Sampling and estimating unknown quantities from samples

Making assumptions

- About the data

- Sampling theory: will help us specify the assumptions upon which our statistical methods rely

Inferences about what and based on what?

- Inferences about the population

- Based on the sample
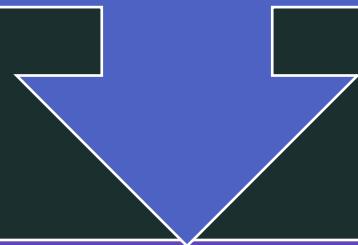
Population: cogsci questions

- All of the undergraduate students at IIITH?
- Undergraduate students in general, anywhere in the world?
- Indians currently living?
- Indians of similar ages to my sample?
- Anyone currently alive?
- Any human being, past, present or future?
- Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?
- Any intelligent being?

What is the population?

Not always clear

This is probably the first assumption you will make: "My study will reveal phenomenon X as it pertains to the population of ____"

Sample

- Different sampling schemes: how you gather a data sample from the population

Simple random sampling without replacement
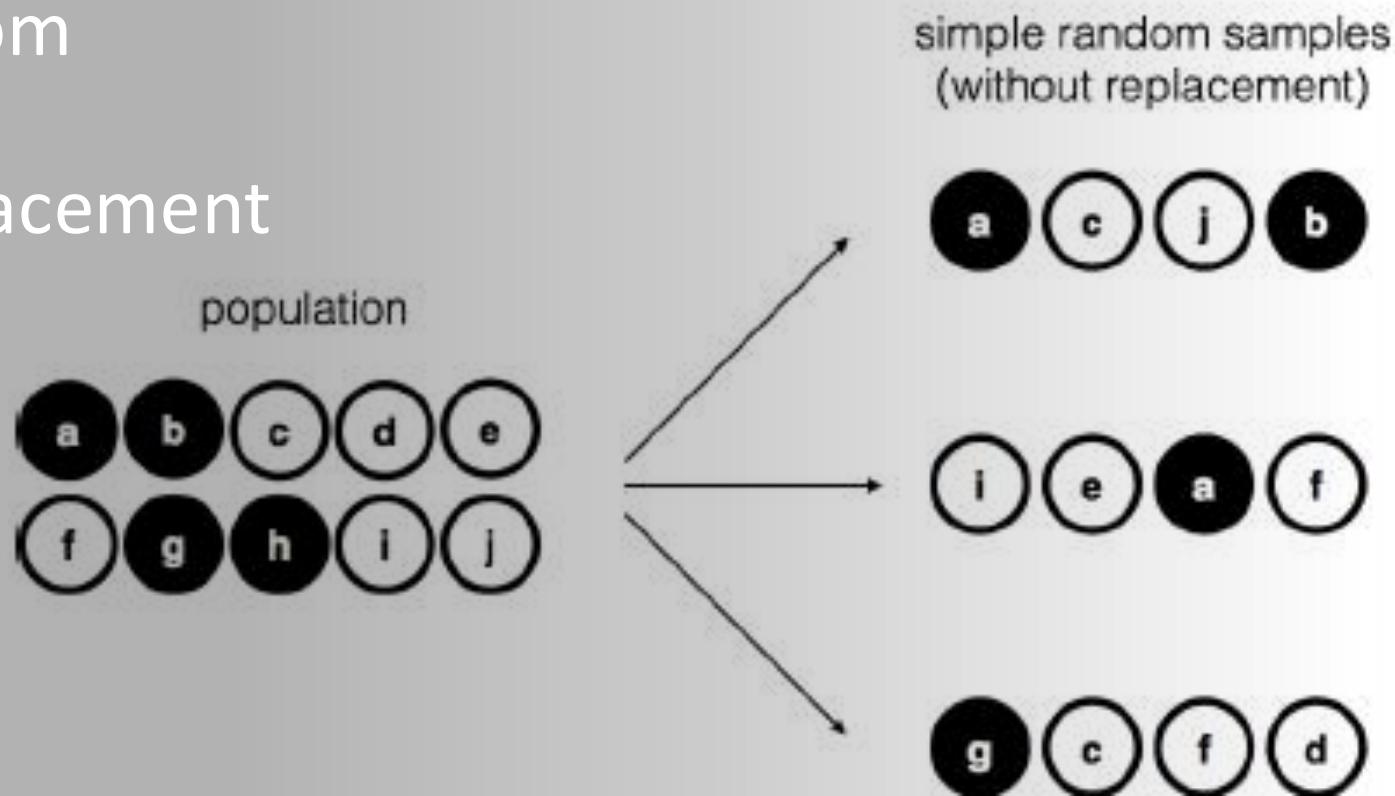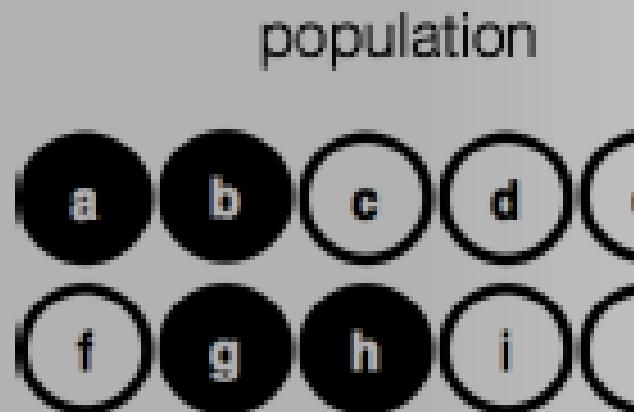
Figure 10.1: Simple random sampling without replacement from a finite population

