Summarizing and Displaying Continuous Data

EDUC 641: Unit 3 Part 1

David D. Liebowitz

UNIVERSITY OF OREGON

Roadmap

Research is a partnership of questions and data		What types of data are collected?	
		Categorical data	**Continuous data**
What kinds of questions can be asked of those data?	**Descriptive questions**	• How many members of class have black hair? • What proportion of the class attends full-time?	• How tall are class members, on average • How many hours per week do class members report studying, on average?
	Relational questions	• Are male-identifying students more likely to study part-time? • Are PrevSci PhD students more likely to be female-identifying?	• Do people who say they study for more hours also think they'll finish their doctorate earlier? • Are computer-literate students less anxious about statistics?

Class goals

- Describe and summarize quantitative data that are continuous
- Describe the purpose and compute the following measures of central tendency: mean, median and mode
- Describe the purpose and compute the following measures of variability: quartiles, inter-quartile range, range, variance and standard deviation
  - Describe conceptually the principles of skewness and kurtosis
- Create visualizations of quantitative data that are continuous using R
  - Includes constructing and interpreting histograms, densities, stem-and-leaf, and box-and-whisker plots

Life expectancy data

Suppose you are working for the World Health Organization and are investigating life expectancy across different regions.

> Using this dataset, we can ask questions like:
>
> - How does life expectancy compare in high-income vs. middle- and low-income countries?
> - Do individuals living in countries with more total years of attendance in school experience, on average, higher life expectancy?

Note this question is framed in an implicitly causal way, but we are not able to answer that. This analysis is purely for educational purposes. Do NOT attempt at home!!! We now return you to your originally scheduled programming...

Before making comparisons between social/economic conditions, we want to start with describing our data, looking at all 183 nations in the dataset.

Our first task: Describe the distribution of regional life expectancy in 2015.

Materials

1. Life expectancy data (in file called life_expectancy.csv)
2. Codebook describing the contents of said data (life_expectancy_codebook.pdf)
3. R script to conduct the data analytic tasks of the unit (EDUC641_7_code.R)

Read in life expectancy

```
who <- read.csv(here("data/life_expectancy.csv")) %>%

  # going to do some data cleaning;
  # first making variable names take a common format
  janitor::clean_names() %>%

  # filtering to focus only on 2015
  filter(year == 2015) %>%

  # selecting only the variables we need
  select(country, status, life_expectancy, schooling) %>%

  # renaming one of the variables
  rename(region = country) %>%

  # rounding life expectancy to nearest year
  mutate(life_expectancy = round(life_expectancy, digits = 0))
```

You do not need to learn how to do this; just a demonstration!

Distributions

Describing a distribution

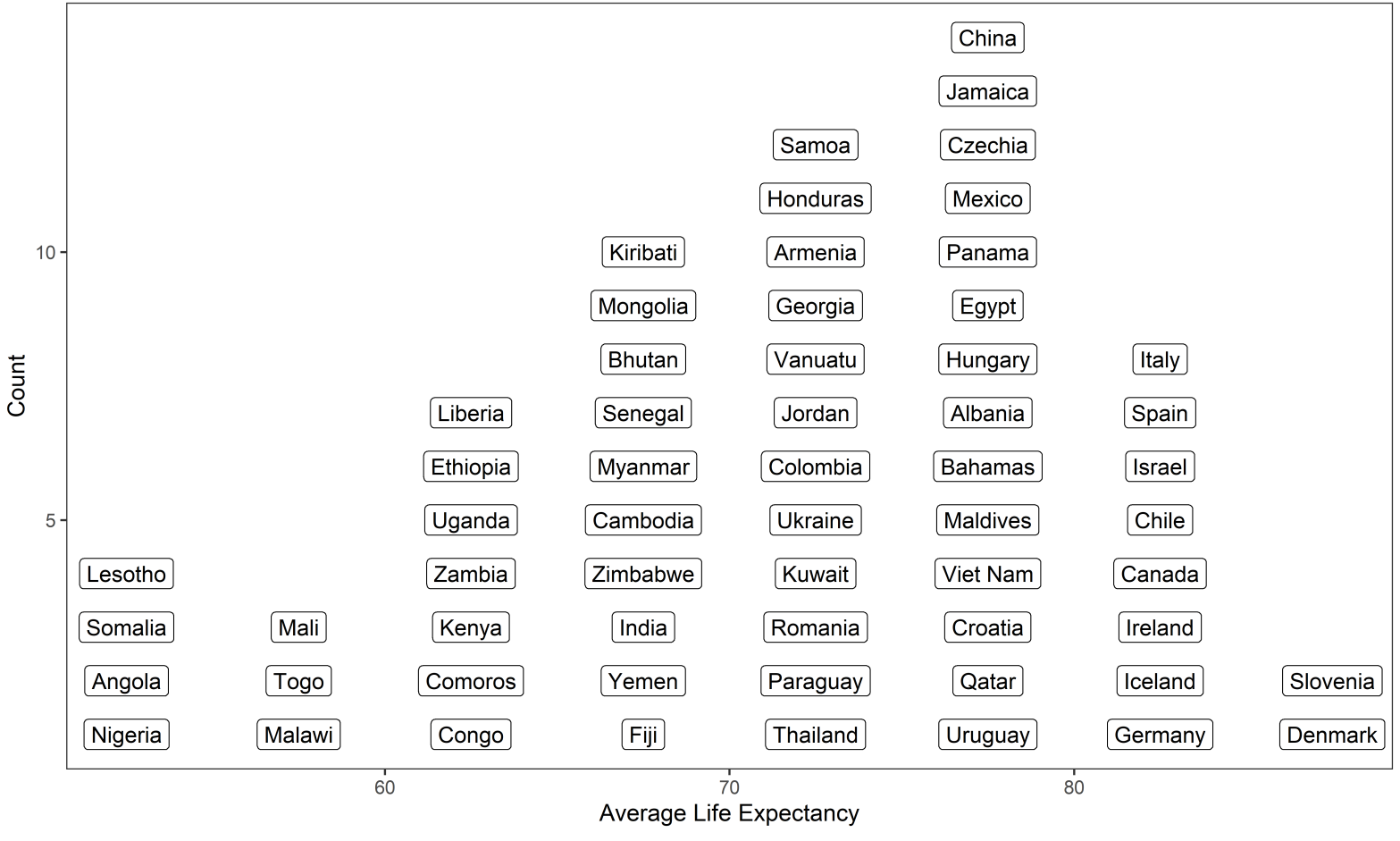
There are many ways to describe the characteristics of a distribution.

In these classes, we are reviewing:

- What is a distribution?
- Measures of Central Tendency
    - Mean
    - Median
    - Mode
- Spread or Variability
    - Variance
    - Standard Deviation
    - Interquartile Range
    - Range
- Measures of Skewness
- Measures of Kurtosis

What is a distribution?

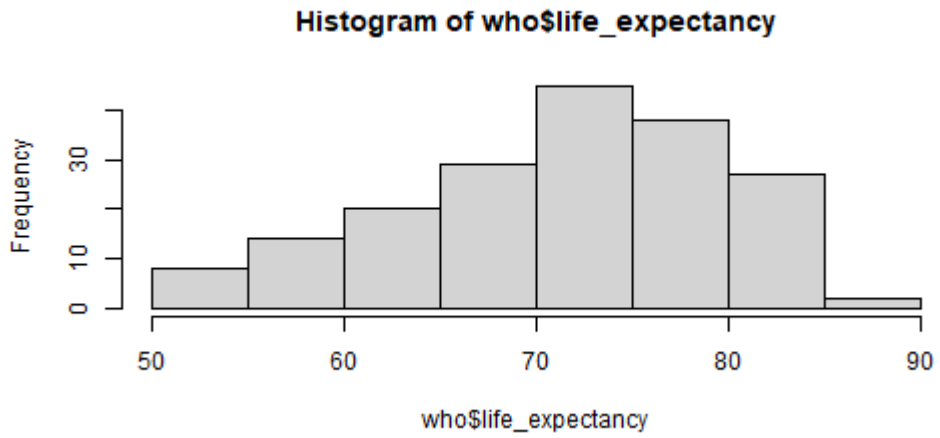
A distribution describes the number of observations of a variable that assume a particular value.



