GPPI PPOL 501–02 & -06: Fall 2014
Course Notes # 12: Inference for Two Population Means
Professor Carolyn Hill

I. Testing whether 2 Hypothesized Population Means are different, using information from 2 Sample Means *(for an interval-ratio variable of interest)*

- In the one-sample hypothesis tests that we've been working with up until now, we obtain information about one *sample* of individuals, organizations, or other unit of analysis. Our goal has been to make some inference to the *population* from which such a sample was drawn.

- We can extend these ideas further: for example, suppose we want to say something about how two different groups (each drawn from samples) differ along some dimension of interest.

- Once again, we're interested in making some generalization about all observations of such groups or types. However, because of cost or other considerations, we may not be able to gather information about *all* the members of *each population* of interest.

- Difference-of-mean tests can be conducted either for (i) independent samples; or for (ii) paired samples. In this class, we'll focus on independent samples. Put a reminder in your head that the test statistic will be different if you ever have a paired sample; and refer to Weiss section 10.5.

- Focusing now on inference of whether two population means are different for independent samples: In a single mean test, we were interested in the sampling distribution of sample means to make our inferences about a single sample to its population. Now we're interested in the *sampling distribution of the* ***difference in sample means***

  IF the population variances were known, the sample statistic would be:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}}$$

- But, as we've discussed, we seldom if ever know the population variance.

  ➔ So now we'll talk about two specific formulas. <u>Both assume that the population variances are unknown</u>.
  ➔ The two differ in this way: One assumes that the population variances are *equal*. The other assumes that the population variances are *unequal.*
  ➔ At the end of this set of notes, we'll discuss what to consider in using either test.

II. Population variances are *equal:* Use a *pooled* standard deviation

For the differences of means where population variances are assumed to be equal:

-- *d.f.* is: $d.f. = (n_1\text{-}1) + (n_2\text{-}1)$

-- and the test statistic is: $t = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}}$

Where the standard error is:

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = s_p * \sqrt{\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)} = \left[\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}} * \sqrt{\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)}\right]$$

Example 1: A researcher is interested in understanding whether families in central cities are significantly larger or smaller than suburban families, as measured by the number of children. Random samples from both types of areas are gathered and the following sample statistics are computed:

Descriptive statistics for number of children in:

	Suburban Families	Central City Families
$\bar{X}$	2.37	2.78
s	0.63	0.95
n	42	37

-- The steps for answering this two-sample question are exactly the same as in the one-sample case. We just use a slightly different null hypothesis format (to reflect the two-sample case), and a different standard deviation formula, noted above:

-- H_0: $\mu_{suburban} = \mu_{centcit} \Leftrightarrow \mu_{suburban} - \mu_{centcit} = 0$
H_1: $\mu_{suburban} \neq \mu_{centcit} \Leftrightarrow \mu_{suburban} - \mu_{centcit} \neq 0$

d.f. = (42+37-2) = 77

Let’s use a 95% confidence level (alpha =0.05, two-sided):

Because 77 d.f. isn't listed in the table, we'll be conservative and select the value for d.f. = 75: $t_{critical} = \pm 1.992$

-- Calculate $\hat{\sigma}_{\bar{X}_1-\bar{X}_2}$ (i.e., the estimated standard deviation of the sampling distribution of the difference in sample means):

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}} * \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \sqrt{\frac{48.7629}{77}} * \sqrt{0.050837} = 0.1794.$$

-- Back to the full test statistic:

$$t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\hat{\sigma}_{\bar{X}_1-\bar{X}_2}} = \frac{(2.37-2.78)-(0)}{0.1794} = -2.285$$

-- Compare this result to the critical value and make a decision: |-2.285| > |±1.992|, so we *reject the null hypothesis* at the 95% confidence level that families in the central city and in the suburbs are the same size. We conclude that there's a statistically significant difference (at the alpha =0.05 level) in the sizes of these families. It's unlikely that we would have seen as much of a difference as we did in these samples, if in fact there was truly no difference between the average family sizes of suburban and rural families.

-- Is the difference statistically significant at the alpha = 0.02 (2-sided) level?

