Comments for MEDB 5501, Week 4

Bad quiz question

A research paper computes a p-value of 0.45. How would you interpret this p-value?

1. Strong evidence for the null hypothesis
2. Strong evidence for the alternative hypothesis
3. Little or no evidence for the null hypothesis
4. Little or no evidence for the alternative hypothesis
5. More than one answer above is correct.
6. I do not know the answer.

Speaker notes

Here’s a quiz question that I proposed in an earlier presentation on p-values and confidence intervals. I wrote the responses without thinking, but then realized

“This is an easy mistake to make.”

and kept using this question. None of the answers listed above are correct, and you will see why in just a little bit.

P-values

- Most commonly reported statistic
  - Also sharply criticized
  - Requires a research hypothesis
- Two alternatives
  - Confidence intervals
  - Bayesian analysis
- What to do when no research hypothesis

Speaker notes

P-values are a fundamental tools used in most research papers, but they are coming under increasing attack in the research community. P-values are an inferential tool and require a research hypothesis. Two alternatives are confidence intervals and Bayesian data analysis.

Much research is not inferential and does not have a formal research hypothesis. It is a mistake to force these studies into a hypothesis testing framework. I will cover what you should do when you do not have a formal research hypothesis.

First, you need to remember some basic definitions from your Statistics 101 class.

What is a population?

- Population: a group that you wish to generalize your research results to. it is defined in terms of
  - Demography,
  - Geography,
  - Occupation,
  - Time,
  - Care requirements,
  - Diagnosis,
  - Or some combination of the above.

Speaker notes

A population is a group that you have an interest in. You want to get a better understanding of this group, so you conduct a research study and wish to generalize the results of that study to the population.

In clinical research, a population is almost always a group of people. There are a few exceptions. Sometimes you want to characterize inanimate objects, such as a group of hospitals or a group of medical devices. But let’s keep the focus on people for now.

A population of people is defined in terms of certain characteristics. Usually it is a combination of these characteristics.

Example of a population

All infants born in the state of Missouri during the 1995 calendar year who have one or more visits to the Emergency room during their first year of life.

Speaker notes

Here is an example of a population. It has many of the characteristics described on the previous slide: demography (infants), geography (born in Missouri), time (born in calendar year 1995, during first year of life) and care requirements (one or more ER visits).

Most times the population is so large that it is difficult to get data on all the individuals of that population.

Here, we actually did have access to the data on all 29,637 infants, but most times you would not be so fortunate.

What is a sample?

- Sample: subset of a population.
- Random sample: every person has the same probability of being in the sample.
- Biased sample: Some people have a decreased probability of being in the sample.
  - Always ask “who was left out?”

Speaker notes

A sample is a subset of a population. Because that population of infants was so large, you decided to collect data on a smaller group, a sample of 100 infants, say.

Statistics, according to one definition is the use of data from samples to make inferences about populations. That may be a bit too narrow a definition, but it does characterize quite a bit of what we statisticians do.

A random sample is a special type of sample. It is chosen in a way to insure that every person in the sample has the same probability of being in the sample.

In contrast a biased sample is one where some people in the population have a decreased chance of being in the sample. Often in a biased sample some people in the population are totally excluded.

An example of a biased sample

- A researcher wants to characterize **illicit drug use in teenagers.** She distributes a questionnaire to students attending a local public high school
- (in the U.S. high school is grades 9-12, which is mostly students from ages 14 to 18.)
- Explain how this sample is biased.
- Who has a decreased or even zero probability of being selected.

Type your ideas in the chat box.

Speaker notes

Here is a scenario where a researcher selects a biased sample. I should note here that this is an example specific to the United States. In Italy, you might talk about a survey distributed to the scuola secondaria di secondo grado.

STOP AND GET STUDENT RESPONSES

There are a variety of responses here. The sample does not include home schooled students, students in private schools, students with chronic diseases that force frequent school absences, and students who have dropped out.

Fixing a biased sample

- Redfine your population
  - Not all teenagers,
    - but those attending public high schools.

Speaker notes

Add note

What is a parameter?

- A parameter is a number computed from a sample.
  - Examples
    - Average health care cost associated with the 29,637 children
    - Proportion of these 29,637 children who died in their first year of life.
    - Correlation between gestational age and number of ER visits of these 29,637 children.
  - Designated by Greek letters (μ, π, ρ)

Speaker notes

Add note

What is a statistic?

- A statistic is a number computed from a sample
  - Examples
    - Average health care cost associated with 100 children.
    - Proportion of these 100 children who died in their first year of life.
    - Correlation between genstational age and number of ER visits of these 100 children.
  - Designated by non-Greek letters ($\bar{X}$, $\hat{p}$, r).

Speaker notes

Add note

What is Statistics?

- Statistics
  - The use of information from a sample (a statistic) to make inferences about a population (a parameter)
    - Often a comparison of two populations

Speaker notes

Add note

Break

- What have you just learned?
  - Populations, samples, parameters, statistics
- What is coming next?
  - The null and alternative hypotheses

Speaker notes

Let me pause here for a second. Are there any questions?

What is the null hypothesis?

- The null hypothesis (H_0) is a statement about a parameter.
- It implies no difference, no change, or no relationship.
  - Example
    - $H_0 : \ \mu = C$ (some constant)
    - Hypothesis involving proportions covered later

Speaker notes

Add note

What is the alternative hypothesis?

- The alternative hypothesis (H_1 or H_a) implies a difference, change, or relationship.
  - Examples
    - $H_1 : \ \mu \neq C$

Speaker notes

Add note

Hypothesis in English instead of Greek

- Only statisticians like Greek letters
  - Translate to simple text
  - For mean and proportion comparisons
    - Safer, more effective
  - For correlations
    - Trend, association

Speaker notes

As a researcher, you should always think about your hypothesis in terms of population parameters, but your writing should use text. Translate the Greek letters to English.

If you have a hypothesis involving a mean or proportion, look for comparative words like “safer” or “more effective”. If your hypothesis involves some type of regression model, you should consider terms like “trend” or “association”.

One-sided alternatives

- Examples
  - $H_1: \ \mu > C$ or
  - $H_1: \ \mu < C$
- Changes in only one direction expected
- Changes in opposite direction uninteresting

Speaker notes

Add note

Passive smoking controversy

- EPA meta-analysis of passive smoking
  - Criticized for using a one-sided hypothesis
  - Samet JM, Burke TA. Turning science into junk: the tobacco industry and passive smoking. Am J Public Health. 2001;91(11):1742–1744.

Speaker notes

Available in html format or PDF format.

Consider a study of the effects of second-hand smoke. These studies always use directional alternatives. From what we know about active cigarette smoking is that it increases the risk of cancer and cardiovascular disease. So there is no reason to expect that passive smoke exposure should be any different than active smoking. Maybe it is less toxic, because of dilution and because the smoking coming off a cigarette from one end is different than the smoke coming off the cigarette from the other end. Fair enough, but there is not reason to believe that things are so different that all of a sudden the smoke becomes protective.

Since there is no scientific basis for a protective effect of passive smoking, it makes sense to test that passive smoking has no effect versus it having an increase in bad outcomes compared to the control group. So your null hypothesis is “not harmful” and your alternative is “harmful”. The beneficial hypothesis is lumped into the null hypothesis, but no one would dare claim that passive smoking was protective.

Actually, the tobacco companies did complain that the use of a directional alternative violated the norms of science. They won in a court battle in North Carolina, but lost on appeal.

As another aside, I was involved with prayer study. We planned this study using a one-sided hypothesis (remote prayer has a positive effect on health). The Institutional Review Board suggested changing this to a two-sided hypothesis (remote prayer has either a positive or a negative effect on health). Thankfully, we did not observe an outcome in the opposite tail as that would have been very difficult to explain.

Break

- What have you just learned?
  - The null and alternative hypotheses
- What is coming next?
  - Decision rules, Type I and II errors

Speaker notes

Let me pause here for a second. Are there any questions?

What is a decision rule? (Example)

- $H_0: \ \mu = C$
- $H_1: \ \mu \neq C$
- t = $(\bar{X} - C)$ / se
- Accept H_0 if t is close to zero.
  - $-2 < t < 2$ or
  - $-Z_{\alpha/2} < t < Z_{\alpha/2}$ or
  - $-t_{\alpha/2;n-1} < t < t_{\alpha/2;n-1}$

Speaker notes

Add note

What is a Type I error?

- A Type I error is rejecting the null hypothesis when the null hypothesis is true
  - False positive
  - Example involving drug approval: a Type I error is allowing an ineffective drug onto the market.
- α = P[Type I error]

Speaker notes

In your research, you specify a null hypothesis (typically labeled H0) and an alternative hypothesis (typically labeled Ha, or sometimes H1). By tradition, the null hypothesis corresponds to no change. When you are using Statistics to decide between these two hypothesis, you have to allow for the possibility of error. Actually, if you are using any other procedure, you should still allow for the possibility of error, but we statisticians are the only ones honest enough to admit this.

A Type I error is rejecting the null hypothesis when the null hypothesis is true.

Consider a new drug that we will put on the market if we can show that it is better than a placebo. In this context, H0 would represent the hypothesis that the average improvement (or perhaps the probability of improvement) among all patients taking the new drug is equal to the average improvement (probability of improvement) among all patients taking the placebo. A Type I error would be allowing an ineffective drug onto the market.

Remember that the hypotheses involve population parameters. Population parameters are impossible to compute. So you can only talk about Type I errors in an abstract sense. You will never know for certain if you have made a Type I error.

Alpha is the probability of a Type I error, and alpha is a value that you can compute. In most studies, researchers work hard to keep the probability of a Type I error low, typically at 5%.

What is a Type II error?

