nur Ausprägungen und keine Häufigkeiten angegeben):

$$H_{rel}(x > 5) = 1 - H_{rel}(x \leq 5) = 1 - H(5) = 0$$
, da es keine Ausprägungen über 5 gibt und somit auch keine relative Häufigkeiten.
$$H_{rel}(x \geq 5) = 1 - H_{rel}(x < 5) = 1 - H_{rel}(x \leq 4) = 1 - H(4)$$

$$H_{rel}(x > 4) = 1 - H_{rel}(x \leq 4) = 1 - H(4)$$
$$H_{rel}(x \geq 4) = 1 - H_{rel}(x < 4) = 1 - H_{rel}(x \leq 3) = 1 - H(3)$$

$$H_{rel}(x > 3) = 1 - H_{rel}(x \leq 3) = 1 - H(3)$$
$$H_{rel}(x \geq 3) = 1 - H_{rel}(x < 3) = 1 - H_{rel}(x \leq 2) = 1 - H(2)$$

$$H_{rel}(x > 2) = 1 - H_{rel}(x \leq 2) = 1 - H(2)$$
$$H_{rel}(x \geq 2) = 1 - H_{rel}(x < 2) = 1 - H_{rel}(x \leq 1) = 1 - H(1)$$

$$H_{rel}(x > 1) = 1 - H_{rel}(x \leq 1) = 1 - H(1)$$
$$H_{rel}(x \geq 1) = 1 - H_{rel}(x < 1) = 1 - H_{rel}(x \leq 0) = 1 - H(0) = 1$$
, da es keine Zahlen kleiner als 1 und somit auch keine relative Häufigkeiten gibt.

Zur Übung könnten relative Häufigkeiten für die Reihen

$$y_i = 11, 23, 34, 54, 55$$
$$z_i = 0.5, 1, 1.5, 2, 2.5$$
$$t_i = 4, 5, 6, 7, 8$$

aufgestellt werden.

Verteilungsfunktion

Tipp 1: Man sollte die Kreise bei den Sprungstellen groß genug zeichnen, damit man bei der Korrektur klar erkennen kann, wo die Funktion springt. Am Ende einer waagerechten (horizontalen) Linie sollte also ein *leerer* Kreis und am Anfang einer horizontalen Linie ein ausgefüllter Kreis gezeichnet werden. Die vertikalen Linien zwischen den Sprungstellen sind nicht notwendig. Sie können gezeichnet werden (ob durchgezogen oder gestrichelt), sind aber nicht wichtig um die volle Punktzahl zu kriegen.

Tipp 2: Weiter sollte die Eigenschaft der Treppenfunktion bezüglich den Grenzwerten beachtet werden. So sollte man immer die untere Grenze bei Null ($\lim_{x\to-\infty} H(x) = 0$) und die obere Grenze bei 1 ($\lim_{x\to\infty} H(x) = 1$) einzeichnen!

```
# beob <- rep(x_i,n_i) Die Information von x und n muss in einem Vektor
# gespeichert werden. cdf_fun <- ecdf(beob) ecdf erstellt einen Vektor für
# die empirische Verteilungsfunktion (ecdf: empirical cumulative
# distribution function) plot(cdf_fun, xlab='Noten', ylab='Empirische
# Verteilungsfunktion',main='')
```

```
beob <- rep(x_i, n_i)
x_leer <- x_i + c(1, 1, 1, 1, 3, 3)
H_leer <- c(0, H_i[1:5])
ggplot(data = as.data.frame(beob), aes(x = beob)) + stat_ecdf() +
    geom_point(data = dat, aes(x = x_i, y = H_i), size = 2) +
    geom_point(data = dat, aes(x = x_i, y = H_leer), shape = 1,
        size = 2) + labs(title = "Treppenfunktion der Beobachtungsreihe 'Noten'",
    y = "Empirische Verteilungsfunktion", x = "Noten") + scale_x_continuous(limits = c(0,
    10), breaks = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)) + annotate("text",
    x = 9.3, y = 0.9, label = "Linie weiter zeichnen!") + annotate("text",
    x = 9.3, y = 0.85, label = "Wichtig!") + annotate("text",
    x = 3.5, y = 0.78, label = "Inbegriffen, voller Kreis") +
    annotate("text", x = 4, y = 0.63, label = "Nicht inbegriffen, leerer Kreis")
```

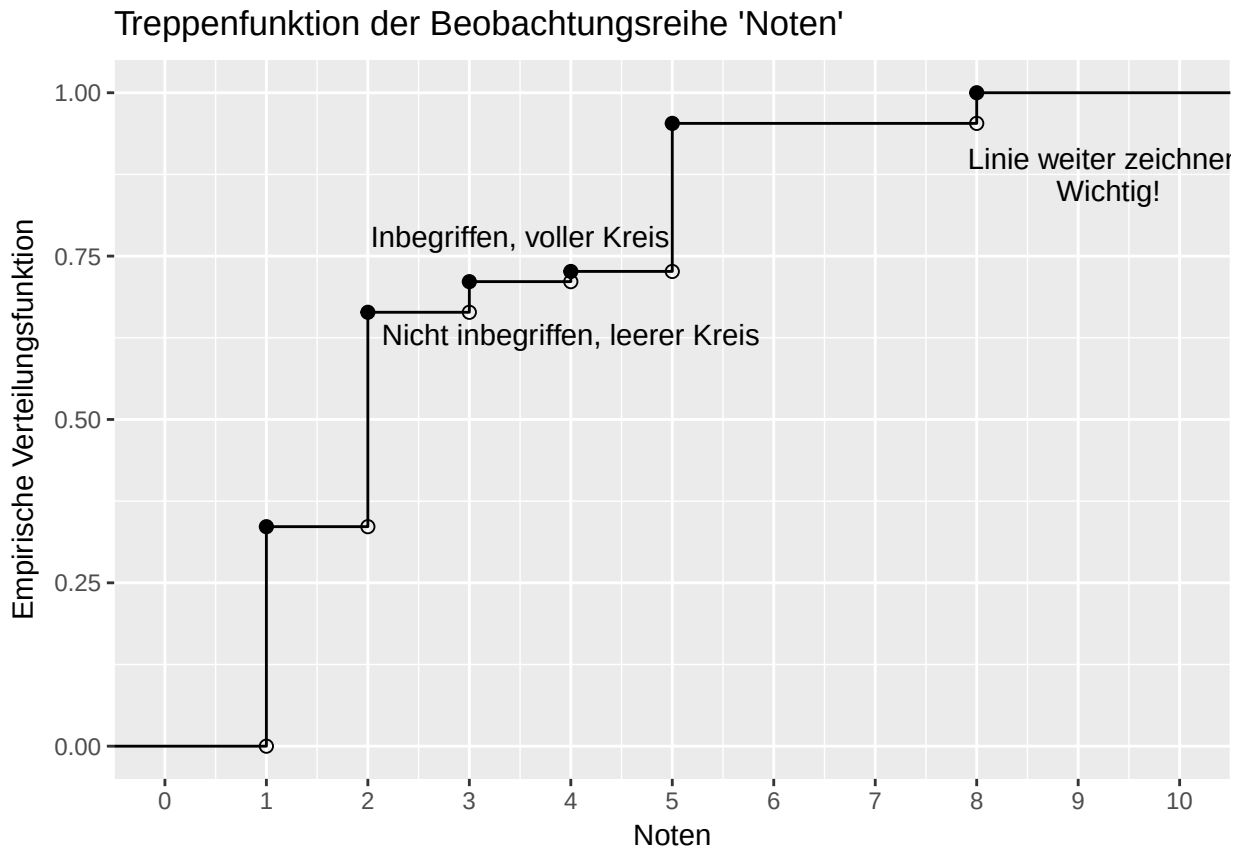


Tabelle von Klassen

Tipp: Bei den Klassen muss klar definiert werden, welche Werte in welcher Gruppe sind. Ist zum Beispiel das Alter von Individuen in den Klassen $18-20; 20-22; 22-24; 24-26$ gegeben, dann müssen die runden und eckigen Klammern angegeben werden, um klar zu stellen, in welche Gruppe man fällt, wenn man genau das Schnittalter hat (20, 22, 24). Konvention für die Vorlesung ist wie folgt: $(18-20]; (20-22]; (22-24]; (24-26]$. Die runde Klammer besagt, dass die Zahl nicht mehr mit aufgenommen wird, bei der Eckigen wird die Zahl noch zur Gruppe hinzugefügt. Formel geschrieben: $\xi_{i-1} < x \leq \xi_i$. x ist immer größer als die untere Grenze, aber kleiner **gleich** der obigen Grenze.

Berechnung der Verteilungsfunktion bei stetigen Merkmalen, die klassifiziert wurden (approximierender Polygonzug)]

Tipp: Um den Punkt ($\overline{H}(x)$) und die Steigung ($\overline{h}(x)$) des approximierenden Polygonzuges zu berechnen, müssen jetzt folgende Formeln verwendet werden:

$$\overline{H}(x) = H_K(\xi_{i-1}) + \overline{h}(x)(x - \xi_{i-1}) \qquad \text{, mit}$$
$$\overline{h}(x) = \frac{H_K(\xi_i) - H_K(\xi_{i-1})}{\xi_i - \xi_{i-1}} = \frac{h_i}{\Delta_i}, \quad i = 1, \ldots, m.,$$

Um die Rechnung zu vereinfachen, sollten dann weiter die benötigten Werte in folgender Tabelle berechnet werden:

```r
ci_im <- seq(150, 190, 10)
ci_i <- ci_im + 10
n_i <- c(15, 41, 46, 23, 3)
h_i <- round(n_i/sum(n_i), 5)
delta_i <- ci_i - ci_im
h_bar <- round(h_i/delta_i, 5)
H_k <- cumsum(h_i)
dat <- data.frame(ci_im, ci_i, n_i, h_i,
    delta_i, h_bar, H_k)
knitr::kable(dat, col.names = c("$(\\xi_{i-1}$;",
    "$\\xi_{i}$]", "$n_{i}$", "$h_{i}$",
    "$\\Delta_{i}$", "$\\overline{h}(x_{i})= \\frac{h_i}{\\Delta_i}$",
    "$H_{K}(x_i)= \\sum_{i|\\xi_i \\leq x} {h_i}$"),
    bootstrap_options = c("striped", "hover",
        "condensed", "responsive"), full_width = F)
```

$(\xi_{i-1}$;	$\xi_i]$	n_i	h_i	Δ_i	$\overline{h}(x_i) = \frac{h_i}{\Delta_i}$	$H_K(x_i) = \sum_{i\|\xi_i \leq x} h_i$
150	160	15	0.11719	10	0.01172	0.11719
160	170	41	0.32031	10	0.03203	0.43750
170	180	46	0.35938	10	0.03594	0.79688
180	190	23	0.17969	10	0.01797	0.97657
190	200	3	0.02344	10	0.00234	1.00001

Es sollte auffallen, dass hier die Höhe einfach geteilt durch 10 ist, da die Breite 'Δ_i' überall genau 10 ist. Dies muss nicht immer der Fall sein! 'Δ_i' berechnet sich, indem man $\xi_i - \xi_{i-1}$ rechnet. Hier ist z.b. $\xi_1 - \xi_{1-1} = \xi_1 - \xi_0 = 160 - 150 = 10$ oder $\xi_5 - \xi_{5-1} = \xi_5 - \xi_4 = 200 - 190 = 10$.

Die Mehrheit der Studierende sind zwischen 170 cm und 180 cm groß, danach zwischen 160 cm und 170 cm. Die kleinste Gruppen bilden Studierende die größer als 190 cm sind. Dies wird (auch) deutlich durch die absoluten Häufigkeiten n_i.

Sei weiter z.B. die Frage, wo der Polygonzug für x = 175 liegt, muss als erstes ermittelt werden, in welche Klasse 175 fällt. Hier in diesem Beispiel fällt 175 in die 3. Klasse, da $170 < 175 \leq 180$. Somit ist $\xi_i = 180$ und $\xi_{i-1} = 170$. Als nächstes muss die Steigung an dieser Stelle berechnet werden, um dann den Punkt selber zu berechnen (= den zugehörigen y-Wert zum x-Wert finden).

$$\begin{aligned}
\overline{h}(x) =& \frac{H_K(180) - H_K(170)}{180 - 170} = \frac{h_3}{\Delta_3} = \frac{0.3203}{10} = 0.03594 \\
\overline{H}(175) =& H_K(\xi_2) + \overline{h}(175)(175 - \xi_2) = H_K(170) + \overline{h}(175)(175 - 170) \\
=& 0.4375 + 0.03594 \cdot 5 = 0.6172
\end{aligned}$$

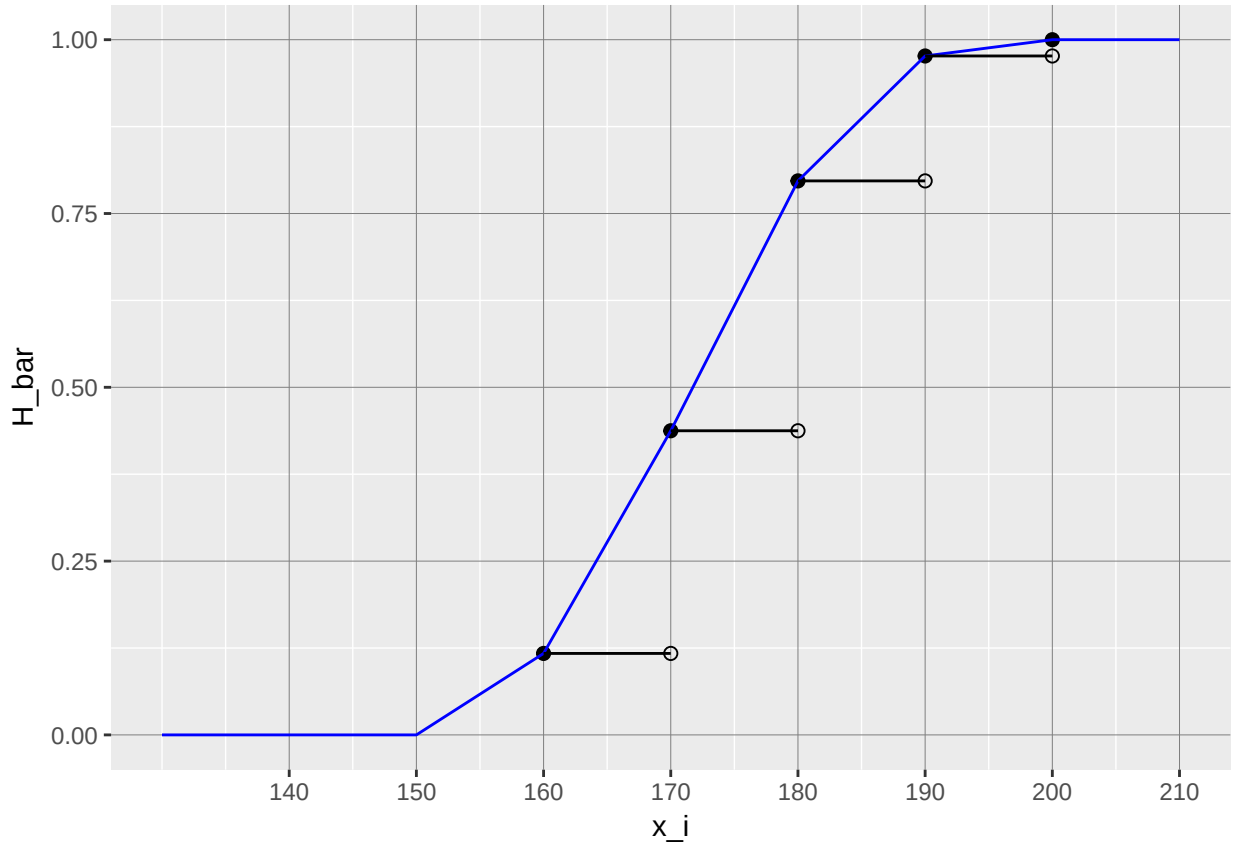
Es sollte klar sein, dass $\overline{H}(175)$ *irgendwo* zwischen $\overline{H}(170) = K_K(170)$ und $\overline{H}(180) = H_K(180)$ liegen sollte. Sobald das Ergebnis nicht zwischen diesen Grenzen liegt, sollte mindestens der Fehler benannt werden. Erste Fehlermöglichkeit könnte sein, dass nicht $H_K(\xi_{i-1})$ sondern $H_K(\xi_i)$ verwendet wurde.

