Descriptives - Redux

Mid Semester Check-in

These first 5 weeks have been jam packed with getting you up to speed using `R`. By now, you should feel comfortable with:

- Importing data
- Manipulating data (getting subsets, using logical operators, knowing the different data classes and when one is more appropriate than another, etc.)
- Plotting data with `ggplot2`

Where we're going

The next 5 weeks are going to be dedicated in getting you up to speed on statistics. Everyone in this class should have taken Psych 300 (Intro Psych Stats) or equivalent. ***If you have not taken either of these classes, you must get in touch with me ASAP.***

What you can expect:

- Refresh your memory about stats. The content presented will hopefully go into slightly more depth (for certain topics) than what you covered in Psych Stats. But it shouldn't be new, per se.
- Applying this theoretical knowledge to practical knowledge in R. You will be expected to know what the outputs mean (e.g., how to interpret them).
- As of now, there will not be any Practice Sets for this unit on statistics. I will update you if there are any changes on this front.
- No HW assignments. Quizzes based on lecture content (10 questions max/quiz). Lecture content longer and denser...but no HW!!

Why are we *not* going to talk about more advanced stats in these few weeks?

This time

- Descriptive Stats (moments of a distribution)
- Bias
- z-scores

Why do we describe data?

- Understand your data
  - There's a lot to learn from descriptive statistics
- Find errors in data entry or collection

Happiness

Examples today are based on data from the 2015 World Happiness Report, which is an annual survey part of the Gallup World Poll.

You should be able to download the dataset from 20: Descriptives lecture.

Get in touch with me ASAP if you cannot download this dataset

`world` Data

Country	Happiness	GDP	Support	Life	Freedom	Generosity	Corrupt
Albania	4.606651	9.251464	0.6393561	68.43517	0.7038507	-0.0823377	0.8847
Argentina	6.697131	NA	0.9264923	67.28722	0.8812237	NA	0.8509
Armenia	4.348319	8.968936	0.7225510	65.30076	0.5510266	-0.1866965	0.9014
Australia	7.309061	10.680326	0.9518616	72.56024	0.9218710	0.3157020	0.3565
Austria	7.076447	10.691354	0.9281103	70.82256	0.9003052	0.0890886	0.5574
Azerbaijan	5.146775	9.730904	0.7857028	61.97585	0.7642895	-0.2226351	0.6155

About the `world` dataset

```
colnames(world)
```

```
## [1] "Country"    "Happiness"  "GDP"        "Support"    "Life"
## [6] "Freedom"    "Generosity" "Corruption"
```

Happiness: “Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?”

GDP: Log gross domestic product per capita

Support: “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”

Life: Healthy life expectancy at birth

Freedom: “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”

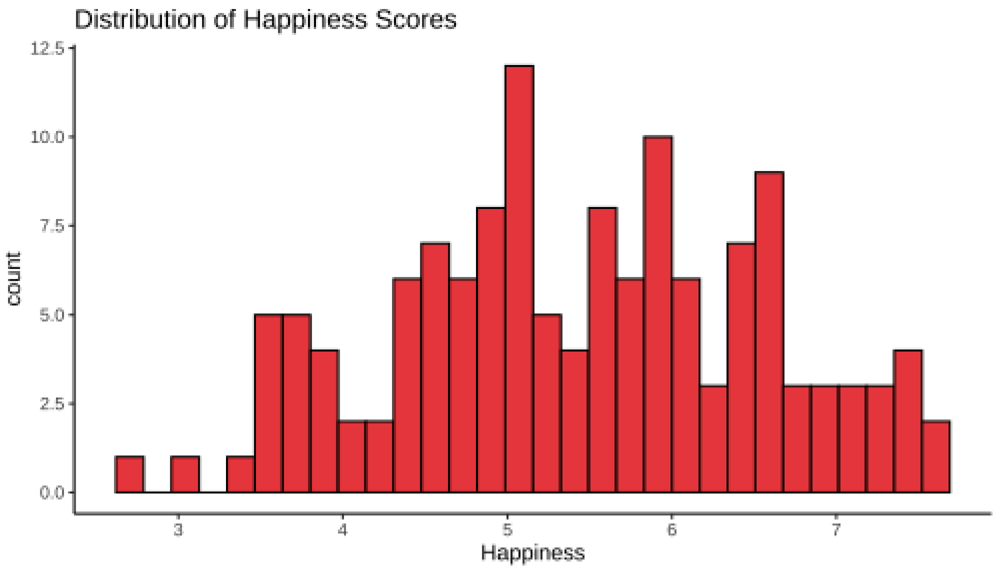
Corruption: “Is corruption widespread throughout the government or not” and “Is corruption widespread within businesses or not?” (average of 2 questions)

