Describing Data Part 1

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.

The dataset is available on our GitHub site in the `data` folder, for those who want to play along at home.

```
world = read.csv("../data/world_happiness_2015.csv")
head(world)
```

```
##       Country Happiness       GDP   Support     Life   Freedom  Generosit
## 1     Albania  4.606651  9.251464 0.6393561 68.43517 0.7038507 -0.0823376
## 2   Argentina  6.697131        NA 0.9264923 67.28722 0.8812237           N
## 3     Armenia  4.348320  8.968936 0.7225510 65.30076 0.5510266 -0.1866965
## 4   Australia  7.309061 10.680326 0.9518616 72.56024 0.9218710  0.3157019
## 5     Austria  7.076447 10.691354 0.9281103 70.82256 0.9003052  0.0890885
## 6  Azerbaijan  5.146775  9.730904 0.7857028 61.97585 0.7642895 -0.2226351
##    Corruption
## 1  0.8847930
## 2  0.8509062
## 3  0.9014622
## 4  0.3565544
## 5  0.5574796
## 6  0.6155525
```

```
names(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?”

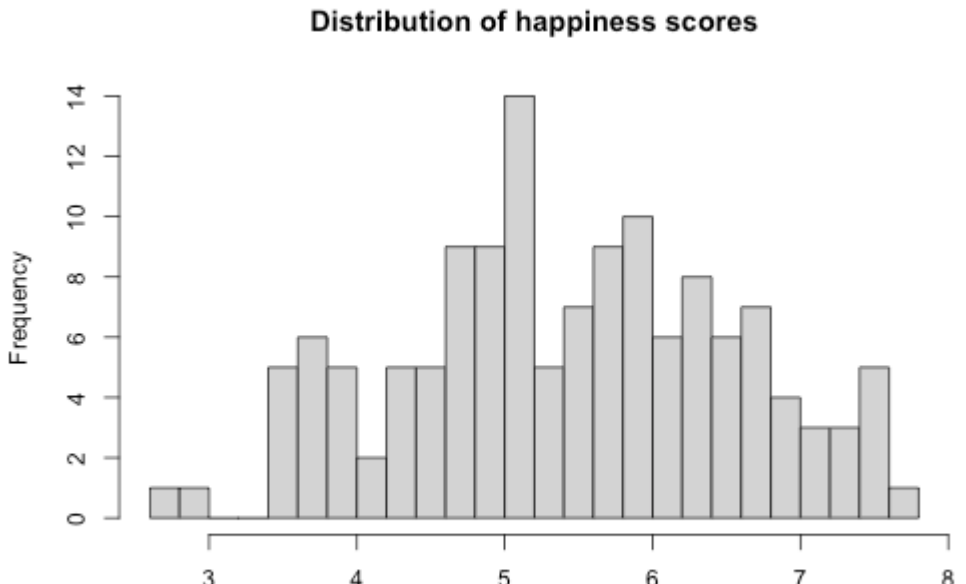
Generosity: “Have you donated money to a charity in the past month?” (residual, adjusting for GDP)

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

Distributions

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

```
hist(world$Happiness, breaks = 30, main = "Distribution of happir
     xlab = "Happiness")
```



Moments of a distribution

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

Mean, μ

- 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
- Can only be used with interval- and ratio-level variables.

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.
  - 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.

Center and spread

- Distributions are most often described by their first two moments, 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.
- *Let's take the average of how far numbers deviates from the mean*

Average deviation

```
x = c(7,7,8,3,9,2)
mean(x)
```

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

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

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

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

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

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

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

Sums of squares

Our solution is to square deviations.

