

## Module 2: Introduction to Statistics

Niko Kaciroti, Ph.D.  
BIOINF 525 Module 2: W16  
University of Michigan

### Topic

- Estimation
  - Point Estimates, Interval Estimates
- Hypothesis testing
  - Type I and Type II Error
- Comparing means
  - Parametric test: t-test
  - Nonparametric test: Wilcoxon test
- Power and sample size calculations

## Topic

- **Estimation**
  - Point Estimates, Interval Estimates
- Hypothesis testing
  - Type I and Type II Error
- Comparing means
  - Parametric test: t-test
  - Nonparametric test: Wilcoxon test
- Power and sample size calculations

## Estimation: Point Estimate

- **Point Estimate.** Let,  $x_1, x_2, \dots, x_n$ , be a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ 
  - A point estimate for a parameter of interest  $\theta$  is a function of the sample, named  $\hat{\theta}$ , which is a good approximation of  $\theta$
- Good estimates are unbiased and have minimum variance (error)
  - Unbiased
    - $E(\hat{\theta}) = \theta$
  - Minimum variance
    - If  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are both unbiased estimates of  $\theta$ , then the one with the smallest error is a better estimate

## Example: Unbiased Estimates for $\mu$ and $\sigma^2$

- **Estimate of  $\mu$ :**  $\bar{x}$  and  $\tilde{x}$  are two unbiased estimates for the mean parameter  $\mu$

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} \quad E(\bar{x}) = \mu \quad se(\bar{x}) = \frac{s}{\sqrt{n}}$$

$$\tilde{x} = \frac{x_1 + x_2}{2} \quad E(\tilde{x}) = \mu \quad se(\tilde{x}) = \frac{s}{\sqrt{2}}$$

Which one is better?

- **Estimate of  $\sigma^2$ :**

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad E(s^2) = \sigma^2 \quad (\text{Unbiased})$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \quad E(\tilde{\sigma}^2) = \frac{n}{n-1} \sigma^2 \quad (\text{Approximately Unbiased})$$

## Interval Estimate

- **Interval Estimate** is derived by using the sample to calculate an interval  $(\hat{\theta}_L, \hat{\theta}_U)$  of possible values for  $\theta$ :  $(\hat{\theta}_L < \theta < \hat{\theta}_U)$ .
  - If  $(\hat{\theta}_{L1}, \hat{\theta}_{U1})$  and  $(\hat{\theta}_{L2}, \hat{\theta}_{U2})$  are two interval estimates with the same credibility for  $\theta$ , the one with the shortest length is better
- **Confidence Interval (CI)** is an interval estimate that is used to indicate the reliability of a point estimate.
  - 95%CI for  $\mu$ :  $[\bar{x} - 1.96 * se(\bar{x}), \bar{x} + 1.96 * se(\bar{x})]$  or  $\bar{x} \pm 1.96 * se(\bar{x})$
  - $(1-\alpha)\%$ CI for  $\mu$ :  $[\bar{x} - z_{\alpha/2} * se(\bar{x}), \bar{x} + z_{1-\alpha/2} * se(\bar{x})]$
  - $(1-\alpha)$  is the confidence level, that is, how frequently the CI contains  $\mu$  if the experiment was repeated many times

Which is wider 95%CI or 99%CI?

## Topic

- Estimation
  - Point Estimates, Interval Estimates
- **Hypothesis testing**
  - **Type I and Type II Error**
- Comparing means
  - Parametric test: t-test
  - Nonparametric test: Wilcoxon test
- Power and sample size calculations

## Hypothesis Testing

- In hypothesis testing usually there is a null hypothesis and an alternative hypothesis. The **null hypothesis** typically corresponds to a default position. It is referred to as  $H_0$ . The **alternative hypothesis** is referred to as  $H_A$
- For example, let  $\mu$  be the mean cholesterol for a population. A researcher wants to know if the mean cholesterol is equal (or bigger) to some value  $\mu_0 (=130)$ ?

