Datenanalyse mit R

6 Beschreibende Statistik: Einzelvariablen

Tobias Wiß, Carmen Walenta und Felix Wohlgemuth

23.04.2020

Vielen Dank für Ihre rege Beteiligung am R-Teil des Kurses !!

Ihre Skripte sind sehr gut und ich bin beigeistert, dass Sie nach kurzer Zeit so gut mit R umgehen können.

All das was Sie in den Übungen geschrieben haben, können Sie direkt für die Seminararbeit verwenden. Passen Sie einfach die Skripte an Ihre Fragestellung an und dann haben Sie schon den R-Teil für die Seminararbeit.

In den letzten Einheiten haben Sie die grundlegenden Funktionen für ein tidy dataset kennen gelernt. In den nächste Einheiten werden wir lernen, wie man die interessanten Informationen aus den Daten bekommt und wie man diese visualisiert.

Neuer Ablauf des R-Teils ab 24.04.

Falls Sie Probleme bei den Übungen haben oder Fragen zu den Folien, dann nutzen Sie bitte die Sprechstunde und das R Forum auf moodle.

Um einen besseren Austausch zu ermöglichen, gibt es jetzt einen neuen Ablauf des R-Teils:

- Die Folien und die Übungsaufgabe der Woche werden am Freitag Abend auf moodle hochgeladen.
- Bitte schauen Sie sich die Folien an und notieren Sie sich alle Unklarheiten und welche Themen genauer erläutert werden sollen.
- Bitte versuchen Sie die Übung zu lösen und notieren Sie sich alle Probleme (Sie können jederzeit Ihre Frage im R Forum stellen und auch Ihre Kolleg*innen mit Ihrem Wissen unterstützen).
- Am **Anschluss der inhatlichen Live-Sitzung zu Familienpolitik am Donnerstag** oder alternativ am **Donnerstag von 11:00 bis 11:45** gibt es dann eine R Sprechstunde über zoom (Link und Passwort auf moodle). Hier besprechen wir dann alle Themen und Fragen. Sie können natürlich auch ohne Fragen an der Sprechstunde teilnehmen.
- Ihr R-Skript für die Übungsaufgabe können Sie bis Freitag 12:00 auf moodle hochladen.
- Das Lösungsskript wird am Freitag um 13:00 auf moodle veröffentlicht.

Neuer Ablauf des R-Teils ab 24.04.

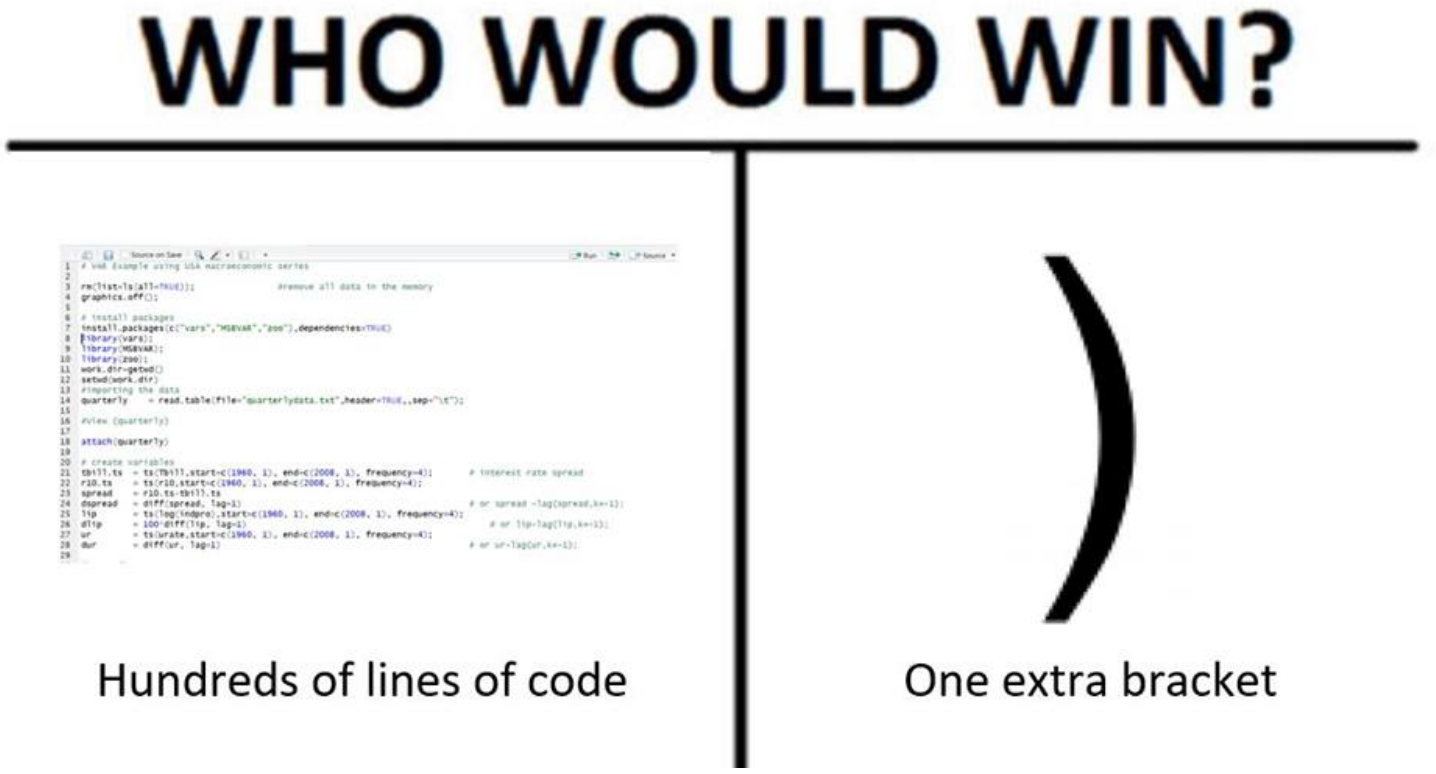
Wann?	Was passiert?
Freitag Abend (Woche vor der Sprechstunde)	neue Folien und Übungsaufgabe auf moodle
Donnerstag 11:00 oder nach der Live-Familienpolitik Sitzung	R-Sprechstunde
Freitag 12:00	Abgabe Übungsaufgabe auf moodle
Freitag 13:00	Lösungsskript wird veröffentlicht

