Introduction to Probability and Statistics

26 June 2020

recap

Last lecture, we discussed:

- Types of studies
- Types of sampling
- Principles of experimental design

recap

Suppose we want to estimate household size, where a “household” is defined as people living together in the same dwelling, and sharing living accommodations.

If we select students at random at an elementary school and ask them what their family size is, **will this be a good measure of household size**? Or will our average be biased? If so, will it overestimate or underestimate the true value?

(*OpenIntro Statistics*, exercise 1.26)

Summarizing data

```r
# load food consumption dataset
food_consumption <-
  readr::read_csv(
    paste0('https://raw.githubusercontent.com/',
           'rfordatascience/tidytuesday/master/',
           'data/2020/2020-02-18/food_consumption.csv')
    )
```

Food Consumption data

Data from the the United Nations on the **annual per-capita** consumption of eleven categories of food for 130 countries.

```
unique(food_consumption$country)
```

```
##   [1] "Argentina"                "Australia"
##   [3] "Albania"                  "Iceland"
##   [5] "New Zealand"              "USA"
##   [7] "Uruguay"                  "Luxembourg"
##   [9] "Brazil"                   "Kazakhstan"
##  [11] "Sweden"                   "Bermuda"
##  [13] "Denmark"                  "Finland"
##  [15] "Ireland"                  "Greece"
##  [17] "France"                   "Canada"
##  [19] "Norway"                   "Hong Kong SAR. China"
##  [21] "French Polynesia"         "Israel"
##  [23] "Switzerland"              "Netherlands"
##  [25] "Kuwait"                   "United Kingdom"
##  [27] "Austria"                  "Oman"
##  [29] "Italy"                    "Bahamas"
##  [31] "Portugal"                 "Malta"
##  [33] "Armenia"                  "Slovenia"
##  [35] "Chile"                    "Venezuela"
##  [37] "Belgium"                  "Germany"
```

Food Consumption data

```
food_consumption %>%
  filter(country == "USA")
```

```
## # A tibble: 11 x 4
##    country food_category            consumption co2_emmission
##    <chr>   <chr>                          <dbl>         <dbl>
##  1 USA     Pork                           27.6          97.8
##  2 USA     Poultry                        50.0          53.7
##  3 USA     Beef                           36.2        1118.
##  4 USA     Lamb & Goat                     0.43         15.1
##  5 USA     Fish                           12.4          19.7
##  6 USA     Eggs                           14.6          13.4
##  7 USA     Milk - inc. cheese            255.          363.
##  8 USA     Wheat and Wheat Products       80.4          15.3
##  9 USA     Rice                            6.88          8.8
## 10 USA     Soybeans                        0.04          0.02
## 11 USA     Nuts inc. Peanut Butter         7.86         13.9
```

Food Consumption data

```r
unique(food_consumption$food_category)
```

```
##  [1] "Pork"                     "Poultry"
##  [3] "Beef"                     "Lamb & Goat"
##  [5] "Fish"                     "Eggs"
##  [7] "Milk - inc. cheese"       "Wheat and Wheat Products"
##  [9] "Rice"                     "Soybeans"
## [11] "Nuts inc. Peanut Butter"
```

Food Consumption data

```r
food_consumption %>%
  filter(food_category == "Rice")
```

```
## # A tibble: 130 x 4
##    country     food_category consumption co2_emmission
##    <chr>       <chr>               <dbl>         <dbl>
##  1 Argentina   Rice                 8.77         11.2
##  2 Australia   Rice                11.0          14.1
##  3 Albania     Rice                 7.78          9.96
##  4 Iceland     Rice                 3.89          4.98
##  5 New Zealand Rice                 9.16         11.7
##  6 USA         Rice                 6.88          8.8
##  7 Uruguay     Rice                11.5          14.7
##  8 Luxembourg  Rice                 4.2           5.37
##  9 Brazil      Rice                32.1          41.1
## 10 Kazakhstan  Rice                 7.32          9.37
## # ... with 120 more rows
```

Descriptive statistics...

