Bericht

R_Projekt_AFS

1. Einleitung

1.1. Der Datensatz

1.1.1. Ursprung

Der Datensatz ist im Rahmen einer Masterarbeit von Gregor Ziegeltrum entstanden. Hierfür wurde öffentlich verfügbare Daten von der Website www.myfitnesspal.com via Web Scraping extrahiert. Gregor Ziegeltrum hat uns ein Teil dieses Datensatzes für unser R-Projekt bereitgestellt.

1.1.2. Inhalt

Kurz gesagt: Der Datensatz enthält das Essverhalten von 221 Personen in Form von ca. 1 Millionen Beobachtungen. Im Detail stellt eine Beobachtung (Zeile), jeweils ein Lebensmittel, welches die jeweilige Person verzerrt hat, dar. Wir finden pro Beobachtung die folgenden Spalten vor:

- **username**: Der Username der Person z.B. “Stephan85” - ein String, der bei ca. 221 Personen als kategorial angenommen werden kann
- **gender**: Das Geschlecht der Person: “m” oder “f” - Eine kategorische Variabel mit genau 2 Optionen
- **location**: Der Ort, in dem die jeweilige Person wohnt, z.B. “Mount Airy, MD” - ein String, der bei ca. 140 verschiedenen Orten als kategorial angenommen werden kann.
- **joined_date**: Das Datum, an dem die jeweilige Person ihren Account auf www.myfitnesspal.de erstellt hat, z.B. “16-05-13” - ein String, der Form DD-MM-YY.
- **age**: Das Alter der Person z.B. 43 - ein Integer.
- **meal_date**: Das Datum, an dem die Person das jeweilige Lebensmittel verzerrt hat, z.B. “12-02-13” - ein String, der Form DD-MM-YY.
- **meal**: Die Mahlzeit, zu der das jeweilige Lebensmittel verzerrt; Standartoptionen sind “breakfast”, “lunch”, “dinner” und “snacks”, allerdings sind auch benutzerdefinierte Eingaben möglich wie “11am - 1pm” - ein String
- **name**: Der Name (inkl. Mengenangabe) des verzerrten Lebensmittels z.B. “Pace - Medium Chunky Salsa, 20 tbsp” - ein String.
- **calories**: Die Kalorien (in Kilokalorien), welche die jeweilige Portion des Lebensmittels enthält, z.B. 100 - ein Integer.
- **carbs(g)**: Die Menge an Kohlenhydraten (in Gramm), welche die jeweilige Portion enthält z.B. 105 - ein Integer.
- **fat(g)**: Die Menge an Fett (in Gramm), welche die jeweilige Portion enthält - ein Integer.
- **protein(g)**: Die Menge an Protein (in Gramm), welche die jeweilige Portion enthält - ein Integer.
- **cholest(mg)**: Die Menge an Cholesterin (in Milligramm), welche die jeweilige Portion enthält - ein Integer.
- **sodium(mg)**: Die Menge an Salz (in Milligramm), welche die jeweilige Portion enthält - ein Integer.
- **sugars(g)**: Die Menge an Zucker (in Gramm), welche die jeweilige Portion enthält - ein Integer.

- **fiber(g)**: Die Menge an Ballaststoffe (in Gram), welche die jeweilige Portion enthält - ein Integer.

Hinweis: Die Nährwertangaben sind hier pro angegebene Portion und nicht pro 100 Gramm.

1.2 Fragestellungen

Folgende Fragestellungen hoffen wir mit Hilfe des Datensatzes zu beantworten:

- Haben Frauen einen sogenannten Sweet Tooth, genauer, gibt es einen Zusammenhang von Gender und Zuckerkonsum?
- Bei welcher Fast-Food-Kette ist die durchschnittliche Mahlzeit am gesündesten/ungesündesten?

Sekundär stellt sich dann natürlich die Frage, was überhaupt eine Mahlzeit gesund oder ungesund gemacht.

2. Explorative Datenanalyse

2.1 Vorbereitung

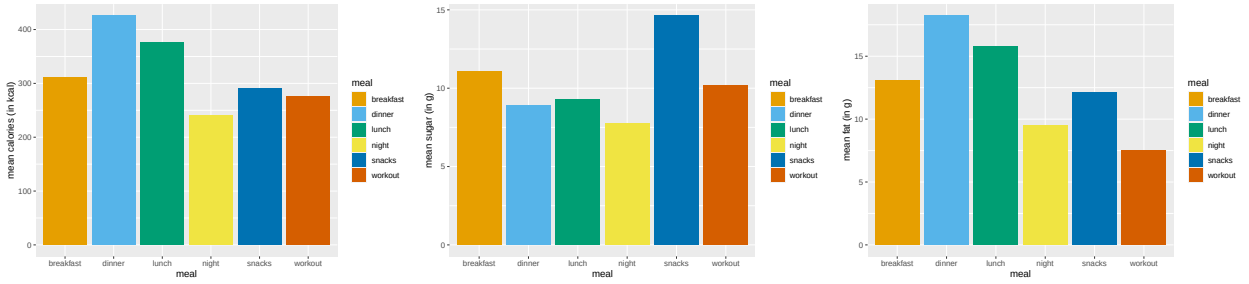
Da der Datensatz von Nutzern generiert wurde und dann durch Web Scraping extrahiert wurde, ist er natürlich nicht perfekt. Deshalb müssen ein paar Säuberungen durchgeführt werden. Die Wichtigsten sind hier zusammengefasst:

- **Beobachtungen mit 'NA' als eine der Nährwerte**: Dies sind ca. 20-30 % der Beobachtungen und stammen in so gut wie jeden Fall daher, dass ein Nutzer zu faul war, eine Null für z.B. Fett einzutragen. Da wir allerdings genügen Beobachtungen haben (ca. 1 Million), lassen wir diese Beobachtungen trotzdem einfach weg.
- **Beobachtungen mit negativen Nährwerten**: Diese sind Tippfehler oder "Korrekturrechnungen" einiger Nutzer und werden natürlich weggelassen (allerdings sind das nur ca. 200 Beobachtungen).
- **Beobachtungen mit zu hohen Nährwerten**: Tippfehler und Korrekturrechnungen gehen leider auch in die andere Richtung (wie wäre es mit einer Mahlzeit mit 20000 Kalorien?). Hierfür entfernen wir die 0,1 % größten Beobachtungen pro Nährwert (Kalorien, Kohlenhydrate, Salz, Fett etc.).
- **Eintragungen mit personalisierten Mahlzeiten**: Den Nutzern war es möglich sich eigene Namen für ihre Mahlzeiten zu erstellen. Das macht es schwer die Variable meal zu untersuchen, weshalb wir Einträge jeweils mit passender Zuordnung in die Mahlzeiten "breakfast", "lunch", "dinner", "night", "workout", "snacks" und "all day" unterteilen. "all day" enthält hier alle restlichen schwer zuweißbaren Einträge.

2.2. Kategoriale Variablen

Als kategoriale Variablen möchten wir im Folgenden meal, date, age, gender und location betrachten. Die Einträge username und joined-date lassen wir aus Datenschutzgründen außen vor. Die Variable name sticht durch die Menge an unterschiedlichen Einträgen hervor, wir werden sie deshalb gesondert in Abschnitt 4 im Bezug auf Fast-Food Ketten betrachten. Zunächst einige exemplarische Übersichten der Nährstoffanteile unterschiedlicher Mahlzeiten. In diesem Abschnitt ist grundsätzlich eine Diskrepanz zwischen bekannten Durchschnittsnährwerten und den hier errechneten Daten zu erkennen. Man kann davon ausgehen, dass der Unterschied zwischen dem zu erwartenden Gesamtkonsum an Kalorien (oder anderen Inhaltsstoffen) pro Tag und den hier dargestellten Daten auf die Inkonsequenz der Nutzer zurück zu führen ist. Die Verteilung ist trotzdem aussagekräftig, da man von einem tagesunahängigen Verhalten der Probanden ausgehen kann. Eine Betrachtung von Absolutwerten ist also weniger sinnvoll, als eine relative Darstellung und die Analyse der Verteilung.