Biased sampling without replacement
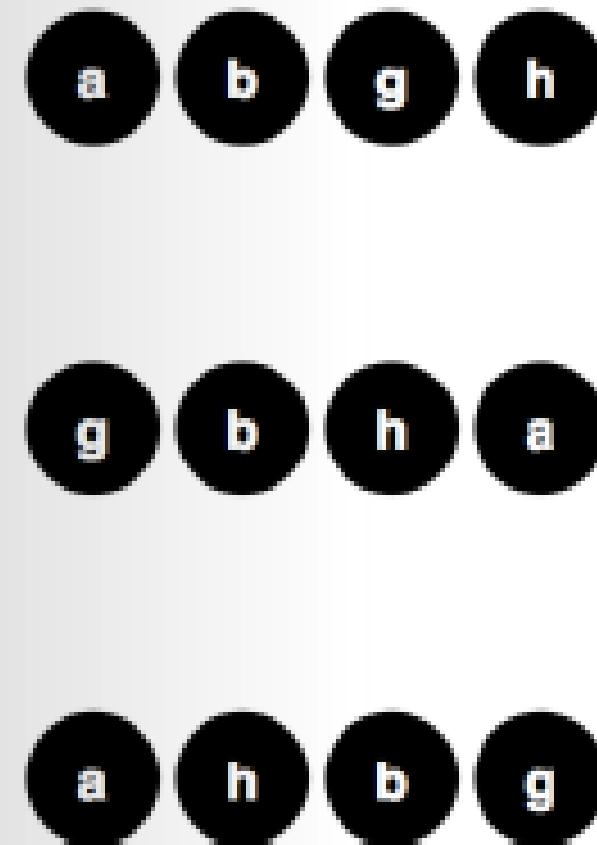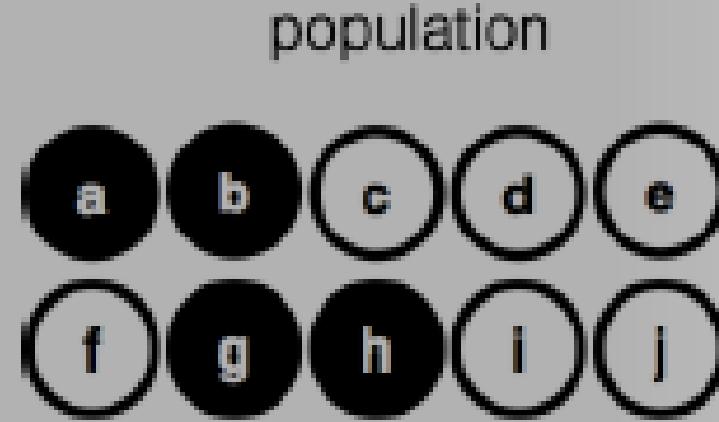
Figure 10.2: Biased sampling without replacement from a finite population

Simple random sampling with replacement
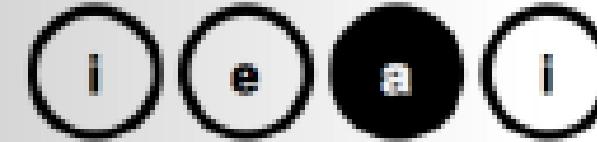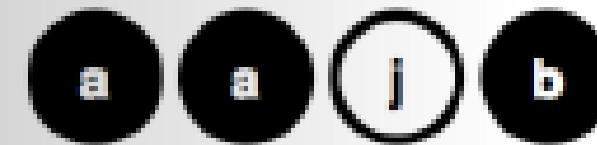
Figure 10.3: Simple random sampling *with* replacement from a finite population

Real world behavioral experiments

Samples with or without replacement?

Most statistical methods assume sampling with replacement

For a large enough number of participants, the difference does not matter too much

Biased vs unbiased (simple random) sampling needs attention though

Stratified Sampling

- A natural strata structure (e.g. schizophrenia or Alzheimer's study)

- Simple random sampling within each strata: why?

- If you attempt to randomly sample from the whole population, you will get skewed numbers across the groups of interest.

- For example: Nielson et al., 2015, *PNAS* – all 9 participants in our fMRI study were female as we recruited by placing ads around our psychology building.

- Solution: oversample from the rare strata to equate the numbers

Convenience Sampling

- A convenient sample, not random usually

- e.g. recruit from the undergrad population of iiith

- Not always a problem but depends on the study question and goals

- Most psychology studies involve convenience sampling, which is why people have realized the importance of at least occasionally doing large N replication studies, using more ecologically valid frameworks to test psychology theories that were developed in the lab.

Snowball sampling

- Typically to be used when you want to recruit hard-to-locate participant groups
- e.g. a study on trans health, you do not have many personal contacts, and recruiting from the whole population might be too expensive if you have to discard the majority of the data
- So you get the few contacts you have, ask them to provide other contacts, and so on = snowball sampling
- Fraught with issues: privacy, ethical, highly non-random samples in ways that are hard to control
- However, this is often the only way you can get a sufficient number of participants for such studies
- Snowball sampling is a type of convenience sampling

What to do when you don't have a random sample

If you know exactly how you sampled and what bias you introduced, there are advanced statistical methods to correct for bias (e.g. in stratified sampling).

Otherwise, if you have a random sample, you only need to worry about randomness in sampling certain features that are relevant for the concept being studied.

Memory study

- Options:
- Sample from the Indian population
- Sample from many different countries but restricted to people born on a Sunday
- Goal: to make conclusions about how memory works in all humans
- Both are random samples, but one is better than the other: where the randomness is in a feature that doesn't matter for the concept being studied given the generalization goal

Population parameters, sample statistics

- Plot b: mean IQ of 98.5, and the standard deviation of 15.9, N = 100 – sample statistics
- An approximation of the population parameters.
- Our goal: how can we estimate population parameters based on sample statistics?
- Also, can we come up with a measure of "confidence" in our estimates?
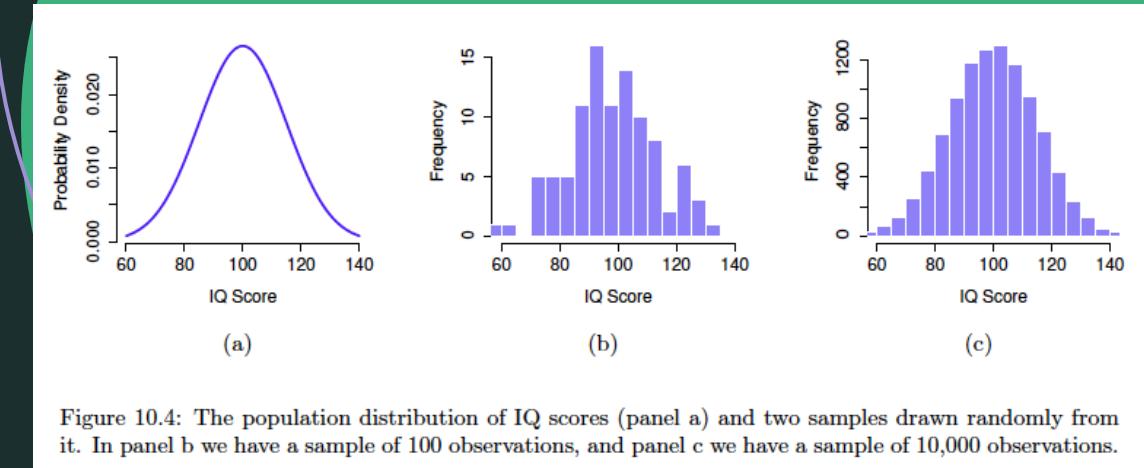
Figure 10.4: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

