Hypothesis Tests for μ

Making Decisions About the True Population Mean

Confidence Intervals vs Hypothesis Tests

- Both confidence and hypothesis tests are used to make inferences and decisions about the population parameters of interest based on samples. They accomplish the same goals.
- Tests or confidence intervals are used more or less in certain subject areas. A general tendency is that CI's are used more in the sciences and tests more in the Social Sciences. However, CI's are becoming more common in the soc sciences.

Hypothesis Test Demonstration

- The terminology for hypothesis tests is very thick. This demonstration should help you remember what we are doing when things get tough.
- I will flip a coin several times. Keep track of the results.

Demonstration Result

- As I got too many tails in a row, about 4 or 5 usually, most of you started to doubt the fairness of the flipping system.
- When we got way too many tails in a row, you knew there was a problem.
- The data got too unusual or atypical to be coming from a fair 50-50 system. So you rejected your notion of fairness and claimed you were being tricked.

Demonstration Result

- Just remember that when data gets too unusual you will doubt your underlying hypothesis about how the world works, just like in our demonstration.
- P(5 tails in a row) = (1/2)^5 = .031, this is about when most people start to object.
- Keeping this example in your mind will help understand things as we continue.

Four Stages of a Hypothesis Test

- 1) Write hypotheses: null hypothesis Ho, and alternative hypothesis, Ha.
- 2) Calculate the test statistic.
- 3) Calculate the p-value.
- 4) Give a conclusion.

Null Hypothesis

- The null hypothesis is denoted Ho.
- This is pronounced "H – not". If you call it "Hoe", I will chuckle at you because it sounds dorky.
- The null hypothesis is a belief about the world, or population. It is typically what we are trying to show is false.

Alternative Hypothesis

- The alternative hypothesis is denoted Ha.
- This hypothesis is pronounced "H-A" or alternative hypothesis. If you say "Hah", I will chuckle because you will sound dorky.
- The alternative hypothesis is a belief about the population that you believe is true and are probably trying to show is true.

Test Statistic

- The test statistic is a convenient summary of the sample data that can be easily used to make decisions about the hypotheses.
- It is often some kind of transform of the data to a more convenient form, like Z-scores.
- Test statistic is calculated under assumption the Ho is true.

P-Value

- The next step in a test is the p-value.
- A p-value is a probability, so it is a number that is between 0 and 1.
- It is a measure of how consistent the sample data is with the null hypothesis.
- There are two definitions of p-value that will be useful to you:

Definitions

- The p-value is the probability of observing a value of the test statistic as extreme or more extreme than the one observed, if the Ho is true.

- The p-value is the probability of observing data like ours if the null hypothesis is true.

Conclusion

- The conclusion gives our final opinion about what the data tells us about our hypotheses about the world. A good conclusion will contain:
- 1) Data consistency/inconsistency with Ho?
- 2) Evidence to doubt Ho?
- 3) Evidence to support Ha?
- 4) Final conclusion about problem, final punch line. (Visit Eiffel Tower!)

An Example

- Tobacco company claims new cigarette is low-tar, less than 5mg of tar per cigarette on average. We work for consumer organization that tests such claims. From our perspective the hypotheses are:
- Ho: mu <= 5.0mg Ha: mu > 5.0mg
- This is because if we believe mu is above 5mg, we will need to take some action against the tobacco company.

Tar Example

- The data for this experiment was a sample mean of x-bar = 5.5 mg, and a standard deviation of sigma=1.2mg/cig. There were n=36 cigs tested.
- Test statistic is calculated assuming the Ho is true. We use a mu value from the null Ho. The test statistic value is Z =

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{5.5 - 5.0}{1.2/\sqrt{36}} = \frac{.5}{.2} = 2.5$$

Tar P-Value

- The p-value is the probability of observing a value of Z more extreme than 2.5 if the Ho was true.
- What is more extreme comes from the Ha hypothesis, so that our p-value is:
- P(Z > 2.5) because our Ha had a greater than (>) sign.
- From table, p-value= .0062.