```
x = c(7,7,8,3,9,2)
mean(x)
```

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

```
deviation = x - mean(x)
deviation^2
```

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

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

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

The sum of squared deviations is referred to as **the Sum of Squares (SS)**

Variance

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

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

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

Variance

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:

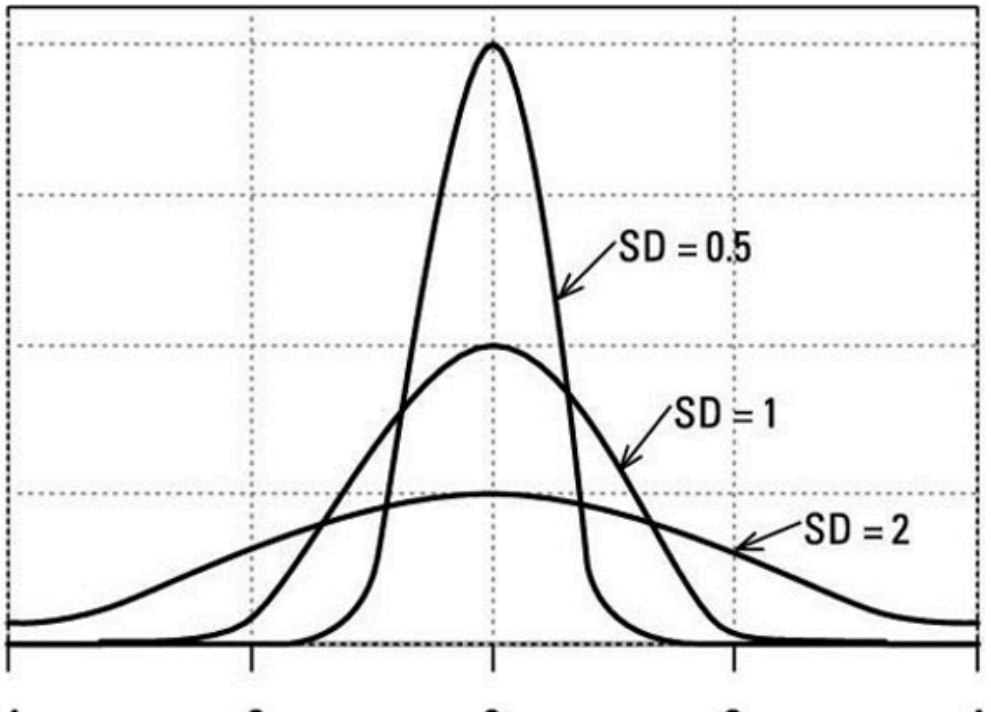
- What the heck does it mean?

Standard Deviation

Standard deviation σ is the square root of the variance.

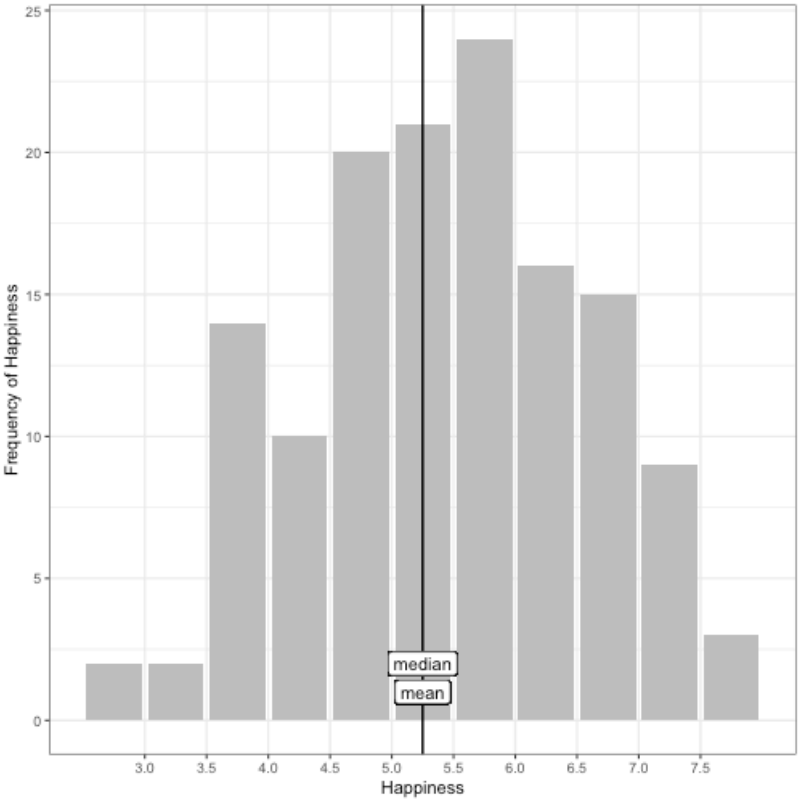
```
sqrt(sum(deviation^2)/length(deviation))
```

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



