Lecture 20: Hypothesis Tests for a Mean with a Known Standard Deviation

Chapter 14

Corinne Riddell (Instructors Alan Hubbard and Tomer Altman)

October 16, 2024

Learning objectives for today

- What is a hypothesis test
- How to conduct a hypothesis test for the mean when the population standard deviation (SD) is known
  - (We note that pretending we know the SD is a toy problem for the purposes of teaching, but we will generalize it to the more practical situation where the SD must be estimated from the data)
  - One-sided hypothesis testing
  - Two-sided hypothesis testing
- Comparing hypothesis testing with confidence intervals

Readings

- Chapter 14 of Baldi and Moore

Last lecture

- We covered how to make a confidence interval (CI) around the sample mean and how to interpret this interval
- Today, we learn a different approach to statistical inference, which is a statistical test against a so-called null hypothesis resulting in p-values
- Statisticians debate about the relative benefits of reporting p-values versus CIs (though both are often reported)
- In public health research, CIs are favored

Hypothesis testing

- Hypothesis testing is the process of using statistics to say how likely something is to be true under certain assumptions
- Step 1: Specify the parameter of interest that addresses the research questions
- Step 2: Specify the null hypothesis, H_0 (pronounced "H naught") of that parameter
- Step 3: Specify the alternative hypothesis, H_A
- Step 4: Set the significance level α
- Step 5: Calculate the test statistic and p-value based upon the null sampling distribution implied by the null hypothesis
- Step 6: Write a conclusion

Example: Inorganic phosphorus levels in the elderly

Levels of inorganic phosphorus in the blood are known to vary among adults following a Normal distribution with the mean $\mu = 1.2$ mmol/l (millimoles per liter) and standard deviation $\sigma = 0.1$ mmol/l.

A study examined inorganic phosphorus in older individuals **to see if its lower among older adults.** Here are the data from 12 people between the ages of 75 and 79:

```
phos <- c(1.26, 1.39, 1.00, 1.00, 1.00, 1.00,
          0.87, 1.23, 1.19, 1.29, 1.03, 1.18)

known_sigma <- 0.1
```

- So the question is: Is there evidence that levels of inorganic phosphorus are lower among older adults?
- We have data on 12 elderly patients and can use this to assess the evidence when $n = 12$

Example: Inorganic phosphorus levels in the elderly

Step 1: Specify the parameter of interest

Often, one uses the mean level to investigate the distribution, so in this case we proposed the mean level of inorganic phoshporous as the parameter of interest.

Step 2: Specify the null hypothesis, H_0

- The null hypothesis assumes that the data is sampled from a distribution for which the underlying mean is indeed equal to 1.2. We will look to see how much evidence there is against the null hypothesis (conversely, how much evidence in favor of the alternative hypothesis).
- Using notation, we write the null hypothesis like this: $H_0 : \mu = 1.2$
- Generally, the null hypothesis is the hypothesis that things are the same

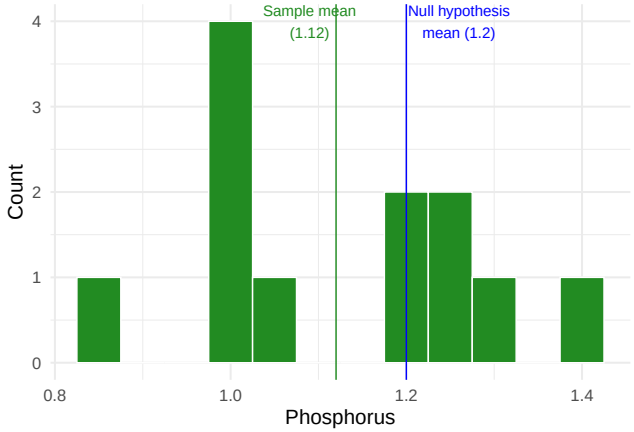
Step 3: Specify the alternative hypothesis, H_A

- Re-read the question. We want to know if the average for elderly patients is lower than the overall average. This dictates the direction of the alternative hypothesis.
- Using notation, we write the alternate hypothesis like this: H_A: $\mu < 1.2$
- This is an example of a **one-sided** alternative hypothesis. The question is specifically interested in knowing if inorganic phosphorus is lower in this population.

Before we conduct the other steps of the test, let's plot the data along with the sample mean $\bar{x}$ and the expected mean under the null hypothesis.