Neuer Ablauf - Ausnahme diese Woche!

Diese Woche (bis 23.04.) läuft wegen der Umstellung nach einer eigenen Logik ab:

- Dienstag Abend Folien auf moodle.
- Mittwoch 12:00 Abgabe Übungsaufgabe 6.
- Mittwoch 13:00 Lösungsskript für Übungsaufgabe 6.
- Donnerstag nach Familienpolitik-Live-Sitzung R-Sprechstunde zu Folien 7 und allen Themen, die Sie besprechen wollen.
- Die heutige Übung 7 ist optional. Sie können Sie jederzeit bis Mittwoch nächste Woche hochladen. Falls Sie keine Übung abgeben, dann zählt das nicht zu Ihren 4 freien Nicht-Agaben. Ich gebe Ihnen gerne Feedback zu Ihrer Abgabe!

R ist sehr empfindlich bei Eingabefehler!



Source: R Memes for Statistical Fiends https://www.reddit.com/r/rstatsmemes/

R ist sehr empfindlich bei Eingabefehler!

- Fall Sie eine Fehlermeldung bekommen und R die Ausführung des Skriptes stoppt, dann fehlt sehr oft eine nicht geschlossene Klammer oder ein Variablennamen oder Funktionsnamen ist falsch geschrieben.
- Die Fehlermeldung ist nicht immer sehr hilfreich, aber sehr oft gibt R die Zeilennummer an wo der Fehler begonnen hat. RStudio hilft mit einem roten X neben der Zeilennummer oder einer rot unterkringelten Klammer.
- Als nächstes hilft leider nur genaues Lesen des Codes und ausbessern.
- Sie können auch die Fehlermeldung googlen. Gerade auf stack overflow gibt es andere User mit dem gleichen Problem. Sie sind sicher nicht allein mit dem Problem!

Übung 6

Die Lösungsskript für die Übung 6 finden Sie ab Mittwoch 13:00 auf moodle.

Das Lösungsskript ist als html-Report abgespeichert. Es beinhaltet alle Befehle / Funktionen und den Output in einem Dokument.

Was haben wir letzte Woche gelernt?

dplyr = Tools für Datenaufbereitung

Source: Chris Vaccaro https://medium.com/@chrisvaccaro_78233/the-absolute-fastest-way-to-learn-r-for-data-science-606ab2b28b7e/

dplyr - Datenaufbereitung

`dplyr` ist Teil des `tidyverse`-Pakets. Am Anfang des Skripts `library(tidyverse)` laden!

- Mit `%>%` können Befehle hintereinander gereiht werden. Vor der ersten `%>%` muss spezifiziert werden auf welchen Dataframe sich die nachfolgende Funktionen beziehen (im Beispiel: `socx_data`).
- Mit `select()` werden Variablen (Spalten) ausgewählt. Der Vorteil von `select()` ist, dass die Auswahl per Variablennamen und nicht per Position funktioniert.
- Mit `filter()`werden Beobachtungen (Zeilen) ausgewählt. Die nachfolgende Behfehle beziehen sich auf die ausgewählten Zeilen.
- Mit `rename()` können Variablen umbenannt werden.
- Mit `mutate()` werden neue Variable erstellt oder existierende Variablen bearbeitet und überschrieben.
- Mit `drop_na()` werden alle Zeilen ohne Wert gelöscht.
- `arrange()` ordnet den Datensatz basierend auf einer oder mehreren Variablen neu an. In der Grundeinstellung wird der Datensatz aufsteigend angeordnet mit – vor dem Variablennamen wird der Datensatz absteigend angeordnet.

logische Operatoren in R

Für die `filter()` und `select()` Befehle werden logische Operatoren benötigt, um R zu sagen welche Variablen oder Zeilen ausgewählt werden sollen.