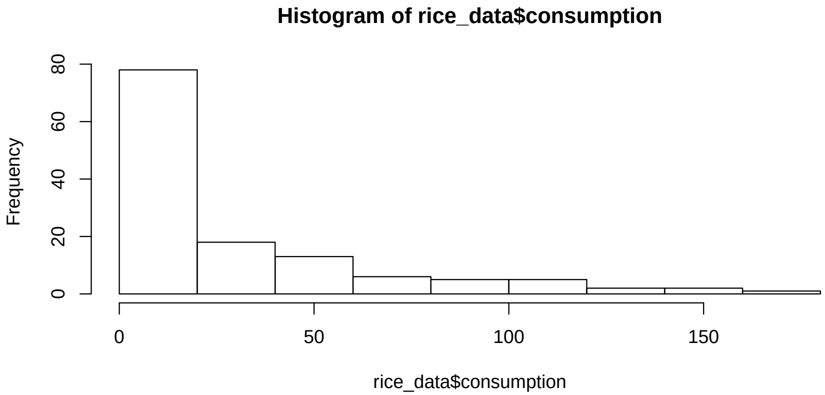
or: how do we make sense of a long list of numbers?

Visualizing distributions

Histogram

Histograms illustrate the **“distribution of data”** for **one** variable at a time, showing which values are relatively more common in our dataset.

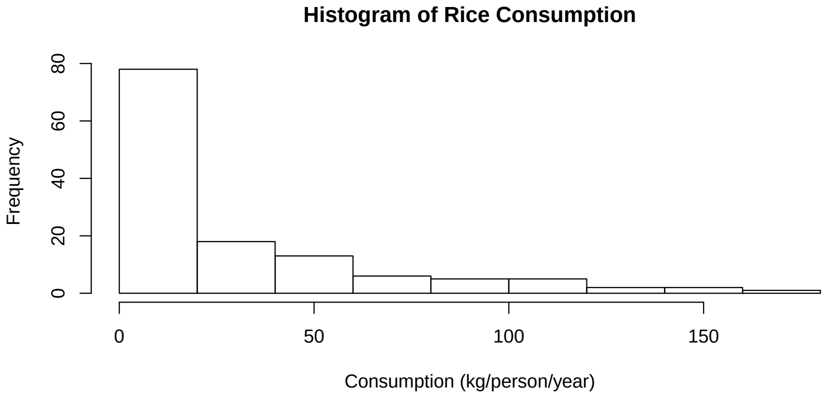
```
rice_data <- food_consumption %>%
  filter(food_category == "Rice")
hist(rice_data$consumption)
```



Histogram

Which range of values is most common in our dataset?

```
hist(rice_data$consumption,
     xlab = "Consumption (kg/person/year)",
     main = "Histogram of Rice Consumption")
```



Histogram

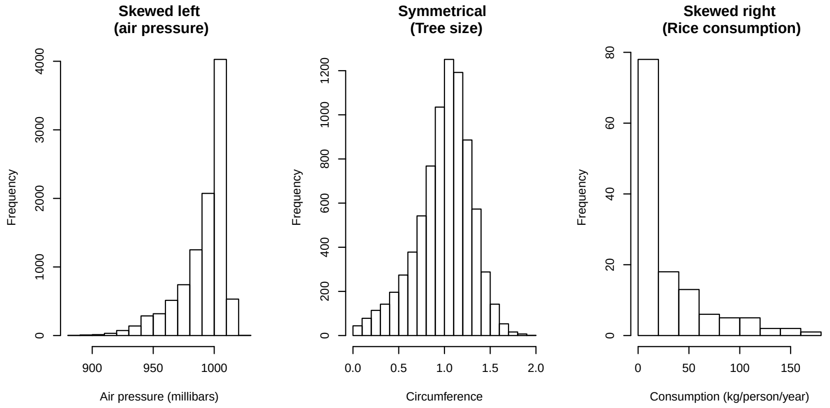
Can we use this plot to make any generalizations about rice consumption habits in the world?

