www.epilepsy.va.gov/Statistics

Statistics in Evidence Based Medicine

Lecture 5: Vocabulary of Inferential Statistics Hypothesis Testing

Rizwana Rehman, PhD
Regional Statistician
Southeast Epilepsy Center of Excellence
Durham VA Medical Center, Durham NC
Rizwana.Rehman@va.gov
(919)286-0411 ext: 5024

Epilepsy Center of Excellence
SOUTHEAST

Inferential Statistics

- The branch of statistics dealing with conclusions, generalizations, predictions, and estimations based on data from small data sets
- Answers the following questions.
  - How far the computed value of a quantity of interest from a small data set be from the broader population's value?

    Confidence Interval
  - Is there a statistically significant difference between the computed quantity (statistic) and the population parameter?

    Hypothesis Testing

Overview

- Review of previous talk
  - confidence intervals
- Hypothesis testing
  - Vocabulary
  - Six steps
  - *p* value
  - Type I & type II errors
  - Power of a test
  - Sample size for a powerful study

Review Sampling Distribution

- Suppose all possible samples of size n from a population are drawn.
- A statistic of interest is calculated from each sample.
- The probability distribution of this series of statistics is called a sampling distribution.
- Sampling distributions of many statistics are similar to theoretical distributions.
  - Sampling distribution of mean is normal.

Review of Standard Error or Sampling Error

- A series of samples drawn from a population shows variation; Variation depends on variation in population and sample size.
- Standard error (SE) is a measure of the variability that a constant (population parameter) would be expected to show during sampling. It provides a measure of how far from the true value the estimate is likely to be.

Review Confidence Intervals

An interval of values bounded by confidence limits within which the true value of a population parameter is stated to lie with a specified probability

CI= statistic critical value standard error

95% C.I for the mean

$$C.I = \bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

Critical value is a number related to the level of confidence we want. The typical values for level of confidence are 90%, 95% & 99%.

Hypothesis Test

- Hypothesis testing addresses the question about whether any observed difference is real or spurious.
- Hypothesis testing involves a null and an alternative hypothesis along with performing a test to determine which hypothesis should be concluded.

Vocabulary of Hypothesis Testing

- Null and Alternative Hypotheses
- One Tailed & Two Tailed Tests
- Test Statistic
- Level of Significance
- Critical Value

Null & Alternative Hypotheses

- A null hypothesis H_o says that there is no difference
- An alternative hypothesis H_A says that there is a difference
- Reject H_o ⟹ Accept H_A
- Can't reject H_o ⟺ Don't accept H_o

Example Hypotheses

- H_o: $\mu = 20$

  H_A: $\mu \neq 20$

  Two tailed test No direction specified

- H_o: $\mu = 20$

  H_A: $\mu > 20$

  One tailed test Only interested in one direction

Difference between one tailed and two tailed tests

Test Statistic

A statistic used to determine whether a hypothesis will be accepted or rejected.