Operatoren	Beschreibung
!	NOT (Gegenteil einer logischen Aussage)
==	gleich
!=	ungleich
<	kleiner
>	größer
<=	kleiner gleich
>=	größer gleich
`%in%`	wählt Elementen aus einer Menge `c()` aus, zB für `filter()`

Zur Verknüpfung von logischen Tests können Sie | für *oder* und & für *und* verwenden.

dplyr - Gruppieren & Zusammenfassen

- Mit `group_by()` wird der Datensatz nach den Werten der ausgewählten Variable gruppiert. Alle folgende Berechnungen werden für jede Gruppe separat durchgeführt
- Mit `summarise()` können unterschiedliche Kennzahlen wie `mean()` oder `min()` und `max()` berechnet werden. **Das behandeln wir heute genauer**.

dplyr - Beispiel

```
library(tidyverse)
# load dataset
socx_data <- read_csv("_raw/SOCX_AGG_31032020142101957.csv") # impor

# clean data
socx_data <- socx_data %>%              # select dataframe and replace
  select(COUNTRY, YEAR, Value) %>%     # select essential variables
  rename(fampol_exp_pct = Value) %>%   # rename Value Variable
  mutate(fampol_exp_pct = fampol_exp_pct / 100) # recode expenditure

# average public expenditure on family policy for Germany (2000 - 20
socx_data %>%                          # select dataframe
  filter(COUNTRY == "DEU") %>%         # select only German data
  summarise(mean_fampol_exp_pct = mean(fampol_exp_pct)) # specify mea
```

dplyr - Beispiel

Die Befehle zeigen sehr gut den Unteschied zwischen == und = in R:

- == ist eine logische Aussage. Hier ist die logische Aussage `COUNTRY == "DEU"` welche mit `TRUE` und `FALSE` beantwortet wird. R wählt alle Zeilen mit `TRUE` aus.
- = bedeutet bei `mutate()`, dass die neue Variable `fampol_exp_pct` aus `fampol_exp_pct / 100` besteht. Hier legt = den Inhalt fest.

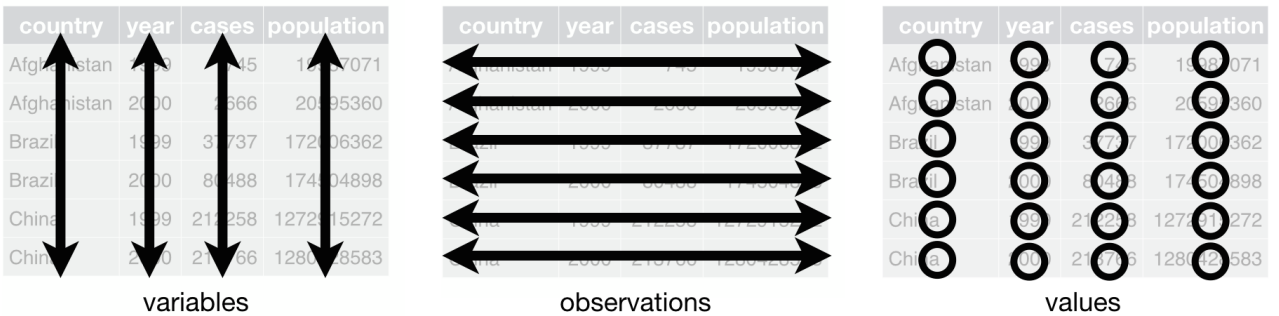
dplyr - Beispiel

```
# average public expenditure on family policy for all countries in sa
mean_fampol_exp_by_country <-  socx_data %>%    # select dataframe ar
  group_by(COUNTRY) %>%                         # group by countries
  summarise(mean_fampol_exp_pct = mean(fampol_exp_pct)) %>% # specify
  print()                                       # print results first

# top 5 countries with most family policy expenditure in 2000
socx_data %>%                     # select dataframe
  filter(YEAR == 2000) %>%        # select only 2000 values
  arrange(-fampol_exp_pct) %>%    # arrange data ascending by family po
  head( n = 5)                    # print first 5 values
```

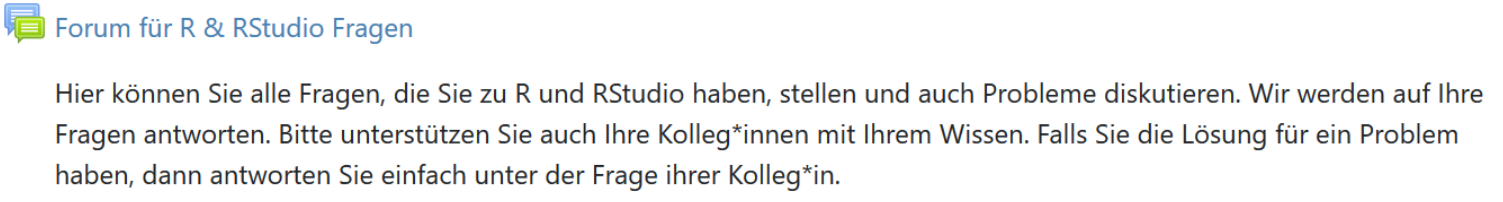
tidy Data

Jede Zeile eine Beobachtung, jede Spalte eine Variable, jede Zelle ein Wert.