II. Population variances are *unequal:* Use a *nonpooled* standard deviation

For the differences of means where population variances are assumed to be unequal:

-- *d.f.* is:

$$d.f. = \left(\frac{\left[(s_1^2/n_1)+(s_2^2/n_2)\right]^2}{\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{n_2-1}} \right)$$

-- and the test statistic is: $t = \dfrac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\hat{\sigma}_{\bar{X}_1-\bar{X}_2}}$

The standard error now has a different formula, however:

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\left(\frac{s_1^2}{n_1}\right)+\left(\frac{s_2^2}{n_2}\right)}$$

- Example 2: **Are college students who live in dorms significantly more or less involved in campus life than students who commute to campus?** Researchers drew random samples of residential students, and commuter students, and measured the average number of hours per week that students devoted to extracurricular activities.

  For the residential students, the sample average was 12.4 activities, with a sample size of 158 and a standard deviation of 2. For the commuter students, the average was 10.2, with a sample size of 173 and a standard deviation of 1.9.

  -- H$_0$: ⇔
  H$_1$: ⇔

  -- Confidence level: let's say 95% ($\alpha = 0.05$), two-sided test.

  -- d.f. = 322.49, so t-crit = ±1.968 (using d.f.=300)

  -- $\sigma_{\bar{X}_r - \bar{X}_c}$ is the standard deviation of the sampling distribution of the difference in sample means (i.e., the standard error for this sampling distribution). When the variances are not equal, we calculate the standard error as:

$$\sigma_{\bar{X}_r - \bar{X}_c} = \sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)} = \sqrt{\frac{2^2}{158} + \frac{1.9^2}{173}} = 0.2149.$$

  -- Don't stop here! We've only calculated the denominator of the *t*-statistic. Now, go back and plug in the standard deviation of the sampling distribution into the denominator of the test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} = \frac{(12.4 - 10.2) - (0)}{0.2149} = 10.237$$

  -- Compare the test statistic to the critical values and make a decision: |10.237|> |±1.968|. So we *reject the null hypothesis* that there is no difference between activity levels of residential and commuter students at the 95% confidence level.

  Q: Can we reject this hypothesis at a higher level of confidence (i.e., a smaller alpha value)? What is the approximate p-value for this test?

IV. Deciding whether to assume that Population Variances are Equal or Unequal

- Often (especially with large samples), you will reach the same substantive conclusion, and often the same statistical conclusion, with either form of this test.

- With smaller samples, however, the version of the test that you use may matter. So how should you make a decision about the equality / inequality of variances?

  → opinions vary on this matter

Idea: In lieu of certain knowledge about the population standard deviations and their equality, you could run a statistical test regarding the equality of standard deviations:

$$H_0: \quad \sigma_{residential} = \sigma_{commuter} \quad \Leftrightarrow \quad \sigma_{residential} - \sigma_{commuter} = 0$$
$$H_1: \quad \sigma_{residential} \neq \sigma_{commuter} \quad \Leftrightarrow \quad \sigma_{residential} - \sigma_{commuter} \neq 0$$

The test statistic for testing this null hypothesis uses an *F* distribution, where the test statistic is calculated as: $F = \frac{s_r^2}{s_c^2}$. Looking at the *p*-value of this test statistic can tell you whether reject, or fail to reject, the null at particular levels of statistical significance.

(this is the idea behind the "rule of 2" that Weiss mentions).

HOWEVER: Moore & McCabe (2002, p. 553) note:

> "Unlike the *t* procedures for means, the *F* test and other procedures for standard deviations are extremely sensitive to nonnormal distributions. This lack of robustness does not improve in large samples. It is difficult in practice to tell whether a significant *F*-value is evidence of unequal population spreads or simply evidence that the populations are not normal….we do not recommend use of inference about population standard deviations in basic statistical practice….It was once common to test equality of standard deviations as a preliminary to performing the pooled two-sample *t* test for equality of two population means. It is better practice to check the distributions graphically, with special attention to skewness and outliers, and to use the software-based two-sample *t* that does not require equal standard deviations."