Visualizing distributions: Histogram

- Groups the data into "bins" and shows the count of observations in each bin
- R automatically creates a sensible bin size, but you can specify number of bins/bin width

```
hist(who$life_expectancy)
```



What can you already say about the shape of this distribution?

Visualizing distributions: Histogram

- Groups the data into "bins" and shows the count of observations in each bin
- R automatically creates a sensible bin size, but you can specify number of bins/bin width

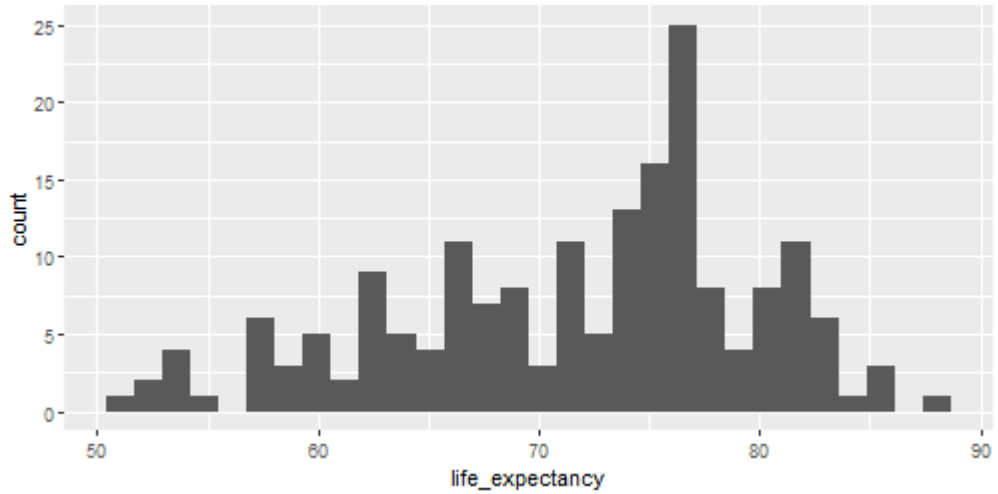
```
hist(who$life_expectancy, breaks = 16)
```

What can you already say about the shape of this distribution?

Visualizing distributions: Histogram

- Groups the data into "bins" and shows the count of observations in each bin
- R automatically creates a sensible bin size, but you can specify number of bins/bin width

```
ggplot(who, aes(life_expectancy)) + geom_histogram()
```

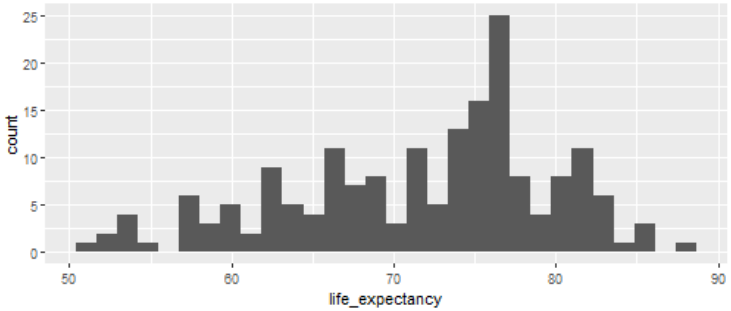


What can you already say about the shape of this distribution?

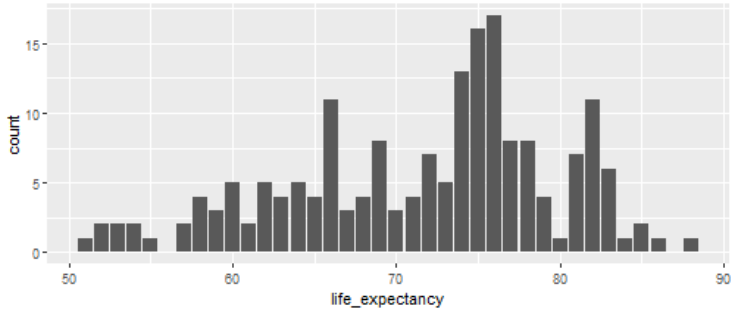
Visualizing distributions: Histogram

- Groups the data into "bins" and shows the count of observations in each bin
- R automatically creates a sensible bin size, but you can specify number of bins/bin width

Histogram



Bar plot



What is the difference between the two plots?

Visualizing distributions: Stem-and-leaf

- The "leaf" represents the last digit of a number (typically the 1s place, unless data need to be rounded to the nearest ten)
- The "stem" contains all other digits of number and serves as a grouping category for observations

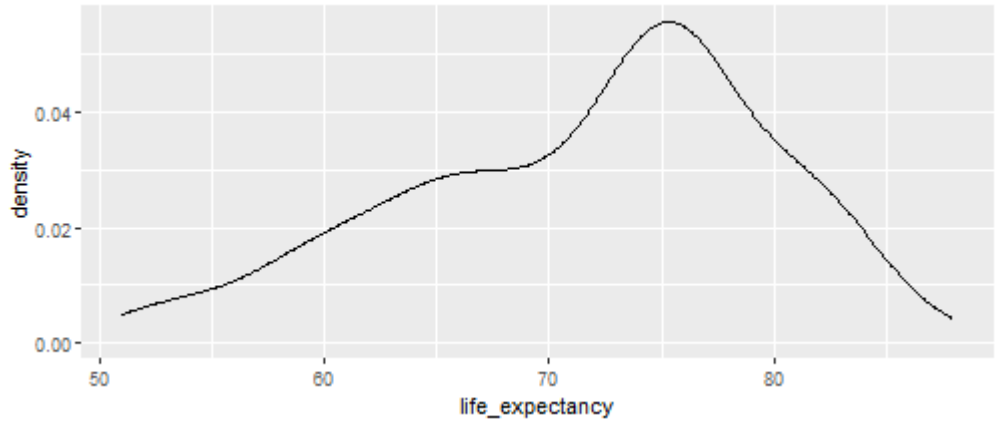
```
stem(who$life_expectancy)
```

```
##
##   The decimal point is at the |
##
##    50 | 0
##    52 | 0000
##    54 | 000
##    56 | 00
##    58 | 0000000
##    60 | 0000000
##    62 | 000000000
##    64 | 000000000
##    66 | 00000000000000
##    68 | 000000000000
##    70 | 0000000
##    72 | 000000000000
##    74 | 0000000000000000000000000000000
##    76 | 000000000000000000000000
##    78 | 000000000000
##    80 | 00000000
##    82 | 0000000000000000
##    84 | 000
##    86 | 0
##    88 | 0
```

Visualizing distributions: Density

- Can think of a density plot as a "smoothed out" histogram
- Values reflect the distribution's probability density function (not essential to understand their calculation)

```
ggplot(who, aes(life_expectancy)) +
  geom_density()
```



- Any point on the curve records the probability that you would observe that value of the variable by randomly drawing one observation from your sample

Measures of central tendency

Mean

- Represents the **average**, or the sum of all observations divided by the number of observations.
- One of the most common forms of central tendency.

> $$mean(71, 73, 76, 78, 79) = \frac{71 + 73 + 76 + 78 + 79}{5} = 75.4$$

Writing this out can be a little tedious. Since our datasets can often have hundreds, thousands or millions of observations, we often use **summation notation**.

Summation notation

Summation (or sigma) notation is used to provide a concise expression for the sum of observations.

$$\sum_{i=i}^{n} x_i$$

- i = index of summation (the unit that the observations are in; e.g., schools, students, numbers)
- n = stopping point (number of observations)
- x_i = summation element (what we are summing)

Our mean formula rewritten using summation notation:

$$\bar{X} = \frac{1}{n}\sum_{i=i}^{n} X_i$$

We use the bar on the X ($\bar{X}$) to indicate that we are calculating the mean value of the variable.

Summation notation

Let's demonstrate using our WHO data of 183 countries to find the mean.

$$\bar{X} = \frac{1}{183}\sum_{i=1}^{183} X_i = \frac{13110}{183} = 71.64$$

The mean national life expectancy for these countries in 2015 is 71.64 years.