Descriptives

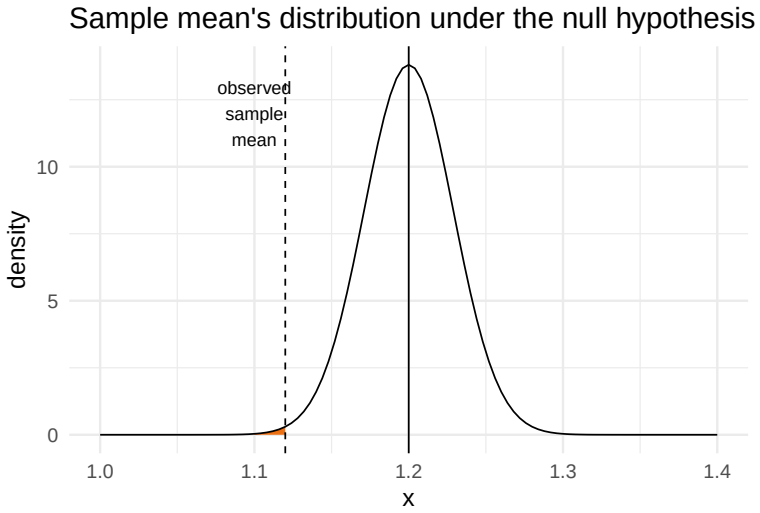
- Does the sample mean *really* differ from the null hypothesis mean?
- Does it look like our data came from a population where $N(\mu = 1.2, \sigma = 0.1)$?



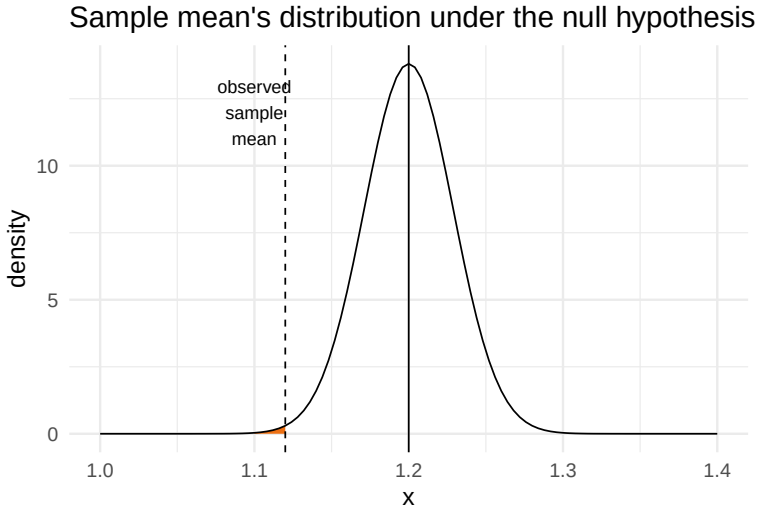
Hypothesis testing

- To conduct the hypothesis test, we first assume the null hypothesis is true

- If the null hypothesis is true, then we know the sampling distribution for the mean, it would have the distribution $\bar{x} \sim N(\mu = 1.2, \sigma = 0.1/\sqrt{12} = 0.0289)$
- So, we can sketch that distribution and add a vertical line where we observe our sample mean $\bar{x} = 1.12$. Then we ask "Assuming the null distribution, how likely are we to observe an $\bar{x}$ as extreme as or more extreme than the one we saw?"



Hypothesis testing



- This probability – of observing an $\bar{x}$ as "extreme or more extreme" than the one we got *if the data are distributed according to the null distribution* – is called the **p-value**.
- **If this probability is very small**, it means that if H_0 is true, there is a very small chance of seeing an $\bar{x}$ of the magnitude that we observed. This would provide **evidence against H_0, in favor of H_A**
- **If this probability is very large**, it means that if H_0 is true, it is very possible that we could see this $\bar{x}$. This can be interpreted as **no evidence against H_0**
- Looking at this plot, does the probability look small or large?

