

# BIOS 600: Principles of Statistical Inference

## Hypothesis Testing about Two Means

Fall 2012

# Reading

- ▶ Pagano and Gauvreau, Chapter 11
- ▶ North et al., “Examining a Comprehensive Model of Disaster-Related Posttraumatic Stress Disorder in Systematically Studied Survivors of 10 Disasters”

# Review: Hypothesis Testing Steps

- ▶ State Claim 1 and Claim 2. Claim 1 states “nothing unusual is happening” and Claim 2 challenges it.

## Review: Hypothesis Testing Steps

- ▶ State Claim 1 and Claim 2. Claim 1 states “nothing unusual is happening” and Claim 2 challenges it.
- ▶ Collect relevant data and summarize it.

# Review: Hypothesis Testing Steps

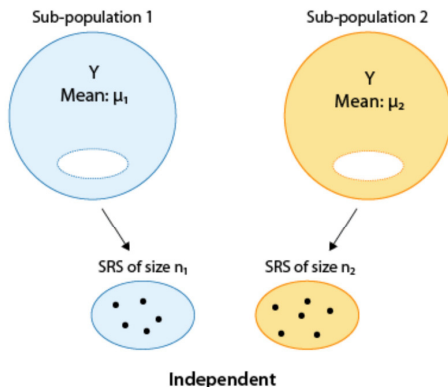
- ▶ State Claim 1 and Claim 2. Claim 1 states “nothing unusual is happening” and Claim 2 challenges it.
- ▶ Collect relevant data and summarize it.
- ▶ Assess how surprising it would be to see data like that *if Claim 1 is really true.*

# Review: Hypothesis Testing Steps

- ▶ State Claim 1 and Claim 2. Claim 1 states “nothing unusual is happening” and Claim 2 challenges it.
- ▶ Collect relevant data and summarize it.
- ▶ Assess how surprising it would be to see data like that *if Claim 1 is really true*.
- ▶ Draw conclusions.

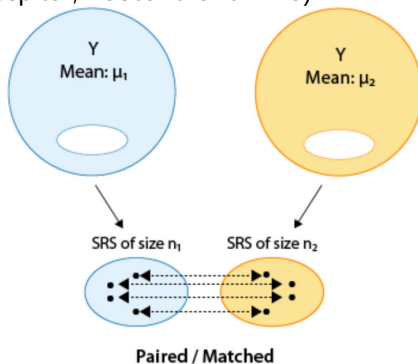
# Independent Samples

The type of t-test we use to compare two means depends on how the samples were obtained. One approach would be to obtain two independent samples and test the equality of means  $\mu_1$  and  $\mu_2$ .



## Paired or Matched Samples

An alternative would be to obtain paired or matched samples and test the equality of means  $\mu_1$  and  $\mu_2$ . Matching could be by person (e.g., before and after measures) or could be a pair of individuals who belong together in another way (e.g., same date of birth in same hospital, husband and wife).





## Interactive Case Study: Impaired Driving

The Department of Motor Vehicles wishes to compare impairment of drivers while texting to impairment after drinking two beers. Describe an independent samples design and a matched pairs design for this question of interest.

## Case Study: Personality or Looks?

A random sample of 25 college students answered the question “What is more important to you – personality or looks?” on a 25 point scale, where 1 indicates personality has maximum importance and looks don’t matter, and 25 indicates that looks have maximum importance and personality plays no role. The researcher was interested in whether gender plays a role in the importance of looks versus personality.

## Case Study: Hypothesis Testing Step 1

The null hypothesis is that males and females place the same emphasis on personality (score is unrelated to gender), while the alternative is that they do not (score is related to gender). Some folks may have prior experience leading them to want to make this a one-sided test, but let's not be so cynical to start!

$$H_0 : \mu_{MALE} = \mu_{FEMALE}$$

$$H_A : \mu_{MALE} \neq \mu_{FEMALE}$$

OR equivalently,

$$H_0 : \mu_{MALE} - \mu_{FEMALE} = 0$$

$$H_A : \mu_{MALE} - \mu_{FEMALE} \neq 0$$

## Case Study: Hypothesis Testing Step 2

The researcher enrolled 239 subjects and obtained complete data on 235, with 85 males and 120 females in the sample. Analyzing the data, we obtain  $\bar{x}_{MALE} = 13.33$ ,  $\bar{x}_{FEMALE} = 10.73$ ,  $s_{MALE} = 4.02$ ,  $s_{FEMALE} = 4.25$ .

## Case Study: Hypothesis Testing Step 2

The researcher enrolled 239 subjects and obtained complete data on 235, with 85 males and 120 females in the sample. Analyzing the data, we obtain  $\bar{x}_{MALE} = 13.33$ ,  $\bar{x}_{FEMALE} = 10.73$ ,  $s_{MALE} = 4.02$ ,  $s_{FEMALE} = 4.25$ .



"I am a marvellous housekeeper. Every time I leave a man, I keep his house."

## Two-sample t-test, independent samples

The two-sample t-test for independent samples is given by

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The degrees of freedom (df) depend on whether or not  $\sigma_1 = \sigma_2$ .

## Equal or Unequal Variances?

The choice of df depends on whether the independent samples have the same, or different, variances. There is a test for this in Stata (**sdtest**) though power may be low in small to moderate sized samples (meaning you could incorrectly be led to use the equal variances test).