Source: https://r4ds.had.co.nz/tidy-data.html

Falls Sie noch Fragen haben, nutzen Sie das **Forum** auf moodle und unterstützen Sie Ihre Kolleg*innen mit Ihrem Wissen!



Nutzen Sie auch unsere **R Sprechstunde**.
Jeden Donnerstag von 11:00 bis 11:45 auf zoom (link finden Sie auf moodle).

Deskriptive Statistik: Einzelvariablen

Datensatz für die heutige Sitzung

- Die folgende Beispiele basieren wieder auf dem OECD SOCX Datensatz der letzten Woche.
- Zusätzlich zu den insgesamten öffentlichen Ausgaben für Familienpolitik, beinhaltet der Datensatz jetzt die Ausgaben für die einzelnen familienpolitischen Instrumente (alle Variablen sind in % BIP).

Datensatz für die heutige Sitzung

```
library(tidyverse)
socx_data <- read_csv("_raw/SOCX_AGG_20042020191205895.csv")
```

```
variable.names(socx_data)
```

```
##  [1] "COUNTRY"
##  [2] "YEAR"
##  [3] "family_cash_allowances_pct_gdp"
##  [4] "family_cash_leave_pct_gdp"
##  [5] "family_cash_other_cash_pct_gdp"
##  [6] "family_cash_total_pct_gdp"
##  [7] "family_service_accomodation_pct_gdp"
##  [8] "family_service_childcare_pct_gdp"
##  [9] "family_service_other_services_pct_gdp"
## [10] "family_service_total_pct_gdp"
## [11] "family_total_total_pct_gdp"
```

Sie finden die Daten auf moodle.

Deskriptive Statistik mit dplyr

`summarise()` ist die Grundfunktion von `dplyr` für die Berechnung von deskriptiven Statistikmaßen:

- `summarise()` fasst eine Variable im Dataframe zu einem oder mehreren Werten zusammen.
- `summarise()` berechnet Lagemaße und auch Streuungsmaße, wie zB[1]:
  - Center: mean(), median()
  - Spread: sd(), IQR(), mad()
  - Range: min(), max(), quantile()
  - Position: first(), last(), nth(),
  - Count: n(), n_distinct()
- In `summarise` können Sie auch eigene Maße berechnen, die auf diesen Funktionen basieren (zB die Differenz des Durschnitts einer Variable und dem Durchschnitt einer zweiten Variable).
- `summarise()` ist Teil des `dplyr`-Pakets, d.h. es ist kompatibel mit `%>%`, berücksichtigt `group_by` und gibt ein Dataframe aus.

[1] In der Lektüre zur heutigen Sitzung werde die Lagemaße und Streuungsmaße nochmal genau erläutert. Bitte schauen Sie in Sauer: S. 103-112 nach, falls Ihnen die Bedeutungen nicht ganz klar sind.

summarise()

Die Schreibweise der `summarise()` Funktion folgt einem simplem System:

- Der verwendete Dataframe wird entweder mit `%>%` vor `summarise()` definiert oder direkt im Befehl an erster Stelle: `summarise(.data, ...)` (diese Logik gilt für alle `tidyverse` Funktionen).
- Als nächstes müssen die Maße definiert werden, die berechnet werden sollen:
  - Zuerst den Namen des Maßes definieren (Sie haben hier volle Freiheit - außer Umlaute) zB: `niedrigste =`
  - Dann die Funktion definieren mit der das Maß berechnet wird `niedrigste = min()`
  - Dann die Variable auswählen auf die sich die Funkion beziehen soll `niedrigste = min(family_service_childcare_pct_gdp)`
  - Nach einem `,` können sie das nächste Maß berechen. Es kann eine komplett andere Funktion sein und sich auf eine andere Variable beziehen.
- Das Ergebnis ist ein Dataframe mit einem Wert pro Maß
- Sie können das Ergebnis abspeichern, dafür muss am Anfang der `%>%`-Kette ein Namen definiert und mit `<-` zugewießen werden: `descriptives <- ...`

summarise() - Lagemaße

Wir arbeiten mit dem `dplyr` Paket und Funktionen aus anderen Paketen des tidyverse. Deshalb laden wir am Anfang das `tidyverse`, wodurch `dplyr` auch gleich geladen wird.

```
library(tidyverse)
```

```
# Die niedrigsten, mittleren und höchsten Ausgaben
# für formale Kinderbetreuung im Jahr 2015 der OECD Länder
descriptives <- socx_data %>%
  filter(YEAR == 2015) %>%
  summarise(niedrigste = min(family_service_childcare_pct_gdp),
            mittlere = median(family_service_childcare_pct_gdp),
            hoechste = max(family_service_childcare_pct_gdp))
print(descriptives)
```

```
## # A tibble: 1 x 3
##   niedrigste mittlere hoechste
##        <dbl>    <dbl>    <dbl>
## 1      0.083      0.6     1.80
```

summarise() - fehlende Werte

Vorsicht:

Die meisten Funktionen für deskriptive Statistik wie `median()`, `min()`, `max()` oder `median()` haben die Grundeinstellung `na.rm = FALSE`. Falls sich in den Daten der ausgewählten Variable ein fehlender Wert befindet, gibt R eine Fehlermeldung (siehe Beispiel auf der)

Die `family_service_accomodation_pct_gdp` Variable existiert nicht für alle Länder, deshalb sind einige NAs in der Variable. R berechnet deshalb die Maße nicht.

```
socx_data %>%
  filter(YEAR == 2015) %>%
  summarise(niedrigste = min(family_service_accomodation_pct_gdp),
            mittlere = median(family_service_accomodation_pct_gdp),
            hoechste = max(family_service_accomodation_pct_gdp))
```

```
## # A tibble: 1 x 3
##   niedrigste mittlere hoechste
##        <dbl>    <dbl>    <dbl>
## 1         NA       NA       NA
```

summarise() - fehlende Werte

Sie können entweder die ganze Zeile mit `drop_na()` löschen oder innerhalb der Funktion `na.rm = TRUE` einstellen, jetzt berücksichtigt die Funktion die fehlende Werte nicht (zB `niedrigste = min(family_service_accomodation_pct_gdp, na.rm = TRUE)`). `na.rm = TRUE` ist zwar sehr praktisch, aber es ist nicht ganz eindeutig welche Zeilen jetzt wegfallen.

```
socx_data %>%
  filter(YEAR == 2015) %>%
  summarise(niedrigste = min(family_service_accomodation_pct_gdp,
                             na.rm = TRUE),
            mittlere = median(family_service_accomodation_pct_gdp,
                              na.rm = TRUE),
            hoechste = max(family_service_accomodation_pct_gdp,
                           na.rm = TRUE))
```

```
## # A tibble: 1 x 3
##   niedrigste mittlere hoechste
##        <dbl>    <dbl>    <dbl>
## 1          0   0.0410    0.445
```

summarise() - fehlende Werte

Mehr Kontrolle hat man, wenn man vorher die Zeilen mit fehlende Werte per `drop_na()` löscht, das kann zB direkt nach `filter()` gemacht werden.

```
socx_data %>%
  filter(YEAR == 2015) %>%
  drop_na(family_service_accomodation_pct_gdp) %>%
  summarise(niedrigste = min(family_service_accomodation_pct_gdp),
            mittlere = median(family_service_accomodation_pct_gdp),
            hoechste = max(family_service_accomodation_pct_gdp))
```

```
## # A tibble: 1 x 3
##   niedrigste mittlere hoechste
##        <dbl>    <dbl>    <dbl>
## 1          0   0.0410    0.445
```

summarise() - Streuungsmaße

Mit `summarise()` können auch mehrere Variablen verglichen werden, zB anhand von Streuungsmaßen.

Sie können zB überprüfen, ob alle Länder im Sample im Jahr 2015 für die unterschiedlichen familienpolitische Instrumente ähnlich viel oder unterschiedlich viel ausgegeben haben.

Haben Barleistungen & Karenz im Jahr 2015 eine kleinere oder größere Streuung als Sachleistungen? Bei welchen Ausgaben ähneln sich die Länder mehr?

summarise() - Streuungsmaße

```
socx_data %>%
  filter(YEAR == 2015) %>%
  summarise(Var_cash = var(family_cash_total_pct_gdp),
            Var_services = var(family_service_total_pct_gdp),
            sd_cash = sd(family_cash_total_pct_gdp),
            sd_services = sd(family_service_total_pct_gdp),
            IQR_cash = IQR(family_cash_total_pct_gdp),
            IQR_services = IQR(family_service_total_pct_gdp))
```

```
## # A tibble: 1 x 6
##   Var_cash Var_services sd_cash sd_services IQR_cash IQR_services
##      <dbl>        <dbl>   <dbl>       <dbl>    <dbl>        <dbl>
## 1    0.327        0.309   0.572       0.556    0.696        0.570
```

Die unterschiedlichen Streuungsmaße (Varianz = `var()`, Standardabweichung = `sd()`, Abstand zwischen dem 75%- und 25%-Quartil = `IQR()`) sind für Sachleistungen (services) niedriger als für Barleistungen & Karenz (cash). D.h.: im Jahr 2015 hatten die OECD Länder ähnlichere Ausgaben für Sachleistungen als für Barleistungen & Karenz.