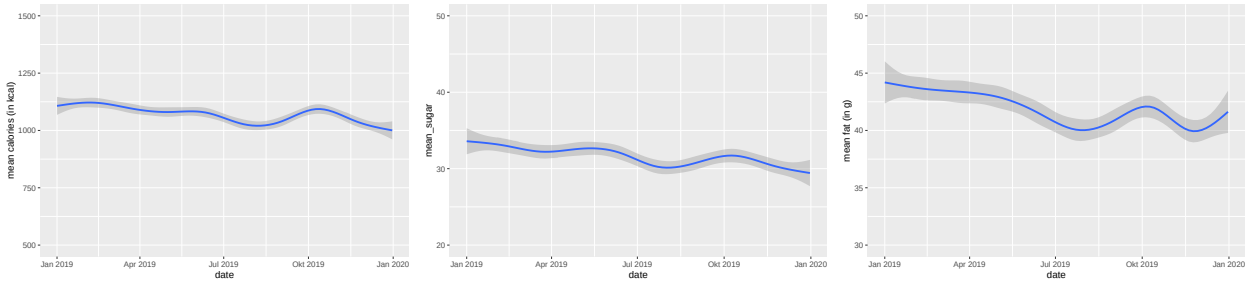
2.2.1 Meal



Es lassen sich also durchaus Unterschiede in den Nährstoffzusammensetzungen einzelner Mahlzeiten identifizieren. Im Vergleich zu der Übersicht der Kalorien fällt bei der Betrachtung der Zuckeranteile auf, dass Frühstück, Workout und insbesondere Snacks überproportinal viel Zucker enthalten. Ein Workout enthält dafür besonders wenig Fett.

2.2.2 Date

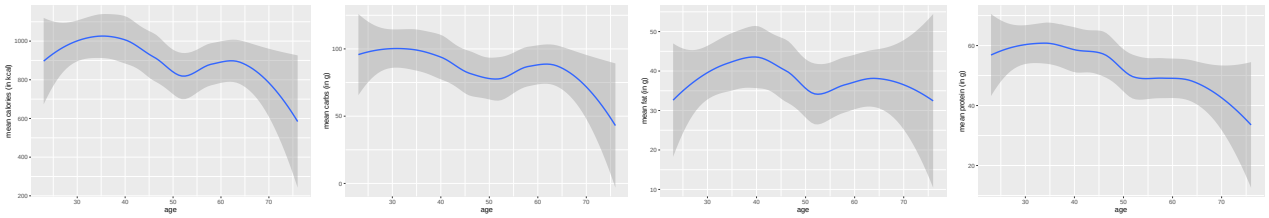
Wir möchten nun die Unterschiede in der Nahrungsaufnahme innerhalb eines Jahres betrachten, hier exemplarisch das Jahr 2019.



Zu erkennen ist ein zeitweises Hoch im Oktober bei allen drei Angaben und ein anschließendes Abfallen zum Ende Jahres bei Kalorien und Zucker. Gesamttendenz ist in allen drei Fällen fallend, trotzdem kann man grundsätzlich von einer Gleichverteilung sprechen.

2.2.3 Age

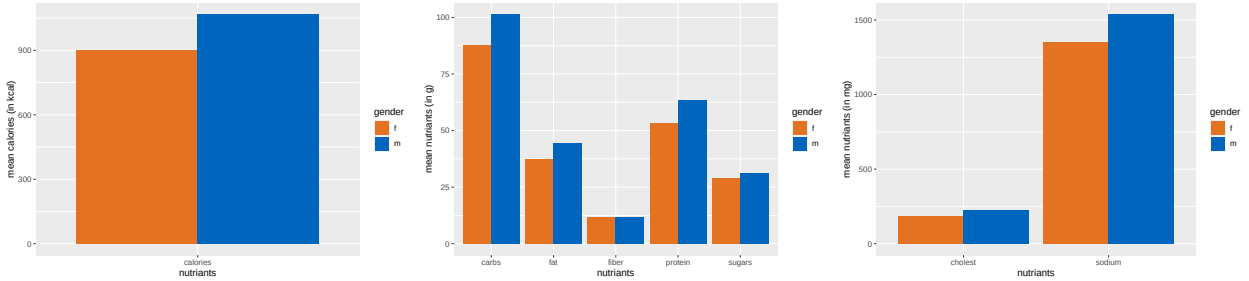
Wir möchten nun das Essverhalten nach Alter gegliedert, anhand einiger exemplarischer Nährwerte, darstellen.



Man beobachtet zwei deutliche Peaks bei 35-40 und 60-65 Jahren für Kalorien und Fett, die bei Kohlenhydraten und Protein nicht bis deutlich abgeschwächt zu sehen sind. Kohlenhydrate und Protein weisen einen fast linearen Abfall auf, während die Werte, Kalorien und Fett, sowohl für jüngere als auch für ältere Menschen niedriger sind.

2.2.4 Gender

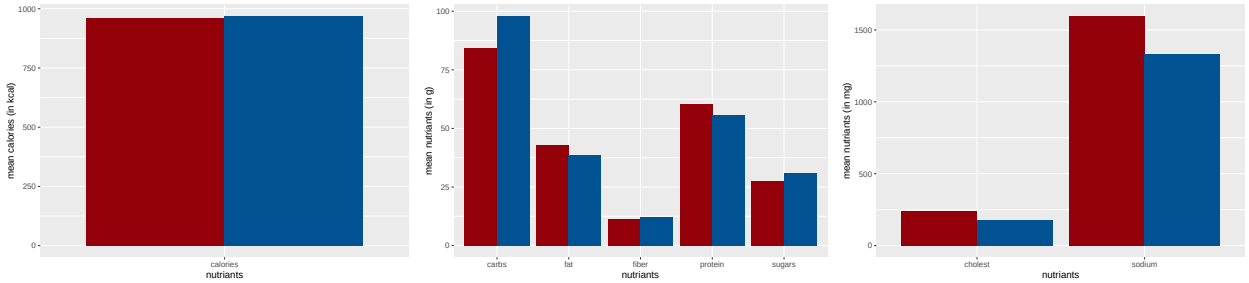
Wir möchten nun auch die Unterschiede im Konsumverhalten von Männer und Frauen beleuchten.



Es ist sowohl für Kalorien, als auch für alle Nährstoffe außer Cholesterin und Ballaststoffen ist ein deutlich höherer Wert bei Männern zu erkennen. Der höchste prozentuale Unterschied ist bei Protein und Fett zusehen, wo Männer 25% bzw. 20% mehr zu sich nehmen.

2.2.5 Location

Nun möchten wir noch den Wohnort der Nutzer als Faktor für Unterschiede im Essverhalten untersuchen. Bei rund 220 Nutzern in den USA ist eine Einteilung in States oder sogar Städte wenig repräsentativ, weshalb wir stattdessen in dieser Betrachtung nur nach Red- und Blue-States trennen, also nach den Wahlentscheidungen der letzten Präsidentschaftswahl.