Histogram

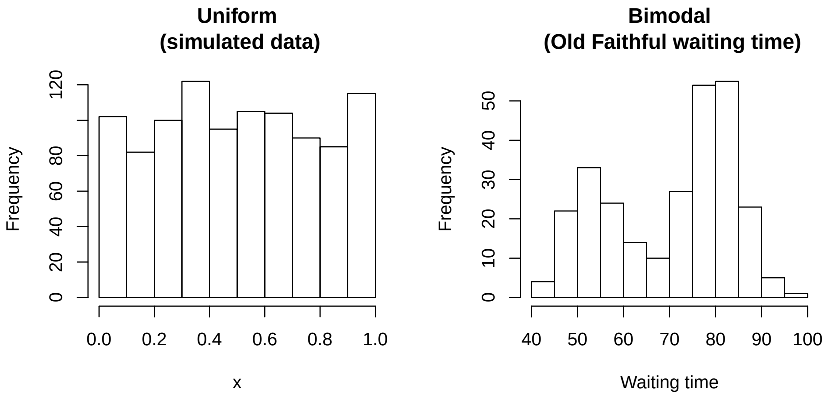
Can we use this plot to make any generalizations about rice consumption habits in the world?

Only if we make statements that are connected to the types of countries that are well represented in the dataset.

Shapes of Histograms



Other Shapes of Histograms



Scatter plot

Scatter plots illustrate the **joint distribution** of **two** variables at a time, showing which values are relatively more common in our dataset.

```r
wheat_data <- food_consumption %>%
  filter(food_category == "Wheat and Wheat Products")
wheat_data
```

```
## # A tibble: 130 x 4
##    country     food_category            consumption co2_emmission
##    <chr>       <chr>                          <dbl>         <dbl>
##  1 Argentina   Wheat and Wheat Products       103.           19.7
##  2 Australia   Wheat and Wheat Products        70.5          13.4
##  3 Albania     Wheat and Wheat Products       139.           26.4
##  4 Iceland     Wheat and Wheat Products        72.9          13.9
##  5 New Zealand Wheat and Wheat Products        76.9          14.7
##  6 USA         Wheat and Wheat Products        80.4          15.3
##  7 Uruguay     Wheat and Wheat Products       109.           20.8
##  8 Luxembourg  Wheat and Wheat Products       103.           19.7
##  9 Brazil      Wheat and Wheat Products        53            10.1
## 10 Kazakhstan  Wheat and Wheat Products        92.3          17.6
## # ... with 120 more rows
```

Joining data tables

In order to produce a scatter plot, we need to match countries' wheat and rice consumption.

```r
grains_data <- wheat_data %>%
  bind_rows(rice_data)
grains_data
```

```
## # A tibble: 260 x 4
##    country     food_category               consumption co2_emmission
##    <chr>       <chr>                             <dbl>         <dbl>
##  1 Argentina   Wheat and Wheat Products          103.           19.7
##  2 Australia   Wheat and Wheat Products           70.5          13.4
##  3 Albania     Wheat and Wheat Products          139.           26.4
##  4 Iceland     Wheat and Wheat Products           72.9          13.9
##  5 New Zealand Wheat and Wheat Products           76.9          14.7
##  6 USA         Wheat and Wheat Products           80.4          15.3
##  7 Uruguay     Wheat and Wheat Products          109.           20.8
##  8 Luxembourg  Wheat and Wheat Products          103.           19.7
##  9 Brazil      Wheat and Wheat Products           53            10.1
## 10 Kazakhstan  Wheat and Wheat Products           92.3          17.6
## # ... with 250 more rows
```

Joining data tables

What we really want is each row to represent a country, and each column to represent a type of food. The `pivot_wider` function comes in handy here:

```
grains_data <- grains_data %>%
  select(-co2_emmission) %>% # remove c02 variable
  pivot_wider(names_from = food_category, values_from = consumption) %>%
  rename(Wheat = `Wheat and Wheat Products`) # make column name shorter
```

Joining data tables

What we really want is each row to represent a country, and each column to represent a type of food.

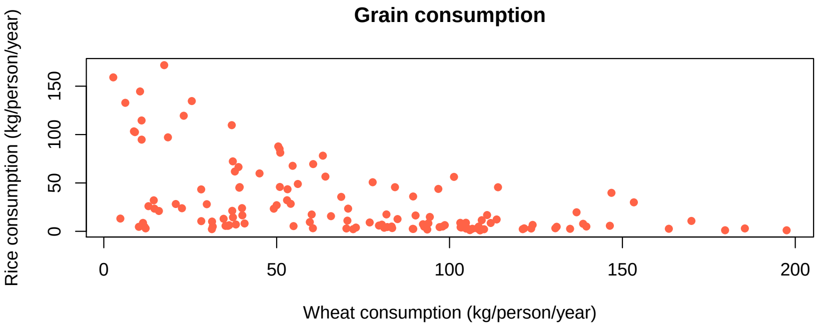
```
head(grains_data)
```

```
## # A tibble: 6 x 3
##   country     Wheat  Rice
##   <chr>       <dbl> <dbl>
## 1 Argentina   103.   8.77
## 2 Australia    70.5 11.0
## 3 Albania     139.   7.78
## 4 Iceland      72.9  3.89
## 5 New Zealand  76.9  9.16
## 6 USA          80.4  6.88
```

