

Biost 517 / Biost 514

Applied Biostatistics I /

Biostatistics I



Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 7:

Two-sample inference: Hypothesis testing and
confidence intervals for differences in means

Motivation: Using Statistics to Answer ...Scientific Questions...

- Scientific questions are often concerned with the relationship between two (or more) variables of interest
- Insight is often provided by using statistical methods to assess the association between variables
 - Response variable** (outcome, dependent variable)
 - Predictor of Interest** (grouping variable, “independent” variable)
 - The scientific question is addressed by comparing the distribution of the response variable across groups that are defined by the predictor/grouping variable

Two Sample Setting



- When the predictor/grouping variable is binary, statistical methods for two-sample inference can be used for detecting differences in the response variable
 - Treatment versus control
 - Smokers and non-smokers
 - Etc.

Test to Compare Two Means



In this lecture we will cover the following:

- Two-Sample Problems and a Comparison to Matched Pairs
- Comparing Two Means when σ is Known
- Comparing Two Means when σ is Unknown
- Pooled Two-Sample t Procedures

Two-Sample Problem



- The goal of two-sample inference is to compare the responses in two groups
- Each group is considered to be a sample from a distinct population
- We will consider the two-sample program setting when responses in each group are independent of those in other other group (in addition to being independent of each other)
- There are statistical methods that can allow for dependencies within and across groups

Two-Sample Problem: Example



For example, suppose we have a SRS of size n_1 drawn from a $N(\mu_1, \sigma_1^2)$ population and an independent SRS of size n_2 drawn from a $N(\mu_2, \sigma_2^2)$ population.

- The first sample might be individuals who receive a treatment in a randomized controlled study and the second sample might be individuals who receive the placebo.
- We are interested in comparing the distribution of an outcome of interest to assess if there is any association between the outcome and treatment
- To do this, it is common to assess evidence that the means of the distributions for the two groups are the same. We can test $H_0 : \mu_1 = \mu_2$ against $H_A : \mu_1 \neq \mu_2$.

Two-Sample vs. Matched Pairs



How is the two-sample problem different from the *matched pairs design*? In a matched pairs design, subjects are matched in pairs and the outcomes are compared within each matched pair, e.g., we compute the difference for each pair.

For the two-sample problem:

1. There is no matching of the units in the two samples.
2. The two samples may be of different size.

We will cover the matched pairs design later in the course.

Comparing Two Means when σ is Known



Suppose we have a SRS of size n_1 drawn from a $N(\mu_1, \sigma_1^2)$ population (with sample mean \bar{x}_1) and an independent SRS of size n_2 drawn from a $N(\mu_2, \sigma_2^2)$ population (with sample mean \bar{x}_2).

Suppose σ_1 and σ_2 are known.

What is the distribution of $\bar{X}_1 - \bar{X}_2$?

What is the mean of $\bar{X}_1 - \bar{X}_2$?

What is the variance of $\bar{X}_1 - \bar{X}_2$?

Distribution of the Difference of Two Sample Means



Distribution of $\bar{X}_1 - \bar{X}_2$

- For sufficiently large sample sizes n_1 and n_2 , it can be shown using the Central Limit Theorem that $\bar{X}_1 - \bar{X}_2$ has a $N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
- The mean of $\bar{X}_1 - \bar{X}_2$ is $\mu_1 - \mu_2$
- The variance of $\bar{X}_1 - \bar{X}_2$ is $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
- The standard deviation of $\bar{X}_1 - \bar{X}_2$ is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Testing Equality of Means when σ_1 and σ_2 are known



The **two-sample z -statistic** is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

- To test the hypothesis $H_0 : \mu_1 = \mu_2$ (or equivalently $\mu_1 - \mu_2 = 0$) , we use