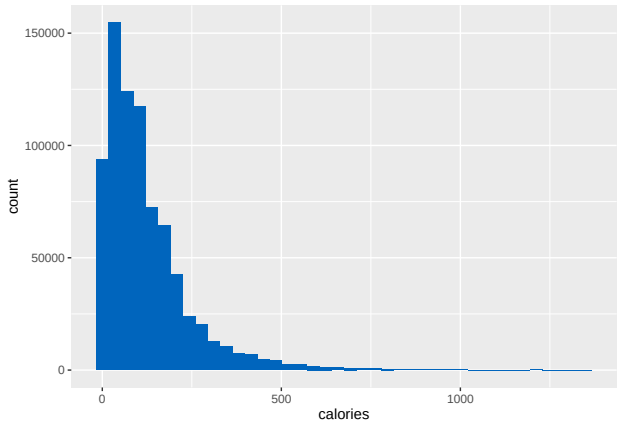


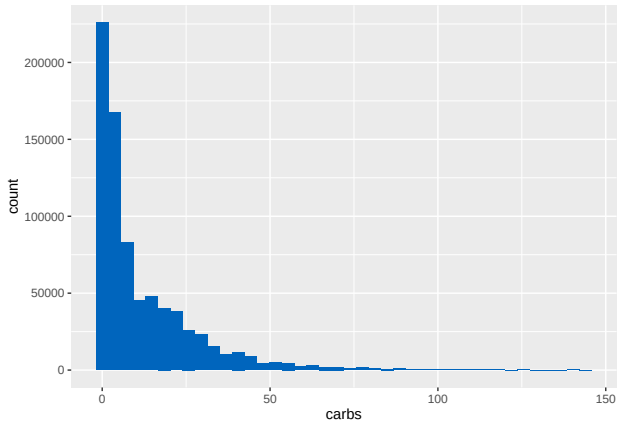
Hier entspricht der rote Balken den Republikanern und der blaue Balken den Demokraten. Zu erkennen ist, dass Wähler beider Parteien etwa gleich viele Kalorien zu sich nehmen. Demokraten tendieren dabei deutlich eher zu Kohlenhydraten bzw. Zucker, während Republikaner mehr Fett und Protein konsumieren.

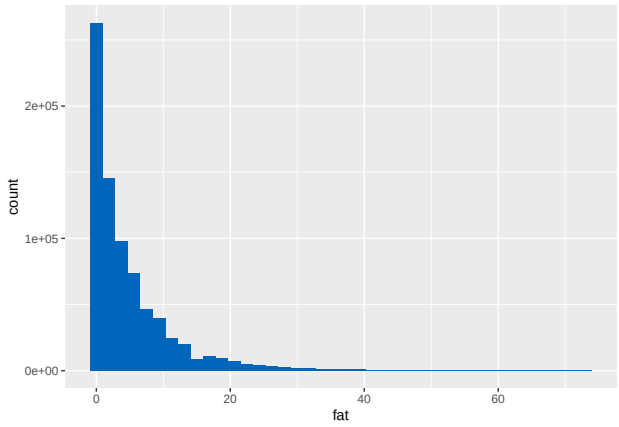
2.3. Kontinuierliche Variablen

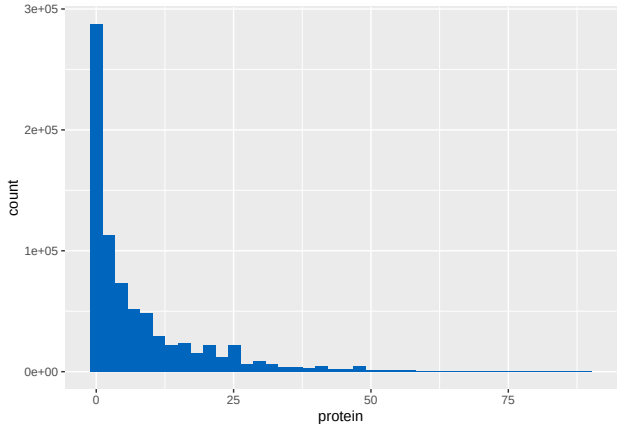
2.3.1. Ein Überblick

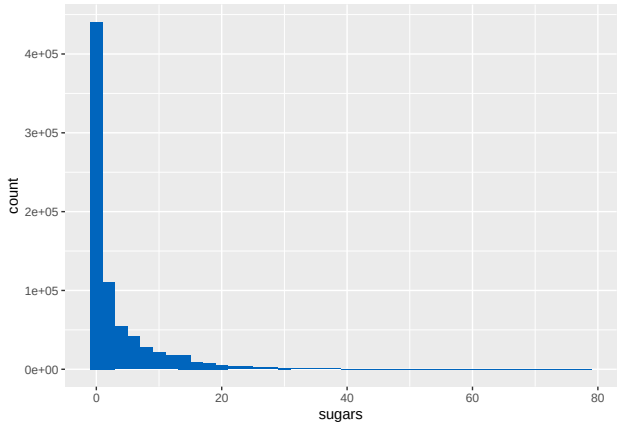
Bei den kontinuierlichen Variablen (unseren Nährwerten) verschaffen wir uns einen Überblick, indem wir die jeweiligen Verteilungen als Histogramme darstellen.

