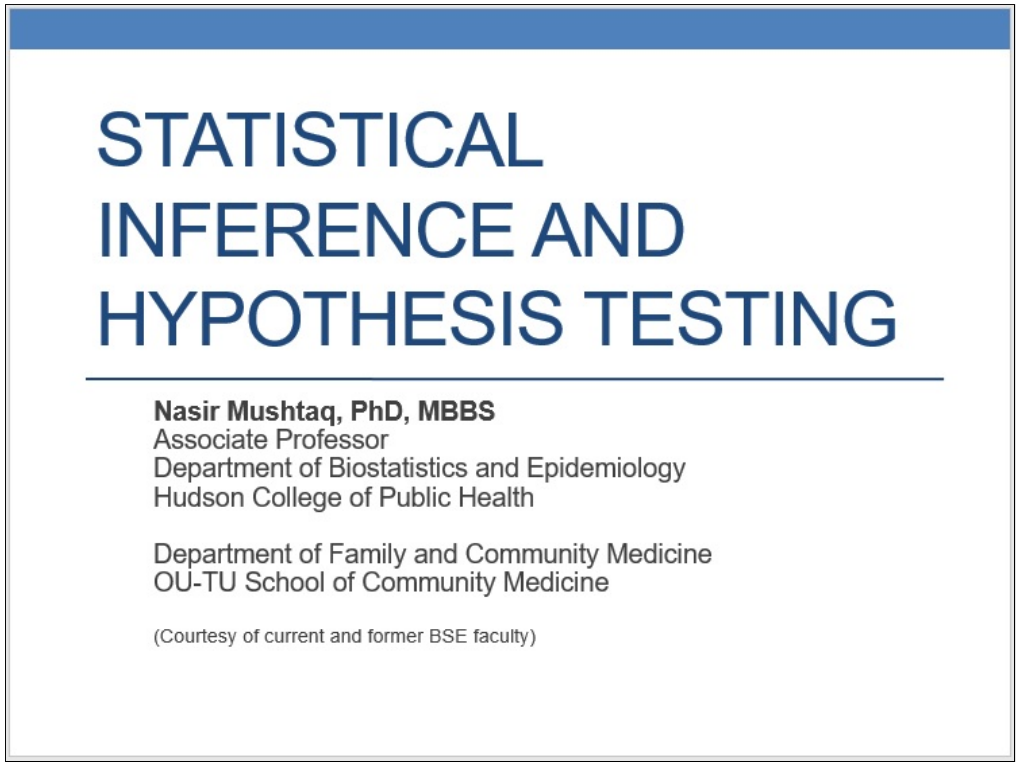
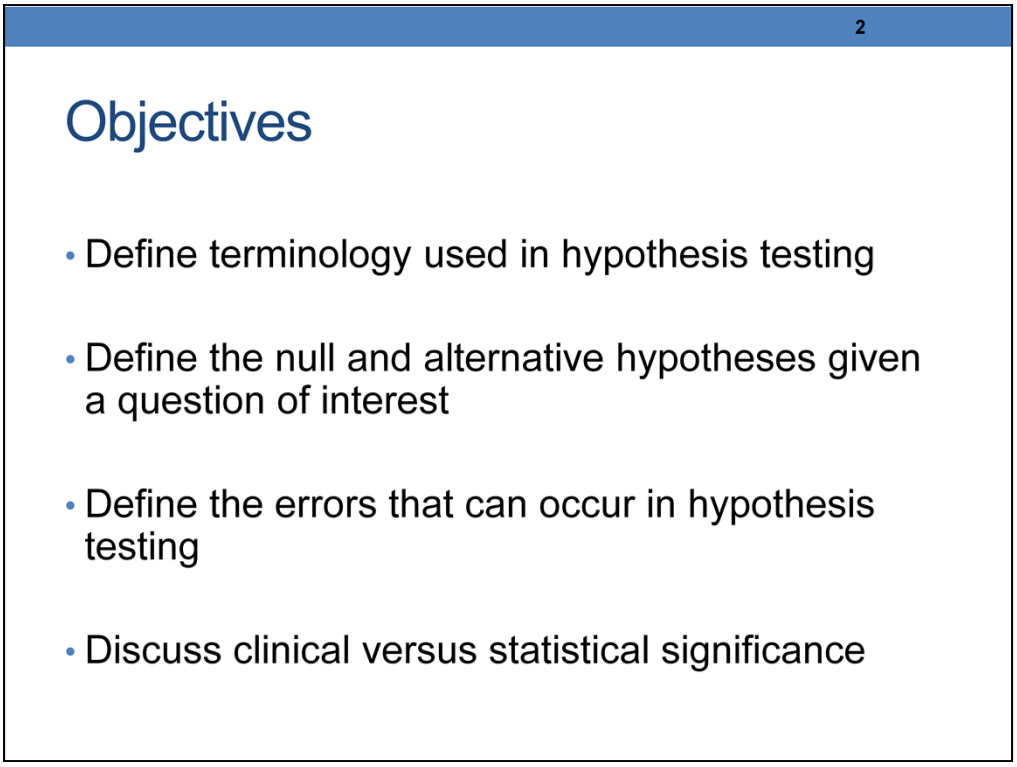


