Module D: Introductory Inference

Slide Deck D3:

The Theory of Hypothesis Testing

The section in which we are introduced to the theory hypothesis testing. Here, we will see three types of hypotheses, the meaning of the p-value, and why alpha is specified by the researcher.

Today's Objectives

By the end of this slidedeck, you should

- identify the research, null, and alternative hypotheses
- calculate the p-value for a given alternative hypothesis
- properly interpret the p-value

The Theory

The theory behind hypothesis testing is

- State the research hypothesis and the null hypothesis
- Determine how much the data support the null hypothesis:
  - Determine the parameter tested
  - Determine appropriate statistic
  - Determine distribution of that statistic assuming the null hypothesis is true
  - Determine how likely it is to observe the statistic (data) *if* the null hypothesis is true
- Interpret that level of support

Definitions: Hypotheses

Definition (Hypothesis)

A **hypothesis** is a testable claim about reality.

Since it is a claim about reality, it concerns some aspect of the population (a parameter). The usual parameters hypothesized about at this level are the mean μ, proportion p, and variance σ^2. BUT we can hypothesize about *any* aspect of the population.

The "generic" parameter is θ.

Since it is a claim about reality, it separates all possible realities into those that are consistent with the hypothesis and those that are incompatible with it.

Definitions: Hypotheses

The most important hypothesis is the one made by the researcher:

Definition (Research Hypothesis)

A **research hypothesis** is a testable claim about reality made by the scientist.

This is the claim that the statistician must eventually come to a conclusion about.

Because we are using statistics to test this hypothesis, we create two statistics-specific hypotheses. These two hypotheses divide all possible realities into two groups, those that are compatible with the research hypothesis and those that are not.

Definitions: Hypotheses

H_R	H_0	H_A
$\theta < \theta_0$	$\theta \geq \theta_0$	$\theta < \theta_0$
$\theta = \theta_0$	$\theta = \theta_0$	$\theta \neq \theta_0$
$\theta > \theta_0$	$\theta \leq \theta_0$	$\theta > \theta_0$
$\theta \leq \theta_0$	$\theta \leq \theta_0$	$\theta > \theta_0$
$\theta \neq \theta_0$	$\theta = \theta_0$	$\theta \neq \theta_0$
$\theta \geq \theta_0$	$\theta \geq \theta_0$	$\theta < \theta_0$

Table: A listing of all possible research hypotheses and their corresponding null and alternative hypotheses.

The symbol θ is a generic symbol for a parameter. It can represent μ, p, σ^2, or *any other* parameter you can imagine.

The symbol θ_0 represents the value claimed by the researcher. It is a number.

Do not memorize this table. Learn what it says about the *relationships* between the research and other hypotheses... and between the null and alternative hypotheses.

What do all of the null hypotheses have in common?
(Why would that be a requirement?) What is the relationship between the null and alternative hypotheses?
(Why does this make sense?)

Definitions: The p-value

Definition (p-value)

The **p-value** is the probability of observing a test statistic this extreme, or more so, *given* the null hypothesis is true.

Definition (p-value)

The **p-value** is the probability of observing data this extreme, or more so, given the null hypothesis is true.

Definition (p-value)

The **p-value** is the amount of support in the data for the null hypothesis.

Definitions: Decision Theory

We now have two of the three parts of the "hypothesis-testing theory" laid out earlier. The last part is making a decision about the research hypothesis based on the data.

Do the data sufficiently support the research hypothesis?

Note that the statement is binary: Yes or No.

Since the decision is binary, we need to determine a 'cut-off' point between what sufficiently supports the research hypothesis and what does not. This threshold is referred to as "the alpha-value."

If the level of support is less than alpha, we reject the null hypothesis. Otherwise, we do not (we fail to reject).

Definitions: Decision Theory

Definition (alpha-value)

The **alpha value** is the Type I error rate claimed by the statistician.

Definition (Type I error)

A **Type I error** occurs when the researcher rejects a true null hypothesis.

