Professor Gia Barboza-Salerno

College of Social Work

The Ohio State University

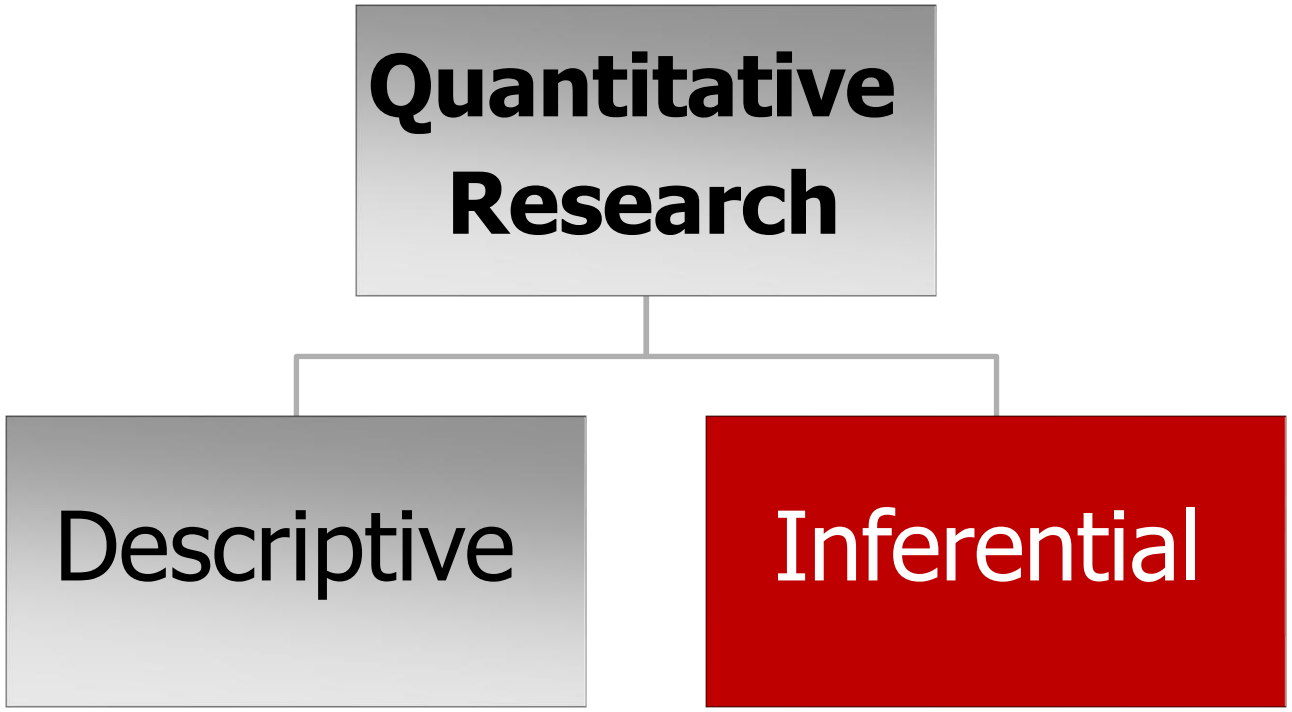
Week 4: Intro to Probability Theory and Sampling Distributions

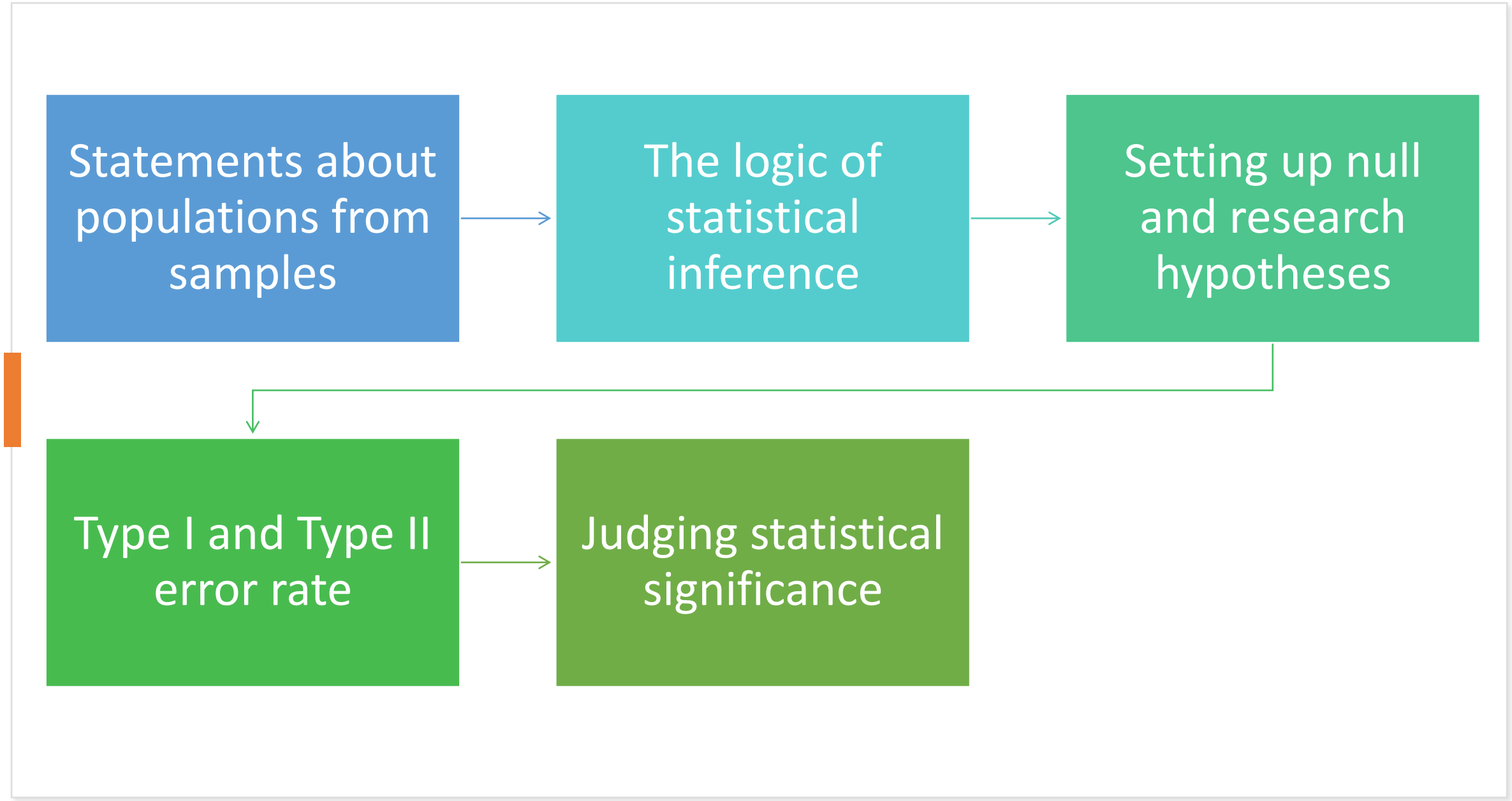
September 12, 2023

Overview

1. Inferential Statistics
2. Sampling Distributions: What are they?
3. Types of errors
4. Intro to probability theory
5. Events and Operations
6. The Binomial Distribution
7. Applications

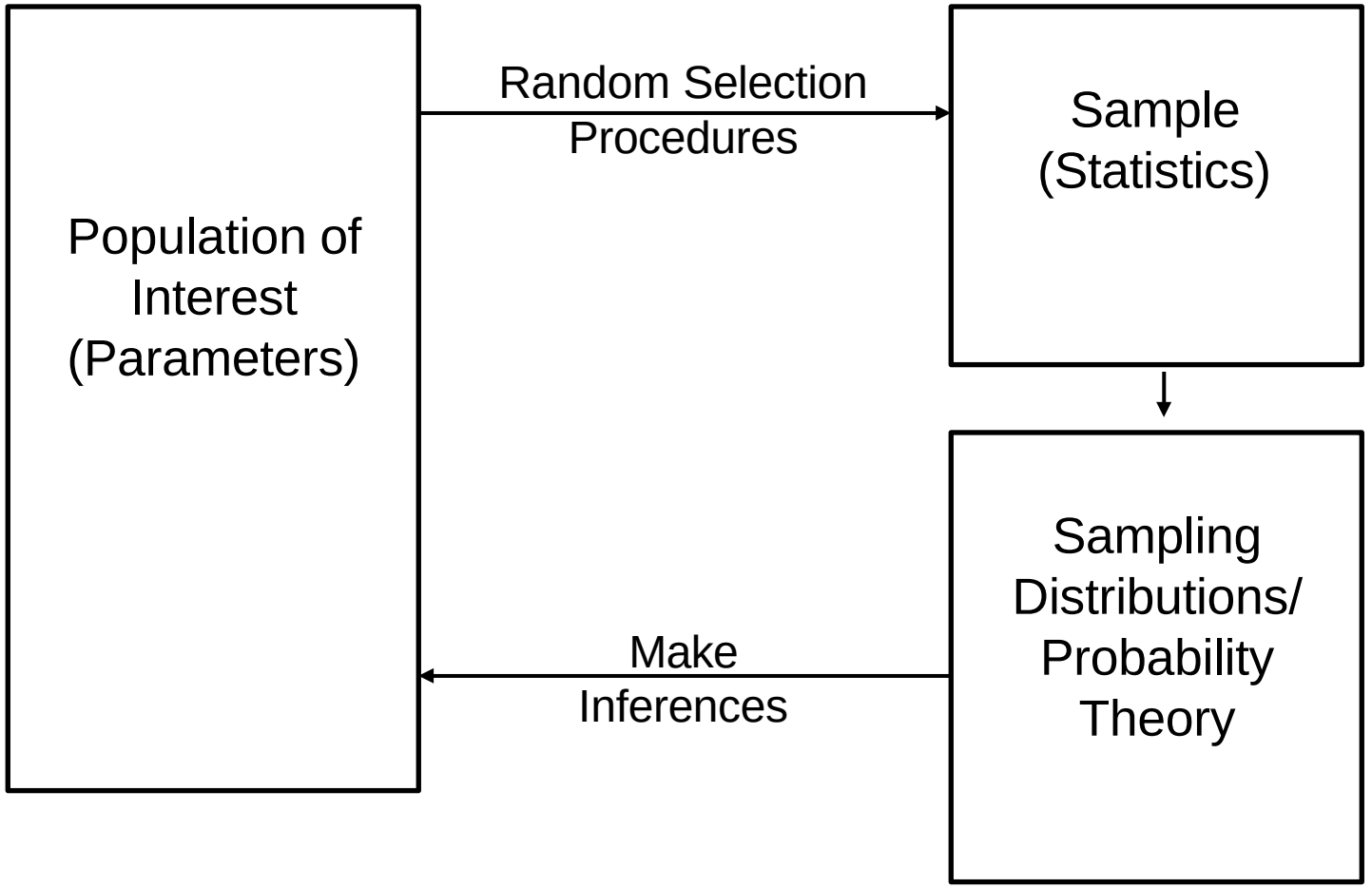
Inferential Statistics





Inferential Statistics: Overview

Inferential Statistics



Inferential Statistics

Random Sampling

EPSM

Purpose

Random selection

- Each element (member, person, or unit of analysis) of the population has an *equal and known probability of selection;* and
- All sample elements are selected *independently* of one another