- A Type II error is accepting the null hypothesis when the null hypothesis is false.
  - False negative result
  - Usually computed at MCD
  - An example involving drug approval: a Type II error is keeping an effective drug off of the market.
- β = P[Type II error]
- Power = $1 - \beta$

Speaker notes

A Type II error is accepting the null hypothesis when the null hypothesis is false. You should always remember that it is impossible to prove a negative. Some statisticians will emphasize this fact by using the phrase “fail to reject the null hypothesis” in place of “accept the null hypothesis.” The former phrase always strikes me as semantic overkill.

Many studies have small sample sizes that make it difficult to reject the null hypothesis, even when there is a big change in the data. In these situations, a Type II error might be a possible explanation for the negative study results.

Consider a new drug that we will put on the market if we can show that it is better than a placebo. In this context, H0 would represent the hypothesis that the average improvement (or perhaps the probability of improvement) among all patients taking the new drug is equal to the average improvement (probability of improvement) among all patients taking the placebo. A Type II error would be keeping an effective drug off the market.

It bears repeating that population parameters are impossible to compute. So you will never know for certain if you have made a Type I error.

Beta is the probability of a Type II error. Beta is a known quantity. Typically researchers try to keep beta small. 10% is a typical value, though in some settings, a Type II error rate as large as 20% could be tolerated.

Power is defined as 1-beta. I will talk more about power in a little bit.

Break

- What have you just learned?
  - Decision rules, Type I and II errors
- What is coming next?
  - p-values

Speaker notes

Let me pause here for a second. Are there any questions?

What is a p-value?

- Let t = ($\bar{X} - C$) / se
- p-value = Prob of sample result, t, or a result more extreme,
  - **assuming the null hypothesis is true**
- Small p-value, reject H_0
- Large p-value, accept H_0

Speaker notes

A p-value is a measure of how much evidence we have against the null hypothesis.

The smaller the p-value, the more evidence we have against H0.

The p-value is also a measure of how likely we are to get a certain sample result or a result “more extreme,” assuming H0 is true.

The type of hypothesis (right tailed, left tailed or two tailed) will determine what “more extreme” means.

Alternate interpretations

- Consistency between the data and the null
  - Small value, inconsistent
  - Large value, consistent
- Evidence against the null
  - Small, lots of evidence against the null
  - Large, little evidence against the null

Speaker notes

There are two interpretations that I feel are more practical. You can think of the p-value as a measure of consistency between the data and the null hypothesis. A small value implies inconsistency. It is very unlikely that you will get a value like you've seen in your sample or a value more extreme under the assumption that the null hypothesis is true. So you should reject that assumption.

On the other hand if the sample results or anything more extreme has a high probability under the assumption that the null hypothesis is true, then you should feel comfortable accepting that assumption.

I have argued that the p-value is a measure of evidence. Some have called it a poor measure of evidence, but I stand by my interpretation.

If the p-value is small, you have lots of evidence against the null hypothesis. If the p-value is large, you have little or no evidence against the null hypothesis.

What the p-value is not (1/2)

- A p-value is NOT the probability that the null hypothesis is true.
  - P[t or more extreme | null] is different than
  - P[null | t or more extreme]
    - P[null] is nonsensical
    - μ is an unknown constant (no sampling error)

Speaker notes

The p-value is a conditional probability, and you always need to be careful about conditional probabilities. It is a probability about a sample result given an assumption about the population result. It is not a probability about a population result given the sample result. There are two reasons for this.

First, you can't reorder a conditional probability. The probability of A given B is almost never the same as the probability of B given A. The example I give for this is the probability of being happy given that you are rich. That's a pretty high number, I hope you'll agree. There are a few rich people who lead miserable lives, but from everything I've seen, most rich people are pretty darn happy. The reverse of this is the probability of being rich given that you are happy. That number is much smaller. Because although I believe that money can buy happiness, a lot of other things can also buy happiness just as well. It's not quite as easy to find happiness if you're poor, but somehow, a lot of poor people find a way to be happy anyway.

A second reason that you can't reverse the order is that you cannot make a probability statement about population parameters. They are numbers computed from the entire population, and are fixed values. You cannot make a probability statement about something that has no sampling error.

Only numbers computed from a sample (i.e., statistics) have sampling error.

What the p-value is not (2/2)

- Not a measure FOR either hypothesis
  - Little evidence **against** the null $\neq$ lots of evidence **for** the null
- Not very informative if it is large
  - Need a power calculation, or
  - Narrow confidence interval
- Not very helpful for huge data sets

Speaker notes

The p-value is not a measure for either hypothesis. It is always a measure against a particular hypothesis. Now when the p-value is small, you can make a strong statement. We have lots of evidence against the null hypothesis. That translates into lots of evidence in favor of the alternative hypothesis.

When the p-value is large, however, you are in a quandary. Little or no evidence against the null hypothesis is not the same as lots of evidence for the null hypothesis.

It’s possible to have little or no evidence against the null and also have little or no evidence against the alternative. This happens whenever you have a really small sample size combined with a lot of noise.

You can’t prove a negative, so the saying goes. Well, you can prove a negative, but you have to work harder at it. A large p-value by itself is not persuasive, but if you combine it with a power calculation done prior to data collection, that’s pretty good evidence in support of the null hypothesis.

You could also combine a large p-value with a narrow confidence interval to support the null hypothesis. I’ll talk about that more in just a bit.

In general, the p-value is not very helpful for large samples. We’re seeing this more and more. Just about everything pops up as statistically significant with these huge data sets, and you can’t use the p-value to separate the important stuff from the trivial stuff. You need to look instead at the magnitude of the sample estimates and calculate how much uncertainty you can remove in your future predictions.

Pop quiz, revisited

A research paper computes a p-value of 0.45. How would you interpret this p-value?

1. Strong evidence **for** the null
2. Strong evidence **for** the alternative
3. Little or no evidence **for** the null
4. Little or no evidence **for** the alternative
5. More than one answer above is correct.
6. I do not know the answer.

Speaker notes

Here’s that pop quiz again. Take a look at it quickly. Note that the p-value is of evidence against the null hypothesis. So each of the first four responses is wrong.

I wrote this question quickly, so shame, shame on me. But I’ve reproduced the example because it illustrates an important point.

Break

- What have you just learned?
  - p-values
- What is coming next?
  - Criticisms of p-values and hypothesis testing

Speaker notes

Let me pause here for a second. Are there any questions?

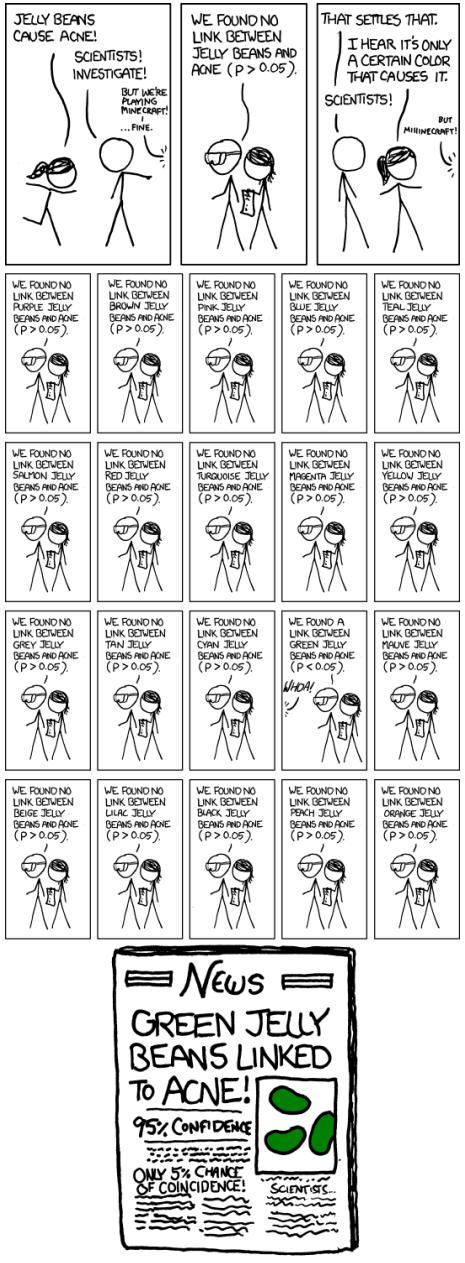


Figure 1: xkcd cartoon about jelly beans and cancer

Speaker notes

This cartoon is impossible to read, but you can find it on the Canvas site or in the readings. Here's a brief run down.

In the first panel, a woman runs up to a man and shouts: Jelly beans cause acne!

The man replies : Scientists! Investigate!

In the second panel, one scientist, holding a clipboard announces: We found no link between jelly beans and acne ($p > 0.05$).

In the third panel, the woman says: I hear it's only a certain color that causes it.

In a bunch of small panels, the scientist with a clipboard reports: We found no link between purple jelly beans and acne ($p > 0.05$).

We found no link between brown jelly beans and acne ($p > 0.05$).

We found no link between pink jelly beans and acne ($p > 0.05$).

The same for blue, teal, salmon, red, and so forth. And then…

We found a link between green jelly beans and acne ($p < 0.05$). An off-screen voice goes: Whoa!

The next six panels show

We found no link between mauve jelly beans and acne ($p > 0.05$).

We found no link between beige jelly beans and acne ($p > 0.05$).

We found no link between lilac jelly beans and acne ($p > 0.05$).

We found no link between black jelly beans and acne ($p > 0.05$).

We found no link between peach jelly beans and acne ($p > 0.05$).

We found no link between orange jelly beans and acne ($p > 0.05$).

At the bottom is a newspaper with the headline: Green Jelly Beans Linked To Acne! 95% Confidence. Only 5% chance of coincidence!