$H_0: \mu = \mu_0$                       Two sided hypothesis testing

$H_A: \mu \neq \mu_0$

$H_0: \mu > \mu_0$                       One sided hypothesis testing

$H_A: \mu \leq \mu_0$

## Hypothesis Testing

- Hypothesis testing is done by collecting data and then quantifying in probability terms how likely it is to have observed these data if  $H_0$  were true
- If the collected data were unlikely to have occurred under the  $H_0$ , then reject  $H_0$
- Quantifying how likely it is to have observed the sample data is done via a statistical test, and based on probability distribution theory (e.g. *t-test*, *Chi-square*, *F-test*)

## Type I Error and Type II Errors

- When testing a hypothesis the unknown parameters are estimated, which will result in an estimation error. Thus, there will be potential errors for rejecting or not rejecting the null hypothesis  $H_0$
- **Type I Error** is the incorrect rejection of a true  $H_0$ 
  - $Pr(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha$
- **Type II Error** is failure to reject a false  $H_0$ 
  - $Pr(\text{Do not Reject } H_0 | H_A \text{ is true}) = \beta$

## Type I Error and Type II Errors

- A **type I error** is a false positive result. For example, showing that a particular gene relates to some disease when that's not the case.  
Or convicting an innocent person
- A **type II error** is a false negative result. An example of type II error would be a failing to show that a particular gene relates to some disease when it really does  
Or failing to convict a guilty person
- All statistical hypothesis tests have a probability of making type I and type II errors. It's important to keep Type I error low

## Type I Error and Type II Errors

- What is the Type I and Type II error for a test that always rejects  $H_0$ ?  



$$Pr(\text{Reject } H_0 | H_0 \text{ is true})=1$$

$$Pr(\text{Do not Reject } H_0 | H_A \text{ is true})=0$$
- What is the Type I and Type II error for a test that never rejects  $H_0$ ?  

$$Pr(\text{Reject } H_0 | H_0 \text{ is true})=0$$

$$Pr(\text{Do not Reject } H_0 | H_A \text{ is true})=1$$
- A good test has both small  $\alpha$  and small  $\beta$ .  $\alpha$  is usually set at 0.05 or 0.01; and  $\beta$  is usually set at 0.2 or 0.1

## Errors When Making Inferences

		True Condition	
		$H_0$ False	$H_0$ True
Decision made from Data Analysis	Reject $H_0$	 Correct Power = $1-\beta$	Type I Error $\alpha$
	Fail to Reject $H_0$	Type II Error $\beta$	 Correct

## Topic

- Estimation
  - Point Estimates, Interval Estimates
- Hypothesis testing
  - Type I and Type II Error
- **Comparing means**
  - **Parametric test: t-test**
  - Nonparametric test: Wilcoxon test
- Power and sample size calculations

## Comparing Means: Parametric Test

### Student's t-test

- Student's t-test is commonly used for comparing means and it is based on the T-distribution. Depending on the hypothesis one of the following three tests are used
  1. One-sample t-test
  2. Paired t-test
  3. Two-sample t-test
- The t-test is a parametric test. It is based on the parametric assumption that the data have a normal distribution

## One-Sample t-test

- Let,  $x_1, x_2, \dots, x_n$ , be an independently normally distributed sample from a population with mean  $\mu$  and sampling variance  $s^2$ . We want to test whether the population mean  $\mu$  is equal to some value  $\mu_0$
- Then, the null and alternative hypotheses are:

$$H_0: \mu = \mu_0$$

vs.

$$H_A: \mu \neq \mu_0$$



## One-Sample t-test

How to build a statistical test for:  $H_0: \mu = \mu_0$  vs.  $H_A: \mu \neq \mu_0$

- First, use  $\bar{x}$  to estimate  $\mu$ , then compare  $\bar{x}$  to  $\mu_0$ . If the difference,  $\bar{x} - \mu_0$ , is large it suggests that  $\mu \neq \mu_0$ . So, reject  $H_0: \mu = \mu_0$  for “large” values of  $\bar{x} - \mu_0$
- How large should  $\bar{x} - \mu_0$  be to reject  $H_0$ ?
  - It depends on the margin of the error,  $se(\bar{x})$ . The larger the error the larger the  $\bar{x} - \mu_0$  should be

Q: Which difference  $\bar{x} - \mu_0$  would be considered “larger”?