Generosity: “Have you donated money to a charity in the past month?” (residual, adjusting for GDP)

Distributions

A **distribution** often refers to a description of the [relative] number of times a variable X will take each of its unique values.

```
ggplot(data = world, aes(x = Happiness)) +
  geom_histogram(bins = 30, fill = "#eb4d4b", color = "black") +
  labs(title = "Distribution of Happiness Scores",
       xlab = "Happiness",
       ylab = "Frequency") +
  theme_classic()
```



Moments of a distribution

1. Mean
2. Variance
3. Skew
4. Kurtosis

Mean

- The **mean** is the average
- The population mean is represented by the Greek symbol μ
- The sample mean is represented by the Latin symbol $\bar{X}$
- Example: a set of numbers is: `7, 7, 8, 3, 9, 2`.

$$\mu = \frac{\Sigma(x_i)}{N} = \frac{7+7+8+3+9+2}{6} = \frac{36}{6} = 6$$

Properties of the mean

- The mean can take a value not found in the dataset
- Fulcrum of the data
- The mean is strongly influenced by outliers
- Deviations from the mean sum to 0
    - Deviation score: $x - \bar{x}$
    - Sum of Deviation Scores: $\Sigma(x - \bar{x})$
    - $\Sigma(x - \bar{x}) = 0$
- Can only be used with interval- and ratio-level variables
    - no nominal or ordinal variables

Mean

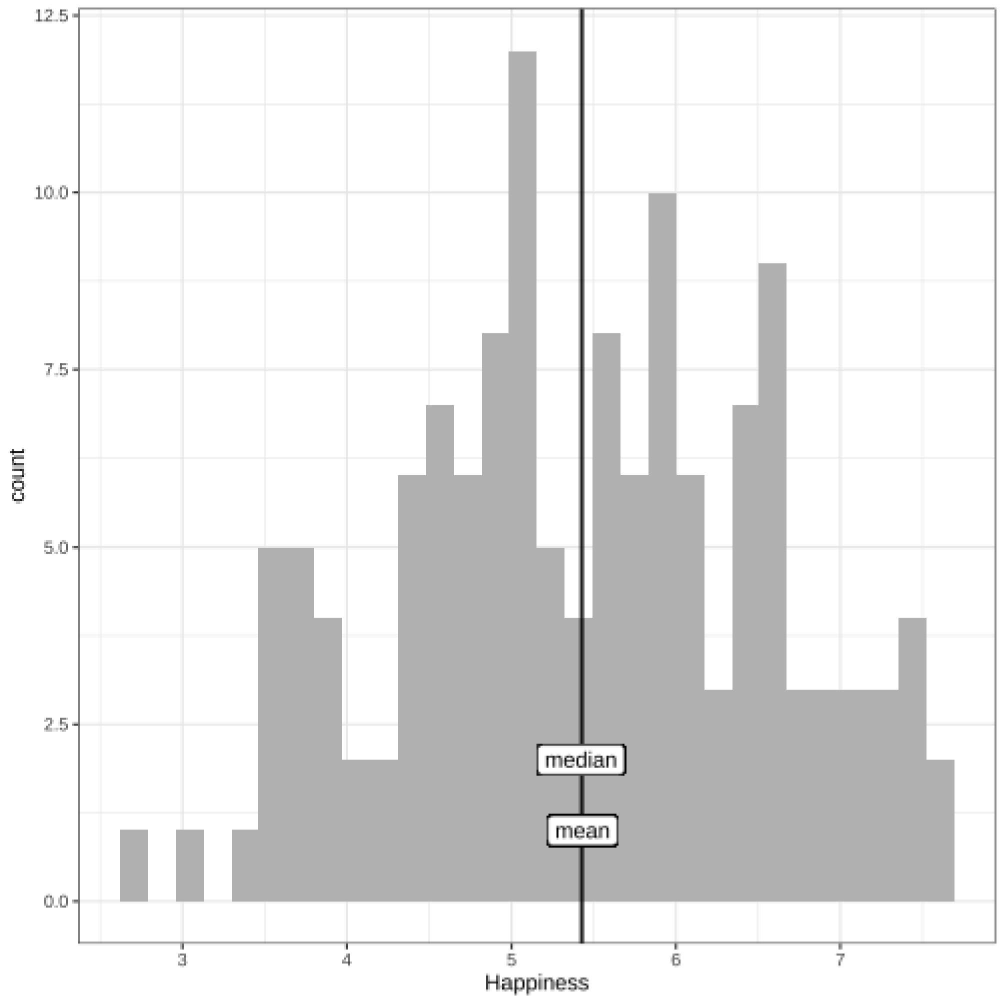
It's important to remember that the mean of a population (or group) may not represent well some (or any) members of the population.