Independent random samples

An independent random sample is a sequence of observations which are not dependent on any other sample or data.
The requirements for ***independent random samples*** are mentioned below:

1. Each member of the population must have an equal chance of selection
2. The members are selected randomly instead of voluntarily selecting themselves

If the data values of one random sample do not affect the data values of the other random sample, then the samples are said to be independent random samples

Examples: samples_handout.pdf

Dilemma facing researchers

Problem: we only have distributions from samples, how are we able to infer something about the population?

In statistical terms: use the distribution of sample scores to make statements about characteristics of the population

Inferential Statistics

- **The Process**
  - Once our sample is selected and our data of interest is collected, the inferential process begins.
  - We use sample statistics (e.g., means, variances, regression coefficients, etc.) to make **"estimates"** about corresponding population parameters.
  - It is critical to understand that our estimates are just "guesses."
- We use theoretical sampling distributions to manage and understand our guesses.

Distributions

Sampling Distributions

- The logic of hypothesis testing depends on something called a sampling distributions
  - We actually know something about theoretical sampling distributions – the normal curve is the most well-known of such distributions.
  - Other distributions we use in inferential work:
    - The binomial distribution
    - The *t* distribution
    - The chi-square distribution
    - The *F* distribution
  - There is a common logic to how these distributions help us understand our inferences.

Distributions

- **Sampling Distributions (cont.)**
  - A sampling distribution is a "frequency distribution" of a sample statistic (e.g., mean, variance/standard deviation)
  - Theoretically, a sampling distribution represents values of the statistic that would be generated from an large number of samples of the same sample size from a given population
  - A sampling distribution is the distribution of a statistic derived from ***multiple*** samples

Distributions

- **Sampling Distributions (cont.)**
  - Many sampling distributions of various statistics are based on and use the properties of theoretical distributions. For instance,
    - If the distribution of the population is normal, then the distribution of the sample mean is also normal.
    - If n is large, then the distribution of the sample means will be approximately normally distributed (Central Limit Theorem)

Parameters and sample statistics

Statisticians use different symbols to distinguish statistics on a population from statistics on a sample

	Mean	Variance	Standard Deviation
Sample distribution	$\bar{X}$	s^2	s
Population distribution	μ	σ^2	σ

A *point estimator* for a population parameter is a statistic that is used to estimate the parameter.

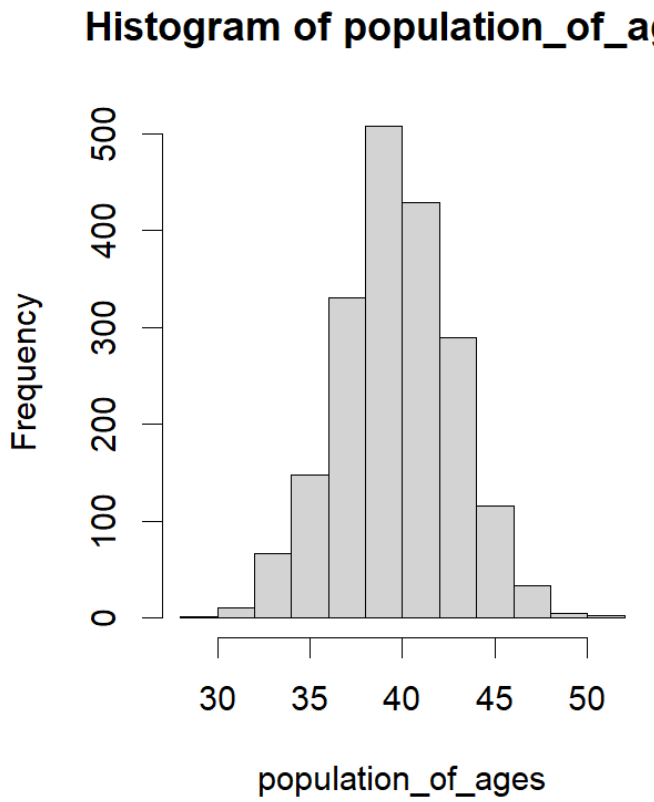
Let's see how this works: Run sampling_distributions.Rmd

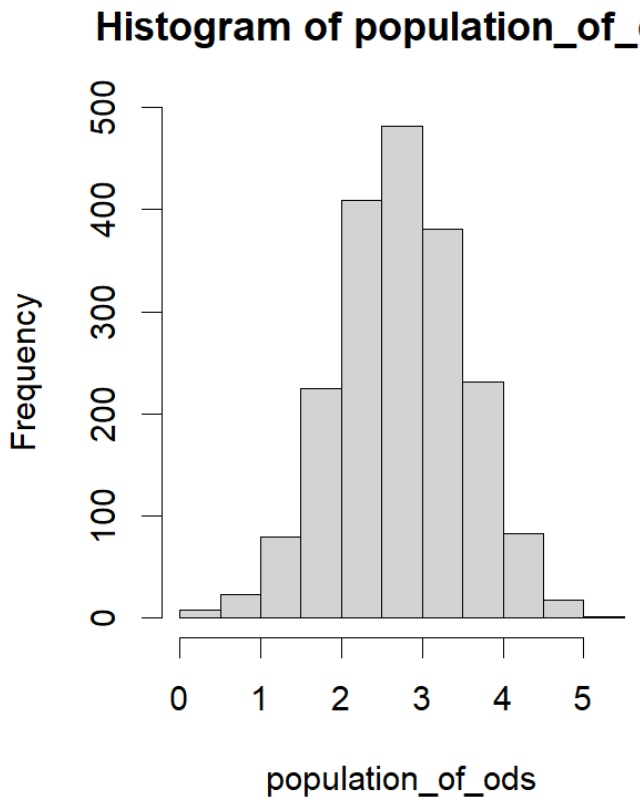
Example: True mean age of first overdose for the population of all persons in Ohio (this data is simulated)

Draw 10 samples of 100 people who overdosed and compute their age

Result is different each time; population parameters remains the same

This reinforces why measures of dispersion are so important





Code walkthrough

```
population_of_ages <- rnorm(1940, mean=39.7, sd=3)
population_of_ods <- rnorm(1940, mean=2.72, sd=.75)

mean(population_of_ages)    39.65878
sd(population_of_ages)      3.113047

mean(population_of_ods)     2.734343
sd(population_of_ods)       0.7820914
```

Simulation Results

1 Sample

Simulation Results

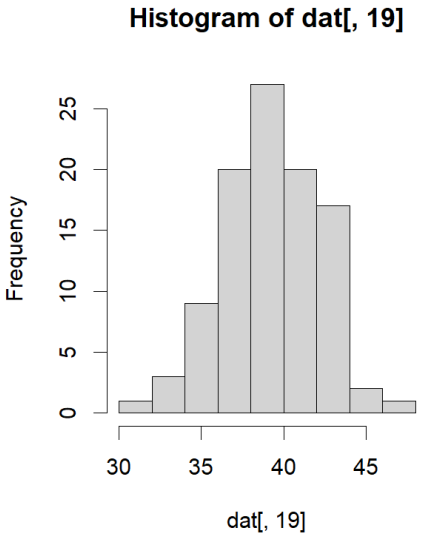
OD_pop_mean	OD_sample_mean	OD_pop_sd	OD_sample_sd
2.73	2.82	0.78	0.74

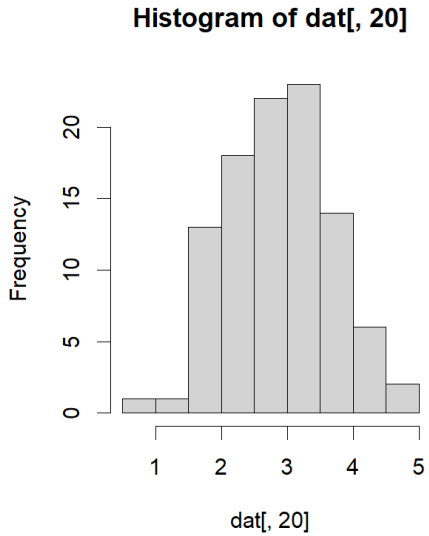
100 Samples

Simulation Results

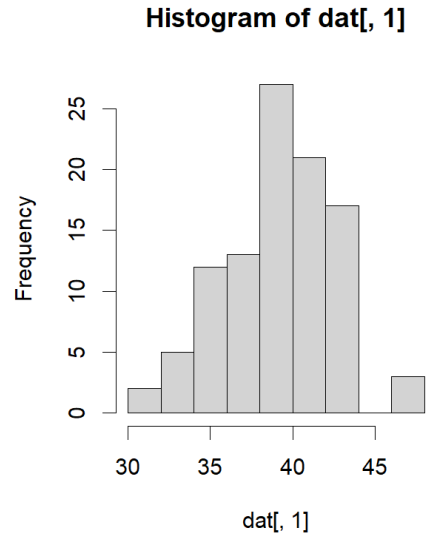
OD_pop_mean	OD_sample_mean	OD_pop_sd	OD_sample_sd
2.73	2.71	0.78	0.76