> **Reminder**:
>
> i = index of summation (the unit that the observations are in; e.g., schools, students, numbers)
>
> n = stopping point (number of observations)
>
> x_i = summation element (what we are summing)

While this isn't critical to understand right now, the reason we index the summation i through n is that we could, in principle, want to summarize only the first 6 countries. Or the 5th through 24th countries. We would write the latter as:

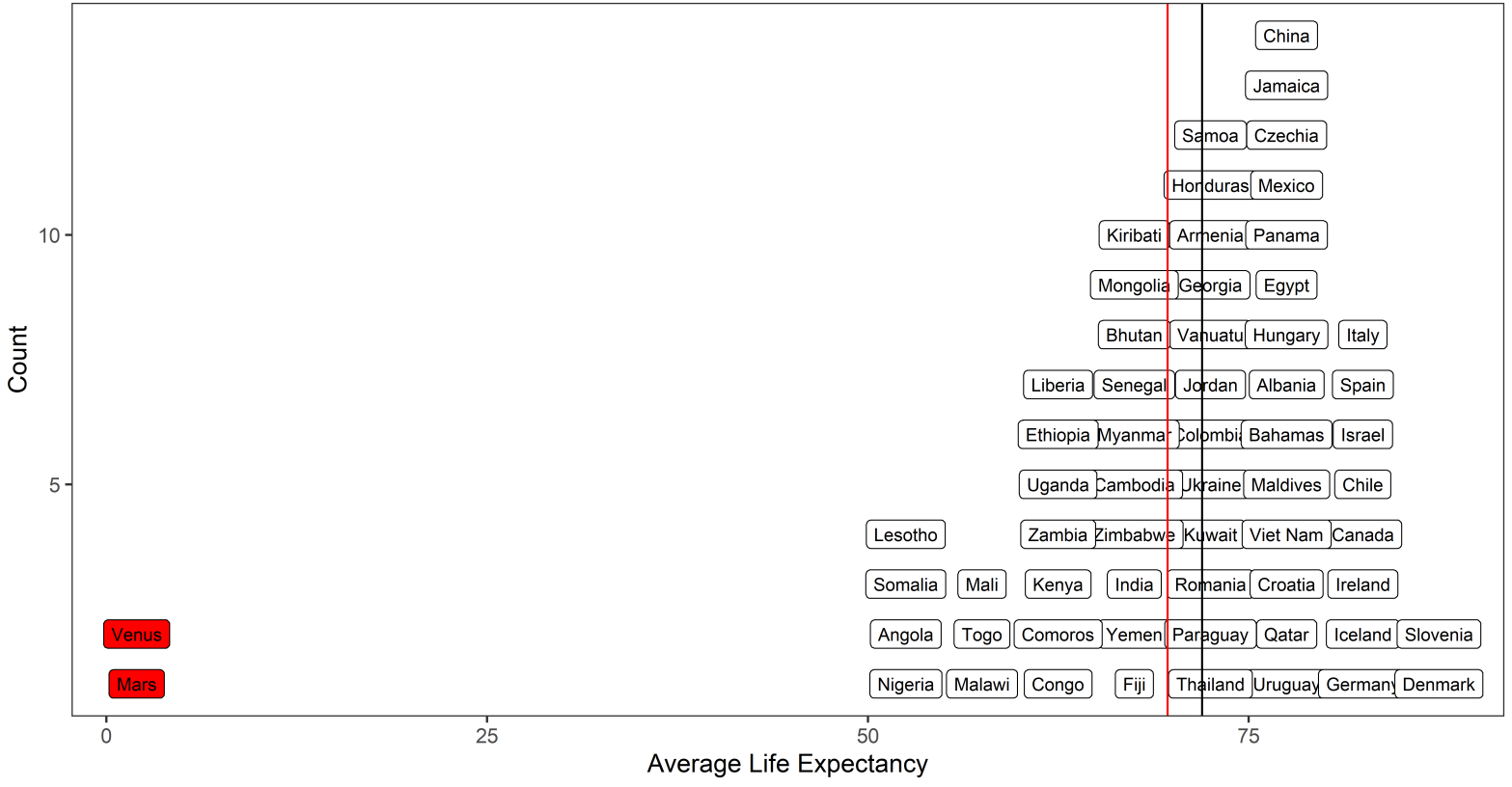
$$\bar{X} = \frac{1}{20}\sum_{i=5}^{24} X_i$$

Mean

- The mean can take a value not found in the data
- The mean represents the "balance point" or "fulcrum" of the distribution
- Deviations from the mean sum to 0
- Outliers, or observations with highly unusual values, can affect this balance point
- Thus, the mean is vulnerable to outliers
- Can only be used with interval- and ratio-scale variables

Vulnerability of the mean

- Humans on Mars and Venus have very low life expectancies (for now).
- Here, having Mars and Venus in this sample of countries substantially shifts our mean away from the "peak" of the distribution and from the mean of the distribution when excluding these two outliers.



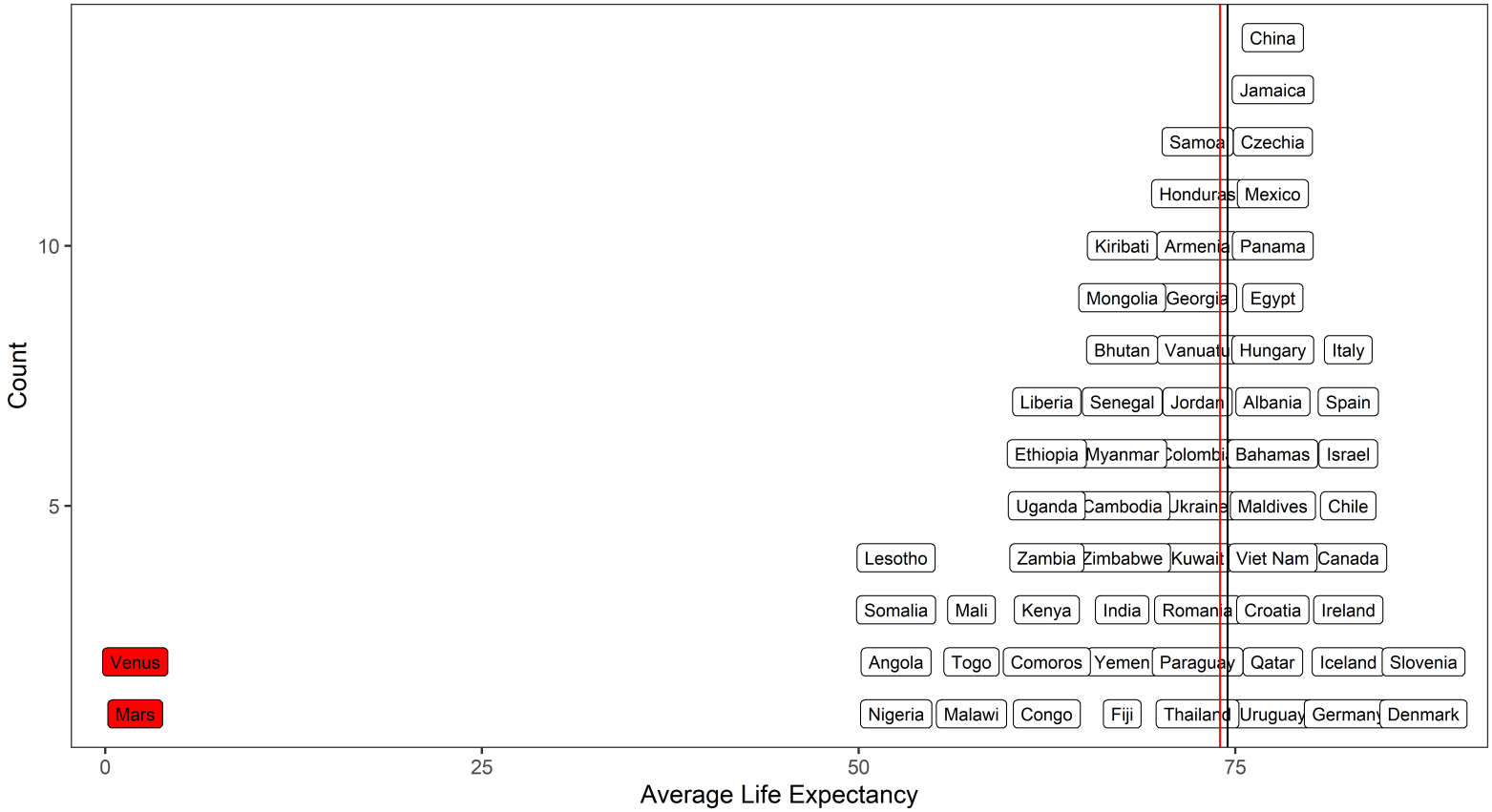
Median

- Represents the **midpoint** of the distribution
- Also called the *50th percentile* of a distribution or *2nd quartile*, meaning half of the observations are above the median and half are below
- Particularly useful for describing the central tendency of skewed distributions (more on this later)
- Can be used with ordinal-, interval-, or ratio-scale variables
- To find the median, arrange the scores in ascending order and identify the middle value. If there are two middle values (i.e., an even number of observations), use the average of the two:

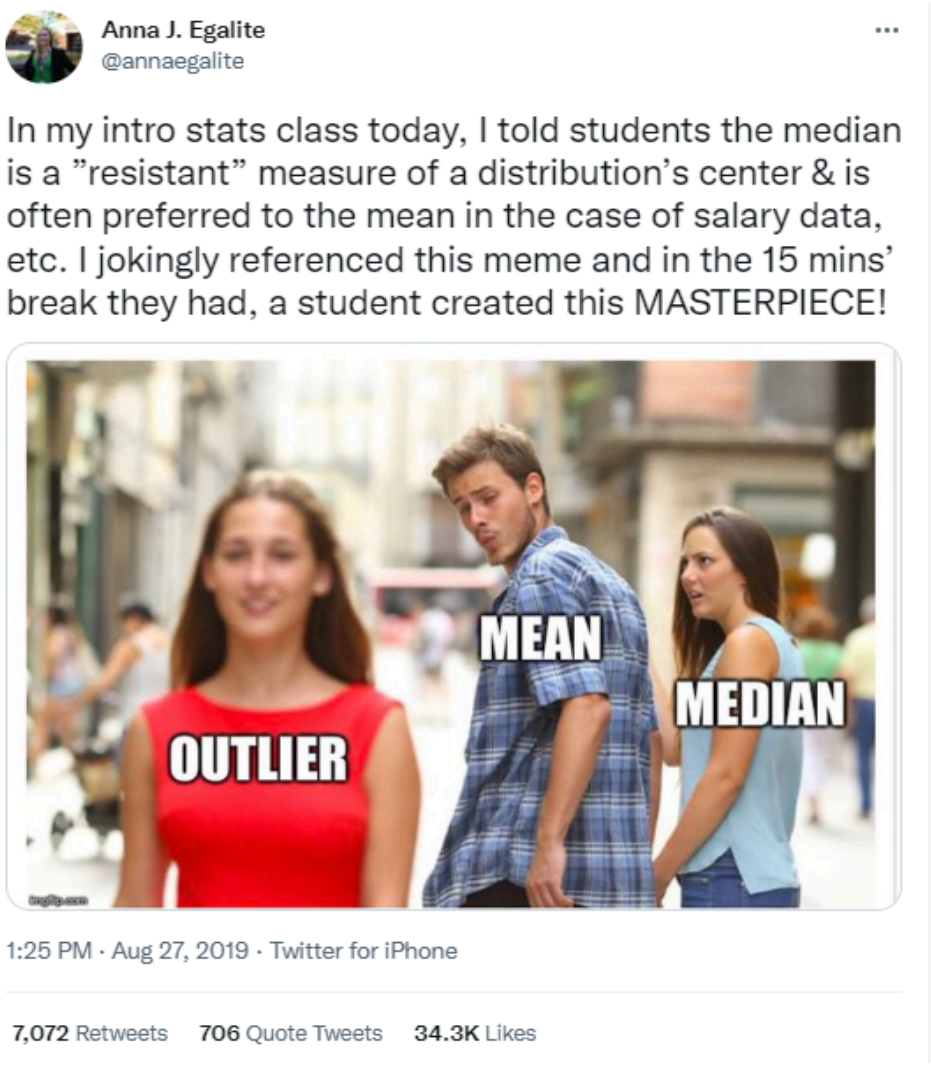
> $$median(71, 73, 76, 78, 79) = 76$$
>
> $$median(71, 73, 76, 77, 78, 79) = 76.5$$

Median

- Unlike the mean, the median is less affected by outliers.



With apologies for reifying gender stereotypes:



Anna J. Egalite
@annaegalite

In my intro stats class today, I told students the median is a "resistant" measure of a distribution's center & is often preferred to the mean in the case of salary data, etc. I jokingly referenced this meme and in the 15 mins' break they had, a student created this MASTERPIECE!

1:25 PM · Aug 27, 2019 · Twitter for iPhone

7,072 Retweets **706** Quote Tweets **34.3K** Likes

Mode

The mode simply refers to the **most frequent value** in the data.

What is the mode of the following data?

> 78, 79, 78, 77, 80, 79, 76, 74, 79, 78, 77, 78, 80, 79, 78, 76, 79, 77, 78, 78

It's easier if we sort the data first.

> 74, 76, 76, 77, 77, 77, **78, 78, 78, 78, 78, 78, 78,** 79, 79, 79, 79, 79, 80, 80

78 is the most frequent value in the data. Therefore, the mode is 78.

Mode

Sometimes there may be more than one mode. Consider the following example:

> 4, 6, 6, 7, 7, 7, **8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9,** 10, 10

Both 8 and 9 are observed seven times. In this case, the data can be considered **bimodal** and both modes would be reported.

In general, the mode--as a single statistic--is rarely used and is particularly vulnerable to idiosyncratic patterns in the data. However, the concept can be helpful in describing the overall shape of a distribution.

Implementing in R

```
mean(who$life_expectancy, na.rm = T)
```

```
## [1] 71.63934
```

```
## If there are missing values of life_expectancy (NA), we will get
## an error. So, we generally want to set na.rm = T which tells R
## to ignore missing values. (By default na.rm = F)

## Are there any missing? No!
sum(is.na(who$life_expectancy))
```

```
## [1] 0
```

```
median(who$life_expectancy)
```

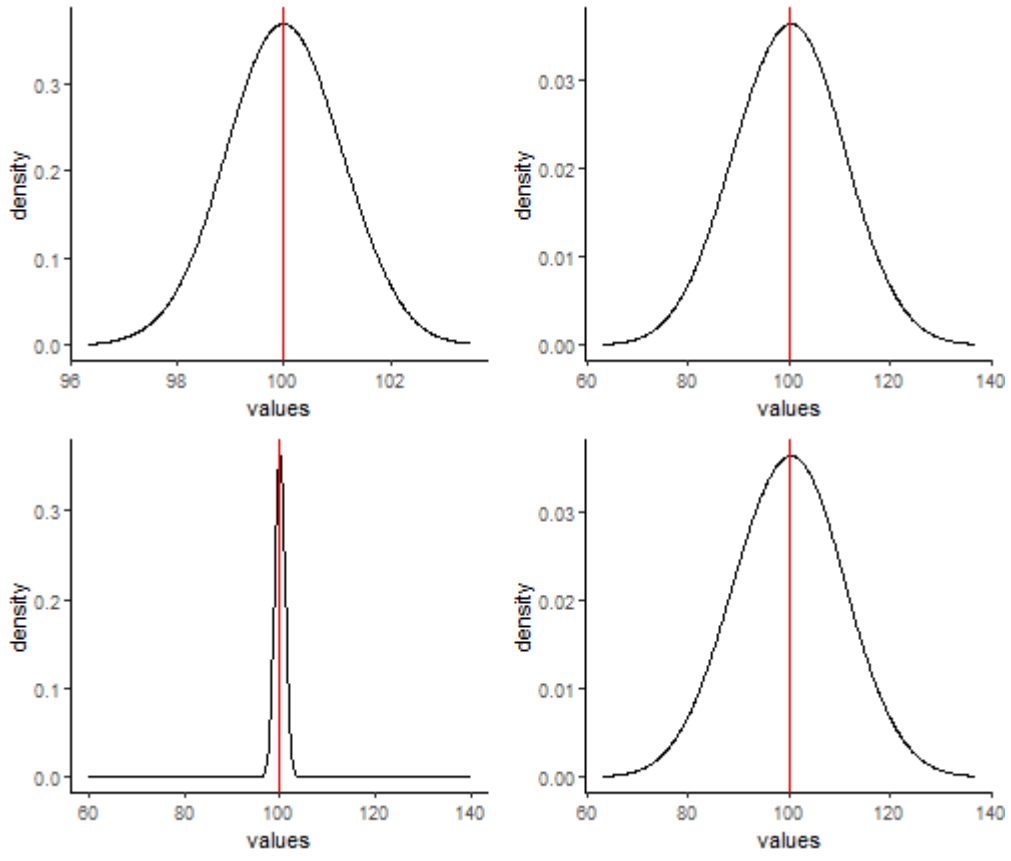
```
## [1] 74
```

Base R does not have an easy way of calculating the mode. There are some work-arounds, but we won't cover them in this class. Importantly the `mode` function does something altogether different.