z, t and chi-square are examples of test statistics

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

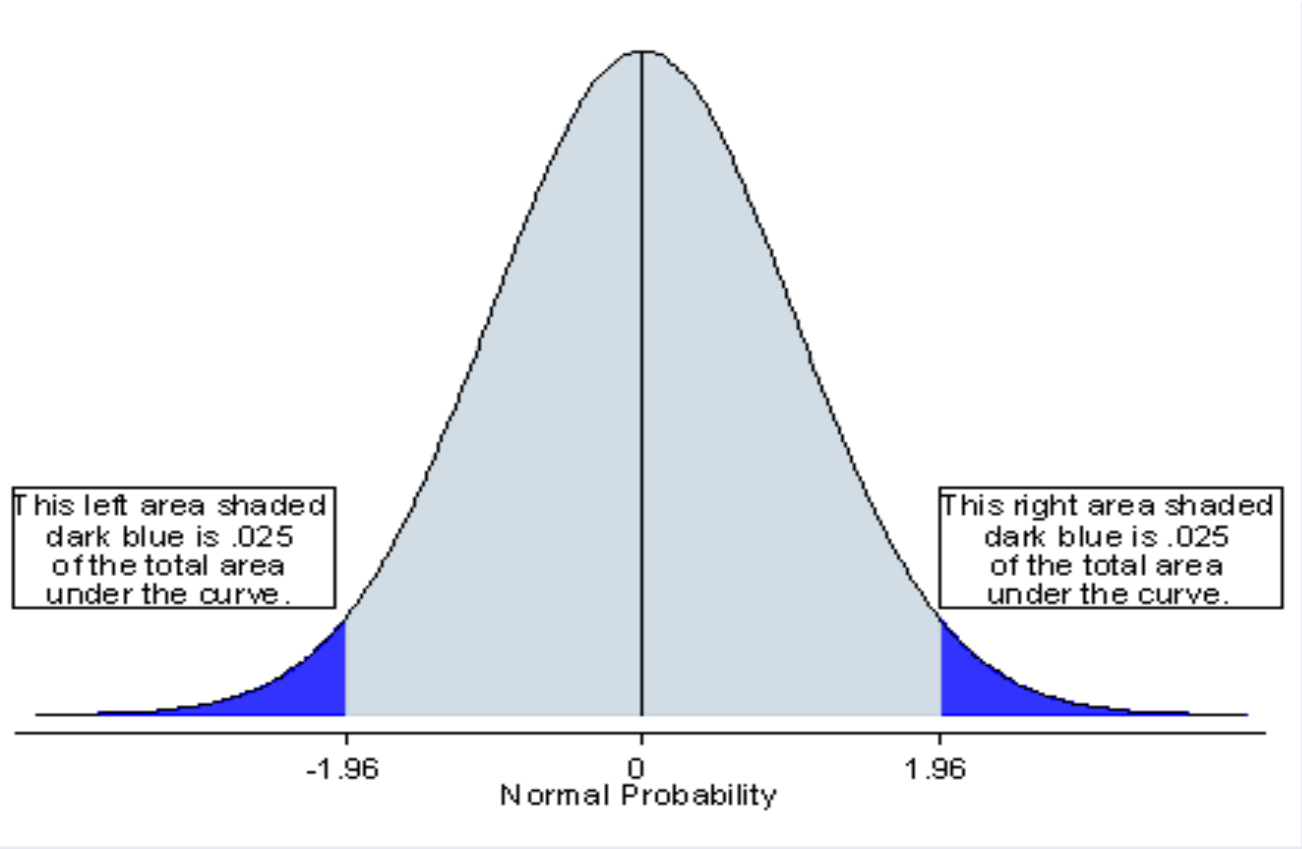
$$\text{Test Statistic} = \frac{\text{Observed value} - \text{Expected value}}{\text{Standard Error}}$$

Level of Significance

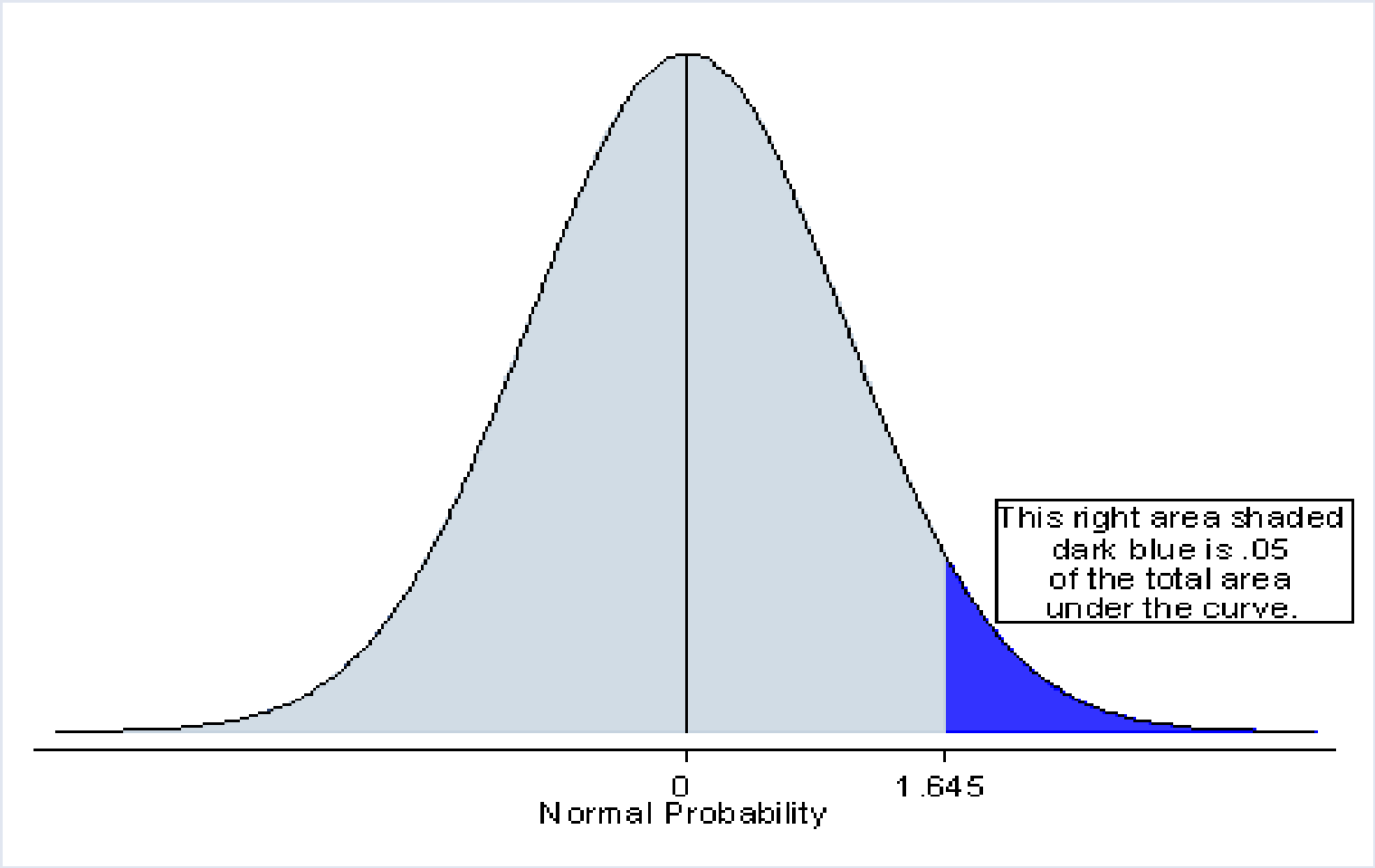
- Gives the probability of incorrectly rejecting the null hypothesis when it is actually true.
- Also known as type 1 error rate
- Traditional values are 0.05, 0.01, 0.001
- Denoted by alpha α