Welcome, this video series is entitled Statistical Inference and Hypothesis Testing. In this first part, we will discuss statistical concepts related to hypothesis testing.



After viewing this video, you will be able to:

Define terminology used in hypothesis testing
Define the null and alternative hypotheses given a research question of interest
Define the errors that can occur in hypothesis testing
Discuss clinical versus statistical significance

Sample vs. Population

- Population describes the hypothetical (and usually) large number of people to whom you wish to generalize
- Sample describes those individuals who are in the study (fraction of the population)
  - The study is only generalizable to the type of patients who are in the study

In practice, we are not able to study an entire population of patients in a given medical research study. Instead, we will study a sample of participants from the target population and will make inference from the sample of participants to the target population at large. Our ability to generalize results from the sample to a given population depends on the sampling selection strategy and the methods of the study. For example, if we only collect information on women in our study sample, we cannot necessarily generalize the results to a broad population of both men and women.

STATISTICAL CONCEPTS IN THE LITERATURE

Now, I will define statistical concepts related to hypothesis testing that you will see in published medical literature.

Example: Intensifying Antihypertensive Treatment

"A sample size calculation indicated that 114 patients per treatment group would be necessary for 90% power to detect a true mean difference in change from baseline of 3 mm Hg in sitting DBP between the two randomized treatment groups. This calculation assumed a two-sided test, α=0.05, and standard deviation in sitting DBP of 7 mm Hg."

Source: *AJH*. 1999;12:691-696

As an example from a clinical trial investigating the effect of single agent or dual agent antihypertensive medication that was published in the American Journal of Hypertension, consider this sample size justification. We are told that a sample size calculation indicated that 114 patients per treatment group would be necessary for 90% power to detect a true mean difference in change from baseline of 3 mm Hg in sitting DBP between the two randomized treatment groups. This calculation assumed a two-sided test, □=0.05, and standard deviation in sitting DBP of 7 mm Hg.

Example: Intensifying Antihypertensive Treatment

"A sample size calculation indicated that 114 patients per treatment group would be necessary for 90% power to detect a true mean difference in change from baseline of 3 mm Hg in sitting DBP between the two randomized treatment groups. This calculation assumed a two-sided test, α=0.05, and standard deviation in sitting DBP of 7 mm Hg."

Source: *AJH*. 1999;12:691-696

After reviewing this lecture, you will be able to interpret power and the assumed alpha level.

