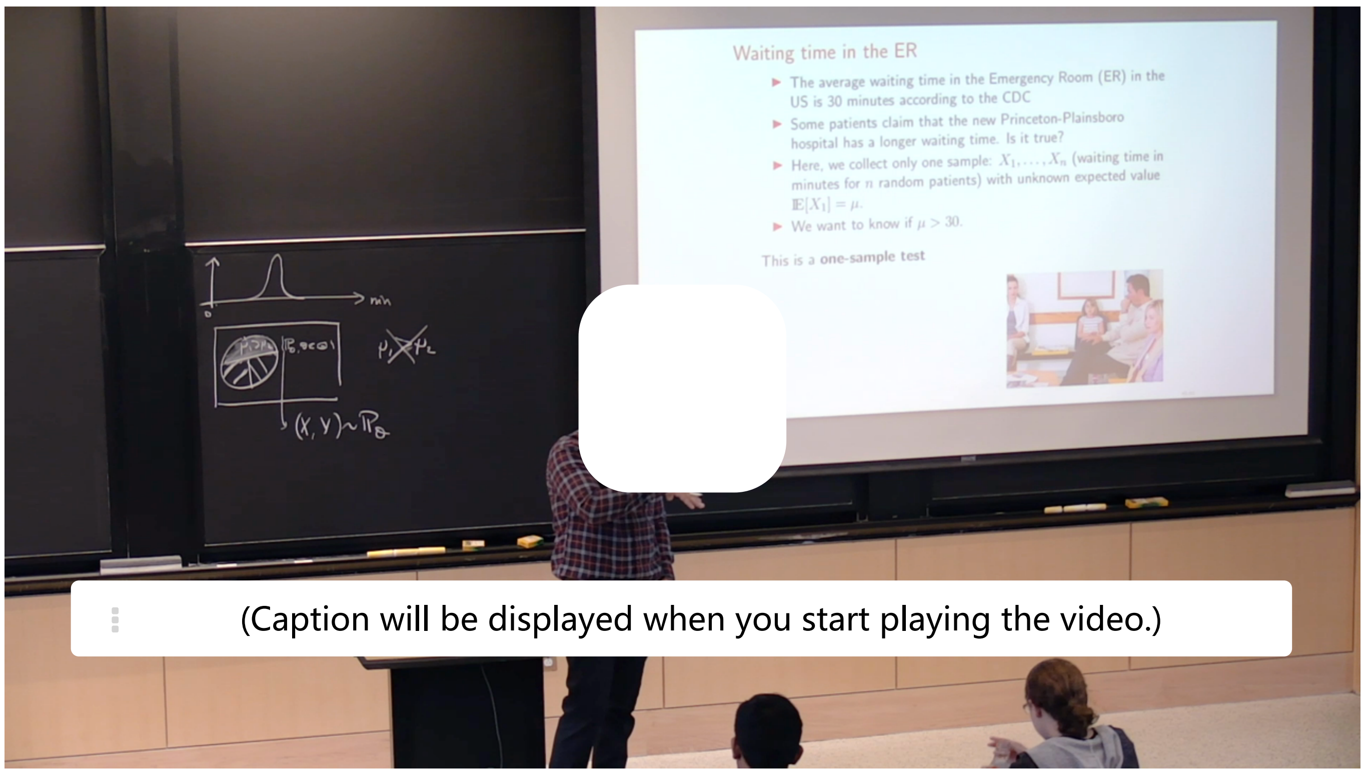
6. Heuristics for One Sample Tests

Heuristics for One Sample Tests



Waiting time in the ER

- The average waiting time in the Emergency Room (ER) in the US is 30 minutes according to the CDC
- Some patients claim that the new Princeton-Plainsboro hospital has a longer waiting time. Is it true?
- Here, we collect only one sample: $X_1, \dots, X_n$ (waiting time in minutes for n random patients) with unknown expected value $\mathbb{E}[X_1] = \mu$.
- We want to know if $\mu > 30$.

This is a **one-sample test**

(Caption will be displayed when you start playing the video.)

Start of transcript. Skip to the end.

So before we go in there, I want to see one last example.

So that was a two sample test.

Let's see an example of a one sample test,

so we have an idea of what it looks like.

And the key thing here is that I actually,

for a one sample test, they typically give you a benchmark.