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \text{ under } H_0$$

The p -value is calculated as before

$$H_A : \mu > \mu_0 \quad P(Z \geq z)$$

$$H_A : \mu < \mu_0 \quad P(Z \leq z)$$

$$H_A : \mu \neq \mu_0 \quad 2P(Z \geq |z|)$$

Confidence intervals for the Difference of Means when σ_1 and σ_2 are known



A $(1 - \alpha)$ CI for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

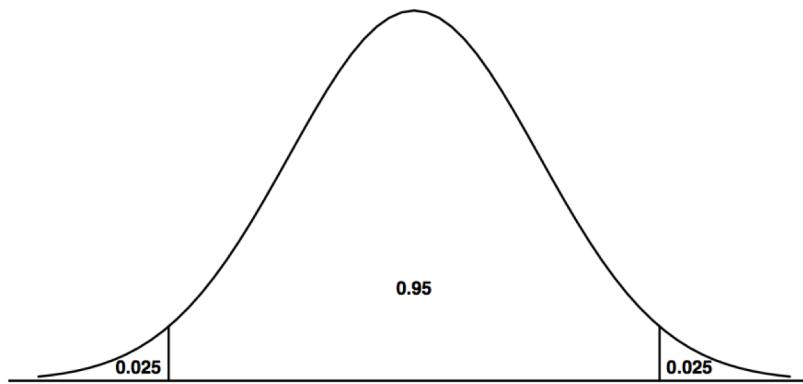
where z^* is the $\alpha/2$ critical value of the standard normal distribution.

Finding critical values for CI for the z-statistic is the same as in the one sample setting



For a given confidence level $(1 - \alpha)$, how do we find z^* ?

Let $Z \sim N(0, 1)$:



$$P(-z^* \leq Z \leq z^*) = (1 - \alpha) \iff P(Z < -z^*) = \frac{\alpha}{2}$$

Thus, for a given confidence level $(1 - \alpha)$, we can look up the corresponding z^* value on the Normal table.

Common z^* values:

Confidence Level	90	95	99
z^*	1.645	1.96	2.576

Comparing Two Means when σ is Unknown



- In most cases, the variance/standard deviation of the outcome variable of interest will be unknown for both populations
- As such, we will need to estimate σ_1 and σ_2 from the sample
- The sampling distribution of the two-sample test statistic will no longer be normally distributed when σ_1 and σ_2 are unknown, as we saw in the one sample case

Two-sample t-test: allowing for heteroscedasticity



Comparing Two Means with σ Unknown: Allowing for Unequal Variances

The **two-sample t -statistic** allowing for **heteroscedasticity**, or unequal variances across groups, is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_k$$

- This statistic does not have an exact t distribution, it has an approximate t_k distribution with $k = \min(n_1 - 1, n_2 - 1)$.
 - Note that statistical software packages use a more complicated formula for a more accurate calculation of the degrees of freedom (which does not have to be an integer)

Two-sample t-test: allowing for heteroscedasticity



- To test the hypothesis $H_0 : \mu_1 = \mu_2$, we use

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \underset{\text{under } H_0}{\sim} t_{(k)}$$

The p -value is calculated as before.

$$H_a : \mu > \mu_0 \quad P(t_{(k)} \geq t)$$

$$H_a : \mu < \mu_0 \quad P(t_{(k)} \leq t)$$

$$H_a : \mu \neq \mu_0 \quad 2P(t_{(k)} \geq |t|)$$

This approximate procedure is conservative. That is, the CIs are longer than the true CI and p -values are larger than the true p -values.

Confidence intervals for the Difference of Means when σ_1 and σ_2 are unknown



- A $(1 - \alpha)$ CI for $\mu_1 - \mu_2$ is given by

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

where t^* is the $\alpha/2$ critical value of the t_k distribution.

- Similar to the one-sample case, the critical value, t^* , is chosen such that $100(1 - \alpha)\%$ of the area under the $t_{(n-1)}$ density lies between $-t^*$ and t^* .
- The area to the right of t^* should be $\frac{\alpha}{2}$ and the area to the left of $-t^*$ should be $\frac{\alpha}{2}$.

Weak Null Hypothesis vs. Strong Null Hypotheses



- For the t-statistic we have introduced that allows for heterogenous variances, we consider $H_0: \mu_1 = \mu_2$ for some outcome variable of interest for two populations, while allowing for $\sigma_1 = \sigma_2$.
- This is a **weak null hypothesis**; as we are only testing one aspect of how the distribution of our variable varies across groups, i.e, differences in the means
- A **strong null hypothesis** would state that the **distribution** of our outcome variable of interest is identical in both groups under the null hypothesis.