Definition (Type I error rate)

The **Type I error rate** is the proportion of the time the researcher commits a Type I error (rejects a true null hypothesis).

Note: The actual Type I error rate may *not* be α. Laboratory Activity F explores this in greater detail.

Definitions: Decision Theory

If there is a Type I error, then there must be (at least) a Type II error.

Definition (Type II error)

A **Type II error** occurs when the researcher fails to reject a false null hypothesis. Its value is symbolized by β.

Both α and β are error rates. So, we would like to minimize both.

- However, decreasing α results in increasing β.
- Furthermore, reducing either to zero results in the other being 100%.
- Fisher suggested that we specify α and let β be whatever it is.

Definitions: Decision Theory (aside)

If there is a Type I error, then there must be (at least) a Type II error. There are, in fact, a couple more (non-standard) types of errors. They do, however, give some more insight into what can go wrong in statistical analysis.

Definition (Type III error)

A **Type III error** occurs when the researcher rejects a false null hypothesis, but for the wrong reason.

Remember: This is not standard terminology. I am including this to give you some insight into where errors can occur and what applied statisticians spend their time investigating.

Definitions: Decision Theory (aside)

Definition (Type III error)

A **Type III error** occurs when the researcher rejects a false null hypothesis, but for the wrong reason.

Causes of Type III errors include:

- wrong test
- aggregation bias
- ecological fallacy
- collinearity among predictors

The key to avoiding Type III errors is to fully understand the statistical tests and their requirements. The wrong test or the wrong interpretation or the wrong assumptions *all* lead to errors.

Exploratory Example CCD1: Coins

Example

I have a Swedish 10 kronor coin. It looks very cool. I wonder, however, if it is biased in favor of getting heads. Using statistical language, my research hypothesis is

$H_R : p > 0.500$

To test this, I flip the coin 100 times and get 58 heads.

Exploratory Example CCD1: Coins

Since the research hypothesis uses "$>$," we know the null uses "$\leq$."

- From the problem set-up, we know that the *number* of heads, X is distributed as

$$X \sim \mathcal{B}in(n = 100,\ p = 0.500)$$

if the coin is fair.

- Also, if the the number of observed heads is too large, we know that the alternative hypothesis is more likely to be correct.

The next step is to calculate the p-value:

The p-value is the probability of observing data this extreme — or more so — given the null hypothesis is true.

Exploratory Example CCD1: Coins

The p-value is the probability of observing data this extreme — or more so — given the null hypothesis is true.

Thus, the p-value is

$$\mathbb{P}\Big[\ X \geq 58 \ \Big|\ X \sim \mathcal{B}in(100,\ 0.500)\ \Big]$$

Here is the calculation of this p-value in `R`:

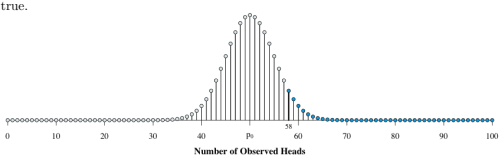
```
1 - pbinom(57, size=100, prob=0.500)
```
= 0.06660531

Thus, the probability of observing data this extreme (or more so), given the null hypothesis is true, is **0.0660531**.

How should we interpret this result??

Exploratory Example CCD1: Coins

This is the probability mass function of the number of heads, X, given the null hypothesis is true.



The dark-shaded values are the probabilities corresponding to values as extreme — or more so — as what was observed, given the distribution of the number of heads.

— Their sum is the p-value.

Exploratory Example CCD2: Cards

Example

I enjoy playing blackjack. The rules are simple, the results are subject only to randomness, and the strategy is constant. In my last visit to my brother, I spent some time at the Mill Casino in North Bend, OR.

However, I lost a lot. I believe that my blackjack dealer is cheating. If everything is fair, then I would expect to have a blackjack 4.75% of the time. I have played $n = 132$ hands and got a blackjack only once. Is this sufficient evidence that the dealer is cheating?