If you are interested in a transcript and a detailed explanation, https://www.explainxkcd.com/wiki/index.php/882:_Significant

What is p-hacking?

- Abuse of the hypothesis testing framework.
  - Run multiple tests on the same outcome
  - Test multiple outcome measures
  - Remove outliers and retest
- Defenses against p-hacking
  - Bonferroni
  - Primary versus secondary
  - Published protocol

Speaker notes

This is an example of p-hacking. You change the testing process to increase the probability of a Type I error (Rejecting the null hypothesis when the null hypothesis is true). This increases the chance of getting a positive result, which you may find desirable, but only by increasing the probability of a false positive result.

Some examples of p-hacking. Run multiple tests on the same outcome measure. Start with the regular t-test, include the t-test that allows for unequal variances, and run two different non-parametric tests, the Wilcoxon-Mann-Whitney test and the sign test. Choose the test with the smallest p-value.

You also might consider multiple outcome measures. Compare the mortality rate, the relapse rate, and the re-hospitalization rate. If any of the three is statistically significant, claim victory.

You could also do this with longitudinal data. Compare pain relief at one hour and at four hours. If you see a difference at one hour, claim that your new medication is faster acting. If you see a difference at four hours, claim that your medication is longer lasting.

You might run a test with the full data set and then with an outlier or two removed. Report for the data set that has the smaller p-value and pretend that this was your original choice all along.

These are only a few of the choices. I don't want to say more because I feel like I'm the devil tempting you.

There are two defenses against p-hacking. Well three if you count being honest. But what I mean is there are two things that you can do that will satisfy others that you are playing fairly.

First, you can adjust your decision rule by using a Bonferroni correction. Bonferroni divides alpha by the number of tests. If you are using three different outcome measures, compare your p-value of 0.0133 instead of 0.05.

Second, you can designate one of your outcome measures as primary. If you achieve statistical significance on your primary outcome, great. The remaining outcome measures are secondary. If you achieve statistical significance on a secondary outcome meausure only, report the results as provisional and requiring independent replication.

You should publish a detailed protocol, either through a clinical trial registry, or now there are journals which accept publications of the research protocols before any data are collected. It's a paper with literature review and methods section, but no results and no discussion section.

Now p-hacking has happened because some people have a skewed view of research. They are interested in using research to promote their own agenda rather than using research to uncover the truth. Perfectly understandable if you are a drug company, but you as an

independent researcher should never try to skew the data. It hurts you and it hurts your patients. You need to adopt a disinterested posture in that you are glad when the research points in one direction and you are glad when it points in the opposite direction, because either way, you know more than you did before and you can treat your patients better because of this knowledge.

Criticisms of hypothesis testing (1 of 4)

- Criticisms of the binary hypothesis
  - Dichotomy is simplistic
  - Point null is never true
  - Cannot prove the null
- Possible remedy
  - H_0 $C - \Delta \leq \ \mu \leq C + \Delta$
  - H_1 $\mu < C - \Delta$ or $\mu > C + \Delta$

Speaker notes

There are many criticisms of hypothesis testing. You need to be aware of these criticisms, but I am not suggesting that you abandon hypothesis testing because of these criticisms.

The first set of criticisms deals with the binary nature of the hypotheses.

I've said many times that all dichotomies are false dichotomies, and I still hold to that. Hypothesis testing is a double dichotomy. You specify only two hypotheses, and you only two choices are accepting the null hypothesis and rejecting the null hypothesis. Shouldn't there be more than two choices?

Let me give an example. I found a couple of articles that studied the safety of vaccines. Now, vaccines are complicated, but let's try to understand the safety issue more clearly. It depends on the benefits of the vaccination combined with the probability that an individual will receive the benefit. Balance that against the harm caused by the vaccine combined with the probability that an individual will experience the harm. So how much harm and how probable does it need to be before you can say that you have a clinically important difference? Complicated, yes, but let's throw in a curve ball. How much do the harms and the probabilities need to be before you warn someone about those harms. How much do the harms and the probabilities need to be before you decide that you shouldn't be using this vaccine? Two very different questions, and two very different thresholds. So why do we force both of them into a single decision point?

Why shouldn't you allow a gray decision? So you could accept the null hypothesis for some values of the sample statistics, reject it for other values, and choose to neither accept nor reject for values intermediate.

Another thing about hypotheses is that a difference of exactly zero never actually occurs in the population. There's no way that, if you averaged a population of all males with particular disease and you averaged a population of all the females with a particular disease that you'd get the two to be exactly the same, even for a disease that has no association with gender. So what is the point of setting up a hypothesis and making a decision about it when you know in your heart of hearts that the null hypothesis is never true.

The other issue with hypothesis testing is that it does not allow you to prove the null hypothesis. If you really wanted to prove the null hypothesis, you have to do all sort of messy gyrations. Wouldn't it be nice to be able to act with the same level of certainty when you accept the null hypothesis as you do when you reject the null hypothesis? The very phrasing that some people use (fail to reject the null hypothesis in place of accept the null hypothesis) shows how convoluted things get.

Many of these criticisms of the binary hypothesis would disappear if we allowed for testing not equality in the null hypothesis but rather testing whether the difference is within some interval. But this is not going to happen anytime soon.

Criticisms of hypothesis testing (2 of 4)

- Criticisms of the p-value
  - Not intuitive, easily misunderstood
  - “results more extreme”
  - Ignores clinical importance
  - Does not measure uncontrolled biases

Speaker notes

The p-value that lies at the heart of hypothesis testing has also been roundly criticized. It seems backwards in more than one way. It is evidence against a hypothesis rather than for a hypothesis. You get lots of evidence against the null hypothesis when the p-value is small and little or no evidence when the p-value is large. And the conditional probability that your p-value represents, the probability of getting sample results or results more extreme given that the null hypothesis is true represents the reverse of what you really want. What you really want is the probability that the null hypothesis is true given the data that you observed.

A lot of people dislike the idea of looking for a probability involving the sample results or results more extreme. Why, they ask, do you want a probability involving results more extreme. You didn’t observe a value more extreme. You observed a single value.

The p-value ignores clinical importance. This would be easy to fix if people used an interval -delta to delta for the null hypothesis as mentioned earlier, but no one is ready to do this.

Finally, the p-value is unaffected by threats to internal validity. If you conduct the study poorly, such as failing to keep information away from patients in a blinded study, or having a large number of protocol violations, that should be reflected in your p-value. But the p-value ignore these problems.

Criticisms of hypothesis testing (3 of 4)

- General criticisms
  - Too hard to reject H0
  - Too easy to reject H0
  - Too reliant on a single study
  - Thoughtless application

Speaker notes

Hypothesis testing has been criticized because it is too hard to reject the null hypothesis, especially for small samples in a noisy setting. The conservative inside of you and me probably thinks this is good. Don't make a choice between two therapies or drugs until you've accumulated sufficient evidence. But failing to act quickly can sometimes force you to pay a price. This relates to the trade-offs that you see all the time between false negative results and false positive results. The hypothesis testing framework is biased towards preventing a false positive result (a Type I error) and it is very difficult to get this framework to work well when a false negative result is worse. This can occur, for example, in a setting with a 100% fatal disease with no known cure.

The opposite problem is also true. It is often too easy to reject the null hypothesis. In this era of big data, you can quickly get millions of data points or more and you have so much precision that any sample statistics even of a trivial size will lead to a statistically significant result. This is not good. Instead of identifying one or two risk factors as being statistically significant, a really big data set will identify hundreds or even thousands of risk factors as being statistically significant.

The p-value is too reliant on a single study and does not consider what previous research has been done.

My biggest criticism of p-values, though, is the thoughtless way in which they are applied. I've written about the p-value receptor inside a scientist's brain.

A research team took ten scientists and placed them inside an fMRI. The fMRI shows which parts of your brain are active as your brain processes different types of information. The scientists were shown a variety of graphs taken from actual peer-reviewed publications.

As you might expect, the part of your brain that activates first when you are presented an image of a graph is your visual cortex. For most graphs, this was quickly followed by an activation of the parietal lobe, the part of your brain responsible for numerical computations.

But some graphs showed a different pattern. If the graph included a p-value, activation of the visual cortex is followed by activation of an area of your amygdala that is as yet poorly understood. The research team called this portion of the amygdala the p-value receptor.

If your p-value receptor is activated and the p-value is larger than 0.05, the p-value receptor sends strong signals to the pain centers of the brain. This is clearly an adaptive behavior. Scientists who routinely produce p-values larger than 0.05 will not survive and reproduce.

If the p-value receptor is activated and the p-value is 0.05 or smaller, the p-value receptor sends strong signals to the pleasure centers of the brain. Again this is an adaptive behavior. But the interesting finding is that there is a dose response effect. The p-value receptor produces about the same level of pleasure stimulation for p-values of 0.05, 0.04, 0.03, and 0.02. But p-values of 0.01 show an increase in stimulation that becomes even strong for p-values of 0.0099 and smaller. Perhaps there is some pattern associated with p-values that have two zeros to the right of the decimal place that is stronger then a p-value with just a single zero.