- Example: André-François Raffray and the French apartment

Other measures of central tendency

- **Median** -- the middle point of the data
  - e.g., in the set of numbers `7, 7, 8, 3, 9, 2`, the median number is 7
  - Must put them in order first (highest to lowest or vice versa), then find the middle point
  - If even number, get the "middle" of the 2 middle-est numbers (average of 2 middle-est)
  - Can be used with ordinal-, interval-, or ratio-level variables
- **Mode** -- the number that most commonly occurs in the distribution.
  - e.g., in the set of numbers above, the mode is 7
  - Can be used with any kind of variable

If Normal, All Relatively Equal



Center and spread

- Distributions are most often described by their mean and **variance**.
- Typically, these moments are the two used in common inferential techniques.
- The mean represents the average score in a distribution. A good measure of spread will tell us something about how the typical score deviates from the mean.
  - $x - \bar{x}$
- Why can't we use the average deviation?

Average deviation

```r
x <- c(7,7,8,3,9,2)
mean(x)
```

```
## [1] 6
```

```r
x - mean(x)
```

```
## [1]  1  1  2 -3  3 -4
```

```r
sum(x - mean(x))
```

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

```r
sum(x - mean(x))/length(x)
```

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

Sum of Squares (SS)

Our solution is to square the deviation scores

```r
x <- c(7,7,8,3,9,2)
mean(x)
```

```
## [1] 6
```

```r
deviation <- x - mean(x)
deviation^2
```

```
## [1]  1  1  4  9  9 16
```

```r
sum(deviation^2)
```

```
## [1] 40
```

Is there any inherent meaning in the Sum of Squares?

Variance

We calculate the average squared deviation: this is our variance, σ^2:

```
# nested functions galore!
sum((x - mean(x))^2)/length(x)
```

```
## [1] 6.666667
```

Good things about variance:

- It's additive.
  - Given two variables X and Y, if I create $Z = X + Y$ then $Var(Z) = Var(X) + Var(Y)$
- Represents all values in a dataset

Bad things about variance:

- What the heck does it mean?

Standard Deviation

Standard deviation σ is the square root of the variance.

```
sqrt(sum((x - mean(x))^2)/length(x))
```

```
## [1] 2.581989
```