Specifically, using statistical language, my research hypothesis (claim) is

$H_R : p < 0.0475$

Exploratory Example CCD2: Cards

Since the research hypothesis uses "$<$," we know the null uses "$\geq$." From the problem set-up, we know that the *number* of blackjacks, X is distributed as

$$X \sim \mathcal{B}in(n = 132,\ p = 0.0475)$$

if the dealer is not cheating.

- Also, if the the number of observed blackjacks is too small, we know that the alternative hypothesis is more likely to be correct.

So, let's calculate the p-value:

The p-value is the probability of observing data this extreme — or more so — given the null hypothesis is true.

Exploratory Example CCD2: Cards

The p-value is the probability of observing data this extreme — or more so — given the null hypothesis is true.

Thus, the p-value is

$$\mathbb{P}\Big[\ X \leq 1 \ \Big|\ X \sim \mathcal{B}in(132,\ 0.0475)\ \Big]$$

Using `R` to perform the calculation:

$$\begin{aligned} \text{p-value} &= \mathbb{P}\Big[\ X \leq 1 \ \Big|\ X \sim \mathcal{B}in(132,\ 0.0475)\ \Big] \\ &= \texttt{pbinom(1, size=132, prob=0.0475)} \\ &= 0.0123027 \end{aligned}$$

Exploratory Example CCD2: Cards

Thus, the probability of observing data this extreme (or more so), given the null hypothesis is true, is **0.0123027**.

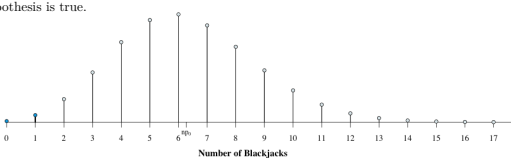
So, how should we interpret this p-value?

Conclusion:

Because the p-value of 0.0123 is less than our alpha-value of 0.05, we should reject the null hypothesis. There is sufficient evidence that the dealer is not being fair.

Exploratory Example CCD2: Cards

This is the probability mass function of the number of blackjacks, X, given the null hypothesis is true.



The dark-shaded values are the probabilities corresponding to values as extreme — or more so — as what was observed, given the distribution of the number of blackjacks.

— Their sum is the p-value.

Exploratory Example CCD3: Dice

Example

I believe that this die is biased against getting a 6. Specifically, using statistical language, my research hypothesis (claim) is

$$H_R : p < 0.1667$$

To test this, I roll the die 100 times and get a 6 a total of nine times.

Exploratory Example CCD3: Dice

Since the research hypothesis uses "<," then the null uses "≥." From the problem set-up, we know that the *number* of sixes, X is distributed as

$$X \sim \mathcal{B}in(n = 100;\ p = 0.1667)$$

if the die is fair.

- Also, if the the number of observed sixes is too small, we know that the alternative hypothesis is more likely to be correct.

So, let's calculate the p-value:

The p-value is the probability of observing data this extreme — or more so — given the null hypothesis is true.

Exploratory Example CCD3: Dice

Thus, it is

$$\mathbb{P}\Big[\ X \leq 9\ \Big|\ X \sim \mathcal{B}in(100,\ 0.1667)\ \Big]$$

The calculation in R is

`pbinom(9, size=100, prob=0.1667)` $= 0.02124964$

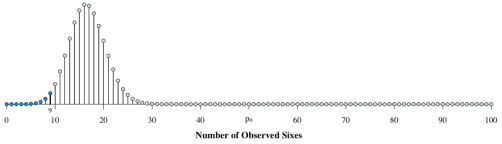
How should we interpret this p-value?

Conclusion:

Because the p-value of 0.0212 is less than our alpha-value of 0.05, we should reject the null hypothesis. There is sufficient evidence that the die is unfair, biased against the 6.

Exploratory Example CCD3: Dice

This is the probability mass function of the number of sixes, X, given the null hypothesis is true.



The dark-shaded values are the probabilities corresponding to values as extreme — or more so — as what was observed, given the distribution of the number of sixes.

