PROBABILITY

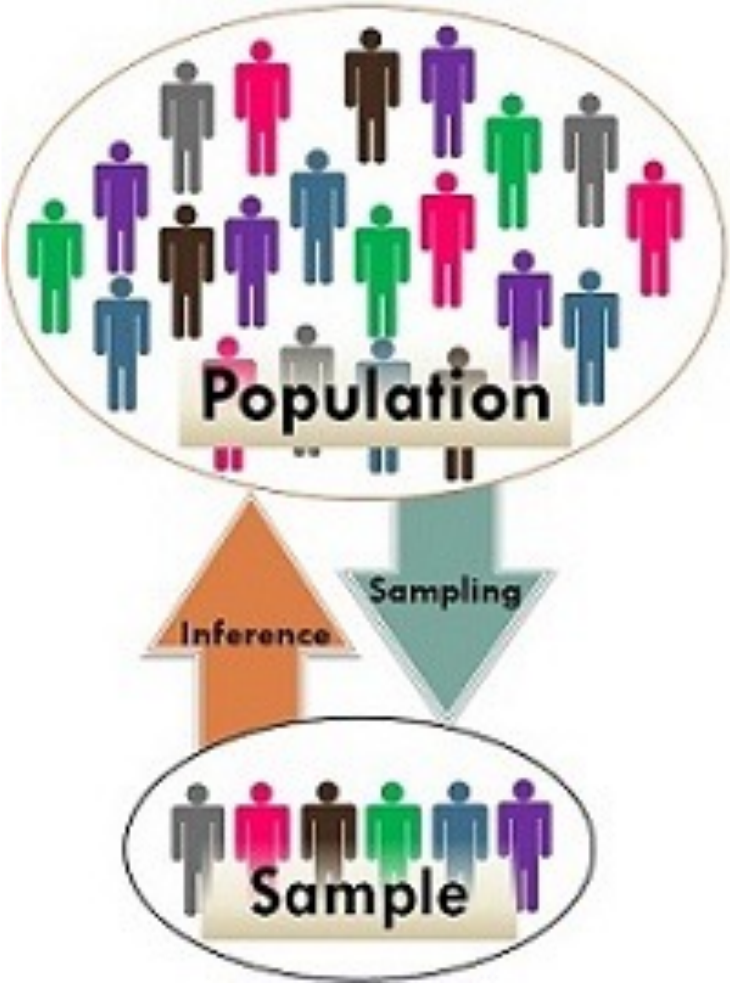
Prof. Wolf
POLS 095
Drake University

POPULATION AND SAMPLES

Population: data for every possible relevant case

Sample: a subset of cases

- Random sample: each member/case in the population has an equal chance of being in the sample
- Samples may also be non-random, which are samples of convenience



M&M'S

There are six colors of M&M's: blue, orange, green, yellow, red, and brown.

What proportion of the population of M&M's are blue?

- 91 are blue out of 302
  - Proportion of M&Ms that are blue = 0.30

I can use a sample of M&M's to make an inference about the proportion of blue M&M's in the population