Conclusion

- Let's do the recipe for the conclusion:
- 1) The data are unusual and unlikely to occur if the Ho was true, so data are inconsistent with Ho.
- 2) There is evidence to doubt Ho.
- 3) There is evidence to support Ha. (Must be !)
- 4) There is evidence that the mean mu is above 5mg, so evidence against the company claim, and this is evidently not a low-tar cigarette.
- DIAL-LAWYERS !!

Example

- Page 439 in text says Tim Kelley's weight on driver license is 187, but we suspect he has gained weight since last license. This means we have:
- Ho: mu = 187 versus Ha: mu > 187
- Sample data weights from this month: 190.5, 189, 195.5, 187 give x-bar 190.5, n=4, and assume sigma=3, and wts are normally distributed.

Example

- Test Statistic:

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{190.5 - 187}{3/\sqrt{4}} = \frac{3.5}{1.5} = 2.33$$

Example: P-Value

- P-value is the probability of observing a test statistic Z value more extreme than 2.33. Because the Ha was greater than, we need to find the chance of a test statistic greater than 2.33, P-value = $P(Z > 2.33) = .0099$.

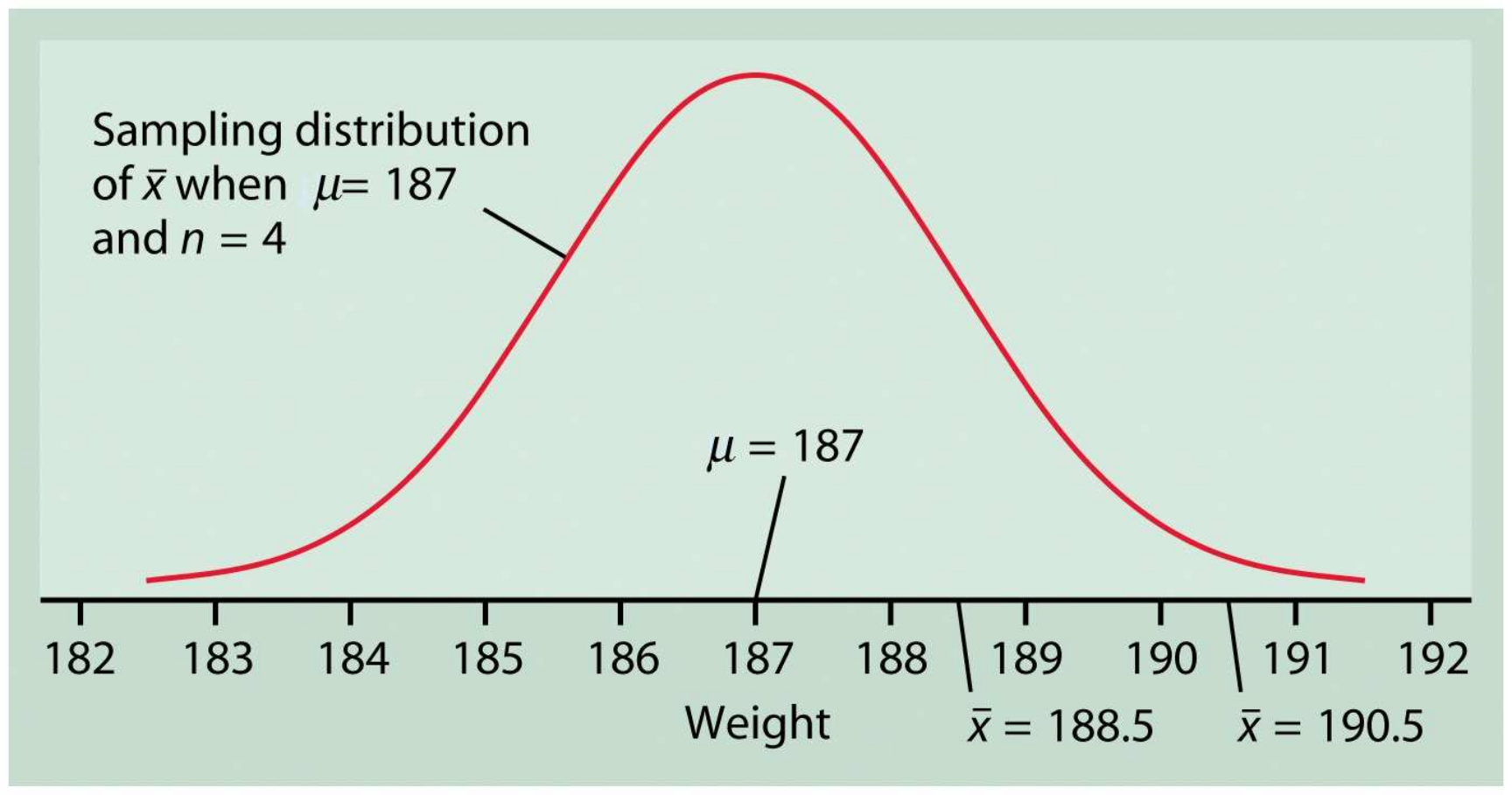
Conclusion

- 1) The data are unlikely to occur if the Ho was true, the data are inconsistent with Ho.
- 2) There is evidence to doubt the Ho.
- 3) There is evidence to support the Ha.
- 4) There is evidence that the average weight mu, is above 187 pounds, and evidence that Tim Kelley has gained weight since the last license.

