2. Summarizing Data with Numbers

Stat 241

Summarizing quantitative data

Generally we are interested in measures of **center** (what is a "typical" value?) and **spread** (are the values clustered close together or more spread out?).

Five Number Summary and Boxplots

For a set of numerical data:

- Median (Q2):
- First Quartile (Q1):
- Third Quartile (Q3):
- Min (Q0):
- Max (Q4):
- Interquartile Range (IQR):
- Five Number Summary:

By Hand Example

1. Compute the numerical and graphical summaries listed for the following data values:
   $3, 5, 2, 3, 7, 5, 8, 6, 7$
   a. Five Number Summary:
   b. IQR:
   c. Boxplot:

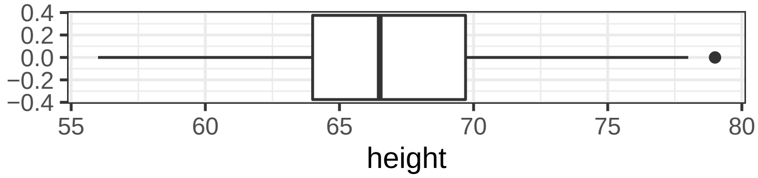
Five Number Summary Using R

Example: Find the Five Number Summary for the heights of the adult children in the data frame Galton.

```
quantile( ~ height, data = Galton)
```

```
##   0%  25%  50%  75% 100%
## 56.0 64.0 66.5 69.7 79.0
```

```
gf_boxplot( ~ height, data = Galton)
```



Boxplots and skew

2. Sketch a boxplot for each of the following situatoins.
    a. Boxplot of a data set skewed to right.

    b. Boxplot of a data set skewed to left.

Mean and Standard Deviation

Some notation

- n = number of data values
- x = variable
- $x_1, x_2, \ldots x_n$ = individual values of the variable x.

Mean

- mean of $x = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
    - "add them all up and divide by how many there are"

Example Computation

3. Compute the mean for $x : 1, 1, 3, 4, 6, 7, 7, 9$.

Things to note about the mean and median

- If the distribution is symmetric...

- If the distribution is skewed to the right...

- If the distribution is skewed to the left...

- If there are outliers in the data...

- If the distribution is not unimodal...

Variance and Standard Deviation

- sample variance of x = $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$
  - "add up squared differences from the mean and divide by 1 less than how many there are"
  - "sample" indicates that we only have data for a sample, not for all individuals; since that is the usual situation, we will drop the word sample.

- sample standard deviation of x = $s_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}}$

- Why do we square the differences?
  - If we do nothing...

  - If we use absolute value...

Small Example:

4. Compute the variance and standard deviation for x: $1, 2, 3, 4, 5$

Computing summary statistics with R

We won't usually compute these statistics by hand. The hand computations above are just to make sure we understand what these numbers are. But R can quickly and easily compute any of these statistics (and more) with the following template.

```
stat( ~ x, data = mydata)
```

We can compute statistics within groups using either of these forms:

```
stat( ~ x | z, data = mydata)
stat( y ~ x, data = mydata)
```

Example: Penguin body mass

```
library(palmerpenguins)
df_stats( ~ body_mass_g, data = penguins)
```

```
##      response  min   Q1 median   Q3  max     mean       sd   n missing
## 1 body_mass_g 2700 3550   4050 4750 6300 4201.754 801.9545 342       2
```

Notice that there are some missing values (penguins for which no body mass was recorded). This causes problems for some of our statistics. We can't really compute the mean of all of the penguins if we don't know all of the values. `df_stats()` lets us know that there are some missing values and that our statistics are computed only on the non-missing values. But the individual statistic functions report `NA` (not available, missing).

```
mean( ~ body_mass_g, data = penguins)
```

```
## [1] NA
```