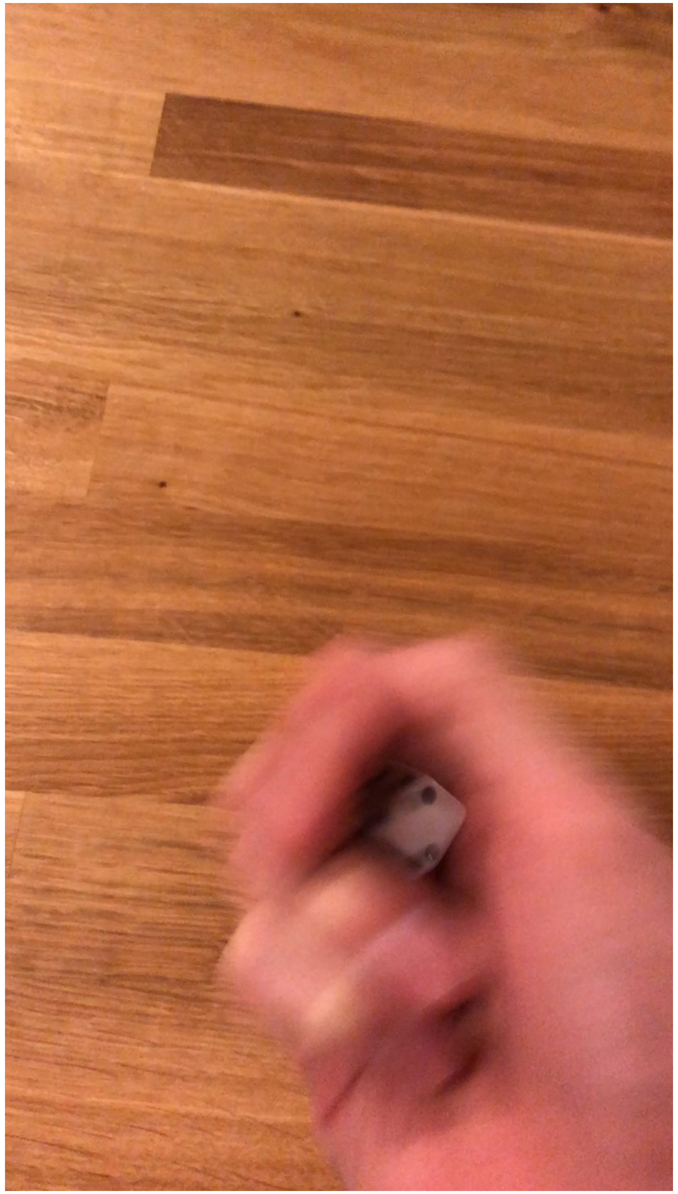
OUTCOMES

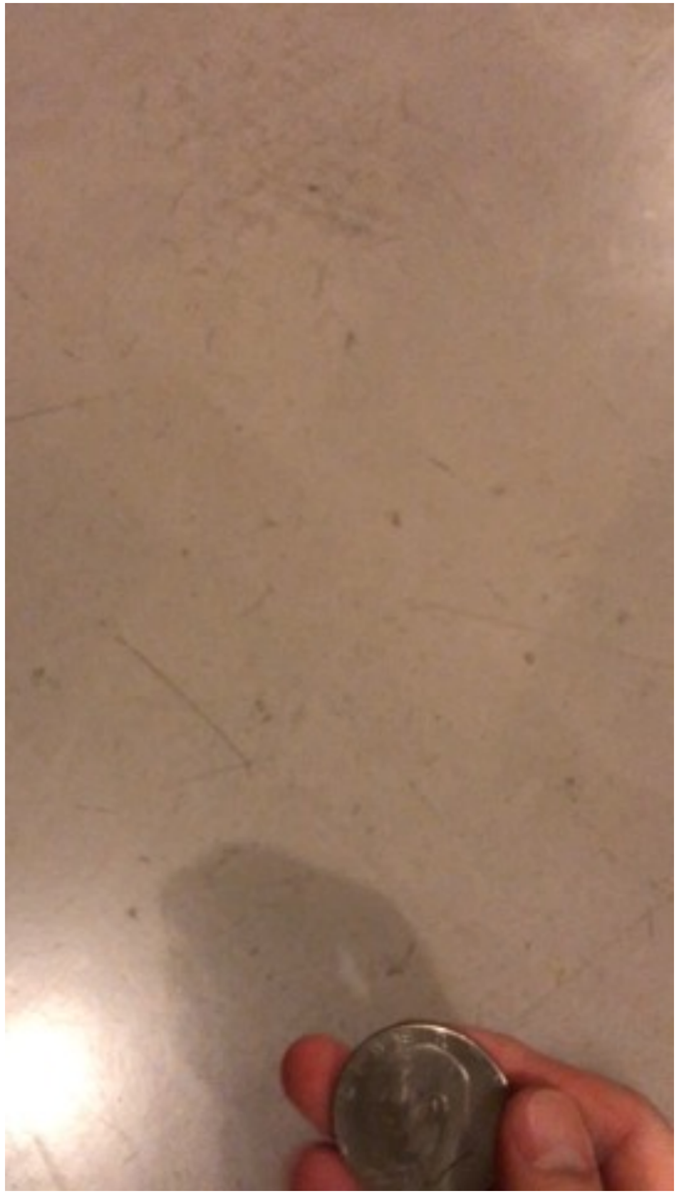
Outcome: result of a random observation

- Any M&M's color

Independent Outcome: if the realization of one outcome doesn't affect the realization of another outcome

- Dice roll or coin flip





KEY PROPERTIES OF PROBABILITY (1)