Scatter plot

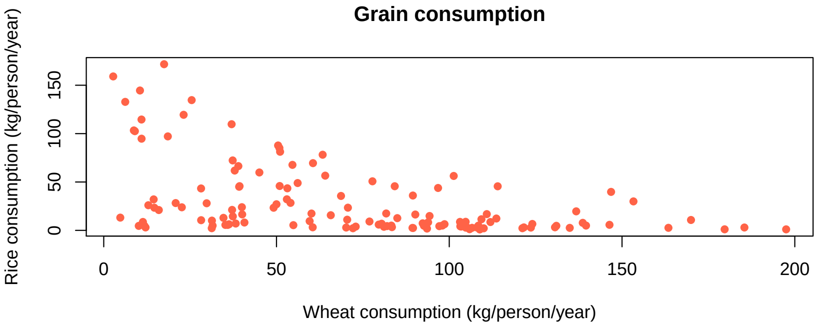
Finally, we can make our scatter plot. What do we observe?

```
plot(grains_data$Wheat, grains_data$Rice,
     xlab = "Wheat consumption (kg/person/year)",
     ylab = "Rice consumption (kg/person/year)",
     main = "Grain consumption",
     pch = 16, col = 'tomato')
```



Scatter plot

For countries in our dataset, there is a **negative association** betweeen rice consumption and wheat consumption. Countries that consume relatively more rice tend consume relatively less wheat.



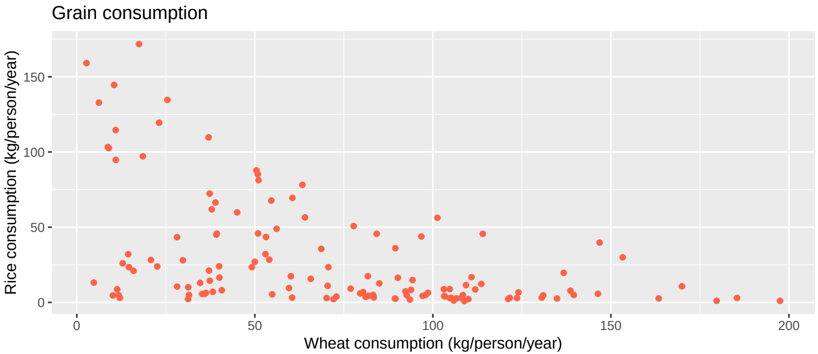
Revisiting the histogram

```r
hist(grains_data$Wheat + grains_data$Rice,
     xlab = "Consumption (kg/person/year)",
     main = "Wheat and Rice consumption")
```

Scatter plot in ggplot2

In case you wanted code to do this in ggplot2.

```
ggplot(grains_data) +
  geom_point(aes(x = Wheat, y = Rice), color = 'tomato') +
  xlab("Wheat consumption (kg/person/year)") +
  ylab("Rice consumption (kg/person/year)") +
  ggtitle("Grain consumption")
```



Summary statistics

Summary statistics

Sometimes we don't want the full **distribution**: we may just want a few numbers that summarize our data.

The full distribution will give us more information, but **summary statistics** take up less space and can speed up decision making.

ex. GPA

Summary statistics

Sometimes we don't want the full **distribution**: we may just want a few numbers that summarize our data.

The full distribution will give us more information, but **summary statistics** take up less space and can speed up decision making.

ex. GPA

Note that these are all computed based on **samples** and that we will focus on statistics for single variables for now.

“Location”

Location statistics

Some statistics describe the **location** of our data.

Crudely speaking, these statistics tell us what possible values of our variable are the most representative/common/important.

Notation

Let $x_1, \ldots, x_n$ denote the n observations of our variable of interest.