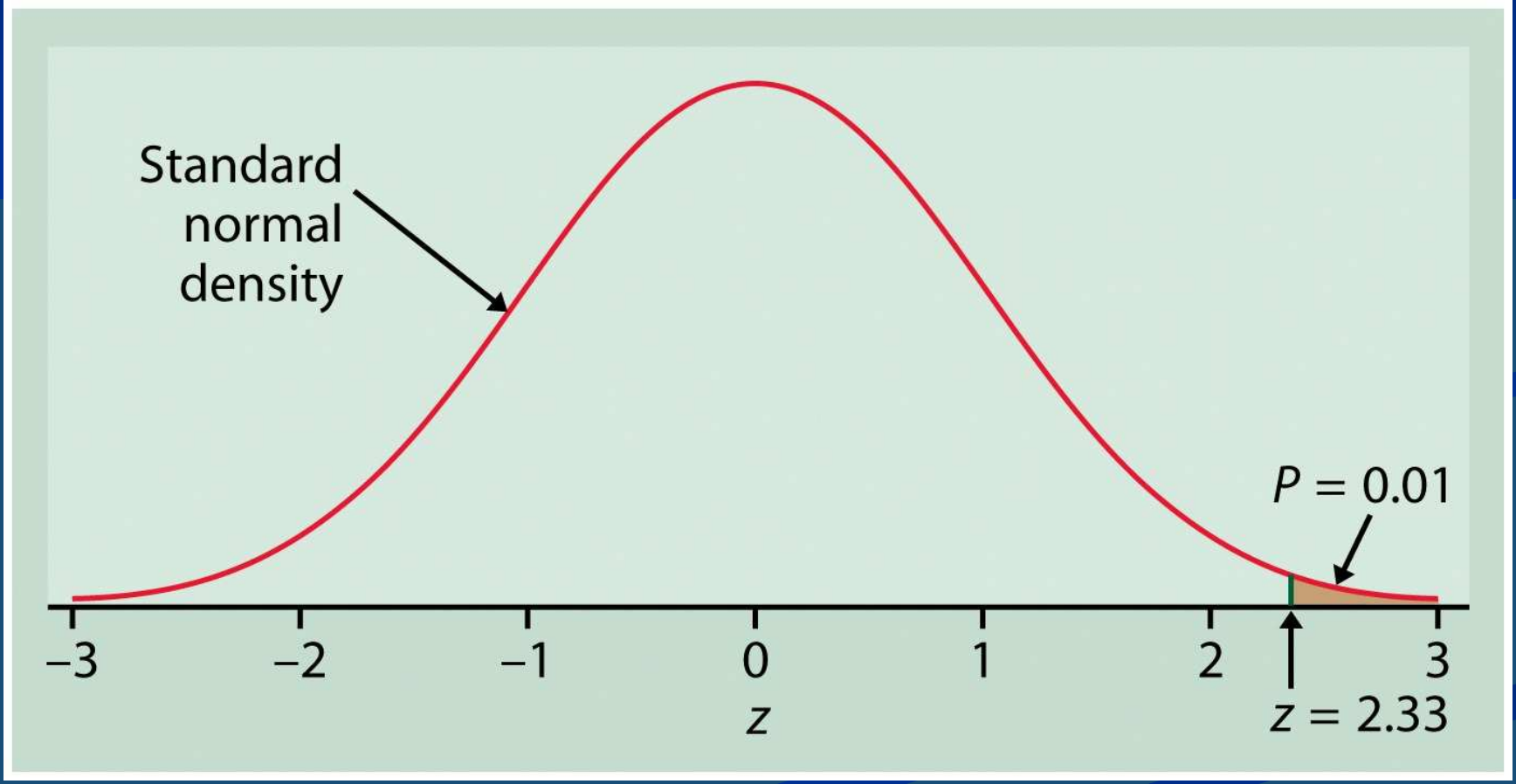
Conclusion Note

- Notice that the first steps are full of jargon, but that the end statement should be clearly understood by anyone reading it. It simply gives the plain English answer to the research question.
- Notice that these examples are just like the coin flipping demonstration in this unit – the data were weird and contradictory to the Ho, and so we begin to doubt the Ho and support Ha.

Weight Example



Weight Example



Psychology Today Example

- Bohrnstedt-Knoke example. Researchers surveyed 2013 readers of Psychology Today and asked them to rate their physical attractiveness to others. Scale: 1 = much less attractive than others, 4= About the same as others, to 7= much more attractive than others.
- Rate your own physical attractiveness on scale of 1 to 7. Turn in slips of paper, NO NAMES PLEASE !

Psychology Today

- Researchers hypothesized that people will rate their own physical attractiveness above average.
- Research hypothesis (Ha): people tend to over-rate their own physical attractiveness.
- This means null hypothesis is that people do not rate themselves above average.
- Ho: mu <= 4, Ha: mu >4.

Psychology Today Example

- Data: x-bar = 4.9, sigma=1.153, n=2013.
- Test Statistic:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{4.9 - 4.0}{1.153 / \sqrt{2013}} = 35.02$$

Psychology Today

- P-value = P(Z > 35.02) = about 0.
- It is basically impossible to get out past 35.02 on the normal scale. Way off the chart !

Psychology Today

- Conclusion:
- 1) Data is unlikely (impossible) to occur if the Ho was true, data inconsistent with the Ho.
- 2) Lots of evidence to doubt the Ho.
- 3) Evidence to support the Ha.
- 4) Evidence that mean rating is above 4 which means evidence that average rating is above average. So evidence people do over-rate themselves, on average. Why do you think?

Stat 1601 Student Self - Ratings

- Class data: x-bar = 4.41, sigma = 1.153, n=79
- Test Statistic: Z= 3.133
- P-Value= .0009
- Evidence stat students rate themselves above average.
- Note that stat students are all above average, and physically very hot, Hot, HOT !
- This is well known on campus. Ask anyone.