Was wurde jetzt berechnet? Bei stetigen Merkmalen verbindet man die vollen Punkte in der Treppenfunktion und $\overline{H}(175)$ ist eben genau der Punkt dieser Verbindungslinie bei $x = 175$. $\overline{h}(x)$ ist die Steigung an dieser Stelle. Sollte man in der Klausur die Tabelle erstellen, muss man die Berechnung der Steigung nicht noch mal extra aufschreiben.

Zur Übung sollte man unter anderem die Punkte $\overline{H}(165) = 0.5976$ und $\overline{H}(185) = 0.707$ berechnen.

Polygonzug richtig Zeichnen

```
x_line <- c(130, 150, 160, 170, 180, 190, 200, 210)
x_i <- x_line[3:7]
x_leer <- x_i[-5] + 10
H_line <- c(0, 0, 0.1172, 0.4375, 0.7969, 0.9766, 1, 1)
H_bar <- H_line[3:7]
H_leer <- H_bar[-5]
y <- rep(H_leer,2)
dat <- data.frame(x_i, H_bar)
dat_leer <- data.frame(x_leer, H_leer)
dat_line <- data.frame(x_line,H_line)
dat_x <- data.frame(x = c(x_i[-5], x_leer), y = y)
ggplot(data=dat, aes(x=x_i, y=H_bar)) +
  geom_point(data=dat, aes(x=x_i, y=H_bar), size = 2) +
  geom_point(data=dat_leer, aes(x=x_leer, y=H_leer), shape = 1,size = 2) +
  geom_line(data=dat_line, aes(x=x_line, y=H_line), color="blue") +
  geom_line(data=dat_x, aes(x=x, y=y, group = y)) +
  scale_x_continuous(limits = c(130,210), breaks = seq(140,210,10)) +
  scale_y_continuous(limits = c(0,1)) +
  theme(panel.grid.major = element_line(size = 0.1, linetype = 'solid',colour = "grey50"))
```




```
#### Commands which should be included in the future (but have problems right now)
#  labs(x=expression(x_i))
#  annotate("text", x = 160, y = 0.50, label = "H[k]^2~(x[i])",parse = TRUE, angle = 58) +
#         label = "Value~is~sigma~R^{2}==0.6 ",parse = TRUE,size=20)
```

```
#'$\\overline{H}_{k}(x_i)$' mit der Steigung '$\\Delta_i$'
```

Die blaue Linie wird durch die Formel $\overline{H}(x_i)$ berechnet, wobei die Steigung $\overline{h}(x)$ ist. Bei den vollen Punkte ist weiter $\overline{H}(x_i) = H_k(x_i)$

Zu beachten ist hier: (i) Die x-Achse ist gekürzt, damit sie nicht zu lang wird und die Zeichung selber besser zu erkennen ist, (ii) die leeren Kreise zeigen die Sprungstellen an, (iii) der erste volle Punkt startet bei 160, und nicht bei 150 und (iv) $\xi_0 = 150$, $\xi_1 = 160$, $\xi_2 = 170$, $\xi_3 = 180$, $\xi_4 = 190$, $\xi_5 = 200$. (v) Weiter ließt sich x_i als x_i und H_bar als $\overline{H}(x)$. In der Kürze konnte ich die *schönere* Schreibweise noch nicht in die Grafik einfügen.

Wichtig bei der Grafik ist, dass der erste volle Punkt bei 160 gezeichnet wird. Warum? Die Beobachtungen der ersten Klasse sind irgendwo zwischen 150 und 160. Es ist keine Information darüber gegeben, wo genau innerhalb dieser Klasse die Beobachtungen sich befinden. Erst bei 160 kann man sich dabei sicher sein, wirklich alle Beobachtungen zu haben und damit bei der Häufigkeit für diese Gruppe zu landen.

Als Annahme ist gegeben, dass die Individuen in der Klasse gleichverteilt sind. Diese Annahme ist der Grund, warum die gerade Linie zwischen 150 und 160 gezeichnet werden darf und somit $\overline{H}(x)$ berechnet werden kann. Wenn eine andere Verteilungsannahme genommen wird, würden die Verbindungslinie zwischen den beiden Punkten anders aussehen. Dies ist für uns hier aber nicht weiter von Bedeutung, da die Gleichverteilung eine aussreichend gute Abschätzung.

Grafiken

Tipp 1: Wenn man wie bei dem Punkt vorher bei klassierten Daten ein Histogramm oder einen Polygonzug erstellt, hilft es die jeweiligen Grenzen mit den Werten für die sehen stehen zu beschriften. Hierbei sollte klar werden, wo man anfangen und wo man aufhören muss. Also was genau ξ_0, ξ_1 u.s.w sind.

Tipp 2: Man sollte die Grafik vom letzten Punkt (hier: 200) an zeichnen. Hierdruch wird klar, dass der *erste* volle Punkt bei 160 und nicht bei 150 liegt.

Tipp 3: Linien sollten mit einem Linear gezeichnet werden, damit klar ist, dass es keine Kurve ist. Ein guter Ersatz für das Linear ist der Studentenausweis.

Kapitel 3

Berechnung des arithmetischen Mittels und des Medians (gilt auch für Quantile)

Wichtig zu beachten ist, dass die Daten hier schon gruppiert sind. Das bedeutet, dass es die unterschiedlichen Merkmalsausprägungen für X häufiger gibt. Die genau Anzahl wird durch n_i angegeben.

```
i <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)
x_i <- c(1, 1.3, 1.7, 2, 2.3, 2.7, 3, 3.3, 3.7, 4, 5)
n_i <- c(6, 3, 6, 6, 2, 2, 6, 2, 3, 1, 12)
sum_n <- c(6, 9, 15, 21, 23, 25, 31, 33, 36, 37, 49)
dat <- data.frame(i, x_i, n_i, sum_n)
knitr::kable(dat, col.names = c("$i$", "$x_{i}$", "$n_{i}$",
    "$\\sum_{j=1}^{i} n_j$"), bootstrap_options = c("striped",
    "hover", "condensed", "responsive"), full_width = F)
```

i	x_i	n_i	$\sum_{j=1}^{i} n_j$
1	1.0	6	6
2	1.3	3	9
3	1.7	6	15
4	2.0	6	21
5	2.3	2	23
6	2.7	2	25
7	3.0	6	31
8	3.3	2	33
9	3.7	3	36
10	4.0	1	37
11	5.0	12	49

Das arithmetische Mittel berechnet sich wie folgt:

$$\begin{aligned}\overline{x} =& \frac{1}{n}\sum_{j=1}^{k} n_j x_j \\ =& \frac{1}{49}(6\cdot 1.0 + 3\cdot 1.3 + 6\cdot 1.7 + 6\cdot 2.0 + 2\cdot 2.3 + 2\cdot 2.7 \\ & + 6\cdot 3.0 + 2\cdot 3.3 + 2\cdot 3.7 + 1\cdot 4.0 + 12\cdot 5.0) \\ =& 2.8939\end{aligned}$$

Man multipliziert jeweils die Merkmalsausprägung von X mit der dazugehörigen absoluten Häufigkeit. Dazugehörigig ist hier, was in der gleichen Zeile steht.

Bevor man den Median berechnet, sollte festgestellt werden, ob n gerade oder ungerade ist, um die Richtige Formel zu verwenden. Hier ist n ungerade. Weiter kann man, um die Rechnung zu vereinfachen, die 4. Spalte in der obigen Tabelle erstellen, um die genauen Position zu finden.

$$\begin{aligned}\tilde{x} &= x_{\left(\frac{(n+1)}{2}\right)} = x_{\left(\frac{(49+1)}{2}\right)} = x_{(50/2)} = x_{(25)} \\ &= 2.7 \text{ , wobei } x_{(25)} \neq 25!\end{aligned}$$

Die 25 im Index bezieht sich auf die Position in der geordneten Zahlenreihe, nicht auf den tatsächlichen Wert.

Wichtig ist hier erstens, dass auch das richtige n gewählt wird. Hier muss n (natürlich) genauso wie oben 49 sein. In den Übungen wurden hier auch andere Zahlen eingesetzt.

Wie kommt man auf die 2.7? Man könnte eine geordnete Zahlenreihe aufschreiben und dann einfach abzählen. Dies kann aber etwas lange dauern, weswegen man einfach die n_i's aufsummieren sollte, bis man bei 25 ankommt. Dies wird hier eben bei 2.7 erreicht. Summiert man z.B. bis 2.3 auf, hat man erst 23 Beobachtungswerte aufsummiert. Dies wird in der vierten Zeile der Tabelle mit $\sum_{j=1}^{i} n_j$ verdeutlicht. Die Einträge sind immer eine Summe bis zu dem angegebenen i. (Selber nachrechnen kann beim Nachvollziehen helfen.)

Die Beobachtungsreihe ist übrigens ein Beispiel, bei dem der Median nicht zwischen Modus und Mittelwert liegt! (Es sollte klar sein, dass der Modus 12 ist.) In der Vorlesung wurden die Lageregeln für statistische Verteilungen aufgestellt, bei denen der Modus nicht am Rand liegt. Hier ist dies nicht der Fall, weswegen u.a. die Ordnung hier unterschiedlich zur Vorlesung ist. (Zur Übung könnte man die Verteilung zeichnen.)

Der Modus ist die Merkmalsausprägung, die am häufigsten vorkommt. In diesem Fall ist es $x_{11} = 5.0$.

Geometrisches Mittel

Tipp 1: Negative Wachstumsrate werden wie folgt aufgeschrieben: Hat man einen negativen Wachstum von $-2\% = -0.02$, so muss man für die Berechnung des geometrischen Mittel den Wert mit 1 addieren, $1 + (-0.02) = 1 - 0.02 = 0.98$. Negative Wachstumsraten haben einen Wert zwischen 0 und 1!

Tipp 2: Wann wird das geometrische Mittel verwendet? Bei Wachstum über die Zeit, bei dem das vorige Wachstum wichtig für den nächsten Wachstum ist. Beispiele sind die wirtschaftliche Entwicklung eines Landes oder Zinsmodelle bei Konten. In der Physik und anderen Naturwissenschaften hat sie noch weitere Anwendungen.

Quartile und Quantile

Es gibt genau drei Quartile: 1. Quartil, 2. Quartil und 3.Quartil. Diese Quartile teilen die Daten in 4 Teile. (i) Kleiner als das 1. Quartil, (ii) zwischen dem 1. und dem 2., (iii) zwischen dem 2. und dem 3. und als letztes (iv) größer als das 3.

Das 1. Quartil ist gleich dem 0.25 Quantil

Das 2. Quartil ist gleich dem 0.5 Quantil, also dem Median

Das 3. Quartil ist gleich dem 0.75 Quantil

Will man ein Quartil berechnen, kann man einfach das dazu gehörige Quantil berechnen. Die Berechnung eines Quantils ist zweigeteilt:

Wenn $n \cdot q$ **keine** natürlich Zahl ist, muss man den $\langle n \cdot q \rangle$ ten Wert finden. Hierbei ist q gleich dem Quantil und die spitzen Klammern stehen für das Aufrunden des Produktes

Wenn $n \cdot q$ **eine** natürlich Zahl ist, muss man den den Mittelwert zwischen $x_{(nq)}$ und $x_{(nq+1)}$ nehmen, also $\frac{1}{2} \cdot (x_{(nq)} + x_{(nq+1)})$.

Wichtig zu beachten ist, dass z.B. $x_{(30)}$ der 30. x-Wert in einer geordneten Beobachtungsreihe präsentiert, nicht unbedingt den Wert 30!

Berechnung des MAA

Tipp: Für die Erklärung erstmal einen Schritt zurück: Für das arithmetische Mittel gilt folgende Gleichung:

$$\overline{x} = \frac{1}{n}\sum_{j=i}^{\mathbf{n}} x_j = \frac{1}{n}\sum_{j=i}^{\mathbf{k}} n_j \cdot x_j = \sum_{j=i}^{\mathbf{k}} h_j x_j \qquad \text{, mit } k \leq n.$$

Hier ist wichtig zu sehen, dass wenn die Daten gruppiert (und somit mindestens n_i oder h_i) angegeben sind, dann summiert man nur bis k, sonst bis n. Das gleiche passiert auch bei der MAA. In der Vorlesung wurde nur die allgemeine Formel mit der Summe bis n präsentiert. Sind die Werte aber gruppiert angegeben, dann kann man die gleiche Umstellung wie bei dem arithmetischen Mittel durchführen:

$$MAA = \frac{1}{n}\sum_{j=i}^{\mathbf{n}} |x_j - \overline{x}| = \frac{1}{n}\sum_{j=i}^{\mathbf{k}} n_j \cdot |x_j - \overline{x}| = \sum_{j=i}^{\mathbf{k}} h_j |x_j - \overline{x}|$$

Das Erkennen der nötigen Umstellung war Teil der Übungsaufgabe. Weiter sind auch alle drei Formeln in der Formelsammlung gegeben.