Sample mean $\overline{x}$

In words: the sum of a collection of numbers divided by the count of numbers

In math:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In R:

```
mean(grains_data$Rice)
```

```
## [1] 29.37515
```

Sample median $\tilde{x}$

In words: the "middle" value of a set of numbers

In math:

$$\text{even number of numbers} : \text{median}(\{1, 2, 3, \mathbf{3}, \mathbf{5}, 8, 8, 9\}) = \frac{3+5}{2} = 4$$

$$\text{odd number of numbers} : \text{median}(\{3, 4, 4, \mathbf{5}, 6, 8, 9\}) = 5$$

In R:

```
median(grains_data$Rice)
```

```
## [1] 11.875
```

Sample mode

In words: the most common value of a variable

In math:

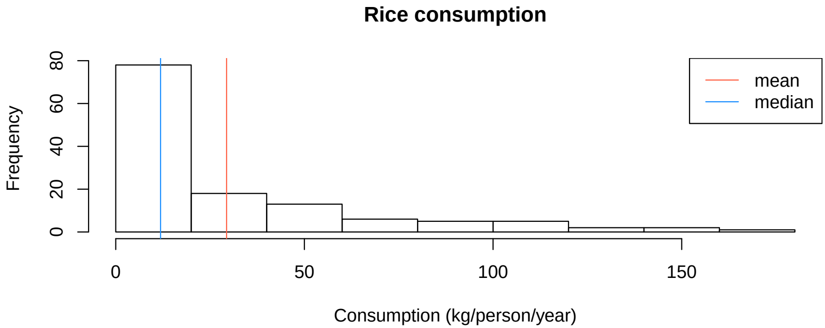
$$\text{mode}(\{1, 2, \mathbf{3}, \mathbf{3}, 5\}) = 3$$

$$\text{mode}(\{1, 2, \mathbf{3}, \mathbf{3}, 5, \mathbf{8}, \mathbf{8}, 9\}) = 3 \text{ and } 8$$

In R: R has no built in mode function!

Revisiting the histogram

```
hist(grains_data$Rice,
     xlab = "Consumption (kg/person/year)",
     main = "Rice consumption")
abline(v = mean(grains_data$Rice), col = 'tomato')
abline(v = median(grains_data$Rice), col = 'dodgerblue')
legend("topright", c("mean", "median"),
       col = c("tomato", "dodgerblue"), lty = 1)
```



Sensitivity to skew/outliers

Imagine that we're all in one classroom (scary!).

A brave student says: *I wonder how many Instagram followers the typical person in this room has.*

Another over-eager student says: *Let's take a sample of 10 people and find out!*

What sample location statistic should we calculate based on our sample?

Sensitivity to skew/outliers

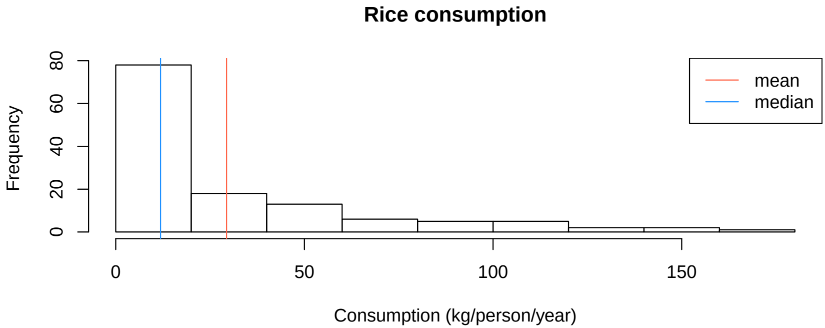
Now, suppose Christiano Ronaldo (225 million), Ariana Grande (191.1 m), The Rock (187.3 m) walk into the room.

They would be **outliers** in our population, meaning that their values would be located abnormally far from the other values.

Suppose they all end up in our sample. What happens to the sample mean? What happens to the sample median?