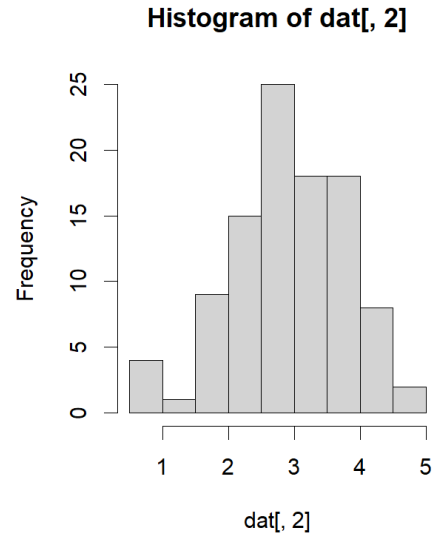
Sample 1

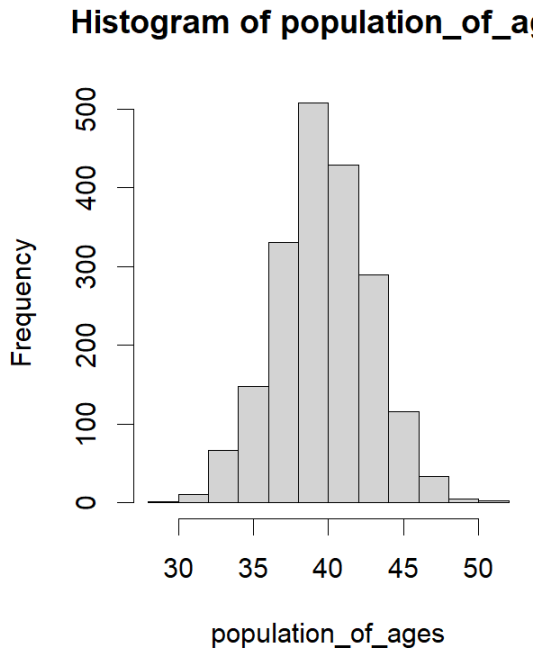


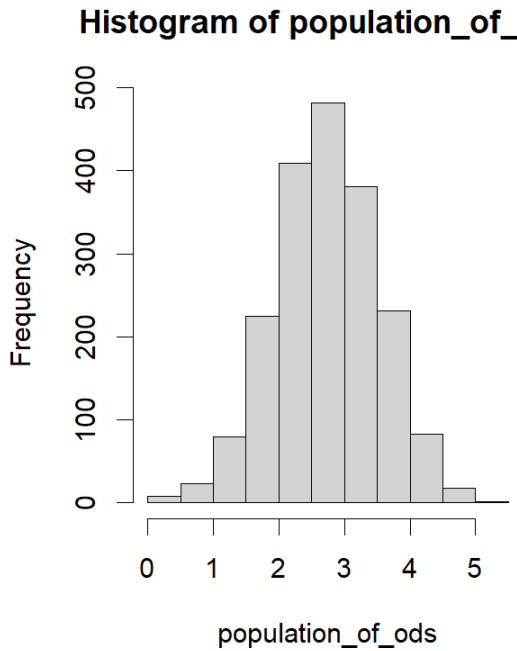


Sample 10









Your turn

Modify the code to take 10 samples of 10

- Note differences

Modify the code to take 1000 samples of 10

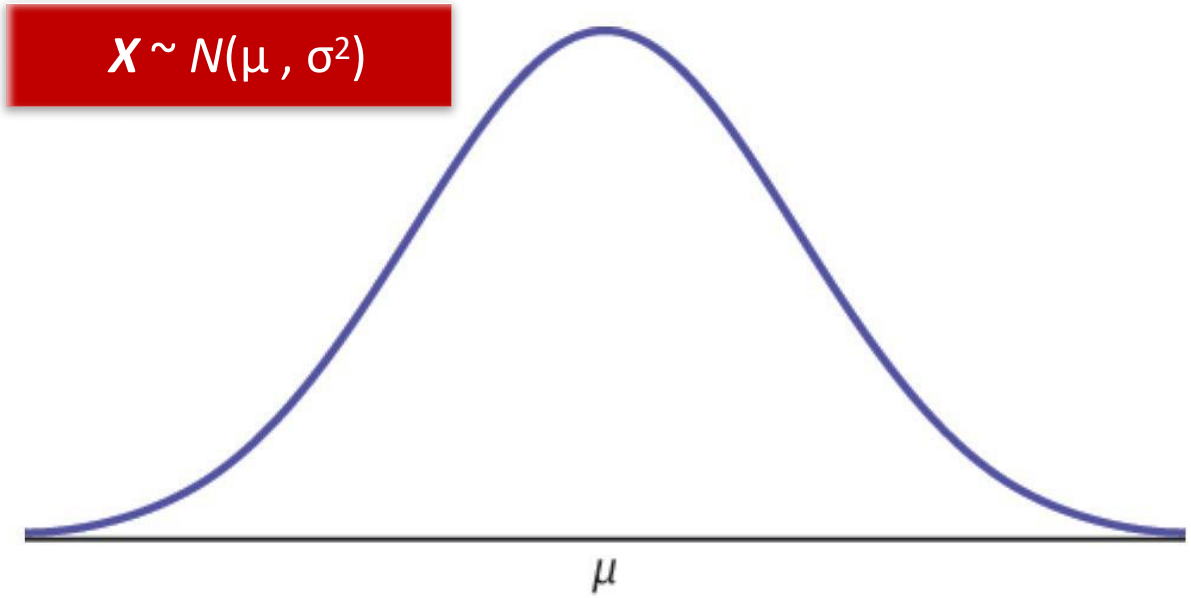
- Note differences

Important Distributions

Binomial
Normal
t distribution
Chi-square distribution
F distribution

The Normal Distribution

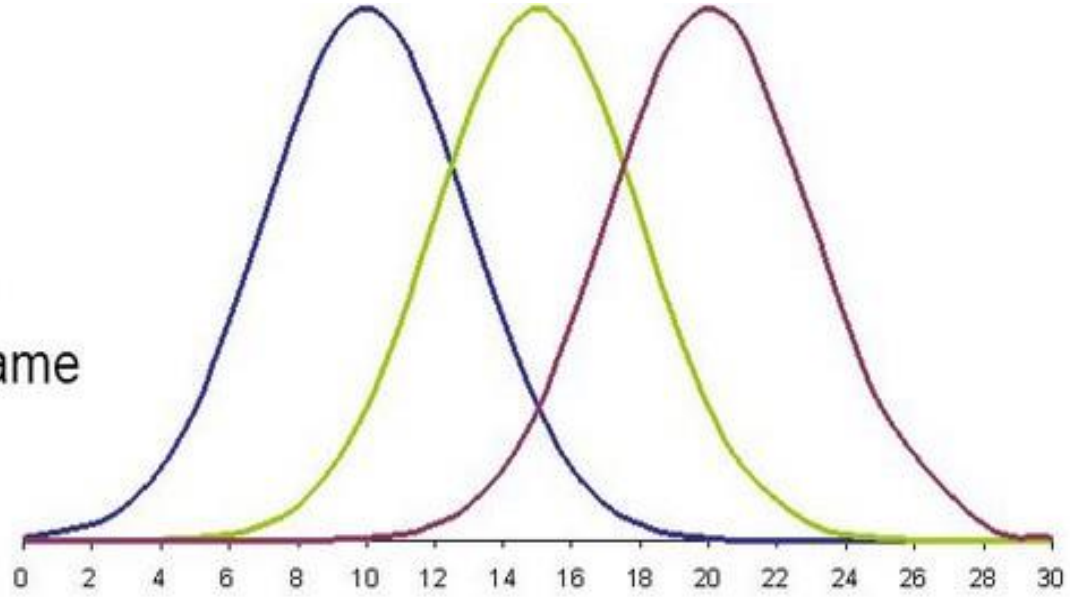
- A symmetric, bell-shaped curve
- Used to describe and understand continuous variables



Distributions

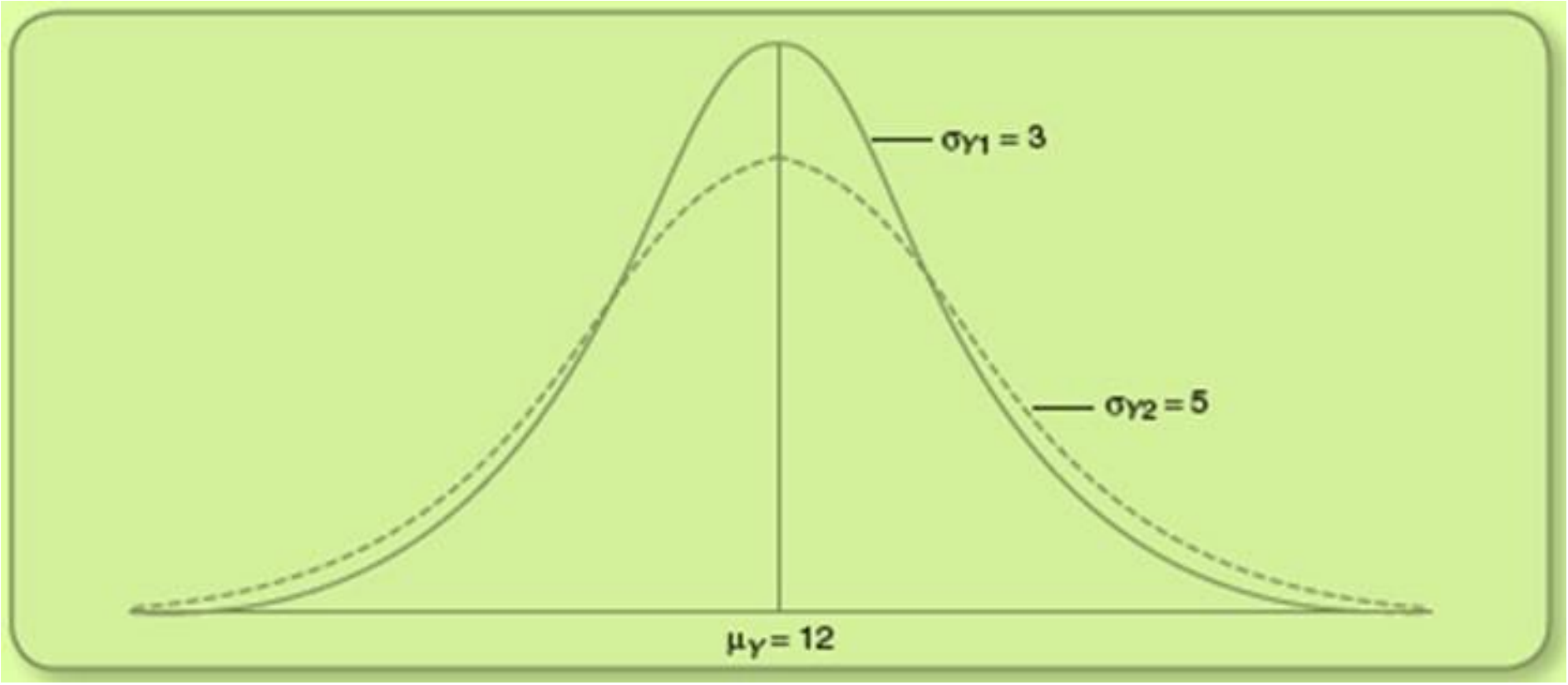
- The normal curve shifts to the left (for a smaller mean) or right (for a larger mean) depending on the mean values.