summarise() - Lagemaße

Streuungsmaße allein zeigen nur das halbe Bild. Die Ausgaben für Sachleistungen sind zwar ähnlicher, aber wie hoch sind Ausgaben für Sachleistungen im Vergleich zu Barleistungen & Karenz? Lagemaße stellen das dar.

summarise() - Lagemaße

```
socx_data %>%
  filter(YEAR == 2015) %>%
  summarise(min_cash = min(family_cash_total_pct_gdp),
            min_services = min(family_service_total_pct_gdp),
            mean_cash = mean(family_cash_total_pct_gdp),
            mean_services = mean(family_service_total_pct_gdp),
            max_cash = max(family_cash_total_pct_gdp),
            max_services = max(family_service_total_pct_gdp))
```

```
## # A tibble: 1 x 6
##   min_cash min_services mean_cash mean_services max_cash max_services
##      <dbl>        <dbl>     <dbl>         <dbl>    <dbl>        <dbl>
## 1    0.074          0.1      1.22         0.940     2.51         2.38
```

Die OECD Länder haben durchschnittlich weniger für Sachleistungen als für Barleistungen & Karenz ausgegeben. D.h.: Ausgaben für Sachleistungen der OECD Länder im Jahr 2015 waren ähnlicher als Barleistungen aber auch niedriger. Es gibt mehr Unterschiede bei Ausgaben für Barleistungen, die aber auch durschnittlich höher als Ausgaben für Sachleistungen sind. Nächste Woche werden wir das mit Boxplots noch besser erkennbar machen.

summarise() & Gruppierungen

Unser Datensatz ist eine Zeitreihe mit Daten von 2000 bis 2015 für jedes OECD Land. Wir können damit auch Veränderungen der Länder zwischen 2000 und 2015 für die unterschiedlichen familienpolitischen Instrumente vergleichen. Mit `group_by()` aus dem `dplyr`-Paket sagen Sie der `summarise()`-Funktion für welche Gruppen die Lage- und Streuungsmaße berechnet werden sollen.

```
socx_data %>%
  filter(YEAR >= 2000 & YEAR <= 2015) %>%
  group_by(COUNTRY) %>%
  summarise(Var_cash = var(family_cash_total_pct_gdp),
            Var_services = var(family_service_total_pct_gdp),
            sd_cash = sd(family_cash_total_pct_gdp),
            sd_services = sd(family_service_total_pct_gdp),
            IQR_cash = IQR(family_cash_total_pct_gdp),
            IQR_services = IQR(family_service_total_pct_gdp))
```

summarise() & Gruppierungen

Show 7 entries Search:

	COUNTRY	sd_cash	sd_services	IQR_cash	IQR_services
1	AUS	0.274	0.108	0.425	0.157
2	AUT	0.227	0.114	0.414	0.206
3	BEL	0.057	0.093	0.088	0.124
4	CAN	0.13	0.024	0.146	0.032
5	CHE	0.044	0.063	0.066	0.056
6	CHL	0.131	0.188	0.159	0.343
7	CZE	0.231	0.041	0.391	0.052

Showing 1 to 7 of 36 entries

Previous 1 2 3 4 5 6 Next

summarise() & Gruppierungen

Es zeigt sich schon, dass die Streuungen zwischen den Ländern sehr unterschiedlich sind.

Es gibt Länder wie die USA die eine geringe Streuung bei Bar- und Sachleistungen haben, d.h. die Ausgaben von Jahr zu Jahr je Sach- und Barleistung waren zwischen 2000 und 2015 ähnlich.

Es gibt aber auch Länder wie Korea mit einer großen Streuung bei Sachleistungen und einer niedrige Streuung bei Barleistungen. In Österreich ist die Streuung gerade umgekehrt.

Die Veränderungen je familienpolitischen Instrument sind zwischen den Ländern sehr unterschiedlich.

Um das besser vergleichen zu können, sollten wir uns noch die Lagemaße anschauen. Und in einem zweiten Schritt die Werte für 2000 und 2015, also am Anfang und Ende der Zeitreihe, vergleichen.

summarise() & Gruppierungen

```
socx_data %>%
  filter(YEAR >= 2000 & YEAR <= 2015) %>%
  group_by(COUNTRY) %>%
  summarise(min_cash = min(family_cash_total_pct_gdp),
            min_services = min(family_service_total_pct_gdp),
            max_cash = max(family_cash_total_pct_gdp),
            max_services = max(family_service_total_pct_gdp),
            mean_cash = max(family_cash_total_pct_gdp),
            mean_services = max(family_service_total_pct_gdp),
            med_cash = median(family_cash_total_pct_gdp),
            med_services = median(family_service_total_pct_gdp)))
```

summarise() & Gruppierungen

Show 7 entries Search:

	COUNTRY	min_cash	min_services	max_cash	max_services
1	AUS	1.74	0.588	2.577	0.914
2	AUT	1.926	0.41	2.584	0.692
3	BEL	1.66	0.719	1.838	1.05
4	CAN	0.743	0.143	1.317	0.235
5	CHE	1.052	0.25	1.218	0.498
6	CHL	0.383	0.421	0.83	0.946
7	CZE	1.17	0.443	1.857	0.575

Showing 1 to 7 of 36 entries

Previous 1 2 3 4 5 6 Next

summarise() & Gruppierungen

Show 7 entries Search:

	COUNTRY	mean_cash	mean_services	med_cash	med_services
1	AUS	2.06	0.71	1.97	0.67
2	AUT	2.28	0.52	2.33	0.45
3	BEL	1.75	0.94	1.74	0.94
4	CAN	0.97	0.2	0.98	0.21
5	CHE	1.16	0.32	1.17	0.3
6	CHL	0.59	0.62	0.6	0.6
7	CZE	1.53	0.51	1.53	0.52

Showing 1 to 7 of 36 entries

Previous 1 2 3 4 5 6 Next

summarise() & Gruppierungen

Die Lagemaße verfestigen das Bild, dass wir schon von der Analyse des Jahres 2015 bekommen haben. Die Spannweite bei den Sachleistungen ist niedriger als die Spannweite der Ausgaben für Barleistungen & Karenz.

Das Minimum und das Maximum darf aber nicht als Start- und Endwert der Zeitreihe von 2000 bis 2015 interpretiert werden. Es kann sein, dass in einem Land die Ausgaben konstant stiegen, in einem anderen Land die Ausgaben konstant sanken oder schwankten. Am besten erkennt man das mit einer Visualiseerung des zeitlichen Verlaufs (kommt in den nächsten Wochen).

Wir schauen uns jetzt die Werte für 2000 und 2015 getrennt an. Daraus kann man erste Indizien ziehen von wo bis wohin sich die Ausgaben pro Land entwickelt haben.

summarise() & Gruppierungen

Um die Daten von 2000 und 2015 vergleichen zu können, generieren wir ein neues Dataframe nur mit den Werten von 2000 und 2015. D.h. wir löschen alle Zeilen außer den zwei Jahren. Damit haben wir den Start- und Endwert der Zeitreihe.

```
socx_2000_2015 <- socx_data %>%
  filter(YEAR == 2000 | YEAR == 2015) %>%
  select(COUNTRY, YEAR, family_cash_total_pct_gdp, family_service_tot
  print()
```

summarise() & Gruppierungen

Show 7 entries Search:

	COUNTRY	YEAR	family_cash_total_pct_gdp	family_services_total_pct_gdp
1	AUS	2000	2.314	0.594
2	AUS	2015	1.74	0.914
3	AUT	2000	2.519	0.41
4	AUT	2015	1.953	0.692
5	BEL	2000	1.731	0.719
6	BEL	2015	1.786	1.037
7	CAN	2000	0.743	0.143

Showing 1 to 7 of 71 entries

Previous 1 2 3 4 5 … 11 Next

summarise() & Gruppierungen

Die Werte für 2000 und 2015 zeigen, dass die Entwicklung zwischen 2000 und 2015 in einigen Ländern angestiegen und in anderen Ländern gesunken sind. Die Streuungsmaße und Lagemaße haben diese unterschiedlichen Entwicklungen nicht dargestellt. Deshalb ist es bei der deskriptiven Analyse immer wichtig auch nochmal in die Daten direkt zu schauen und ausgewählte Werte genauer zu betrachten.

(Wir können die die Entwicklung nochmals genauer anschauen und die Differenz zwischen dem Wert im Jahr 2015 und 2000 berechnen, das ist aber ein wenig komplizierter. Bei interesse kann ich das gerne in den nächsten Wochen zeigen).

Relative Häufigkeiten

Um herauszufinden ob die Länder in einem Jahr mehr für Sachleistungen oder Barleistungen & Karenz ausgeben, können wir schauen in welchem Verhältnis diese zu den Gesamtausgaben stehen.

Dazu wird mit `mutate()` zwei neue Variablen erstellt, in denen das Verhältnis der zwei familienpolitischen Instrumenten pro Land für das Jahr 2015 berechnet wird (ohne `filter(YEAR==2015)` würde R das für jedes Jahr berechnen). Der neue Datensatz beinhaltet nur Daten für das Jahr 2015, also eine Zeile ist gleich ein Land im Jahr 2015.

```
socx_2015 <- socx_data %>%
  filter(YEAR == 2015) %>%
  mutate(cash_prop = (family_cash_total_pct_gdp / family_total_total_
  mutate(services_prop = (family_service_total_pct_gdp / family_total
```

Hier kann `group_by(COUNTRY)` weggelassen werden, weil R die neuen Werte für jede Zeile berechnet und das ist automatisch ein Land im Jahr 2015.