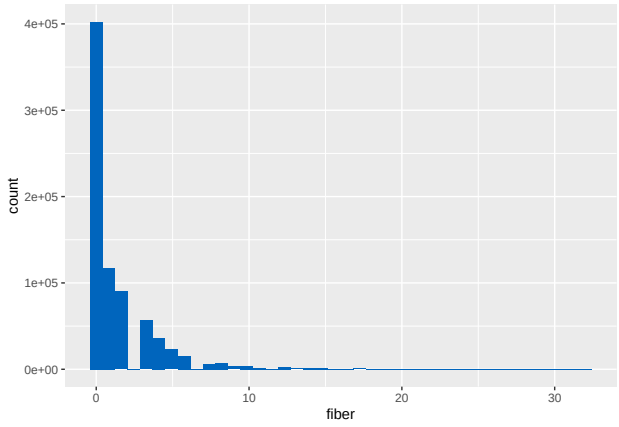

count
150000
100000
50000
0
0
500
1000
calories

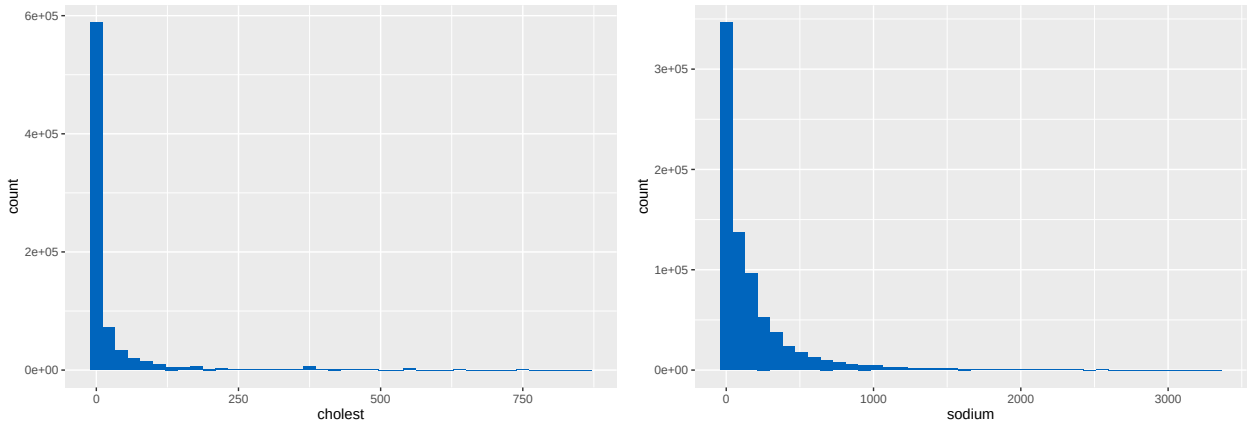

count
200000
150000
100000
50000
0
0
50
100
150
carbs


count
2e+05
1e+05
0e+00
0
20
40
60
fat


count
3e+05
2e+05
1e+05
0e+00
0
25
50
75
protein


count
4e+05
3e+05
2e+05
1e+05
0e+00
0
20
40
60
80
sugars


count
4e+05
3e+05
2e+05
1e+05
0e+00
0
10
20
30
fiber



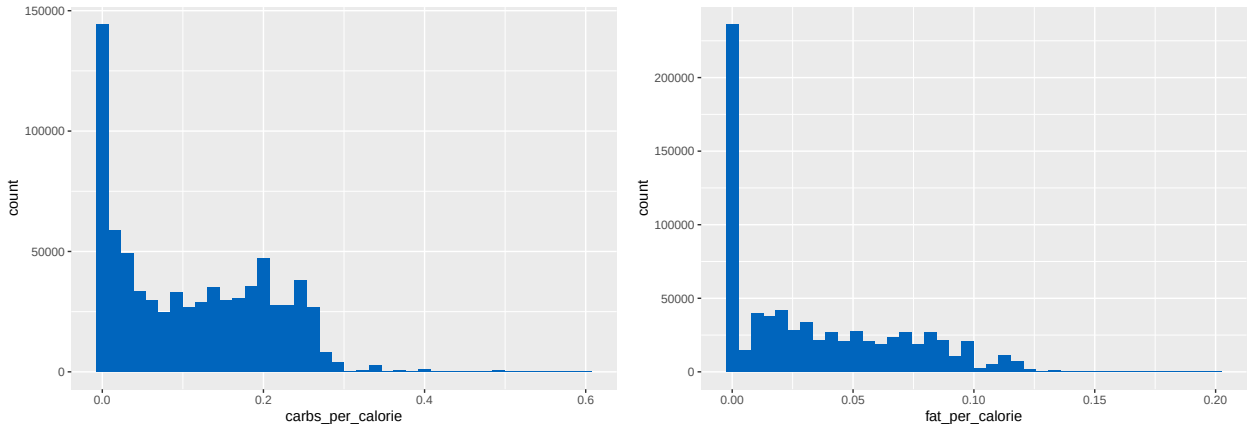
Alle Histogramme sind durch ein Maximum bei 0 und ein daraufhin schnelles Abfallen charakterisiert. Wir wissen zwar jetzt, wie die jeweiligen Daten verteilt sind, aufgrund der stark variierenden Portionsgrößen können wir allerdings noch nicht sehen "wie salzig" oder "wie fettig" die verspeisten Lebensmittel waren. Damit werden wir uns jetzt beschäftigen.

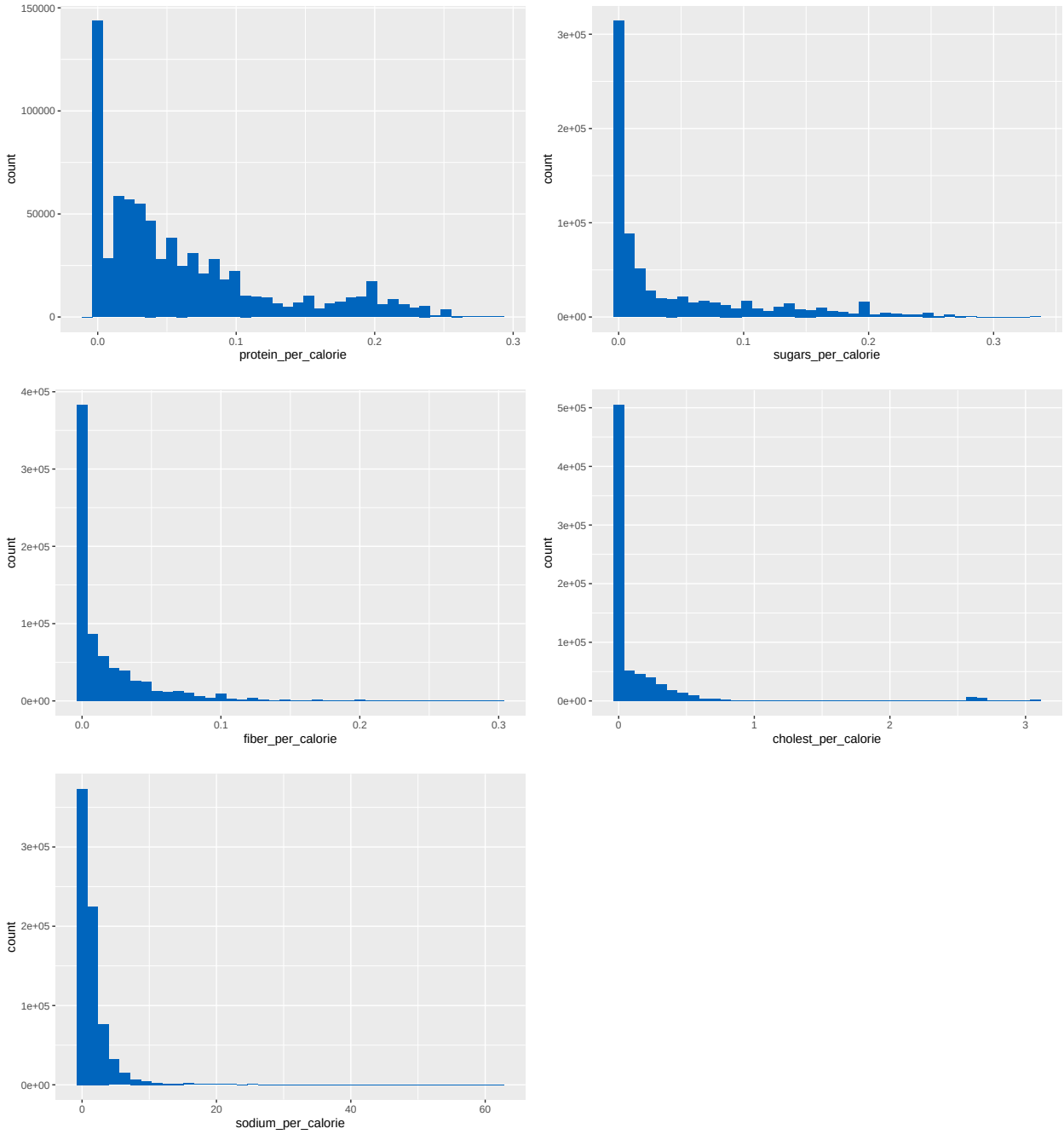
2.3.1 Relative Nährwerte

Diesem Abschnitt liegt die folgende Überlegung zugrunde: Jeder Mensch ist pro Tag ungefähr die gleiche Menge an Kalorien. Der tägliche Konsum an Fett, Zucker, Salz und anderen Nährstoffen verteilt sich als auf diese konstante Menge an Kalorien. Möchten wir also herausfinden, wie "süß", "salzig" oder "fettig" eine Ernährung ist, so betrachten wir, wie viel Fett, Zucker oder Salz pro Kalorien zu sich genommen werden.