(a) 10 with  $se(\bar{x})=1$  or (b) 100 with  $se(\bar{x})=100$

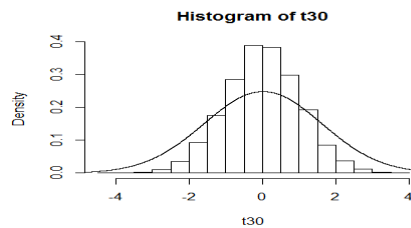
## One-Sample t-test

- Calculate the difference relative to the error,  $\frac{\bar{x}-\mu_0}{se(\bar{x})}$ , if the ratio is large, then reject  $H_0: \mu = \mu_0$
- Next, define in probability terms, what values of  $\frac{\bar{x}-\mu_0}{se(\bar{x})}$  are considered large to reject  $H_0$

## Using T-distribution for Testing One-sample t-test

- If  $H_0$  is true, then the following test statistics has a T-distribution with  $n-1$  df:

$$t = \frac{\bar{x} - \mu_0}{se(\bar{x})} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$



- T-distribution is symmetric around 0, and very large/small values are less likely to occur. So, if  $H_0$  was true one would expect to see  $t$  (or  $t$ -values) that are not too far from 0

## Using T-distribution for Testing One-sample t-test

- If  $t$ -value is large,  $|t| \geq C$ , then  $H_0$  is rejected ( $\mu \neq \mu_0$ ).  $C$  is referred to as critical value and is defined such that  $Pr(|T| \geq C) = \alpha$  (usually  $\alpha = 0.05$ , for which  $C \approx 2$ )

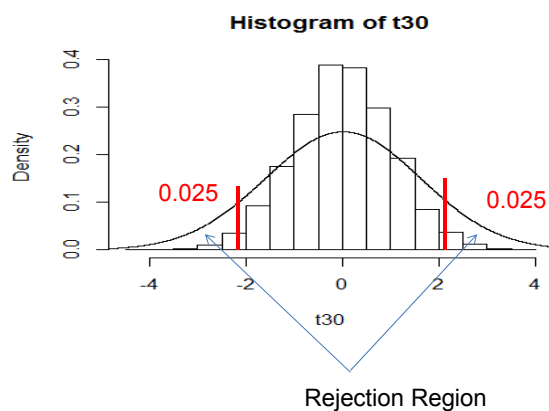
- If  $|t\text{-value}| > 2$  reject  $H_0$

- Alternatively,  $p$ -value is used as a criteria for rejecting  $H_0$ 
  - $p$ -value is defined as  $Pr(|T| \geq t\text{-value}) = p\text{-value}$ . For any  $t$ -value the  $p$ -value is calculated using  $t$ -distribution tables or statistical programs

- If  $p\text{-value} < 0.05$  reject  $H_0$

## T-distribution and the Rejection Region

- Let  $t = \frac{\bar{x} - \mu_0}{se(\bar{x})}$ . The rejection region, is the set of t-values for which  $H_0$  is rejected. The area under the curve for the values in the rejection region is the Type I error and equals to 0.05

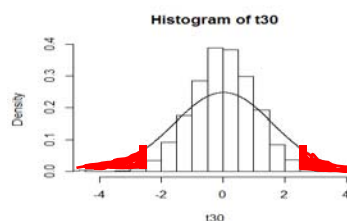


## Meaning of the p-value

- A  $p$ -value is *not* the probability that  $H_0$  is true. On the contrary, the  $p$ -value is based on the assumption that  $H_0$  is true
- A  $p$ -value is an estimate of the probability that the observed result, or a more extreme result, could have occurred by chance, *if  $H_0$  were true*

*E.g. if  $t=2.5$  then*

$$p\text{-value} = \Pr(|T| \geq 2.5 | H_0) = .018$$



- The  $p$ -value is a measure of the credibility of the null hypothesis. A large  $p$ -value ( $> 0.05$ ) means that there is no compelling evidence to reject  $H_0$

## Example of One-Sample t-test

(TROPHY Data)

- $H_0$ : Mean of LDL = 130 vs.  $H_A$ : Mean of LDL  $\neq$  130

$$t = \frac{\bar{x} - 130}{se(\bar{x})} = \frac{\bar{x} - 130}{s/\sqrt{n}}$$

- mean(LDL)=123.06
- sd(LDL)=32.1
- n=length(LDL)=255

$$t = (123.06 - 130) / (32.1 / \sqrt{255}) = -3.453 \text{ (Reject } H_0)$$

## Example of One-Sample t-test in R

(TROPHY Data)