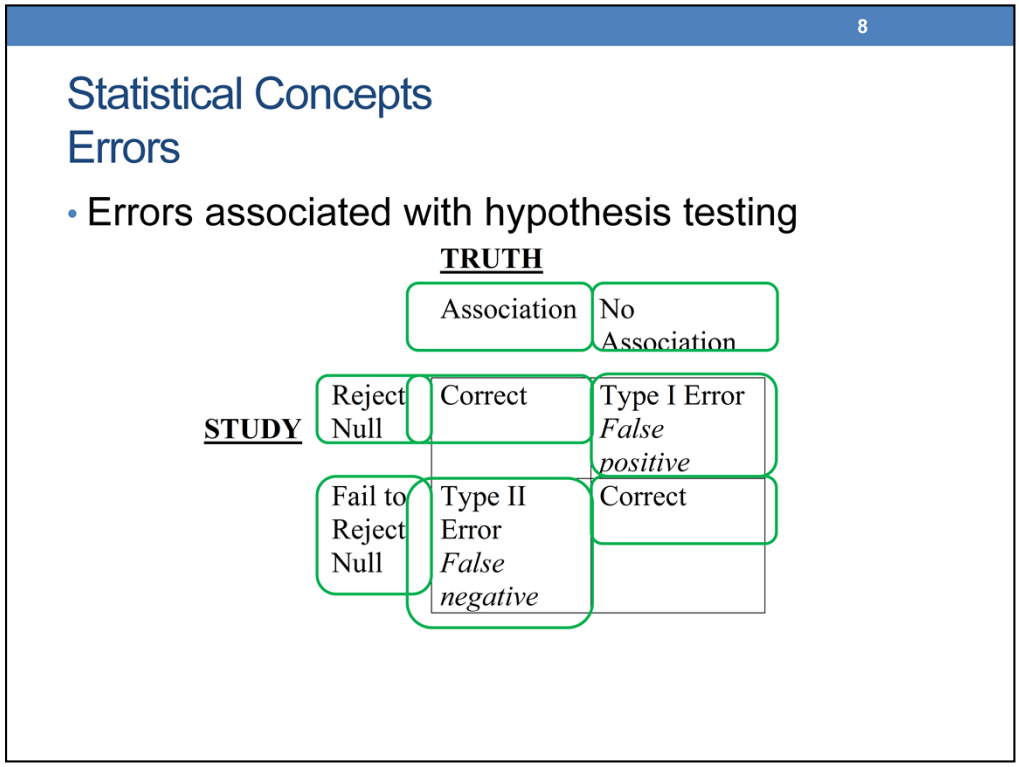
Statistical Concepts: Hypotheses

- Null hypothesis: H_0
  - Typically a statement of no treatment effect
  - Assumed true until evidence suggests otherwise
  - Example: H_0: No difference in mean DBP between treatment groups
- Alternative: H_A
  - Reject null hypothesis in favor of alternative hypothesis
  - Often two-sided
  - Example: H_A: mean DBP differs between treatment groups

In hypothesis testing, when we aim to detect intervention effects or associations between exposures and outcomes, meaning, superiority settings, we have two statistical hypotheses in mind, the Null Hypothesis and the Alternative Hypothesis. In a superiority setting, the Null Hypothesis is a statement of no treatment effect or no association and is assumed true until we find evidence to suggest otherwise. Relative to our study example, the Null Hypothesis will be that the mean diastolic blood pressure is the same in the treatment groups.

The Alternative Hypothesis is the hypothesis that we hope to find evidence in favor of. In practice, we will reject the Null Hypothesis in favor of the Alternative Hypothesis. The Alternative Hypothesis is typically two-sided, which means that we are interested in detecting differences between groups in either a positive or negative direction. Relative to our study example, the Alternative Hypothesis will be that the mean diastolic blood pressure differs between the treatment groups.

Note that in statistical hypothesis testing, we cannot prove that a given alternative hypothesis is true. Instead, we assume that the null is true and determine if there is sufficient evidence to reject the null based on the observed data.



When conducting hypothesis tests, we collect a sample of data and then make a decision or form inferences based on the sample of data. When making a decision, we either make a correct decision or make an error relative to the true status. In truth, there either is an association or is no association between, say, treatment and outcome. Based on our data, we make a decision to either reject the null and conclude that there is an association or we fail to reject the null hypothesis and conclude that there is no significant indication of an association.

If there is an association between the intervention and response, and we reject the null, we have made a correct decision.

If there is no association between the intervention and response, and we fail to reject the null, we have made a correct decision.