— Their sum is the p-value.

Example 1: WacDnalds

Example

WacDnalds claims that the weight of a quarter-pounder hamburger patty (before cooking) is 4 ounces, with a standard deviation of $\sigma = 1$ ounce. In symbols, this is

$$H_R : \mu = 4$$

To test this, we weigh a stack of $n = 25$ patties and find that the *total weight* is only 94 ounces. Does this offer sufficient evidence that WacDnalds is incorrect in their claim?

Example 1: WacDnalds

We seek to calculate

$$\mathbb{P}[\ T \leq 94 \text{ or } T \geq 106\ ]$$

where

$$T \sim \mathcal{N}(100,\ \sigma = 1\sqrt{25})$$

This calculation is

$$\text{p-value} = \texttt{2 * pnorm(94, m=100, s=5)} = 0.2301393$$

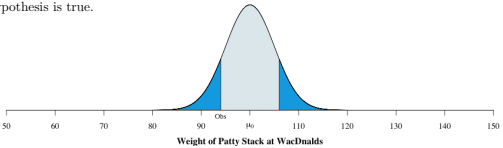
What is the proper interpretation of this calculation?

Conclusion:

Because the p-value of 0.2301 is greater than our alpha-value of 0.05, we cannot reject the null hypothesis. We do not know if the average patty weight is greater than, less than, or exactly 4oz at WacDnalds.

Example 1: WacDnalds

This is the probability mass function of the patty stack weight, T, given the null hypothesis is true.



The dark-shaded region is the probability corresponding to values as extreme — or more so — as what was observed, given the distribution of the total patty weight.

— The area is the p-value.

Example 2: WacDnalds, II

Example

WacDnalds claims that the weight of a quarter-pounder hamburger patty (before cooking) is 4 ounces, with a standard deviation of $\sigma = 1$ ounce. In symbols, this is

$$H_R : \mu = 4$$

To test this, we weigh a stack of $n = 25$ patties and find that the *average* weight is only 3.76 ounces. Does this offer sufficient evidence that WacDnalds is incorrect in their claim?

Note: This is essentially the same problem as the previous one. Here, we are looking at the distribution of the sample mean instead of the sample sum. Thus, we would hope that the conclusion is the same.

Example 2: WacDnalds, II

From the problem description, we need to calculate

$$\mathbb{P}\left[\, \overline{X} \leq 3.76 \text{ or } \overline{X} \geq 4.24 \,\right]$$

where

$$\overline{X} \sim \mathcal{N}(4;\ \sigma = 1/\sqrt{25})$$

This calculates as

$$\text{p-value} = \texttt{2 * pnorm(3.76, m=4, s=0.2)} = 0.2301393$$

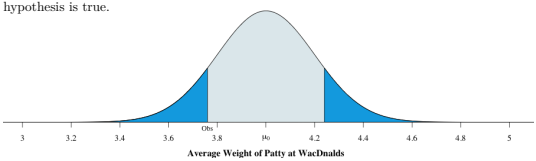
What is the proper interpretation of this calculation?

Conclusion:

Because the p-value of 0.2301 is greater than our alpha-value of 0.05, we cannot reject the null hypothesis. We do not know if the average patty weight is greater than, less than, or exactly 4oz at WacDnalds.

Example 2: WacDnalds, II

This is the probability mass function of the average patty weight, $\overline{X}$, given the null hypothesis is true.



The dark-shaded region is the probability corresponding to values as extreme — or more so — as what was observed, given the distribution of the average patty weight.

— The area is the p-value.

Example 3: MgRonalds

Example

A completely different restaurant, MgRonalds, claims that the weight of its pounder hamburger patty (before cooking) is *at least* 16 ounces, with a standard deviation of $\sigma = 1$ ounce. In symbols, this is

$$H_R : \mu \geq 16$$

To test this, we weigh a stack of $n = 100$ patties and find that the average weight is 15.9 ounces. Does this offer sufficient evidence that MgRonalds is incorrect in their claim?