Here the means are different (μ = 10, 15, and 20) while the standard deviations are the same (σ = 3).



Distributions

The larger the standard deviation is, the wider and flatter the curve is, why?



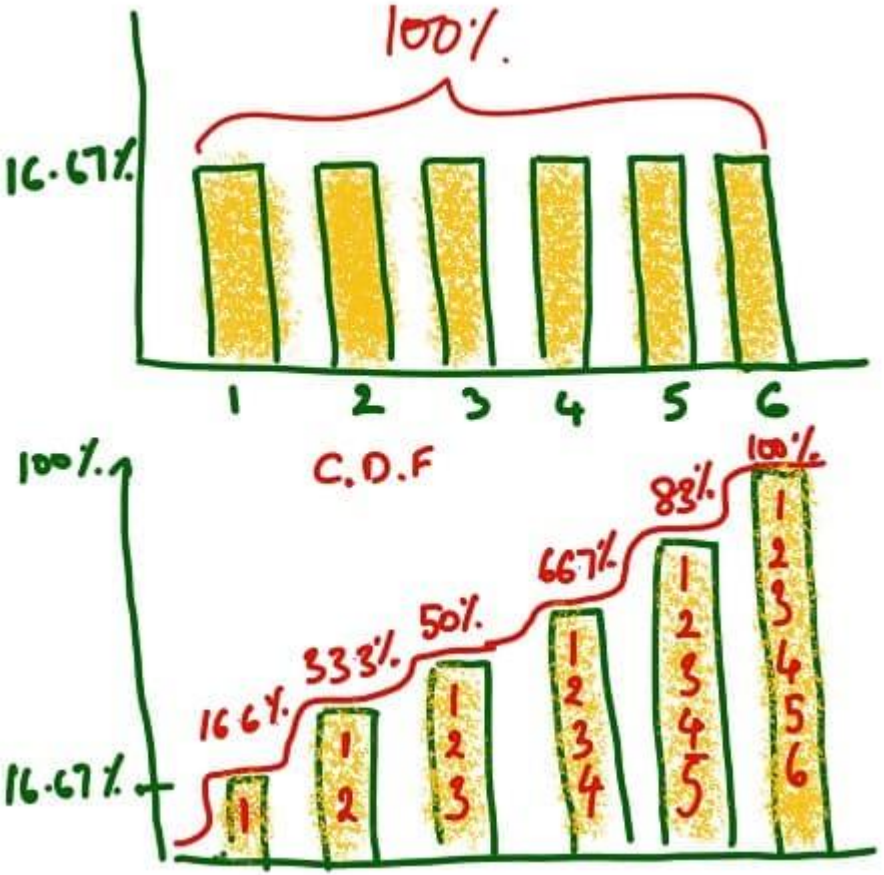
Conceptual Note: PDF v CDF

The **probability density function** (PDF) is the probability that a random variable, say *X*, will take a value **exactly equal** to *x:* $P(X = x)$

- **Example**: The probability of getting a 3 in one toss of a fair die → $P(X = 3)$=1/6 or .1667?

The **cumulative distribution function** (CDF) is the probability that a random variable, say *X*, will take a value equal to or less than *x:* $P(X \leq x)$

- **Example**: The probability of getting a 3 in one toss of a fair die → $P(X \leq 3)$ = .50?



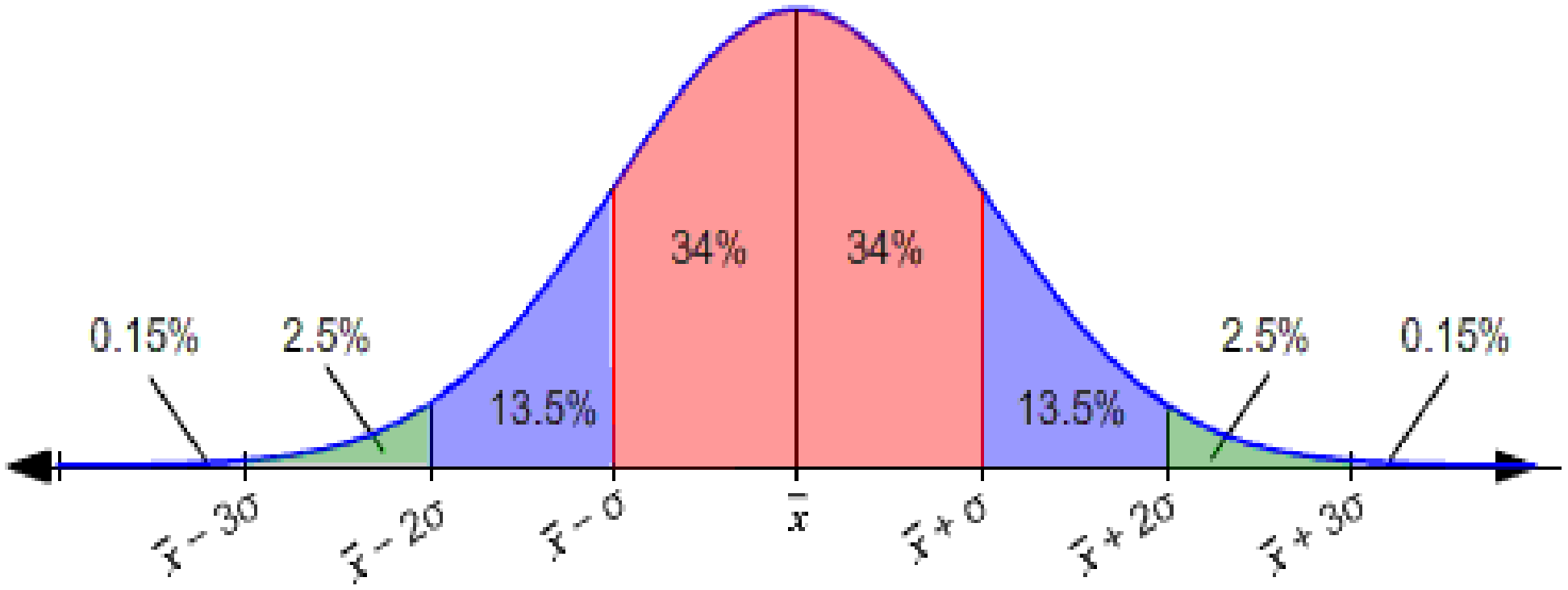
- **Normal Curve Properties**
  - It is a probability distribution with the density function:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

  - The probability is the area under the curve (need calculus)
  - The area under the curve must equal 1.00.
  - Standard deviation units prescribe known areas under the normal curve.
    - ± 1 SD: 68% of the values in the distribution
    - ± 2 SD: 95% of the values in the distribution
    - ± 3 SD: 99% of the values in the distribution

Distributions

Areas Under The Normal Curve: The Empirical Rule



Distributions

The Empirical Rule is just an approximation and only works for certain values. What if you want to find the probability for *x* values that are not integer multiples of the standard deviation?

- **The Standard Normal Distribution**
  - The score units of the standard normal distribution are called ***z-scores*** and are in **standard deviation units**.