1. Probability ranges from 0 to 1.
   - 0 = impossible
   - 1 = absolutely certain to happen
   - Ex. Probability of rolling a 4 on a six sided die is 1/6 (0.167), probability of flipping a head on a fair, two sided coin is 1/2 (0.50)

KEY PROPERTIES OF PROBABILITY (2)

2. Sum of all possible outcomes totals to 1 *(and only 1)*
   - Coin flip: 1/2 it's heads, 1/2 it's tails.
     - 1/2 + 1/2 = 2/2 = 1
       - Or, 0.5 (heads) + 0.5 (tails) = 0.5 + 0.5 = 1
   - Rolling a 1, 2, 3, 4, 5, or 6:
     - 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 6/6 = 1

KEY PROPERTIES OF PROBABILITY (3)

3. If outcomes are independent, the probability of outcomes occurring is equal to the product of them occurring individually.
   - What are the chances of flipping three heads in a row?
     - HHH= ½ X ½ X ½ = 1/8
       - 0.5*0.5*0.5=0.125

KEY PROPERTIES OF PROBABILITY (4)

4. If outcomes are not independent, then any one outcome occurring is conditional on previous outcomes
   - Ex. Drawing a blue M&M
     - Probability of drawing a blue M&M = 91/302 = 0.3013
     - Draw 1 = blue M&M
     - Now there are only 90 blue M&Ms left out of a total of 301 M&Ms
     - Probability that I draw another blue M&M on the second draw = 90/301 = 0.2990

RELEVANCE FOR SOCIAL SCIENCE

Political scientists typically work with samples, not populations

The rules of probability are key to identifying which relationships are “statistically significant”

Main point: We use probability theory to decide whether the patterns of relationships we observe in a sample could have occurred by chance

Standard Error

Review: statistics and parameters

WE ARE INTERESTED IN PARAMETERS		WE USE STATISTICS TO ESTIMATE THEM	
Population mean	μ	Sample mean	$\bar{Y}$ or $\bar{X}$
Population variance	σ^2	Sample variance	s^2
Pop. st. deviation	σ	Sample st. dev.	s
Pop. Proportion	p	Sample proportion	$\hat{p}$

What is the Standard Error?

Standard Error: Standard deviation of the sampling distribution of a statistic

$$\frac{s}{\sqrt{n}}$$

s = standard deviation of the sample

n = number of observations of the sample

Today: Standard error of a mean

Future: Standard error of estimates in a regression

Calculate the standard deviation

Find the standard deviation: 3, 12, 10, 12, –4

1. Find the *mean (μ)*: $\bar{X} = \frac{3 + 12 + 10 + 12 + (-4)}{5} = \frac{33}{5} = 6.6$
2. Find the *deviations*, subtracting mean from each observed value:

3 – 6.6	12 – 6.6	10 – 6.6	12 – 6.6	–4 – 6.6
–3.6	5.4	3.4	5.4	–10.6

3. Find the *squared deviations*.

(–3.6) (–3.6)	(5.4)(5.4)	(3.4)(3.4)	(5.4)(5.4)	(–10.6)(–10.6)
12.96	29.16	11.56	29.16	112.36