Weak Null Hypothesis vs. Strong Null Hypotheses



- If the weak null is rejected (e.g. there is a difference in means across groups using a t-test that allows for heteroscedasticity) we can also reject the strong null (e.g. there is a difference in the distribution across groups)
- But if our tests fails to reject the the weak null, this alone is not sufficient to "accept" the strong null
 - while there may not be differences in means across groups, there but there could be differences in variances across groups so the strong null hypothesis that the distributions are the same across groups is not true
- Formally, the strong null implies the weak null, but the weak null doesn't imply the strong null

Variance of Difference of Same Means: Assuming Equal Variances



- For the two-sample t-test that allows for heterogenous variances, we considered a weak null hypothesis of $H_0 : \mu_1 = \mu_2$ while allowing for $\sigma_1 \neq \sigma_2$.
- For a strong null hypothesis we would assume that the distributions are the same in the two populations, such that $H_0 : \mu_1 = \mu_2$ and $\sigma_1 = \sigma_2 = \sigma$
- We have that under the strong null hypothesis, the variance of $\bar{X}_1 - \bar{X}_2$ is

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

and the SD is $\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Two-Sample Pooled Variance: Assuming Equal Variances



- In practice, we don't know σ^2 and will need to estimate it. If the strong null hypothesis is true, we can improve our estimate of σ^2 by *pooling* the two samples together when estimating the variance as follows:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- So, using a pooled estimate of σ^2 , the variance of $\bar{X}_1 - \bar{X}_2$ under the strong null hypothesis is estimated to be $s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$ and the SD is estimated to be $s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Two-sample t-test: presuming homoscedasticity



The t statistic is then

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

- A $(1 - \alpha)$ CI for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where t^* is the $\alpha/2$ critical value of the $t_{(n_1+n_2-2)}$ distribution

- To test the hypothesis $H_0 : \mu_1 = \mu_2$, we use

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \text{ under } H_0$$

The p -value is calculated as before.

Two-sample t-test: presuming homoscedasticity



The **two-sample t -statistic** that presumes **homoscedasticity**, or equal variances across groups, is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \text{ under } H_0$$

- To test the hypothesis $H_0 : \mu_1 = \mu_2$ presuming homoscedasticity, we use

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \text{ under } H_0$$

- The p -value is calculated as before.

$$H_a : \mu > \mu_0 \quad P(t_{(n_1+n_2-2)} \geq t)$$

$$H_a : \mu < \mu_0 \quad P(t_{(n_1+n_2-2)} \leq t)$$

$$H_a : \mu \neq \mu_0 \quad 2P(t_{(n_1+n_2-2)} \geq |t|)$$

Confidence intervals for the Difference of Means Presuming homoscedasticity



- A $(1 - \alpha)$ CI for $\mu_1 - \mu_2$ when presuming homoscedasticity is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where t^* is the $\alpha/2$ critical value of the $t_{(n_1+n_2-2)}$ distribution

Two-Sample Example



Weight gains (in kg) of babies from birth to age one year are measured. All babies weighed approximately the same at birth. Group A are babies that were breast fed and Group B are babies that were given formula.

Group A	5	7	8	9	6	7	10	8	6
Group B	9	10	8	6	8	7	9		

Assume that the samples are randomly selected from independent normal populations. Is there any difference between the true means of the two groups?

- i) Assume $\sigma_1 = \sigma_2 = 1.5$ is known. What type of test is used?
- ii) Assume σ_1 and σ_2 are unknown and unequal. What type of test is used?
- iii) Assume σ_1 and σ_2 are unknown but equal. What type of test is used.

Two-Sample Example



State the hypothesis:

$$H_0 : \mu_1 = \mu_2 \quad H_a : \mu_1 \neq \mu_2$$

where μ_1 is the true population mean of the Group A
and μ_2 is the true population mean of Group B.