Stat 1601 Results

Sample Statistics for rating

N	Mean	Std. Dev.	Std. Error
79	4.41	1.16	0.13

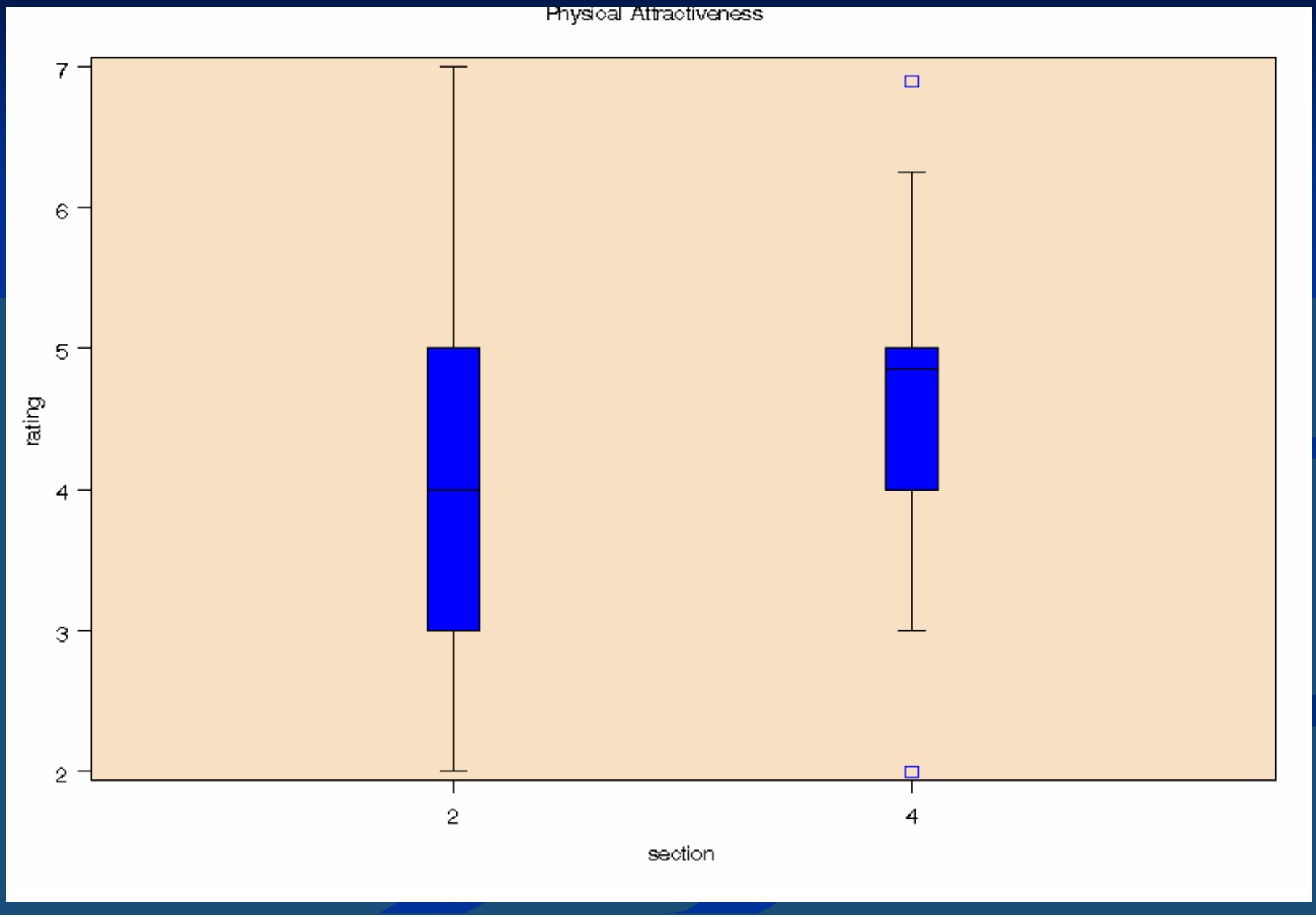
Null hypothesis: Mean of rating <= 4
Alternative: Mean of rating > 4

Z Statistic	Prob > Z
3.133	0.0009

95% Confidence Interval for the Mean

Lower Limit	Upper Limit
4.15	4.66

Stat 1601 Results



Physical Attractiveness

- Notice that if we would use any other value for mu-not in the null hypothesis other than 4, say 2.74, the test statistic becomes even bigger and we still end up doubting the Ho. No matter what we put for mu-not we get the same conclusion.
- Shows that we only need to use values on the border between Ho and Ha.