```
world %>% ggplot(aes(x = Happiness)) +
  geom_bar(fill = "gray") +
  scale_x_binned() +
  geom_vline(aes(xintercept = mean(Happ
  geom_vline(aes(xintercept = median(Ha
  geom_label(aes(x = mean(Happiness), y
  geom_label(aes(x = median(Happiness),
  labs(y = "Frequency of Happiness") +
  theme_bw()
```

In a normal distribution, the mean, median, and mode are all relatively equal.

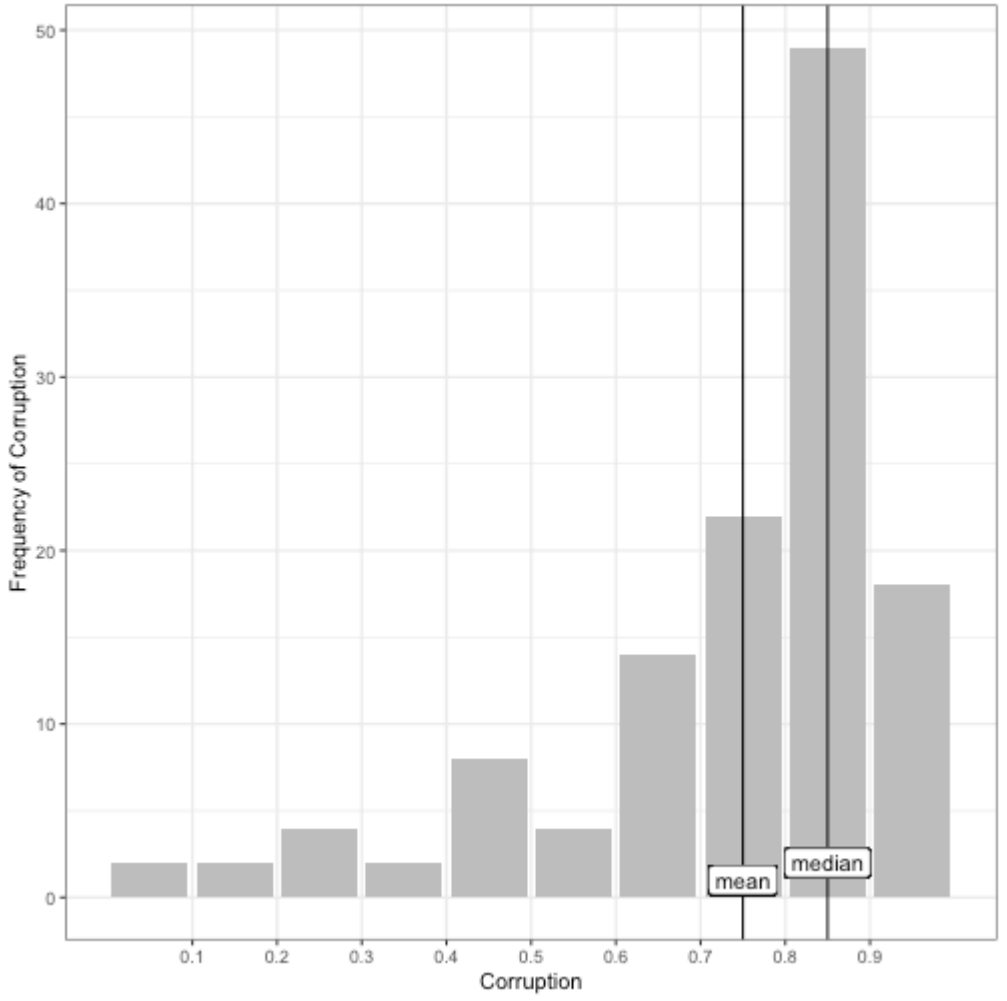


Skew and Kurtosis

Moments 3 and 4 of a distribution are **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; $+$ kurtosis
  - Perfect = mesokurtic
  - Too flat = platykurtic (as in platypus!); $-$ kurtosis

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



In a skewed distribution, both the mean and median get pulled away from the mode. The mean is pulled further.

Moments of a distribution

Where do the names come from?

1. First moment: Mean

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

2. Second moment: Variance

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

3. Third moment: Skew

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

4. Fourth moment: Kurtosis

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

Problems