The law of large numbers

Previous slide: plot c with a greater N (10k) provided a closer approximation to the population parameters
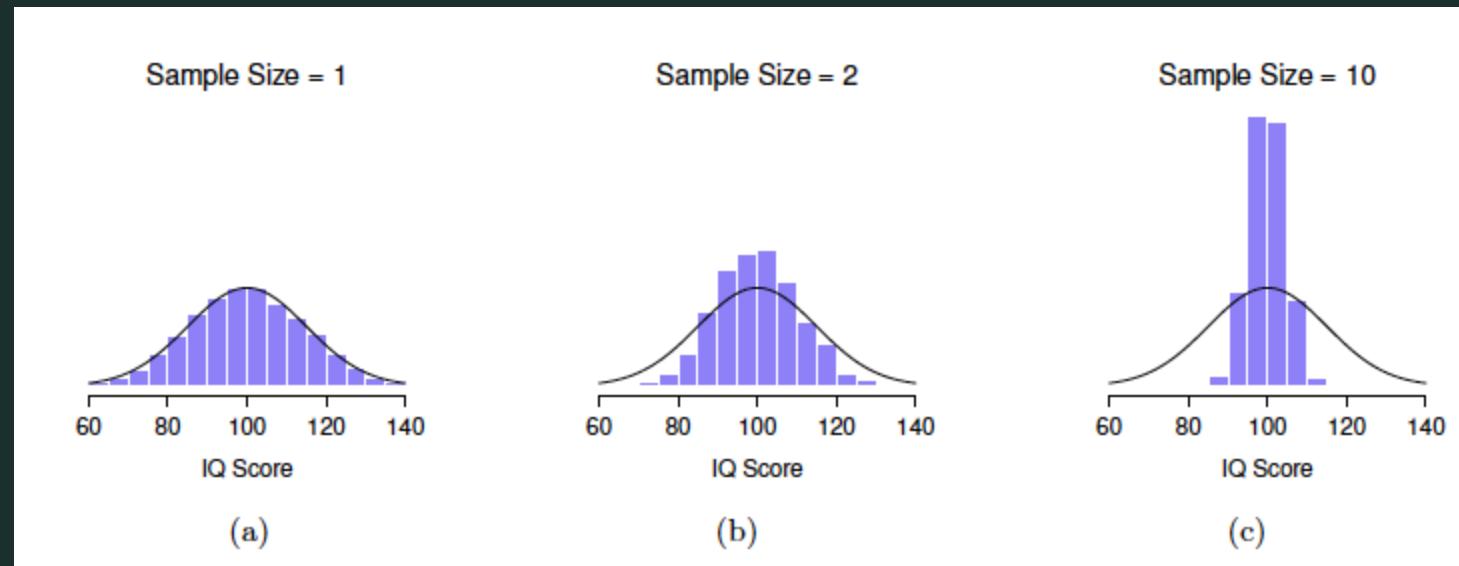
This law of large numbers applies to many statistics, but easiest to demonstrate is as a law of averages (sampling distribution of the mean, which we saw in the probability distribution lecture)

Revisiting the central limit theorem

Table 10.1: Ten replications of the IQ experiment, each with a sample size of $N = 5$.

	Person 1	Person 2	Person 3	Person 4	Person 5	Sample Mean
Replication 1	90	82	94	99	110	95.0
Replication 2	78	88	111	111	117	101.0
Replication 3	111	122	91	98	86	101.6
Replication 4	98	96	119	99	107	103.8
Replication 5	105	113	103	103	98	104.4
Replication 6	81	89	93	85	114	92.4
Replication 7	100	93	108	98	133	106.4
Replication 8	107	100	105	117	85	102.8
Replication 9	86	119	108	73	116	100.4
Replication 10	95	126	112	120	76	105.8

10k sample means

The black line is the true population distribution
What observation do can you make about the mean of any single sample and how that relates to the population mean across different values of sample size?

Other observations

- The mean of the sampling distribution is the same as the mean of the population

- The standard deviation of the sampling distribution (i.e., the standard error) gets smaller as the sample size increases

- The shape of the sampling distribution becomes normal as the sample size increases

Standard error of the mean

- SEM

- Sampling distribution of the mean

- The standard deviation of the sampling distribution (the standard error), or the standard error of the mean (SEM, in this case) relates to the population standard deviation sigma as

$$\text{SEM} = \frac{\sigma}{\sqrt{N}}$$

IQ test in a small village of Andhra

- We're not sure the true population mean is 100 (as "defined" by the test makers)

- We need to provide a best guess about the population mean based on say 50 villagers who agreed to take the test

- I conduct the test and the mean in this sample of 50 comes out to be 97

- What is my best guess about the population mean?

- CLT, sampling distribution of the mean exercises earlier --> my best guess is 97!