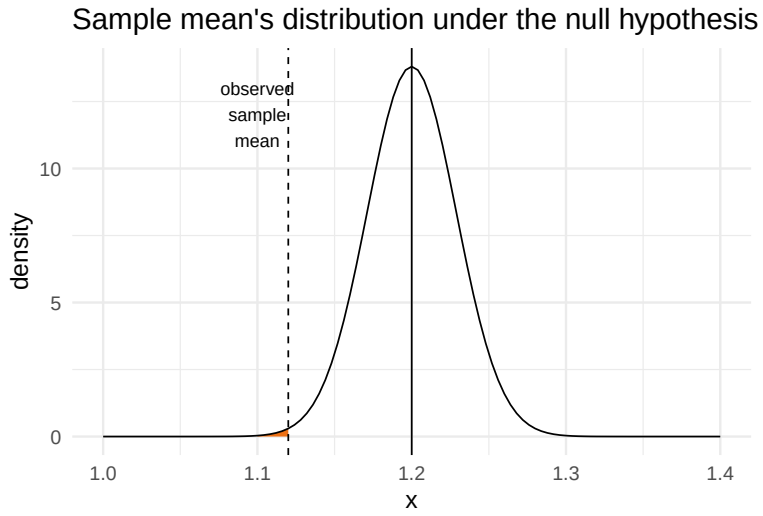
Step 4: Significance level α

- The α level is the desired type I error rate. This is the tolerable probability of false positives (rejection of the null when it is true) in the theoretical context of repeated experiments (optional).
- Where do we draw the line between "small" and "large" probabilities?

- Traditionally, we draw the line at 5%, or 0.05. This is known as the **significance level**, and we write $\alpha = 0.05$
- The value of α is arbitrary (why not 2%, or 10%, or some other value?)
- In experimental studies, like an RCT, we need to choose a significance level in advance. This prevents us from changing how we interpret the results of the test based on the results we get.
- In observational studies, hypothesis testing is controversial. We will talk more about this soon.
- In class, we will often ask you to calculate the probability, but not always to compare it to 5%. This is because we want you to become familiar with interpreting these probabilities rather than comparing them to an arbitrary cut off.

Step 5: Calculate the test statistic and p-value

What is the probability of observing the sample mean we observed *or lower* if the null hypothesis were true?




```
pnorm(q = 1.12, mean = 1.2, sd = 0.1/sqrt(12))
```

```
## [1] 0.002791808
```

The p-value is equal to 0.0028, or 0.28%.

Step 5: Calculate the test statistic and p-value

- In the previous calculation, we used R to calculate the p-value using the sampling distribution under H_0 directly
- Another way to do this, is to first calculate the z-score and use R to calculate the p-value for the z-score
- The benefit of this approach is that you have a sense of which z-scores are small or large

The z-score for $\bar{x}$ is:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

It equals: $z = \frac{1.12-1.2}{0.1/\sqrt{12}} = -2.771281$

- The z-score in this setting is known as the **one sample z test statistic**.

We can use R to find the probability of observing this z-score or less:

```
pnorm(q = -2.771281, mean = 0, sd = 1)
```

```
## [1] 0.002791811
```

This is the same as the previous calculation.

Step 6: Interpret your findings

Step 6:

- Thus, under the null hypothesis, this is a 0.28% chance of observing a sample mean at least as small as what we saw
- This is a very tiny probability, and suggests that H_0 may not be true and that there is evidence in favor of H_A

Phosphorus example, continued:

Suppose instead that for this example, we chose a sample such that $\bar{x} = 1.17$.

- Steps 1 & 2: What are the null and alternative hypotheses?
  - H_0:
  - H_a:
- Sketch the distribution under the null hypothesis and add a line at $\bar{x}$.
- Step 4: Calculate the one-sample z statistic and the probability of observing an observation of 1.17 or lower.

```
# The original p-value
pnorm(q = 1.12, mean = 1.2, sd = 0.1/sqrt(12))
```

```
## [1] 0.002791808
```

```
# The new one?
pnorm(q = 1.17, mean = 1.2, sd = 0.1/sqrt(12))
```

```
## [1] 0.1493488
```

```
pnorm(q=(1.17-1.2)/(0.1/sqrt(12)),mean=0,sd=1)
```

```
## [1] 0.1493488
```

- Step 6: Do you think you could observe this $\bar{x}$ if the null hypothesis were true?

Definitions

Test Statistic

- Measures how far the sample mean diverges from the null hypothesized mean H_0
- Large values of the statistic show that the sample mean is far from what we would expect if H_0 were true
- The one sample z test statistic is $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

P-value

- The probability, assuming that H_0 is true, that the sample mean would take a value at least as extreme (in the direction of H_a) as that actually observed
- The smaller the p-value, the stronger the evidence against H_0 provided by the data
- We never conclude that either the null hypothesis or alternative hypothesis is true, only that there is no evidence against the null hypothesis or "strong evidence in favor of the alternative hypothesis"

So far:

- We have considered a one-sided H_a
- We computed the p-value using the Normal distribution of the sampling mean, and the standardized Normal distribution for the z-score