α=1-level of confidence

A Two Tailed Test (H_A: $\mu \neq 20$) at Significance Level 0.05



A One Tailed Test (H_A: $\mu > 20$) at Significance Level 0.05



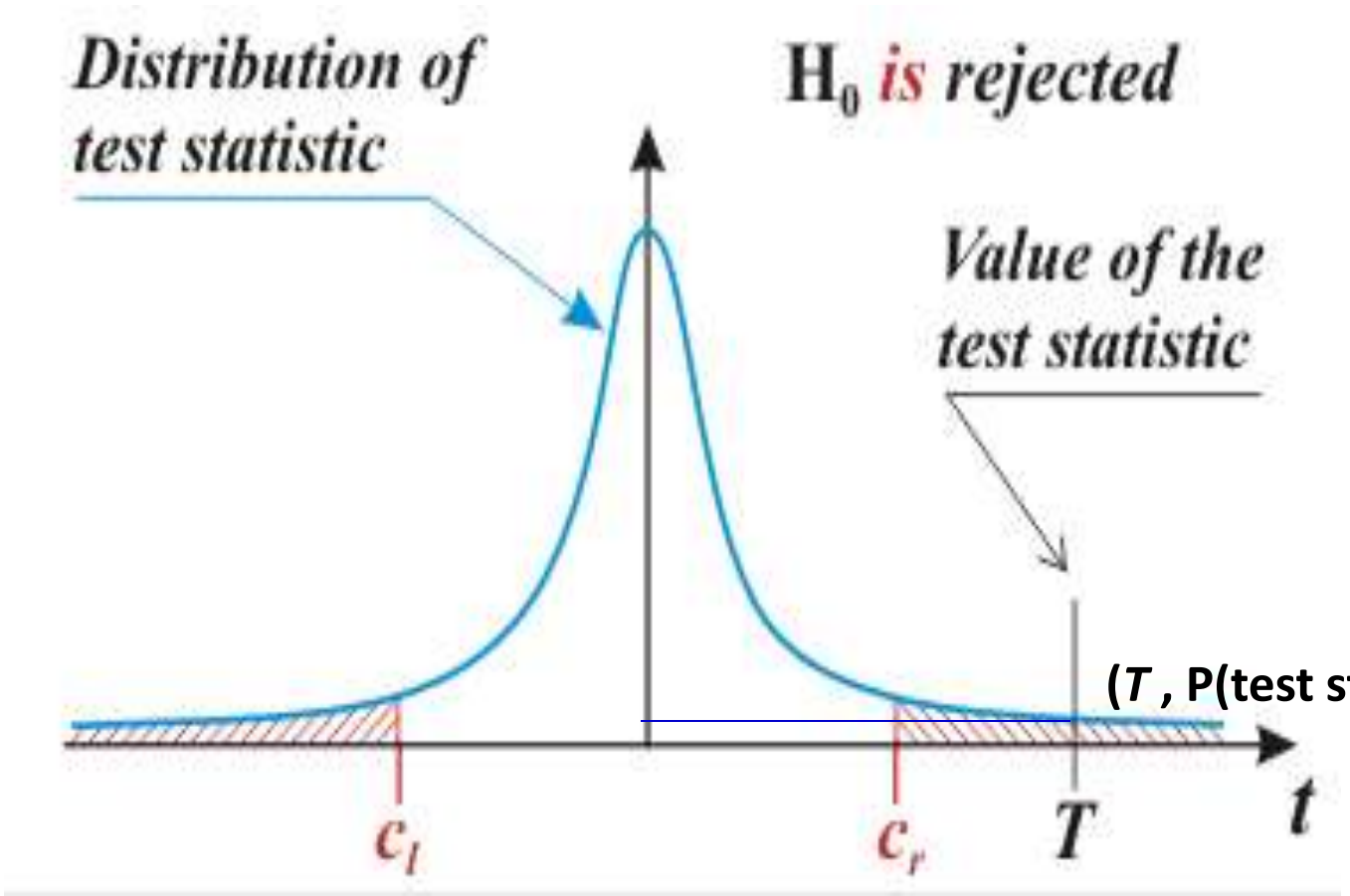
Critical Value

- Each test statistic has a distribution. This distribution is divided into acceptance and rejection regions. Critical value is the dividing line between the regions.
- In hypothesis testing, the value of some test statistic is compared with a critical value; the null hypothesis is rejected when the value of test statistic falls within the rejection region.
- The critical value depends on the significance level and whether the test is one-or two-sided.

Critical Value and Degrees of Freedom

- Critical value for many test statistics is computed from degrees of freedom d.f
- d.f = the number of values in the sample **minus** the number of population parameters that must be estimated from sample data
- The formula for d.f depends on the test statistic, the number of samples etc.
  - one sample t test **n-1** d.f
  - two sample t test $\mathbf{n_1+n_2-2}$ d.f
  - chi square test for independence **(# of rows-1)(# of columns-1)** d.f

p Value and Test Statistic



For a two tailed test *p* = 2 P(test statistic *T*)

Six Steps of Hypothesis Testing

- Construct null and alternative hypotheses
- Decide on an appropriate test statistic
- Select the level of significance
- Determine the critical value
- Compute the test statistic and compare with critical value
- State the conclusion about rejecting or not rejecting the null hypothesis

p Value

- *p* value is related to a hypothesis test. It is the probability of obtaining a result as extreme as (or more extreme than) the one observed, if the null hypothesis is true.
- Probability that the observed result is due to chance alone.
- *p* value is calculated after the statistical test has been performed.

$$p < \alpha \Rightarrow \text{reject } H_o$$

Reading and Reporting *p* Value

- $p = 0.05$ is moderate evidence against null hypothesis, $p = 0.01$ is moderate to strong, $p = 0.001$ strong to very strong.
- When computer output is 0.000, then report *p* value as $p < 0.001$.
- Don't report $p > 0.05$ as n.s, instead report exact *p* value.
- Don't report more than two significant figures for *p* value.

Misinterpretation of *p* Value

- *p* value is not the probability that the null hypothesis is true.
- *p* = 0.05 does not mean there is a 5% chance of type I error.