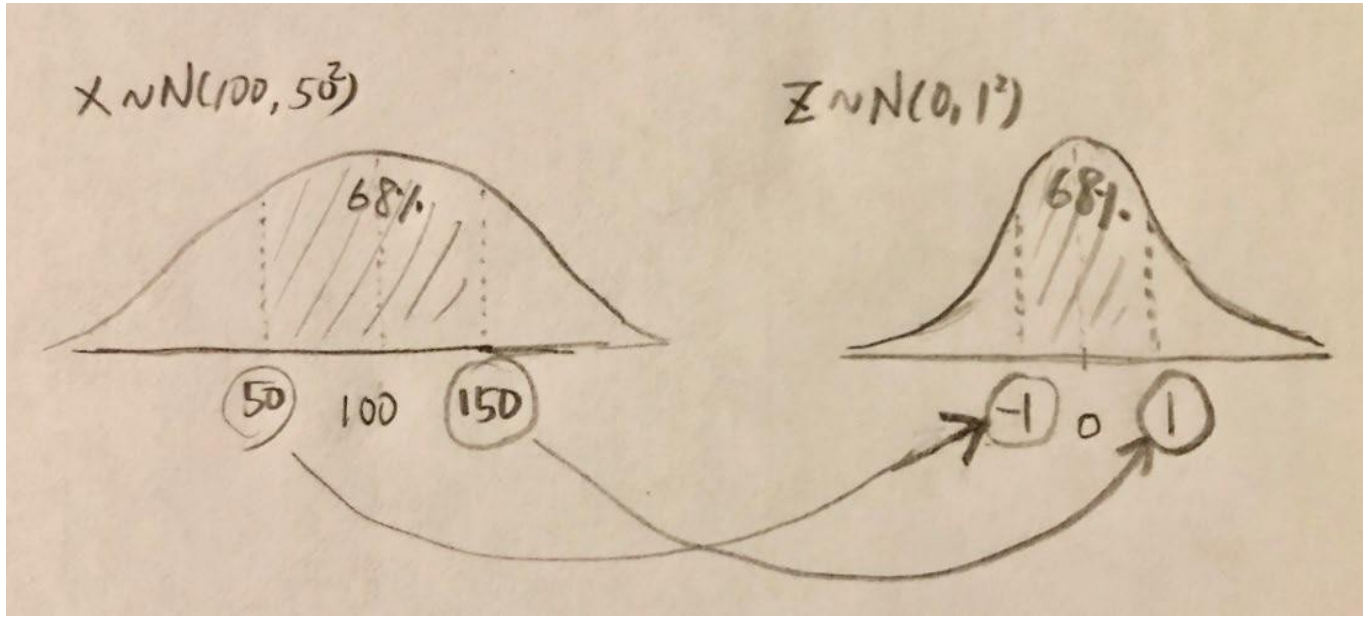
For X ~ $N(\mu, \sigma^2)$

$$z = \frac{x - \mu}{\sigma}$$

Standard deviations from population mean

Population Standard deviation

Population mean



Distributions

- Interpretations: The z-score tells you *how many standard deviations* the value x is *above or below the mean*.
  - The sign
    - Positive z scores: the value is above the mean
    - Negative z scores: the value is below the mean
  - The magnitude: # of standard deviations from the mean
    - e.g.) z=2.5 means that x is 2.5 standard deviations above the mean

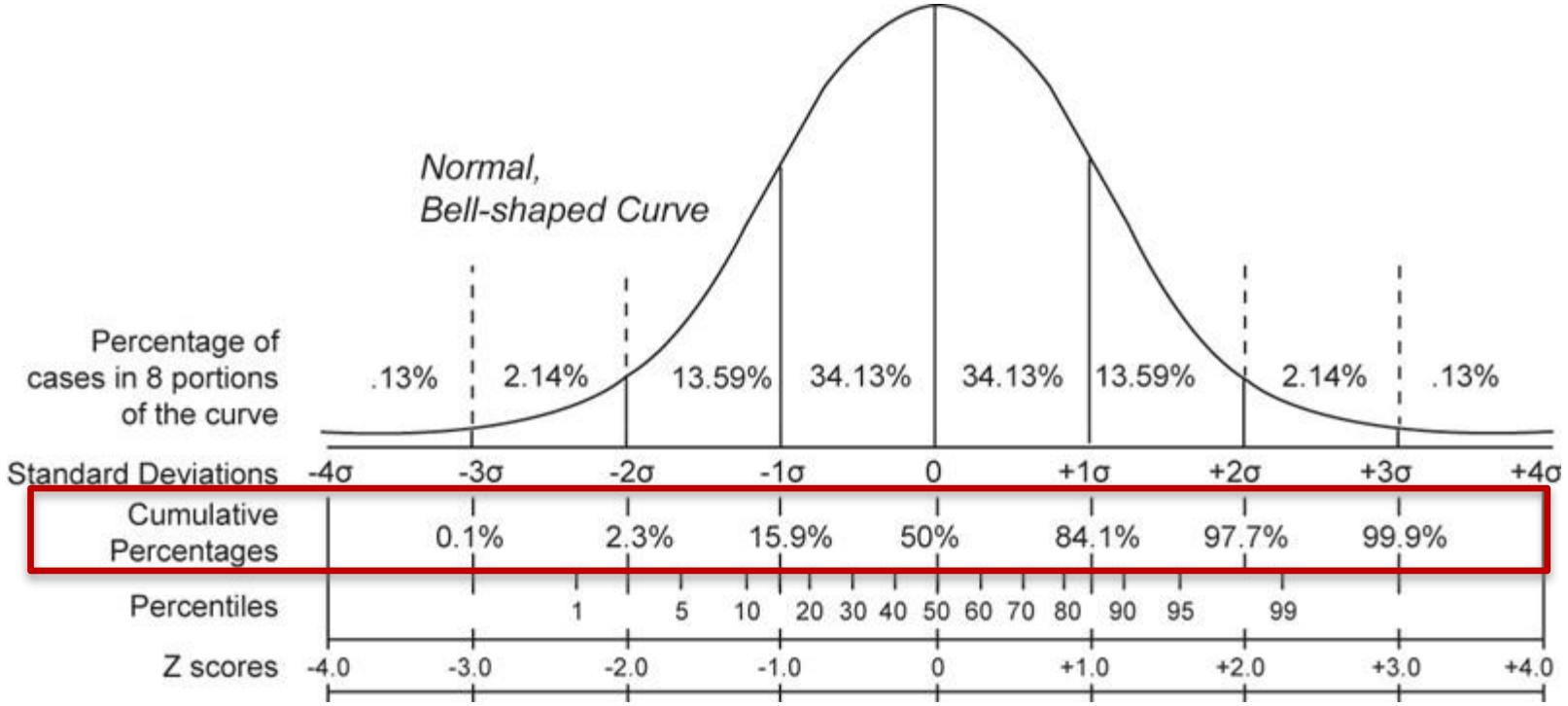
- **Example**: For a normally distributed random variable, $X \sim (5, 6^2)$, what is the z score for $x = 17$?

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

Old school: look up value in "back of the book"
New school: use R to compute

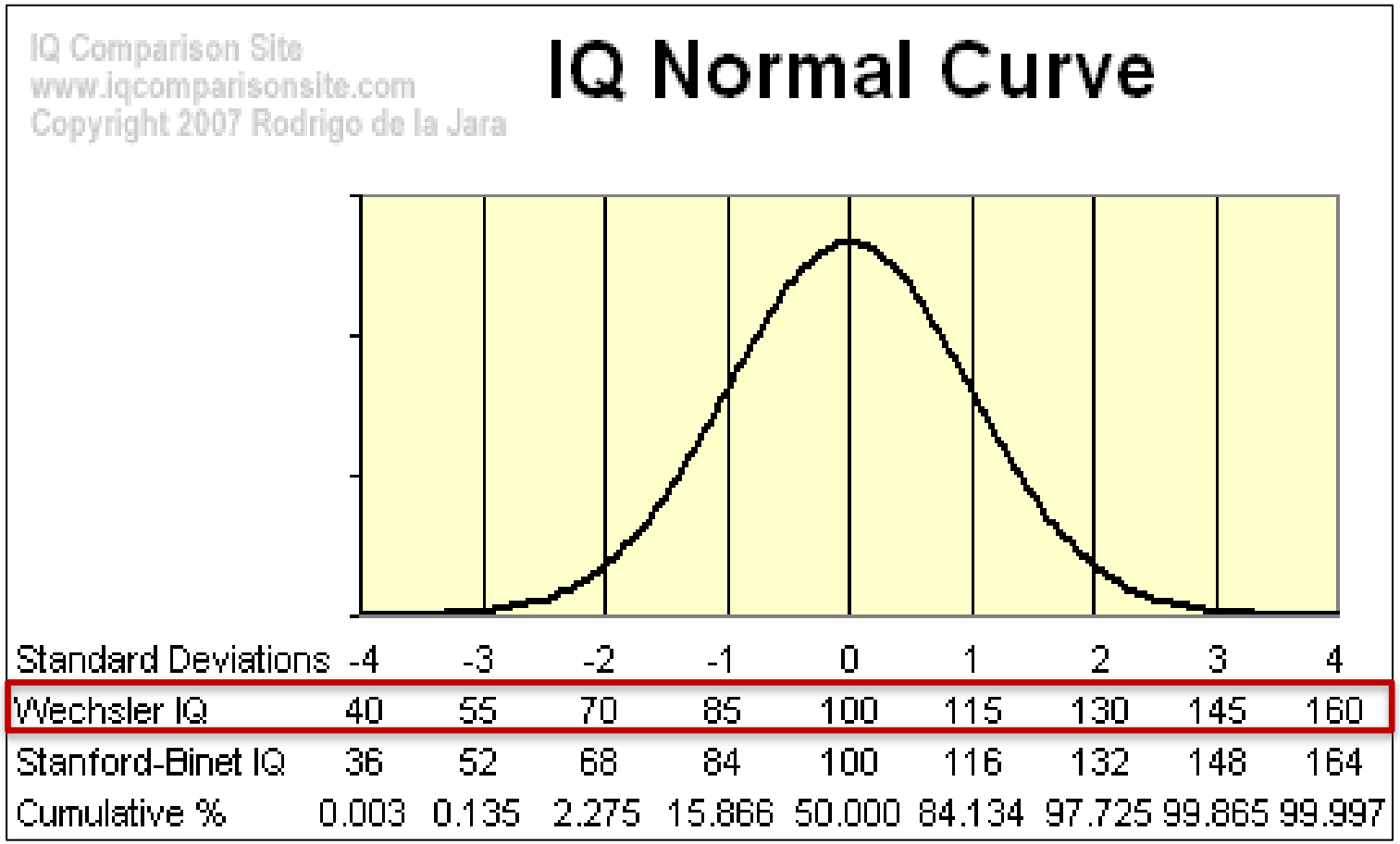
Distributions

- In practice, we can use z-scores for
  - Determining percentile ranks of individual scores
  - Comparing scores from different distributions



Distributions

- Determining percentile ranks of individual scores (cont.)