Relative Häufigkeiten

Show 7 entries Search:

	COUNTRY	cash_prop	services_prop
1	AUS	65.56	34.44
2	AUT	73.84	26.16
3	BEL	63.27	36.73
4	CAN	84.86	15.14
5	CHE	70.94	29
6	CHL	44.44	55.61
7	CZE	73.68	26.27

Showing 1 to 7 of 35 entries

Previous 1 2 3 4 5 Next

Relative Häufigkeiten

Um das Verhältnis in Zusammenhang zu den anderen Ländern zu setzen können wir das arithmetische Mittel berechnen und dann im zweiten Schritt schauen, um wie viel Prozentpunkte jedes Land vom OECD Durschnitt abweicht.

Wir müssen das Jahr nicht filtern, da unser neuer Datensatz `socx_2015` nur Werte für 2015 beinhaltet.

```
socx_2015 %>%
  summarise(mean_cash_prop =  mean(cash_prop),
            mean_services_prop = mean(services_prop))
```

```
## # A tibble: 1 x 2
##   mean_cash_prop mean_services_prop
##            <dbl>              <dbl>
## 1           56.0               44.0
```

Durschnittlich gaben die OECD Länder mehr für Barleistungen & Karenz aus als für Sachleistungen.

Relative Häufigkeiten

Um wie viel Prozentpunkte weichen die Länder von diesem OECD Durchschnitt ab?

```
socx_2015 <- socx_2015 %>%
  mutate(cash_prop_diff = cash_prop - mean(cash_prop)) %>%
  mutate(services_prop_diff = services_prop - mean(services_prop))
```

Hier kann `group_by(COUNTRY)` weggelassen werden, weil R die neuen Werte für jede Zeile berechnet und das ist automatisch ein Land im Jahr 2015.

Relative Häufigkeiten

Show 7 entries Search:

	COUNTRY	cash_prop_diff	services_prop_diff
1	AUS	9.57	-9.56
2	AUT	17.84	-17.84
3	BEL	7.27	-7.27
4	CAN	28.87	-28.86
5	CHE	14.95	-15
6	CHL	-11.55	11.61
7	CZE	17.68	-17.73

Showing 1 to 7 of 35 entries

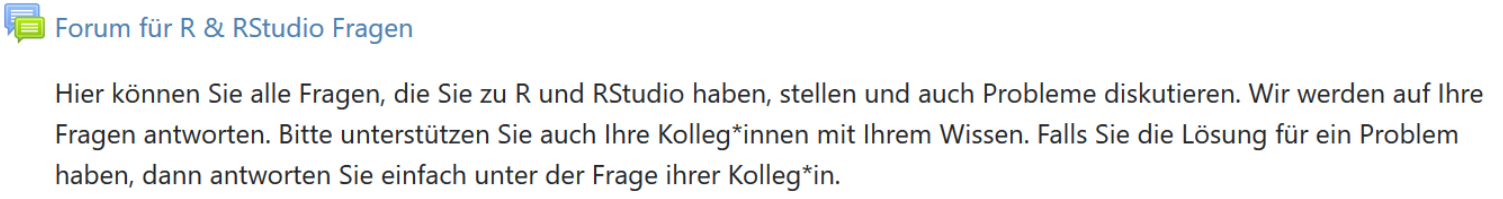
Previous 1 2 3 4 5 Next

Übung 7 - optional

Die Übung 7 für diese Woche ist freiwillig. Ich gebe Ihnen gerne Feedback auf Ihre Abgabe, aber Sie müssen nichts abgeben. Falls Sie nichts abgeben, zählt das nicht zu Ihren 4 freien Nicht-Abgaben.

- Laden Sie den Datensatz "SOCX_AGG_20042020191205895.csv" von moodle und importieren Sie diesen in R (Vergessen Sie nicht die notwendigen Pakete am Anfang des Skripts zu laden).
- Wählen Sie eine Variable Ihrer Wahl.
- Filtern Sie den Datensatz, so dass die Variable nur für ein Jahr Ihrer Wahl ausgegeben wird.
- Berechnen Sie ein passendes Lagemaß und ein passendes Streuungsmaß für dieses Jahr.
- Verwenden Sie `group_by` und einen Zeitraum Ihrer Wahl (Sie müssen dafür den ungefilterten Datensatz verwenden).
- Berechnen Sie für den Zeitraum ein Lagemaß und ein Streuungsmaß für jedes Land getrennt.
- Laden Sie ihr R-Skript bis Mittwoch 29.04. 12:00 auf moodle hoch.

Falls Sie noch Fragen haben, nutzen Sie das **Forum** auf moodle und unterstützen Sie Ihre Kolleg*innen mit Ihrem Wissen!



Nutzen Sie auch unsere **R Sprechstunde**.
Jeden Donnerstag von 11:00 bis 11:45 auf zoom (link finden Sie auf moodle).