- *p* = 0.05 does not mean that the results would replicate if the study were repeated.
- *p* > 0.05 does not mean there is no difference between groups.

Confidence Interval and *p* Value

When a 95% C.I is constructed for a population parameter such as mean, then all values in C.I are considered plausible values of parameter. When the value of parameter specified by the null hypothesis is outside the 95% C.I the null hypothesis is rejected. In this case $p < 0.05$.

Confidence Intervals & Hypothesis Testing are Complimentary

- C.I provides information about statistical significance as well as the direction and strength of the effect.
- Hypothesis testing using *p* value allows for a rapid decision for statistical significance, however this is an overly simplistic approach.
- With some known facts confidence intervals can be constructed from *p* values and vise versa. The two statistical concepts are complimentary.

Confidence Interval or P-Value?

Type I & Type II Errors

- Rejecting the null hypothesis when it is true is called type I error.

$$\alpha = \text{Probability of type I erorr} = P(\text{rejecting } H_0 \mid H_0 \text{ is true})$$

- Failure to reject the null hypothesis when it is false is called type II error.

$$\beta = \text{Probability of type II error} = P(\text{not rejecting } H_0 \mid H_0 \text{ is false})$$

Type I and type II errors are not independent. If one increases the other decreases and vice versa

Illustration of Errors

		The true situation is	
		H_0 is true	**H_A is true**
Action	Reject H_0	Type I error	Correct decision
	Can't reject H_0	Correct decision	Type II error

		The true situation is	
		Not guilty (H_0)	**Guilty (H_A)**
Action	Jury finds guilty (Accept H_A)	Type I error	Correct decision
	Jury finds not guilty (reject H_A)	Correct decision	Type II error

Another Way to Remember

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True Positive
Fail to reject null hypothesis	Correct outcome True Negative	Type II error False negative

Power of a Test

- Making the right decision
- P(rejecting the null hypothesis when it is actually false)
- power= 1- β

		Null Hypothsis	
		is false	is true
Test result	is significant	Power	Type I error
	is not significant	Type II error	

Power Facts

- Increasing α increases power.
- A one tailed test is more powerful.
- A larger sample size boosts the power.
- More tests on a data set, less powerful the study.

