Data collection and data wrangling

The ASTA team

Contents

1 Data

1.1 Data example

We use data about pengiuns from the R package palmerpenguins

```
pingviner <- palmerpenguins::penguins
pingviner
```

```
## # A tibble: 344 x 8
##    species island    bill_length_mm bill_depth_mm flipp~ body~ sex    year
##    <fctr>  <fctr>              <dbl>         <dbl>  <int> <int> <fct> <int>
##  1 Adelie  Torgersen           39.1          18.7    181  3750 male   2007
##  2 Adelie  Torgersen           39.5          17.4    186  3800 fema~  2007
##  3 Adelie  Torgersen           40.3          18.0    195  3250 fema~  2007
##  4 Adelie  Torgersen           NA            NA       NA    NA <NA>   2007
##  5 Adelie  Torgersen           36.7          19.3    193  3450 fema~  2007
##  6 Adelie  Torgersen           39.3          20.6    190  3650 male   2007
##  7 Adelie  Torgersen           38.9          17.8    181  3625 fema~  2007
##  8 Adelie  Torgersen           39.2          19.6    195  4675 male   2007
##  9 Adelie  Torgersen           34.1          18.1    193  3475 <NA>   2007
## 10 Adelie  Torgersen           42.0          20.2    190  4250 <NA>   2007
## # ... with 334 more rows
```

2 Summaries and plots of qualitative variables

2.1 Tables of qualitative variables

- The main function to make tables from a data frame of observations is `tally()` which tallies (counts up) the number of observations within a given category. E.g:

```
tally(~species, data = pingviner)
```

```
## species
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

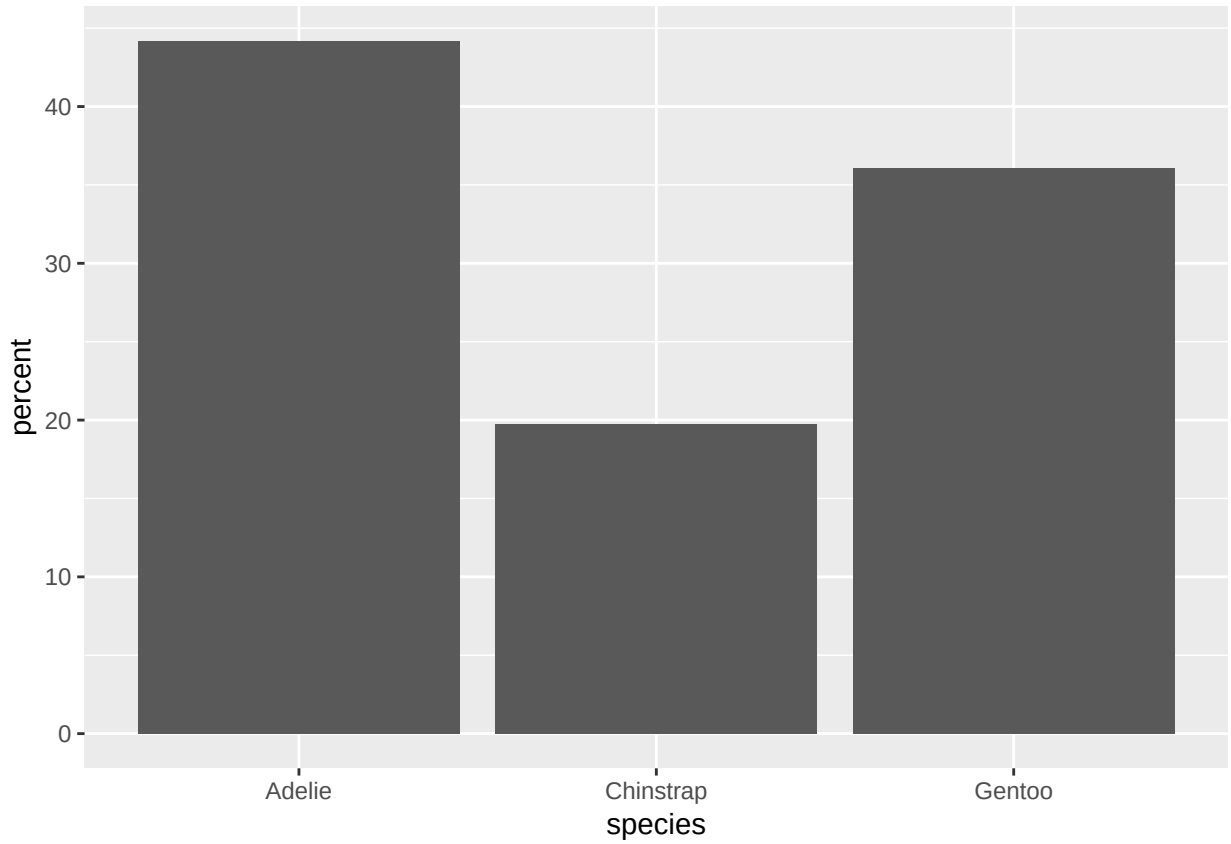
```
tally(species ~ island, data = pingviner)
```