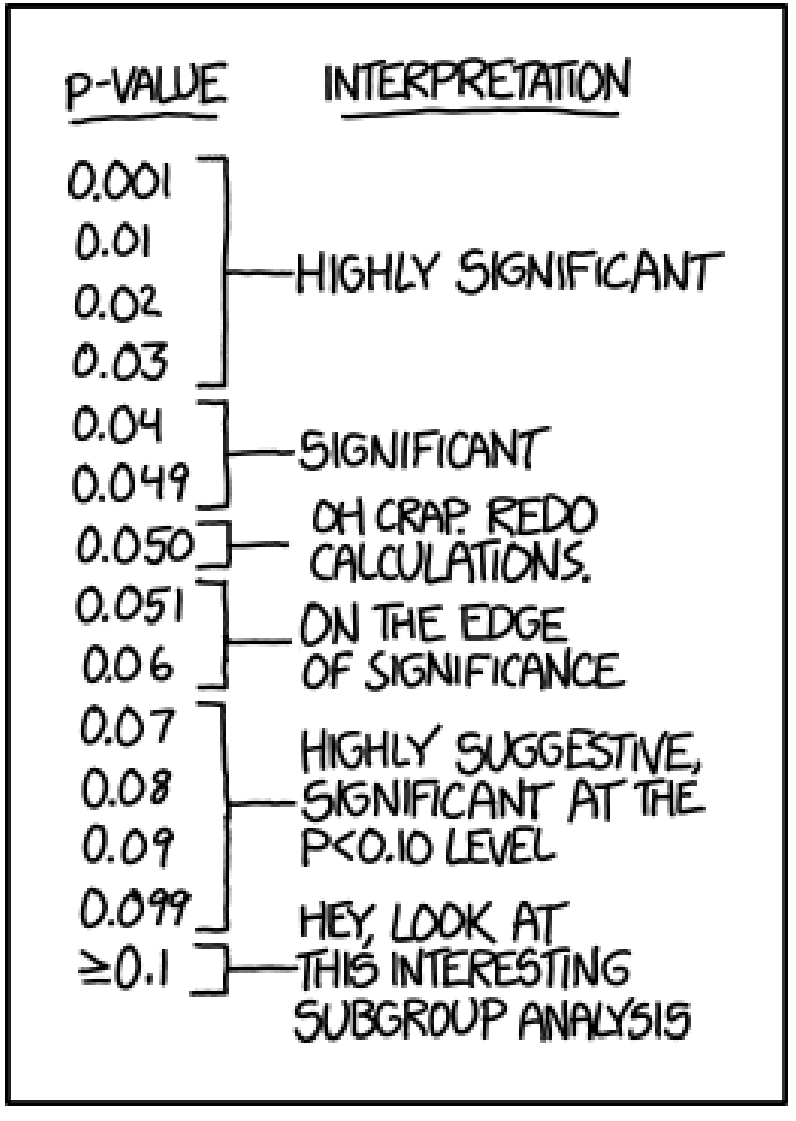
The scientists also examined the effect of p-values reported in scientific notation. There was an increase in latency when the p-value receptor is fed a p-value in scientific notation. This probably represents an attempt to decode the scientific notation. But p-values in scientific notation with exponents of -4 or smaller showed an eventual spike in activation of the pleasure centers of the brain that are comparable to those achieved during orgasm.

The research also noted a second important effect of the p-value receptor. Once the p-value receptor is stimulated, the entire cerebral cortex, the portion of your brain associated with logic and complex thinking, is immediately shut down. This insures that a scientist's brain will focus only on the pleasure or pain associated with the p-value and will ignore the power of the study, the magnitude of the treatment effect, and other unimportant issues.

The researchers suggest that statisticians who want to earn more consulting income and insure repeated business should do their best to produce only p-values that stimulate the pleasure centers of the brain.

This is from my blog: http://blog.pmean.com/scientist-brain/

Criticisms of hypothesis testing (4 of 4)



Speaker notes

I have to end with another cartoon. Both of these cartoons were drawn by Scott Munro, the creator of the xkcd comic series.

This is a cartoon showing a table of p-values with various labels.

0.001, 0.01, 0.02, and 0.03 are labeled “Highly Significant”.

0.04 and 0.049 are labeled “Significant”.

0.050 is labeled “Oh crap. Redo calculations.” This is because no one knows what exactly to do with a p-value that is right on the boundary.

0.051, 0.06 are labeled “On the edge of significance”. If these values are on the edge of significance then the 0.04 and 0.049 should share that same label. But no one uses these hedging terms unless they are on the “wrong” side of 0.05.

0.07, 0.08, 0.09, 0.099 are labeled “Highly suggestive, relevant at the $p<0.10$ level”. I personally don’t have a complaint here, but others consider this a post hoc modification and a form of p-hacking.

≥0.1 is labeled “Hey, look at this interesting subgroup analysis”. This is a reference to p-hacking. If your primary p-value is not statistically significant, hunt for some other p-values.

What should you do if you do not have a hypothesis to test?

- Descriptive statistics
  - Include confidence intervals
- Qualitative data analysis

Speaker notes

Add note

Break

- What have you just learned?
  - Criticisms of p-values and hypothesis testing
- What is coming next?
  - Review confidence intervals

Speaker notes

Let me pause here for a second. Are there any questions?

Standardization of a statistic

General form of standardization

- $Z \; or \; t = \frac{statistic - parameter}{se(statistic)}$

Specific standardization for the mean

- $Z \; or \; t = \frac{\bar{X} - \mu}{se(\bar{X})}$

Speaker notes

Convert this to a confidence interval

- $P[-t(\alpha/2; n-1) < \frac{\bar{X}-\mu}{se(\bar{X})} < t(\alpha/2; n-1)] = 1-\alpha$
- $P[-t(\alpha/2; n-1)se(\bar{X}) < \bar{X}-\mu < t(\alpha/2; n-1)se(\bar{X})]$
- $P[\bar{X} - t(\alpha/2; n-1)se(\bar{X}) < \mu < \bar{X} + t(\alpha/2; n-1)se(\bar{X}$

If n > 30

- $P[\bar{X} - Z(\alpha/2)se(\bar{X}) < \mu < \bar{X} + Z(\alpha/2)se(\bar{X})] \approx 1-\alpha$

Speaker notes

Intepretation

If n < 30, we have 1-α level of confidence that the population mean lies between

- $\bar{X} - t(\alpha/2; n-1)se(\bar{X})$ and
- $\bar{X} + t(\alpha/2; n-1)se(\bar{X})$

If n > 30, we have 1-α level of confidence that the population mean lies between

- $\bar{X} - Z(\alpha/2)se(\bar{X})$ and
- $\bar{X} + Z(\alpha/2)se(\bar{X})$

Speaker notes

The formula for the confidence interval for the population mean

Other forms of the confidence interval

- Simpler (too simple?)
  - $\bar{X} - 2\ se(\bar{X})$ and
  - $\bar{X} + 2\ se(\bar{X})$ and
- Use $t(\alpha/2; n-1)$ even if n > 30
- Do not use these alternate forms for your homework.

Original Article

Using ChatGPT for Writing Articles for Patients' Education for Dermatological Diseases: A Pilot Study

Abstract

Background: Patients' education is a vital strategy for understanding a disease by patients and proper management of the condition. Physicians and academicians frequently make customized education materials for their patients. An artificial intelligence (AI)-based writer can help them write an article. Chat Generative Pre-Trained Transformer (ChatGPT) is a conversational language model developed by OpenAI (openai.com). The model can generate human-like responses. **Objective:** We aimed to evaluate the generated text from ChatGPT for its suitability in patients' education. **Materials and Methods:** We asked the ChatGPT to list common dermatological diseases. It provided a list of 14 diseases. We used the disease names to converse with the application with disease-specific input (e.g., write a patient education guide on acne). The text was copied for checking the number of words, readability, and text similarity by software. The text's accuracy was checked by a dermatologist following the structure of observed learning outcomes (SOLO) taxonomy. For the readability ease score, we compared the observed value with a score of 30. For the similarity index, we compared the observed value with 15% and tested it with a one-sample *t*-test. **Results:** The ChatGPT generated a paragraph of text of 377.43 ± 60.85 words for a patient education guide on skin diseases. The average text reading ease score was 46.94 ± 8.23 ($P < 0.0001$), and it indicates that this level of text can easily be understood by a high-school student to a newly joined college student. The text similarity index was higher (27.07 ± 11.46%, $P = 0.002$) than the expected limit of 15%. The text had a "relational" level of accuracy according to the SOLO taxonomy. **Conclusion:** In its current form, ChatGPT can generate a paragraph of text for patients' educational purposes that can be easily understood. However, the similarity index is high. Hence, doctors should be cautious when using the text generated by ChatGPT and must check for text similarity before using it.

Keywords: *Article, artificial intelligence, ChatGPT, dermatologists, software*

Himel Mondal, Shaikat Mondal[1], Indrashis Podder[2]

Department of Physiology, All India Institute of Medical Sciences, Deoghar, Jharkhand, [1]Department of Physiology, Raiganj Government Medical College and Hospital, West Bengal, [2]Department of Dermatology, College of Medicine and Sagore Dutta Hospital, Kolkata, West Bengal, India

Figure 3: Excerpt from Mondal 2023

Speaker notes

Here is an article from the Indian Dematology Online Journal. You can find it using the Pubmed ID of 37521213.

Table 1: Characteristics of generated text (patient education guide on common skin diseases) from ChatGPT

Variable	Mean	Standard deviation	First quartile	Median	Third quartile	One-sample *t*-test
Sentences (*n*)	23.07	3.6	20.75	22	25.5	-
Words (*n*)	377.43	60.85	322.75	374.5	429.25	-
Words/sentence	16.39	1.31	15.35	16.25	17.78	-
Syllables/word	1.67	0.06	1.6	1.7	1.7	-
Grade level	10.53	0.73	10	10.4	10.95	-
Ease score	46.94	8.23	39.15	47.6	53.6	<0.0001*
Overall similarity (%)	27.07	11.46	19	29	35.5	0.002†
Internet (%)	19	11.39	6	18	31	-
Publication (%)	3.86	6.02	0	1	4.75	-
Students' paper (%)	18.79	9.22	13.75	20	23.5	-

-: Not required. *One-sample *t*-test by comparing with ease score of 30. †One-sample *t*-test by comparing with an overall similarity of 15%

Figure 4: Table 1 from Mondal 2023

Speaker notes

This is Table 1 from the paper.