Multiple Tests Decrease Power

- Suppose 20 independent true null hypotheses are tested at α=0.05 for each test on a data set then the combined value of α is 0.64
- To get true value of α = 0.05, B*onferroni* correction can be made. If there are k independent hypotheses, then the actual type I error is α / k

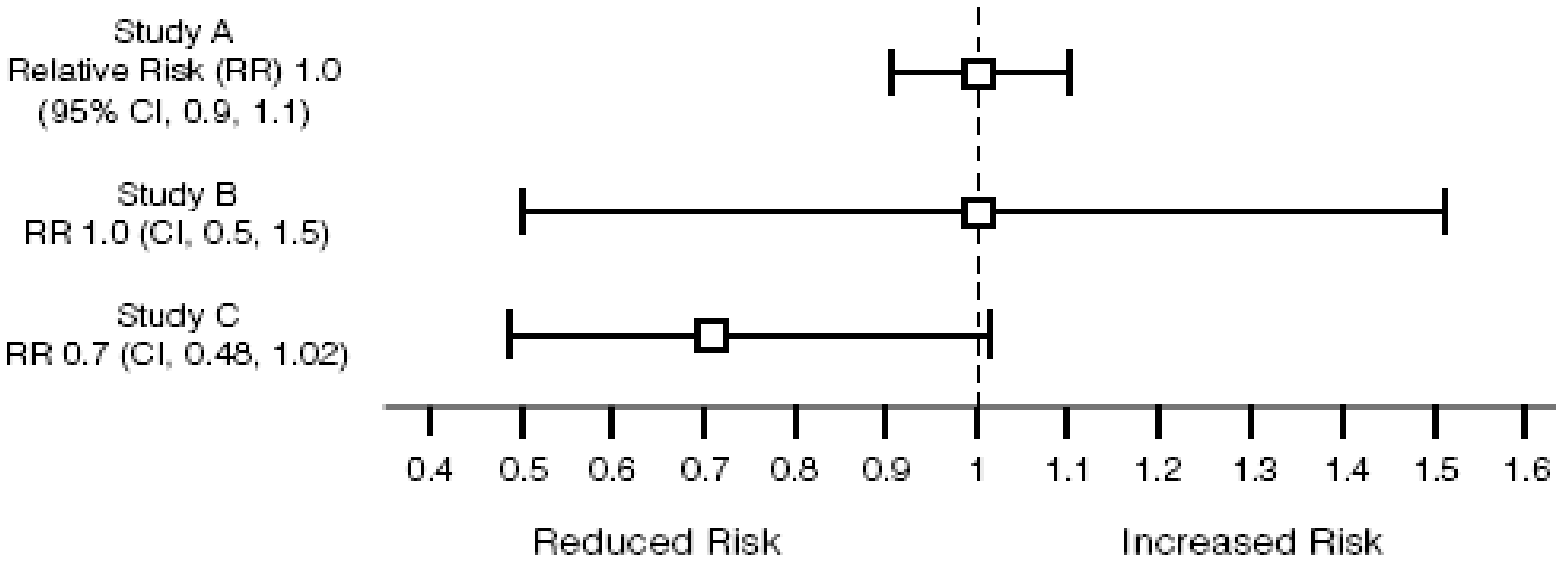
α decreases, β increases, power decreases

Confidence Intervals & Power

Study A: Low type II error, powerful

Study B: Type II error in either direction

Study C: A large type II error, weak power



Primer on type I and type II errors: Effective Clinical Practice, Nov/Dec 2001

Sample Size for a Powerful Study

- Enough persons should be included in a study to detect a significant difference if it exists.
- In calculation of sample size consider the following.
  - Variability in the data
  - Effect size (clinical significance)
  - Significance level
  - Power
  - Sidedness of the test

www.epilepsy.va.gov/Statistics

Questions/Comments

Rizwana.Rehman@va.gov

(919) 286-0411 ext: 5024

Next lecture's highlights

Research questions about one group