Hypothesis Test Example

- Ho: 23 star papers, 1 plain
- Ha: 1 star, 23 plain
- Data:
- P-Value=P(1 plain | Ho) = 1/24 = .042
- P-Value=P(1 star | Ho) = 23/24 = .96
- P-Value=P(2 plain | Ho) = 0/ # ways = 0
- A plain or two plains give lots of evidence to doubt the Ho.

Shakespeare Example

- Video clip on new Bill Shakespeare poem.
- Ho: Bill S. wrote poem, Ha: Bill S not write it
- Test Statistic is based on number of new words not used before in a Bill S poem. Value: Z= approx 1.0.
- P-Value = about .15
- Conclusion: 1) Data is consistent with Ho.
- 2) No evidence to doubt Ho.
- 3) No evidence to support Ha.
- 4) No evidence that Bill did not write poem, so no reason to doubt that he wrote it. He might not have written it, but we have no evidence to say he didn't ! Included in complete works.

Stylemetry Authorship

- Recently computer analysis techniques have been used to establish authorship of several disputed documents. An example is the Federalist papers. Although they were published anonymously, the author of 73 of these was determined to be John Jay (5) and the rest divided between Alexander Hamilton and James Madison. There were twelve that were left open to question. Using frequency of usage of the small filler words, they found overwhelming evidence favoring Madison as the author of all twelve disputed papers.
- A second example deals with an unfinished novel by Jane Austin when she died in 1817. A skilled author completed the novel and had it published. Although she duplicated the style she failed to duplicate the subconscious habits of detail. When these habit patterns were examined, the difference was clearly evident.

Authorship Situations

- The Federalist Papers - who wrote the essays whose authorship is unknown?
- The Epistle to the Hebrews - was this written by St Paul (as traditionally believed?), or are modern scholars right to dispute this?
- The Acts of the Apostles - is there evidence to support the traditional ascription of this to the author of Luke's Gospel?
- Plato's Dialogues - can stylometric analysis shed light on their probable order of composition?
- Aristotle's Ethics - did the chapters which are common to the Nicomachean and the Eudemean Ethics originate as part of the Nicomachean Ethics, or as part of the Eudemean Ethics?
- The Shakespeare controversy - how plausible is the claim that the plays traditionally ascribed to Shakespeare were really written by Bacon (or perhaps by Marlowe)?

Radon Detectors

- Manufacturer wants to determine if home radon detectors are reliable enough to sell. Detectors exposed to known source of 105 picocuries.
- Null hypothesis: Ho: mu =105 versus Ha: mu != 105 (not equal to 105). Why not equal?
- If the actual average reading was over 105 it would mean the detectors were over-stating the radon risk, and we would freak out homeowners. Bad idea.

Radon Detectors

- If the average reading was under 105, the detectors would be understating the risk and homeowners would think things were fine when in fact there was much radon risk. Ooops! Not Good!
- This is why the Ha is not equal to 105, we care about values of mu too much or too little compared to 105.

Radon Detectors

- Data: x-bar 104.133, sigma=9, n=12.
- Test Statistic: Z= -.33.
- P-Value=Probability of observing Z value more extreme than -.33. More extreme is more negative than -.33, AND also more positive than .33 because this would also be further from center of 0 than -.33. So…
- P-Value = P(Z<-.33) + P(Z> .33) =
- =2*P(Z > |.33|) = 2(.3707) = big chance

Radon Conclusion

- Conclusion: 1) Data is likely to occur if Ho was true, so data is consistent with Ho.
- 2) No evidence to doubt Ho.
- 3) No evidence to support Ha.
- 4) No evidence to suggest the mean reading is anything other than 105picocuries. So no evidence of a problem with the detectors, so go ahead and sell them.