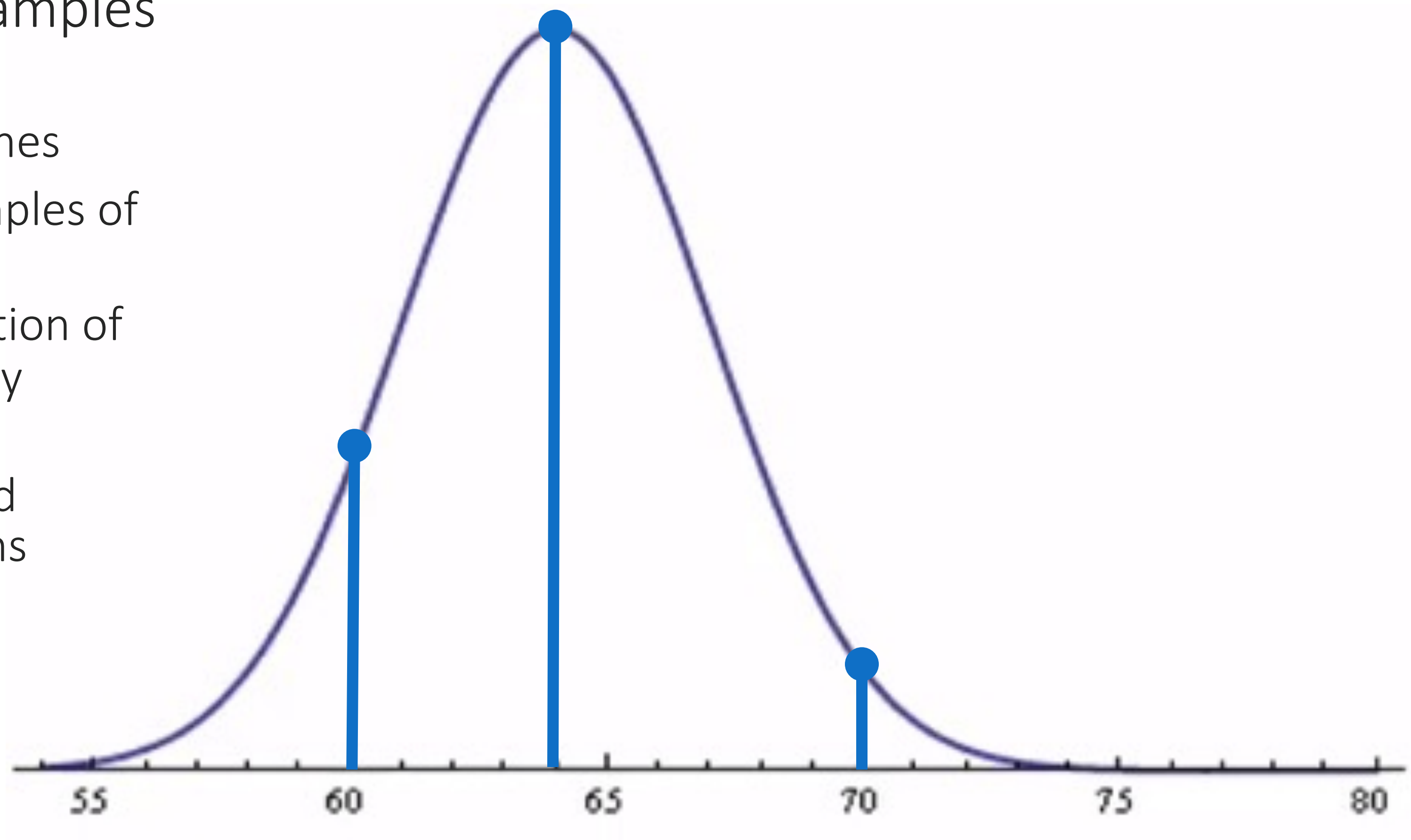
4. Sum the squared deviations: 195.2
5. Find the average (the *variance:* σ^2): 195.2/5 = 39.04
6. Take the square root: $Standard\ deviation = \sigma = \sqrt{39}.04 = {\sim}6.25$ How would we find the standard error?

$Standard\ error = \frac{\sigma}{\sqrt{n}}$, so $\frac{6.25}{\sqrt{5}} = {\sim}2.8$

Standard Error Example

Take an infinite number of samples of American women

- Population mean height: 64 inches
- Take an infinites number of samples of American women
- Central Limit Theorem: distribution of sample mean heights is normally distributed
- The sample error is the standard deviation of these sample means



Standard Error Example

- Sample 100 American women
  - Sample mean: 65 inches
  - Sample standard deviation: 6 inches
  - Standard error: $\frac{s}{\sqrt{n}} = \frac{6}{\sqrt{100}} = \frac{6}{10}$=0.6

Confidence Intervals

Probability, Population Parameters, and Sample Statistics

Central Limit Theorem enables us to make inferences about population parameters from a sample

But our sample might differ from the population due to chance (sampling error)

- Our sample statistics may not be equal to the underlying population parameter.
- The *confidence interval* establishes the likely range for a population parameter.

What is a Confidence Interval?