```
##            island
## species     Biscoe Dream Torgersen
##   Adelie        44    56        52
##   Chinstrap      0    68         0
##   Gentoo       124     0         0
```

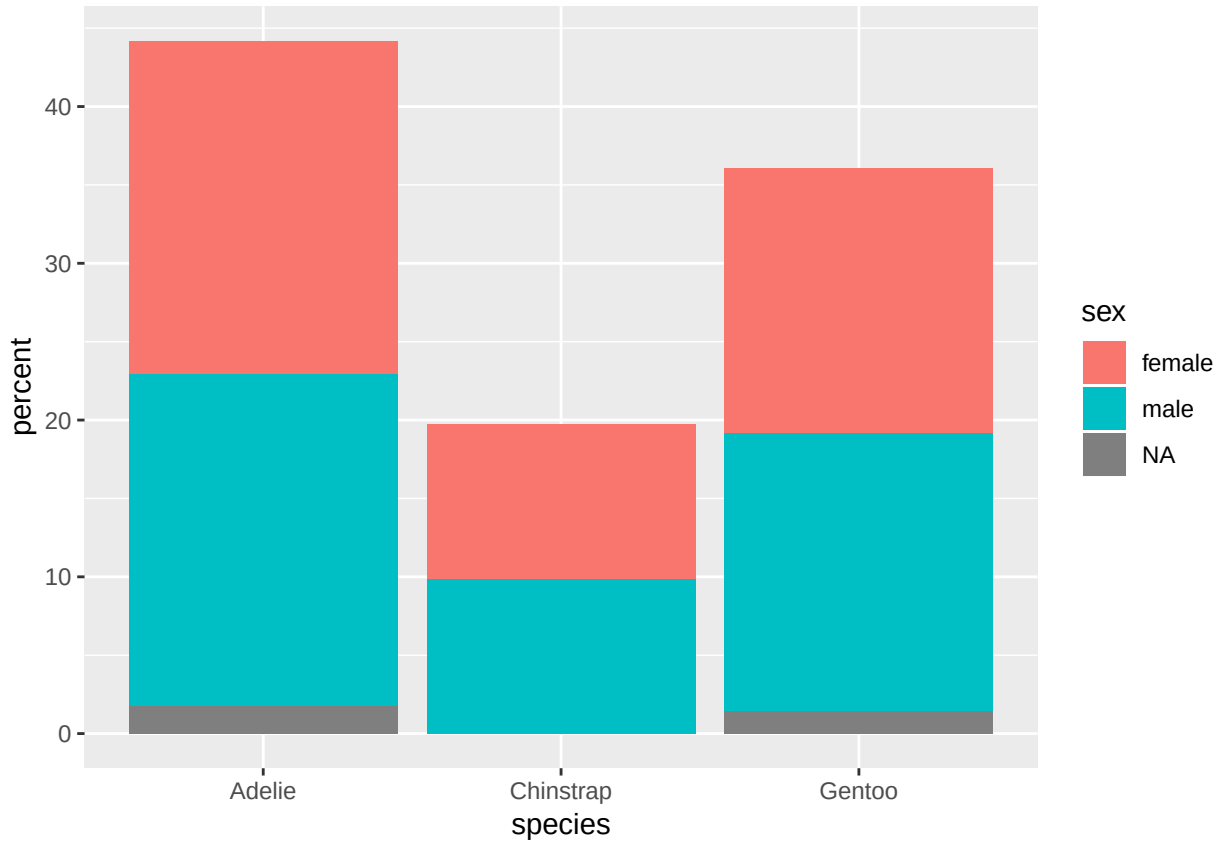
2.2 Plots of qualitative variables

- The main plotting functions for qualitative variables are `gf_percents()` and `gf_bar()`. E.g:

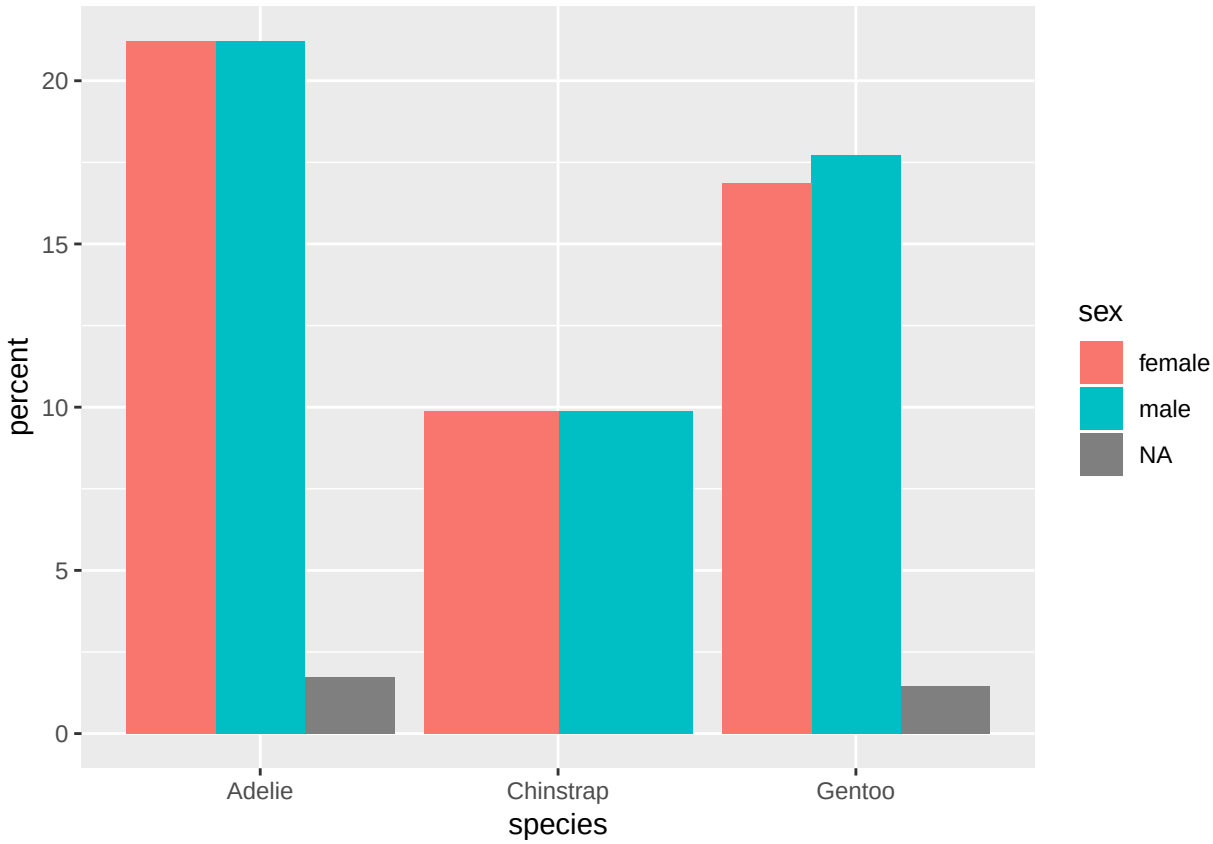
```
gf_percents(~species, data = pingviner)
```