```
descriptives = describe(world)
descriptives = descriptives %>%
  mutate_if(is.numeric, round, 2)

kable(descriptives) %>%
  kable_styling(bootstrap_options = "striped", font_size = 14)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Country*	1	136	68.50	39.40	68.50	68.50	50.41	1.00	136.00	135.00	0.00	-1.23	3.38
Happiness	2	136	5.43	1.11	5.42	5.44	1.20	2.70	7.60	4.90	-0.07	-0.69	0.10
GDP	3	121	9.22	1.16	9.45	9.28	1.20	6.61	11.43	4.82	-0.43	-0.78	0.11
Support	4	135	0.80	0.12	0.83	0.81	0.12	0.43	0.99	0.55	-0.88	0.11	0.01
Life	5	135	63.12	7.46	64.64	63.73	7.35	43.74	76.04	32.30	-0.67	-0.34	0.64
Freedom	6	132	0.75	0.13	0.78	0.76	0.16	0.40	0.98	0.58	-0.45	-0.60	0.01
Generosity	7	120	0.00	0.16	-0.03	-0.01	0.15	-0.28	0.46	0.74	0.59	-0.27	0.01
Corruption	8	125	0.73	0.20	0.81	0.77	0.12	0.09	0.96	0.87	-1.48	1.53	0.02

How do we know if there is a problem and what should we

Problems

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Country*	1	136	68.50	39.40	68.50	68.50	50.41	1.00	136.00	135.00	0.00	-1.23	3.38
Happiness	2	136	5.43	1.11	5.42	5.44	1.20	2.70	7.60	4.90	-0.07	-0.69	0.10
GDP	3	121	9.22	1.16	9.45	9.28	1.20	6.61	11.43	4.82	-0.43	-0.78	0.11
Support	4	135	0.80	0.12	0.83	0.81	0.12	0.43	0.99	0.55	-0.88	0.11	0.01
Life	5	135	63.12	7.46	64.64	63.73	7.35	43.74	76.04	32.30	-0.67	-0.34	0.64
Freedom	6	132	0.75	0.13	0.78	0.76	0.16	0.40	0.98	0.58	-0.45	-0.60	0.01
Generosity	7	120	0.00	0.16	-0.03	-0.01	0.15	-0.28	0.46	0.74	0.59	-0.27	0.01
Corruption	8	125	0.73	0.20	0.81	0.77	0.12	0.09	0.96	0.87	-1.48	1.53	0.02

- mean and median are very different
- skew and kurtosis are large (|value| > 1)

There are several approaches that could be taken to detecting and dealing with non-normality:

- Overall tests of normality (e.g., Kolmogorov-Smirnov, Shapiro-Wilk tests)
  - Tests of extremity for a particular moment
  - $$SE_{skew} = \sqrt{\frac{6n(n-1)}{(n-1)(n+2)(n+3)}}$$
- Implication?
- Use procedures that are immune to the problem.