We can fix this two ways:

```
mean( ~ body_mass_g, data = penguins, na.rm = TRUE)
```

```
## [1] 4201.754
```

```
df_stats( ~ body_mass_g, data = penguins, mean)
```

```
## Warning: Excluding 2 rows due to missing data [df_stats()].
```

```
##      response     mean
## 1 body_mass_g 4201.754
```

Let's end this example by comparing the body mass of the different penguin species. Notice the different output formats below.

```
mean( ~ body_mass_g | species, data = penguins, na.rm = TRUE)
```

```
##    Adelie Chinstrap    Gentoo
##  3700.662  3733.088  5076.016
```

```
df_stats( ~ body_mass_g | species, data = penguins, mean)
```

```
## Warning: Excluding 2 rows due to missing data [df_stats()].
```

```
##      response   species     mean
## 1 body_mass_g    Adelie 3700.662
## 2 body_mass_g Chinstrap 3733.088
## 3 body_mass_g    Gentoo 5076.016
```

Example: HELPrct

The data frame HELPrct contains data on an experiment testing a treatment of substance abusers. It is too large to conveniently display the entire data set, but we dan see the top few rows of the data using `head()`.

```
head(HELPrct)
```

```
##   age anysubstatus anysub cesd d1 daysanysub dayslink drugrisk e2b female
## 1  37            1    yes   49  3        177      225        0  NA      0
## 2  37            1    yes   30 22          2       NA        0  NA      0
## 3  26            1    yes   39  0          3      365       20  NA      0
## 4  39            1    yes   15  2        189      343        0   1      1
## 5  32            1    yes   39 12          2       57        0   1      0
## 6  47            1    yes    6  1         31      365        0  NA      1
##      sex g1b homeless i1 i2 id indtot linkstatus link       mcs      pcs pss_fr
## 1   male yes   housed 13 26  1     39          1  yes 25.111990 58.41369      0
## 2   male yes homeless 56 62  2     43         NA <NA> 26.670307 36.03694      1
## 3   male  no   housed  0  0  3     41          0   no  6.762923 74.80633     13
## 4 female  no   housed  5  5  4     28          0   no 43.967880 61.93168     11
## 5   male  no homeless 10 13  5     38          1  yes 21.675755 37.34558     10
## 6 female  no   housed  4  4  6     29          0   no 55.508991 46.47521      5
##   racegrp satreat sexrisk substance treat avg_drinks max_drinks
## 1   black      no       4   cocaine   yes         13         26
## 2   white      no       7   alcohol   yes         56         62
## 3   black      no       2    heroin    no          0          0
## 4   white     yes       4    heroin    no          5          5
## 5   black      no       6   cocaine    no         10         13
## 6   black      no       5   cocaine   yes          4          4
##   hospitalizations
## 1                3
```

```
## 2               22
## 3                0
## 4                2
## 5               12
## 6                1
```