Beispielhaft hier die Mathenote von Studierende (aus dem Fragebogen) die in einem zufälligen Monat geboren sind:

```
i <- c(1, 2, 3, 4, 5, 6, 7)
x_i <- c(1.3, 2.7, 3, 3.7, 4, 5.3, 6)
n_i <- c(1, 2, 3, 1, 3, 1, 1)
h_i <- round(n_i/sum(n_i), 4)
x_mean <- sum(x_i * n_i)/sum(n_i)
x_demeand <- round(abs(x_i - x_mean), 4)
dat <- data.frame(i, x_i, n_i, h_i, x_demeand)
knitr::kable(dat, col.names = c("$i$", "$x_{i}$", "$n_{i}$",
    "$h_i$", "$\\mid x_i - \\overline x \\mid$"), bootstrap_options = c("striped",
    "hover", "condensed", "responsive"), full_width = F)
```

i	x_i	n_i	h_i	$\mid x_i - \overline{x} \mid$
1	1.3	1	0.0833	2.2583
2	2.7	2	0.1667	0.8583
3	3.0	3	0.2500	0.5583
4	3.7	1	0.0833	0.1417
5	4.0	3	0.2500	0.4417
6	5.3	1	0.0833	1.7417
7	6.0	1	0.0833	2.4417

Zur Wiederholung: $n = \sum_{j=1}^{n} n_j = 12$ und $h_i = n_i/n$. H_i wird hier nicht berechnet, da es keine Relevanz für die Aufgabe hat. $\overline{x} = 3.5583$.

$$
\begin{aligned}
MAA_1 &= \frac{1}{n}\sum_{j=i}^{n}|x_j-\overline{x}| \\
&= \frac{1}{12}(|1.3-3.5583|+|2.7-3.5583|+|2.7-3.5583|+|3-3.5583|+|3-3.5583|+|3-3.5583| \\
&+|3.7-3.5583|+|4-3.5583|+|4-3.5583|+|4-3.5583|+|5.3-3.5583|+|6-3.5583|) \\
&= \frac{1}{12}(2.2583+0.8583+0.8583+0.5583+0.5583+0.5583+0.1417+0.4417+0.4417+0.4417+1.7417+2.4417) \\
&= \frac{1}{12}(8.4414) = 0.70345
\end{aligned}
$$

$$
\begin{aligned}
MAA_2 &= \frac{1}{n}\sum_{j=i}^{k} n_j\cdot|x_j-\overline{x}| \\
&= \frac{1}{12}(|1.3-3.5583|+2\cdot|2.7-3.5583|+3\cdot|3-3.5583|+|3.7-3.5583| \\
&+3\cdot|4-3.5583|+|5.3-3.5583|+|6-3.5583|) \\
&= \frac{1}{12}(2.2583+2\cdot0.8583+3\cdot0.5583+0.1417+3\cdot0.4417+1.7417+2.4417) \\
&= \frac{1}{12}(8.4414) = 0.70345
\end{aligned}
$$

$$
\begin{aligned}
MAA_3 &= \sum_{j=i}^{k} h_j\cdot|x_j-\overline{x}| \\
&= (0.0833\cdot|1.3-3.5583|+0.1667\cdot|2.7-3.5583|+0.2500\cdot|3-3.5583|+0.0833\cdot|3.7-3.5583| \\
&+0.2500\cdot|4-3.5583|+0.0833\cdot|5.3-3.5583|+0.0833\cdot|6-3.5583|) \\
&= (0.0833\cdot1.95+0.1667\cdot0.55+0.2500\cdot0.25+0.0833\cdot0.45 \\
&+0.2500\cdot0.75+0.0833\cdot2.05+0.0833\cdot2.75) \\
&= 0.70345
\end{aligned}
$$

Man sieht, dass alle drei Rechenarten zum gleich Ergebnis kommen, weil $MAA_1 = MAA_2 = MAA_3$ ist. Welche Formel verwendet wird, ist freigestellt, allerdings ist die 2. und 3. meiner Einschätzung nach kürzer. Hilfreich ist es weiter, die Rechnung in den Bestragsstrichen einfach mit in die Tabelle aufzunehmen (Spalte 5), damit die übersicht besser ist. (Ich persönlich würde die zweite Formel verwenden.)

IQA und MQA

Tipp: Der Interquartilsabstand (IQA) bezieht sich auf den Abstand des 3.Quartils vom 1.Quartil. Das 1.Quartil ist gleich dem 25% Quantil und das 3.Quartil gleich dem 75% Quantil: $Q_1 = x_{0.25}$ und $Q_3 = x_{0.75}$. Somit ist $IQA = Q_3 - Q_1 = x_{0.75} - x_{0.25}$ Um also Q_1 und Q_3 auszurechnen, muss man die Formel für $x_{0.25}$ und $x_{0.75}$ anwenden. Je nachdem ob $n \cdot q$ eine ganze Zahl ist oder nicht, ist die Formel unterschiedlich:

$$
x_{\mathrm{q}} = \begin{cases} x_{([qn]+1)} = x_{\langle nq\rangle} & \text{falls } qn \notin \mathbb{N} \\ \frac{1}{2}\cdot(x_{(nq)} + x_{(nq+1)}) & \text{falls } qn \in \mathbb{N} \end{cases}
$$

Nimmt man das Beispiel der Noten welches auch für die MAA benutzt wurde, ergibt sich $n \cdot q = 12 \cdot 0.25 = 3$ und $n \cdot q = 12 \cdot 0.75 = 9$. Somit sind beides ganze Zahlen und die zweite Formel muss verwendet werden:

$$
\begin{aligned}
x_{0.25} &= \frac{1}{2} \cdot (x_3 + x_4) = \frac{1}{2} \cdot (2.7 + 3) = 2.85 = Q_1 \\
x_{0.75} &= \frac{1}{2} \cdot (x_9 + x_{10}) = \frac{1}{2} \cdot (4 + 4) = 4 = Q_3 \\
\Rightarrow IQA &= Q_3 - Q_1 = 4 - 2.85 = 1.15 \\
\Rightarrow MQA &= \frac{IQA}{2} = 1.15/2 = 0.575
\end{aligned}
$$

Vergleicht man den MAA mit dem MQA sieht man, dass der MQA kleiner ist. Die sollte an den Rändern 1.3 und 6 liegen.

Berechnung der Varianz

Tipp 1:

$$
s_x^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \bar{x}^2 = \frac{1}{n}\sum_{j=1}^{k} n_j x_j^2 - \overline{x}^2 = \sum_{j=1}^{k} h_j x_j^2 - \overline{x}^2 = \overline{x^2} - \bar{x}^2
$$

Die verkürzte Berechnung ($\overline{x^2} - \bar{x}^2$) ist weniger anfällig für Fehler als die lange Berechnung ($\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$). Ein Grund dafür ist, dass man mehr Rechnungen in einzelne Zwischenschritte machen kann. Hier sei jetzt nur die kurze Schreibweise beispielhaft angegeben. Es werden wieder die Daten von der MAA benutzt. Da die Daten dort gruppiert angegeben sind, wird weiter nur eine der drei kurzen Berechnungen angewendet. Es sollte klar sein, dass die Formel für die komplette Reihe mit der Summe von i bis n und die Formel mit h_i zu gleichen Ergebnissen führt.

$$
s_x^2 = \frac{1}{n}\sum_{j=1}^{k} n_j x_j^2 - \overline{x}^2 = \overline{x^2} - \overline{x}^2
$$

Zu erst wird hierfür das arithmetische Mittel berechnet:

$$
\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{j=1}^{k} n_j x_j = \frac{1}{12}(1 \cdot 1.3 + 2 \cdot 2.7 + 3 \cdot 3 + 1 \cdot 3.7 + 3 \cdot 4 + 1 \cdot 5.3 + 1 \cdot 6) \\
&= \frac{1}{12}(42.7) = \frac{427}{120} = 3.5583
\end{aligned}
$$

Als nächstes berechnet man dann das arithmetische Mittel für x_j^2. Die Gleichung ist ähnlich zu der obigen, nur das hier die x-Werte zuallererst quadriert werden müssen!

$$
\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{j=1}^{k} n_j x_j = \frac{1}{12}(1 \cdot 1.3^2 + 2 \cdot 2.7^2 + 3 \cdot 3^2 + 1 \cdot 3.7^2 + 3 \cdot 4^2 + 1 \cdot 5.3^2 + 1 \cdot 6^2) \\
&= \frac{1}{12}(169.05) = \frac{1127}{80} = 14.0875
\end{aligned}
$$

Dann kann jetzt die Varianz selber berechnet werden:

$$
s_x^2 = \frac{1}{n}\sum_{j=1}^{k} n_j x_j^2 - \overline{x}^2 = \overline{x^2} - \overline{x}^2 = 14.0875 - \left(\frac{427}{120}\right)^2 = 1.4258
$$

Tipp 2:

In der Formel für gruppierte Daten, $s_x^2 = \frac{1}{n}\sum_{j=1}^{k} n_j x_j^2 - \overline{x}^2$ ist das Quadratzeichen bei dem x. Ein beliebter Fehler ist es, dass nicht x_j quadriert wird, sondern n_j, was natürlich falsch und ohne Sinn ist. Es gibt keinen Grund, die Häufigkeit einer Merkmalsausprägung zu quadrieren.

Zerlegung in Intra- und Intergruppen Varianz

Tipp: Ein Beispiel, um die Formel zu verdeutlichen. Es sei der Mittelwert und die Varianz für das Gewicht nach Geschlecht für die Studierende gegeben. Die 51 Männer sind im Schnitt 76.52 kg schwer, mit einer Varianz von 143.0416. Die 70 Frauen sind im Schnitt 63.35 kg leicht, mit einer Varianz von 112.5808. Jetzt soll die Gesamtvarianz berechnet werden. Da es nur 2 Gruppen gibt, ist $m = 2$. Insgesamt haben $n = 51 + 70 = 121$ Studierende ihr Gewicht angegeben. Folgende Formel muss hierfür verwendet werden:

$$\Rightarrow \underbrace{s^2_{\text{ges}}}_{\substack{\text{Gesamt-}\\\text{varianz}}} = \underbrace{\frac{1}{n}\sum_{j=1}^{m} n_j s_j^2}_{\substack{\emptyset\text{-Varianz innerhalb}\\\text{der Teilgruppen}}} + \underbrace{\frac{1}{n}\sum_{j=1}^{m} n_j(\overline{x}_j - \overline{x})^2}_{\substack{\text{Varianz zwischen den Teil-}\\\text{gruppen (bzw. ihren Mittelw.)}}} = \frac{1}{n}\sum_{j=1}^{m} n_j s_j^2 + \frac{1}{n}\sum_{j=1}^{m} n_j \overline{x}_j^2 - \overline{x}^2$$

Die Berechnung wird wieder für die einzelnen Teile durchgeführt:

$$\frac{1}{n}\sum_{j=1}^{m} n_j s_j^2 = \frac{1}{121} \cdot (51 \cdot 143.0416 + 70 \cdot 112.5808) = 125.4196$$

Zu beachten ist hier, dass im Text schon direkt die jeweiligen Varianzen gegeben sind. Sollte nur die Standardabweichung gegeben sein, muss diese natürlich vorher (oder in der Gleichung) quadriert werden. (In der Übung war dies der Fall.)\
Weiter mit dem zweiten Teil:

$$\frac{1}{n}\sum_{j=1}^{m} n_j \overline{x}_j^2 = \frac{1}{121} \cdot (51 \cdot 76.52^2 + 70 \cdot 63.35^2) = 4789.639714$$

Für den dritten Teil muss die Formel für den Gesamtmittelwert benutzt werden, da der Wert nicht gegeben ist. Die Formel ist eigentlich die selbe, nur dass die unterschiedlichen Mittelwerte nicht quadriert werden:

$$\overline{x}_{ges} = \frac{1}{n}\sum_{j=1}^{m} n_j \overline{x}_j = \frac{1}{121} \cdot (51 \cdot 76.52 + 70 \cdot 63.35) = 68.901$$

$$\Rightarrow \overline{x}^2_{ges} = 4747.3467$$

Dann können alle drei Teile zusammengefasst werden:

$$\begin{aligned} s^2_{\text{ges}} &= \frac{1}{n}\sum_{j=1}^{m} n_j s_j^2 + \frac{1}{n}\sum_{j=1}^{m} n_j \overline{x}_j^2 - \overline{x}^2 \\ &= 125.4196 + 4789.639714 - 4747.3467 \\ &= 167.7126 \end{aligned}$$

Dispersionsindex und Diversität

Tipp 1: Beide Streuungsmaße sind normiert zwischen 0 und 1, also $0 \leq P \leq 1$ und $0 \leq D \leq 1$. Ergebnisse die größer 1 oder kleiner 0 sind, sind somit nicht möglich. Auch hier gilt wieder, dass wenn ein solches Ergebnis herauskommt, dass man mindestens sagen sollte, dass es falsch ist. (Mindestens weil man eigentlich den Fehler finden sollte.) \ In der Vorlesungsfolie war bezüglich der Diversität ein Fehler. Hier muss H_i und nicht h_i verwendet werden. Sonst handelt sich um eine *einfache* Summe mit einem Vorfaktor.

Tipp 2: Bei dem Dispersionsindex ist ein Teil der Formel wie folgt: $\left(1 - \sum_{j=1}^{m} h_j^2\right)$. Hier muss erst die komplette Summe berechnet werden und dann von 1 abgezogen werden. Um dies zu verdeutlichen, kann die Formel auch wie folgt schreiben: $\left(1 - \left[\sum_{j=1}^{m} h_j^2\right]\right)$. Ich würde hier aber die zweite Berechnung empfehlen, d.h. $\frac{m}{m-1} \sum_{j=1}^{m} h_j(1 - h_j)$. (Tendenziell, wenn eine Formel gegeben ist, und auf den Seiten danach die Formel anders angegeben ist, ist die letztere Berechungsart eine Vereinfachung somit zu bevorzugen.)