- $H_0$ : Mean of LDL = 130 vs.  $H_A$ : Mean of LDL  $\neq$  130

- Code in R: `t.test(LDL, m=130)`

- The output for this t.test is:

One Sample t-test data: LDL

`t = -3.4529, df = 254, p-value = 0.0006494`  $\longrightarrow$  `p < .05, Reject  $H_0$`   
 alternative hypothesis: true mean is not equal to 130  
`95 percent confidence interval: 119.1000 127.0176`  $\longrightarrow$  Does not include 130,  
 sample estimates: mean of x 123.0588 hence Reject  $H_0$

## Paired t-test

- Let,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  be a sample of pairs independently normally distributed with mean  $(\mu_1, \mu_2)$  and sampling standard deviation  $(s_1, s_2)$ . Here  $x_i$  and  $y_i$  are data on the same subject, for example,  $x_i$  is the BP measured before taking some medication and  $y_i$  is the BP 2 years after.
- These type of data are used to test whether there is a treatment effect by comparing the before  $(\mu_1)$  and after treatment  $(\mu_2)$  BP values.
- The null and the alternative hypotheses are:

$$H_0: \mu_1 = \mu_2$$

vs.

$$H_A: \mu_1 \neq \mu_2$$

## Paired t-test

- Paired t-test is derived using one-sample t-test on  $d_i = y_i - x_i$
- Calculate  $d_i = y_i - x_i$ , then  $d_1, d_2, \dots, d_n$ , are independently normally distributed, with mean  $\mu_d = \mu_2 - \mu_1$ . If there were no treatment effect  $\mu_d$  should be zero

- Then testing

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

is the same as testing:

$$H_0: \mu_d = 0$$

$$H_A: \mu_d \neq 0$$

- Use the one-sample t-test on  $d_1, d_2, \dots, d_n$

## Example of Paired t-test in R

(TROPHY Data)

- Let  $\mu_1$  and  $\mu_2$  be the mean of DBP for patients in treatment group at baseline (before treatment) and 2 years later (after treatment):
- $H_0: \mu_2 - \mu_1 = 0$  vs.  $H_A: \mu_2 - \mu_1 \neq 0$
- `t.test(DBP24,DBP0,paired=T)`

Paired t-test data: DBP0 and DBP24

`t = -6.712, df = 126, p-value = 5.862e-10`  $\longrightarrow$   $p < .05$ , Reject  $H_0$

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -4.686208 -2.552063

sample estimates: mean of the differences -3.619135

Does not contain 0,  
hence Reject  $H_0$

## Two-Sample t-test

- Let,  $x_1, x_2, \dots, x_{n1}$ , and  $y_1, y_2, \dots, y_{n2}$  be two samples independently normally distributed with mean  $\mu_1, \mu_2$  and sampling variance  $s_1, s_2$ . For example  $x$  is the sample for a treated group and  $y$  is the sample for a control (placebo) group. We want to test whether the means between two groups are equal or not. The null and alternative hypotheses are:

$$H_0: \mu_1 = \mu_2$$

Vs.

$$H_A: \mu_1 \neq \mu_2$$

## Two-Sample t-test

- If  $H_0$  is true, then the following test statistics has a T-distribution:

$$t\text{-value} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where,  $\bar{x} = \frac{\sum x_i}{n_1}$ ,  $\bar{y} = \frac{\sum y_i}{n_2}$ ,  $s_1^2 = \frac{\sum (x_i - \bar{x})^2}{n_1 - 1}$ , and  $s_2^2 = \frac{\sum (y_i - \bar{y})^2}{n_2 - 1}$ . The same argument described for the one-sample t-test applies here:

– **If |t-value| > 2 reject  $H_0$**

– **If p-value < .05 reject  $H_0$**

Note: If  $s_1$  and  $s_2$  are the same the t-test with equal variance is used.

