Exam 2 Overview SST 115

November 2025

This exam will be similar to the first exam. There will be some general knowledge questions, a section on probability, a section on hypothesis testing (probably a z-test or t-test), a section that has a confidence interval (probably using a t-distribution). There will be a less mathy section that talks about sampling methods and possibly identifying parts of an experiment.

1 Executive Summary

- Be able to do a z-test or t-test
- Be able to do a confidence interval (z or t)
- Know your 3 assumptions (random, IID, normal or large n)
- Be able to do basic probability calculations (eg dice rolls)
- Give an example of different sampling methods (simple random sampling, stratified, etc..)
- Be able to identify biases in surveys (eg undercoverage)
- Be able to identify parts of an experiment (eg factor)

2 Probability

- Define probability
  - Be comfortable with the basic notation, eg $P(X \geq 2)$
- Empirical vs Theoretical vs Subjective probability
  - Empirical: Based off of data
    - Law of Large Numbers: Empirical probability will approach the "true" probability
  - Theoretical: Based off of a mathematical model
  - Subject: Based off of vibe checks
- Identify a uniform distribution and use to find probabilities
  - Eg what's the probability of rolling a 3 or 4 on a 4-sided dice (called a d4)?

- Use a uniform distribution to find a different distribution
    - Eg what is the probabilities associated with the sum of two d4's? Ie if we roll two 4-sided dice and add them up what is the probabilities of the possible outcomes?
    -

Outcome	2	3	4	5	6	7	8
Probability	1/16	2/16	3/16	4/16	3/16	2/16	1/16

- Marginal vs Conditional Probability
    - Be comfortable calculating these
    - Marginal: The probabilities of an event, ignoring other categories/variables
        - Sampling 100 Americans (men and women) and reporting the proportions (probabilities) an American is in a political party
    - Conditional: The probabilities of an event within a subpopulation.
        - Phrased as Variable A given Variable B
        - Eg sampling 100 Americans (men and women) and reporting the proportions (probabilities) an American WOMAN is in a political party
        - So "women" is our subpopulation and the variable we are looking at is the political party of the women

3 Permutation Test

- Gives exact probabilities and not approximations
    - Eg there is only 3 ways in 100,000 that I could deal a hand to myself that nice
- Won't be tested nearly as in-depth as the z-/t-tests but still worth knowing
- What *is* a null and alternative hypothesis?
    - Alternative hypothesis is what we are trying to show
        - Eg this jury wasn't picked randomly
        - Eg the proportion of adventures that Indiana Jones saves the world is great than .75
    - Null hypothesis is everything the alternative isn't and is usually the not fun "default"
        - Eg the jury was randomly selected
        - Eg the proportion of adventures Inidiana saves the world is less than or equal to .75
        - Eg our new medication isn't different than the old medication
- A permutation test....
    1. writes out all possible outcomes
    2. finds the proportion that fulfill some crtierion

3. Can assign an exact probability we see an (EVENT/Success)
4. Based on if that probability is way low we can call bs

- Eg In a business of 40 people with 5 employees related to the owner, what is the probability all three promotions would go to the owner's relatives if promotions were handed out randomly?
- No, you will not need to know how to calculate the above
- Useful when possible but computationally slow and intensive

4 Normal Distribution

You will not need to code in the exam and relevant information will be given for any values that you'll need for the normal distribution. Focus mostly on...

- Being able to sketch a normal distribution and...
  - indicate the mean and variance for the distribution
  - Shade in parts that correspond to probabilities..
    - Eg P(X < -1) is everything to the left of -1 on the normal distribution
    - Eg P(X > -1) is everything to the right of -1
    - Eg P(-1 < X < 3) is everything between -1 and positive 3
- Understand that many distributions are not actually normal but kinda look normal
  - This is why normal approximations work
  - Eg binomial is X heads in Y coin flips, and is usually bell shaped
- Be able to standardize a variable:

$$\frac{\text{(Observation)} - \text{(Population Mean)}}{\text{(standard deviation)}}$$

- And used standardized variables to compare apples to oranges
  - Eg this apple's standardized weight vs this orange's standardized sugar concentration is valid
  - Standardized values give how many standard deviations above/below the mean the observation is
  - Eg the apple is 2 standard devaitions above the mean weight of apples but this orange is only 1.5 st. dev. above the mean sugar concentration

5 Sampling Distributions

- Be able to define a sampling distribution
  - The long term behavior of a sample statistic, ie if we take a sample and recorded the mean and did it again a million times
  - Standard error: the standard deviation of the sampling distribution (idk why it gets a special name)
- Understand the sampling distribution talks about some summary statistic from the data and not the data itself
  - It won't help you much understand where your next observation will be
- Population vs Sample vs Sampling distributions
  - Population dist is everyone; usually unknown but hopefully looks like the sample
  - Sample dist is the data we collected; only fully known distribution we have
  - Sampling dist is the behvior of the summary statistic; we know it if our assumptions are met
  - Parameter: numeric summary of the population; unknown but hopefully similar to the sample statistic
  - Statistic: numeric summary of our data (our sample); we know this value

6 Z-tests and t-tests

Making this one section since they are sooo similar.