Confidence interval for Ease Score

- Calculations
  - $se(\bar{X}) = \frac{8.23}{\sqrt{14}} = 2.19956$
  - $t(0.025, 13) = 2.16$
  - $46.94 - 2.16 \times 2.19956 = 42.18895$
  - $46.94 + 2.16 \times 2.19956 = 51.69105$
- We are 95% confident that the population mean reading age of ChatGPT education guides is between 42 and 52.

Speaker notes

The ease score represents ease of reading with larger values representing better readability. AQ value of 30 was cited as representing difficult to read text based on a review of editorials in popular Indian medical journals.

Confidence interval for overall similarity

- Calculations
  - $se(\bar{X}) = \frac{11.46}{\sqrt{14}} = 3.062814$
  - $t(0.025, 13) = 2.16$
  - $27.07 - 2.16 \times 7.234762 = 20.45432$
  - $27.07 + 2.16 \times 3.062814 = 33.68568$
- We are 95% confident that the population mean similarity is between 20% and 34%.

Speaker notes

The similarity score is used for plagiarism checks. A value of 15% or more is taken by some resources as concerning.

Effect sizes

- Cohen's d = $\frac{\bar{X}-C}{se(\bar{X})}$
- Useful for
  - Systematic overviews
  - Intermediate calculation in sample size formulas
- Critcisms
  - Clinical relevance requires units of measure
  - Small, medium, large are arbitrary labels

Speaker notes

The effect size for a sample mean is calculated by subtracting the hypothesized mean, C, from the sample mean and dividing by the standard error of the mean.

The use of effect sizes is controversial. Monica Gaddis loves them, other faculty in our department love them. I hate them.

They are okay for systematic overviews and as an intermediate calculation in sample size formulas.

But I do not like to see them reported in journal articles. Using terms like small, medium, or large is very arbitrary and does not translate well from one discipline to another. I also believe that unitless quantities provide no help in deciding whether a result has clinical relevance.

There’s a joke I tell about this. A store displays a huge banner at their entrance. The banner reads “Big sale! All prices reduced by half a standard deviation.”

Even though I dislike effect sizes, I feel the obligation to teach them because they are used by many researchers.

Break

- What have you just learned?
  - Review confidence intervals testing
- What is coming next?
  - SPSS examples

Speaker notes

Let me pause here for a second. Are there any questions?

Baseball data dictionary (1/2)

This file was downloaded from the DASL (Data and Story Library) website. There are no details about who created the data set or what permissions are allowed. Educational uses of this data are probably allowed under the Fair Use provisions of U.S. Copyright Law.

This is a tab delimited data file. There are 50 rows and 2 columns of data.

The first variable is the sample number (1 to 50). The second variable is the circumference of the baseball in inches. The variable names are included at the top of the data.

Speaker notes

I want to show some SPSS analyses. The first relates to a dataset on baseballs.

Baseball data dictionary (2/2)

The first variable is the sample number (1 to 50). The second variable is the circumference of the baseball in inches. The variable names are included at the top of the data.

The standard sized baseball, according to Wikipedia and other sources on the Internet is 9 to 9.25 inches. There are no missing values in this data set.

This data dictionary was written by Steve Simon on 2023-09-10 and is placed in the public domain.

Please be sure to skip past this documentation while importing the data.

Speaker notes

I want to show some SPSS analyses. The first relates to a dataset on baseballs.

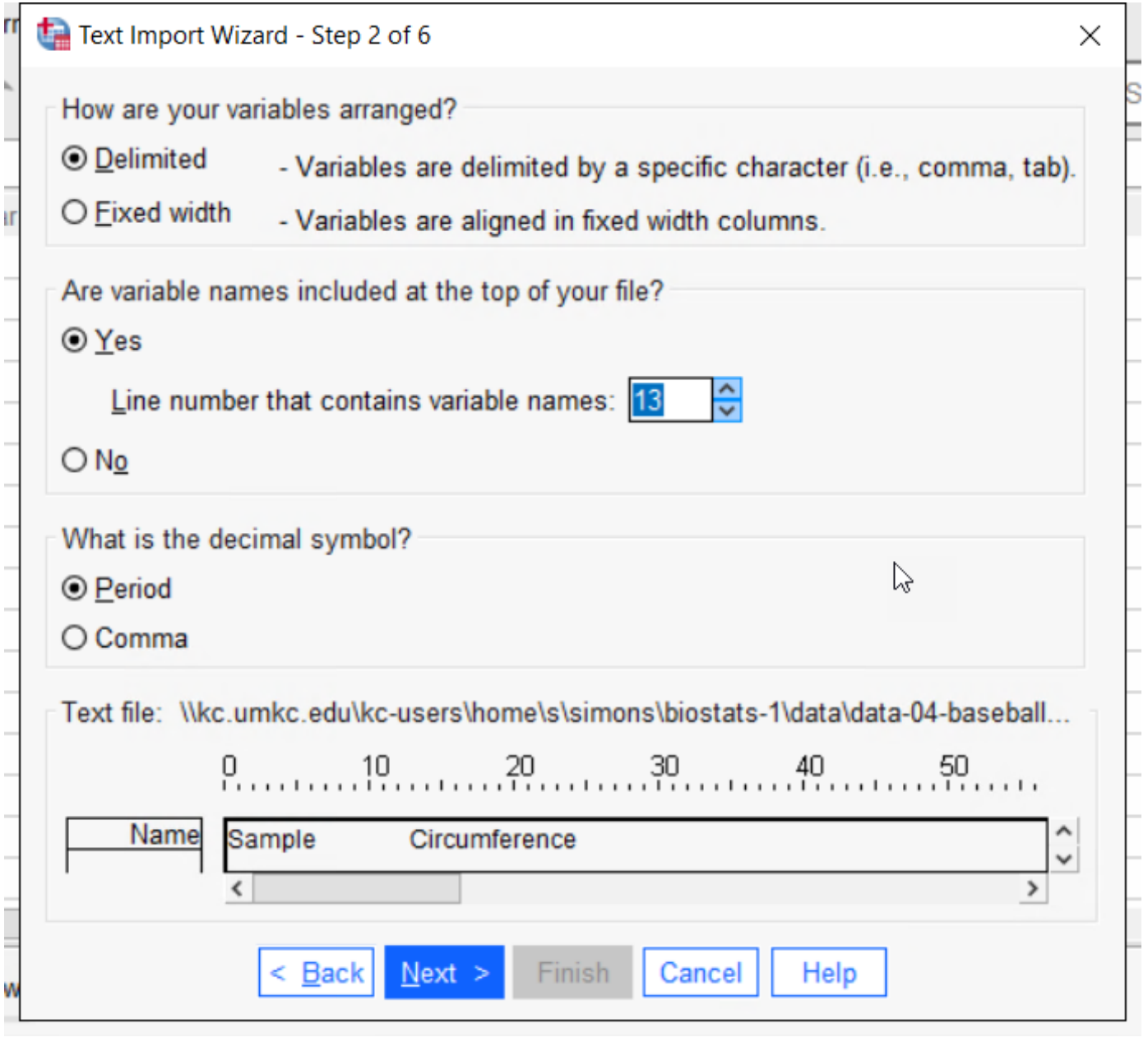


Figure 5: SPSS import dialog box

Speaker notes

Add note.