## Example of Two-Sample t-test in R

(TROPHY Data)

- Let  $\mu_1$  and  $\mu_2$  be the mean of LDL for patients in treatment group and placebo group respectively
- $H_0: \mu_1 = \mu_2$  vs.  $H_A: \mu_1 \neq \mu_2$
- Code in R: `t.test(LDL~Trt)`

Welch Two Sample t-test data: LDL by Trt

t = -0.4579, df = 246.327, p-value = 0.6474 → p > .05 Do not Reject  $H_0$

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -9.768556 6.083147 → If contains 0, do not Reject  $H_0$

sample estimates: mean in group 1 mean in group 2

122.1339            123.9766

## Topic

- Estimation
  - Point Estimates, Interval Estimates
- Hypothesis testing
  - Type I and Type II Error
- **Comparing means**
  - Parametric test: t-test
  - **Nonparametric test: Wilcoxon test**
- Power and sample size calculations

## Nonparametric Test for Comparing Means

- The Student's t-test is a parametric test, it assumes that data are normally distributed, which may or may not be true
- Nonparametric test is an alternative to the t-test for comparing means. Nonparametric test does not assume normality for the data
- If the data are not normally distributed and have large outliers, a nonparametric test is preferred



## Nonparametric Test for Comparing Means Wilcoxon test

- Wilcoxon test is the most common used among nonparametric tests. As with t-test, there are three Wilcoxon tests related to a specific  $H_0$ 
  1. One-sample test: Wilcoxon sign rank test
  2. Paired test: Wilcoxon sign rank test
  3. Two-sample test: Wilcoxon sum rank test (aka Mann-Whitney test)

## Wilcoxon Test for Comparing Means From Two Samples

- Let's consider the following data on two samples:
  - X sample: **2,5,3,6**
  - Y sample: 3,7,9,8,9
- The  $H_0$  states that the data for the X and Y samples come from the same distribution with the same mean ( $\mu_x = \mu_y$ )
- If  $H_0$  is true then “X > Y” or “Y > X” are equally likely to occur:

$$Pr(X > Y) = Pr(Y > X)$$

That is, smaller (or larger) values are equally likely to be from either X or Y sample

## Wilcoxon Test for Comparing Means From Two Samples

- X sample: **2,5,3,6**
- Y sample: 3,7,99,8,9
- Rank all the data from smallest to largest:
  - Rank all data: **2** **3** 3 **5** **6** 7 8 9 99
  - Assign ranks: **1** **2.5** 2.5 **4** **5** 6 7 8 9
  - (In case of a tie, assign an average rank to each data point)
- Calculate  $R_1 = 12.5$  sum of ranks from X sample.  $R_2 = 32.5$  sum of ranks from Y sample
- If the low ranking is dominated by one sample (e.g.  $R_1 < R_2$ ), that is evidence against  $H_0$ , which is rejected using the p-value  $< 0.05$  criteria

## Example of Wilcoxon two-sample test in R (TROPHY Data)

- Let  $\mu_1$  and  $\mu_2$  be the mean of HDL for patients in the treatment group and the placebo group respectively
- $H_0: \mu_1 = \mu_2$  vs.  $H_A: \mu_1 \neq \mu_2$
- Code in R: `wilcox.test(HDL~Trt)`

Wilcoxon rank sum test with continuity correction

HDL by Trt W = 7574, **p-value = 0.3471**  
 alternative hypothesis: true location shift is not equal to 0

## Compare t-test vs. Wilcoxon test

- X sample: **2,5,3,6**
- Y sample: 3,7,99,8,9

- **wilcox.test(X,Y)** Wilcoxon rank sum test with continuity correction

data: x and y W = 2.5, **p-value = 0.0851**

- **t.test(X,Y)** Welch Two Sample t-test



data: x and y t = -1.1459, df = 4.02, **p-value = 0.3154**

- Usually Wilcoxon is better to use if n is small and there are outliers.  
If data are normally distributed, the t-test is better

## Topic

- Estimation
  - Point Estimates, Interval Estimates
- Hypothesis testing
  - Type I and Type II Error
- Comparing means
  - Parametric test: t-test
  - Nonparametric test: Wilcoxon test
- **Power and sample size calculations**

## Power of a Statistical Test

		True Condition	
		$H_0$ False	$H_0$ True
Decision made from Data Analysis	Reject $H_0$	 Correct Power = $1-\beta$	Type I Error $\alpha$
	Fail to Reject $H_0$	Type II Error $\beta$	 Correct