This 30 here is the reference, so if you

视频
下载视频文件

字幕
下载 SubRip (.srt) file
下载 Text (.txt) file

Another Example: Modeling the Height of the U.S. Population I

1/1 point (graded)

You have access to U.S. census data for the height of individuals from the year 1920. The dataset shows that the average height of the U.S. was 5.5 feet. For simplicity, let's assume that the 1920 dataset included the heights of *all* people residing in the U.S. at that time.

Your goal as a statistician is to provide a response to the **question of interest**:

"**Were people in the U.S. taller in 2018 than in 1920?**".

The company that you work for has limited resources, so you will not be able to survey the entire U.S. population, but you still would like to assess the heights of individuals in the U.S. Therefore, you decide to take the following sampling approach:

Pick 1 million people (with replacement, for simplicity) randomly from the U.S. population and record their heights. Let X_i denote the random variable equal to the height of the i-th person chosen. Assume that any particular individual's height does not influence anyone else's and that there is a common underlying distribution which describes the random variables $X_1, \dots, X_n$.

Which mathematical property of $X_1, \dots, X_n$ most accurately captures all assumptions made in the previous paragraph?

○ $X_1, \dots, X_n$ all have the same distribution, but some of them are correlated.

○ $X_1, \dots, X_n$ are independent, but may not all have the same distribution.

● The random variables $X_1, \dots, X_n$ are iid. ✓

Solution:

We first examine the correct choice and then look at the incorrect choices in order.

- The third choice "The random variables $X_1, \ldots, X_n$ are iid." is correct. Since we are assuming that a person's height will not affect any other peron's height, it makes sense to impose that the X_i's are mutually independent. Moreover, since we stated that there is an underlying distribution describing $X_1, \ldots, X_n$, this is the same as saying that the X_i's are identically distributed.

- The first and second choices "$X_1, \ldots, X_n$ are independent, but may not all have the same distribution." and "$X_1, \ldots, X_n$ all have the same distribution, but some of them are correlated.", respectively, are incorrect because either would contradict the iid assumption.

提交 你已经尝试了1次（总共可以尝试2次）

Answers are displayed within the problem

Modeling the Height of the U.S. Population II

1/1 point (graded)

Continuing from the problem above, your goal is to answer the question of interest

"**Were people in the U.S. taller in 2018 than in 1920?**"

You do so by sampling 10^6 individuals labeled $1, 2, \ldots, 10^6$ chosen randomly from the U.S. population. Let X_i denote the height of the i -th individual. We will treat X_i as a random variable, and use the sample $X_1, \ldots, X_n$ to answer the question of interest.

In addition to the initial modeling assumptions on $X_1, \ldots, X_n$ discussed in the previous problem, we further assume:

- X_i is Gaussian;
- $\mathrm{Var}\,(X_i) = 1.3$.

These assumptions were derived by fitting the data from the 1920 census.

Having established these assumptions, we decide on the following protocol for answering the question of interest. If $\mu = \mathbb{E}\,[X_i] > 5.5$ (and the goal of this lecture is to tackle the question "Is $\mu > 5.5$?"), then we respond by "Yes, the 2018 U.S. population was taller as a whole than the 1920 population". Otherwise, we respond by "No."

Which of the following are **true** statements regarding the two additional assumptions above? (Choose all that apply.)

☑ They place restrictions on the different possible distributions that $X_1, \ldots, X_n$ could follow. ✓

☑ For the purposes of hypothesis testing, it allows us to interpret the question of interest as a very specific mathematical question about the mean of X_i. ✓

✓

Solution:

We examine the choices in order.

- "They place restrictions on the different possible distributions that $X_1, \ldots, X_n$ could follow." is correct. There are many possible distributions that $X_1, \ldots, X_n$ could follow, but we have specifically assumed that $X_1, \ldots, X_n \overset{iid}{\sim} N\,(\mu, 1.3)$ where the parameter μ is unknown.

- "For the purposes of hypothesis testing, it allows us to interpret the question of interest as a very specific mathematical question about the mean of X_1." is correct. Originally, the question "**Were people in the U.S. taller in 1920 or 2018?**" is not precise enough to be well-posed mathematically. However, in making the above assumptions, we were able to focus on some specific property of $X_1, \ldots, X_n$ that can be rigorously tested. The new, more specific question that we have to answer is now:

  "Is the true, unknown parameter μ that describes the mean height of the 2018 U.S. population larger than 5.5 or smaller than 5.5?"

  Remark: We will not be able to answer this question directly because, from practicality constraints, we cannot sample the *entire* U.S. population. Rather, we will use our sample of 1 million individuals to statistically infer, with quantified error, what the answer to the above question should be.