Comparing location statistics

```r
hist(grains_data$Rice,
     xlab = "Consumption (kg/person/year)",
     main = "Rice consumption")
abline(v = mean(grains_data$Rice), col = 'tomato')
abline(v = median(grains_data$Rice), col = 'dodgerblue')
legend("topright", c("mean", "median"),
       col = c("tomato", "dodgerblue"), lty = 1)
```



Comparing location statistics

Statistic	Mean	Median	Mode
Pros	- unique value - good for inference	- unique value - robust to outliers	- easy to explain - can compute for numerical and categorical variables
Cons	- sensitive to outliers - only numerical variables	- only numerical variables	- may not be unique/meaningful

Other location statistics

Quartiles split your data up into quarters. If you think of the median as the second quartile, then the first quartile is the median of the first half of the data and third quartile is the median of the second half of the data.

So, 25% of the data fall below the first quartile and 75% fall below the third quartile.

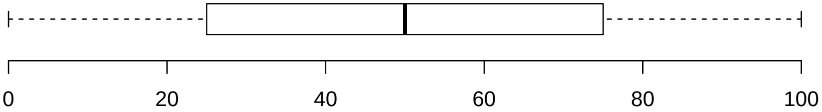
Boxplots

We can use boxplots to summarize distributions and visualize the quartiles:

The edges of the box represent the first and third quartile

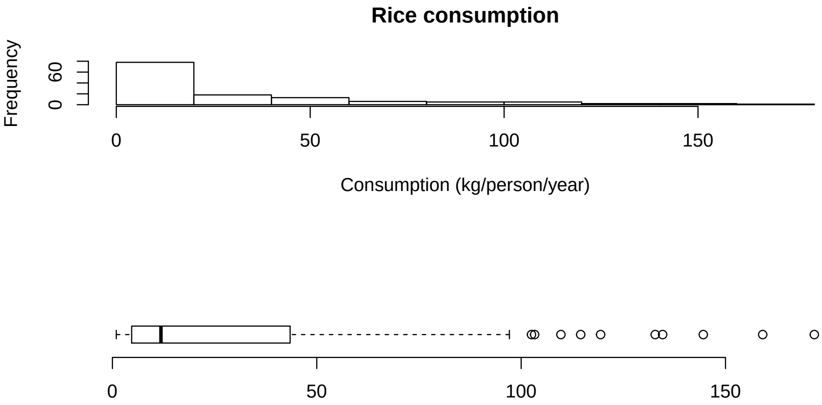
The line inside the box is the median

The edges of the "whiskers" are usually the maximum and minimum

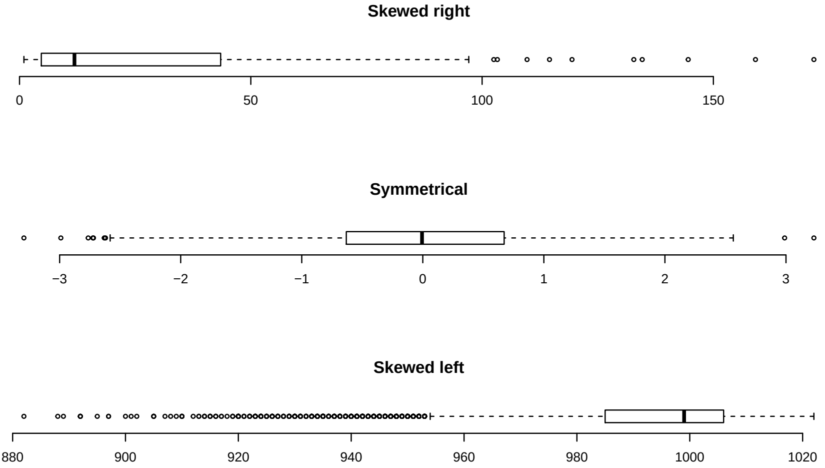


Boxplots

Returning to our rice consumption data (the points to the right are **outliers**):



Example Boxplots



“Spread”

Spread statistics

Some statistics describe the **location** of our data.

Crudely speaking, these statistics tell us how far our data values tend to be from each other.

Range

In words: the difference between the biggest data value and the smallest data value

In math:

$$\text{range}(x_1, \ldots, x_n) = \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n)$$

In R:

```
range(grains_data$Rice)
```

```
## [1]   0.95 171.73
```

Sample Variance

Let the **sample deviation** be the distance of an observation from its sample mean. So, for now, we compute the sample deviation of x_i as $x_i - \overline{x}$.

We define **sample variance** as the average squared sample deviation of the observations.