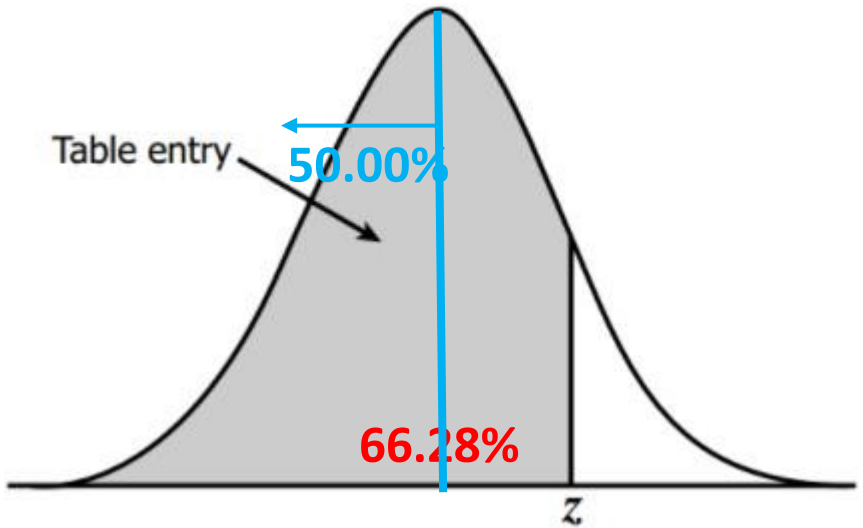


Standard Deviations	-4	-3	-2	-1	0	1	2	3	4
Wechsler IQ	40	55	70	85	100	115	130	145	160
Stanford-Binet IQ	36	52	68	84	100	116	132	148	164
Cumulative %	0.003	0.135	2.275	15.866	50.000	84.134	97.725	99.865	99.997

Distributions

- **Example**: What is the percentile of $z = 0$ in the standard normal distribution?
- **Example**: What is the percentile of $z = 0.42$ in the standard normal distribution?

z	.00	.01	.02	.03
0.0	.5000	.5040	.5080	.5120
0.1	.5398	.5438	.5478	.5517
0.2	.5793	.5832	.5871	.5910
0.3	.6179	.6217	.6255	.6293
0.4	.6554	.6591	.6628	.6664
0.5	.6915	.6950	.6985	.7019



To find the *z*-score associated with a *p*-value in R, we can use the qnorm() function, which uses the following syntax:

```
qnorm(p, mean = 0, sd = 1,
lower.tail = TRUE, log.p = FALSE)
```

qnorm(.5) retuns 0 (why)
qnorm(0.6628) returns .420

To find the *p*-value associated with a *z*-score in R, we can use the pnorm() function, which uses the following syntax:

```
pnorm(q, mean = 0, sd = 1,
lower.tail = TRUE, log.p = FALSE)
```

pnorm(0) retuns ??
pnorm(.420) returns ??

CAUTION

Draw the distribution lest you make a mistake
The area by default is the left side of the distribution
That means, the probability that is being returned is the area covering everything to the left of the distribution

Distributions

- Comparing scores from different distributions (cont.)
  - Suppose we are interested in comparing a high school student's relative position on four needs assessment scales–school connection, stress, support, and efficacy.
  - First we transform each scale to a z-score scale and compare the student's z scores for the four assessments. For instance, the student might have
    - school connection z-score = -.89
    - Stress z-score = .01
    - Support z-score = .09
    - Efficacy z-score = .21

Central Limit Theorem

Central Limit Theorem

- **Central Limit Theorem (for sample means)**
  - If you draw random samples of size *n*, then as *n* increases, the random variable ($\bar{X}$) (i.e., sample mean) tends to be **normally distributed** as follows:

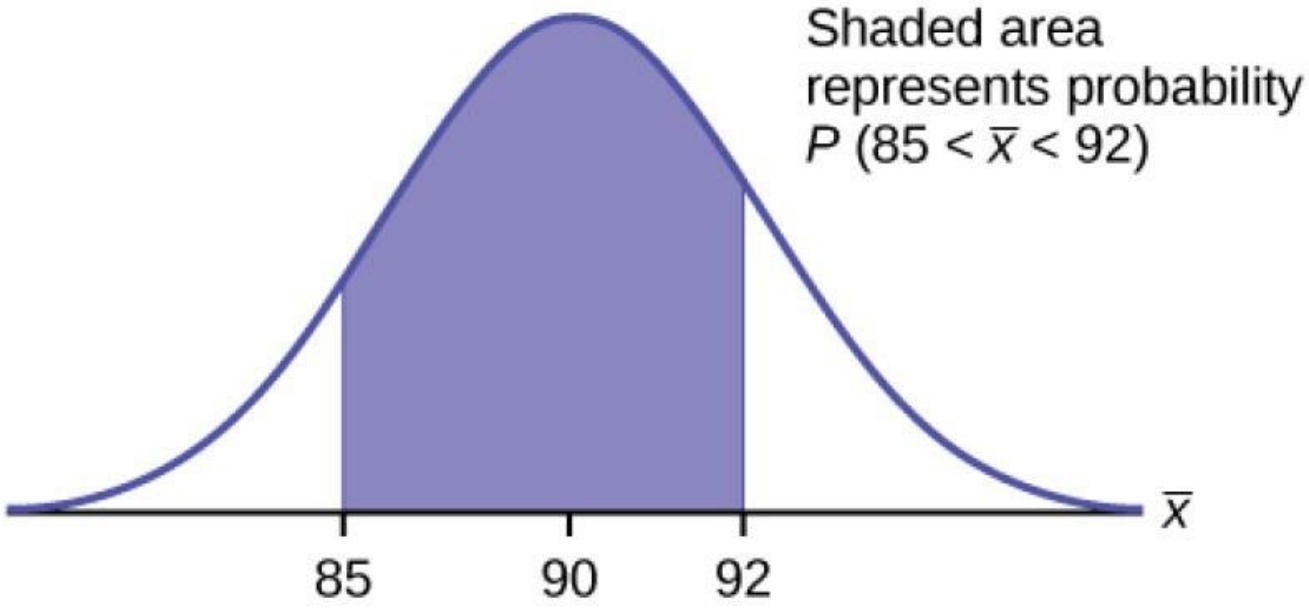
$$\bar{X} \sim N\left(\mu_x, \left(\frac{\sigma_x}{\sqrt{n}}\right)^2\right)$$

  - **Example**: If you randomly select *n* = 9 samples from a population with mean $\bar{X} = 90$ and SD $\sigma_X = 15$ in their IQ scores. By CLT, the average IQ score of the sample will be distributed as follows:

$$\bar{X} \sim N\left(90, \left(\frac{15}{9}\right)^2\right)$$

Central Limit Theorem

- Example – cont.: Given the CLT, we can compute the probability of having an average score in certain range (e.g., 85-92).



Central Limit Theorem

- **Why Does CLT Matter?**
  - For parametric hypothesis tests of the mean (e.g., t-test)
    - If your sample size is large enough, you can use these hypothesis tests even when your data are not normally distributed.
  - For precision for estimates
    - With a larger sample size, your sample mean is likely to be close to the population mean.

Your turn

The length of a human pregnancy is normally distributed with a mean of 272 days with a standard deviation of 9 days (Bhat & Kushtagi, 2006).

1. State the random variable.
2. Find the probability of a pregnancy lasting more than 280 days.
3. Find the probability of a pregnancy lasting less than 250 days.
4. Find the probability that a pregnancy lasts between 265 and 280 days.
5. Find the length of pregnancy that 10% of all pregnancies last less than.
6. Suppose you meet a woman who says that she was pregnant for less than 250 days. Would this be unusual and what might you think?

We will draw it all out, then run the markdown file zscores.Rmd

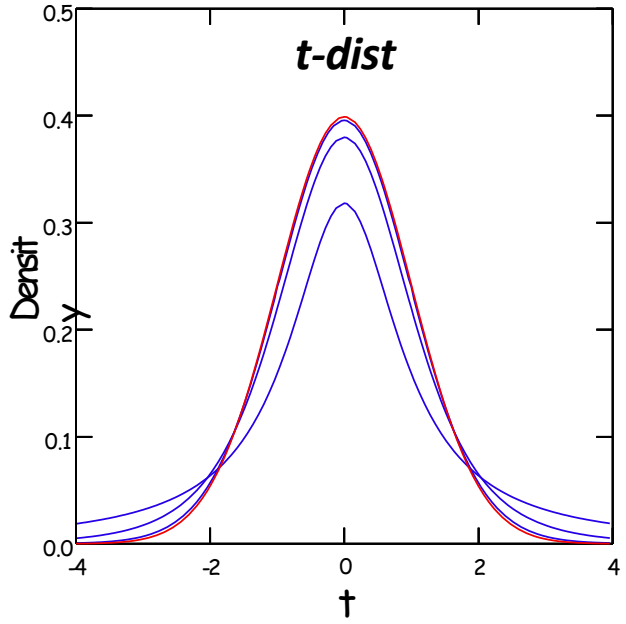
Your turn

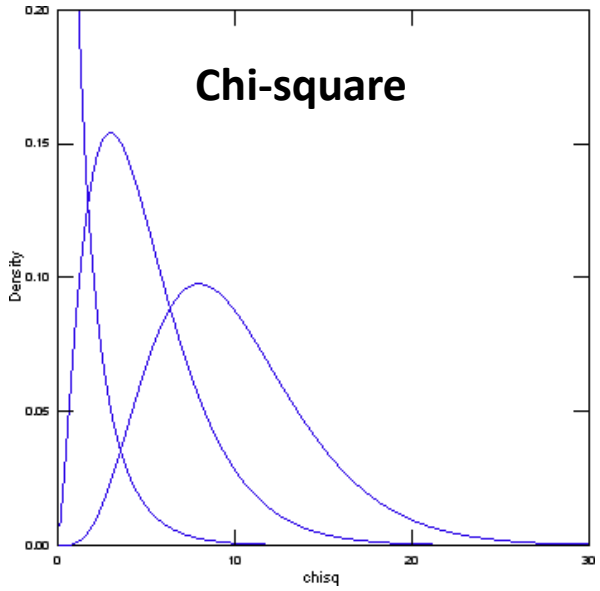
The mean mathematics SAT score was 514 with a standard deviation of 117. Assume the mathematics SAT score is normally distributed.