That's a bit ugly because there are so many variables. Let's try `glimpse()` instead

```
glimpse(HELPrct)
```

```
## Rows: 453
## Columns: 30
## $ age              <int> 37, 37, 26, 39, 32, 47, 49, 28, 50, 39, 34, 58, 58...
## $ anysubstatus     <int> 1, 1, 1, 1, 1, 1, NA, 1, 1, 1, NA, 0, 1, 1, 1, 1, ...
## $ anysub           <fct> yes, yes, yes, yes, yes, yes, NA, yes, yes, yes, N...
## $ cesd             <int> 49, 30, 39, 15, 39, 6, 52, 32, 50, 46, 46, 49, 22,...
## $ d1               <int> 3, 22, 0, 2, 12, 1, 14, 1, 14, 4, 0, 3, 5, 10, 2, ...
## $ daysanysub       <int> 177, 2, 3, 189, 2, 31, NA, 47, 31, 115, NA, 192, 6...
## $ dayslink         <int> 225, NA, 365, 343, 57, 365, 334, 365, 365, 382, 36...
## $ drugrisk         <int> 0, 0, 20, 0, 0, 0, 0, 7, 18, 20, 8, 0, 0, 0, 0, 0,...
## $ e2b              <int> NA, NA, NA, 1, 1, NA, 1, 8, 7, 3, NA, NA, NA, 1, N...
## $ female           <int> 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0,...
## $ sex              <fct> male, male, male, female, male, female, female, ma...
## $ g1b              <fct> yes, yes, no, no, no, no, yes, yes, no, no, no, no...
## $ homeless         <fct> housed, homeless, housed, housed, homeless, housed...
## $ i1               <int> 13, 56, 0, 5, 10, 4, 13, 12, 71, 20, 0, 13, 20, 13...
## $ i2               <int> 26, 62, 0, 5, 13, 4, 20, 24, 129, 27, 0, 13, 31, 2...
## $ id               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16,...
## $ indtot           <int> 39, 43, 41, 28, 38, 29, 38, 44, 44, 44, 34, 11, 40...
## $ linkstatus       <int> 1, NA, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0...
## $ link             <fct> yes, NA, no, no, yes, no, no, no, no, no, no, no, ...
## $ mcs              <dbl> 25.111990, 26.670307, 6.762923, 43.967880, 21.6757...
## $ pcs              <dbl> 58.41369, 36.03694, 74.80633, 61.93168, 37.34558, ...
## $ pss_fr           <int> 0, 1, 13, 11, 10, 5, 1, 4, 5, 0, 0, 13, 13, 1, 1, ...
## $ racegrp          <fct> black, white, black, white, black, black, black, w...
## $ satreat          <fct> no, no, no, yes, no, no, yes, yes, no, yes, no, ye...
## $ sexrisk          <int> 4, 7, 2, 4, 6, 5, 8, 6, 8, 0, 2, 0, 1, 4, 8, 3, 4,...
## $ substance        <fct> cocaine, alcohol, heroin, heroin, cocaine, cocaine...
## $ treat            <fct> yes, yes, no, no, no, yes, no, yes, no, yes, yes, ...
## $ avg_drinks       <int> 13, 56, 0, 5, 10, 4, 13, 12, 71, 20, 0, 13, 20, 13...
## $ max_drinks       <int> 26, 62, 0, 5, 13, 4, 20, 24, 129, 27, 0, 13, 31, 2...
## $ hospitalizations <int> 3, 22, 0, 2, 12, 1, 14, 1, 14, 4, 0, 3, 5, 10, 2, ...
```

`inspect()` computes some summary statistics for each variable instead of showing the raw data (for a few rows). Give it a try:

```
inspect(HELPrct)
```

We can get summary statistics about age for each of the three abuse groups.

```
df_stats(age ~ substance, data = HELPrct)
```