Kapitel 4

Tabelle zur Erstellung der Lorenzkurve

Tipp: Die Berechnung der Anteile von Merkmalssumme und Häufigkeiten werden in folgender Tabelle noch mal verdeutlicht:

```
i <- c(1, 2, 3, 4, 5)
n_i <- c(42, 48, 24, 10, 4)
x_i <- c(200, 250, 350, 450, 550)
n_x <- n_i * x_i
MS_i <- cumsum(n_x)
M_i <- round(MS_i/sum(n_x), 4)
h_i <- round(n_i/sum(n_i), 4)
H_i <- cumsum(h_i)
dat <- data.frame(i, n_i, x_i, n_x, MS_i, M_i, h_i, H_i)
knitr::kable(dat, col.names = c("$i$", "$n_{i}$", "$x_{i}$", "$n_{i} \\cdot x_i$",
    "$MS_i = \\sum_{j=1}^i n_i \\cdot x_i$", "$M_i = MS_i / MS$",
    "$h_i$", "$H_i$"), bootstrap_options = c("striped", "hover", "condensed",
    "responsive"), full_width = F)
```

i	n_i	x_i	$n_i \cdot x_i$	$MS_i = \sum_{j=1}^{i} n_i \cdot x_i$	$M_i = MS_i/MS$	h_i	H_i
1	42	200	8400	8400	0.2366	0.3281	0.3281
2	48	250	12000	20400	0.5746	0.3750	0.7031
3	24	350	8400	28800	0.8113	0.1875	0.8906
4	10	450	4500	33300	0.9380	0.0781	0.9687
5	4	550	2200	35500	1.0000	0.0312	0.9999

Normalerweise sind n_i und x_i gegeben. Mithilfe dieser beiden Informationen, kann man alle weiteren Werte berechnen. Es gab aber auch eine Aufgabe, in der Merkmalssumme und relative Häufigkeit gegeben ist. Man sollte sich anfangs immer vergewissern, welche Werte gegeben sind.

Um die Merkmalssumme zu berechnen, ist hier mit der dritten Spalte ein weiterer Zwischenschritt aufgenommen. In der dritten Spalte wird einfach für jede Zeile das Produkt von x_i und n_i genommen. **Dies ist aber noch nicht die Merkmalssumme!** Die Merkmalssumme, MS_i, wird erst in der 4. Spalte berechnet! Hier wird für jede Zeile kumulative gearbeitet! Die zeigt die Summe von j bis i, $\sum_{j=1}^{i}$, an. So wird für das MS_i immer die Zeile selber und alle Zeilen zuvor, aber auch nur zuvor, aufsummiert. Deswegen ist für die erste Zeile das Ergebnis einfach gleich dem Wert in der dritten Spalte, 8400, da es keine Werte vorher gibt (außer 0). In der nächsten Spalte dann aber, ist die Rechnung wie folgt:

$$\begin{aligned} MS_2 &= \sum_{j=1}^{2} n_j \cdot x_j = n_1 \cdot x_1 + n_2 \cdot x_2 \\ &= 42 \cdot 200 + 48 \cdot 250 = 8400 + 12000 \\ &= 20400 \end{aligned}$$

Woher weiß man, dass man hier nur die ersten beiden Zeilen aufaddieren soll? Hier hilft die erste Spalte. Dort ist nämlich das i definiert. i ist hierbei die Anzahl an Zeilen. Wenn $i = 2$ ist, müssen eben genau die ersten beiden Zeilen aufsummiert werden. Die Rechnung für die weiteren Zeilen sollten dann klar sein, um es aber deutlich zu machen:

$$
\begin{aligned}
MS_3 &= \sum_{j=1}^{3} n_j \cdot x_j = n_1 \cdot x_1 + n_2 \cdot x_2 + n_3 \cdot x_3 \\
&= MS_2 + n_3 \cdot x_3 \\
&= 20400 + 24 \cdot 350 = 20400 + 8400 \\
&= 28800
\end{aligned}
$$

$$
\begin{aligned}
MS_4 &= \sum_{j=1}^{4} n_j \cdot x_j = n_1 \cdot x_1 + n_2 \cdot x_2 + n_3 \cdot x_3 + n_4 \cdot x_4 \\
&= MS_3 + n_4 \cdot x_4 \\
&= 28800 + 10 \cdot 450 = 28800 + 4500 \\
&= 33300
\end{aligned}
$$

$$
\begin{aligned}
MS_5 &= \sum_{j=1}^{5} n_j \cdot x_j = n_1 \cdot x_1 + n_2 \cdot x_2 + n_3 \cdot x_3 + n_4 \cdot x_4 + n_5 \cdot x_5 \\
&= MS_4 + n_5 \cdot x_5 \\
&= 33300 + 4 \cdot 550 = 33300 + 2200 \\
&= 35500
\end{aligned}
$$

Das man

$$n_1 \cdot x_1 + n_2 \cdot x_2 + n_3 \cdot x_3 + n_4 \cdot x_4 + n_5 \cdot x_5$$

zu

$$MS_4 + n_5 \cdot x_5$$

abkürzen kann sollte klar sein. Bei

$$MS_4 =_1 \cdot x_1 + n_2 \cdot x_2 + n_3 \cdot x_3 + n_4 \cdot x_4$$

summiert man schon alles bis Zeile 4 auf. Wenn dann MS_5 berechnen muss, muss man auf MS_4 nur noch die neue Zeile 5 hinzuaddieren.

Für die Lorenzkurve muss allerdings noch der Anteil der Merkmalssumme (also ein Wert zwischen 0 und 1) berechnet werden. Für den Anteil muss man nur noch die jeweiligen Merkmalssummen (MS_i) durch die Gesamtmerkmalssumme ($MS = \sum_{j=1}^{k} = MS_5$) geteilt werden. Hier wird die Summe bis k genommen, da k immer der letzte Wert der Zahlenreihe i ist. In diesem Fall ist $k = 5$. Somit muss in der letzten Zeile bei M_i immer eine 1 stehen! Die Rechnung ist wie folgt:

$$
\begin{aligned}
M_1 &= \frac{MS_1}{MS} = 8400/35500 = 0.2366 \\
M_2 &= \frac{MS_2}{MS} = 20400/35500 = 0.5746 \\
M_3 &= \frac{MS_3}{MS} = 28800/35500 = 0.8113 \\
M_4 &= \frac{MS_4}{MS} = 33300/35500 = 0.9380 \\
M_5 &= \frac{MS_5}{MS} = 35500/35500 = 1
\end{aligned}
$$

Die Lorenzkurve steigt von (0,0) hin zu (1,1). Dies kann als Hilfe genommen werden, um nicht zu vergessen, dass die Merkmalssumme immer die vorigen Werte mit beinhaltet. Dadurch ist nämlich gegeben, dass die Merkmalssumme für jede neue Merkmalsausprägung steigt. Die Werte können in der vorigen Tabelle in der sechsten Spalte gefunden werden.

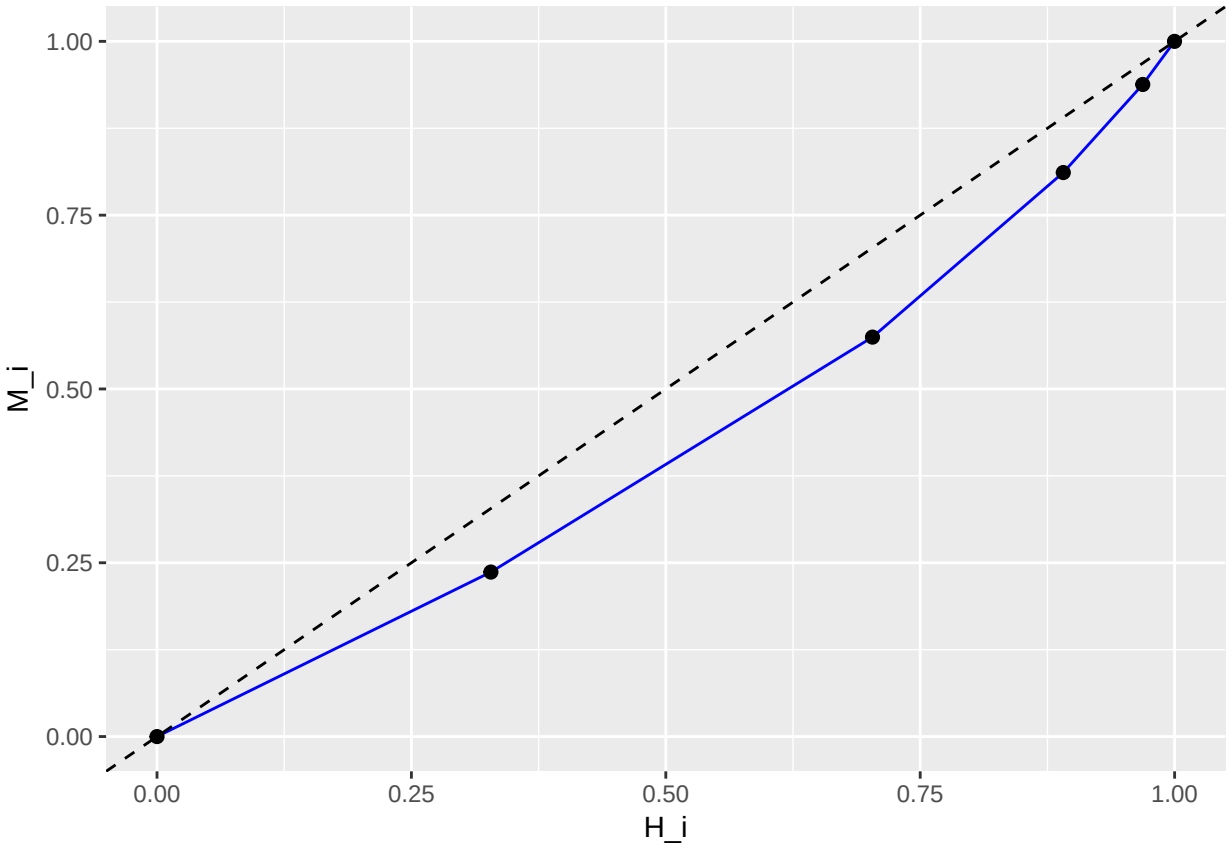
Als letztes muss für die Lorenzkurve noch der relative Anteil des Merkmalsträger berechnet werden. Dies ist nichts anderes als die empirische Verteilungsfunktion H_i aus Kapitel 2. Sie ist die Summe aller relativen Häufigkeiten h_i. Wer nicht weiß, wie man H_i berechnet, sollte dringend Kapitel 2 wiederholen.

Es sollte klar sein, dass für alle i gilt, dass $M_i \leq H_i$ ist. Hierdurch ist gewährleistet, dass die Lorenzkurve nicht oberhalb der Winkelhalbierenden ist (was sie per Definition nicht sein kann). In welchem Fall ist $M_i = H_i$? Immer für $i = 0$ und für das letztmögliche i (hier 5). Weiter kann es den Sonderfall der perfekten Gleichheit geben, bei dem die Lorenzkurve auf der Winkelhalbierenden liegt.

Zeichnen der Lorenzkurve

Tipp: Wenn man die Tabelle richtig berechnet hat, sollte das Zeichnen der Lorenzkurve kein Problem mehr sein. Der Vollständigkeit aber wird Sie in Abbildung **??** hier trotzdem aufgenommen:

```
dat0 <- rbind(0,dat) #Der Startpunkt 0 wird hinzugefügt.
ggplot(dat0, aes(x=H_i, y=M_i)) +
  geom_line( color = "blue") +
  geom_abline(intercept=0, slope=1, linetype="dashed") +
  geom_point(size=2)
```



Die gestrichelte Linie ist die Winkelalbierende, die die Lorenzkurve per Konstruktion nicht überschreiten kann. Da Lorenzkurve hier ziemlich nahe einer glatten Kurve ist, habe ich noch Punkte hinzugefügt, um die Zeichnung zu verdeutlichen. Bis auf den ersten Punkt bei (0,0), sind alle weiteren Werte aus der Tabelle ablesbar. Weiter muss aufgrund der kumulierten Werte die Steigung der Lorenzkurve mit jedem neuen Punkt zunehmen (steiler werden). Wenn diese Eigenschaft nicht zutrifft, die Lorenzkurve nicht bei (1,1) endet oder die Winkelhalbiernde schneidet, sollte klar sein, dass in der Rechnung ein Fehler vorliegt.

Tipp 2:

Weiter sollte man wissen, dass je kleiner die Fläche zwischen der Lorenzkurve und der Winkelhalbierende ist, desto kleiner auch die Konzentration (und damit die Ungleichverteilung) ist. Liegt eine Kurve monotom über einer anderen Kurve, ist dort die Ungleichheit kleiner. Schneiden sich zwei Kurven, ist eine einfache Aussage selten möglich. Hier sollte dann der Gini-Koeffizient mit betrachtet werden.

Tipp 3:

Die Lorenzkurve **muss** unterhalb der Winkelhalbierenden liegen. Weiter darf ein neuer Punkt weiter rechts auf der x-Achse (Hier: H_i), immer oberhalb des vorigen Punktes sein. Die blau Linie darf also niemals, von links nach rechts gesehen, nach unten gehen. Dies gilt, da M_i durch die Summen die vorigen Werte beinhaltet.

Berechnung des Gini-Koeffizienten

Tipp: Der Gini-Koeffizient steht für die Fläche zwischen der Winkelhalbierende und der Lorenzkurve. Sie ist ein relatives Maß, da die jeweilige Konzentrationsfläche durch die größtmögliche Konzentrationsfläche geteilt wird:

$$Gini = \frac{\text{Konzentrationsfläche}}{\text{größtmögl. Konzentrationsfläche}} = \frac{K}{K_{\max}}$$

Hierbei ist

$$K_{max} = \frac{1}{2} - \frac{1}{2n} = \frac{n}{2n} - \frac{1}{2n} = \frac{n-1}{2n}$$
$$(K_{max})^{-1} = \frac{2n}{n-1}$$

$(K_{max})^{-1}$ ist die Inverse von K_{max}. Es macht Sinn diese zu nehmen, da es man dadurch bei der Berechnung des Gini's ein Produkt und keine Division hat. Die Formel für K beruht auf einer Vereinfachung von Berechnungen von Dreiecken und Trapezen und wird hier nicht weiter erkl"art und einfach gegeben:

$$K = \frac{1}{2} - \sum_{i=1}^{k} \frac{1}{2} (M_{i-1} + M_i) \cdot h_i.$$

Bei K sowie K_{max} steht die 1/2 am Anfang für die Winkelhalbierende, der Teil danach für die jeweilige Kurve.