Comparing distributions

These distributions have the same central tendency (mean = 100).

How are they different?



Measures of variability

Variability

- Central tendency is only one component of a distribution.
- What if we want to know how much variation there is in life expectancy across nations?
    - How much does the typical observation deviate from the mean?
- **Measures of Variability:**
    - Range
    - Interquartile Range
    - Variance
    - Standard Deviation

Range

- Represents the difference between the highest value and lowest value in the dataset.
- Can provide a rough estimate of spread

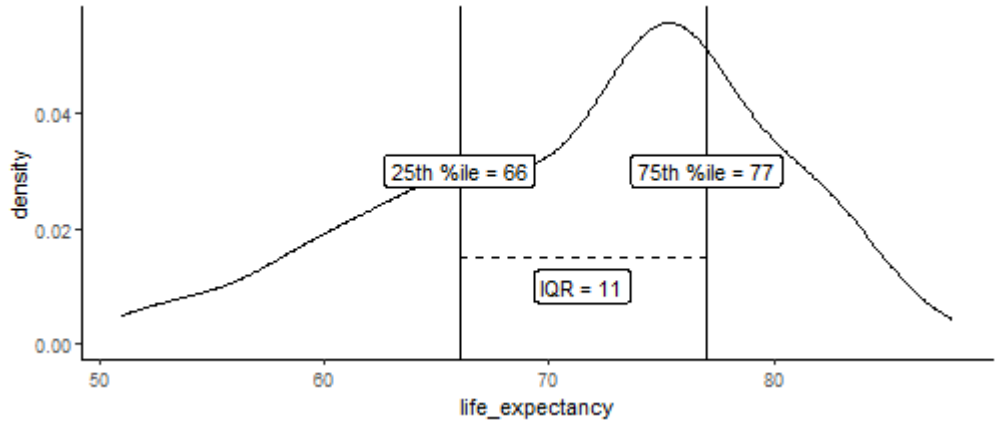
$$range(71, 73, 76, 78, 79) = 79 - 71 = 8$$

- Very vulnerable to outliers. One observation can make a big difference!

$$range(55, 73, 76, 78, 79) = 79 - 55 = 24$$

Interquartile Range (IQR)

- Represents the difference between the 1st quartile (25th percentile) and the 3rd quartile (75th percentile).
- Percentile refers to the percentage of observations that fall at or below that score.
    - e.g., 25% of observed national life expectancies fall at or below 66 years of age.
- IQR summarizes the range of the most commonly observed values in the dataset.



Implementing in R

```
range(who$life_expectancy)
```

```
## [1] 51 88
```

```
IQR(who$life_expectancy)
```

```
## [1] 11
```

```
quantile(who$life_expectancy)
```

```
##   0%  25%  50%  75% 100%
##   51   66   74   77   88
```

```
# By default will give quartiles (25 percentiles)
# Can change with 'probs' sub-command
# e.g, quantile(who$life_expectancy, probs = seq(0,1,0.1))
# will give the values of each decile btwn 0 and 1 (10 pp)
```

Variance

- Represents the average squared deviation (let's call this s^2) of each observation from the mean.[1]

$$s^2 = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{N}$$

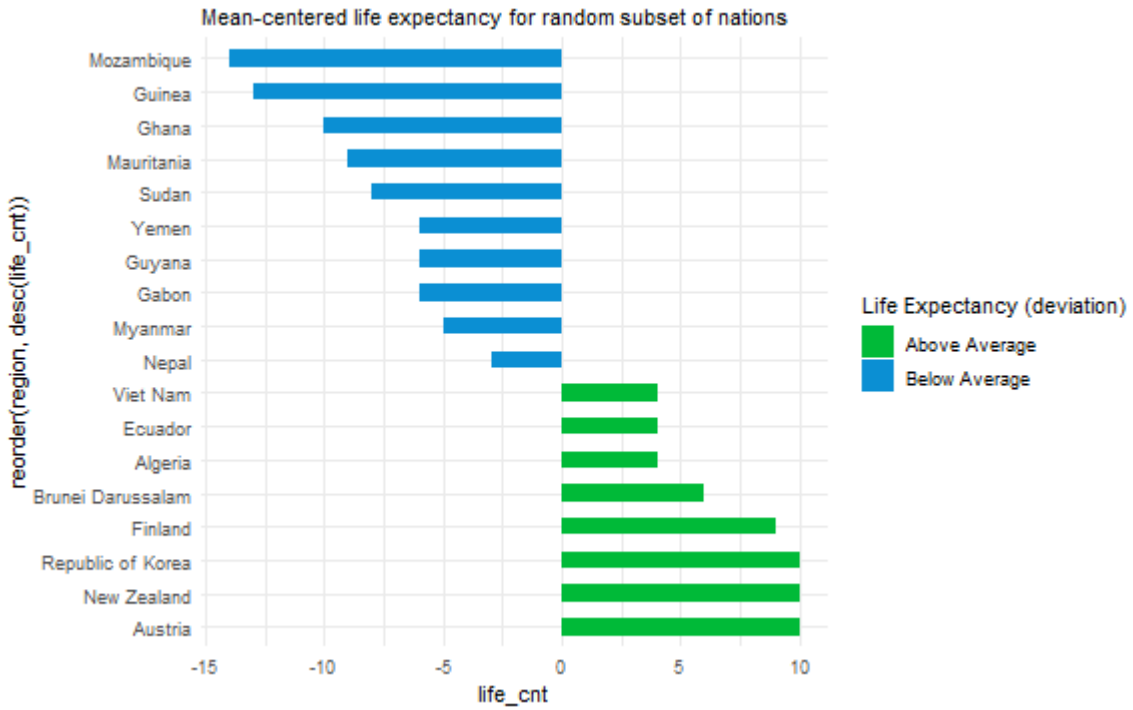
- *Why do we square the deviation?*

Let's unpack what this equation does...

[1] This is actually not quite right. When calculating a sample statistic of the variance or standard deviation, the denominator in the above equation is actually *N*-1. We will learn why when we get to *degrees of freedom* in the next unit.