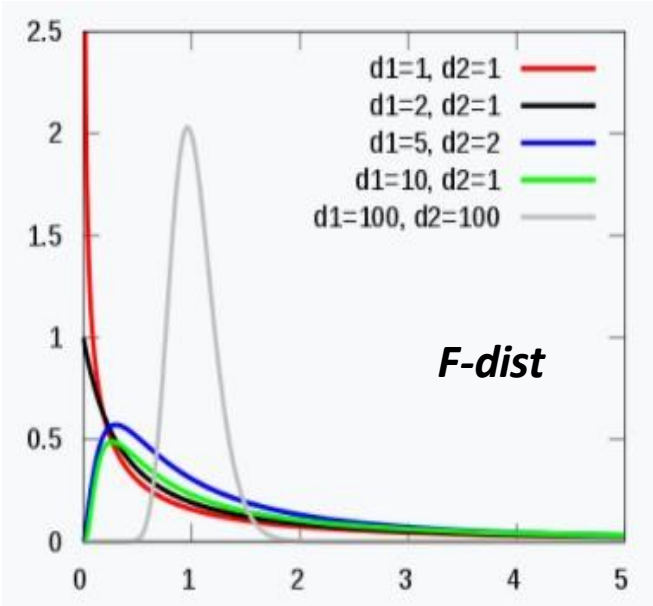
1. State the random variable.
2. Find the probability that a person has a mathematics SAT score over 700.
3. Find the probability that a person has a mathematics SAT score of less than 400.
4. Find the probability that a person has a mathematics SAT score between a 500 and a 650.
5. Find the mathematics SAT score that represents the top 1% of all scores.

Distributions

- **Other Distributions**
  1) *t* distribution
  2) Chi-square distribution
  3) *F* distribution







Key Takeaways

1. For valid inferences, we need data, representative of the population, collected via random sampling.
2. We can borrow from statistical theories to obtain sampling distribution of sample statistics.
3. Some examples of distributions include binomial, normal, *t*, *chi-square*, and *F* distributions.
4. Central Limit Theorem enables us to use a normal distribution as the sampling distribution for the sample mean (when *n* is large), which is essential in hypothesis testing.

The Logic of Statistical Inference

Hypothesis Testing

Null or **statistical hypothesis**, denoted by H_0

Scientific, alternative, or **research hypothesis**, denoted by H_A or H_1

What is Hypothesis Testing?

Hypothesis testing: A procedure, **based on sample evidence and probability theory**, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Null Hypothesis H_0: A statement about the value of a population parameter that is ***assumed*** to be true for the purpose of testing

Example: Twenty percent of all juvenile offenders are disciplined by the criminal justice system.

Alternative Hypothesis H_a: A statement about the value of a population parameter that is assumed to be true ***if the Null Hypothesis is rejected*** during testing.

Example: The percentage of juvenile offenders who are disciplined is not equal to 20%

Hypothesis Generation

These two hypotheses, designated H_O and H_A (or H_1), are mutually exclusive exhaustive- the don't overlap

Properties of the Null

- Specifies ***no difference*** or no change from a standard or theoretical value
- Always specifies something about a particular ***population*** parameter
- Used in constructing a sampling distribution
- For the subsequent quantitative work, the null hypothesis is assumed to be true

Properties of the Alternative

- Usually stated in general terms
- Mutually exclusive - no overlap with H_O
- Can be directional or non-directional

Directional vs. Non-directional H_As

Non-directional H_A - usually stated as “does not equal” or “is different than”

Directional H_A - stated as “greater than” or “less than”

note that a non-directional hypothesis is equal to the two directional hypotheses “greater than” or “less than”

The Null Hypothesis

A statement of no relationship or no difference, denoted H_0:

In practice, in statistics, we make decisions about hypotheses in relation to the null hypothesis rather than the research hypothesis

This is because the null hypothesis states that the parameter in which we are interested is a particular value.

For example, your null hypothesis H_0 might be that there is no difference between sexual minority and heterosexual youth in average number of ACEs, or, put differently, that the difference is equal to zero.

Error in Hypothesis Testing

Whenever we rely on sample statistics to make statements about population parameters, we must always accept that our conclusions are tentative

This means that when we test hypotheses in research, we generally do not ask whether a hypothesis is true or false. Why?

The question is: can we make an inference, or draw a conclusion, about our hypotheses based on what we know from a sample?

Risk of error and Levels of Significance

In statistical inference, we assess the risk of making a wrong decision about the population parameter in reference to Type I error.

- The language is “reject” or “fail to reject” the null
- The amount of error we are willing to risk is called the significance level of a test of statistical significance
- The decision to reject or not is based on a sample statistic
- The significance criterion is denoted alpha

The estimate of the risk of Type I error that is associated with rejecting the null hypothesis in a test of statistical significance (based on a sample statistic) is called the observed significance level and is ordinarily represented by the symbol *p*.

Types of Decisions and Errors

	Decision	
State of Nature (reality)	**Fail to reject H_0**	**Reject H_0**
H_0 is true	Correct decision $(1 - \alpha)$	Type I error (α)
H_0 is false	Type II error (β)	Correct decision $(1 - \beta)$ = power

Classic Example: Decision making in the courtroom

Hypo: D is on trial for murder. The jury hears the evidence in the case and must decide guilt or not guilt:
What is H0? Put in terms of “no difference”
What is H1?

Example

- A man is on trial for murder;
- The null hypothesis
- What are the possible decisions

Hypothesis testing framework		
	Person Not guilty	**Person is guilty**
Jury Decides Person Not guilty	Correct	Type I error
Jury Decides Person Guilty	Type II error	Correct

H_0: D is not guilty

H_1: D is guilty

	Decision	
State of Nature	**Fail to reject H_0**	**Reject H_0**
H_0 is true	Correct decision $(1-\alpha)$ D is not guilty Jury finds D not guilty Good result	Type I error (α) D is not guilty Jury finds D guilty Not good result
H_0 is false	Type II error (β) D is guilty Jury finds D not guilty Not good result	Correct decision (1-β) D is guilty Jury finds D guilty Good result

Statistical versus Practical Significance

- Statistical significance
  - Determined through null hypothesis significance testing
- Practical significance
  - Measured by effect size
- Legal significance
  - Measured by the weight of the evidence
  - Beyond a reasonable doubt
  - Probable cause

What types of errors do we potentially make?
Test H_0: Sexual minority and heterosexual youth experience the same number of ACEs, i.e. no difference

Hypothesis testing framework		
Decision/Truth	Null is true	Null is false
Fail to reject Null hypothesis	There is no difference between sexual min. and heterosexual youth, and we conclude that there is no difference	Type II (or beta) error There is a difference between sexual min. and heterosexual youth, and we conclude that there is not a difference
Reject Null hypothesis	Type I (or alpha) error There is a no difference between sexual min. and heterosexual youth, but we conclude that there is a difference	There is a difference between sexual min. and heterosexual youth, and we conclude that there is a difference

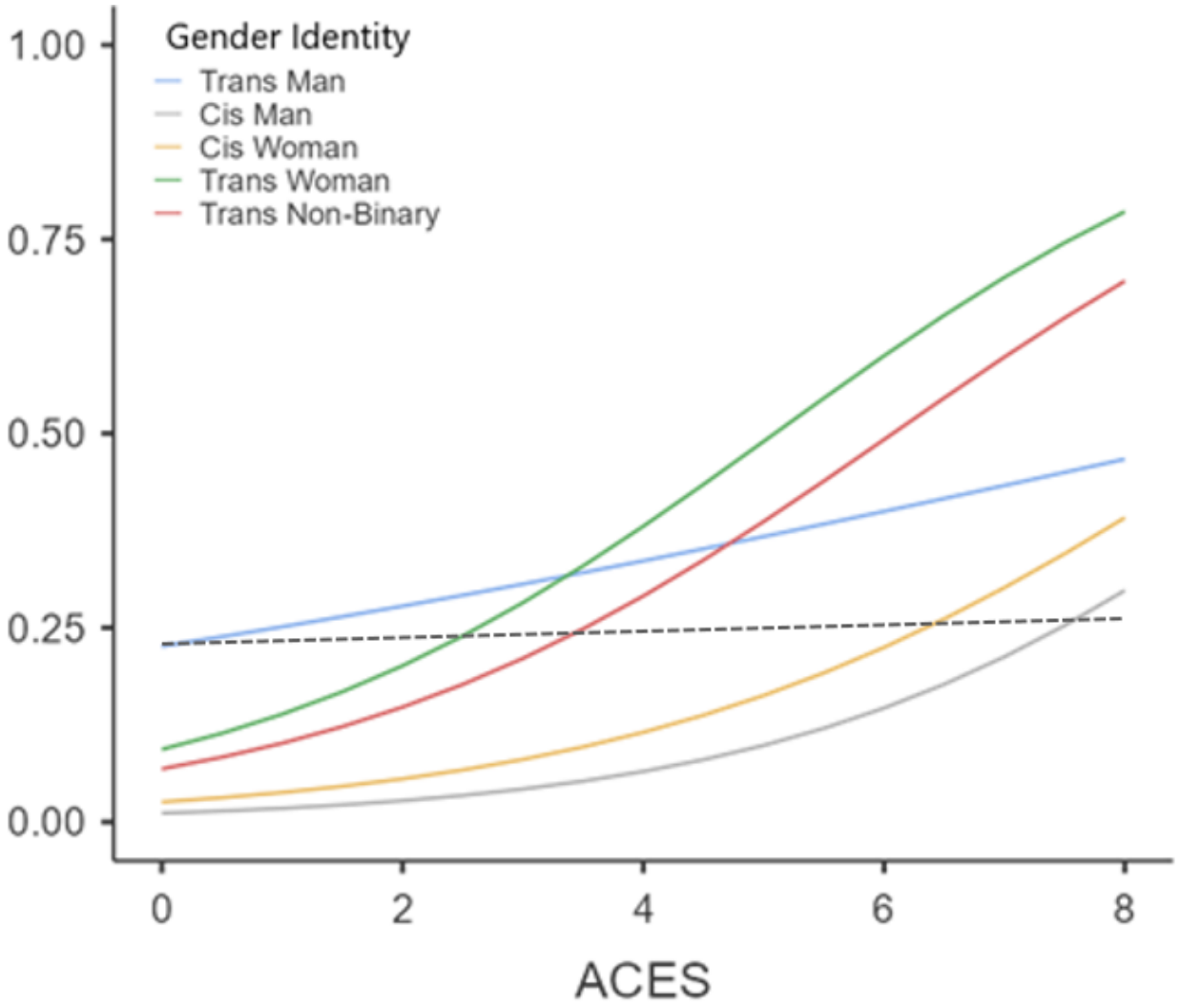


Fig. 2. Predicted probability of suicide attempt by ACEs CR across gender identity.
Notes. Predicted probabilities were generated from moderation model of ACEs CR and gender identity controlling for age, race, sexual identity, sex at birth, income, poverty level and education. Dotted line shows similarity between 0 ACEs and 8 ACEs.