Example 3: MgRonalds

We seek to calculate

$$\mathbb{P}\left[\ \overline{X} \leq 15.9\ \right]$$

where

$$\overline{X} \sim \mathcal{N}(16;\ \sigma = 1/\sqrt{100})$$

This calculation is

$$\text{p-value} = \texttt{pnorm(15.9, m=16, s=0.10)} = 0.1586553$$

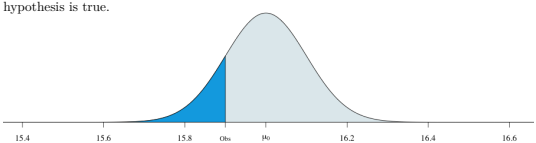
What is the proper interpretation of this calculation?

Conclusion:

Because the p-value of 0.1587 is greater than our alpha-value of 0.05, we cannot reject the null hypothesis. We do not know if the average patty weight is greater than, less than, or exactly 16oz at MgRonalds.

Example 3: MgRonalds

This is the probability mass function of the average patty weight, $\overline{X}$, given the null hypothesis is true.



Average Weight of Patty at MgRonalds

The dark-shaded region is the probability corresponding to values as extreme — or more so — as what was observed, given the distribution of the average patty weight.

— The area is the p-value.

Example 4: MgRonalds, II

Example

MgRonalds claims that the number of Calories in a MgPork is at most 350, with a standard deviation of $\sigma = 50$ Calories. In symbols, this is

$$H_R : \mu \leq 350$$

To test this, we perform a calorimetry test on a stack of $n = 100$ MgPorks and find that the average Calories is 343. Is there sufficient evidence that MgRonalds is incorrect in their claim?

Example 4: MgRonalds, II

We are asked to calculate $\mathbb{P}\left[\ \overline{X} \geq 343\ \right]$, where

$$\overline{X} \sim \mathcal{N}(350,\ \sigma = 50/\sqrt{100})$$

This calculation is

$$\text{p-value} = \texttt{1-pnorm(343, m=350, s=5)} = 0.9192433$$

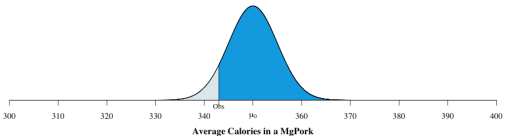
What is the proper interpretation of this calculation?

Conclusion:

Because the p-value of 0.9192 is greater than our alpha-value of 0.05, we cannot reject the null hypothesis. We do not know if the average Calories in a MgPork is greater than, less than, or exactly 350 Cal.

Example 4: MgRonalds, II

This is the probability mass function of the average Calories in the MgPork, $\overline{X}$, given the null hypothesis is true.



The dark-shaded region is the probability corresponding to values as extreme — or more so — as what was observed, given the distribution of the average Calories.

— The area is the p-value.

Today's Objectives

Now that we have concluded this lecture, you should be able to

- identify the research, null, and alternative hypotheses
- calculate the p-value for a given alternative hypothesis
- properly interpret the p-value

Today's R Functions

Here is what we used the following R functions:

- `pnorm(x, m,s)`
- `pbinom(x, size,prob)`

Supplemental Activities

The following activities are currently available from the STAT 200 website to give you some practice in performing hypothesis tests.

- SCA 9a
- SCA 9b
- SCA-11
- SCA-12
- SCA-13
- SCA-21
- SCA-22
- SCA-23

Source: `https://www.kvasaheim.com/courses/stat200/sca/`

In addition to the SCAs, there is **Laboratory Activity F**, which also looks at hypothesis testing.
Source: `https://www.kvasaheim.com/courses/stat200/labs/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding the theory of hypothesis testing:

- Hawkes Learning: Chapters 8 (and 9)
- Intro to Modern Statistics: Chapters 11–15
- R for Starters: Chapters 5 (and 6)

- Wikipedia: Hypothesis Testing