Skew and Kurtosis

Skewness = asymmetry

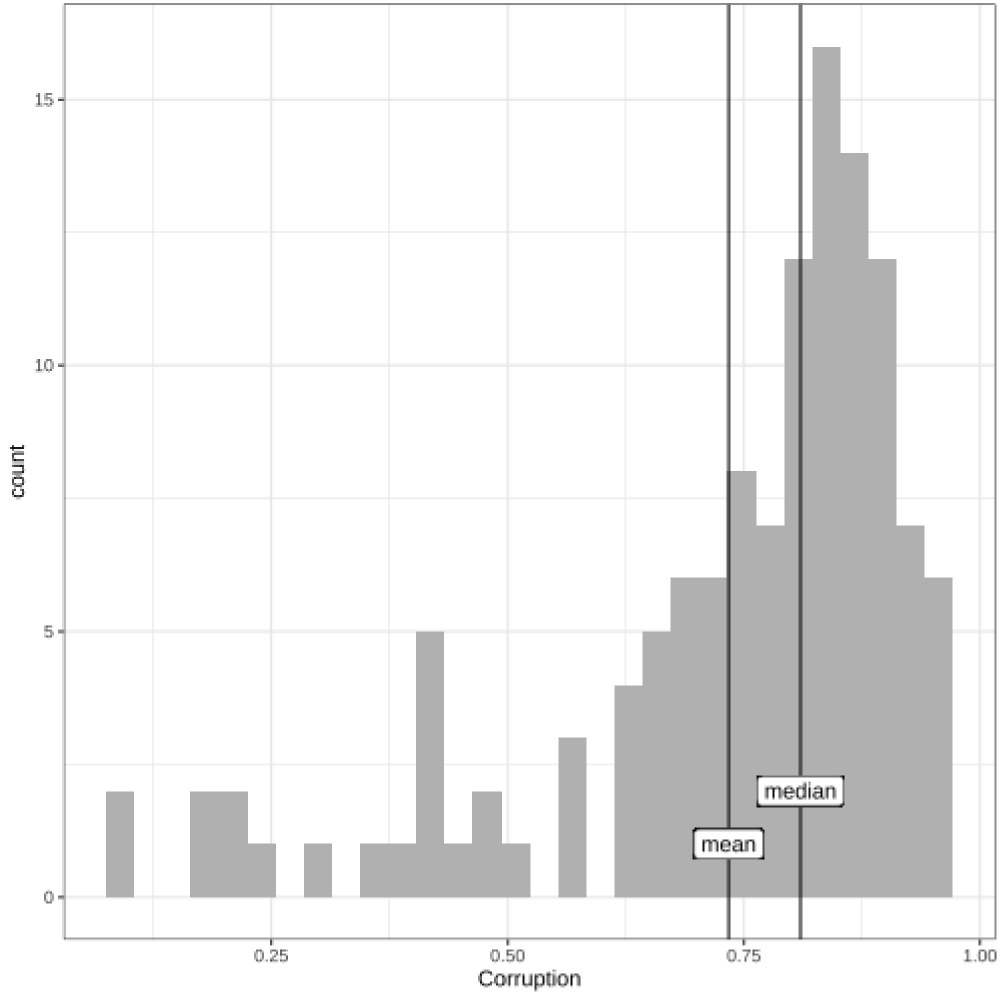
- Negative skew = tail pointed towards the negative values (left)
- Positive skew = tail pointed towards the positive values (right)

Kurtosis = pointyness

- Too pointy = leptokurtic
- Perfect = mesokurtic
- Too flat = platykurtic (as in platypus!)

Most inferential statistics assume distributions are not skewed and are mesokurtic.

If Skewed...



Bias

Populations versus Samples

Why are these different?

```r
# how we calculated variance "by hand" in this lecture
sum((x - mean(x))^2)/length(x)
```

```
## [1] 6.666667
```

```r
# R's default variance function
var(x)
```

```
## [1] 8
```

Populations versus Samples

The value that represents the entire *population* is called a **parameter**.

- Population parameters are represented with Greek letters (μ , σ)

We collect samples to estimate the properties of populations; the value that represents a *sample* is called a **statistic**.

- Sample statistics are represented with Latin letters (x, $\bar{x}$, s).

Bias: An estimator is biased if its expected value and the true value of the parameter are different.

- Our estimates of standard deviation & variance (in the formulas up until now) are biased
- They *under*estimate variability in the population