Um die Berechnung einfacher zu gestalten, macht es Sinn, die einzelnen Werte der Summe in der Tabelle der Lorenzkurve zu berechnen.

```
dat$g <- c(0, 0.5 * (dat[2:nrow(dat), 6] + dat[1:(nrow(dat) - 1),
    6]) * dat[2:nrow(dat), 7])
dat$g[1] <- 0.5 * dat$M_i[1] * dat$h_i[1]
dat <- round(dat, 4)
knitr::kable(dat, col.names = c("$i$", "$n_{i}$", "$x_{i}$", "$n_{i} \\cdot x_i$",
    "$MS_i = \\sum_{j=1}^i n_i \\cdot x_i$", "$M_i = MS_i / MS$",
    "$h_i$", "$H_i$", "$\\frac{1}{2} (M_{i-1} + M_i) h_i$"), bootstrap_options = c("striped",
    "hover", "condensed", "responsive"), full_width = F)
```

i	n_i	x_i	$n_i \cdot x_i$	$MS_i = \sum_{j=1}^{i} n_i \cdot x_i$	$M_i = MS_i/MS$	h_i	H_i	$\frac{1}{2}(M_{i-1} + M_i)h_i$
1	42	200	8400	8400	0.2366	0.3281	0.3281	0.0388
2	48	250	12000	20400	0.5746	0.3750	0.7031	0.1521
3	24	350	8400	28800	0.8113	0.1875	0.8906	0.1299
4	10	450	4500	33300	0.9380	0.0781	0.9687	0.0683
5	4	550	2200	35500	1.0000	0.0312	0.9999	0.0302

Der übersicht wegen habe ich die Tabelle hier nochmal komplett aufgenommen. In der Übung und der Klausur reicht es, wie auch in Kapitel 2 die Tabelle einmal komplett auszuschreiben. Für $\frac{1}{2}(M_{i-1} + M_i)h_i$ muss man immer den aktuellen Wert des Anteils an der Merkmalssumme und den Vorigen verwenden. In der ersten Zeile ist der vorige Wer $M_0 = 0$, da die Lorenzkurve im Punkt (0,0) startet. Folgend wird die Rechnung nochmal verdeutlich:

$$\text{Zeile } i = 1 : \frac{1}{2}(M_0 + M_1)h_1 = \frac{1}{2} \cdot (0 + 0.2366) \cdot 0.3281 = 0.0388$$
$$\text{Zeile } i = 2 : \frac{1}{2}(M_1 + M_2)h_2 = \frac{1}{2} \cdot (0.2366 + 0.5746) \cdot 0.3750 = 0.1521$$
$$\text{Zeile } i = 3 : \frac{1}{2}(M_2 + M_3)h_3 = \frac{1}{2} \cdot (0.5746 + 0.8113) \cdot 0.1875 = 0.1299$$
$$\text{Zeile } i = 4 : \frac{1}{2}(M_3 + M_4)h_4 = \frac{1}{2} \cdot (0.8113 + 0.9380) \cdot 0.0781 = 0.0683$$
$$\text{Zeile } i = 5 : \frac{1}{2}(M_4 + M_5)h_5 = \frac{1}{2} \cdot (0.9380 + 1) \cdot 0.0312 = 0.0302$$

Für den Gini muss man jetzt nur noch die obigen Werte zusammenrechnen und von 0.5 abziehen und dann noch mit der Inversen von K_{max} multiplizieren:

$$
\begin{aligned}
K_{max} &= \frac{128-1}{128 \cdot 2} = \frac{127}{256} \\
Gini &= K/K_{max} = (K_{max})^{-1} \cdot K \\
&= \frac{256}{127} \cdot (\frac{1}{2} - (0.0388 + 0.1521 + 0.1299 + 0.0683 + 0.0302)) \\
&= \frac{256}{127} \cdot (\frac{1}{2} - 0.4193) = \frac{256}{127} \cdot 0.0807 \\
&= 0.1627
\end{aligned}
$$

Tipp bezüglich K_{max}: Per Konstruktion ist K_{max} immer etwas kleiner als 0.5. Da der Zähler nur um 1 verringert wird und der Nenner mal zwei genommen wird. Dadurch ist der Nenner immer ein kleines bisschen größer als doppelt so groß. Daraus ergibt, dass die Umkehrung (oder Inverse) von K_{max}, $(K_{max})^{-1}$, immer knapp größer als 2 sein sollte.

Interpretation des Gini-Koeffizienten

Tipp: Je größer die Fläche zwischen Lorenzkurve und Winkelhalbierende ist, desto größer ist der Gini. Eine große Fläche sollte mit einem klaren Knick der Lorenzkurve einhergehen. Beides sind ein Zeichen dafür, dass es eine große Ungleichheit gibt. Dies bedeutet, dass es eine Konzentration bei (prozentual gesehen) wenigen Individuen gibt. D.h., eine große Konzentration bedeutet eine große Ungleichheit, eine große Fläche zwischen Lorenzkurve und Winkelhalbierende und einen großen Gini.

Kapitel 5

Formelle Schreibweise der Randhäufigkeit

$n_{ij} = H_{abs}(X = x_i \cap Y = y_j)$

$n_{i\cdot} = H_{abs}(X = x_i \cap Y = \sum_{j=1}^{l} y_j) = \sum_{j=1}^{l} n_{ij} = (n_{i1} + n_{i2} + \cdots + n_{il}) = n_{i\cdot}$

Berechnung der Randverteilung

Tipp: Die Randverteilung $n_{i\cdot}$ und $n_{\cdot j}$ sind die Summen der zugehörigen Spalten oder Zeilen

```
dat <- data.frame(y_1 = c(56, 33), y_2 = c(78, 91))
rownames(dat) <- c("x_1", "x_2")
dat <- rbind(dat, colSums(dat, dims = 1))
dat <- cbind(dat, rowSums(dat, dims = 1))
rownames(dat)[3] <- "Summe"
colnames(dat)[3] <- "Summe"
dat
```

	y_1	y_2	Summe
x_1	56	78	134
x_2	33	91	124
Summe	89	169	258

Notiz: Die Zahlen sind für dieses Beispiel zufällig ausgedacht, da es hier nur um die Berechnung der Randhäufigkeiten geht. Die Zahlen 1 und 2 sind jeweils zwei Merkmalsausprägungen, die nicht weiter definiert sind.

Man addiert also wie folgend:

$\sum_{j=1}^{2} n_{1j} = n_{11} + n_{12} = 56 + 78 = 134 = n_{1\cdot}$

i ist hier mit 1 festgesetzt und die Summe geht für j von 1 bis 2. Somit müssen diese zwei Werte zusammen gerechnet werden. Am Ende muss die Randhäufigkeit unabhängig von j sein, also einen Punkt im Index haben. Der Grund hierfür ist, dass wir über all j aufaddiert haben.

$\sum_{j=1}^{2} n_{2j} = n_{21} + n_{22} = 33 + 91 = 124 = n_{2\cdot}$

$\sum_{i=1}^{2} n_{i1} = n_{11} + n_{21} = 56 + 33 = 89 = n_{\cdot 1}$

$\sum_{i=1}^{2} n_{i2} = n_{12} + n_{22} = 78 + 91 = 169 = n_{\cdot 2}$

Berechnung von Kontingenztabellenzellen bei gegebenen Randhäufigkeiten

Man kann natürlich die Rechnung auch *umdrehen* und anhand der Randhäufigkeiten auf die Kontingenzzellen schließen. Sei folgende Tabelle gegeben:

```
dat <- data.frame(y_1 = c(56, 33, 4), y_2 = c(78, 91, 1), y_3 = c(34,
    56, 67))
rownames(dat) <- c("x_1", "x_2", "x_3")
dat <- rbind(dat, colSums(dat, dims = 1))
dat <- cbind(dat, rowSums(dat, dims = 1))
rownames(dat)[4] <- "Summe"
colnames(dat)[4] <- "Summe"
dat[3, 1] <- "."
dat[2, 2] <- "."
dat[1, 3] <- "."
dat[1, 4] <- "."
```

```
dat[4, 2] <- "."
dat
```

	y_1	y_2	y_3	Summe
x_1	56	78	.	.
x_2	33	.	56	180
x_3	.	1	67	72
Summe	93	.	157	420

Die Punkte stehen hierbei für fehlende Werte.

Da $n_{.1} = 93 = n_{11} + n_{21} + n_{31}$ ist, und n_{31} unbekannt ist, muss man die Gleichung umstellen. $n_{31} = n_{.1} - n_{11} - n_{21} = 93 - 33 - 56 = 4$

Selbiges gilt für die weiteren Werte:

$n_{13} = n_{.3} - n_{23} - n_{33} = 157 - 67 - 56 = 34$

$n_{1.} = n - n_{3.} - n_{2.} = 420 - 72 - 180 = 168$

$n_{.2} = n - n_{.1} - n_{.3} = 420 - 93 - 157 = 170$

$n_{22} = n_{.2} - n_{12} - n_{32} = 170 - 1 - 78 = 91$

Fragestellung für bedingte Verteilung

(i) Wie viele Prozent haben höchstens ein PKW, **bedingt drauf**, dass sie zu einem 3-Personen-Haushalt gehören?

(ii) Wie viele Prozent **der** 3-Personen-Haushalte haben höchstens einen PKW?

Bei der Fragestellung 'Wie viel Porzent haben höchstens ein PKW **und** gehören zu einem 3-Personen-Haushalt wird nach der unbedingten Verteilung gefragt.

Unbedingte und bedingte relative Verteilung

Tipp: Sollte nach der relativen Verteilung gefragt sein, handelt es sich immer um die unbedingte Verteilung

$$h_{ij} = n_{ij}/n \qquad .$$

Nur wenn extra das Wort *bedingt* verwendet wird, muss die bedingte Verteilung

$$H_{rel}(X = x_i|Y = y_j) = h_{i|y_j} = n_{ij}/n_{.j}$$

oder vertauscht

$$H_{rel}(Y = y_j|X = x_i) = h_{j|x_i} = n_{ij}/n_{i.}$$

berechnet werden. Man teilt durch die Randverteilung des Merkmals, auf das bedingt wird.

Man beachte, dass der Index der Bedingung auch der Index ist, der bei der Teilgesamtheit steht, durch die man teilt.

Bedingt

Tipp: Bedingt man auf etwas, bedeutet das, dass man durch die Bedingung dividiert. Somit zeigt der Index nach dem Bedingtzeichen |, durch was geteilt werden muss.

Bedingte relative Verteilung für zwei Merkmale

Tipp: In der Übung 7 gab es die Aufgabe bezüglich der bedingten Verteilung des Frauenanteils für zwei Kategorien. Hier wäre die Formale Schreibweise wie folgt:

$$H_{rel}(Y = y_1 + y_2 | X = x_1 + x_2) = \frac{n_{12} + n_{22}}{n_{1.} + n_{2.}}$$

Man dividiert durch die Randverteilungen des Einkommens X für die ersten beiden Zeilen, da diese die Zeilen für Geringverdiener sind. Selbiges gilt für den Zähler, dass dort die beiden Merkmalsausprägungen addiert werden müssen.

Richtige Schreibweise der bedingten Verteilung und des bedingten Mittelwertes

Tipp: Bezogen auf die Übung 7, Aufgabe 2 a. Dort gibt es die Tabelle bezüglich des Alters und die Anzahl an Kinder. Welche untere Schreibweise gewählt wird, ist egal, eine dieser sollte aber gewählt werden, bevor das Ergebnis angegeben wird. Die Anzahl der Kinder sei beschrieben mit Y und geht von 0 bis 3. Das Alter ist beschrieben mit X und hat als dritten Eintrag die 37. Dann gibt es folgende Möglichkeiten die bedingte Verteilung aufzuschreiben:

$$H_{rel}(Y = 0 | X = 37) = H_{rel}(Y = y_1 | X = x_3) = h_{1|x_3}$$

Für den Mittelwert gibt es folgende Schreibweisen:

$$\overline{y}|(X = 37) = \overline{y}_{|x_3}$$

Mindestens eine dieser beiden Schreibweisen für den Mittelwert von Y, bedingt darauf, dass das Merkmal X gleich 37 ist, muss in der Klausur aufgeschrieben werden.

Wann ist eine Korrelation nahe 0?

Tipp: Hierfür gibt es keine klare Regel. Vorlesungskonvention wird es sein, dass eine Korrelation klein ist, wenn der absolute Wert der Korrelation $|r_{XY}| < 0.1$ ist. Alles darüber ist eine mittlere oder starke Korrelation (ab 0.7).

Anzahl an Individuen bei statistische Unabhängigkeit

Die Formel für statistische Unabhängigkeit geht wie folgt:

$$h_{ij} = h_{i.} \cdot h_{.j}$$

Die Multiplikation der Randverteilungen gibt die relative Häufigkeit bei statistischer Unabhängigkeit an. Um nun berechnen, wie viele Individuen es sein müssen, damit die Verteilung unabhängig ist, muss die Multiplikation der Randverteilungen noch mit der Gesamtzahl n multipliziert werden, um von der relativen Häufigkeit auf die absolute Häufigkeit zu kommen:

$$h_{i.} \cdot h_{.j} \cdot n = \text{ Anzahl an Individuen bei statischer Unabhängigkeit}$$

Invarianz

Tipp: Die Invarianz bezieht sich auf die Skaleninvarianz. Es geht hierbei darum, ob ein Maß gleich bleibt, wenn die Variablen die zur Berechnung benutzt wurden, geändert /umskaliert werden. Das Konzept ist in Kapitel 3 schon aufgetreten. Die Invarianz bezieht sich auf die Verschiebungseigenschaft und die Homogenität eines Maßes. \
Mittelwert:

$$
\begin{aligned}
&y_i = x_i + a, \qquad i = 1, \ldots, n \quad \rightarrow \quad \overline{y} = \overline{x} + a \\
&y_i = b \cdot x_i, \qquad i = 1, \ldots, n \quad \rightarrow \quad \overline{y} = b \cdot \overline{x} \\
&\quad \text{Zusammengefügt ergibt dies dann} \\
&y_i = b \cdot x_i + a, \quad i = 1, \ldots, n \quad \rightarrow \quad \overline{y} = b \cdot \overline{x} + a
\end{aligned}
$$

Varianz:

$$
\begin{aligned}
&y_i = x_i + a, \qquad i = 1, \ldots, n \quad \rightarrow \quad s_Y^2 = s_X^2 \\
&y_i = b \cdot x_i, \qquad i = 1, \ldots, n \quad \rightarrow \quad s_Y^2 = b^2 \cdot s_X^2 \\
&\quad \text{Zusammengefügt ergibt dies dann} \\
&y_i = b \cdot x_i + a, \quad i = 1, \ldots, n \quad \rightarrow \quad s_Y^2 = b^2 \cdot s_X^2
\end{aligned}
$$

Korrelation:

$$
\left.
\begin{aligned}
U &= a_1 + b_1 X, \quad b_1 \neq 0 \\
V &= a_2 + b_2 Y, \quad b_2 \neq 0
\end{aligned}
\right\} \Longrightarrow r_{UV} = r_{XY}
$$

Bei der Korrelation muss man sich, da es ein Maß für den Zusammenhang zweier Variablen ist, angucken, wie sich das Maß ändert, wenn zwei Variablen geändert werden. Hier ist der Fall, dass wenn man X und Y um eine beliebige Konstante und einen beliebigen Skalar ungleich Null ändert, die Korrelation gleich bleibt.