Diese Methode ist natürlich auch nicht perfekt: Es gibt, wie wir in 2.3.1. gesehen haben, viele Beobachtungen, die keine Kalorien enthalten. Dies ist z.B. ein Teelöffel Salz ("Salt - Salt, 0.25 tsp(s)"), ungesüßter Tee ("Teas' Tea - Jasmine Green Tea (Unsweetened), 8 oz (240ml)") oder auch Fehler beim Eintragen (z.B. "Totinos - Party Pizza - Triple Pepperoni, 0 pizza"). Diese Beobachtungen filtern wir für die Bestimmung der relativen Nährwerte heraus. Zusätzlich entfernen wir wieder die 0,1 % größten Beobachtungen pro relativen Nährwert, um Tippfehlern entgegenzuwirken.

Sonst werden die relativen Nährwerte einer Beobachtung einfach dadurch bestimmt, indem wir die Menge des Nährstoffes durch die Kalorien teilen. Wir plotten die Resultate wieder als Histogramme:





Wieder haben alle Histogramme einen starken Ausschlag bei 0 gemeinsam, allerdings können wir nun auch feinere Unterschiede ablesen. So sind Kohlenhydrate und Fett außerhalb der 0 fast gleich verteilt (bis jeweils 0.25 bzw. 0.1 - dazu kommen wir gleich), während die bei den restlichen Nährstoffen wieder ein schnelles Abfallen zu beobachten ist. Es scheint also so, als würde sowohl kohlenhydrathaltige bzw. fettige als auch kohlenhydratarme bzw. fettfreie Lebensmittel auf dem Speiseplan stehen, während bei den anderen Nährstoffen die Häufigkeit mit der Konzentration stetig abnimmt.

Wir können jetzt auch einen "Sanity-Check" auf unseren Daten durchführen. Die Kalorien, die ein Gramm Kohlenhydrate, Fett, Protein und Zucker enthält, sind bekannt. Dadurch ergibt sich für jeder dieser relativen Nährwerte eine natürliche Obergrenze (ein Lebensmittel kann maximal nur aus Fett oder nur aus Zucker bestehen):

- **Kohlenhydrate:** *"Carbohydrate consumed in food yields 3.87 kilocalories of energy per gram for simple sugars,[18] and 3.57 to 4.12 kilocalories per gram for complex carbohydrate in most other foods."* (Quelle: https://en.wikipedia.org/wiki/Carbohydrate). Ein Gramm pure Kohlenhydrate hätten dann einen relativen Nährwert von maximal $1/3.5 \approx 0.29$.
- **Zucker:** Nach der obigen Quelle liegt die Obergrenze als bei $1/3.87 \approx 0.26$.
- **Fett:** *Each gram of fat when burned or metabolized releases about 9 food calories (37 kJ = 8.8 kcal).* (Quelle: https://en.wikipedia.org/wiki/Fat). Die Obergrenze ergibt sich als $1/8.8 \approx 0.11$.
- **Protein:** *As a fuel, proteins provide as much energy density as carbohydrates: 4 kcal (17 kJ) per gram;"* (Quelle: https://en.wikipedia.org/wiki/Protein_(nutrient)). Wieder berechnen wir die Obergrenze als ungefähr $1/4 = 0.25$

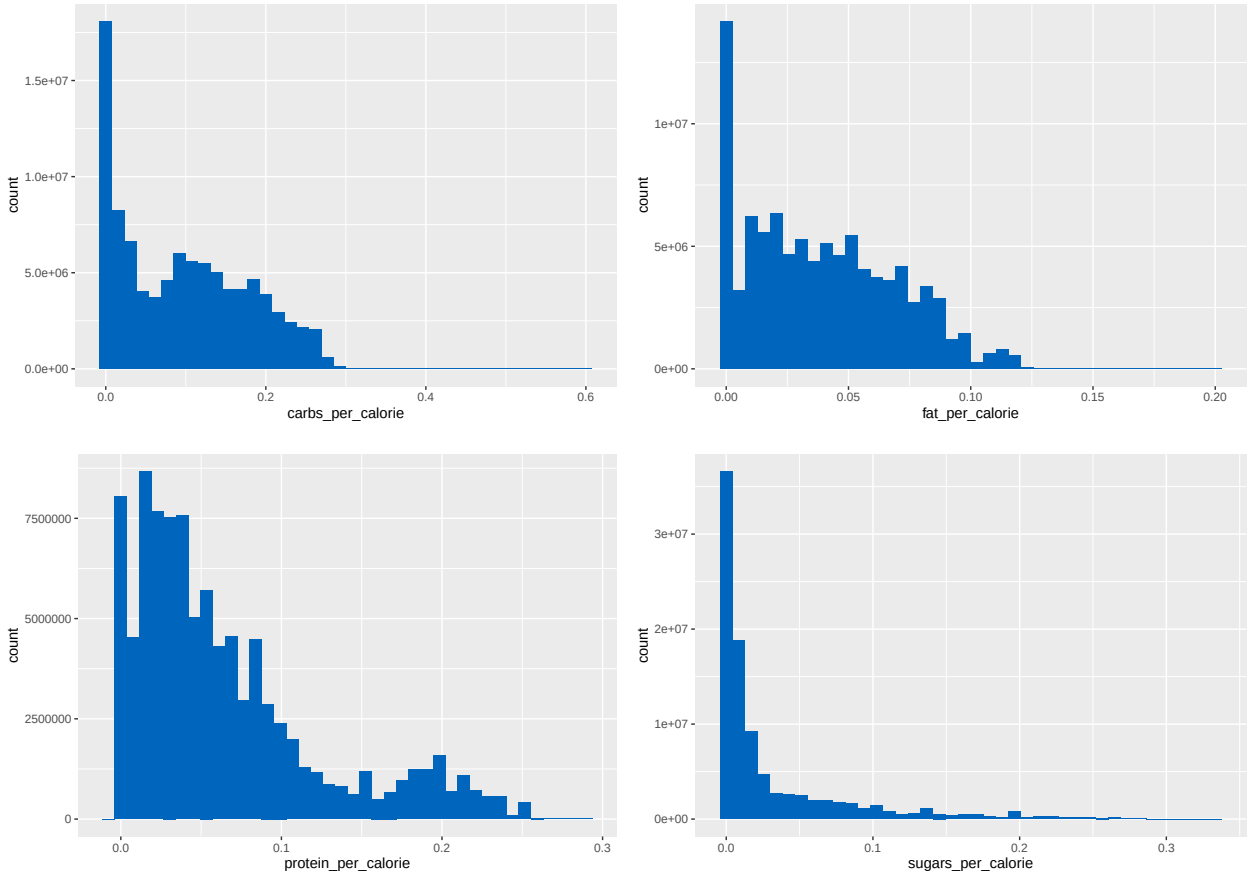
Ein kurzer Blick auf unsere Grafiken ergibt, dass sich die Obergrenzen genau mit unseren Daten decken.