Estimating population mean from the sample mean

| Symbol	What.is.it	Do.we.know.what.it.is
$\bar{X}$	Sample mean	Yes calculated from the raw data
μ	True population mean	Almost never known for sure
$\hat{\mu}$	Estimate of the population mean	Yes identical to the sample mean
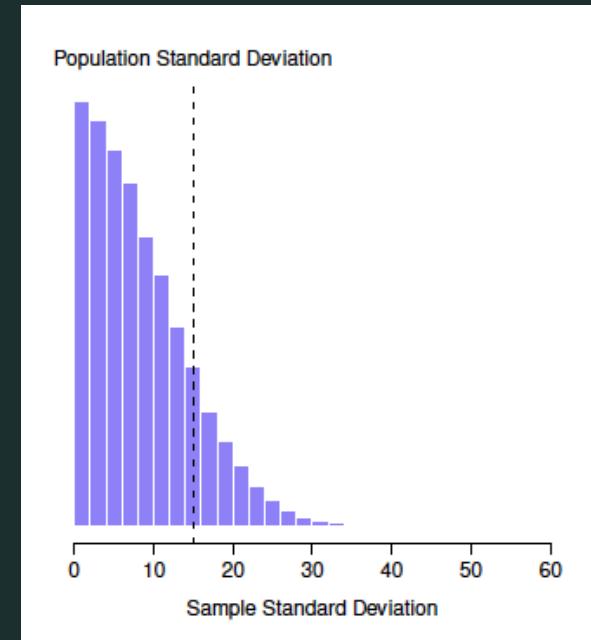
Remember: this only works when CLT applies, so need a sufficiently large sample size, otherwise your estimate will not be very accurate

How about population standard deviation?

- Say N = 1, IQ = 120 (IIITH student)

- What is your best guess about the IIITH mean IQ?

- 120 is the best guess you can make based on your data, you wouldn't be very confident but you can make a guess

- However, what is the population standard deviation?

- No idea! With our sample of 1, the standard deviation is 0 but it would not make any sense to say that about the population as we know it is going to be a wrong guess, so can't say this is the best guess possible
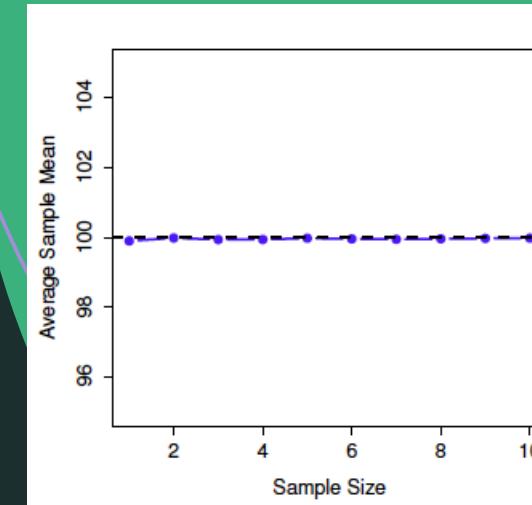
How about population standard deviation?

- N = 2, say s.d. (of the sample) = 8.5

- Intuition: the sample s.d. is a biased estimator of the population s.d.
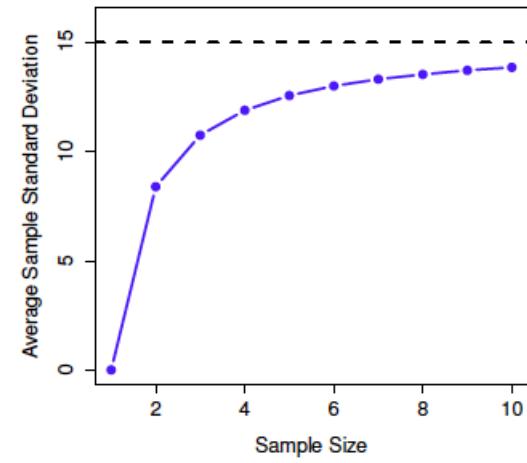
Intuition: Sample s.d. is a biased estimator of population s.d.

- Demonstrate this?

- Simulate N = 10, N = 100, etc?

- s.d. is systematically smaller than the population s.d.

Biased and unbiased estimators

The sample mean is an unbiased estimator of the population mean

The sample s.d. is a biased estimator of the population s.d.

How do we fix this bias?

Sample variance

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

- Also a biased estimator (of the population variance)
- A minor tweak in the formula can make it an unbiased estimator

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

- This is what the R var function calculates, not the sample variance but the unbiased estimator (dividing by N-1 instead of N)
- Similarly for the s.d.

Reporting

- When you calculate the sample s.d. (dividing by N), it should be referred to as the sample s.d.

- When dividing by N-1, this is an unbiased estimator of the population s.d. -- I.e., your best guess about the population s.d. parameter!

- Many people use the unbiased estimator (the output of R std and var functions) and refer to them as the sample s.d. and sample variance

- This is technically incorrect

Estimating the population s.d. and variance: Summary

Symbol	What is it?	Do we know what it is?
s	Sample standard deviation	Yes, calculated from the raw data
σ	Population standard deviation	Almost never known for sure
$\hat{\sigma}$	Estimate of the population standard deviation	Yes, but not the same as the sample standard deviation

Symbol	What is it?	Do we know what it is?
s^2	Sample variance	Yes, calculated from the raw data
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes, but not the same as the sample variance

Standard normal distribution

- If you know the population mean and s.d., you can normalize your variable:

$$\frac{X - \mu}{\sigma}$$

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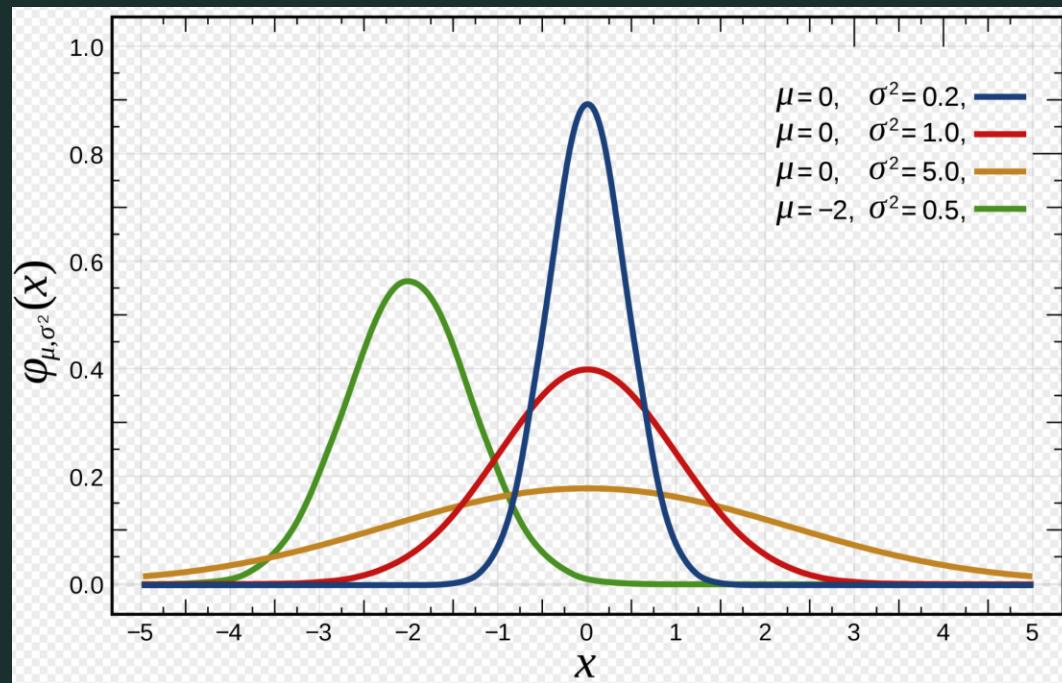
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Mu = 0, var = 1
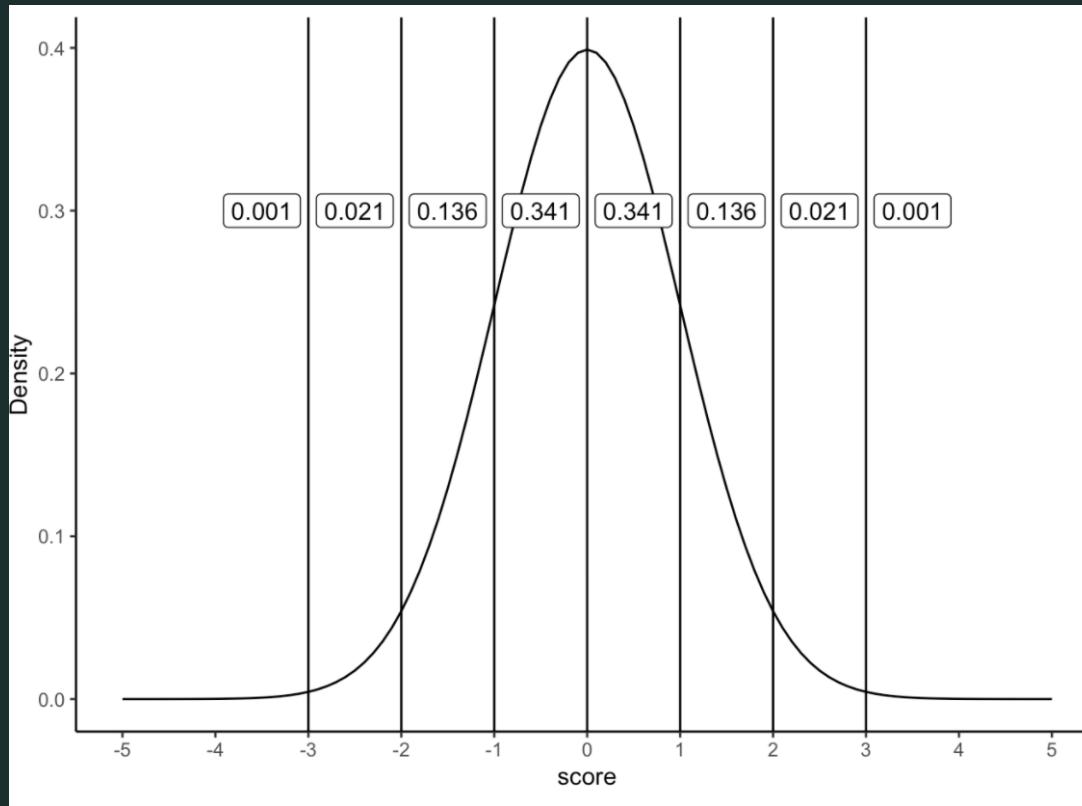
$$\varphi(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$$

Standard normal distribution

The one in red is the standard normal distribution

Normal distributions:
Approx: 68% of the values lie within 1 s.d. of the mean
95% of the values lie within 2 s.d. of the mean
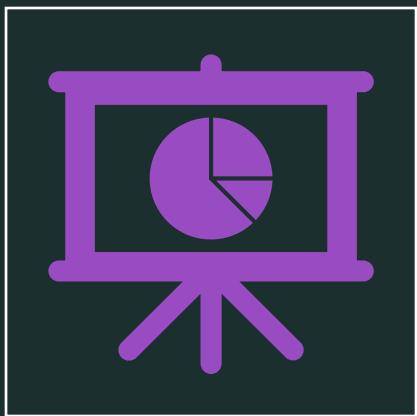
The standard normal quantiles

p	z_p
0.80	1.281 551 565 545
0.90	1.644 853 626 951
0.95	1.959 963 984 540
0.98	2.326 347 874 041
0.99	2.575 829 303 549
0.995	2.807 033 768 344
0.998	3.090 232 306 168

p	z_p
0.999	3.290 526 731 492
0.9999	3.890 591 886 413
0.99999	4.417 173 413 469
0.999999	4.891 638 475 699
0.9999999	5.326 723 886 384
0.99999999	5.730 728 868 236
0.999999999	6.109 410 204 869

Confidence intervals

Ok, so now you've made a guess about the population parameters from your data sample

How confident are you about your guess? (recall, that this probability need not be an intuitive probability, like we discussed, it is a frequentist probability)

Confidence Intervals (CIs)

- My best guess of the mean IQ of IIITH students is 120 based on a sample of 100

- The 95% confidence interval is 110–130

- Compared to the 95% CI of 100-140

- The margin of error is lower in the former case

- Intuition: you can reduce margins of error by using more people in your sample

How do we construct CIs?

- Assume true population mean = mu and s.d. = sigma

- We know from the central limit theorem that the sampling distribution of the mean is normal, and that for Normal distributions, 95% of the values lie within 1.96 (had approximated it to 2 earlier) standard deviations from the mean.