## Power of a Statistical Test

- **Power** is the probability of detecting a difference, when it exists, for a given Type I error, say  $\alpha=.05$ 
  - Power= $Pr(\text{Reject } H_0 | H_A \text{ is true})$
  - or
  - Power= $1-Pr(\text{Do not Reject } H_0 | H_A \text{ is true})=1-\beta$
- A good test should have sufficient Power, usually 80% or 90%

## What Impacts Power?

(e.g. two sample t-test)

- The two-sample t-test which tests  $H_0: \mu_1 = \mu_2$  is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{se(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

→ large  
→ small

- Power =  $Pr(\text{Reject } H_0 / H_A \text{ is true}) = Pr(|t| > C_\alpha)$ . Thus the larger the value of t the bigger the Power.
- t is large when:
  - $\bar{x}_1 - \bar{x}_2$  (or  $\mu_1 - \mu_2$ ) is large.
  - $se(\bar{x}_1 - \bar{x}_2)$  is small (more precision)
    - $s_1$  and  $s_2$  are small
    - $n_1$  and  $n_2$  are large (large sample)

## Examples: Which Experiment Will Have the Larger t-value?

	$\bar{x}_1$	$\bar{x}_2$	$s_1$	$s_2$	$n_1$	$n_2$	t-value
Exp 1a	40	50	30	30	100	100	?
Exp 1b	40	50	30	30	200	200	
Exp 1c	40	50	30	30	300	300	?
Exp 2a	40	47	30	30	100	100	
Exp 2b	40	47	30	30	200	200	
Exp 2c	40	47	30	30	300	300	
Exp 3a	40	50	20	20	100	100	
Exp 3b	40	50	20	20	200	200	
Exp 3c	40	50	20	20	300	300	

$$H_0: \mu_1 = \mu_2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{se(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

## Examples: Which Experiment Will Have the Larger t-value?

	$\bar{x}_1$	$\bar{x}_2$	$s_1$	$s_2$	$n_1$	$n_2$	t-value
Exp 1a	40	50	30	30	100	100	-1.67
Exp 1b	40	50	30	30	200	200	<b>-2.36</b>
Exp 1c	40	50	30	30	300	300	<b>-3.33</b>
Exp 2a	40	47	30	30	100	100	-1.17
Exp 2b	40	47	30	30	200	200	-1.65
Exp 2c	40	47	30	30	300	300	<b>-2.33</b>
Exp 3a	40	50	20	20	100	100	<b>-2.5</b>
Exp 3b	40	50	20	20	200	200	<b>-3.54</b>
Exp 3c	40	50	20	20	300	300	<b>-5</b>

$$H_0: \mu_1 = \mu_2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{se(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

## Power Calculations

- To calculate Power (using R or other programs) for comparing the means between two groups the following input is required:
  - Number of samples in each group :  $(n_1, n_2)$
  - Difference in the mean you want to detect:  $\delta = \mu_1 - \mu_2$  or  $(\bar{x}_1 - \bar{x}_2)$
  - Standard deviation of each group:  $(s_1, s_2)$
  - Type I error:  $(\alpha = 0.05)$

## Power Calculations

$\alpha=0.05$	$\bar{x}_1$	$\bar{x}_2$	$s_1$	$s_2$	$n_1$	$n_2$	Power
Exp 1a	40	50	30	30	100	100	?
Exp 1b	40	50	30	30	200	200	
Exp 1c	40	50	30	30	300	300	
Exp 2a	40	47	30	30	100	100	
Exp 2b	40	47	30	30	200	200	
Exp 2c	40	47	30	30	300	300	
Exp 3a	40	50	20	20	100	100	
Exp 3b	40	50	20	20	200	200	
Exp 3c	40	50	20	20	300	300	

## Power Calculations

$\alpha=0.05$	$\bar{x}_1$	$\bar{x}_2$	$s_1$	$s_2$	$n_1$	$n_2$	Power
Exp 1a	40	50	30	30	100	100	65%
Exp 1b	40	50	30	30	200	200	91%
Exp 1c	40	50	30	30	300	300	98%
Exp 2a	40	47	30	30	100	100	38%
Exp 2b	40	47	30	30	200	200	64%
Exp 2c	40	47	30	30	300	300	81%
Exp 3a	40	50	20	20	100	100	94%
Exp 3b	40	50	20	20	200	200	>99%
Exp 3c	40	50	20	20	300	300	>99%