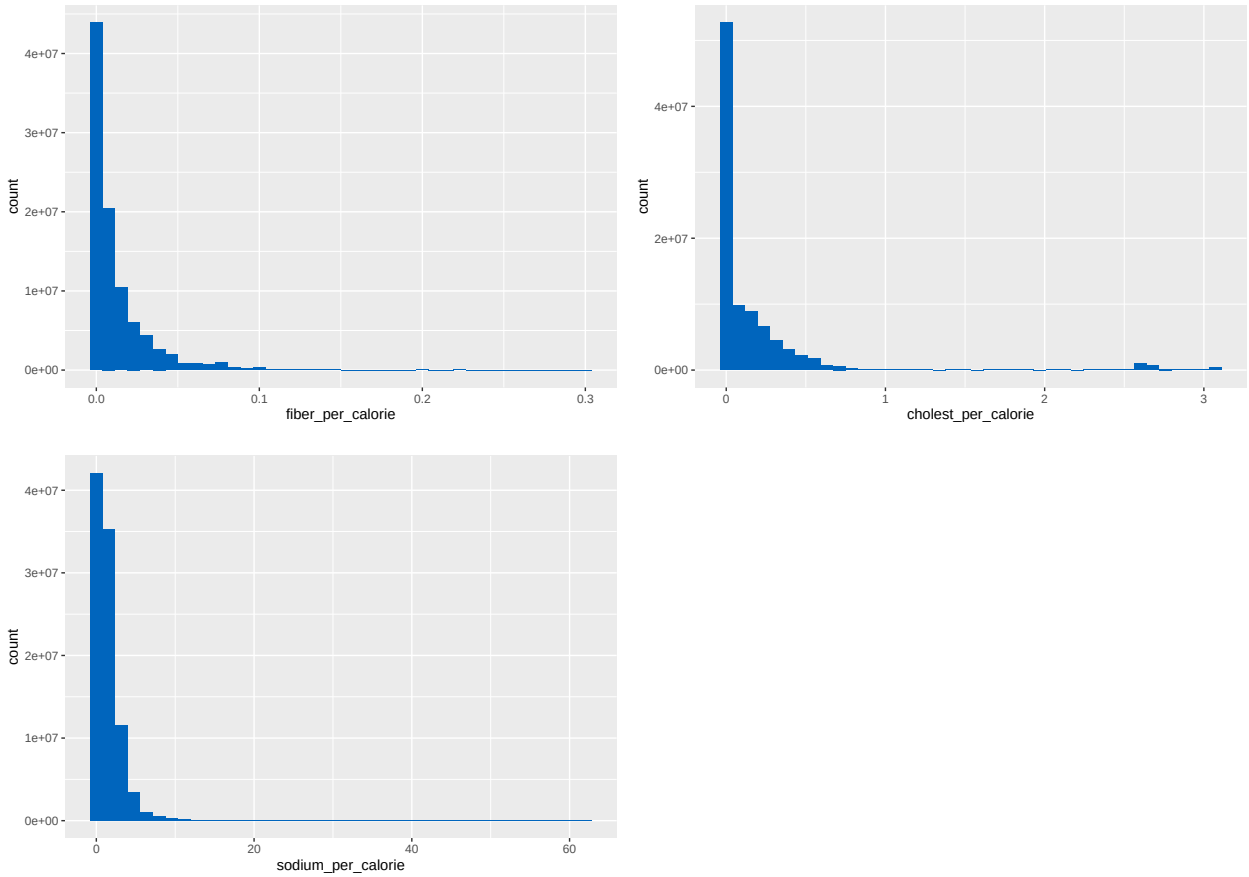
2.3.2. Gewichtete Histogramme

Bis jetzt ging in jedem Histogramm jede Beobachtung mit dem gleichen Gewicht ein. Ist es allerdings wirklich sinnvoll, die eine Schokopraline mit dem Schweinebraten zum Mittagessen gleichzusetzen? Eine gute alternative wäre, dass jede Beobachtung nicht mit dem Gewicht 1, sondern der Anzahl seiner Kalorien mit eingeht. Dies ist eine einzigartige Möglichkeit des Datensatzes, da dieser ja das komplette Essverhalten von 220 Nutzern darstellt.

Bei den Plots der absoluten Nährwerte (2.3.1) ist eine Gewichtung natürlich weniger sinnvoll. Wenn wir die Plots der relativen Nährwerte (2.3.2) gewichten, so können wir besser beurteilen, wie "salzig", "fettig" oder "süß" die durchschnittliche Kalorie ist.

Hier also die Plots aus (2.3.2) gewichtet nach Kalorien:





Über alle Plots hinweg (allerdings ist der Effekt verschieden stark ausgeprägt) sieht man eine Abschwächung der Extrema (d.h. sowohl eine Abschwächung am linken als auch am rechten Rand). Besonders gut sieht man das bei den Proteinen: Die durchschnittliche Kalorie scheint zwar proteinarm zu sein, allerdings auch nicht "proteinfrei" zu sein.

3. Methodik

3.1. Konfidenzintervalle

Im Teil vier werden wir Konfidenzintervalle zum Konfidenzniveau 95%. Dazu nutzen wir die Berechnungsmethode, die im Vorlesungsskript auf Seite 176 im Kapitel Parameterschätzer hergeleitet wurde:

$$C(x) = \left(\overline{x_n} - t_{n-1,\,97.5\%} \sqrt{\frac{s_n^2}{n}}, \overline{x_n} + t_{n-1,\,97.5\%} \sqrt{\frac{s_n^2}{n}} \right)$$

. Dabei bezeichne $\overline{x_n}$ das empirische Mittel, s_n^2 die Stichprobenvarianz und $t_{n-1,\,97.5\%}$ die Quantilsfunktion der t_{n-1}-Verteilung ausgewertet bei 97.5%.

Dabei berechnet diese Methode die Konfidenzintervalle für den Erwartungswert einer Normalverteilung mit unbekannter Varianz bei n unabhängigen Stichproben. Es bleibt also zu argumentieren, dass wir die Stichproben als Produktmodell von Normalverteilten Zufallsvariablen auffassen können

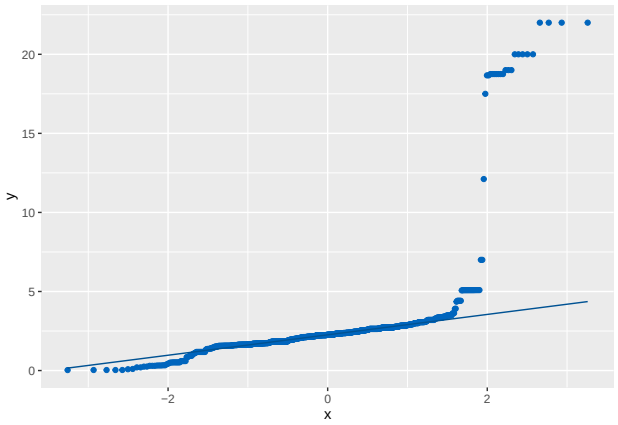
3.1.1. Zur Unabhängigkeit und identischen Verteilung

Wir wenden die Bereichsschätzer auf einzelne Gerichte bei einer festen Fast-Food Kette an und hier bei auf immer denselben Wert, wie z.B. Fett pro Kalorie. Wir wollen das Mittel von zum Beispiel Fett pro Kalorie einer durchsschnittlichen Mahlzeit aus allen verkauten Mahlzeiten bei einer festen Fast-Food Kette schätzen. Da wir annehmen in unserer Datenbank sind durchschnittlichen US-Bürger die ihre gegessenen Mahlzeiten gewissenhaft eintragen, ist auch die Annahme der Unabhängigkeit und identischen Verteilung der z.B Fett pro Kalorie Werte unterschiedlicher Mahlzeiten in unserer Datenbank bei einer festen Fast-Food Kette angebracht.