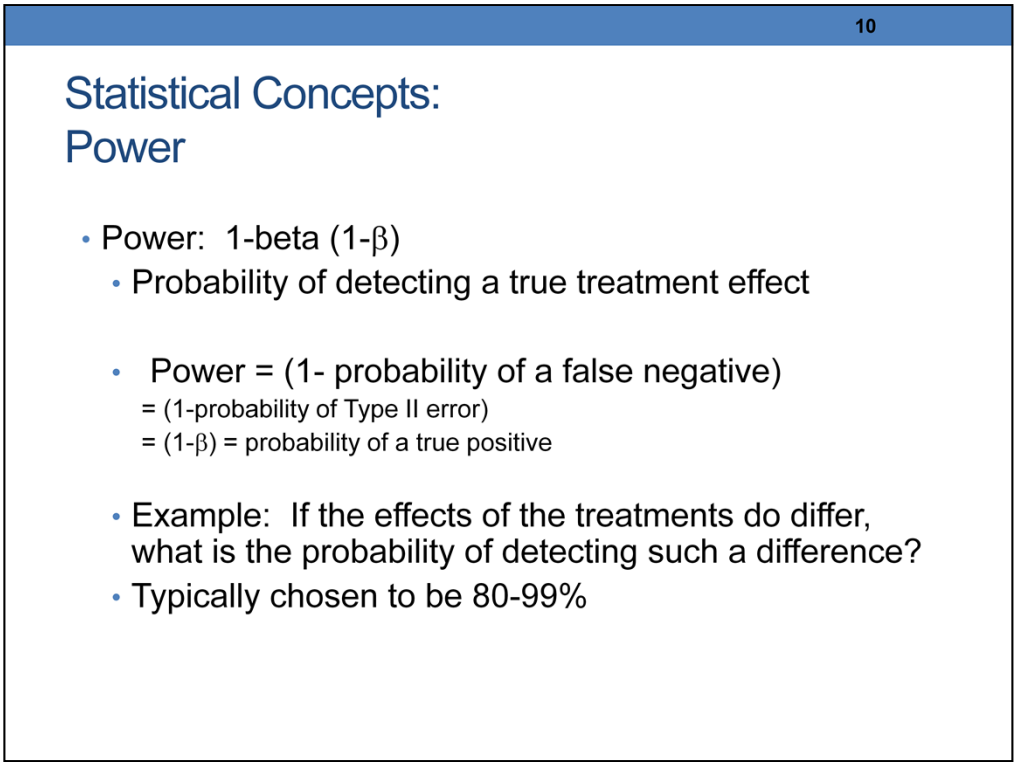
In both other cases, we make an error. If there is no association between the intervention and response, and we reject the null, we have committed a false positive error, which we refer to as a Type I error. Alternatively, if there is an association between the intervention and response, and we fail to reject the null, we have failed to detect a true intervention effect and have committed a false negative error, which we refer to as a Type II error.

Statistical Concepts: Significance Level

- Significance level: α
  - Probability of a Type I error
  - Probability of a false positive
  - Example: If the effect on DBP of the treatments do not differ, what is the probability of incorrectly concluding that there is a difference between the treatments?
  - Typically chosen to be 5%, or 0.05

Let's look more closely at the two types of errors that we may commit in hypothesis testing.

If there is no association between the intervention and response, and we reject the null, we have committed a false positive error, which we refer to as a Type I error. We denote the probability of a type I error as alpha. In terms of our example study, the alpha level addresses the question, "If the effect on DBP of the treatments do not differ, what is the probability of incorrectly concluding that there is a difference between the treatments?". In practice, the probability of a Type I error should be low and is typically set to be 5%. Meaning, we often test hypotheses using an alpha level of 0.05.



The other type of error that can be made in hypothesis testing is a Type II error. If there is an association between the intervention and response, and we fail to reject the null, we have failed to detect a true intervention effect and have committed a false negative error, which we refer to as a Type II error.

The probability of a Type II error is related to the Power of the study. The Power of the study is the probability of rejecting the null hypothesis when the alternative hypothesis is true. This is a correct decision; if there is a true treatment effect, Power is the probability of detecting that treatment effect. The Power can be calculated as one minus the probability of a Type II error, where the probability of a Type II error is denoted as beta.