- We defined the p-value and introduced the concept of the test statistic
- Next: Two-sided alternative hypotheses

Hypothesis testing with a two-sided hypothesis

Suppose you are interested in the Aspirin content in a sample of pills. You have been told that the population of Aspirin tablets is Normally distributed and has a known standard deviation of 5 mg. Furthermore, you have an SRS of $n = 10$ pills and found that your sample mean equals 326.9 mg. You need to detect if there is evidence that the average Aspirin content in your sample is different from the population average and would be concerned if it appears to be higher or lower. Here:

$$H_0 : \mu = 325mg$$

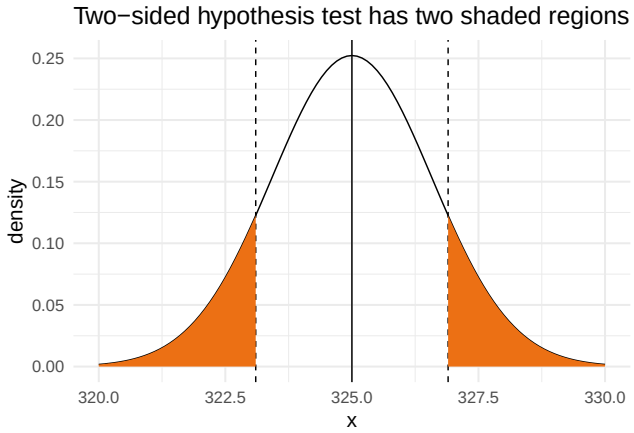
$$H_A : \mu \neq 325mg$$

This is a **two-sided alternative** hypothesis because we're interested in knowing if the sample mean appears to be higher or lower.

Hypothesis testing with a two-sided hypothesis

- The steps for conducting a two-sided test are very similar to the one-sided test
- The only difference is how the probability is calculated. Here, we are interested in knowing the probability of observing the $\bar{x}$ we saw or a more extreme value **in either direction**. How does this change our plot?
- First, calculate the sampling distribution, assuming the null hypothesis is true:
  - $\mu = 325$
  - $\sigma_s = s = \sigma/\sqrt{n} = 5/\sqrt{10} = 1.581139$
- Sketch the sampling distribution and add a vertical line at $\bar{x}$. Shade the region corresponding to H_a. **For a two-sided hypothesis, add another vertical line that is the same distance as $\bar{x}$ is from μ but on the other side of the distribution.**

Two-sided alternative: add vertical lines at $\bar{x}$ and the equivalent distance from the null on the other side of the distribution



P-value calculation for the hypothesis test with a two-sided hypothesis

```
pnorm(q = 326.9, mean = 325, sd = 1.581139, lower.tail = F)
```

```
## [1] 0.1147466
```

```
pnorm(q = 326.9, mean = 325, sd = 1.581139, lower.tail = F)*2
```

```
## [1] 0.2294932
```

- Why do we need to multiply the probability by 2?
- Why do we need to set `lower.tail=F` in this example?

Interpret this probability. Does it provide evidence against the null hypothesis or, to the contrary, does this observation seem to follow under the null hypothesis?

Alternative step: calculate the one sample z test, then the p-value

$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ $z = \frac{326.9-325}{5/\sqrt{10}} = 1.201666$

```
pnorm(q = 1.201666, lower.tail = F)*2
```

```
## [1] 0.229493
```

```
2*pnorm(-1.201666,mean=0,sd=1,lower.tail=T)
```

```
## [1] 0.229493
```

```
pnorm(-1.201666,mean=0,sd=1,lower.tail=T)+pnorm(q = 1.201666, lower.tail = F)
```

```
## [1] 0.229493
```

Conclusion: There is a 22.9% chance of observing a sample mean of this value or more extreme (in either direction) under the null hypothesis, or about a 1 in 5 chance. Thus, this sample mean could be chosen under the null hypothesis and there is no evidence against the null.