Caution !

- A big p-value as in the last example (over .1 say), only means we have no reason to bad-mouth the Ho.
- It does NOT mean we have proved the Ho is true !!
- The next example will hopefully illustrate the idea.

OJ

- Consider the OJ trial, the Ho was Ho: OJ Innocent, Ha: OJ Guilty.
- Test Statistic = compilation of evidence
- P-value = big chance, over .1. Jury thought incriminating evidence such as this could easily occur with LA police investigating.
- Big p-value means jury did not have enough evidence to dump Ho, and support Ha, and convict OJ.

OJ Continued

- Ho: OJ Innocent , Ha: OJ Guilty
- So a big p-value means not enough evidence to convict.
- It does not mean we proved that the Ho was true!!! Did the jury prove that OJ was not the killer?
- No, they simply weren't convinced he did it, they did not prove he didn't do it.
- A big p-value does not necessarily mean Ho is true, it might indeed be true, but can't conclude this.

What is a small p-value?

- Most of scientific world uses .05 as the cut-off.
- P-values less than .05 are considered small chances and those above .05 are large.
- This gets ridiculous when p-values = .0501 or .049999. P-values are in shades of gray, not black and white.

What I Expect

- If a p-value is below .05, you need to conclude it is a small chance and evidence to doubt Ho.
- If a p-value is above .1, you need to conclude it is a big chance and no evidence to doubt Ho.
- For p-values between .05 and .1 you have the option of concluding either the chance is big or small. I will accept a correctly argued position.

Is mean body temp 98.6 degrees?

- Ho: mu=98.6 vs Ha: mu != 98.6 degrees
- Z = (98.12 – 98.6)/ (.63/sqrt(93)) = -7.35
- P-Value=P(Z< -7.35) + P(Z>7.35) =
- = 2*P(Z<-7.35) = 2* zero = 0
- Conclusion:

Is Mean Body Temp 98.6 degs?

$$Ho: \mu = 98.6$$

$$Ha: \mu \neq 98.6$$

$$Z = \frac{98.12 - 98.6}{\frac{.63}{\sqrt{93}}} = -7.35$$

$$P - Value = 2 * P(Z < -7.35) = 0$$

Body Temp Conclusion

- The data are very unusual if the Ho was true, the data are very inconsistent with the Ho.
- There is evidence to doubt the Ho.
- There is evidence to support the Ha.
- There is evidence the mean body temp is not 98.6 degrees. Because the sample mean, xbar is less than 98.6, we have evidence the true mean body temp is below 98.6.

Consequences of Incomplete Conclusion

- Dear Unemployed,

Hit the bricks dude! What a pathetic answer to our question. We pay you statisticians big money. The least you could do is be kind enough to answer our questions. If we have evidence to doubt the Ho, fine, but which way does the evidence go you lunkhead!

This is the last time we hire anyone from Podunk State University. Next time we'll recruit hard and hire someone from an elite school like UMM where the stat professors teach the students how to present conclusions!

Sincerely,

Your Boss

PS. Would you like fries with that?

Cancer Therapy Example

- Old treatment gives average remaining life of mu=4.2 years. New therapy experiment gives x-bar=4.5 years, sigma=1.1 years, n=80 patients.
- Believe new therapy prolongs life.
- Ho: mu=4.2 vs Ha: mu > 4.2
- Z= [4.5-4.2] / [1.1 / sqrt(80)] = 2.44
- P-Value=P(Z>2.44) = .0073 (yippeee) =.007

Cancer Conclusion

- 1) The data are unusual if Ho was true, data is inconsistent with the Ho.
- 2) There is evidence to doubt Ho.
- 3) There is evidence to support the Ha.
- 4) There is evidence the mean remaining life is above 4.2 years, and so evidence the new therapy has longer average life than old treatment.