Rangkorrelation

Bei der Rankkorrelation wird die Korrelation aufgrund der Ränge berechnet, nicht aufgrund der tatsächlichen (hier:) Noten. Sobald man die Ränge aufgestellt hat, sollte man deswegen im besten Fall die Noten durchstreichen, damit man sie nicht doch in die Rechnungen aufnimmt.

```
i <- c("ter Stegen", "Ginter", "Mustafi", "Rüdiger", "Hector", "Kimmich",
    "Rudy", "Goretzka", "Draxler", "Stindl", "Werner")
Kicker <- c(3.5, 2.5, 3, 3, 3, 3.5, 3.5, 3, 2.5, 2.5, 2)
Sport1 <- c(2.5, 3.5, 2.5, 2, 3.5, 2, 2, 3, 2, 2, 2)
RangKicker_1 <- rank(Kicker, ties.method = "first")
RangKicker_2 <- rank(Kicker, ties.method = "average")
RangSport1_1 <- rank(Sport1, ties.method = "first")
RangSport1_2 <- rank(Sport1, ties.method = "average")
n <- length(Kicker)  # Man könnte hier auch jede andere Variable nehmen
dat <- data.frame(i, Kicker, Sport1, RangKicker_1, RangKicker_2, RangSport1_1,
    RangSport1_2)
dat
```

i	Kicker	Sport1	RangKicker_1	RangKicker_2	RangSport1_1	RangSport1_2
ter Stegen	3.5	2.5	9	10.0	7	7.5
Ginter	2.5	3.5	2	3.0	10	10.5
Mustafi	3.0	2.5	5	6.5	8	7.5

i	Kicker	Sport1	RangKicker_1	RangKicker_2	RangSport1_1	RangSport1_2
Rüdiger	3.0	2.0	6	6.5	1	3.5
Hector	3.0	3.5	7	6.5	11	10.5
Kimmich	3.5	2.0	10	10.0	2	3.5
Rudy	3.5	2.0	11	10.0	3	3.5
Goretzka	3.0	3.0	8	6.5	9	9.0
Draxler	2.5	2.0	3	3.0	4	3.5
Stindl	2.5	2.0	4	3.0	5	3.5
Werner	2.0	2.0	1	1.0	6	3.5

Die Variablen RangKicker_1 und RangSport1_1 geben den Rank von oben nach unten vor. Hierbei werden Spieler mit der gleichen Note einfach im Ranking weiter gezählt. Aus diesen Spalten ergeben sich dann die Spalten RangKicker_2 und RangSport1_2. Hier werden die Ränge der Spieler, die die selbe Benotung erhalten haben, gemittelt. Für die Rechnungen wird *Kicker* mit K abgekürzt und *Sport1* mit S. Welche Abkürzung man wählt, ist aber zweitrangig. Man könnte auch einfach bei X und Y bleiben.

Beim *Kicker* ist die beste Note eine 2 und kommt einmal vor. Die nächstbeste Note ist die 2.5 welche 3 mal vertreten ist. Somit wird aus dem Rang 2, 3 und 4 für die Spieler Ginter, Draxler und Stindle der Rang zur 3, dem Durchschnitt der Ränge. Selbiges zieht sich für *Kicker* und *Sport1* durch.

Die Formel für die Rangkorrelation ist wie folgend:

$$r_{XY}^{Sp} = r_{rg(X),rg(Y)} = \frac{c_{XY}^{SP}}{s_X \cdot s_Y} = \frac{\frac{1}{n}\sum_{j=1}^{n} rg(x_j) \cdot rg(y_j) - \overline{rg}_X \cdot \overline{rg}_Y}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} rg(x_j)^2 - \overline{rg}_X^2 \cdot \frac{1}{n}\sum_{j=1}^{n} rg(y_j)^2 - \overline{rg}_Y^2}}$$

$$= \frac{\overline{rg}_{XY} - \overline{rg}_X \cdot \overline{rg}_Y}{\sqrt{\overline{rg^2}_X - \overline{rg}_X^2 \cdot \overline{rg^2}_Y - \overline{rg}_Y^2}}$$

$$\text{mit } \overline{rg}_X = \overline{rg}_Y = \frac{n+1}{2}$$

Um jetzt die Rangkorrelation nach Spearman ausreichen zu können, wird die Rechnung in einzelne Teile vorgerechnet.

Als erstes wird der Zähler berechnet. Es sollte klar sein, dass der Zähler die Kovarianz ist, nur das hier nicht die eigentlichen Ausprägungen genommen werden, sondern die Ränge des Merkmals. Wenn Sie schon berechnet hat, kann man sie ja auch verwenden. Es folgt als erstes die Berechnung in *R*, dann die normale Textansicht.

```
RangKS_mean <- round(sum(RangKicker_2 * RangSport1_2)/length(RangKicker_2),
    4)
RangKS_cov <- round(RangKS_mean - ((length(RangKicker_2) + 1)/2) *
    ((length(RangKicker_2) + 1)/2), 4)
RangKicker_sd <- round(sqrt(sum(RangKicker_2^2)/length(RangKicker_2) -
    ((length(RangKicker_2) + 1)/2)^2), 4)
RangSport1_sd <- round(sqrt(sum(RangSport1_2^2)/length(RangSport1_2) -
    ((length(RangSport1_2) + 1)/2)^2), 4)
RangKS_cor <- round(RangKS_cov/(RangKicker_sd * RangSport1_sd), 4)
RangKS_mean
```

```
## [1] 36.2955
```

```
RangKS_cov
```

```
## [1] 0.2955
```

```
RangKicker_sd
```

```
## [1] 3.0302
```

```
RangSport1_sd
```

```
## [1] 2.8841
```

```
RangKS_cor
```

```
## [1] 0.0338
```

Für die Kovarianz wird erst der jeweilige Mittelwert berechnet:

Der Mittelwert ist beim Rangkorrelation immer genau gleich:

$\overline{rg_K} = \overline{rg_S} = \frac{n+1}{2} = 6$

Als nächstes kommt der erste Teil des Zählers, $\overline{rg}_{KS}$ dran.

$\overline{rg}_{KS} = \frac{1}{n} \sum_{j=1}^{n} rg(k_j) \cdot rg(s_j) = \frac{1}{11}$ $(10 \cdot 7.5 + 3 \cdot 10.5 + 6.5 \cdot 7.5 + 6.5 \cdot 3.5 + 6.5 \cdot 10.5 + 10 \cdot 3.5 + 10 \cdot 3.5 + 6.5 \cdot 9 + 3 \cdot 3.5 + 3 \cdot 3.5 + 1 \cdot 3.5)$

$= \frac{1}{11}$ $(75 + 31.5 + 48.75 + 22.75 + 68.25 + 35 + 35 + 58.5 + 10.5 + 10.5 + 3.5)$

$= \frac{1}{11}$ (399.25)

$= 36.2954545$

$c_{KS}^{SP} = \overline{rg}_{KS} - \overline{rg}_K \cdot \overline{rg}_S = 36.2954545 - 36$

$= 0.2954545$

Als nächstes muss der Nenner berechnet werden.

$\overline{rg^2}_K = \frac{1}{11}$ $((10)^2 + (3)^2 + (6.5)^2 + (6.5)^2 + (6.5)^2 + (10)^2 + (10)^2 + (6.5)^2 + (3)^2 + (3)^2 + (1)^2)$

$= \frac{1}{11} \cdot$ (497)

$= 45.1818182$

$s_K = \sqrt{45.1818182 - 36}$

$= 3.0301515$

$\overline{rg^2}_S = \frac{1}{11}$ $((7.5)^2 + (10.5)^2 + (7.5)^2 + (3.5)^2 + (10.5)^2 + (3.5)^2 + (3.5)^2 + (9)^2 + (3.5)^2 + (3.5)^2 + (3.5)^2)$

$= \frac{1}{11} \cdot$ (487.5)

$= 44.3181818$

$s_S = \sqrt{44.3181818 - 36}$

$= 2.8841258$

$\Rightarrow \frac{c_{KS}^{SP}}{s_K \cdot s_S} = \frac{0.2955}{3.0302 \cdot 2.8841} = 0.0338$

Statistische Unabhängigkeit

Die Formel für die statistische Unabhängigkeit ist $h_{ij} = h_{i.} \cdot h_{.j}$. Weiter wurde gesagt, dass wenn zwei Variablen statistisch unabhängig sein, sie keine Kovarianz und dadurch keine Korrelation haben. Egal ob das Merkmal metrisch oder ordinal ist. (Gilt auch für den Kontingenzkoeffizienten und nominalen Merkmalen.)

Hat man also in einer Aufgabe schon berechnet, dass die Kovarianz oder Korrelation ungleich 0 ist, muss man die obige Rechnung zur Überprüfung nicht mehr durchführen! Man kann direkt mit der Korrelation ungleich 0 argumentieren.

Unabhängigkeit

Wenn es eine Korrelation gibt, sind zwei Merkmale nicht unabhängig, also abhängig. Somit muss, sobald man eine Korrelation ausgerechnet hat, die Unabhängigkeit nicht mehr Formal überprüft werden. Man kann direkt mit der Korrelation argumentieren.

Abhängigkeit

Erst wenn der Kontingenzkoeffizient gleich 0 ist, spricht man von einer fehlenden Abhängigkeit. Werte sind nach Vorlesungskonvention gleich 0, wenn die 4. Kommastelle gleich Null ist. Sollte also ein $K^* = 0.0001$ herauskommen, sprechen wir von einer sehr schwachen Abhängigkeit.

Berechnung des Mittelwertes von Y

Wenn zwei Variablen X und Y gegeben sind und man den Mittelwert (und andere Maße) berechnen müssen, sollte man bei den Namen bei dem jeweiligen Buchstaben bleiben. Dies bedeutet, dass der Mittelwert von $X = \overline{x}$ ist. Der Mittelwert von $Y = \overline{y}$ Für die Berechnungen nimmt man immer den kleinen Buchstaben der Abkürzung des Merkmales.

Die quadratische Kontingenz χ^2

Die quadratische Kontingenz ist laut Formular eine quadratische Maßzahl. Sollte also ein Wert rauskommen, z.B. $\chi^2 = 3.3259$, muss man beim Einsetzen für K^* diesen Wert **nicht** quadrieren!

Korrigierter Kontingenzkoeffizient

K^* beschreibt nur eine Abhängigkeit, keine Korrelation. Der Unterschied daran ist, dass man der Korrelation eine Richtung geben kann. Es ist somit ein Schritt mehr Information als die Abhängigkeit.

Da K^* nur eine Abhängigkeit angibt, ist es zwischen 0 und 1 beschränkt: $0 \leq K^* \geq 1$.

Bivariate Verteilung

```
height_i <- round(c(165, 167, 180, 164, 172, 160, 185, 180, 177, 190,
    168, 175, 180, 184, 175, 180, 194, 173, 179, 179, 170, 178, 180,
    164, 182, 180, 163, NA, 163, 161, 160, 168, 175, 173, 171, 153,
    162, 160, 173, 177, 171, 158, 180, 170, 160, 170, 172, 175, 190,
    169, 174, 167, 188, 173, 182, 189, 164, 162, 184, 178, 185, 182,
    177, 179, 160, 174, 173, 169, 168, 182, 165, 170, 165, 182, 182,
    173, 164, 166, 184, 167, 184, 168, 189, 172, 172, 175, 169, 174,
    158, 160, 175, 168, 163, 163, 168, 163, 172, 172, 167, 170, 164,
    166, 182, 191, 172, 160, 150, 180, 182, 164, 160, 187, 175, 170,
    195, 158, 185, 180, 173, 159, 170, 160, 175, 163, 187, 182, 180,
    167, 179), -1)
weight_j <- round(c(52, 64, 73, 68, NA, 43, 74, 75, 61, 82, 74, 65,
    86, 74, 95, 81, 96, 70, 53, 80, NA, 64, 94, 65, 80, 71, 49, NA,
    65, 70, 60, 60, 65, 64, 72, 60, 62, NA, 57, 74, 68, 54, 70, NA,
    62, 53, 62, 82, 71, 68, 58, 60, 125, 70, 68, 78, 62, 51, 68, 65,
    85, 96, 72, 65, 63, 60, 69, 52, 57, 79, 100, 65, 80, 81, 70, 75,
    65, 63, 101, 66, 74, 55, 85, NA, 59, 65, 52, NA, 60, 65, 62, 54,
    65, 67, 58, 80, 60, 65, 53, 58, 53, 62, 75, 87.5, 65, 65, 52,
    69, 70, 70, 60, 63, 60, 65, 82, 54, 70, 72, 69, 48, 68, 80, 72,
    70, 75, 84, 110, NA, 73), -1)
dat <- data.frame(height_i, weight_j)
dat <- dat[complete.cases(dat), ]
```

```r
tab <- addmargins(table(height_i, weight_j))
tab
```

```
##         weight_j
## height_i  40  50  60  70  80  90 100 110 120 Sum
##      150   0   1   1   0   0   0   0   0   0   2
##      160   1   7  13   5   3   0   1   0   0  30
##      170   0   5  20  10   1   0   0   0   0  36
##      180   0   1   9  17  10   2   3   1   0  43
##      190   0   0   1   1   4   1   1   0   1   9
##      200   0   0   0   0   1   0   0   0   0   1
##      Sum   1  14  44  33  19   3   5   1   1 121
```

Die Tabelle zeigt eine Kontingenztabelle mit Gewicht (*weight*) und Körpergröße (*height*).