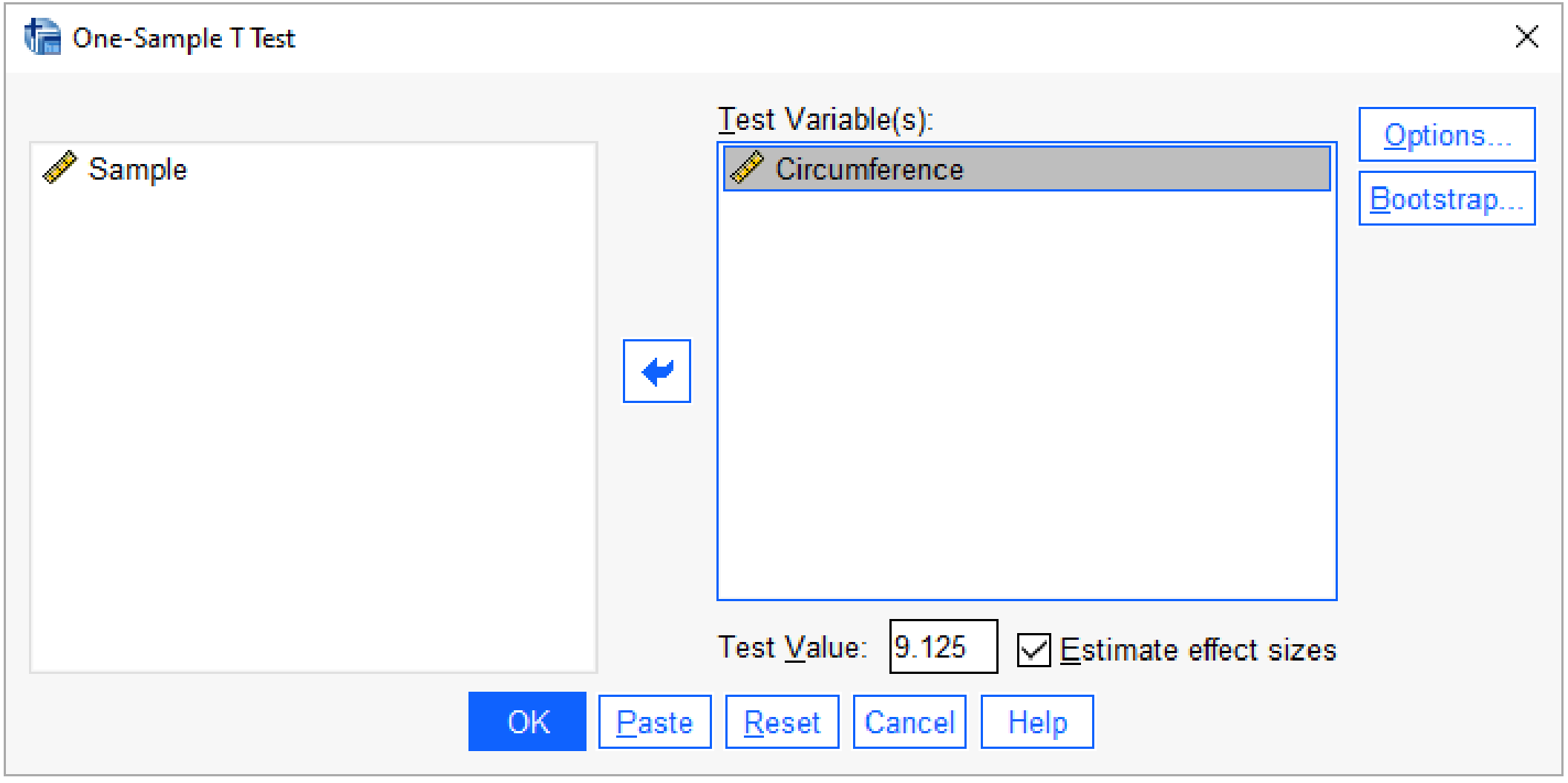


Figure 6: SPSS one-sample t-test dialog box

Speaker notes

Add note.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Circumference	50	9.11754	.049415	.006988

Figure 7: SPSS one-sample t-test output (1/3)

Check that $\frac{0.049415}{\sqrt{50}} = 0.006988.$

Speaker notes

Ask for some help to check the calculation for the standard error.

Approximate confidence interval

- Note that $se(\bar{X}) \approx 0.007$
  - 9.118 - 0.014 = 9.104
  - 9.118 + 0.014 = 9.132

Speaker notes

You can get an approximate 95% confidence interval using the plus or minus two standard errors rule. The standard error is roughly 0.007. Two standard errors would be roughly 0.014. Subtracting and adding from the mean provides an approximate confidence interval of 9.104 to 9.132.

One-Sample Test

Test Value = 9.125

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Circumference	-1.068	49	.145	.291	-.007460	-.02150	.00658

Figure 8: SPSS one-sample t-test output (2/3)

Check that $\frac{9.11754-9.125}{0.006988} = -1.068.$

Speaker notes

Add note.

Converting the SPSS confidence interval

- We are 95% confident that $\mu - 9.125$ is between -.02150 and 0.00658.
  - Add 9.125 to both sides.
  - -0.02150 + 9.125 = 9.10350
  - 0.00658 + 9.125 = 9.13158
- Always round at the end.
  - We are 95% confidence that the population mean circumference is between 9.10 and 9.13.

One-Sample Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Circumference	Cohen's d	.049415	-.151	-.429	.129
	Hedges' correction	.050187	-.149	-.422	.127

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

Figure 9: SPSS one-sample t-test output (3/3)

Check that $\frac{9.11754-9.125}{0.049415} = -0.151.$

Speaker notes

I am not a big fan of effect sizes, but I am in a minority perspective here. Life is too short to worry about Cohen’s d versus the Hedges correction. Both indicate a small effect size (less than 0.2).

BMI data dictionary (1/2)

This file is included as part of the base package of R and was converted by Steve Simon to a text file. There are no details about who created the data set. The code for R is published under an open source license, and the datasets included with R are presumably covered by the same license.

This is a tab delimited data file. There are 15 rows and 3 columns of data.

Speaker notes

Here is the first half of the data dictionary for the BMI data.

BMI data dictionary (1/2)

The first variable is the sample number (1 to 15). The second variable is the height of an adult female (inches). The third variable is the weight (pounds). The variable names are not included at the top of the data.

This data dictionary was written by Steve Simon on 2023-09-10 and is placed in the public domain.

Please be sure to skip past this documentation while importing the data.

Speaker notes

Here is the second half of the data dictionary for the BMI data.

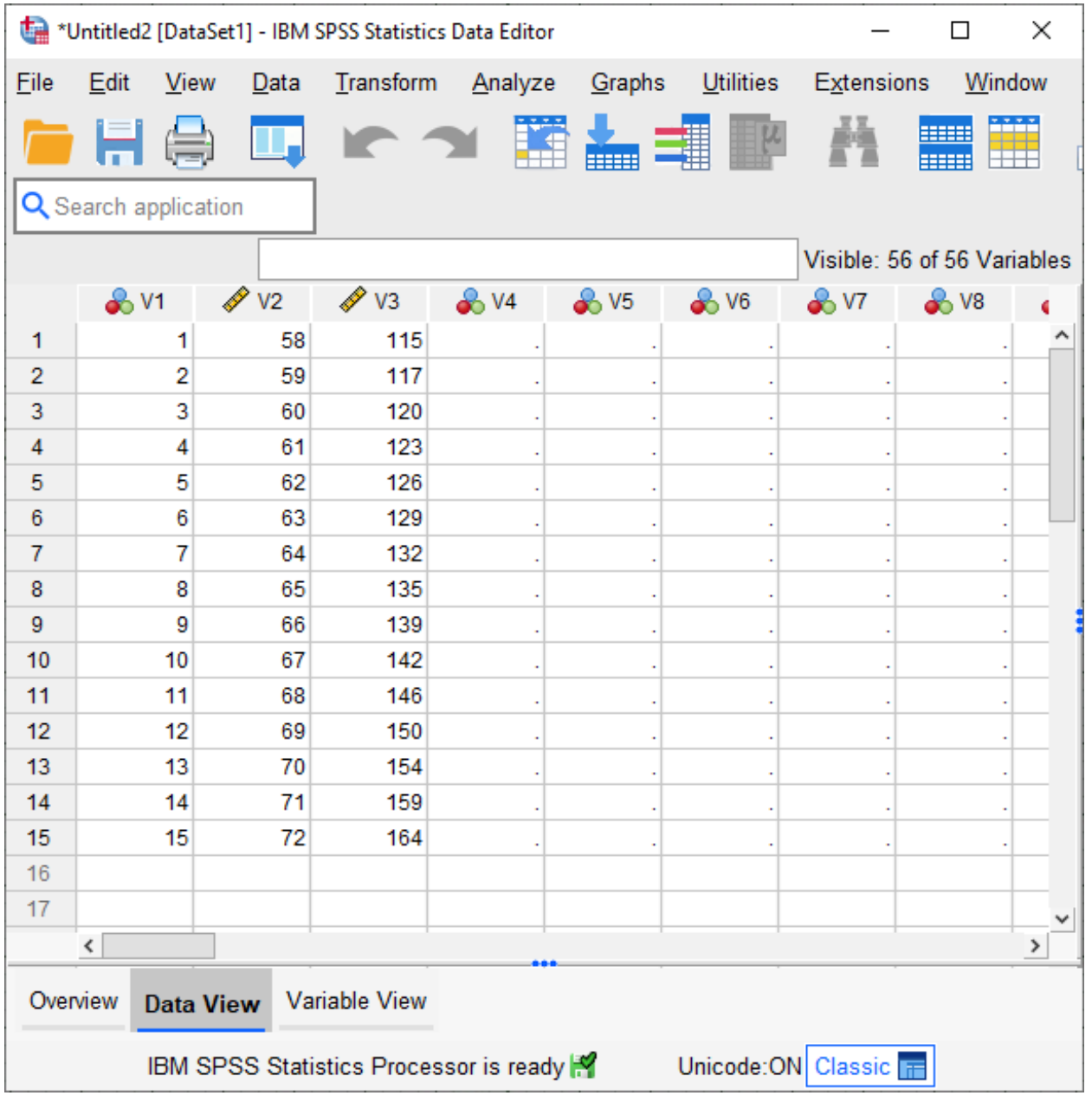


Figure 10: SPSS data after importing

Speaker notes

Add note.