In R: `Power.t.test(n=100,delta=10,sd=30)` → Power 0.65

## How to Increase Power?

- Compare groups with larger difference in their means
- Choose conditions with less variability (small  $s_1, s_2$ )
- Use larger sample sizes (large  $n_1, n_2$ )
  - The larger the  $n$  the larger the Power, still  $n$  is constrained by the budget.
  - How large should  $n$  be?
- Sample size calculation is used to calculate how large  $n$  should be for a study

## Sample Size Calculations

- The sample size calculations can be seen as the reverse of Power calculations
- In Power calculations
  - The sample size is fixed, say  $n=100$ , then one calculates the Power
- In sample size calculations
  - The Power is fixed, say 80%, then one calculates the  $n$  needed to achieve 80% Power
- Sample size calculation is important when designing new studies
  - If  $n$  is too small, it will result in not enough Power to detect differences
  - If  $n$  is too large, it will result in more than sufficient Power; hence a waste of resources



## Sample Size Calculations

$\alpha=0.05$	$\bar{x}_1$	$\bar{x}_2$	$s_1$	$s_2$	$n_1$	$n_2$	Power
Exp 1a	40	50	30	30	?	?	80%
Exp 1b	40	50	30	30			90%
Exp 2a	40	47	30	30			80%
Exp 2b	40	47	30	30			90%
Exp 3a	40	50	20	20			80%
Exp 3b	40	50	20	20			90%

To calculate sample size

Need:  $\text{delta} = \bar{x}_1 - \bar{x}_2$   
 sd=  
 alpha=  
 Power=

## Sample Size Calculations

$\alpha=0.05$	$\bar{x}_1$	$\bar{x}_2$	$s_1$	$s_2$	$n_1$	$n_2$	Power
Exp 1a	40	50	30	30	142	142	80%
Exp 1b	40	50	30	30	190	190	90%
Exp 2a	40	47	30	30	289	289	80%
Exp 2b	40	47	30	30	387	387	90%
Exp 3a	40	50	20	20	64	64	80%
Exp 3b	40	50	20	20	85	85	90%

To calculate sample size

Power.t.test(delta=10,sd=30,power=.8)

Need:  $\text{delta} = \bar{x}_1 - \bar{x}_2$   
 sd=  
 alpha=  
 Power=

↓  
 n=142.24

## Summary Points

**Point Estimate for  $\theta$ :** Unbiased  $E(\hat{\theta}) = \theta$   
 Minimum Variance (error)  $se(\hat{\theta})$  is small.

**95%CI for  $\mu$ :**  $[\bar{x} - 1.96 \cdot se(\bar{x}), \bar{x} + 1.96 \cdot se(\bar{x})]$

**Hypothesis testing:**  $H_0$ : (Default status)  $\mu_1 = \mu_2$  vs.  $H_A$ :  $\mu_1 \neq \mu_2$

**Type I Error:** Incorrect rejection of a true  $H_0$   
 $Pr(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha$

**Type II Error:** Failure to reject a false  $H_0$   
 $Pr(\text{Do not Reject } H_0 | H_A \text{ is true}) = \beta$

**p-value:** Measures the credibility of the null hypothesis.  
 A small *p-value* (e.g.,  $p < 0.05$ ) results in rejection of  $H_0$

## Summary Points

### Test for Mean Comparison

Hypothesis	Parametric test	Nonparametric test
$H_0: \mu = \mu_0$ (One sample)	One-sample t-test	Wilcoxon sign rank test
$H_0: \mu_{\text{before}} = \mu_{\text{after}}$ (Matched pairs)	<i>Paired test based on one-sample</i>	
$H_0: \mu_1 = \mu_2$ (Two independent samples)	Two-sample t-test	Wilcoxon sum rank test (Mann-Whitney test)

## Summary Points

- **Power Analysis**

Calculates what is the statistical power (*i.e.* probability) to reject  $H_0$  in favor of a given  $H_A$ , for a given sample size  $n$  and Type I error. A statistical test is considered to have good Power if:

$$\text{Power} = \Pr(\text{Reject } H_0 \mid H_A \text{ is true}) \geq 80\%$$

- **Sample Size Calculation**

Calculate the required sample size for a study, to be able to reject the  $H_0$  in favor of a given  $H_A$ , with a given Power (80%) and for a given Type I error  $\alpha=0.05$ .