In our example study, power answers the question "If the effects of the treatments do differ, what is the probability of detecting such a difference?".

In practice, we want the Power of a study to be high. We typically design our studies to have Power of at least 80%.

Example

- A complete deck of cards (ignoring jokers) is made up of 26 black cards and 26 red cards
- Null Hypothesis: deck of cards is complete (black/red)
- Alternative Hypothesis: deck of cards is incomplete based on card color (black/red)
- Experiment: draw 10 cards from the shuffled deck without replacement and observe the color of the card

When conducting a hypothesis test to determine whether there is significant evidence based on the observed data that will lead us to reject the null hypothesis, we calculate probabilities called p-values. To help us understand the calculation and interpretation of a p-value, let's consider a simple example.

Consider a study where we want to determine if a deck of cards is fair.

A complete, or fair, deck of cards (ignoring jokers) is made up of 26 black cards and 26 red cards.

Our null hypothesis states that the deck of cards is complete (black/red).

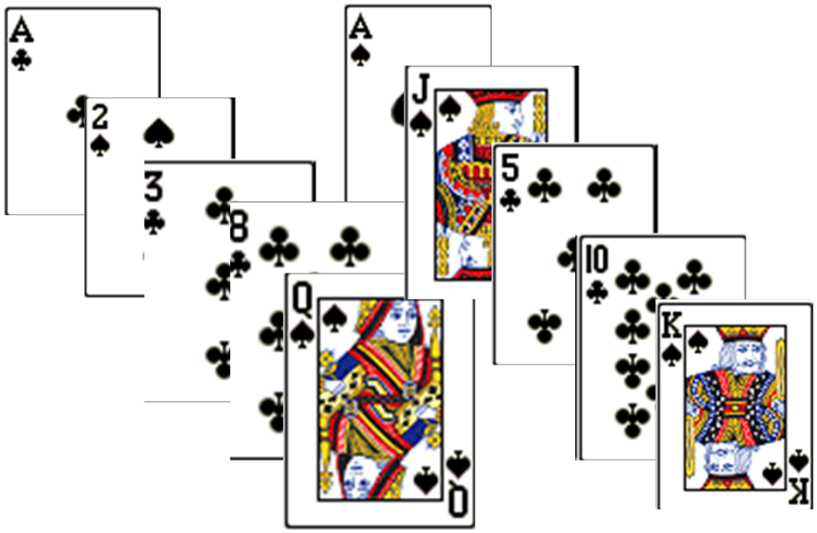
The alternative hypothesis is that the deck of cards is incomplete or unbalanced based on card color (black/red).

Consider a study where we randomly draw 10 cards from the shuffled deck without replacement and observe the color of the card.

If the deck is fair, we would expect, on average across multiple replications of this

study, to see 5 black and 5 red cards.

Observed Data



In the first 5 draws, we observed 5 black cards. This seems unusual and would not be very likely if the deck was in fact balanced. We are beginning to doubt the claim that the deck is fair.

From 10 draws, we have 10 black cards. This is even more unusual and leads us to doubt that the deck is fair.

Even if the deck was fair, we could possibly draw 10 black cards after 10 draws; it is possible, but not very likely. We can determine the probability of observing 10 black cards out of 10 draws by calculating a probability under the assumption of a fair deck.

Example

- Among the 10 cards, all are black.
- What is the probability of observing 10 black cards among 10 drawn cards under the assumption that the deck is fair (balanced between black and red)?
  - $\frac{26}{52} \times \frac{25}{51} \times \frac{24}{50} \times \frac{23}{49} \times \frac{22}{48}$ =0.025 [5 cards, all black]
  - $\frac{26}{52} \times \frac{25}{51} \times \frac{24}{50} \times \frac{23}{49} \times \frac{22}{48} \times \frac{21}{47} \times \frac{20}{46} \times \frac{19}{45} \times \frac{18}{44} \times \frac{17}{43}$ =0.00034 [10 cards, all black]
- Drawing 10 black cards among the 10 cards drawn is not likely to occur if the deck was fair.
- Data such as these were not very likely to have arisen under the null hypothesis

We can calculate the probability of observing 10 black cards among 10 drawn cards under the assumption that the deck is fair (balanced between black and red).