```
gf_percents(~species, fill = ~sex, data = pingviner)
```



```
gf_percents(~species, fill = ~sex, data = pingviner, position = position_dodge())
```



3 Summaries of quantitative variables

3.1 Percentiles

- **The pth percentile** is a value such that about $p\%$ of the sample lies below or at this value and about $(100-p)\%$ of the sample lies above it.
- To calculate a percentile, first sort data in increasing order. For the `bill_length_mm` of `Gentoo` penguins it is:

  $$y_{(1)} = 40.9, y_{(2)} = 41.7, y_{(3)} = 42, \ldots, y_{(n)} = 59.6.$$

  Here the number of observations is $n = 123$ (omitting any `NA`s).
- Find the 5th percentile (i. e. $p = 5$):
  - The observation number corresponding to the 5-percentile is $N = \frac{123 \cdot 5}{100} = 6.15$.
  - So the 5-percentile lies between the observations with observation number $k = 6$ and $k + 1 = 7$. That is, its value lies somewhere in the interval between $y_6 = 42.8$ and $y_7 = 42.9$.
  - One of several methods for estimating the 5-percentile from the value of N is defined as:

    $$y_{(k)} + (N - k)(y_{(k+1)} - y_{(k)})$$

    which in this case states

    $$y_6 + (6.15 - 6)(y_7 - y_6) = 42.8 + 0.15 \cdot (42.9 - 42.8) = 42.81$$

3.2 Measures of center of data: Mean and median

- A number of numerical summaries can be retrieved using the `favstats` command:

```
favstats(bill_length_mm ~ species, data = pingviner)
```

```
##     species min Q1 median Q3 max mean  sd   n missing
## 1    Adelie  32 37     39 41  46   39 2.7 151       1
## 2 Chinstrap  41 46     50 51  58   49 3.3  68       0
## 3    Gentoo  41 45     47 50  60   48 3.1 123       1
```

- The observed values of `bill_length_mm` are $y_1 = 46.1$, $y_2 = 50, \ldots, y_n = 49.9$, where there are a total of $n = 123$ values (omitting any `NA`s). As previously defined this constitutes a **sample**.
- **mean** $= 48$ is the **average** of the sample, which is calculated by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

  We may also call $\bar{y}$ the **(empirical) mean** or the **sample mean**. It is calculated using `mean()` in **R**.
- **median** $= 47$ is the 50-percentile, i.e. the value that splits the sample in 2 groups of equal size. It is calculated using `median()` in **R**.
- An important property of the **mean** and the **median** is that they have the same unit as the observations (e.g. millimeter).

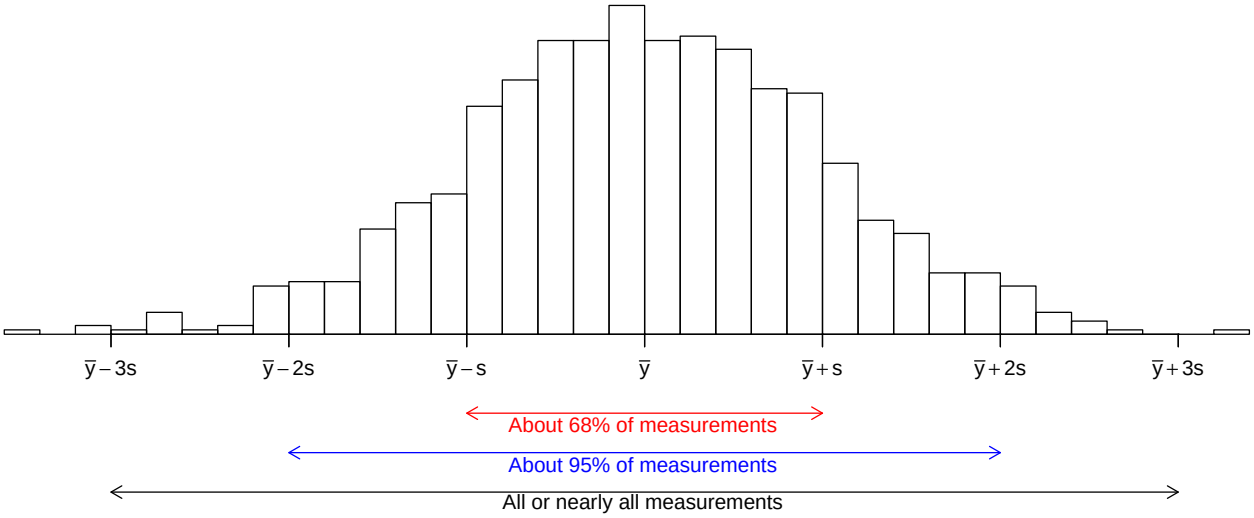
3.3 Measures of variability of data: range, standard deviation and variance

- The **range** is the difference of the largest and smallest observation (`range()` in **R**).
- The **(empirical) variance** (`var()` in **R**) is the average of the squared deviations from the mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

- **sd** $=$ **standard deviation** $= s = \sqrt{s^2}$ (`sd()` in **R**).
- Note: If the observations are measured in mm, the **variance** has unit mm^2 which is hard to interpret. The **standard deviation** on the other hand has the same unit as the observations.
- The standard deviation describes how much data varies around the (empirical) mean.

3.3.1 The empirical rule



If the histogram of the sample looks like a bell shaped curve, then

- about 68% of the observations lie between $\bar{y} - s$ and $\bar{y} + s$.
- about 95% of the observations lie between $\bar{y} - 2s$ and $\bar{y} + 2s$.
- All or almost all (99.7%) of the observations lie between $\bar{y} - 3s$ and $\bar{y} + 3s$.

4 Target population and random sampling

4.1 Population parameters

- When the sample size grows, then e.g. the mean of the sample, $\overline{y}$, will stabilize around a fixed value, μ, which is usually unknown. The value μ is called the **population mean**.
- Correspondingly, the standard deviation of the sample, s, will stabilize around a fixed value, σ, which is usually unknown. The value σ is called the **population standard deviation**.
- Notation:
  - μ (mu) denotes the population mean.
  - σ (sigma) denotes the population standard deviation.

Population	Sample
μ	$\overline{y}$
σ	s

4.1.1 A word about terminology

- **Standard deviation**: a measure of variability of a population or a sample.
- **Standard error**: a measure of variability of an estimate. For example, a measure of variability of the sample mean.

4.2 Aim of statistics

- Statistics is all about "saying something" about a population.
- Typically, this is done by taking a random sample from the population.
- The sample is then analysed and a statement about the population can be made.
- The process of making conclusions about a population from analysing a sample is called **statistical inference**.

4.3 Random sampling schemes

Possible strategies for obtaining a random sample from the target population are explained in Agresti section 2.4:

- **Simple sampling: each possible sample of equal size equally probable**
- Systematic sampling
- Stratified sampling
- Cluster sampling
- Multistage sampling
- ...

5 Biases

5.1 Types of biases