- Check for yourself using qnorm(p = c(.025, .975))

How do we construct CIs?

$$\mu - (1.96 \times \text{SEM}) \leqslant \bar{X} \leqslant \mu + (1.96 \times \text{SEM})$$
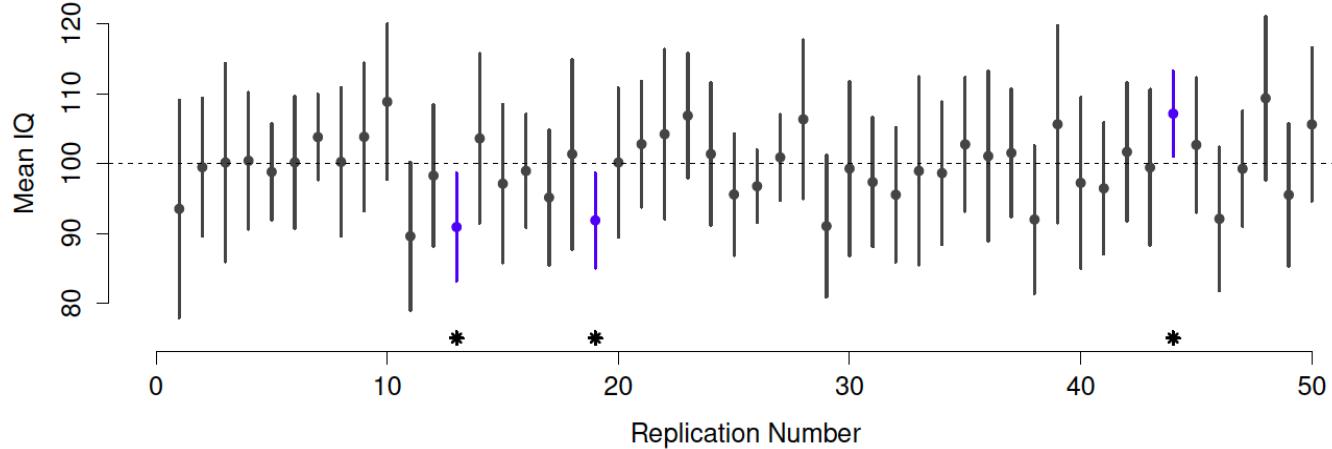
$$\bar{X} - (1.96 \times \text{SEM}) \leqslant \mu \leqslant \bar{X} + (1.96 \times \text{SEM})$$

- SEM is used because note that we are referring to the sampling distribution of the mean, so the s.d. that matters is the standard error of the mean. $\text{SEM} = \dfrac{\sigma}{\sqrt{N}}$

- "The population mean has a 95% chance of falling into this range." -- the textbook says this when it introduces this concept of a 95% coverage area in the Normal distribution but see the final section on interpretation
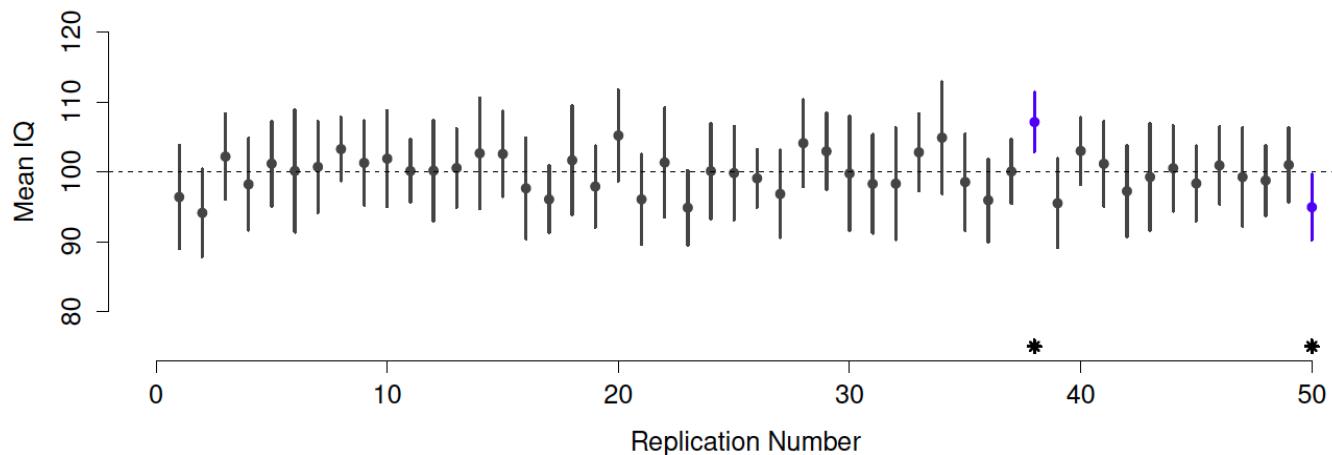
The confidence interval changes from sample to sample

- So if I say I'm 95% confident true population mean IQ lies between 110-130, and then the very next sample, I say I'm 95% confident the true population mean IQ lies between 100-140, there is something quirky about this.

- Also, note that the calculation on the previous slide used the standard error, we do not know the population s.d.

- What it works out to be is that if you repeat this procedure many times, 95% of the confidence intervals you construct would be expected to contain the population mean – this is the correct interpretation of a 95% CI.

Sample Size = 10

Sample Size = 25

Frequentist CI

- A population mean is not a repeatable random variable

- Repeatable: very important for frequentist probability interpretation

- What is repeatable is the CI (in different samples)

- So a frequentist is not allowed to make probabilistic statements about the probability of the population mean (I.e., there is a 95% chance the population mean lies in a certain range) but is allowed to make probabilistic statements about the CI across many samples (I.e., that 95% of such CIs will contain the true population mean).

Does the interpretation matter practically?

- The Bayesian version = credible intervals (will be covered in the last lecture)

- Under some conditions, credible intervals and frequentist CIs can look very different. So the interpretation differences matter in these cases.

An additional issue

In the SEM formula, we used the population s.d. but we do not know the population s.d.!