The mean is more affected than the median by extreme values. If the data are severely skewed or there are extreme outliers, inferential statistics might be affected. There are several remedies:

- Transform the data
- Exclude the outliers
- Use a trimmed mean (e.g., eliminate upper and lower 10%; “robust statistics”)
- Use the median (not susceptible to extreme values)

Pros & Cons to each of these; **be careful**

Bias and efficiency

Population versus sample

For those following along at home:

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

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

```
var(x)
```

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

Population versus sample

- The value that represents the entire population is called a **parameter**.
    - We collect samples to estimate the properties of populations; the statistic that represents a sample is called a **statistic**.
    - Population parameters are represented with Greek letters (μ , σ).
    - Sample statistics are represented with Latin letters (M , $\bar{X}$, s).

Bias and efficiency

- In deciding about different ways to estimate a parameter (e.g., central tendency), it is important to consider bias and efficiency (and sometimes consistency).
- **Bias**: An estimator is biased if its expected value and the true value of the parameter are different.
- **Efficiency**: Of two alternative estimators, the more efficient one will estimate the parameter with less error for the same sample size.

Bias and efficiency

Robust statistics sacrifice efficiency to control possible bias.

Variance (and standard deviation) are *biased* estimators when applied to samples.

- Using the formulas we've described, these statistics will *underestimate* variability in the population.