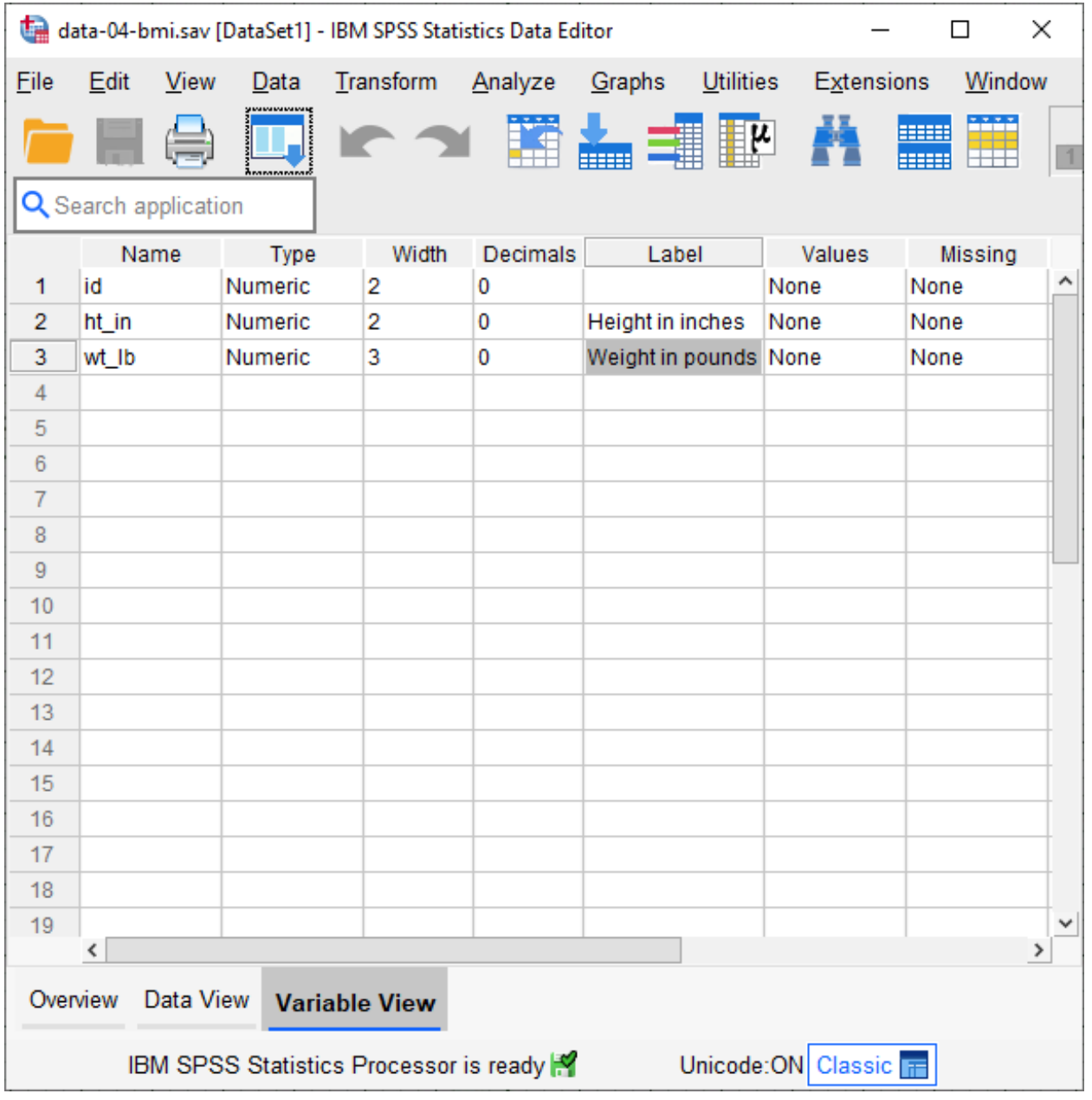


Figure 11: SPSS data after variable name change

Speaker notes

Add note.

Converting height to meters

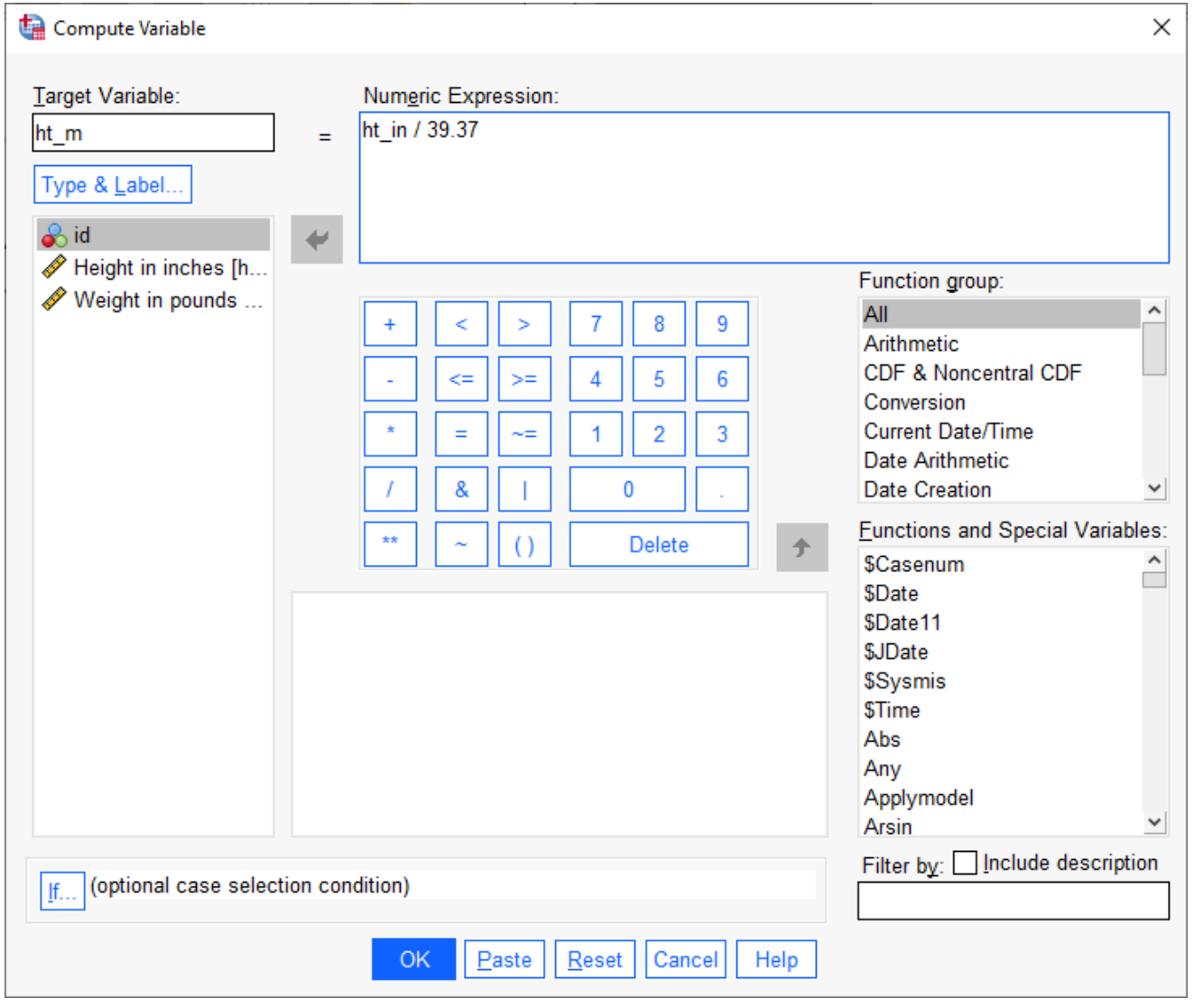


Figure 12: SPSS dialog box for converting from inches to meters

Speaker notes

Add note.

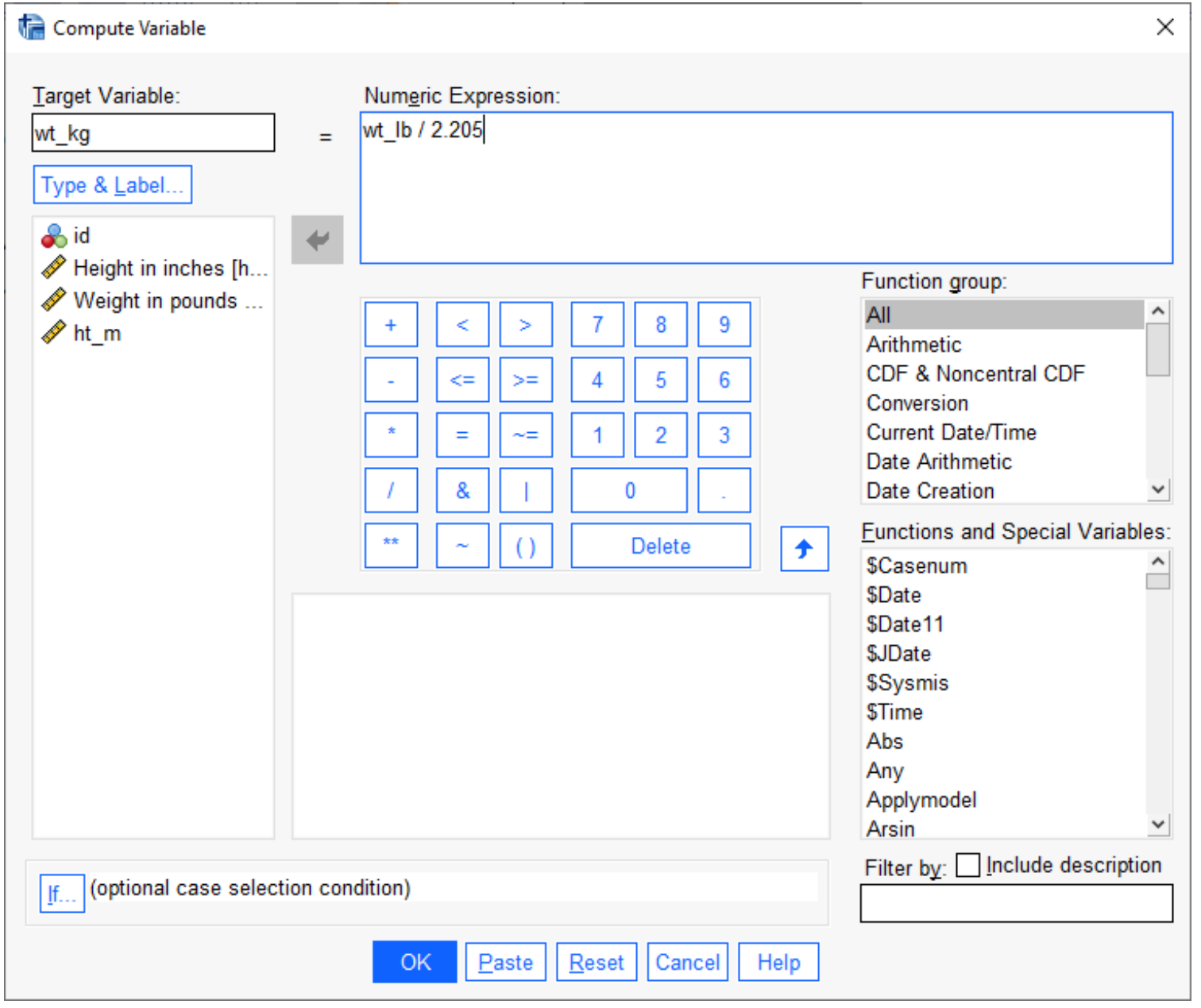


Figure 13: SPSS dialog box for converting from pounds to kilograms

Speaker notes

Add note.