Variance

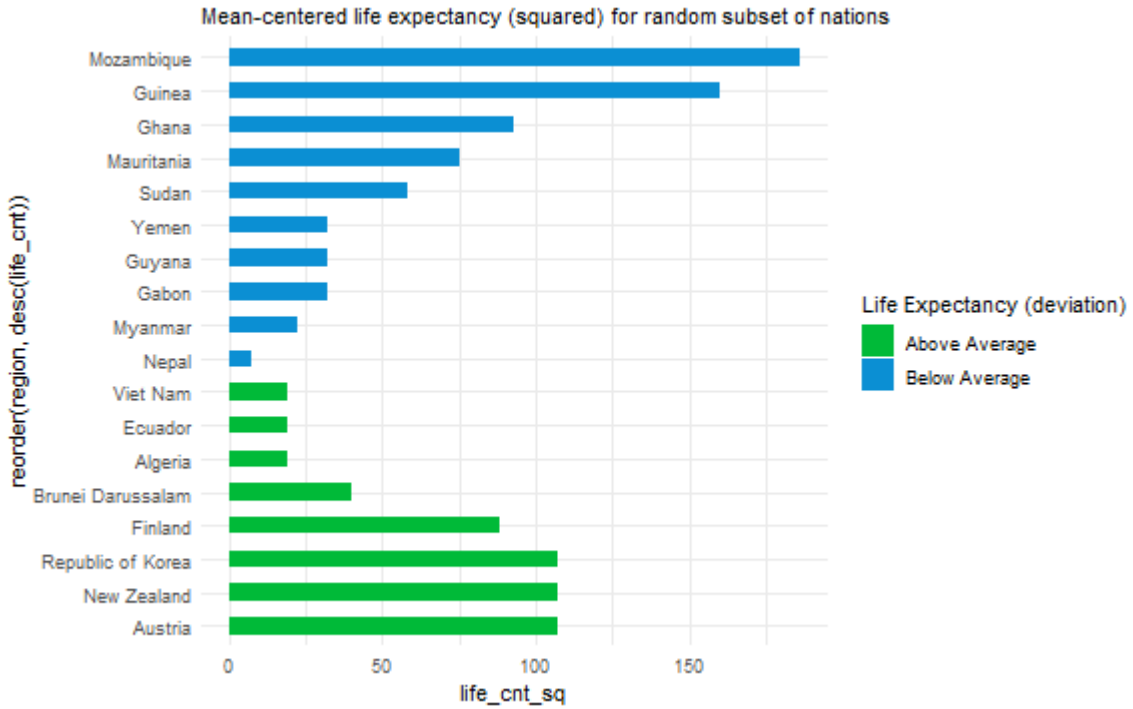
Here are some national life expectancies centered around the mean. Their value represents their difference relative to the sample mean.



Some observations fall below the mean so their deviation is a negative value. If we took the average of our positive and negative deviations, we'd just get zero!

Variance

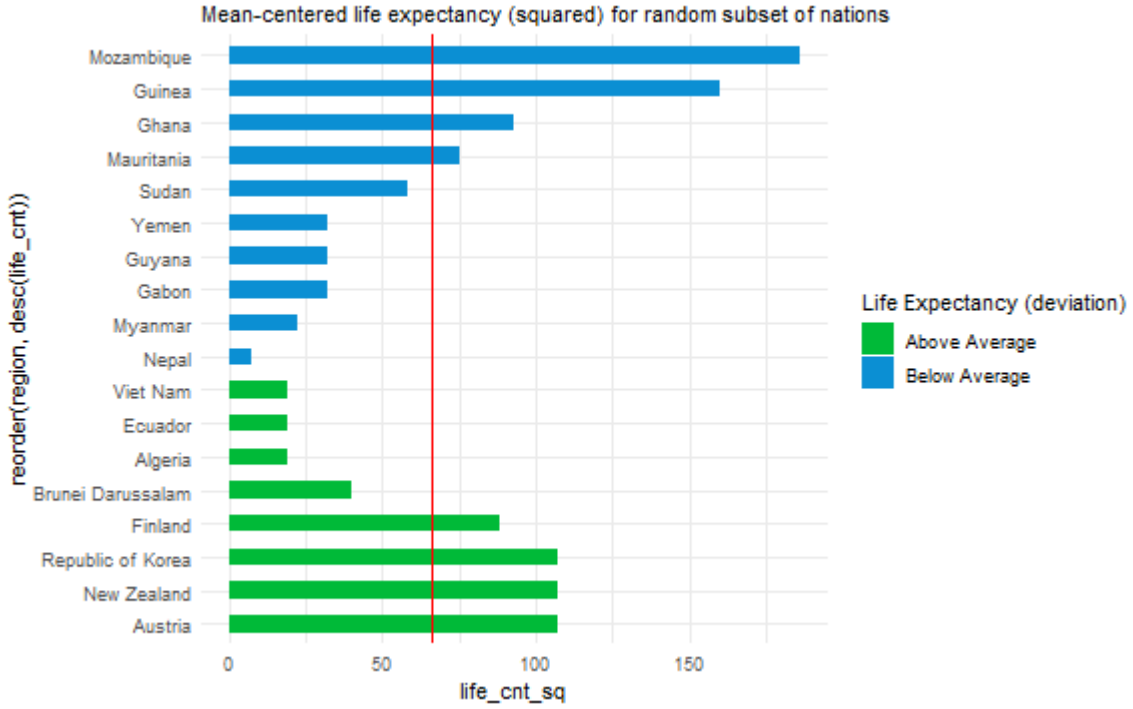
Since any number squared is positive, let's try squaring our deviations.



Now we can take the average of our squared deviations and get a non-zero value!

Variance

Our variance (or average squared deviation) is approximately 66.43.



By itself, a variance of 66.43 is not very meaningful to us. It does not reflect the raw units of our original scale (it is in squared years...what does that mean???).

Standard deviation

- If we take the square root of our sample variance, we can summarize average deviations from the mean back onto our original scale.
- The standard deviation represents the **positive square root of the variance**.
  Standard deviation =

$$s = \sqrt{s^2}$$

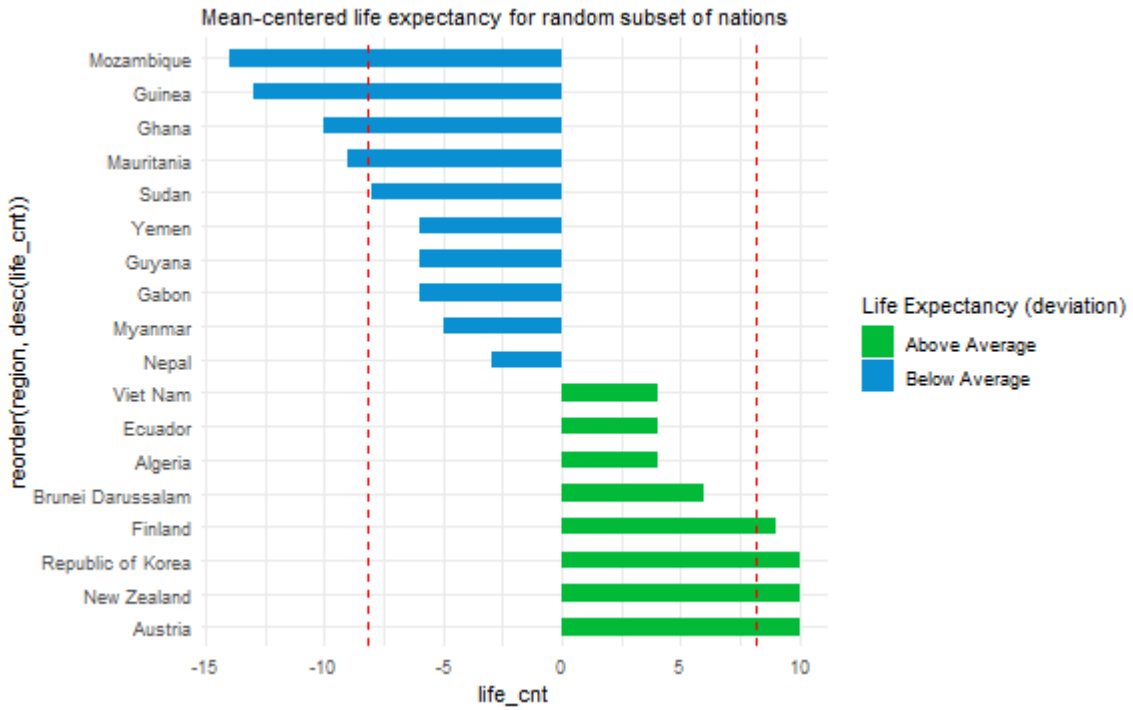
  Standard deviation =

$$\sqrt{66.43} = 8.15$$

The standard deviation of life expectancy across countries in 2015 was approximately 8.15 *years*.

Standard deviation

Here is our original plot with our standard deviation (*SD*) plotted below and above the mean (i.e., −1 *SD*, +1 *SD*).



The standard deviation of life expectancy in 2015 was approximately 8.15 *years*.