提交 你已经尝试了1次（总共可以尝试1次）

Answers are displayed within the problem

Certainty of a One-Sample Hypothesis Test

1/1 point (graded)
As above, the question of interest is "**Were people in the U.S. taller in 1920 or 2018?**".

As above, you decided to answer this question using the following strategy and assumptions:

- Sample 1 million individuals labeled $1, 2, \ldots, 10^6$ randomly from the 2018 U.S. population.
- Model the height of the i-th individual as a random variable X_i and make the assumption, based on 1920 data, that $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1.3)$ where μ is an unknown parameter.

This allowed us to specify a precise way to answer the initial question of interest:

- If $\mu = \mathbb{E}[X_i] > 5.5$, then you would conclude that the U.S. population is taller in 2018 than it was in 1920 and report "Yes" to the question of interest. Otherwise, you would say "No."

Suppose you access samples $X_1, \ldots, X_{10^6}$ from the 2018 U.S. population and observe that the **sample mean** $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is much larger than 5.5.

Can you conclude with 100 $\%$ certainty that $\mu > 5.5$? (Equivalently, can you know for sure that the answer to the question of interest is "Yes" ?)

Choose the correct answer that also has a correct explanation.

○ Yes, because there are only 10^6 people in the 2018 population to begin with.

○ Yes, because we have carefully chosen the 10^6 individuals so that their sample mean agrees with the true mean μ.

● No, because if, by chance, we chose a 'bad sample' (for example, the million tallest individuals in the U.S.), then the true parameter μ may be much smaller than $\overline{X}_n$ and even much smaller than 5.5. ✓

○ No, because the sample mean $\overline{X}_n$ is a biased estimator of the true mean.

Solution:

We handle the choices in order.

- "Yes, because there are only 10^6 people in the 2018 population to begin with." is incorrect. The U.S. population is currently roughly 325 million, which is significantly larger than the number of samples we will access.
- "Yes, because we have carefully chosen the 10^6 individuals so that their sample mean agrees with the true mean μ." is also incorrect. Since we are sampling individuals randomly from the entire U.S. population, we did not use any specific information about population when choosing our sample.
- "No, because if, by chance, we chose a 'bad sample' (for example, the million tallest individuals in the U.S.), then the true parameter μ may be much smaller than $\overline{X}_n$ and even much smaller than 5.5." is correct. In general, the sample mean may have large fluctuations about the true mean, so it is entirely possible that $\overline{X}_n > 5.5$ while $\mu < 5.5$.
- "No, because the sample mean $\overline{X}_n$ is a biased estimator of the true mean." is not the best choice, because the reason it gives is false. By linearity of expectation, the sample mean is an unbiased estimator of the true mean: $\mu = \mathbb{E}[\overline{X}_n]$.

Remark: In general, it is not possible to answer hypothesis testing questions with 100 $\%$ certainty. However, you will see later in this lecture how to quantify this inherent uncertainty.

提交 你已经尝试了1次（总共可以尝试2次）

Answers are displayed within the problem

Aside: Accessing a Global Data-Set

1/1 point (graded)

As above,the **question of interest** is "**Were people in the U.S. taller in 2018 than in 1920?**".

In the problem, you consider the following two approaches to answer this question.

Approach 1:

- Access the entire 2018 U.S. population of ≈ 325 million people.
- Compute the average μ of the entire data set.
- If $\mu > 5.5$ ", then the answer to the question of interest is "Yes". Otherwise, the answer is "No".

Approach 2:

- Sample 10^6 people labeled $1, 2, \ldots, 10^6$ at random from the 2018 U.S. population.
- Model the heights of the i-th individual as a random variable X_i and make the assumption, based on 1920 data, that $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1.3)$ where μ is an unknown parameter.
- If $\mu = \mathbb{E}[X_1] > 5.5$, then you would conclude that the U.S. population from 2018 is taller as a whole than the 1920 population and report "Yes" to the question of interest. Otherwise, you would say "No."

Which of the following is a potential **disadvantage** of using **Approach 1** vs. **Approach 2**?

○ In Approach 1, we obtain μ exactly.

● In Approach 1, we have to invest the time, money, and overall resources to assess the heights of the entire 2018 U.S. population of ≈ 325 million people. ✓

○ In Approach 1, we are not working with a restricted data set (e.g. a limited sample of the population), so we don't have to worry about errors stemming from choosing a 'bad sample' (for example, a sample consisting entirely of outliers)

Solution:

We handle the choices in order.