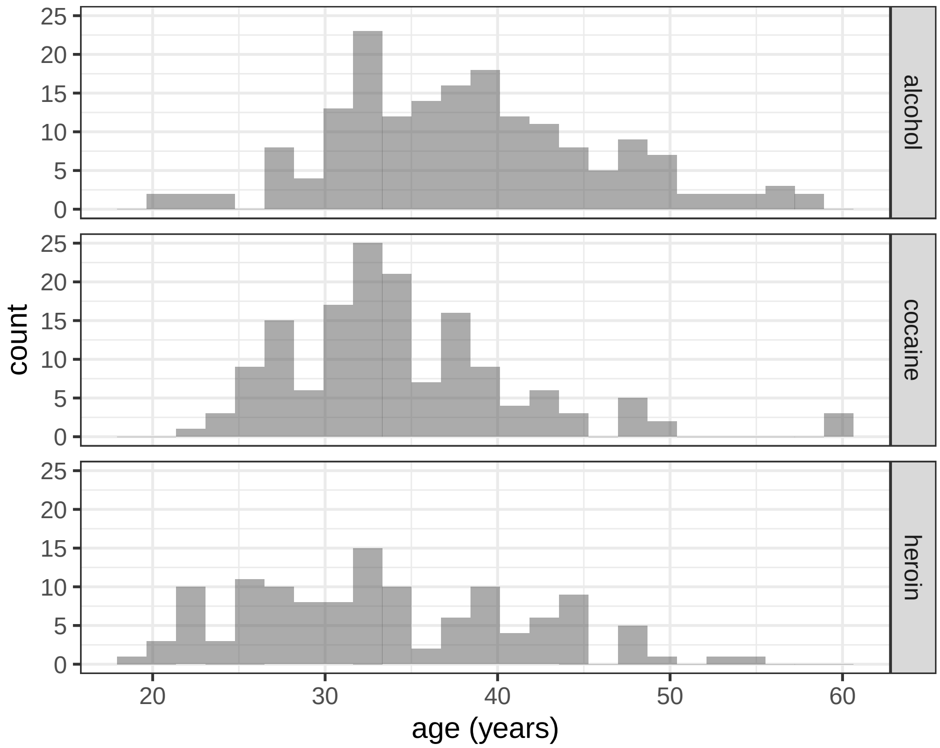
```
##   response substance min Q1 median    Q3 max     mean       sd   n missing
## 1      age   alcohol  20 33   38.0 43.00  58 38.19774 7.652272 177       0
## 2      age   cocaine  23 30   33.5 37.25  60 34.49342 6.692881 152       0
## 3      age    heroin  19 27   33.0 39.00  55 33.44355 7.986068 124       0
```

```
df_stats( ~ age | substance, data = HELPrct)
```

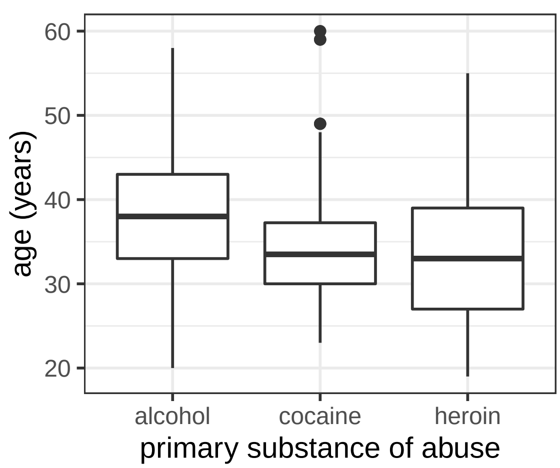
```
##   response substance min Q1 median    Q3 max     mean       sd   n missing
## 1      age   alcohol  20 33   38.0 43.00  58 38.19774 7.652272 177       0
## 2      age   cocaine  23 30   33.5 37.25  60 34.49342 6.692881 152       0
## 3      age    heroin  19 27   33.0 39.00  55 33.44355 7.986068 124       0
```

Histograms and boxplots by groups

```
gf_histogram( ~ age | substance ~ ., data = HELPrct)
```




```
gf_boxplot( age ~ substance, data = HELPrct)
```



Quantitative vs Categorical Variables

A quantitative variable is a variable that takes on numerical values in such a way that it makes sense to perform arithmetic operations on its values. A categorical variable is a variable whose values are categories. Body mass (in grams) is a quantitative variable;
`species` is a categorical variable.

Sometimes numbers are used to represent categories. Using numbers to represent categories does not turn a categorical variable into a quantitative variable, but it can confuse R, which will treat the variable as quantitative if we use numbers. We can explicitly tell R we have a categorical variable using `factor(x)`. (Factor is R's lingo for a categorial variable, it has nothing to do with factoring numbers.)