Group A	5	7	8	9	6	7	10	8	6
Group B	9	10	8	6	8	7	9		

$$\bar{x}_1 = 7.33 \quad \bar{x}_2 = 8.14$$

$$s_1 = 1.58 \quad s_2 = 1.35$$

$$n_1 = 9 \quad n_2 = 7$$

Two-Sample Example: Assume σ is known and equal for the two populations



- i) Assume $\sigma_1 = \sigma_2 = 1.5$ is known. Then, the two-sample z statistic is

$$\begin{aligned} z &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_1 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{7.33 - 8.14}{1.5 \times \sqrt{\frac{1}{9} + \frac{1}{7}}} = -1.07 \end{aligned}$$

The two-sided p -value is

$$2P(Z \geq |z|) = 2P(Z \geq 1.07) = 0.28$$

where $Z \sim N(0, 1)$.

So there is no difference between the true population mean of these two group at the significance level 0.1.

Two-Sample Example: Assuming σ is unknown and allowing for heteroscedasticity

ii) Assume σ_1 and σ_2 are unknown and allow them to be unequal (heteroscedasticity). Then, the two-sample t statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{7.33 - 8.14}{\sqrt{\frac{1.58^2}{9} + \frac{1.35^2}{7}}} = -1.10$$

The two-sided p -value is

$$2P(T \geq |z|) = 2P(T \geq 1.10) = 0.31$$

where $T \sim t_6$.

Example: Confidence Interval Assuming σ is unknown and allowing for heteroscedasticity



A 90% confidence interval for $\mu_1 - \mu_2$ is given by

$$\begin{aligned}(\bar{x}_1 - \bar{x}_2) &\pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\&= (7.33 - 8.14) \pm 1.94 \times \sqrt{\frac{1.58^2}{9} + \frac{1.35^2}{7}} \\&= (-2.23, 0.61)\end{aligned}$$

where $P(|T| < t^*) = 0.90$. That is, $P(T > t^*) = 0.05$

Two-Sample Example: Assuming σ is unknown and assuming homoscedasticity



iii) Assume σ_1 and σ_2 are unknown but equal.

The pooled two-sample estimator of σ is

$$\begin{aligned}s_p &= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \\&= \sqrt{\frac{(9 - 1) \times 1.58^2 + (7 - 1) \times 1.35^2}{9 + 7 - 2}} \\&= 1.54\end{aligned}$$

Thus, the pooled two-sample t statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{7.33 - 8.14}{1.54 \sqrt{\frac{1}{9} + \frac{1}{7}}} = -1.04$$

The two-sided p -value is given by

$$2P(T \geq |t|) = 2P(T \geq 1.04) = 0.32 \quad \text{where } T \sim t_{14}.$$

Example: Confidence Interval assuming σ is unknown and assuming homoscedasticity

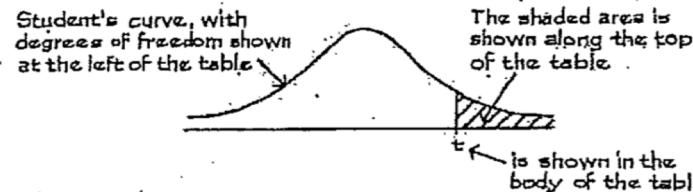


A 90% confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned}(\bar{x}_1 - \bar{x}_2) &\pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\&= (7.33 - 8.14) \pm 1.76 \times 1.54 \times \sqrt{\frac{1}{9} + \frac{1}{7}} \\&= (-2.18, 0.56)\end{aligned}$$

Where $P(|T| < t^*) = 0.90$. That is, $P(T > t^*) = 0.05$, where $T \sim t_{14}$.