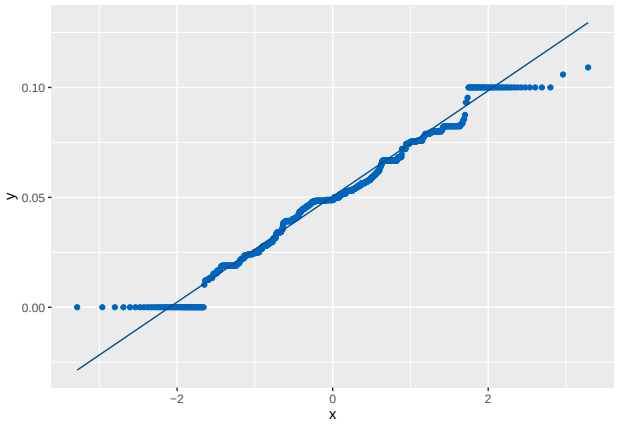
3.1.2. Zur approximativen Normalverteilung

Wir betrachten Normalquantilplots:

```
data_fast_food %>% filter(name == "TacoBell") %>%ggplot(aes(sample = sodium_per_calorie)) +
  stat_qq(color = c_1) + stat_qq_line(color = c_2)
```




```
data_fast_food %>% filter(name == "Wendy's") %>%ggplot(aes(sample = fat_per_calorie)) +
  stat_qq(color = c_1) + stat_qq_line(color = c_2)
```



Hier haben wir einmal Salz pro Kalorie bei allen eingetragen Mahlzeiten von Taco Bell und einmal Fett pro Kalorie bei den Gerichten in unserer Datenbank, die bei Wendy's gekauft wurden in Normalquantilplot eingetragen. Die geringen Abweichungen von einer Gerade, bestätigen unsere Annahme, dass die Daten approximativ Normalverteilt sind.

3.2. Hypothesentest

Wir möchten in Part 4 den Unterschied im Zuckerkonsum von Männern und Frauen betrachten und im Speziellen die Hypothese, dass Frauen am Tag im Schnitt mehr Zucker pro Kalorie essen, untersuchen. Dazu müssen wir zunächst einen Test wählen dessen Vorraussetzungen von unseren Daten erfüllt werden können.

Da wir insbesondere keine Gleichheit von Varianz in den Daten von Männern und Frauen annehmen können, möchten wir, statt dem aus der Vorlesung bekannten Zweistichproben-t-test, einen Welch-t-test verwenden.

Donald W. Zimmermann schreibt "... the Welch test is superior to the t test when variances are unequal."[2]

Dieser hat die die Teststatistik

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S^2_{X,n}}{n} + \frac{S^2_{Y,m}}{m}}}$$

hierbei bezeichnen $\overline{X}, \overline{Y}$ empirische Mittel und $S^2_{X,n}, S^2_{Y,m}$ empirische Stichprobenvarianz zu $(x_1, \ldots, x_n)$ bzw. $(y_1, \ldots, y_m)$ [1].

Vorraussetzungen die es für eine Anwendung zu überprüfen gilt sind die Unabhängigkeit der einzelnen Stichproben und die (approximative-)Normalverteilung innerhalb einer Stichprobengruppe.

3.2.1. Zur Unabhängigkeit

Wir betrachten Mittelwerte des Konsums von Zucker pro Kalorie verschiedener Personen. Wir gehen davon aus, dass diese Personen sich nicht kennen und dadurch ihr Essverhalten nicht beeinflussen, damit erhalten wir also insbesondere Unabhängigkeit der Stichproben der errechneten Mittelwerte.

3.2.2. Zur approximativen Normalverteilung

Für die Anwendung des Welsh-t-Tests ist grundsätzlich vorrausgesetzt, dass die zu vergleichenden Stichprobengruppen jeweils normalverteilt sind, dies können wir zunächst nicht vorraussetzen. Wir betrachten zur Untersuchung zunächst Kolmogorov-Smirnov-Tests [3] auf Normalität

```
ks.test(data_male$mean,"pnorm",mean = mean(data_male$mean),
        sd = sd(data_male$mean))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_male$mean
## D = 0.061282, p-value = 0.8795
## alternative hypothesis: two-sided
```

```
ks.test(data_female$mean,"pnorm",mean = mean(data_female$mean),
        sd = sd(data_female$mean))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data_female$mean
## D = 0.076614, p-value = 0.4111
## alternative hypothesis: two-sided
```

Insbesondere erhalten wir mit Berry-Esseen, dass für größere Mengen an Beobachtungen unter der Annahme eines endlichen dritten Moments, für einen t-test gleiche Bedingungen wie unter Annahme von Normalität gelten [4].

Die dritten Momente unserer Beobachtungsvektoren möchten wir nun noch schätzen

```
moment(data_male$mean,3,absolut = TRUE,central = TRUE)
```

```
## [1] 8.90103e-06
```

```
moment(data_female$mean,3,absolut=TRUE,central = TRUE)
```

```
## [1] 6.555547e-06
```

Wir wollen auf Grund dieser Beobachtungen in Zukunft von einer ungefähren Normalverteilung unserer Beobachtungsvektoren ausgehen.

4. Ergebnise und Schlussfolgerungen

4.1. Fast-Food

4.1.1. Fast-Food vs. Food

Wir vergleichen das Essen in unseren Datensatz, das mit dem Namen der Fast-Food-Ketten McDonalds, Taco Bell, Burger King, Wendy's und Subway gekennzeichnet wurde, sowie allgemeine Mahlzeiten in unserem Datensatz, mit dem empfohlenen Konsum in Gramm beziehungsweise Miligramm pro Kilokalorie.

Die Empfehlungswerte pro Kilokalorie berechnen wir aus den Empfehlungswerten pro Tag, laut folgenden Websites, jeweils geteilt durch die empfohlenen tägliche Kalorienzufuhr von 2000 Kilogramm.

https://www.nhs.uk/live-well/eat-well/what-are-reference-intakes-on-food-labels/

https://www.codecheck.info/hintergrund/tagesbedarf

https://www.ucsfhealth.org/education/increasing-fiber-intake

Wir versuchen also die Frage zu beantworten, ob man seinen Tagesbedarf an Kalorien durch Fast-Food decken sollte, wenn man eine gesunde Ernährung anstrebt.

```
##                      Fast Food   All Food   Recommended
## Sugars per calorie  0.03059901  0.04293114 0.045
## Fat per calorie     0.04498173  0.03444028 0.035
## Fiber per calorie   0.009621767 0.01848685 0.015
## Carbs per calorie   0.1041154   0.1108783  0.13
## Protein per calorie 0.04549365  0.06151535 0.035
## cholest per calorie 0.1644612   0.1523607  "no data"
## sodium per calorie  2.210319    1.640317   1.2
```

Wir können festhalten, dass Fast-Food laut unseren Datensatz zwar weniger Zucker enthält, aber auch mehr Fett und Salz, sowie weniger Protein und Ballaststoff als eine durschnittliche Mahlzeit in unserem Datensatz. Insbesondere konnten wir die Abhängigkeit jener angesprochenen Variablen von der Kategoriellen Variable identifizieren, welche Mahlzeiten die Kategorie Fast Food und allgemeine Mahlzeit zuordnet.

Weiter werden bei Fast-Food Mahlzeiten die empfohlenen Werte zu Fett und Salz, um jeweils 25% und 80% überschritten. Andererseits enthält das durschnittliche Fast-Food Gericht, nur 64% der empfohlenen relativen Menge an Ballaststoffen und 76% der relativen Menge an Kohlenhydraten. Bei Zucker-, Cholesterin-, und Proteinwerten Weisen die Daten keine gesundheitliche Problematik auf.