Populations versus Samples

Variance

Population

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

Sample

$$s^2 = \hat{\sigma}^2 = \frac{\Sigma(X_i - \bar{X})^2}{N - 1}$$

Standard Deviation

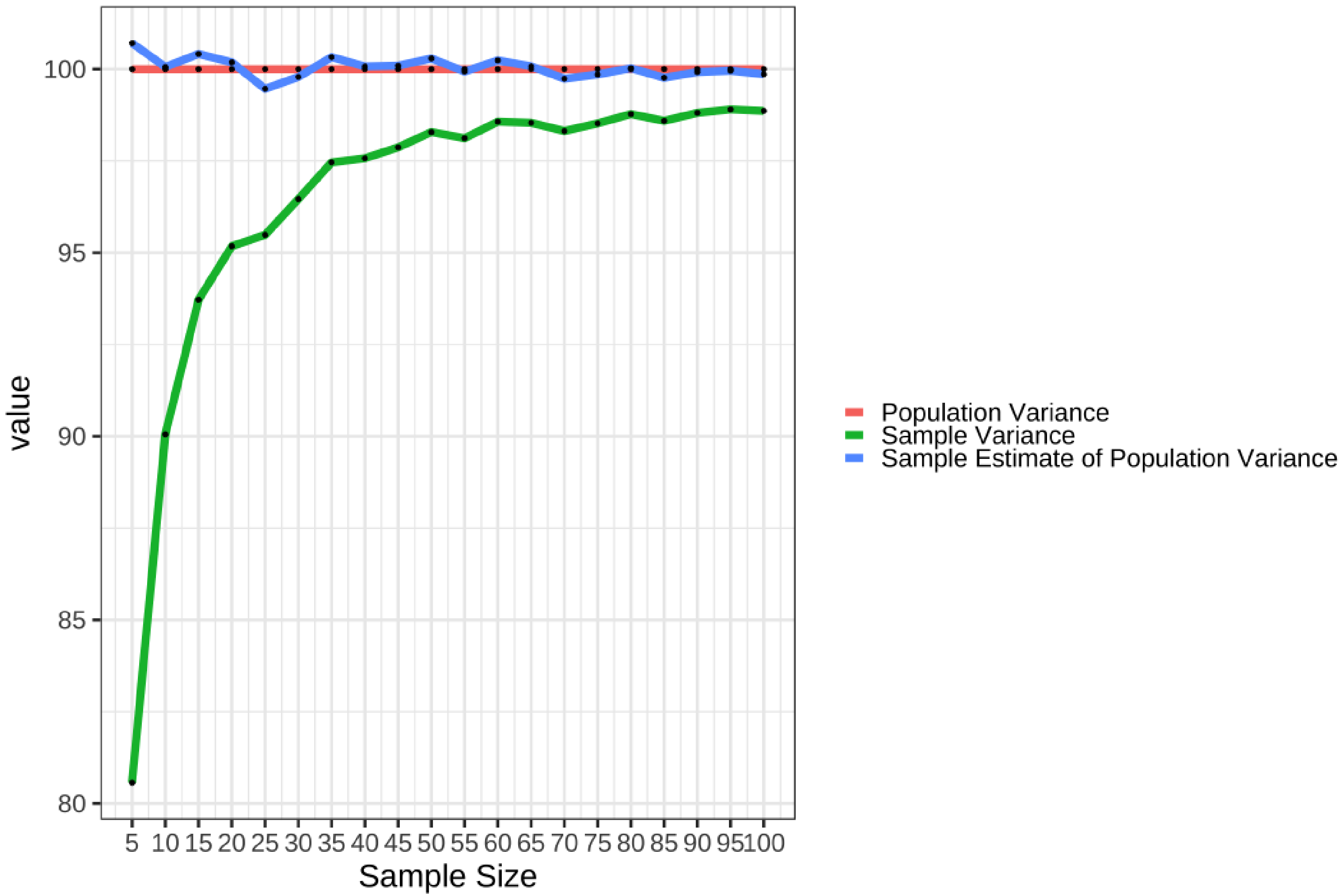
Population

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

Sample

$$s = \hat{\sigma} = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N - 1}}$$

Simulating Bias



Standardized scores

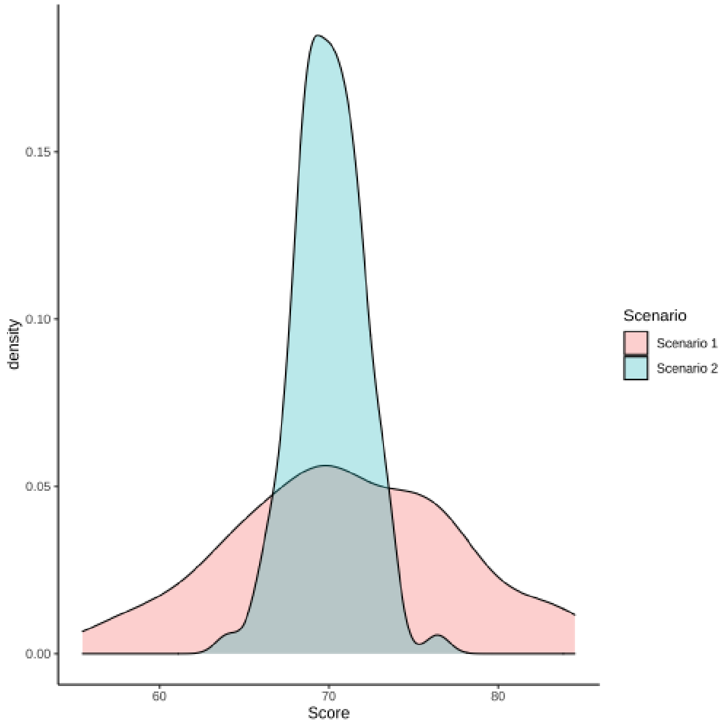
Why not use raw scores for everything?