HOWEVER, Weiss has slightly different guidance (p. 457):

> "If you are reasonably sure that the populations have nearly equal standard deviations, use a pooled *t*-procedure; otherwise, use a nonpooled *t*-procedure."

Weiss points out (p. 457):

> In theory, the pooled *t*-test requires that the population standard deviations be equal, but what if they are not? The answer depends on several factors. If the population standard deviations are not too unequal and the sample sizes are nearly the same, using the pooled *t*-test will not cause serious

> difficulties. If the population standard deviations are quite different, however, using the pooled *t*-test can result in a significantly larger Type I error probability than the specified one.
>
> In contrast, the nonpooled *t*-test applies whether or not the population standard deviations are equal. Then why use the pooled *t*-test at all? The reason is that, if the population standard deviations are equal or nearly so, then, on average, the pooled *t*-test is slightly more powerful; that is, the probability of making a Type II error is somewhat smaller.

V. HYPOTHESIS TESTING FOR TWO- SAMPLE MEANS IN STATA

- To show examples of how to do various kinds of hypothesis tests in STATA, we'll use the same NLS survey that we used earlier:

- Let's go back to the IQ score as a variable of interest. Now, instead of making inferences about the mean IQ of the entire population represented by this NLS sample, we'll ask the question:

Are IQ scores different (on average) for persons who had a library card in their home when they were 14 years old, compared to those persons who did not?

- H_0: $\mu_{IQ,WITHLIBCARD} = \mu_{IQ,WITHOUTLIBCARD} \Leftrightarrow \mu_{IQ,WITHLIBCARD} - \mu_{IQ,WITHOUTLIBCARD} = 0$

  H_1: $\mu_{IQ,WITHLIBCARD} \neq \mu_{IQ,WITHOUTLIBCARD} \Leftrightarrow \mu_{IQ,WITHLIBCARD} - \mu_{IQ,WITHOUTLIBCARD} \neq 0$

- Confidence level: let's say 90% ($\alpha = 0.10$), two-sided test. Given the sample sizes, we use a critical value of ± 1.645. Reminder: for an assumption of unequal population variances, degrees of freedom are calculated as:

$$d.f. = \left(\frac{\left[(s_1^2/n_1) + (s_s^2/n_2)\right]^2}{\frac{(s_1^2/n_1)^2}{(n_1 - 1)} + \frac{(s_s^2/n_2)^2}{n_2 - 1}} \right)$$

- Compute the test statistic: $t = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\hat{s}_{\bar{X}_1 - \bar{X}_2}}$

$$\hat{s}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{(n_1)} + \frac{s_2^2}{(n_2)}} = \sqrt{\frac{15.64602^2}{496} + \frac{14.822^2}{1559}} = 0.79652801.$$

Now, go back and plug in the standard deviation of the sampling distribution into the denominator of the test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{s}_{\bar{X}_1 - \bar{X}_2}} = \frac{(96.381 - 104.399) - (0)}{0.79652801} = -10.0661$$

Compare the test statistic to the critical value and make a decision: |-10.0661|> |±1.645|. So we *reject the null hypothesis* of no difference between of IQ scores between persons who had a library card in their homes when they were 14 years old, and persons who did not have a card, at the 90% confidence level.

Q: Can we reject this hypothesis at a higher level of confidence (i.e., a smaller alpha value)?

```
. ttest iq, by (libcrd14) unequal
Two-sample t test with unequal variances
```

- Alternatively, if we were to assume equal population variances, we would calculate the s.e. in this way:

$$s_{\bar{X}_1 - \bar{X}_2} = \left[\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}} * \sqrt{\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)} \right] = 15.0245905 * 0.05155158 = 0.7745414$$

and our *d.f.* = $(n_1-1)+(n_2-1)$ = (496-1)+(1559-1) = 2,053 d.f.

$t_{critical}$ = ± 1.96 for a 95% confidence level

Back to the test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{s}_{\bar{X}_1 - \bar{X}_2}} = \frac{(96.381 - 104.399) - (0)}{0.7745414} = -10.351867$$

Compare to the critical value: |-10.351867| > |±1.96|. Therefore, reject the null hypothesis at the 95% confidence level that there is no difference between the average IQ scores of youth who did and didn’t have library cards in the home.

```
. ttest iq, by (libcrd14)
Two-sample t test with equal variances
```

Question: Would you expect the p-value to be relatively large or relatively small, given this t-statistic?