So we have to use an estimate

We also know that the SE (i.e., the s.d of the sampling dist) changes as the sample size from all the simulations we did

What is a probability distribution that looks very much like the Normal distribution but has a dependence on sample size?

The T-distribution!

So instead of using the standard normal quantiles, we will use quantiles from the T-distribution

```
N <- 10000   # suppose our sample size is 10,000
qt( p = .975, df = N-1)   # calculate the 97.5th quantile of the t-dist
```

1.960201
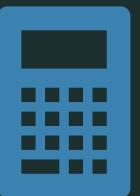
```
N <- 10   # suppose our sample size is 10
qt( p = .975, df = N-1)   # calculate the 97.5th quantile of the t-dist
```

[1] 2.262157

Captures the intuition that with smaller sample sizes, our margin of error should be larger

Summary

Basic ideas about samples, sampling and populations

Statistical theory of sampling: the law of large numbers, sampling distributions and the central limit theorem

Estimating means and standard deviations

Estimating a confidence interval

Quiz 1

- Conceptual

- No coding

- Chapters 1-10 except for the R coding chapters in "Learning statistics with R" + all lecture materials and supplemental readings if any covered until the quiz.

The Scientific Method

- A set of principles about the appropriate relationship between ideas and evidence.
    1. Develop theories (i.e. ideas)
    2. Derive hypotheses from the theories and test them (i.e., evidence)
    3. Modify your theory based on evidence
    4. Repeat 2-3 or if required restart at 1

Theory and Hypothesis

- Theory: A GENERAL hypothetical explanation of a natural phenomenon
- Hypothesis: A SPECIFIC falsifiable prediction made by a theory

Determine if the following are theories/hypotheses:

1. If we give plant A acid while we give plant B water then plant B will grow to be taller than plant A.

2. Any two particles of matter attract one another with a force directly proportional to the product of their masses and inversely proportional to the square of the distance between them.

3. If I throw this ball at 2 m/s at an angle of 45 deg, it will take 0.3s to hit the ground.

Theory vs. Hypothesis

- What's the difference?
  - Hypothesis: specific prediction for a single event
    - "If I throw this ball at 2 m/s at an angle of 45 deg, it will take 0.3s to hit the ground"
  - Theory: framework for understanding a larger phenomenon
    - ie: theory of gravity
    - Can derive many hypotheses from a theory

Falsifiability

"No amount of experimentation can ever prove me right, but a single experiment can prove me wrong"
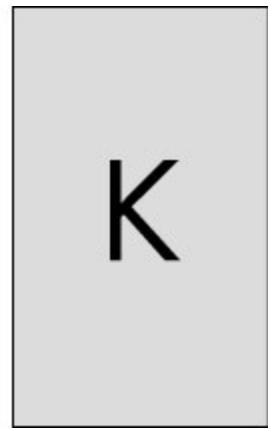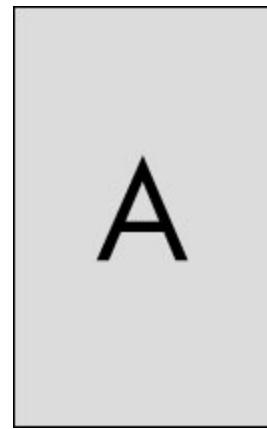
   - Albert Einstein

The importance of falsifying in theory testing

e.g. Hypothesis (derived from some theory): All swans are white.

Observation/Evidence: Observed 100 swans and all were white.
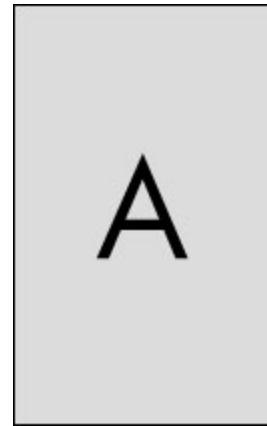

A. Theory proven?

B. Theory not disproven?

If a card has an odd number on one side, then it has a vowel on the other side.
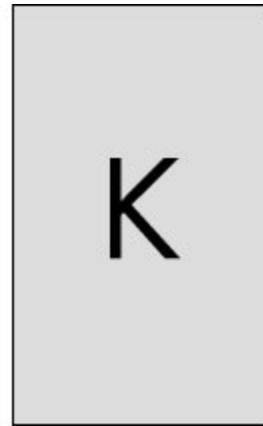
Flip the two best cards to test your hypothesis

If a card has an odd number on one side, then it has a vowel on the other side
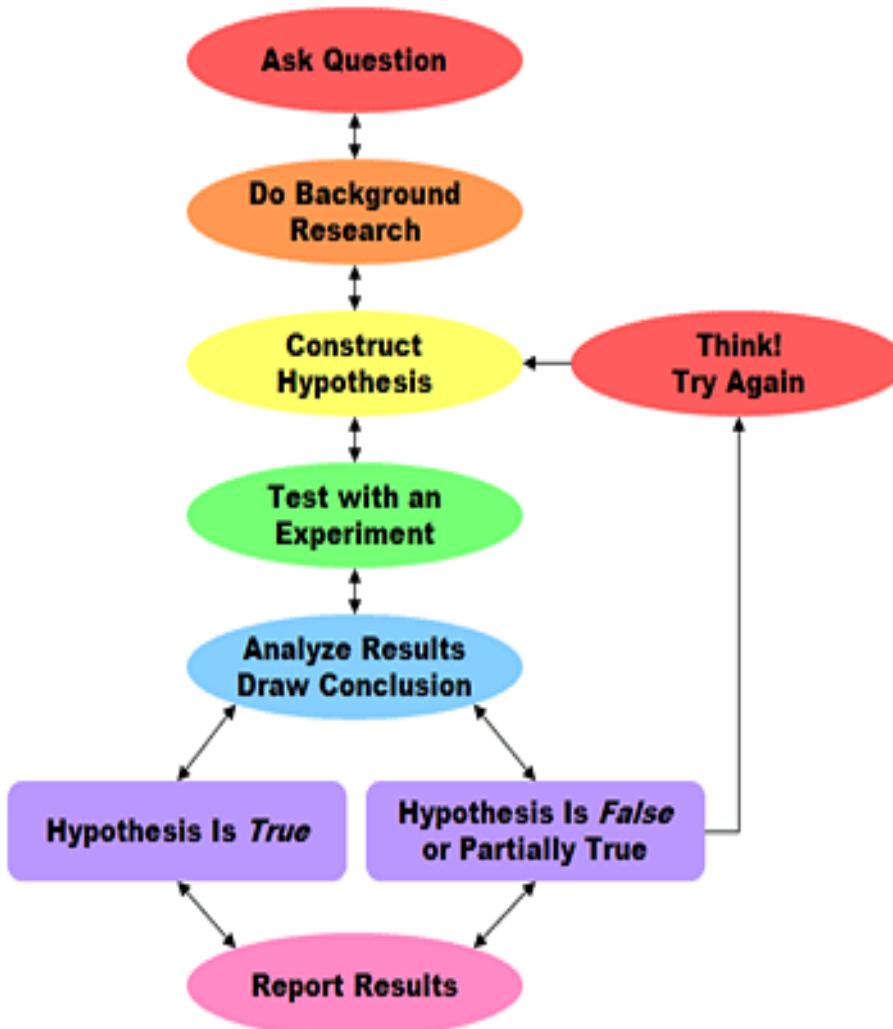
|			
A	K	2	7
[2]	[3]	[D]	[L]