For the first 5 draws, the probability is the product of observing a black card on each draw.

The probability of observing a black card on the first draw is 26/52. Then, if a black card is drawn, there are 25 black cards remaining among 51 cards. Then, if two black cards have been drawn, there are 24 black cards remaining among 50 cards and so on.

The probability of observing 5 black cards out of 5 draws from a fair deck is 0.025 – not very likely.

Considering all 10 draws, the probability of observing 10 black cards out of 10 draws from a fair deck is 0.00034 – even less likely.

Observing 10 black cards out of 10 draws is not likely under the assumption that the deck is fair. The observed data would lead us to reject the claim that

the deck is fair.

P-value

- The probability of obtaining a difference at least as extreme as that obtained, provided the two groups are really equal (null hypothesis is true)
- The probability that an observed difference in outcome is due to chance alone.
- *Statistical Significance*: If the p-value of the calculated statistic is less than the alpha set in advance by the researcher (usually 0.05), then we can conclude the groups are different.

A p-value is the probability of obtaining a difference at least as extreme as that observed, provided the two groups are really equal (null hypothesis is true).

This is a conditional probability that is calculated under the assumption that the null hypothesis is true. In other words, this is the probability that an observed difference in outcome between groups is due to chance alone.

A Statistically Significant Difference is defined as a difference or association that is not likely to arise under the null hypothesis. We set a threshold of 5% to define what is "not very likely". The probability threshold used to define statistical significance is called the alpha level.

If the p-value of the calculated statistic is less than the alpha set in advance by the researcher (usually 0.05), then we can conclude the groups are different.

Statistical Significance

- *P* value $\leq \alpha$ implies statistical significance.
- Statistical significance: ability to state that the observed difference (or association) in outcome is not due to chance alone.
- Statistical significance is necessary for clinical significance but says nothing about the magnitude of the effect.

If the p-value is less than or equal to the alpha level, we declare the result to be statistically significant, meaning, not likely to be observed under the null hypothesis or due to chance alone. A significant result leads us to reject the null hypothesis.

Note that statistical significance does not necessarily mean clinical significance. A small p-value does not provide information about the magnitude of the effect that could provide information about the clinical importance of the difference.

Clinical Significance

- Smallest clinically important difference between two treatments
- Based on clinical judgment of the magnitude of the difference
- Clinician should take into account the side effects, long-term complications, and other costs of the two treatments

Clinical significance is defined as the smallest clinically important difference between two treatment groups.

Clinical significance is based on clinical judgment of the magnitude of the difference that is important to patients or would be large enough to change practice.

Clinicians should take into account the side effects, long-term complications, and other costs of the two treatments when determining a clinically-significant threshold.

Statistical vs. Clinical Significance

- Not all statistically significant differences are clinically significant!
- Confidence intervals can address both clinical and statistical significance

Note that not all statistically significant differences are clinically significant. For example, with a very large sample size, we would have sufficient power to detect small effect sizes that may not be clinically important.

In order for an effect to be clinically significant, the estimate does need to be statistically significant (e.g., different from 0 or no difference between groups).

In practice, confidence intervals provide information about both clinical significance as well as statistical significance.

Summary

- Utilize hypothesis testing to make decisions (fail to reject the null hypothesis or reject the null hypothesis)
- Error in hypothesis testing: type I and type II errors
- Not all statistically significant results are clinically important; however, to be clinically significant, a result does need to be statistically significant

In summary, we have discussed key principles related to hypothesis testing.

We discussed how to utilize hypothesis testing to make decisions regarding research hypotheses. We will either fail to reject the null hypothesis or reject the null hypothesis in favor of the alternative hypothesis.

In practice, we will either make a correct decision regarding the hypotheses or will commit one of two errors: a type I error where we reject a true null hypothesis (i.e., declare a difference to be statistically significant when it is not; a false positive error) or we fail to reject a false null hypothesis (i.e., fail to detect a true difference between groups; a false negative result).

Finally, when interpreting results, keep in mind that not all statistically significant results are clinically important; however, to be clinically significant, a result does need to be statistically significant