Confidence Interval: Probability of the likely value of a population parameter

- Confidence interval for the mean:
  - *What is the average height of American women?*
  - *95% Confidence interval: 65 inches, ± 1.2 inches*
- Confidence interval for a proportion (sometimes called "Margin of Error"):
  - *What percentage of the American electorate approves the of the president's job performance?*
  - *95% Confidence: 44% with a margin of error of ± 2.5%*

How to Calculate the Confidence Interval for a Sample Mean

1. Decide desired level of confidence: the standard is 95%
2. Calculate the sample mean: $\overline{X}$
3. Calculate standard deviation (σ) of the sample
4. Calculate the standard error: $\frac{s}{\sqrt{n}}$
5. Calculate the confidence interval:
   - 95% level of confidence:
     - $\overline{X} \pm 2 * \frac{s}{\sqrt{n}}$

Note: technically we use 1.96 instead of 2 to calculate a 95% confidence interval, but you can follow the "rule of 2."

Confidence Interval Example

Poll of 100 Drake alumni: How much did you donate to the university in 2020?

$\bar{X}$ = \$700

s = \$600

n = 100

What is the mean donation of *all* Drake alumni?

Confidence interval at 95% level: $\bar{X} \pm 2 * \frac{s}{\sqrt{n}}$

$\bar{X}$ = \$700, $\frac{s}{\sqrt{n}} = \frac{600}{\sqrt{100}} = \frac{600}{10} = 60$

95% Confidence interval = 700 $\pm$ 2*60 = 700 $\pm$ 120 (580, 820)

With 95% certainty, we can say that the mean donation of all Drake alumni lies between \$580 and \$820

Margin of Error Example

Poll of 200 prospective Iowa voters in the 2022 gubernatorial election:

Kim Reynolds: 55%, Democratic Candidate 45%, Margin of error: 7%

*Can we say that Reynolds is really ahead of a prospective (generic) Democrat?

Reynolds support = 0.55 $\pm$ 0.07 = between 0.48 and 0.62, 48% and 62% of Iowans

Democrat support = 0.45 $\pm$ 0.07 = between 0.38 and 0.52, 38% and 52% of Iowans

If the intervals overlap then we are unsure that Reynolds is really ahead!

How to Tighten Confidence Interval?

Tightening confidence interval = reducing standard error

$$\frac{s}{\sqrt{n}}$$

Two ways that the standard error might be reduced:

1. Reduce size of the numerator: decrease standard deviation
   - *Not under control of the researcher!*
2. Increase the size of the denominator: increase sample size
   - *Under control of researcher*
   - *But can be costly with diminishing returns – decide on "good enough" standard error*
   - *To cut the sampling error in half (e.g. cut margin of error of poll from 2pts to 1pt), what needs to happen to the sample size? Sample size needs to be quadrupled! Rule of thumb!*

Table 6-5 Terms and Symbols and the Roles They Play in Inference

Term or symbol (pronunciation)	What it is or what it does	What role it plays in sampling and inference
μ ("mew")	Population mean	Usually μ is unknown and is estimated by $\bar{x}$.
N	Population size	
σ ("sigma")	Population standard deviation	Measures variation in a population characteristic. The variation component of random sampling error.
Z score	Converts raw deviations from μ into standard units	Defines the tick marks of the normal distribution; 68 percent of the distribution lies between $Z = -1$ and $Z = +1$; 95 percent of the distribution lies between $Z = -1.96$ and $Z = +1.96$.
$\bar{x}$ ("x bar")	Sample mean	Sample statistic that estimates μ.
n	Sample size	The sample size component of random sampling error is $\sqrt{n}$.
s	Sample standard deviation	Substitutes for σ as the variation component of random sampling error when σ is unknown.
Standard error of the sample mean	Measures how much $\bar{x}$ departs, by chance, from μ	Random sampling error. Equal to $\sigma/\sqrt{n}$, if σ is known. Equal to $s/\sqrt{n}$, if σ is unknown.
95 percent confidence interval	The interval in which 95 percent of all possible values of $\bar{x}$ will fall by chance	Defined by $\bar{x} \pm 1.96$ standard errors in normal estimation. Can usually be determined by rule of thumb: $\bar{x} \pm 2$ standard errors in all estimation.
p	Proportion of a sample falling into one value of a nominal or ordinal variable	Sample estimate of a population proportion.
q	Proportion of a sample falling into all other values of a nominal or ordinal variable	Equal to $1 - p$.
Standard error of a sample proportion	Measures how much p departs, by chance, from a population proportion	Defined by $\sqrt{pq}\,/\sqrt{n}$. Ordinarily can be applied in finding the 95 percent confidence interval of p, using normal estimation.