The conclusion to reject the null hypothesis is the same conclusion we reached above, using the version of the test that did not assume equal standard deviations. With large samples, you'll usually reach the same conclusion no matter which standard error formula you use. With smaller samples, however, the version of the test that you use may matter.

- So far, in all of our tests we've obtained relatively large t-values. How about looking at another question: Do average IQ scores differ between persons who were married in 1976 and those who were not?

H_0: $\mu_{IQ,MARR76} = \mu_{IQ,NOTMARR76} \Leftrightarrow \mu_{IQ,MARR76} - \mu_{IQ,NOTMARR76} = 0$

H_1: $\mu_{IQ,MARR76} \neq \mu_{IQ,NOTMARR76} \Leftrightarrow \mu_{IQ,MARR76} - \mu_{IQ,NOTMARR76} \neq 0$

```
. use "J:\STATA datasets\card1.dta", clear
. do "C:\DOCUME~1\gppilab\LOCALS~1\Temp\STD01000000.tmp"
/* see creation of marr76 from Course Notes #9*/
```

EXAMPLE 1a

```
. summarize iq if libcrd14==.
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          iq |         6    97.66667    21.39782         64        125

. summarize iq if libcrd14==0
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          iq |       496    96.38105    15.64602         50        137

. summarize iq if libcrd14==1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          iq |      1559     104.399     14.8217         51        149


. ttest iq, by (libcrd14) unequal
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     496    96.38105     .702527    15.64602    95.00075    97.76135
       1 |    1559     104.399    .3753833     14.8217    103.6627    105.1353
---------+--------------------------------------------------------------------
combined |    2055    102.4637    .3398909    15.40797    101.7972    103.1303
---------+--------------------------------------------------------------------
    diff |           -8.017925     .796528               -9.581465   -6.454386
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -10.0661
Ho: diff = 0                     Satterthwaite's degrees of freedom =  797.357

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000


. ttest iq, by (libcrd14)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     496    96.38105     .702527    15.64602    95.00075    97.76135
       1 |    1559     104.399    .3753833     14.8217    103.6627    105.1353
---------+--------------------------------------------------------------------
combined |    2055    102.4637    .3398909    15.40797    101.7972    103.1303
---------+--------------------------------------------------------------------
    diff |           -8.017925    .7745415               -9.536894   -6.498956
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -10.3518
Ho: diff = 0                                     degrees of freedom =     2053

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

EXAMPLE 1B

```
. summarize iq if marr76==.

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          iq |         1          75           .         75         75

. summarize iq if marr76==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          iq |       563    102.8455    17.12055         53        149

. summarize iq if marr76==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          iq |      1497    102.3193    14.72705         50        145

. ttest iq, by(marr76) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     563    102.8455     .721545    17.12055    101.4282    104.2627
       1 |    1497    102.3193    .3806316    14.72705    101.5727    103.0659
---------+--------------------------------------------------------------------
combined |    2060    102.4631    .3396469    15.41563     101.797    103.1292
---------+--------------------------------------------------------------------
    diff |            .5261654    .8157865               -1.074918    2.127249
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.6450
Ho: diff = 0                     Satterthwaite's degrees of freedom =  892.349

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7404         Pr(|T| > |t|) = 0.5191          Pr(T > t) = 0.2596

. ttest iq, by (marr76)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     563    102.8455     .721545    17.12055    101.4282    104.2627
       1 |    1497    102.3193    .3806316    14.72705    101.5727    103.0659
---------+--------------------------------------------------------------------
combined |    2060    102.4631    .3396469    15.41563     101.797    103.1292
---------+--------------------------------------------------------------------
    diff |            .5261654    .7622282               -.9686536    2.020984
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.6903
Ho: diff = 0                                     degrees of freedom =     2058

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7550         Pr(|T| > |t|) = 0.4901          Pr(T > t) = 0.2450
```