- Z-test is used when the population variance is known (rare)
- t-test is used when we have to estimate the pop. variance (by using our sample variance)
  - degrees of freedom = n-1 for a single mean
  - The t-distribution is wider because we have to estimate two things instead of one (mean and var vs just the mean)
- Assumptions for the sampling distribution
  - Random: Our data was collected randomly if it's a survey or the treatments were randomly assigned for an experiment
  - Independent and Identically Distributed:
    - Independent: Knowing something about one observation doesn't tell you anything about the next observation

      - Genetically related subjects
      - Measuring the same things over time
    - Identically Distributed
      - No a priori belief that one observation will be different than another observation (outside of random noise)
      - Basically there is no lurking variable that might be messing up our analysis
      - Eg Testing high jump distances given different shoes and break participants into two groups....but one group is half made up of geriatrics while the other is the local college's basketball team
  - Population is normal or n is large
    - Don't know population's shape but we do know the sample's shape so use that instead
    - How large n needs to be depends on the shape of the sample distribution, see slide 8 of z-test notes
    - Large n invokes the Central Limit Theorem which says a sampling distribution will approach normality if the sample size is large
- Understand how strong a p-value is classified by
  - See slide 31 of z-tests finished notes
- Using a p-value be comfortable writing a decision
  - You *will* lose points if you say reject/fail to reject
  - Strength of evidence
- Define a p-value
  - Probability we'd see a sample statistic as or more extreme than what we saw, given the null hypothesis is true
- Understand the limits of a p-value
  - Doesn't give the probability the null/alt hypothesis is true
  - A large p-value doesn't mean the null is true

7 Confidence Intervals

- Same assumptions as for hypothesis testing
- Uses either a normal (z) value or a t-dist value; depending if we know our population variance (normal) or not (t-dist)
- Allows us to find a range of reasonable values for where we think the true parameter is

- Define “confidence”
  - It’s the coverage rate for our intervals...90% confidence means about 90% of our confidence intervals will contain the parameter
- Interpret confidence intervals
- Confidence intervals talk about the location of the mean
  - They do NOT talk about where we think the next observation will be (it exists but is called a prediction interval; not discussed)

8 Sampling Methods

First sampling...

- Difference between prospective and retrospective sampling
  - prospective follows things over time
  - retrospective collects info from the past
  - prospective is preferential to retrospective but can be too costly/unreasonable
- Simple Random Sample
  - Give a brief definition
    - everyone is equally likely to be picked
  - Or give a small example
- Stratified Sampling
  - Give a brief definition
    - We do a simple random sampling within each subpopulation
  - Or give a small example
  - Why this might be picked
- Clustered Sampling
  - Give a brief definition
    - We randomly sample tiny tiny subpopulations (cluster) and record all subjects in that cluster
  - Or give a small example
  - Why this might be picked
- Snowball Sampling

  - Brief definition or example
    - One participant introduces you/leads you to the next participant
  - Loses independence in your sample which is bad

- Convenience Sample
  - define or give an example
    - You sample the easiest subjects
    - Eg I text my friends from undergrad who were in the same political club their views on (political thing)....easy to get the example and I can already tell you what the survey results will be
  - Often leads to undercoverage and a lack of independence
  - Is bad

- Know your biases
  - Undercoverage: some subjects will never be chosen for your sample
  - Wording of Questions: questions made too confusing
  - Response Bias: People are taking your survey for a reason and that might be problematic/eg zealous on a politcal topic
  - Non-Response Bias: People refuse to take your survey....is there a reason for this? Eg the survey topic is too "heavy" and makes students uncomfortable or brings up painful feelings

9 Experiments

- What makes something an experiment vs an observational study (assignment of treatments!!!!!!!)
- Identify an experiment's...
  - Factor(s)
  - Levels of that/those factor(s)
  - Treatment: a combinations of all factors
  - Experimental Unit: Thing a treatment is applied to
  - Observational Unit: Thing that we measure and record in our notebooks/spreadsheets
- Understand that the number of experimental units acts as an upper limit on how much information is in the experiment
  - Imagine four cows split into two diet treatment groups
  - We can weigh the cows every hour for a month

  - Observational unit: a cow at a given time point (eg cow 2 weighed 350lbs on Nov 25th at 2pm)
  - Experimental unit: a cow (eg cow 2 is assigned diet A)
  - At the end of the day...we have 4 cows and that's all we have; not a lot of information

- Argue for or against causality in experiments. You do not need to agree with my position so long as you make a rational, well thought out argument

10 Pretentious Questions To Expect

1. I'm testing you on these because these are used as litmus tests to grade people are statisticians. A couple questions cannot encapsulate's someones understanding of statistics but since you will be judged for them I might as well "teach the test"

2. What's a p-value?

3. What's the difference between st. deviation and st. error?
   - st deviation is the square root of variance
   - st error is the st. deviation of the sampling distribution

11 Table

Test Name	σ^2 known?	Assumptions	Relevant Distribution	Test Statistic	Conf. Interval
z-test	yes	(Random), (IID), (Large n or the pop is normal)	Normal	$\frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$	$\bar{x} \pm z_{1-\alpha/2}\sqrt{\sigma^2/n}$
t-test	no	(Random), (IID), (Large n or the pop is normal)	t-distribution w/ df = n-1	$\frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}}$	$\bar{x} \pm t_{df=n-1}\sqrt{s^2/n}$