Wie viele Menschen wiegen 60 kg und und sind 170 cm groß?

$n_{ij} = n_{33} = 20$

Wie viele Menschen wiegen hier 60 kg?

$n_{\cdot j} = n_{\cdot 3} = 44$

Wie viele Menschen sind 170 cm groß?

$n_{i\cdot} = n_{3\cdot} = 36$

Wie viele Menschen wiegen 60 kg, bedingt darauf, dass sie 170 cm groß sind?

$h_{j|x_i} = h_{3|x_3} = n_{33}/n_{3\cdot} = 20/36 = 0.5556$

Wie viele Menschen sind 170 cm groß, bedingt darauf, dass sie 60 kg wiegen?

$h_{i|y_j} = h_{3|y_3} = n_{33}/n_{\cdot 3} = 20/44 = 0.4545$

Was ist der Mittelwert des Gewichts, bedingt darauf, dass die Individuen 170 cm groß sind?

$\overline{y}_{|x_3} = \frac{1}{n_{3\cdot}} \sum_{j=1}^{l} n_{3j} \cdot y_j = \sum_{j=1}^{l} h_{j|x_3} y_j$

Erste Formel:

$$\begin{aligned}
\frac{1}{n_{3\cdot}} \sum_{j=1}^{l} n_{3j} \cdot y_j &= \frac{1}{36} \cdot (0 \cdot 40 + 5 \cdot 50 + 20 \cdot 60 + 10 \cdot 70 + 1 \cdot 80 + 0 \cdot 90 + 0 \cdot 100 + 0 \cdot 110 + 0 \cdot 120) \\
&= \frac{1}{36} \cdot (0 + 250 + 1200 + 700 + 80 + 0 + 0 + 0 + 0) \\
&= \frac{1}{36} \cdot (2230) \\
&= 61.9444
\end{aligned}$$

Zweite Formel:

$$\begin{aligned}
\sum_{j=1}^{l} h_{j|x_3} y_j &= \left(\frac{0}{36} \cdot 40 + \frac{5}{36} \cdot 50 + \frac{20}{36} \cdot 60 + \frac{10}{36} \cdot 70 + \frac{1}{36} \cdot 80 + \frac{0}{36} \cdot 90 + \frac{0}{36} \cdot 100 + \frac{0}{36} \cdot 110 + \frac{0}{36} \cdot 120 \right) \\
&= 0 + 6.9444444 + 33.3333333 + 19.4444444 + 2.2222222 + 0 + 0 + 0 + 0 \\
&= 61.9444
\end{aligned}$$

Normalerweise bevorzuge ich immer die Formeln mit absoluten Häufigkeiten. Bei den bedingten Mittelwerten, wenn man vorher schon die bedingten relativen Häufigkeiten berechnen musste, würde ich zu der zweiten Formel raten um Zeit zu sparen und um nicht mehr darauf achten zu müssen, welche genauen absoluten Werte benutzt werden müssen.

Was ist der Mittelwert der Körpergröße, bedingt darauf, dass die Individuen 60 cm wiegen?

$\overline{x}_{|y_3} = \frac{1}{n_{\cdot 3}} \sum_{i=1}^{k} n_{i3} \cdot x_i = \sum_{i=1}^{k} h_{i|y_3} x_i = 169.0909$

Was ist die Varianz der Körpergröße, bedingt darauf, dass die Individuen 60 cm wiegen?

$s^2_{x_{|y_3}} = \overline{x^2_{|y_3}} - \overline{x_{|y_3}}^2 = 3886.1111 - (61.9444)^2 = 49.0024086$

Wie viele Studierende sind kleiner als 170?

$\sum_{i=1}^{2} n_{i\cdot} = 32$

Wie viele Prozent der Studierende mit 80 kg sind größer als 180 cm groß?

(4+1)/19

Wie viele Prozent der Studierende mit mehr als und inbegriffen 80 kg sind größer als 180 cm groß?

(4+1+1+1+1)/(19+3+5+1+1)

Kapitel 6

Regression

KQ steht für *Kleinste Quadrate*. Hiermit bezieht man sich auf die Minimierung der quadratischen Abweichungen e_i. Da das Quadrat von e_i, e_i^2 ein Quadrat ergibt, wird nicht einfach nur die Abweichung e_i minimiert, sondern eben das Quadrat von e_i. Diese Minimumierung wird dann aufsummiert, wodurch man auf die Summe der quadratischen Abweichungen (SQA) kommt. *KQ* und *SQA* meinen somit das Gleiche. In der Literatur wird der Begriff *KQ* eher verwendet.

Regressionsgerade:

$$y_i = a + b \cdot x_i + e_i$$

Gefittete Werte:

$$\hat{y}_i = a + b \cdot x_i$$

Schreibt man, ob mit eingesetzten Parametern oder nicht,

$$y_i = a + b \cdot x_i,$$

ist dies falsch!

Vermischung höheres Polynom und multiple Regression: Sei Y der Lohn eines Individuums, X_1 die Erfahrung und X_2 die Anzahl an Schulbildung (in Jahren). Folgende Regressionsmodel ist dann m?glich:

$$y = a + b_{11}x_1 + b_{12}x_1^2 + b_{21}x_2 + b_{22}x_1^2$$

Regressionsgleichung

Die Regressionsgleichung hat immer die gleiche Form. Der Standard ist hierbei:

$$y_i = a + b \cdot x_i + e_i \quad .$$

y_i ist die abhängige Variable, die zu erklären ist. Sie ist abhängig, weil Sie von der rechten Seite der Gleichung beeinflusst wird.

x_i ist die unabhängige Variable, die y_i erklären soll. Sie ist unabhängig, weil es für Sie keine eine Gleichung gibt, die sie erklärt.

a ist der Achsenabschnitt, bei dem die Regressionslinie die y-Achse bei $x = 0$ überschreitet.

b ist die Steigung der Regressionsgleichung. Sie besagt, wenn man bei der x-Achse einen Schritt nach rechts geht, wie weit man nach oben oder nach unten gehen muss.

$a + b \cdot x_i = \hat{y}_i$, genannt die gefitteten Werte. Sie repräsentieren die Werte die Y hätte, wenn es keine Abweichungen geben. Somit liegen die Punkte genau auf der Regressionslinie, senkrecht unter den (wahren) Y-Werten.

e_i ist die Abweichung, die zwischen y_i und $\hat{y}_i$ liegt. Hierdurch wird erlaubt eine Gerade zu zeichnen, auch wenn die (wahren) Punkte gar nicht genau auf einer Gerade liegen. Vergisst man diesen Teil der Gleichung, bedeutet es, dass alle Werte perfekt auf einer Geraden liegen, was in der Realität unrealistisch ist. e_i wird auch als Störterm, Noise, oder Innovation bezeichnet, je nachdem in dem welchem fachlichen Bereich man sich befindet.

Erklärung einer Steigung

Tipp: Man muss die Einheiten der jeweiligen Variablen bei der Interpretation beachten. An sich steht b für die Steigung von X von einer Einheit für einen Änderung (Antieg oder Abstieg) um b bei Y. Diese eine Einheit ist aber abhängig davon, wie die jeweiligen Variablen gemessen sind. Wenn eine Variable z.B. in Tausend Euro gemessen ist, also eine Einheit = 1000 Euro. Dann muss dies in der Interpretation auch berücksichtigt werden. Als eine Steigung um 1000 Euro von X wird mit einer Steigung um b in Y beobachtet.

Zeichnung der Regressionsgerade

Tipp 1: Die gefittete Werte liegen immer senkrecht unter- oder oberhalb der wahren Werte auf der Regressionsgerade. Dies liegt daran, dass die quadratische Abweichung e_i^2 zwischen y_i und $\hat{y}_i$ so gering wie möglich gehalten wird.

Tipp 2: Die x-Achse und y-Achse sollten immer gleich skaliert werden.

Tipp 3: Man kann die (X, Y)-Werte eintragen, die gefitteten Werte ausrechnen, diese mit aufnehmen und die Regressionserade zeichnen. Da man allerdings weiß, dass die Regressionsgerade immer durch den Mittelpunkt $(\overline{x}, \overline{y})$ verläuft, kann man diesen eintragen und eine Gerade mithilfe des Achsenabschnitts zeichnen.

Stata-Ouput

Der *Stata*-Output wird in einem zusätzlichem Dokument erklärt. Für Statistik I ist nur wichtig, dass man weiß, wo man a, b, R^2 und n ablesen kann. Der Output in R sieht (leider) etwas anders aus, wobei die Koeffizienten sehr ähnlich präsentiert werden.

Beispielhaft ein *R*-Output:

```
set.seed(2101990)
x <- rnorm(20, 0, 1)
a <- 5
b <- 3
e <- rnorm(20, 0, 1)
dat <- data.frame(xvar = x, yvar = a + b * x + e)
reg <- lm(data = dat, yvar ~ xvar)
summary(reg)
```

```
##
## Call:
## lm(formula = yvar ~ xvar, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3792 -0.8799  0.1086  0.6309  2.1723
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5785     0.2250   20.35 7.13e-14 ***
## xvar          2.7839     0.2168   12.84 1.68e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9933 on 18 degrees of freedom
## Multiple R-squared:  0.9016, Adjusted R-squared:  0.8961
## F-statistic: 164.9 on 1 and 18 DF,  p-value: 1.682e-10
```

```
cor(dat$xvar, dat$yvar)^2
```

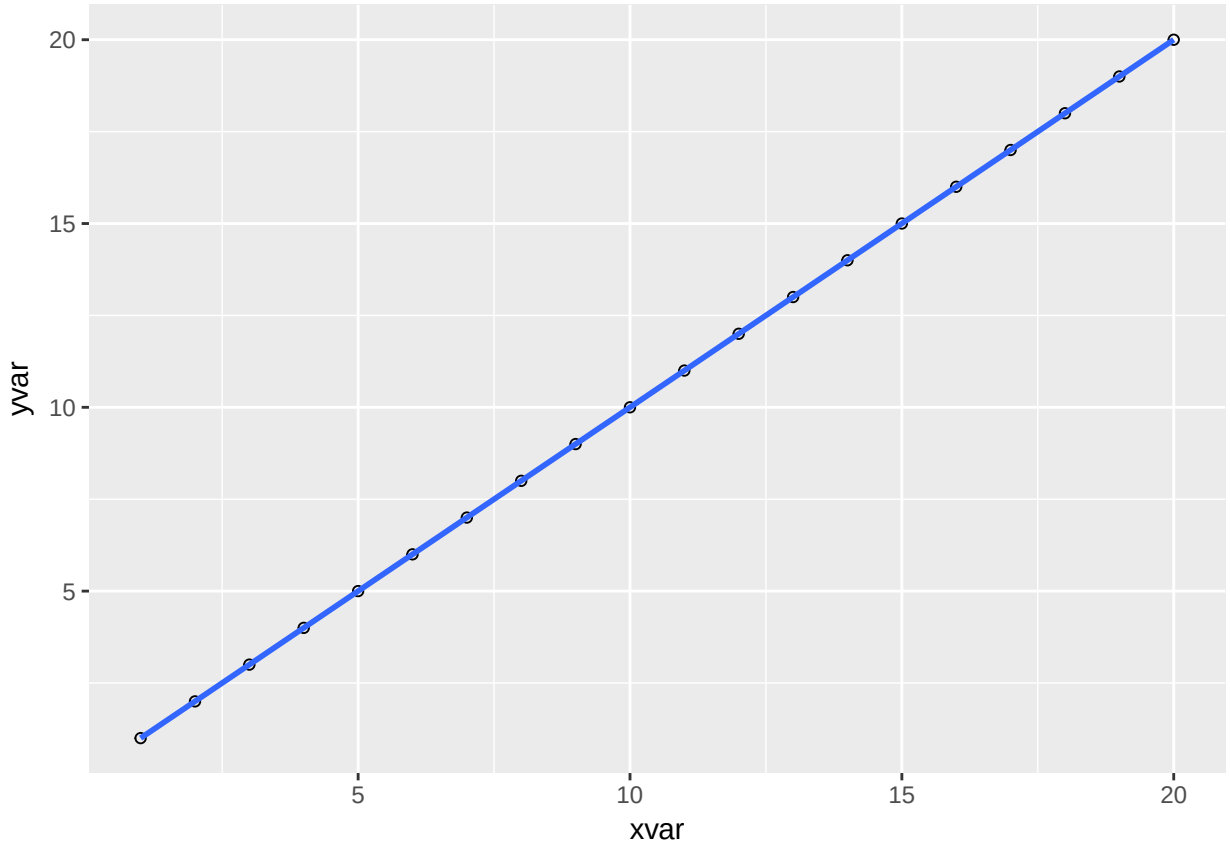
```
## [1] 0.9015912
```

Der Steigungskoeffizient b ist hier gleich 2.7838922, der Achsenabschnitt ab gleich 4.5784827 und das Bestimmtheitsmaß ist gleich 0.9016 und im Output unter *Multiple R-squared* zu finden.

