Numerical Summaries

Summarizing data in R

- Have seen `summary` (5-number summary of each column). But what if we want:
  - a summary or two of just one column
  - a count of observations in each category of a categorical variable
  - summaries by group
  - a different summary of all columns (eg. SD)
- To do this, meet pipe operator %>%. This takes input data frame, does something to it, and outputs result. (Learn: `Ctrl-Shift-M`.)
- Output from a pipe can be used as input to something else, so can have a sequence of pipes.
- Summaries include: mean, median, min, max, sd, IQR, quantile (for obtaining quartiles or any percentile), n (for counting observations).
- Use our Australian athletes data again.

Packages for this section

```
library(tidyverse)
```

Summarizing one column

- Mean height:

```
athletes %>% summarize(m=mean(Ht))
```

m
180.104

or to get mean and SD of BMI:

```
athletes %>% summarize(m=mean(BMI),s=sd(BMI))
```

m	s
22.95589	2.863933

Quartiles

- `quantile` calculates percentiles ("fractiles"), so we want the 25th and 75th percentiles:

```
athletes %>% summarize( Q1=quantile(Wt, 0.25),
                        Q3=quantile(Wt, 0.75))
```

Q1	Q3
66.525	84.125

Creating new columns

- These weights are in kilograms. Maybe we want to summarize the weights in pounds.
- Convert kg to lb by multiplying by 2.2.
- Create new column and summarize that:

```
athletes %>% mutate(wt_lb=Wt*2.2) %>%
  summarize(Q1_lb=quantile(wt_lb, 0.25),
            Q3_lb=quantile(wt_lb, 0.75))
```

Q1_lb	Q3_lb
146.355	185.075

Counting how many

for example, number of athletes in each sport:

```
athletes %>% count(Sport)
```

Sport	n
BBall	25
Field	19
Gym	4
Netball	23
Row	37
Swim	22
T400m	29
Tennis	11
TSprnt	15
WPolo	17

Counting how many, variation 2:

Another way (which will make sense in a moment):

```
athletes %>% group_by(Sport) %>%
  summarize(count=n())
```

Sport	count
BBall	25
Field	19
Gym	4
Netball	23
Row	37
Swim	22
T400m	29
Tennis	11
TSprnt	15
WPolo	17

Summaries by group

- Might want separate summaries for each "group", eg. mean and SD of height for males and females. Strategy is `group_by` (to define the groups) and then `summarize`:

```
athletes %>% group_by(Sex) %>%
  summarize(mean_Ht = mean(Ht), sd_Ht = sd(Ht))
```

Sex	mean_Ht	sd_Ht
female	174.5940	8.242203
male	185.5059	7.903487

Count plus stats

- If you want number of observations per group plus some stats, you need to go the `n()` way:

```
athletes %>% group_by(Sex) %>%
summarize(n = n(), mean_Ht = mean(Ht), sd_Ht = sd(Ht))
```

Sex	n	mean_Ht	sd_Ht
female	100	174.5940	8.242203
male	102	185.5059	7.903487

- This explains second variation on counting within group: "within each sport/Sex, how many athletes were there?"

Summarizing several columns

- Standard deviation of each (numeric) column:

```
athletes %>% summarize(across(where(is.numeric), \(x) sd(x))) -> d
```

```
d
```

RCC	WCC	Hc	Hg	Ferr	BMI	SSF	%Bfat	LBM	Ht	Wt
0.4579764	1.800549	3.662989	1.362451	47.50124	2.863933	32.56533	6.189826	13.0702	9.734495	13.92557

- Median and IQR of all columns whose name starts with H:

```
athletes %>% summarize(across(starts_with("H"),
                              list(med = \(x) median(x),
                                   iqr = \(x) IQR(x))))
```

Hc_med	Hc_iqr	Hg_med	Hg_iqr	Ht_med	Ht_iqr
43.5	4.975	14.7	2.075	179.7	12.175

by group

```
athletes %>%
  group_by(Sex) %>%
  summarize(across(starts_with("H"),
                   list(med = \(x) median(x),
                        iqr = \(x) IQR(x))))
```

Sex	Hc_med	Hc_iqr	Hg_med	Hg_iqr	Ht_med	Ht_iqr
female	40.6	4.025	13.5	1.600	175.00	8.675
male	45.5	2.575	15.5	0.975	185.55	11.250