Implementing in R

```
# Variance
var(who$life_expectancy)
```

```
## [1] 66.42965
```

```
# Standard deviation
sd(who$life_expectancy)
```

```
## [1] 8.150439
```

You try

Given the following set of observed value (75, 74, 66, 78, 73, 78), calculate the:

- Mean
- Median
- Mode
- Range
- Variance
- Standard deviation

Statistical notation

Statistics usually have two notational forms:

- Population statistics (Greek alphabet) - Hypothetical values in the full universe of all possible values
- Sample statistics (Roman alphabet) - Actual values in our research sample

Sample Statistics - Roman Alphabet

- Mean - $\bar{x}$
- Standard Deviation - s
- Variance - s^2

Population Statistics - Greek Alphabet

- Mean - μ
- Standard Deviation - σ
- Variance - σ^2

Statistical notation

Often times we use sample statistics to *estimate* population statistics.

An estimated value is denoted with a "hat."

Sample Statistics $\rightarrow$ Population Estimates

- Mean - $\bar{x} \rightarrow \boldsymbol{\hat{\mu}}$
- Standard Deviation - $s \rightarrow \boldsymbol{\hat{\sigma}}$
- Variance - $s^2 \rightarrow \boldsymbol{\hat{\sigma}^2}$

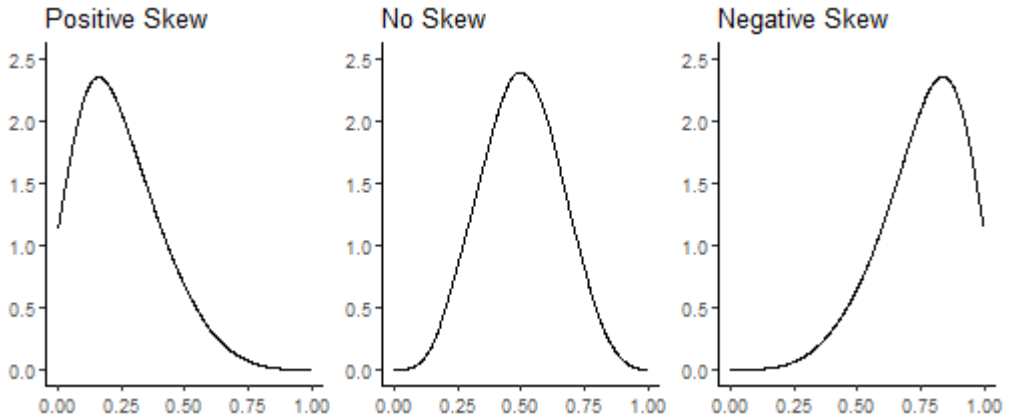
Population Statistics - Greek Alphabet

- Mean - μ
- Standard Deviation - σ
- Variance - σ^2

These Greek letters are conventions that statisticians have used over the years. There are many other of these notational conventions (for example, when to use capital- or lower-case Greek letters). There's no inherent reason why they are used (e.g., you could use ψ to represent the mean), but we tend to use the conventional symbols to reduce cognitive load. You'll want to get familiar with some of the most commonly used ones yourself.

Distributional shape: Skewness

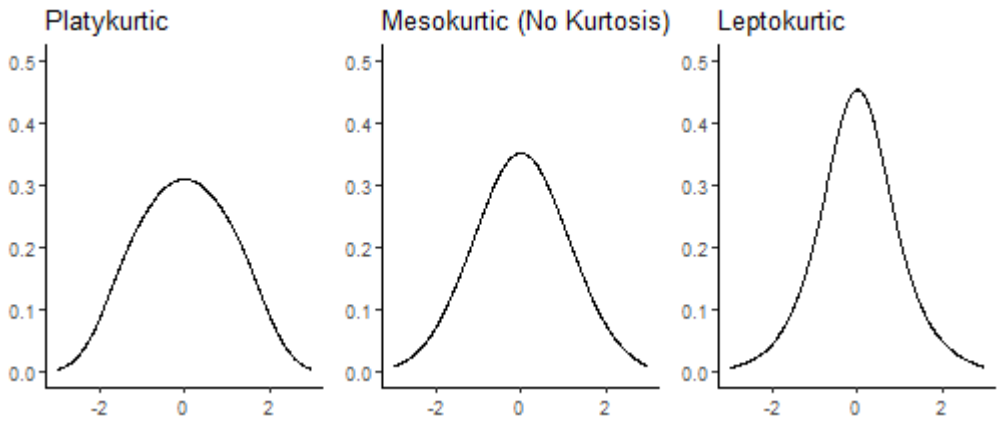
- Not all distributions in the real world are symmetrical. Often they have asymmetry or skewness.
- Skewness describes how much a distribution is "bunched up" to the right or left.



- Positive Skew - Distribution is skewed to the right with a "positive-pointing finger."
- Negative Skew - Distribution is skewed to the left with a "negative-pointing finger."

Distributional shape: Kurtosis

- Kurtosis describes how much the values concentrate around the center (the "pointiness" of a distribution)
  - Leptokurtic distributions have more values around the center and have a "taller" peak.
  - Platykurtic distributions do not have a prominent peak and have a flatter top.



Moments of a distribution

Each of these statistics we've covered describe a different "moment of a distribution."

1. Mean

$$\mu = \frac{\Sigma(x_i)}{N}$$

2. Variance

$$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$$

3. Skewness

$$skewness(X) = \frac{1}{N\sigma^3}\Sigma(X_i - \mu)^3$$

4. Kurtosis

$$kurtosis(X) = \frac{1}{N\sigma^4}\Sigma(X_i - \mu)^4 - 3$$

We do not need to know how to calculate skew and kurtosis but we do need to have a sense of how to interpret them.

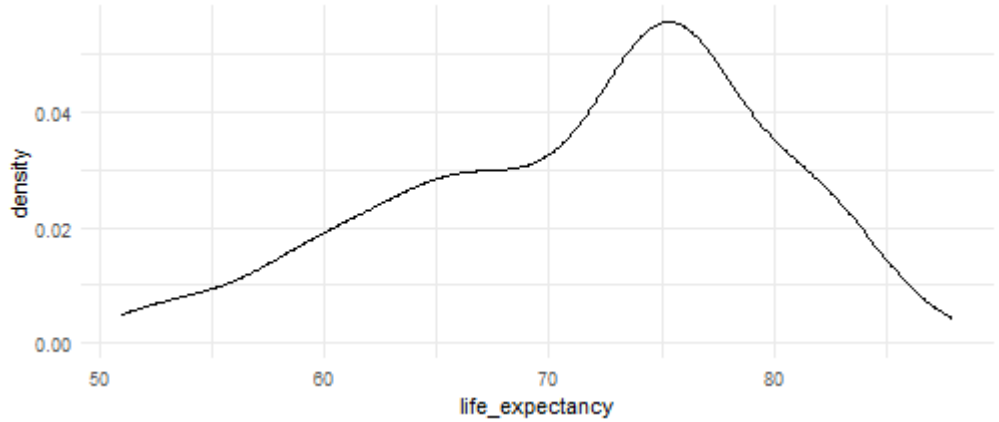
Evaluating skew and kurtosis

- Skewness and kurtosis can be evaluated visually and summarized quantitatively.
  - Positive skew (skewed to the right) $\rightarrow$ Skewness > 0
  - Negative skew (skewed to the left) $\rightarrow$ Skewness < 0
  - Leptokurtic $\rightarrow$ Kurtosis > 0
  - Platykurtic $\rightarrow$ Kurtosis < 0

Note: In this course we will focus visual analysis for severe skew or kurtosis. Future courses will discuss how to correct for skew or kurtosis, as necessary.