Sample Variance

In words: the "average" squared deviation of the observations

In math:

$$\mathrm{Var}(x_1, \dots, x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Note: we divide by $n - 1$ to get a better estimator

In R:

```
var(grains_data$Rice)
```

```
## [1] 1393.116
```

Sample Standard Deviation

In words: the square root of "average" squared deviation of the observations

In math:

$$\mathrm{SD}(x_1, \ldots, x_n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

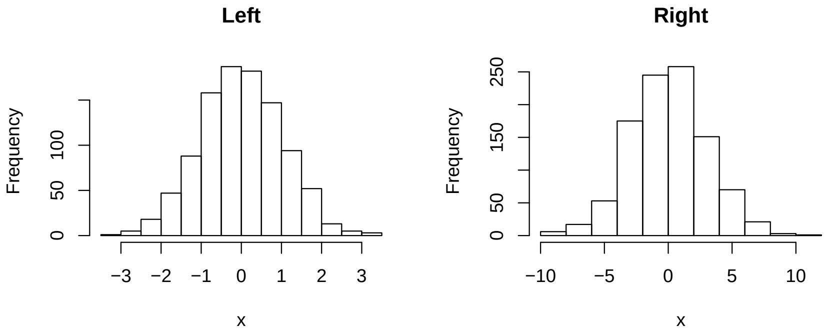
Note: we divide by $n-1$ to get a better estimator

In R:

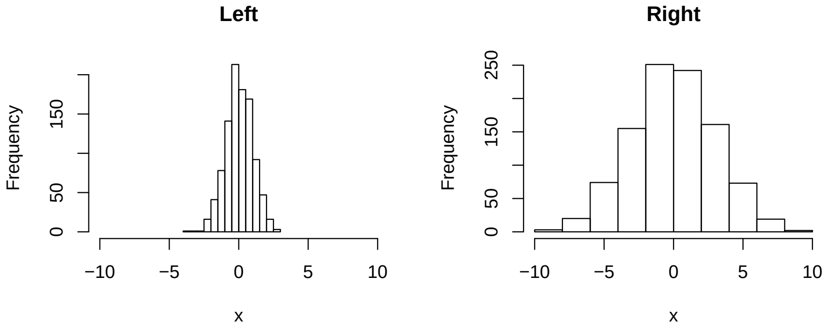
```
sd(grains_data$Rice)
```

```
## [1] 37.32447
```

Which data has the higher standard deviation?

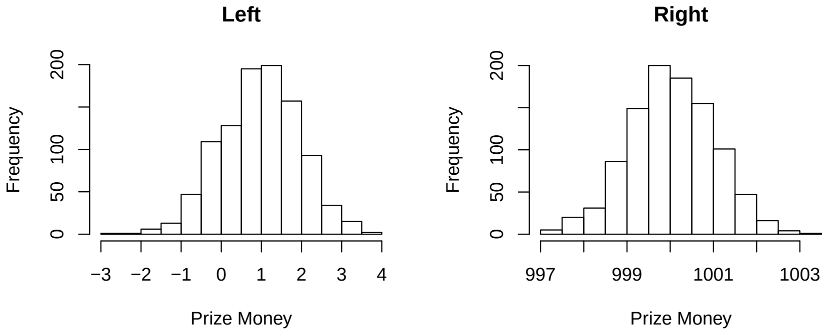


Drawn on the same scale:



Context matters

Context matters a lot when discussing standard deviation. These two (fake) datasets have the same standard deviation:



Coefficient of variation

In words: the standard deviation divided by the mean

In math:

$$\mathrm{CV}(x_1, \ldots, x_n) = \frac{\mathrm{SD}(x_1, \ldots, x_n)}{\overline{x}}$$

In R:

```
sd(grains_data$Rice) / mean(grains_data$Rice)
```

```
## [1] 1.270614
```

recap

Today, we discussed:

Visualizing distributions

Location statistics

Spread statistics

recap

If you only take one thing away from today:

All of the summaries/statistics we discussed today were calculated without making **any assumptions** about our population.

recap

If you only take one thing away from today:

All of the summaries/statistics we discussed today were calculated without making **any assumptions** about our population.

All were based entirely on our **sample**. In the next lecture we'll talk more about how we can use **sample statistics** to estimate **population parameters**.