Erklärungsgehalt einer Regression anhand von drei Beispielen mit unterschiedlichen Abweichungen

Tipp: Der Erklärungsgehalt einer Regression wird durch das Bestimmtheitsmaß R^2 bestimmt. Der Erklärungsgehalt bezieht sich darauf, wie viel von der Varianz der gefitten Werte $s^2_{\hat{y}}$ die Varianz der wahren Werte s^2_y erklären kann: $R^2 = \frac{s^2_{\hat{y}}}{s^2_y}$. Was bedeutet es, dass man eine Variation mit einer anderen Variation erklärt? Hierbei geht es um die Gleichheit der gefitteten und den wahren Werten.

```
set.seed(02101990)
x <- seq(1,20,1)
a <- 0
b <- 1
e <- 0
dat <- data.frame(xvar = x,
                  yvar = a + b*x + e) # xvar = yvar
ggplot(dat, aes(x=xvar, y=yvar)) +
    geom_point(shape=1) + # Use hollow circles
    geom_smooth(method=lm, # Add linear regression line
                se=FALSE) # Don't add shaded confidence region
```




```
reg <- lm(dat$yvar ~ dat$xvar)
```

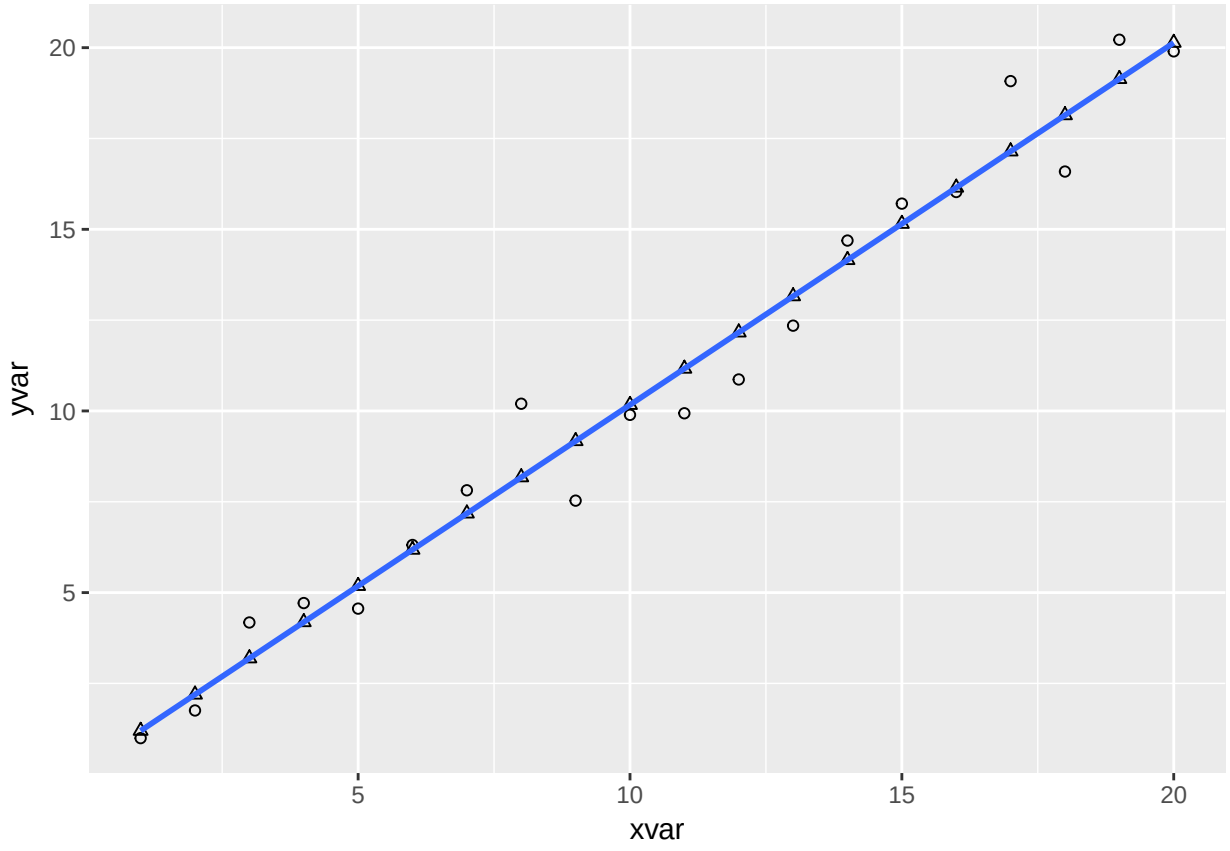
In dem obigen Beispiel sind die Werte auf einer Linie mit der Steigung b gleich 1. Die Korrelation ist gleich 1, da es sich bei den Variablen X und Y um die gleichen Variablen handelt. Somit ist auch das Verhältnis der Standardabweichungen gleich. Hier kann man X perfekt mit Y erklären (was kein Wunder ist, da Y und X genau gleich sind). Je weiter

sich jetzt X von Y entfernt, desto weniger wird man Y mit X erklären können. Da $\hat{y} = a + b \cdot x$ ist, sinkt einer Ekrlärung von X die Erklärung von $\hat{Y}$. In der obigen Gleichung ist somit $y_i = a + b \cdot x + e_i$ mit $e_i = 0$. Somit ist $y_i = a + b \cdot x = \hat{y}_i$. Dies ist ein seltener Spezialfall! Es sollte klar sein, dass hier das Bestimmtheitsmaß gleich 1 ist.

Wenn das Bestimmtheitsmaß gleich 1 ist, muss somit die Variation von Y (gemessen durch s^2_Y) und von $\hat{Y}$ (gemessen durch $s^2_{\hat{Y}}$) gleich sein. Dadurch gibt es keine erklärende Variation von e_i (gemessen durch s^2_E), da es keine Abweichungen gibt. Weiter kann man daraus folgern, dass der Steigungskoeffizeint b gleich 1 sein muss. Dies wird auch von dem Koeffizienten von *dat$xvar* deutlich. Warum muss b gleich 1 sein? Wenn $R^2 = r^2_{XY} = 1$ ist, dann muss auch $\sqrt{r^2_{XY}} = r_{XY} = 1$ sein. Da hier die Variablen X und Y identisch sind, sind damit auch die Standardabweichungen identisch und das Verhältnis $\frac{s_Y}{s_X} = 1$. Daraus folgt: $b = r_{XY} \cdot \frac{s_Y}{s_X} \iff 1 = 1 \cdot 1 = 1$

Im nächsten Beispiel wird eine geringe Abweichung eingeführt, wodurch e_i nicht mehr gleich Null ist und $\hat{Y} \neq Y$ ist.

```
set.seed(02101990)
x <- seq(1,20,1)
a <- 0
b <- 1
e <- rnorm(20,0,1)
# Durch e ungleich 0 wird eine Streuung eingebaut, wodurch x nicht mehr genau gleich y ist.
dat <- data.frame(xvar = x,
                  yvar = a + b*x + e)
reg <- lm(dat$yvar ~ dat$xvar)
dat$fitted <- reg$fitted.values
dat$i <- seq(1,20,1)
ggplot(dat, aes(x=xvar, y=yvar)) +
  geom_point(shape=1) + # Use hollow circles
  geom_point(aes(x=xvar, y=fitted), shape=2) + # Include fitted values
  geom_smooth(method=lm, # Add linear regression line
              se=FALSE) # Don't add shaded confidence region
```



```
reg <- lm(dat$yvar ~ dat$xvar) # Regression from Y on X to get the coefficients
```

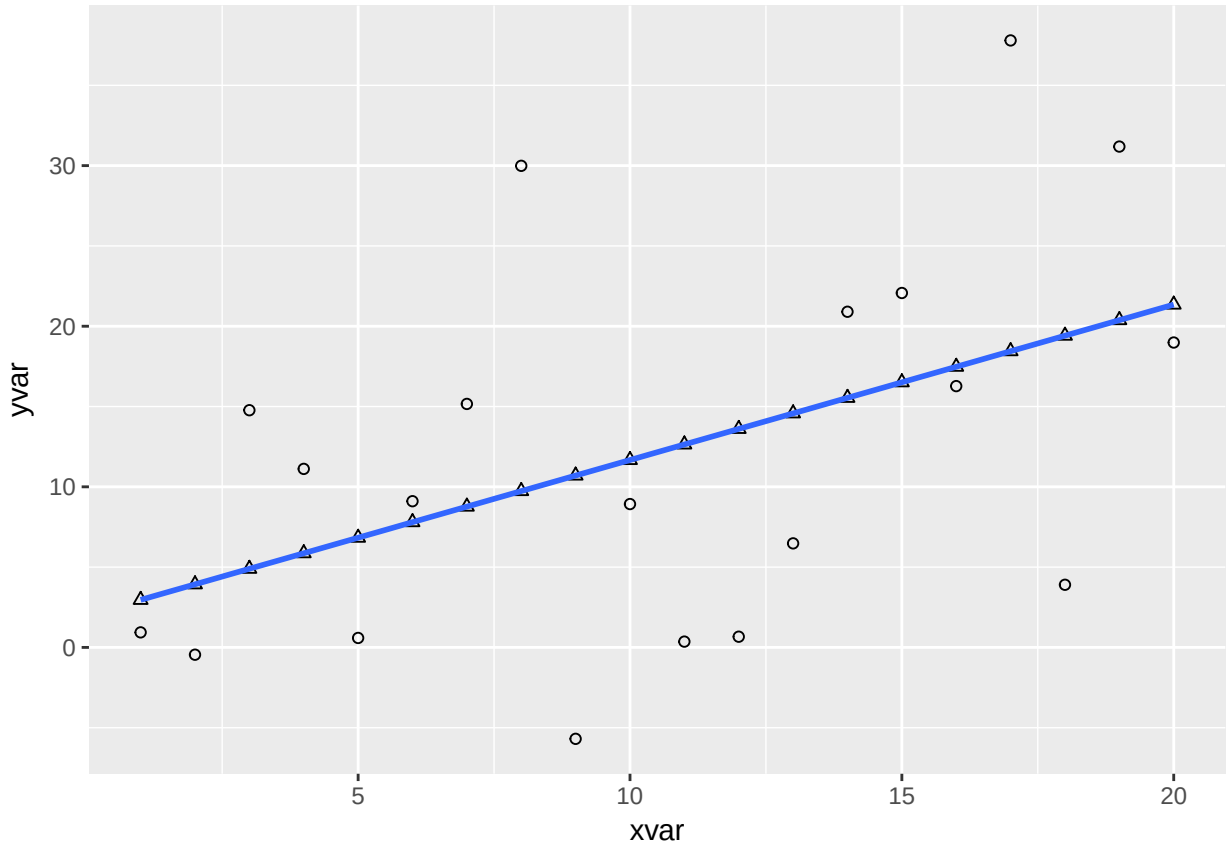
Hier gibt es Abweichungen, weswegen $\hat{Y}$ Y nicht mehr komplett erklären kann. Man kann hier also nicht mehr nur $\hat{y}$ angucken, um alles über y zu erfahren Die Erklärung sollte in diesem Fall aber trotzdem noch groß sein, aber nicht mehr gleich 1 ist. Genau gesagt, ist das Bestimmtheitsmaß R^2 hier gleich 0.9692146.

Die (X, Y)-Paare sind in dem Schaubild die runden und leeren Kreise. Die Dreiecke stehen für $(X, \hat{Y})$-Paare. Somit ist der senkrechte Unterschied zwischen Y und $\hat{Y}$ die Abweichung e.

Die Steigung hier ist 0.9967658 und somit nicht mehr genau 1.

Im nächsten Beispiel wird die Abweichung zur Verdeutlichung nochmal erhöht.

```
set.seed(02101990)
x <- seq(1,20,1)
a <- 0
b <- 1
e <- rnorm(20,0,10)
# Die Streuung hier wird von 1 auf 10 erhöhrt. Somit unterscheiden sich x und y ein Stück mehr.
dat <- data.frame(xvar = x,
                  yvar = a + b*x + e)
reg <- lm(dat$yvar ~ dat$xvar)
dat$fitted <- reg$fitted.values
dat$i <- seq(1,20,1)
ggplot(dat, aes(x=xvar, y=yvar)) +
  geom_point(shape=1) + #Use hollow circles
  geom_point(aes(x=xvar, y=fitted), shape=2) + # Include fitted values
  geom_smooth(method=lm, # Add linear regression line
              se=FALSE) # Don't add shaded confidence region
```



```r
reg <- lm(dat$yvar ~ dat$xvar) # Regression von Y auf X
```

Das Bestimmtheitsmaß ist jetzt gleich 0.2288179. Die größeren Abweichungen (Abstand zwischen dem runden Kreis und dem dazugehörigen Dreieck) führen zu einer geringeren Erklärung von X (und somit $\hat{Y}$) für Y. Die Steigung ist immer noch nahe 1, aber wieder nicht genau eins, sondern gleich 0.9676581. Dies liegt an der großen Streuung von der Abweichung e.

Interpretation des Steigungskoeffizienten

Die Erweiterung des Steigungskoeffizienten im Verlgeich zum Korrelationskoeffizienten ist es, dass man mehr Informationen als die Richtung und Stärke eines Linearenzusammenhanges hat.

Mit dem Steigungskoeffizienten kann man auch sagen, um welchen Wert sich eine Variablen erhöhen sollte, wenn eine andere Variable um eine Einheit vergrößert wird.

Wird also nach der Interpretation des Wertes des Steigungskoeffizienten b gefragt, sollte man nicht nur auf die generelle Abhängigkeit eingehen, sondern auch genauen Einfluss um b von X auf Y beschreiben.

In der Vorlesung gab es folgendes Beispiel für die Interpretation:

Eine Erhöhung der Fernsehdauer um eine Stunde, geht für die 10 Individuen mit einer durchschnittlichen Veränderung der Dauer der Tiefschlafphase um b Stunden einher.

Wichtig hier ist, dass man von einer **durchschnittlichen** Erhöhung spricht, da die Formeln alle auf Mittelwerten beruhen.

Es bedeutet also, dass Kinder, die eine Stunde länger ferngesehen haben, eine kürzere Tiefschlafphase haben. Es bedeutet **nicht**, dass wenn Kinder weniger Stunden fernsehen, dann auch unbedingt besser schlafen. Die Interpreation muss hier rein deskriptiv sein! Es werden Korrelationen erklärt, keine Kausalitäten.

Weitere Beispiele:

Regression von Y auf X Einkommen → Höhe der Ausgaben für Molkereiprodukte:

Eine Erhöhung des Einkommens um tausend Euro pro Jahr, geht für die 5 Individuen mit einer durchschnittlichen positiven Veränderung der Ausgaben für Molkereiprodukte um 1.2 *Euro im Monat einher.*

Regression von X auf Y Menge der Molkereiprodukte → Einkommen.

Eine Erhöhung der Molkereiprodukte um einen Euro pro Monat, geht für die 5 Individuen mit einer durchschnittlichen positiven Veränderung des Einkommens um 600 *Euro pro Jahr einher.* (0.6 tausend Euro = 600 Euro.)

Man sollte immer deutlich machen, dass man eine Korrelation *beobachtet*, und keine Zusammenhang mit der Satzstruktur *Je . . . , desto. . . .* Zweiteres wird meistens die Interpretation für einen kausalen Effekt, nicht aber für die Korrelation. Die Interpretation sollte sich auf die beobachteten Individuen beschränken und nicht verallgemeinert werden.