- "In Approach 1, we obtain μ exactly." is incorrect because this is an **advantage** of Approach 1. If we could determine μ with absolute certainty, then we would be able to avoid doing any statistical modeling and could respond to the question of interest with complete confidence.
- "In Approach 1, we have to invest the time, money, and overall resources to assess the heights of the entire 2018 U.S. population of ≈ 325 million people." is correct. The problem with Approach 1 is that in some cases, it may not be practically feasible to gather all of the data needed to carry out this approach. In such cases, we must resort to statistical modeling to use a limited data set to make inference on these types of questions of interest.
- "In Approach 1, we are not working with a limited data set (e.g. a limited sample of the population), so we don't have to worry about errors stemming from choosing a 'bad sample' (for example, a sample consisting entirely of outliers)" is incorrect because this is an *advantage* of using the entire data set. The previous problem explores the inherent uncertainty that comes in when using a limited sample to answer a hypothesis testing question.

Remark: While for this problem, computational efficiency was the main disadvantage of Approach 1, in real-life examples, there is an even more basic issue with such an approach. Practically speaking, it may be **impossible** to get access to the entire data-set. In such situations, we would need to resort to something more similar to Approach 2. In this case, we sample n individuals from a much larger population and use their statistical information to make inference about statistical properties of the entire population. This will be our approach in tackling hypothesis testing questions.

提交 你已经尝试了1次（总共可以尝试2次）

Answers are displayed within the problem

Hypothesis Testing vs. Parameter Estimation (Optional)

0 points possible (ungraded)

As above, your goal is to answer the question of interest "**Were people in the U.S. taller in 2018 than in 1920?**".

As in previous problems, you know that in 1920, the heights of the U.S. population were distributed (approximately) like a Gaussian with mean 5.5 and variance 1.3. In addition to imposing that $X_1, \ldots, X_{10^6}$ are iid, you also made the assumptions that

- the heights $X_1, \ldots, X_{10^6}$ 2018 are also distributed like a Gaussian, and
- the variance of X_1 is 1.3

Since we made no assumptions about the mean $\mu := \mathbb{E}[X_1]$, we will treat μ as an unknown parameter.

The goal of this unit is to learn how to answer questions similar to the following:
Is $\mu > 5.5$, or is $\mu \leq 5.5$?

This is a basic example of a **hypothesis testing** question.

Which of the following are **true statements** regarding **hypothesis testing** as exemplified above and **parameter estimation** as discussed in previous lectures?
(Choose all that apply.)

☑ In the above hypothesis testing set-up and in the models in the previous lectures on parameter estimation, we make the assumption that our data is iid from some unknown distribution. ✓

☑ When carrying out parameter estimation, we are interested in coming up with an estimator $\hat{\mu}$ that we want to be close to the true parameter μ. ✓

☑ When performing hypothesis testing (as above), we are **not** necessarily interested in finding an estimator for μ. Rather, our goal is to decide whether or not the true parameter μ lies in a certain region. ✓

☐ When performing hypothesis testing, our main goal is to come up with a good approximation of the true parameter.

✓

Solution:

We examine the choices in order.

- "In the above hypothesis testing set-up and in the models considered in the previous unit on parameter estimation, we make the assumption that our data is iid from some unknown distribution." is correct. In the parameter estimation unit, for all statistical models we assumed that our sample consisted of iid random variables. In the U.S. heights example, we have also made the assumption that the heights $X_1, \ldots, X_n$ are iid.
- "When carrying out parameter estimation, we are interested in coming up with an estimator $\hat{\mu}$ that we want to be close to the true parameter μ." is correct. The main goal of parameter estimation is to come up with some approximation for the unknown true parameter using the sample $X_1, \ldots, X_n$.
- "When performing hypothesis testing (as above), we are **not** necessarily interested in finding an estimator for μ. Rather, our goal is to decide find out if mu has some particular property (for example, whether or not mu lies in a certain region)." is correct. The question stated above is: **Is $\mu > 5.5$ OR is $\mu \leq 5.5$?**.To answer this question, we do not necessarily need to come up with an estimator for the true parameter. It would be enough to decide whether or not μ lies in some particular region (in this case, the interval $(5.5, \infty)$.
- "When performing hypothesis testing, our main goal is to come up with a good approximation of the true parameter." is incorrect. As elaborated upon in the previous bullet, this is not the goal of hypothesis testing. Rather, we want to decide if the true parameter has some particular property (e.g. whether it lies in a particular region or not).

提交 你已经尝试了2次（总共可以尝试3次）

Answers are displayed within the problem