Scenario 1:

- Mean = 70, N = 100
- Range = 0 - 100
- You get a 71. Good?

Scenario 2:

- Mean = 70, N = 100
- Range = 65 - 75
- You get a 71. Good?



Problem with Raw Scores

- Raw scores are only meaningful in the *context* of the distribution
- What distirbution are you looking at? Patients, controls, patients + controls etc.?
- What does the distribution itself look like? Skinny, fat?
- A raw score can't take all of this into consideration! What does an exam score of `71` mean *in context*?

z-scores

- Raw scores are in the original metric's units (exam points, heigh in inches etc.)
- z-scores are in units of standard deviation; aka "standardized scores"
- Interpretation: distance from the mean, in standard deviations.

Formula:

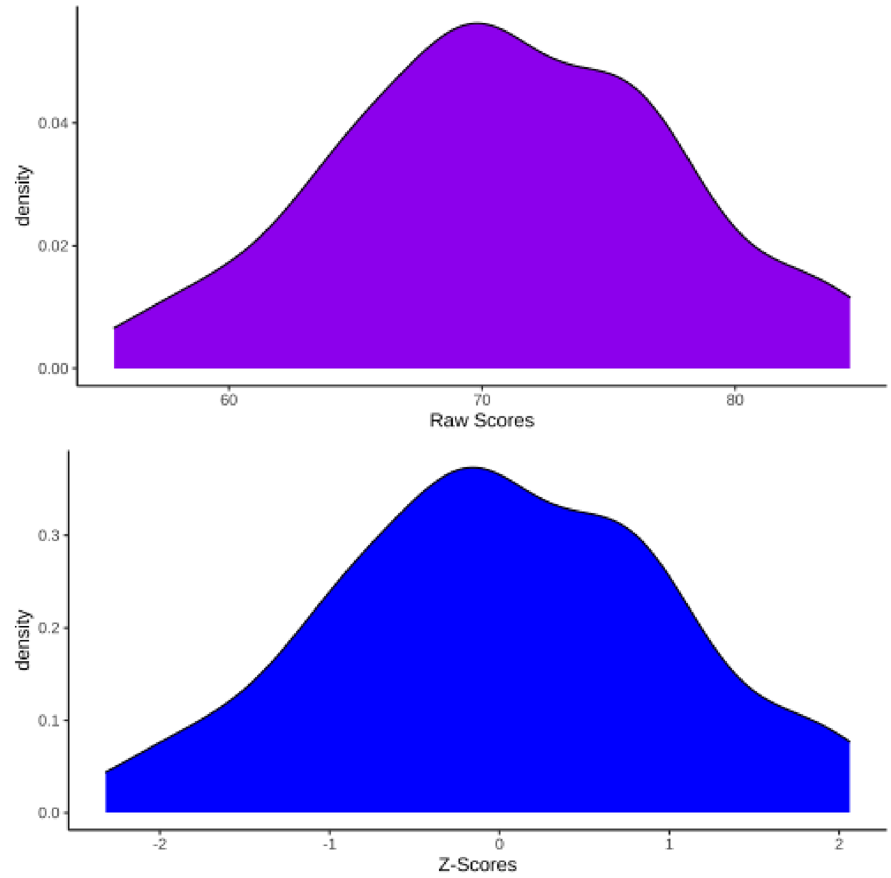
$$z = \frac{x_i - \bar{x}}{s}$$

z-scores

Step 1: Take an entire set of raw scores (x_i)