Student's *t* Distribution



Degrees of freedom	25%	10%	5%	2.5%	1%	0.5%
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17
11	0.70	1.36	1.80	2.20	2.72	3.11
12	0.70	1.36	1.78	2.18	2.68	3.05
13	0.69	1.35	1.77	2.16	2.65	3.01
14	0.69	1.35	1.76	2.14	2.62	2.98
15	0.69	1.34	1.75	2.13	2.60	2.95
16	0.69	1.34	1.75	2.12	2.58	2.92
17	0.69	1.33	1.74	2.11	2.57	2.90
18	0.69	1.33	1.73	2.10	2.55	2.88
19	0.69	1.33	1.73	2.09	2.54	2.86
20	0.69	1.33	1.72	2.09	2.53	2.85
21	0.69	1.32	1.72	2.08	2.52	2.83
22	0.69	1.32	1.72	2.07	2.51	2.82
23	0.69	1.32	1.71	2.07	2.50	2.81
24	0.68	1.32	1.71	2.06	2.49	2.80
25	0.68	1.32	1.71	2.06	2.49	2.79

R: Two-sample t-test allowing for heteroscedasticity

- Can also perform a two-sample t-test in R using the same **t.test()** function in R that is also used for the one sample t test, except that you provide two variable to the function
- Consider a data object named “*data1*” that contains a response variable named *respvar* and a predictor variable named *predvar* that is a factor predictor variable with two levels giving the corresponding groups.
- To perform a two-sample t-test of equality of means for A response variable *respvar* and predictor variable *predvar* defining the two groups the following commands can be used:

```
t.test (respvar~predvar, data=data1)
```

- The default **t.test()** function in R conducts a t-test allowing for unequal variances with a two-sided alternative hypothesis

R: Two-sample t-test allowing for heteroscedasticity

- The output from the **t.test()** function in R provides
 - Point estimates of the means for each of the two groups
 - p-values for a t-test that the means for the two groups are equal (i.e., that the difference in means is equal to 0)
 - 95% confidence intervals for the difference in means for the two groups
- Similar to the one-sample t-test in R, can also do a one-sided tests with the alternative option by using the “alternative=less” or “alternative=greater” options. Note that the default option is “alternative=“two.sided”, so you don’t need to specify this in the function if you want to do a two-sided test.

R: Two-sample t-test presuminghomoscedasticity....

- To conduct a t-test that presumes homoscedasticity, you must specify this using the “var.equal=TRUE” option in the t.test() function as follows:

```
t.test(respvar~predvar, var.equal=TRUE, data=data1)
```

Example: Cholesterol in Diabetic and Non-Diabetic Elderly Men Revisited

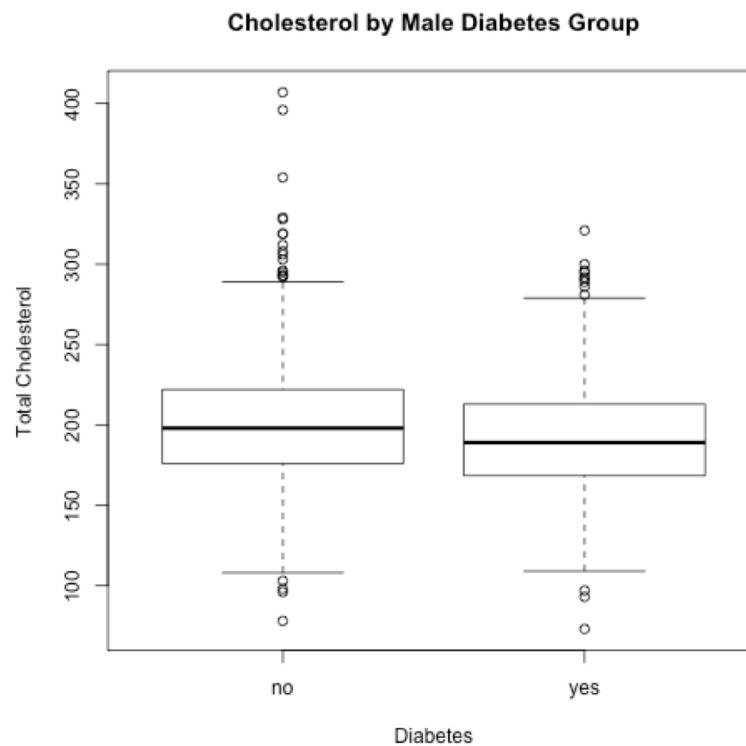
- In lecture 6, we investigated the distribution of total cholesterol levels in a study of Cardiovascular Health Study of elderly adults aged 65 years, 382 men diagnosed with diabetes at the time of enrollment and 1694 elderly men who were not diabetic at the time of enrollment.
- Here, we will perform a statistical test to assess if the data provide evidence that the mean cholesterol level in elderly diabetic men is different than mean cholesterol level in men who are not diabetic.
- We are interested in evaluated the following two Hypotheses:

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

Boxplots of Cholesterol by Diabetes Group

```
> inflamm=read.table("DataSets/inflamm.txt",header=T)
> inflamm$diabetes<-with(inflamm,ifelse(diab2==1,"yes","no"))
> inflamm$diabetes<-as.factor(inflamm$diabetes)
> plot(cholest~diabetes,main="Cholesterol by Diabetes Group",xlab="Diabetes",
ylab="Total Cholesterol",data=inflamm)
```



Testing Differences in Mean Cholesterol by Diabetes Group: allowing for heteroscedasticity



- T-test allowing for heteroscedasticity (or unequal variances) for cholesterol in the two groups defined by diabetes prognosis at the time of enrollment

```
> t.test(cholest~diabetes,data=subset(inflamm,male==1))
```

Welch Two Sample t-test

```
data: cholest by diabetes
t = 3.9221, df = 541.28, p-value = 9.908e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.101886 12.333515
sample estimates:
mean in group no mean in group yes
 199.6730          191.4553
```

Testing Differences in Mean Cholesterol by Diabetes Group: allowing for heteroscedasticity



- Alternatively, you could provide the `t.test()` function with the cholesterol values for each of the two groups to perform the two-sample t-test:

```
> cholestmalediab<-with(inflamm,cholest[diab2==1&male==1])
> cholestmaleN0diab<-with(inflamm,cholest[diab2==0&male==1])
> t.test(cholestmaleN0diab,cholestmalediab)
```

Welch Two Sample t-test

```
data: cholestmaleN0diab and cholestmalediab
t = 3.9221, df = 541.28, p-value = 9.908e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.101886 12.333515
sample estimates:
mean of x mean of y
199.6730 191.4553
```

Cholesterol Example: Scientific reporting of two-sample inference

- Scientific reporting of the results:
 - In our sample, elderly men without diabetes had a mean cholesterol of 199.7 mg/dl and standard deviation of 35.2 mg/dl and elderly men with diabetes had a mean cholesterol of 191.46 mg/dl with standard deviation 37.3 mg/dl. The difference in mean cholesterol in the sample is 8.24mg/dl, and a 95% CI suggests that the data are consistent with the elderly men with diabetes having true mean cholesterol that is between 4.1 mg/dl and 12.3 mg/dl higher than elderly diabetic men. Because the two sided P value is $P < .005$ for a two-sample t-test that allows for unequal variances, we reject the null hypothesis that mean cholesterol in elderly diabetic men is the same as meal cholesterol in elderly men who are not diabetic.
- **Question: What is a plausible explanation for why elderly men with diabetes have lower mean cholesterol than elderly men who do not have diabetes in this sample?**

Example : Testing Mean Differences in Cholesterol presuming homoscedasticity

```
> t.test(cholest~diabetes, var.equal=TRUE, data=subset(inflamm, male==1))
```

Two Sample t-test

data: cholest by diabetes

t = 4.0645, df = 2072, p-value = 4.992e-05

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

95 percent confidence interval:

4.252724 12.182676

sample estimates:

mean in group no mean in group yes

199.6730	191.4553
----------	----------

Final comments on assumptions for two-sample t-test and confidence intervals

- You are strongly advised to *almost always* use two-sample t-tests and two-sample confidence intervals that are obtained *without* requiring the strong assumption of equal variances across groups
- When a two-sample t-test yields different inference when making a strong assumption of homogenous variances versus allowing for unequal variances across groups, the assumption of equal variances across groups is likely wrong.
- Note that one should choose the analysis to conduct prior to analyzing the data to avoid a data driven analysis, and assuming equal variances in the groups is scientifically illogical in nearly all settings: we are interested in comparing the means for the two groups yet somehow we know that they have the same variance???
- Remember that a two-sample t-test allowing for heterogenous variances is an appropriate test when variances across groups are unequal as well as when the variances are equal. So it is appropriate in both settings.