Summarizing data

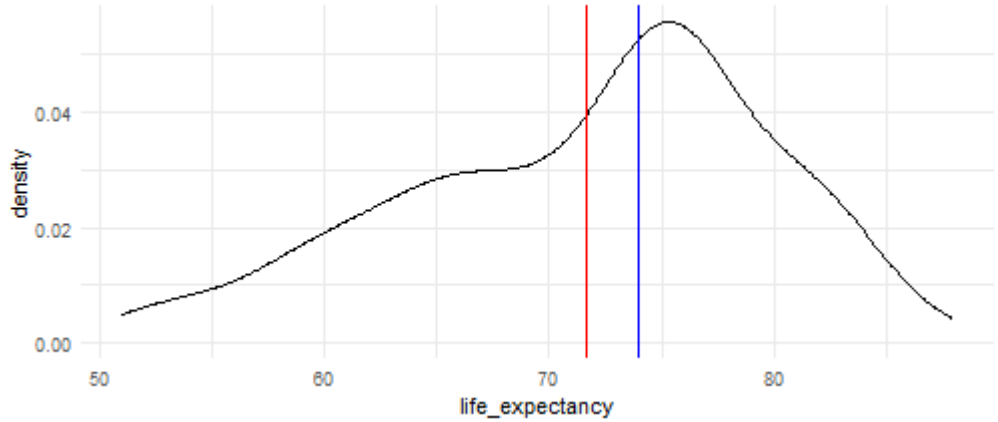
- Here is a density plot of the WHO life expectancy data for 2015.
- How would you characterize the distribution?



- What kind of skew does this distribution have?
- Is there severe skew?
- Is there severe kurtosis?

Summarizing data

The distribution of life expectancy has a slightly negative skew and no excess kurtosis. The skewness and kurtosis do not appear to be severe.



The mean national life expectancy in 2015 was **71.62** years.

The median national life expectancy in 2015 was **74** years.

- Due to the slight skew, the median appears to better capture the "center" of our distribution.

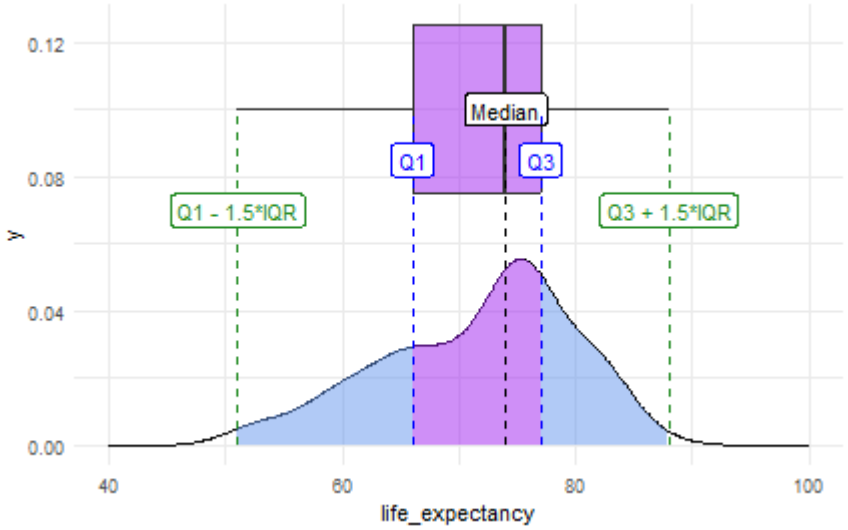
Other visualizations

Although histograms and density plots are useful for evaluating the shape of a distribution, they do not efficiently summarize the central tendency and variance.

Box-and-whisker plot (aka "boxplot")

- Provides five helpful numbers of a distribution
  - Q1, Q2 (Median), Q3
  - Lower Fence (Q1 - 1.5*IQR)
  - Upper Fence (Q3 + 1.5*IQR)

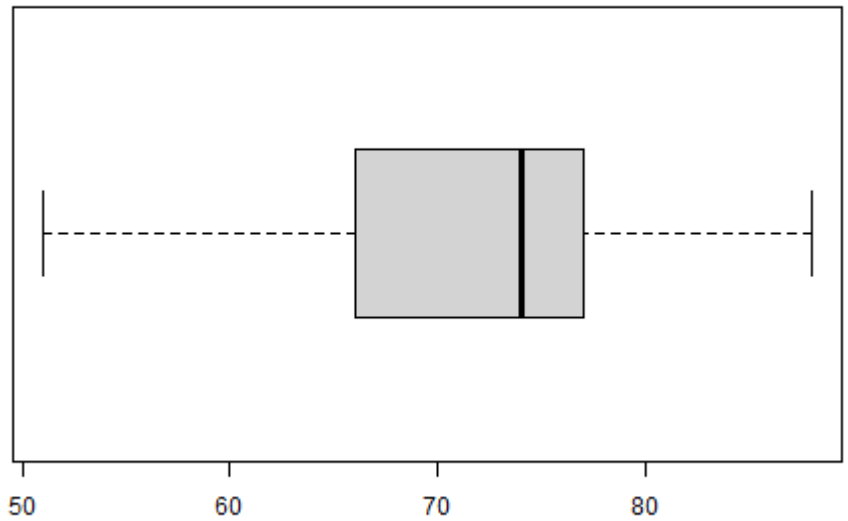
Here is a sample boxplot of our WHO data overlain on a density plot.



Boxplots

A basic boxplot can quickly capture the median, the interquartile range and any outlying values (we don't have any of these in our data).

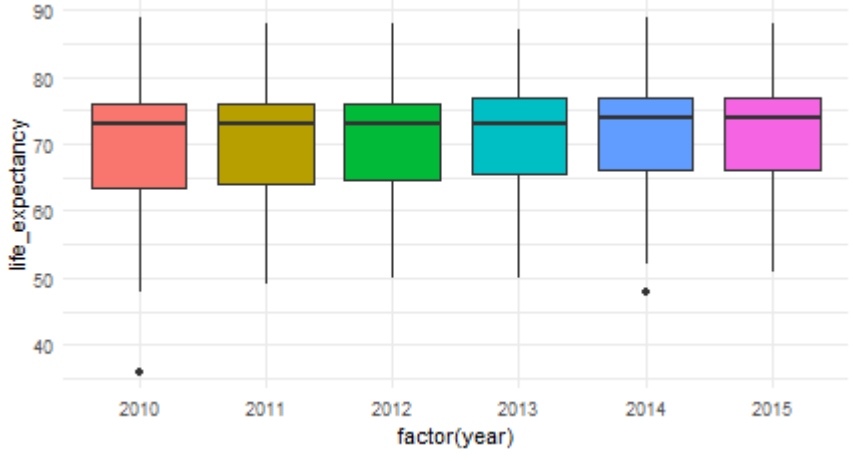
```
boxplot(who$life_expectancy, horizontal=T)
```



Boxplots

However, we frequently visualize distributions across several groups for comparison.

```
ggplot(data = who_more, aes(factor(year), life_expectancy,
                            fill = factor(year))) +
        geom_boxplot() +
        theme_minimal() + theme(legend.position = "none")
```



> It appears the life expectancy of Q1 increases more each year than the median or Q3. We might hypothesize that lower-life-expectancy countries made bigger gains from 2010-2015 than countries with higher life expectancies. **But more exploration would be needed to confirm this!**

Summing up

So we have characterized some features of the average national life expectancy among countries in the WHO database. However, we are still interested in asking more complex questions such as:

> - How does life expectancy compare in high- vs. middle- and low-income countries?
> - Do individuals living in countries with more total years of attendance in school experience, on average, higher life expectancy?

That is where we will turn to in the next unit! But first, we need to learn more about how to transform distributions and make statistical inferences!

Synthesis and wrap-up

Class goals

- Describe and summarize quantitative data that are continuous
- Describe the purpose and compute the following measures of central tendency: mean, median and mode
- Describe the purpose and compute the following measures of variability: quartiles, inter-quartile range, range, variance and standard deviation
  - Describe conceptually the principles of skewness and kurtosis
- Create visualizations of quantitative data that are continuous using R
  - Includes constructing histograms, densities, stem-and-leaf, and box-and-whisker plots

To Dos

Reading

- LSWR Chapter 5: descriptive statistics

Optional follow-up

- Complete Module 8 (dataframes) in R Bootcamp
- Complete Module 10 (data management) in R Bootcamp

Quiz

- Covering Unit 3 on Oct. 31

Assignments

- Assignment #3 due November 7, 11:59pm

Complete midterm SES! If >=75% of the class does so, you all receive 1 extra percentage point on your grade!