Definition: Significance level and statistically significant

- The **significance level**, α, is an arbitrary threshold that can be used when reporting whether a p-value is "statistically significant"
- If the p-value is as small or smaller than α, people say that the data are statistically significant at level α
- Many researchers dislike the use of a significance level because it is a completely arbitrary cutpoint. It is much better to report the exact value of the p-value than to only report if the p-value is statistically significant or not

Definition: Significance level and statistically significant

- For example, both p-value = 0.03 and p-value = 0.004 are "significant at $\alpha = 0.05$", but the latter provides more evidence against the null hypothesis. Thus it is more informative to report the p-value you calculated than to only make a statement regarding statistical significance.
- If a finding is "statistically significant" this does not mean it is "clinically significant", or of a meaningful magnitude. For example, you might find that $\bar{x} = 1.19$ is statistically different from $\mu = 1.2$ (say at $\alpha = 0.05$), but the estimated difference is only 1.2-1.19 = 0.01. This might not be a meaningful difference. Whether the difference is of a meaningful magnitude in practice is not determined by the data, but based on judgement of decision-makers:
  - What is a meaningful reduction in depressive symptoms associated with a new, very expensive treatment vs. a current cheaper treatment?
  - What is a meaningful increase in survival after taking a very expensive cancer treatment? A few days? A few months?
  - What is a meaningful reduction in preterm birth, when the average rate to begin with is 9.8%?

Formalizing what we mean by "test statistic"

The Z test for a population mean: Draw an SRS from a Normal(μ, σ) population, where μ is unknown, but σ is known. A test statistic and a p-value are obtained to test the null hypothesis that μ has a specified value:

H_0:

$$\mu = \mu_0$$

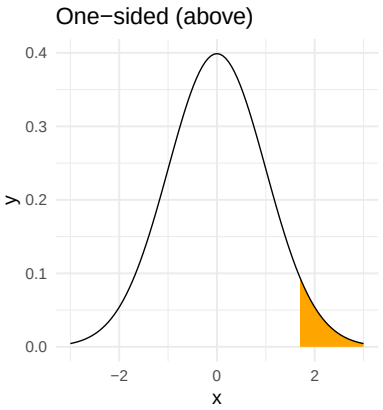
The one-sample z test statistic is:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

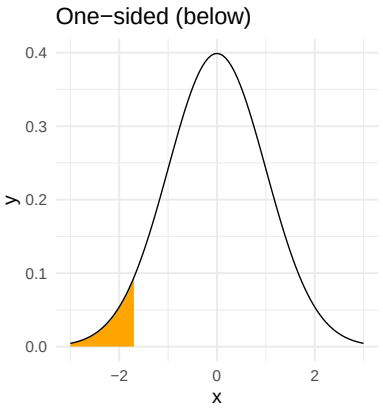
Z test for a population mean (continued)

As the variable Z follows the standard Normal distribution (i.e., N(0,1)), the p-value is represented below for upper tail one-sided, lower tail one-sided, and two-sided tests:

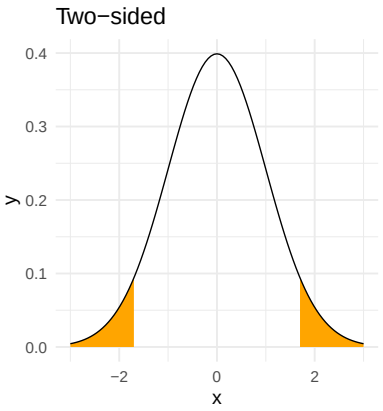
H_a: $\mu > \mu_0$ is $P(Z \geq z)$



H_a: $\mu < \mu_0$ is $P(Z \leq -z)$



H_a: $\mu \neq \mu_0$ is $2 \times P(Z \geq |z|)$



Note that the textbook has an incorrect expression for H_a: $\mu < \mu_0$

Relationship between confidence intervals and test statistics

- A two-sided test statistic at significance level α can be carried out from a confidence interval with confidence level $C = 1 - \alpha$
- If a 95% confidence interval does not contain the null value, this implies that the p-value for the test that $H_0 : \bar{x} = \mu$ has a p-value $< 5\%$
- If a 99% confidence interval does not contain the null value, what does this imply about the p-value for the corresponding two-sided test?

Example: Relationship between confidence intervals and test statistics

Last class, we calculated the 95% CI for the mean height of girls, based on a sample of girls in a Midwestern school district. This CI was from 100.56 to 111.12.

Suppose you wanted to test whether the mean height was different from $H_0 : \mu = 113$ cm. This mean height is outside of the 95% CI, so we know that the p-value corresponding to the two-sided hypothesis test would be $< 5\%$.

Relationship between confidence intervals and parameters

- CIs and p-values provide similar information, because you can deduce directly whether a test will be $<$ 0.05 from a 95% CI
- In fact, a 95% CI can be also derived as the set of parameter values such that two-sided tests fail to reject at $\alpha = 0.05$
- However, if you only know a p-value you cannot derive the CI
- The CI is better because it puts a range around the value of the parameter