Type I error

- Building a sampling distribution provides a method for defining our risk of a Type I error
- We cannot construct a sampling distribution every time we want to test a hypothesis, why?
- Fortunately, (or unfortunately) there is another method we can use
- We rely on known sampling distributions based on probability theory

Hypothesis Testing in a Probabilistic Framework

Intro to probability theory

The magazine Discover once had a special issue on "Life at Risk." In an article, Jeffrey Kluger describes the risks of making it through one day:

> *Imagine my relief when I made it out of bed alive last Monday morning. It was touch and go there for a while, but I managed to scrape through. Getting up was not the only death-defying act I performed that day. There was shaving, for example; that was no walk in the park. Then there was showering, followed by leaving the house and walking to work and spending eight hours at the office. By the time I finished my day – a day that also included eating lunch, exercising, going out to dinner, and going home – I counted myself lucky to have survived in one piece.*

Is this writer unusually fearful? No, why?

Intro to probability theory

- ”There is not a single thing you can do ... that isn’t risky enough to be your last.”
- Examples
  - 1 out of 2 million people will die from falling out of bed.
  - 1 out of 400 will be injured falling out of bed
  - 1 out of 77 adults over 35 will have a heart attack this year
  - Black Americans are twice as likely to be shot by the police compared to Whites (https://www.washingtonpost.com/graphics/investigations/police-shootings-database/)
  - The average American faces a 1 in 13 risk of suffering some kind of injury in home that necessitates medical attention
  - 1 out of 32 risk of being the victim of some violent crime, 1 out of 14 risk of having property stolen this year
- These are all probabilities that were calculated from counts of reported accidents

Probability: An assessment of risk

There is a modest amount of theory we can cover. I will introduce some basic concepts of probability theory and then present some applications. This is meant to give you some basic skills by which to challenge what you read.

A number that lies between 0 and 1 that measures uncertainly of a particular event
The probability of an outcome is simply

$$P(outcome) = \frac{A}{S}$$

where A = the number of ways an event can occur and S = the total number of outcomes

The Sample Space

- A **Sample Space** is a list of all possible outcomes of an experiment
- Example: List the possible sample space for 1 toss of a fair coin.
- S = {*H*, *T*}
- **Example**: List the possible sample space for 2 tosses of a fair coin.

Harder: The Sample Space

- List all possible outcomes of 1 toss of a fair die
- List all possible outcomes of 1 toss of two fair die and then compute the probability of observing a total score of 6? At least 6? At most 12?

Assigning Probabilities

- Any probability that is assigned must fall between 0 and 1
- The sum of the probabilities across all outcomes must be equal to 1
- An outcome will be assigned a probability of 0 if one is sure that that outcome will never occur
- Likewise, if one assigns a probability of 1 to an event, then that event must occur all the time.

Formal Rules of Probability

- Numbers called probabilities are assigned to outcomes in the sample space such that the sum of the numbers over all outcomes is equal to one
- Suppose that the sample space for our random experiment is S
- We denote an event by a letter, say A, where A is a subset of S
- Events can be described as follows
  - $A \cap B$ is the event that both A and B occur, i.e. the intersection of two events
  - $A \cup B$ is the event that either A or B occur, i.e. the union of two events
  - A^c is the event that A does not occur (called the complement of A)

Illustration

Suppose I choose a student at random from class and record the month he, she or they were born. The student could be born during any of the 12 months.

- List the sample space of the experiment
- Let A = the student was born during the last half of the year and B = the student is born during a month that is four letters long.
- Describe the following events in words and find each
  - $A \cap B$
  - $A \cup B$
  - A^c and B^c

The Fair Coin

- Example: In football, a coin toss⇒ at what point do you become suspicious that the coin is unfair?
- Statistical inference provides a ***systematic way*** of determining risk
- The coin toss can be thought of as a test statistic
  - What is H_0
  - What is H_1
  - What is the Type I error of the test statistic?
  - What is the Type II error?
  - What percentage of the time are you willing to be wrong?

The Fair Coin

- How can we calculate the risk of a Type I error associated with a specific outcome in a test of statistical significance also called the observed significance level?
  - One way is to determine how often a fair coin would be that the Patriots do not always get the ball
- Ultimately you are interested in determining the likelihood of getting X heads in a very large number of ***samples or trials*** of a fair coin (i.e., tosses in 16 football games for 10 years)
  - The resulting distribution is the sampling distribution
  - The shape of the distribution of the number of heads in many trials is called a binomial distribution
- Let's consider what thousands of samples of $X = 10$ tosses of a fair coin looks like...

The Multiplication Rule

- In order to estimate the risk of a Type I error in the case of a series of tosses of a fair coin, we can use the multiplication rule
- The MR is based on the assumption of independence: each event in a sample is independent of every other event
- What does this mean in terms of a coin toss? Is this assumption reasonable? When does this assumption not hold?
- To truly understand we need to foray into probability theory both conceptually and theoretically

The Binomial Sampling Distribution

To calculate the probability of observing X results in n trials use the Binomial formula

The binomial coefficient is defined by the next expression:

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

The binomial formula is defined by :

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where X = the number of successes, n = the number of trials, p = the probability of success on each trial

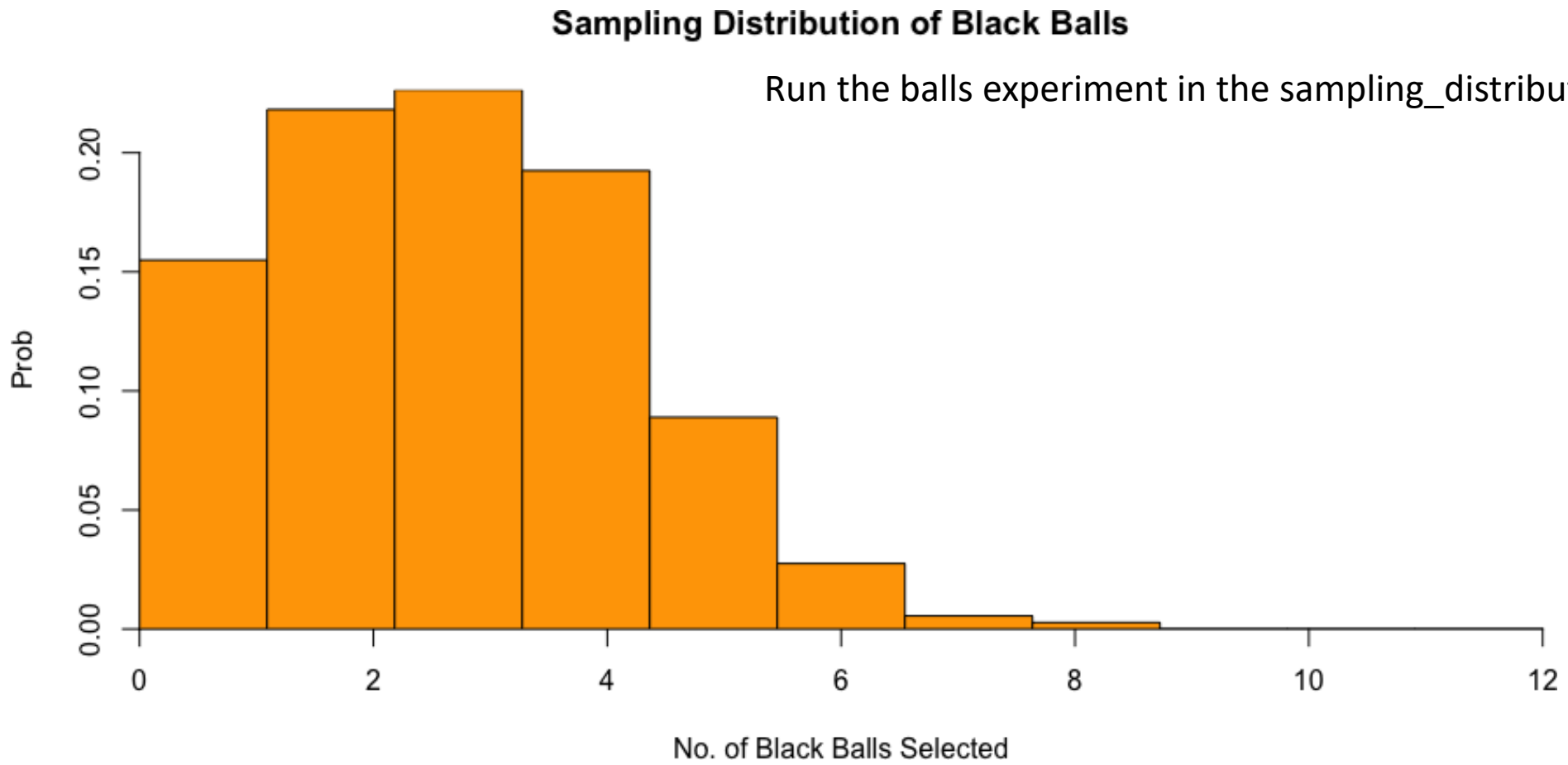
Examples

- Calculate the probability of each of the following:
  1. Two tails in two tosses of a fair coin.
  2. Two heads in three tosses of a fair coin
  3. Four heads in four tosses of an unfair coin where the probability of a head is 0.75
  4. Three sixes in three rolls of a fair die
  5. Five fours in five rolls of an unfair die where the probability of a four is 0.25.

Example

- Suppose one has a bowl with 4 black balls and 2 white balls. Draw two balls at random. What are the possibilities? What is the probability of getting 2 white balls?
- Suppose now the bowl has 100 balls, 25 balls are black and 75 balls are white. Select 12 balls at random. What is the probability that you get 12 white balls?
- How many black balls would you expect to see in your sample of $N = 12$?

Results of Experiment



Run the balls experiment in the sampling_distributions.Rmd

No matter how many times you run this, you will never see a high probability of getting 12 black balls. Think about the implications for this…