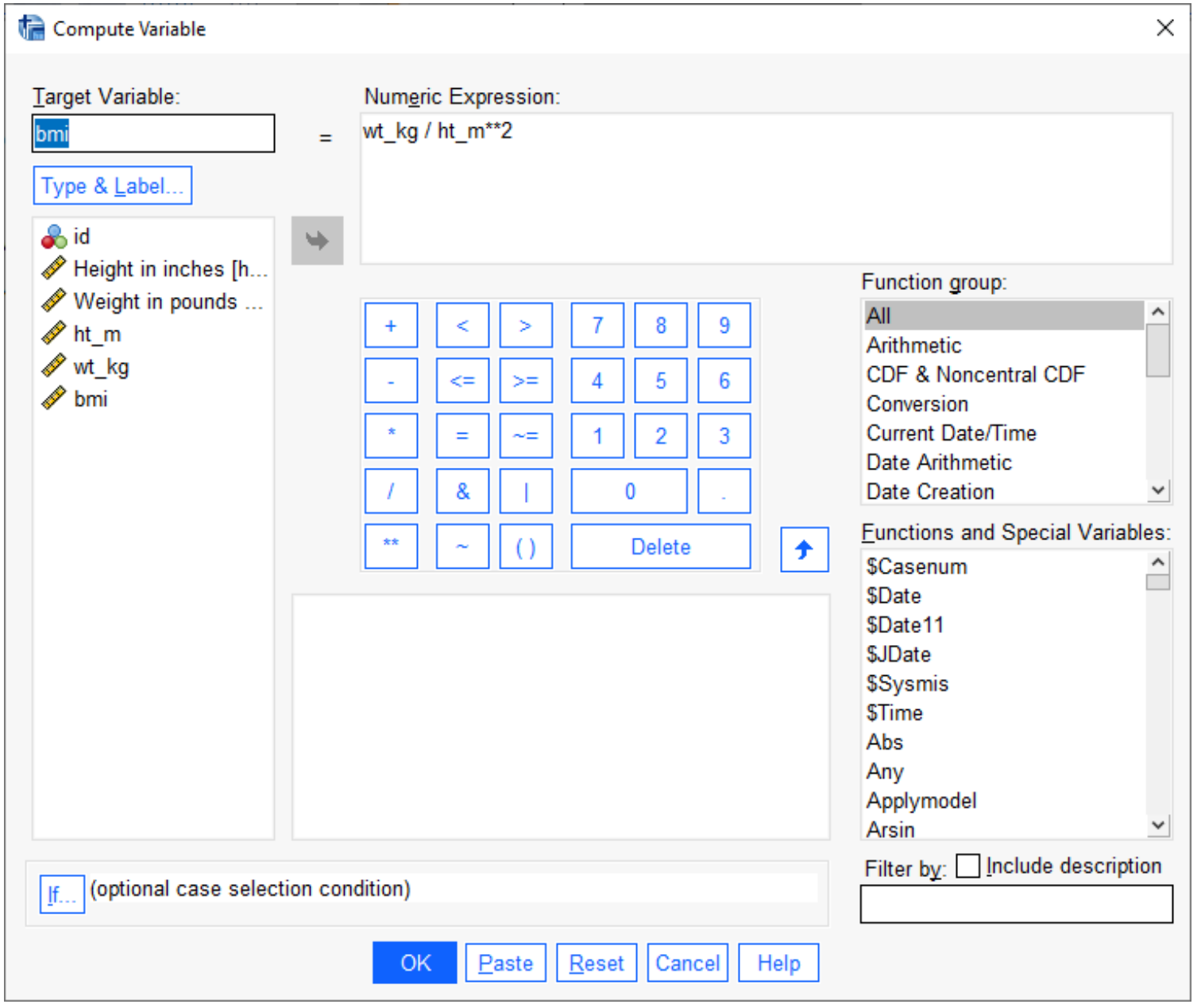


Figure 14: SPSS dialog box for BMI

Speaker notes

Add note.

Select Analyze | Compare Means and Proportions | One-sample T Test from the SPSS menu.

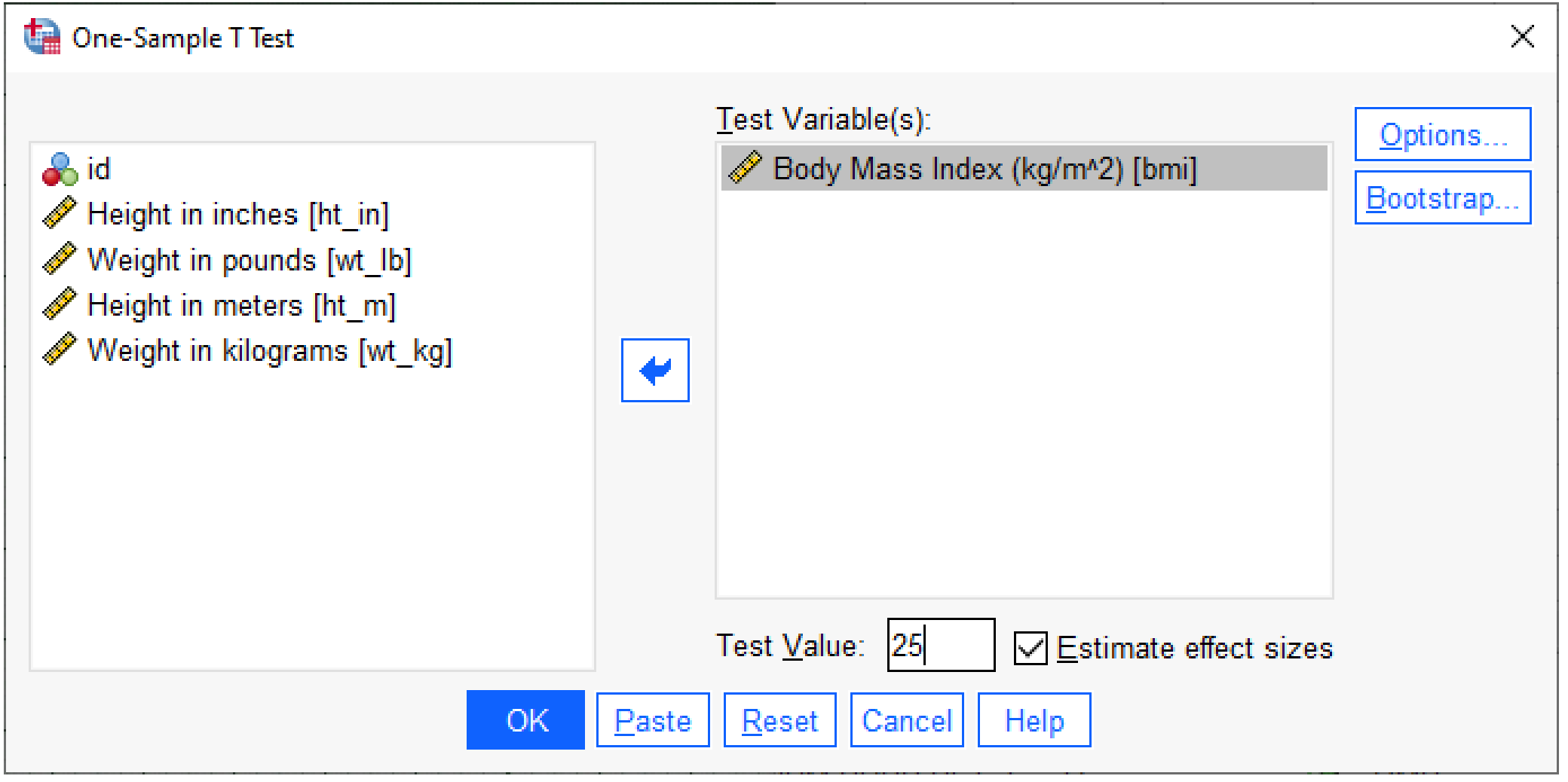


Figure 15: SPSS dialog box for one-sample t-test

Speaker notes

Add note.

Output from a one-sample t-test (1/3)

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Body Mass Index (kg/m^2)	15	22.7227	.61824	.15963

Figure 16: SPSS output from one-sample t-test (1/3)

Speaker notes

Add note.

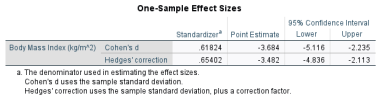
One-Sample Test

	Test Value = 25						
	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Body Mass Index (kg/m^2)	-14.266	14	<.001	<.001	-2.27730	-2.6197	-1.9349

Figure 17: SPSS output from one-sample t-test (2/3)

Speaker notes

Add note.



One-Sample Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	Upper
Body Mass Index (kg/m^2)	Cohen's d	.61824	-3.684	-5.116	-2.235
	Hedges' correction	.65402	-3.482	-4.836	-2.113

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

Figure 18: SPSS output from one-sample t-test (3/3)

Summary

- What have you learned?
  - Populations, samples, parameters, statistics
  - The null and alternative hypotheses
  - Decision rules, Type I and II errors
  - p-values and criticisms
  - Confidence intervals
  - SPSS examples

Speaker notes

Add note.