Agresti section 2.3:

- Sampling/selection bias
  - Probability sampling: each sample of size n has same probability of being sampled
    - Still problems: undercoverage, groups not represented (inmates, homeless, hospitalized, ...)
  - Non-probability sampling: probability of sample not possible to determine
    - E.g. volunteer sampling
- Response bias
  - E.g. poorly worded, confusing or even order of questions
  - Lying if think socially unacceptable
- Non-response bias
  - Non-response rate high; systematic in non-responses (age, health, believes)

5.2 Example of sample bias: United States presidential election, 1936

(Based on Agresti, this and this.)

- Current president: Franklin D. Roosevelt
- Election: Franklin D. Roosevelt vs Alfred Landon (Republican governor of Kansas)
- Literary Digest: magazine with history of accurately predicting winner of past 5 presidential elections

5.2.1 Results

- Literary Digest poll: Landon: 57%; Roosevelt: 43%
- Actual results: Landon: 38%; Roosevelt: 62%
- Sampling error: 57%-38% = 19%
  - Practically all of the sampling error was the result of **sample bias**
  - Poll size of $>$ 2 mio. individuals participated – extremely large poll

---

5.2.2 Problems (biases)

- Mailing list of about 10 mio. names was created
  - Based on every telephone directory, lists of magasine subscribers, rosters of clubs and associations, and other sources
  - Each one of 10 mio. received a mock ballot and asked to return the marked ballot to the magazine
- "respondents who returned their questionnaires represented only that subset of the population with a relatively intense interest in the subject at hand, and as such constitute in no sense a random sample ... it seems clear that the minority of anti-Roosevelt voters felt more strongly about the election than did the pro-Roosevelt majority" (*The American Statistician*, 1976)
- Biases:
  - Sample bias
    - List generated towards middle- and upper-class voters (e.g. 1936 and telephones)
    - Many unemployed (club memberships and magazine subscribers)
  - Non-response bias
    - Only responses from 2.3/2.4 mio out of 10 million people

5.3 Example of response bias: Wording matters

New York Times/CBS News poll on attitude to increased fuel taxes

- "Are you in favour of a new gasoline tax?" - 12% said yes.
- "Are you in favour of a new gasoline tax to decrease US dependency on foreign oil?" - 55% said yes.
- "Do you think a new gas tax would help to reduce global warming?" - 59% said yes.

5.4 Example of response bias: Order of questions matter

US study during cold war asked two questions:

1 "Do you think that US should let Russian newspaper reporters come here and sent back whatever they want?"

2 "Do you think that Russia should let American newspaper reporters come in and sent back whatever they want?"

The percentage of yes to question 1 was 36%, if it was asked first and 73%, when it was asked last.

5.5 Example of survivior bias: Bullet holes of honor

(Based on this.)

- World War II
- Royal Air Force (RAF), UK
  - Lost many planes to German anti-aircraft fire
- Armor up!
  - Where?
  - Count up all the bullet holes in planes that returned from missions
    - Put extra armor in the areas that attracted the most fire

---

- Hungarian-born mathematician Abraham Wald:
  - If a plane makes it back safely with a bunch of bullet holes in its wings: holes in the wings aren't very dangerous
    - **Survivorship bias**
  - Armor up the areas that (on average) don't have any bullet holes
    - They never make it back, apparently dangerous

Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.80

(See also this xkcd)