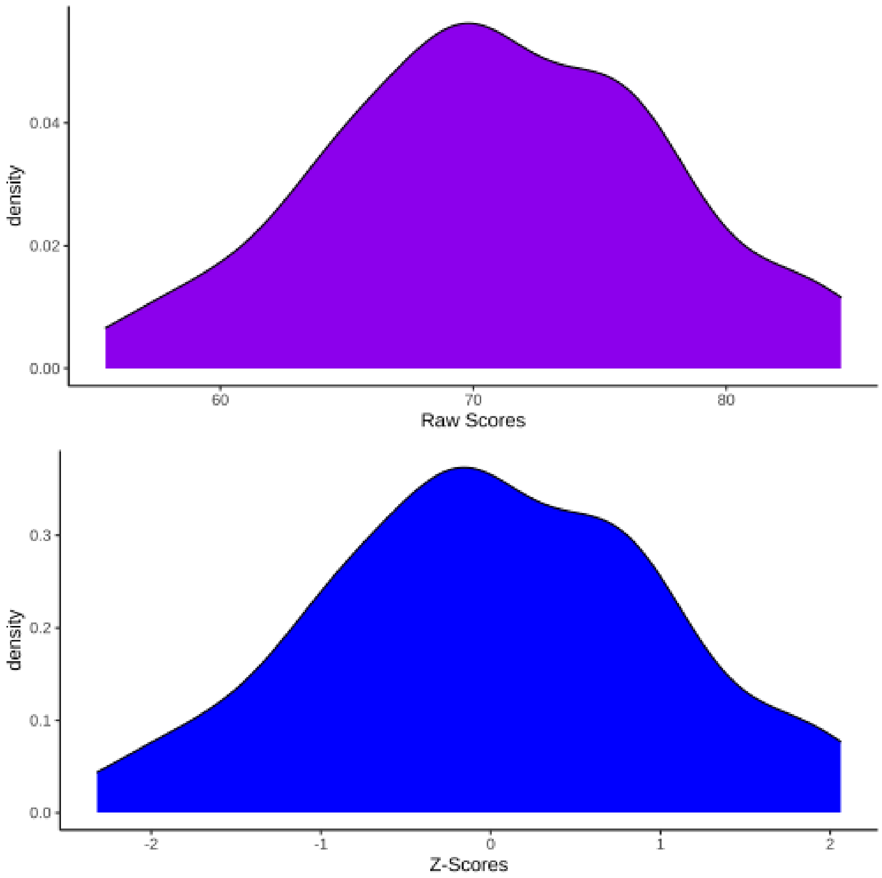
Step 2: Convert them into z-scores

Step 3: Now look at the distribution of z-scores



Properties of z-scores

- $\bar{x} = 0$
- $s = 1$



z-scores

$$z = \frac{x_i - \bar{x}}{s}$$

Why is this useful?

- Compare across scales and unit of measures
- More easily identify extreme data

Which variable has outliers?

```
psych::describe(world, fast =T)
```

```
##              vars   n  mean   sd   min   max range   se
## Country         1 136   NaN   NA   Inf  -Inf  -Inf   NA
## Happiness       2 136  5.43 1.11  2.70  7.60  4.90 0.10
## GDP             3 121  9.22 1.16  6.61 11.43  4.82 0.11
## Support         4 135  0.80 0.12  0.43  0.99  0.55 0.01
## Life            5 135 63.12 7.46 43.74 76.04 32.30 0.64
## Freedom         6 132  0.75 0.13  0.40  0.98  0.58 0.01
## Generosity      7 120  0.00 0.16 -0.28  0.46  0.74 0.01
## Corruption      8 125  0.73 0.20  0.09  0.96  0.87 0.02
```

Which variable has outliers?

```
world %>%
  mutate_if(is.numeric, scale) %>%
  psych::describe(., fast =T)
```

```
##             vars   n mean sd   min  max range   se
## Country        1 136  NaN NA   Inf -Inf  -Inf   NA
## Happiness      2 136    0  1 -2.46 1.96  4.42 0.09
## GDP            3 121    0  1 -2.26 1.91  4.17 0.09
## Support        4 135    0  1 -2.99 1.51  4.50 0.09
## Life           5 135    0  1 -2.60 1.73  4.33 0.09
## Freedom        6 132    0  1 -2.64 1.71  4.34 0.09
## Generosity     7 120    0  1 -1.80 2.90  4.70 0.09
## Corruption     8 125    0  1 -3.20 1.14  4.34 0.09
```