Population versus sample

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

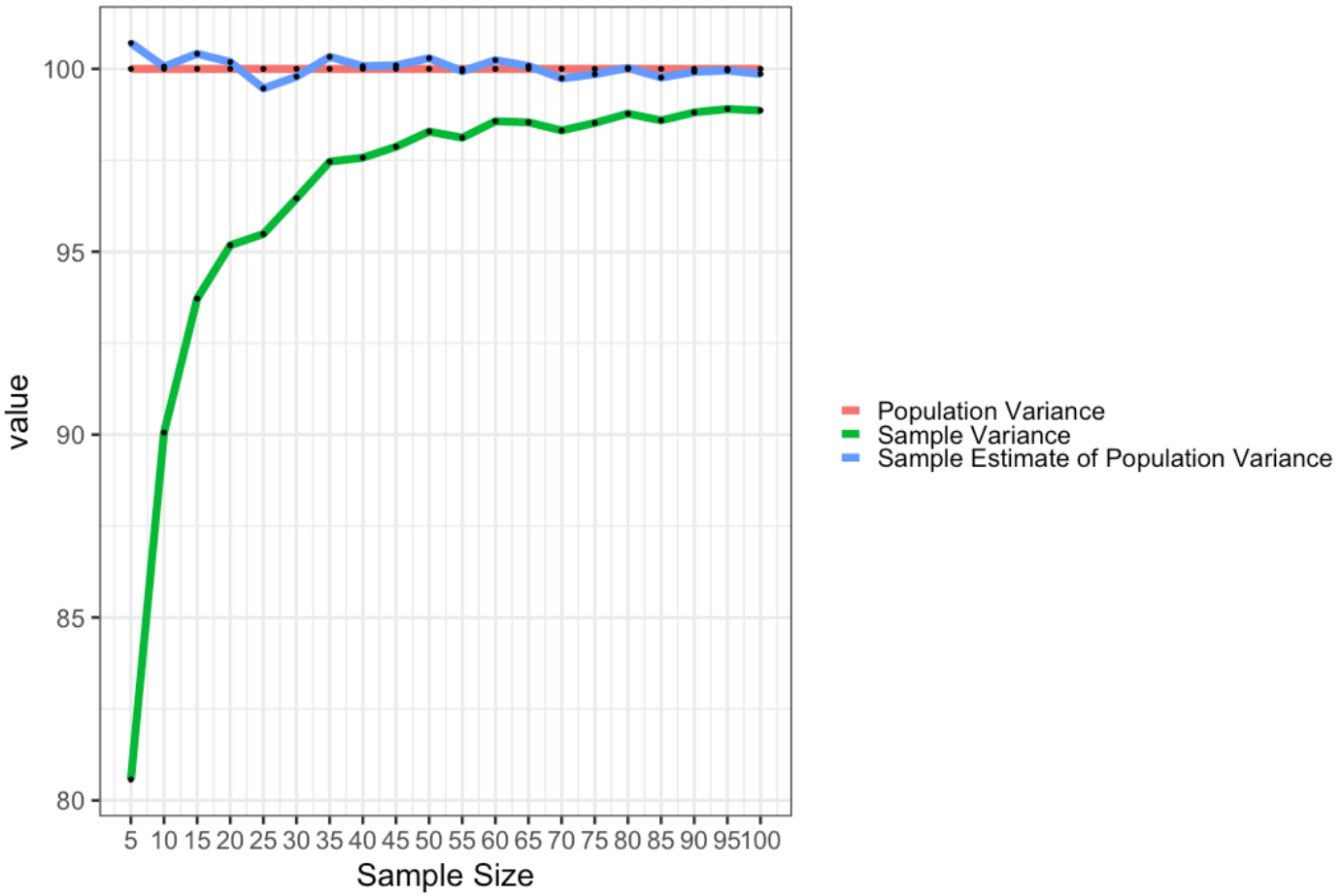
Population

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

Sample

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

Simulation


100
95
90
85
80
value
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
Sample Size
Population Variance
Sample Variance
Sample Estimate of Population Variance

Standardized scores

Standardized scores (z-scores)

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

Scores interpreted as distance from the mean, in standard deviations.

Properties of z-scores

- $\Sigma z = 0$
- $\Sigma z^2 = N$
- $s_z = \frac{\Sigma z^2}{n}$

Standardized scores (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 median   min   max range  skew kurtosis
## Country        1 136   NaN   NA     NA   Inf  -Inf  -Inf    NA       NA
## Happiness      2 136  5.43 1.11   5.42  2.70  7.60  4.90 -0.07    -0.69 0
## GDP            3 121  9.22 1.16   9.45  6.61 11.43  4.82 -0.43    -0.78 0
## Support        4 135  0.80 0.12   0.83  0.43  0.99  0.55 -0.88     0.11 0
## Life           5 135 63.12 7.46  64.64 43.74 76.04 32.30 -0.67    -0.34 0
## Freedom        6 132  0.75 0.13   0.78  0.40  0.98  0.58 -0.45    -0.60 0
## Generosity     7 120  0.00 0.16  -0.03 -0.28  0.46  0.74  0.59    -0.27 0
## Corruption     8 125  0.73 0.20   0.81  0.09  0.96  0.87 -1.48     1.53 0
```

Which variable has outliers?

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

```
##             vars   n mean sd median   min  max range  skew kurtosis   se
## Country        1 136  NaN NA     NA   Inf -Inf  -Inf    NA       NA   NA
## Happiness      2 136    0  1  -0.01 -2.46 1.96  4.42 -0.07    -0.69 0.09
## GDP            3 121    0  1   0.19 -2.26 1.91  4.17 -0.43    -0.78 0.09
## Support        4 135    0  1   0.20 -2.99 1.51  4.50 -0.88     0.11 0.09
## Life           5 135    0  1   0.20 -2.60 1.73  4.33 -0.67    -0.34 0.09
## Freedom        6 132    0  1   0.20 -2.64 1.71  4.34 -0.45    -0.60 0.09
## Generosity     7 120    0  1  -0.21 -1.80 2.90  4.70  0.59    -0.27 0.09
## Corruption     8 125    0  1   0.38 -3.20 1.14  4.34 -1.48     1.53 0.09
```

Next time...

describing relationships amongst 2 variables