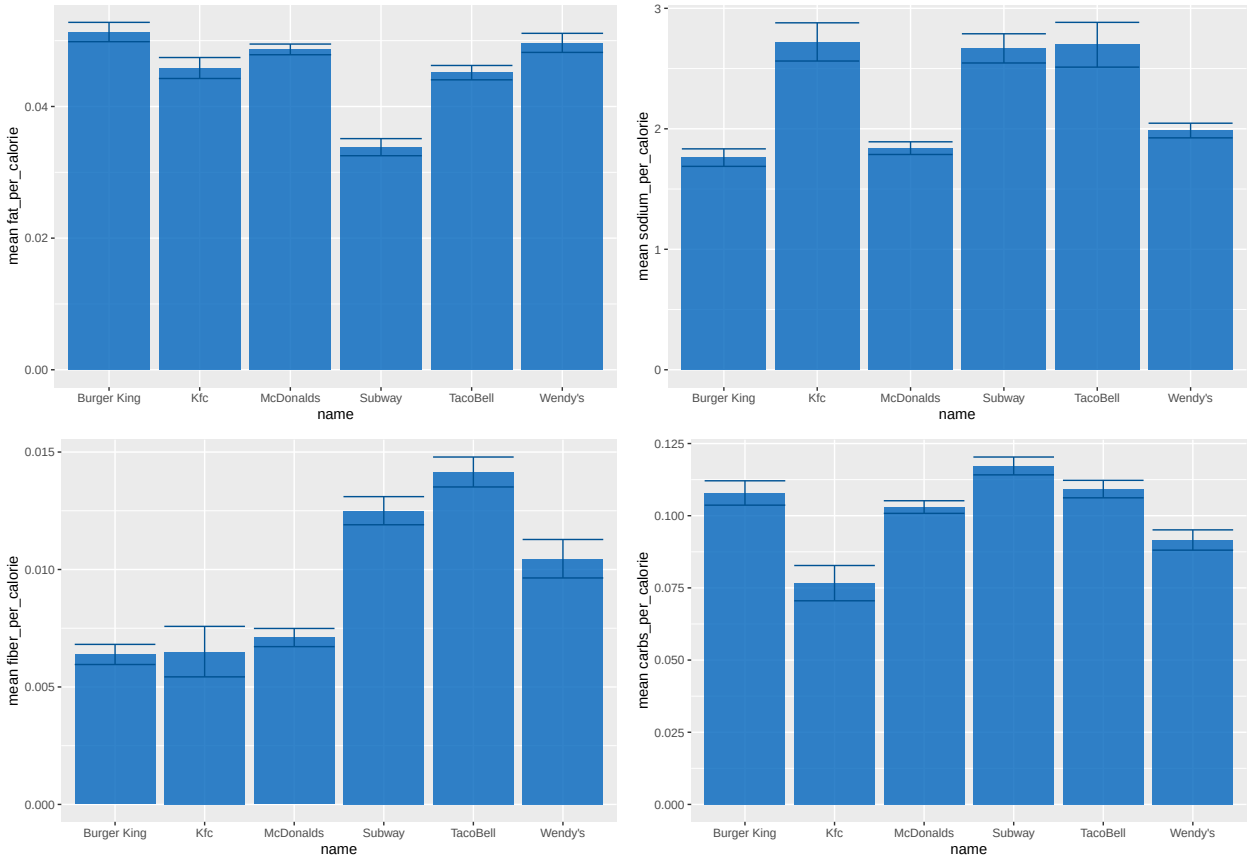
Die ursprüngliche Frage, ob der Kalorienbedarf durch Fast-Food gedeckt werden sollte, können wir, in Hinblick auf die beiden obigen Vergleiche mit nein beantworten.

Wir sehen weiter das Fast-Food weniger Ballaststoff und Kohlenhydrate hat und daher weniger sättigend ist. Das führt dazu, dass weit mehr als der Kalorien-Bedarf gegessen wird.

Die Fast-Food Ketten im Vergleich

Nach dem vorherigen Abschnitt unterscheidet sich ein Fast-Food Gericht und eine allgemeine Mahlzeit am signifikantesten bezüglich enthaltenen Fett, Salz, Ballastsoff und Kohlenhydraten.

Wir wollen nun für die Variablen Gramm Fett pro kcal, Milligramm Sodium pro kcal, Gramm Ballaststoff pro kcal und Gramm Kohlenhydrate pro kcal den Erwartungswert mithilfe des empirischen Mittels schätzen, wobei wir dies für die Gerichte bestimmter Fast Food Ketten einzeln machen. Ziel ist es zu Bewerten, bei welchen Fast Food Ketten das gesündeste Essen gegegessen wird.



Hierbei zeichen wir 95%-Konfidenzintervalle für den Erwartungswert, berechnet wie in Teil 3 beschrieben.

Nur bei Subway überschreiten die durchschnittlichen Fett pro Kalorien Werte nicht den empfohlenen Wert. Alle anderen Fast Food Ketten liegen mehr als 29% über diesem Wert.

Kfc, Taco Bell und Subway überscheiten die empfohlenen Werte an Miligramm Sodium pro Kalorie um mehr als das doppelte. Bei den übrigen Fast Food Ketten wird dieser Wert, um mehr als 50% überschritten.

Insgesamt lässt sich zwar auf Grund von unterschiedlichen Abschneiden bezüglich verschiedener Variablen nicht sagen bei welcher Fast-Food-Kette am gesündetsten gegessen wird, aber unsere Daten sprechen dafür, dass bei KFC die ungesündesten Mahlzeiten verkauft werden. Hier sollte aber angemerkt werden, dass wir zu Kfc nur 448 Beobachtungen haben, werden es bei McDonalds beispielsweise 2527 sind.

Die berechneten 0.95 Konfidenzintervalle machen unsere Vergleiche, trotz unterschiedlicher Datenlage zu verschiedenen Fast Food Ketten, aussagekräftig.

4.2. Sweet tooth

Wir möchten uns mit der Frage beschäftigen, ob Frauen grundsätzlich süßer essen als Männer. Hierzu betrachten wir für jede Person in unserem Datensatz den mittleren Zuckerkonsum pro Kalorie pro Tag und möchten diese nun geschlechterspezifisch Vergleichen.

Wie schon in Abschnitt 3 dargelegt möchten wir hierzu einen Hypothesentest anwenden.

```
t.test(data_female$mean,data_male$mean,alternative="greater",conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_female$mean and data_male$mean
## t = 1.6249, df = 161.83, p-value = 0.05306
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -6.886653e-05           Inf
## sample estimates:
##  mean of x  mean of y
## 0.04404281 0.04023542
```

Wir erhalten also einen Wert von p=.05306, also insbesondere ist $p > \alpha = 0.05$. Damit kann die Nullhypothese, dass Frauen weniger Zucker pro Kalorie essen, nicht verworfen werden.

Es ist also keine Aussage über einen Unterschied im Konsumverhalten von Männern und Frauen möglich.

Literatur

[1] Welch, B. L. (1947). *The generalization of "Student's" problem when several different population variances are involved* Biometrika. 34(1–2): 28–35

[2] Zimmermann, D.W. (2004). *A note on preliminary tests of equality of variances* British Journal of Mathematical and Statistical Psychology 57(Pt 1): 173-81

[3] Kolmogorov, A. (1933). *Sulla determinazione empirica di una lgge di distribuzione.* Inst. Ital. Attuari, Giorn. 4: 83-91

[4] Lehmann, E.L. (1998). *Elements of Large-Sample Theory* Springer: 192