Example: Kids feet

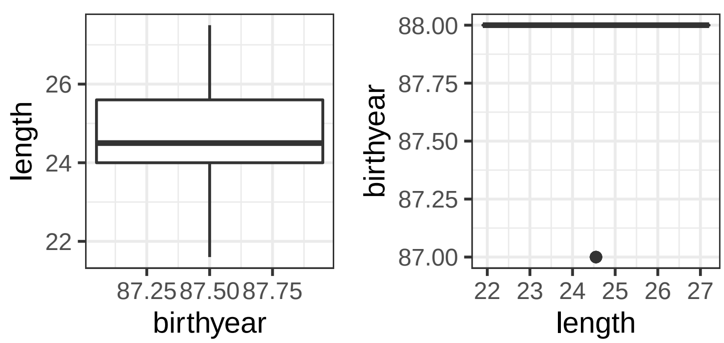
The data frame `KidsFeet` contains data on thirty-nine 4th grade children

```
head(KidsFeet)
```

```
##     name birthmonth birthyear length width sex biggerfoot domhand
## 1  David          5        88   24.4   8.4   B          L       R
## 2   Lars         10        87   25.4   8.8   B          L       L
## 3   Zach         12        87   24.5   9.7   B          R       R
## 4   Josh          1        88   25.2   9.8   B          L       R
## 5   Lang          2        88   25.1   8.9   B          L       R
## 6 Scotty          3        88   25.7   9.7   B          R       R
```

5. Which variables are quantitative and which are categorical?
6. What sorts of plots are appropriate to look at the distribution of a quantitative variable? Make a few of these for some of the quantitative variables.
7. What sorts of plots are appropriate to look at the distribution of a categorical variable? Make a few of these for some of the categorical variables.
8. Are birth month and birth year categorical or quantitative?
9. These plots probably surprise you. Can you figure out what R is doing? How can you fix them?

```
gf_boxplot(length ~ birthyear, data = KidsFeet) |
gf_boxplot(birthyear ~ length, data = KidsFeet)
```



10. Compute the mean foot length separately for boys and girls.
11. Create several different plots showing the distribution of foot lengths separately for boys and girls. Whic one do you prefer? Why?

EXERCISES

2.1 In the data frame `HELPrct`, the variable `avg_drinks` (also called `i1`) gives the average number of drinks per day, `substance` gives the kind of substance abuse, and `sex` gives the gender of the subject.

a. Create histograms comparing the average number of drinks per day by gender.
b. Repeat (a) with boxplots.
c. Compute the means and standard deviations for the daily number of drinks for male and for females.

d. Based on the information in (c), would you conclude that, on average, the males in the study consume more alcohol than the females?

e. For which gender is there more variability in alcohol consumption?
f. Create frequency histograms comparing the alcohol consumption by type of substance abuse.

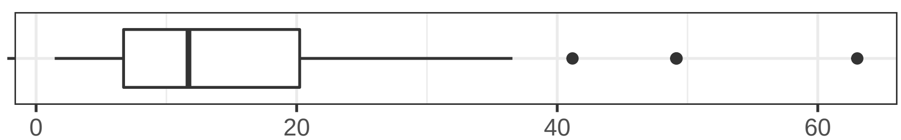
g. Is the distribution of alcohol consumption symmetric for those whose are alcohol abusers? If it is skewed, in which direction is it skewed?
h. Compute the median alcohol consumption for each of the three types of abuse.

2.2 Compute the mean, variance and standard deviation for the data set $x : 1, 5, 3, 7, 9$ by HAND.

2.3 Create a set of 6 numbers in the range 0 -– 10 (inclusive) that will have the largest possible standard deviation.

2.4 Below is the boxplot of a data set.

```
set.seed(1234)
gf_boxplot( "" ~ rgamma(50, shape = 1.5, scale = 10)) %>%
  gf_labs(y = "", x = "")
```



a. What is the median of the distribution?
b. What is the IQR for the distribution?
c. Is the mean larger or smaller than the median? How do you know?

2.5 From the data frame HELPrct, create a bargraph that shows how homelessness (`homeless`) is distributed with respect to (`sex`). Produce the version where the bars are adjacent (dodged). Should you use counts or proportions? Why? If proportions, what should the denominator be?

Based on the picture, should you conclude that males and females in the study are homeless at roughly the same rate or at different rates?