- How many of you chose:
  - A?
  - 2?
  - 7?
  - K?

- Confirmation bias if you pick A
- To falsify the hypothesis, you need to pick cards 7 and K.
- Falsified if you find a consonant on the other side of 7.
- Falsified if you find an odd number on the other side of K.
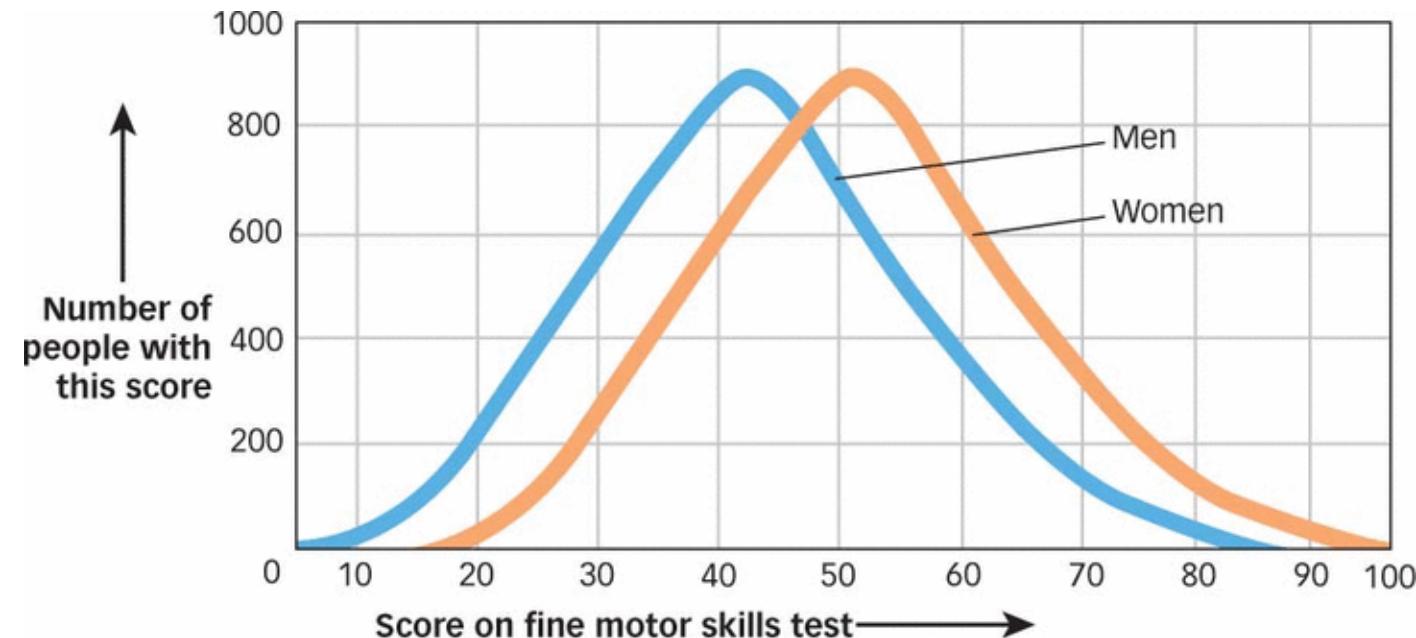
The Scientific Method – A Review

- Provides a logical framework for examining scientific questions.

- Allows other researchers to replicate studies.

Normal Distributions

- Graphic Representations
  - Frequency Distributions
  - **Normal (Gaussian) Distributions**

Hypothesis testing

- Are the means different or do they come from the same population the same mean?
- Consider a test that we *think* is 99% accurate (i.e., the null hypothesis).
- Get 100 people and administer the test. Knowing the truth, let's say we observe that it was actually accurate in 98/100 people (i.e., our sample). The Q is should we believe our "null hypothesis" that the test is 99% accurate?
- How about if our sample shows 97/100 is accurate? Etc etc.

Example

Alessandra designed an experiment where subjects tasted water from four different cups and attempted to identify which cup contained bottled water. Each subject was given three cups that contained regular tap water and one cup that contained bottled water (the order was randomized). She wanted to test if the subjects could do better than simply guessing when identifying the bottled water.

Her hypotheses were $H_0 : p = 0.25$ vs. $H_a : p > 0.25$ (where p is the true likelihood of these subjects identifying the bottled water).

The experiment showed that 20 of the 60 subjects correctly identified the bottle water. Alessandra calculated that the statistic $\hat{p} = \dfrac{20}{60} = 0.\bar{3}$ had an associated P-value of approximately 0.068.

QUESTION A (EXAMPLE 1)

What conclusion should be made using a significance level of $\alpha = 0.05$?

Choose 1 answer:

Ⓐ Fail to reject H_0

Ⓑ Reject H_0 and accept H_a

Ⓒ Accept H_0

https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-categorical-proportions/idea-significance-tests/a/p-value-conclusions