If the variances are equal, then we use a pooled estimate of  $s^2$ , and the degrees of freedom are given by  
 $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ .

If the variances are unequal, the degrees of freedom are difficult to derive, and the Satterthwaite approximation is used (software will do this for you, though the approximation is given in the text).

## Case Study: Hypothesis Testing Step 3

Carrying out the two sample t-test for independent samples with unequal variances, we get  $t = 4.4$  and  $df = 186.9$ , with a corresponding p-value  $< 0.001$ .



## Case Study: Hypothesis Testing Step 4

We conclude that there is a difference in importance placed on personality and looks by gender. Males place more emphasis on looks than women do.



## Case Study: CHNS

Suppose we wish to compare the BMI (body mass index) of smokers and non-smokers in the China Health and Nutrition Study. We specify

$$H_0 = \mu_{SMOKER} = \mu_{NONSMOKER}$$

$$H_A = \mu_{SMOKER} \neq \mu_{NONSMOKER}$$

First, we test whether the variances of BMI are the same in our sample of study participants from 2009.

```
. sdtest bmi, by(smokeyn)
```

Variance ratio test

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
1	6460	23.22975	.0457256	3.675154	23.14011	23.31938
2	2323	23.02818	.0691524	3.332972	22.89257	23.16378
combined	8783	23.17643	.0382931	3.588742	23.10137	23.2515

ratio = sd(1) / sd(2)

f = 1.2159

Ho: ratio = 1

degrees of freedom = 6459, 2322

Ha: ratio < 1

Ha: ratio != 1

Ha: ratio > 1

Pr(F < f) = 1.0000

2\*Pr(F > f) = 0.0000

Pr(F > f) = 0.0000

## Case Study: China Data

Now we carry out the t-test with unequal variances.

```
. ttest bmi, by(smokeyn) unequal
```

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Intervall]	
1	6460	23.22975	.0457256	3.675154	23.14011	23.31938
2	2323	23.02818	.0691524	3.332972	22.89257	23.16378
combined	8783	23.17643	.0382931	3.588742	23.10137	23.2515
diff		.2015711	.0829028		.0390407	.3641015

[illegible]

Ha: diff < 0	Ha: diff != 0	Ha: diff > 0
Pr(T < t) = <b>0.9925</b>	Pr( T  >  t ) = <b>0.0151</b>	Pr(T > t) = <b>0.0075</b>

## What is our conclusion?

# Paired samples

Samples are often paired for a variety of reasons

- ▶ Measurements are taken on a single subject at two distinct points in time (e.g., baseline and follow-up in NYC sugar drinks study)

Pairing can control for unwanted sources of variation that might otherwise influence the results of a comparison. Matching within subject (e.g., baseline and follow-up) is a powerful way to eliminate subject-specific factors.

## Paired samples

Samples are often paired for a variety of reasons

- ▶ Measurements are taken on a single subject at two distinct points in time (e.g., baseline and follow-up in NYC sugar drinks study)
- ▶ Subjects may be matched so that members of each pair are as much alike as possible with respect to important characteristics like age and gender (e.g., matched case-control study)

Pairing can control for unwanted sources of variation that might otherwise influence the results of a comparison. Matching within subject (e.g., baseline and follow-up) is a powerful way to eliminate subject-specific factors.

## Paired sample t-test

The paired sample t-test is pretty easy to carry out. All we do is to create a new outcome variable,  $d$ , that contains the differences in outcomes between members of a pair. Then we analyze the differences  $d$  using the usual one-sample t-test.

## Case Study: CHNS

Investigators wish to study whether the BMI of CHNS participants in 1991 and 2009 is the same. They have measured the BMI of each adult in this sample in both years, so the pairing is of repeated measures within subject.

```
. ttest bmi91=bmi09
```

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
bmi91	3679	21.09459	.050877	3.085931	20.99484	21.19434
bmi09	3679	23.25621	.0569194	3.452435	23.14461	23.36781
diff	3679	-2.16162	.0485223	2.94311	-2.256754	-2.066487

```
mean(diff) = mean(bmi91 - bmi09)          t = -44.5490
Ho: mean(diff) = 0                        degrees of freedom = 3678
```

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

We conclude that BMI is not the same in 1991 and 2009. In 2009, the BMI of the subjects is 2.16 (2.07, 2.26) units greater on average.

## In Practice: PTSD in Disaster Survivors

Let's look at part of the analysis from the North et al. reading for today. How does the relevant portion of the methods section look?

than 5. We compared continuous variables with dichotomous variables using  $t$  tests, substituting Satterthwaite analysis in cases of unequal variance. We set statistical significance level at  $\alpha = 0.05$ .



## In Practice: PTSD in Disaster Survivors

From the paper, it looks like postdisaster PTSD prevalent cases number 163 with 648 not having PTSD, with these groups compared in Table 2. We'll examine some rows of Table 2.

	Disaster-Related PTSD, % (No./Total No.)	PTSD, Mean (SD)	No PTSD, Mean (SD)	Significance
<b>Demographics</b>				
Age, y		42.5 (11.9)	46.7 (15.2)	$t = 3.78, P < .001$

Which version of the t-test was used for age? Try `ttesti 163 42.5 11.9 648 46.7 15.2, unequal` and `ttesti 163 42.5 11.9 648 46.7 15.2` to see. Was this the right one to use?

## In Practice: PTSD in Disaster Survivors

Next look at injury count. Surely this is not a normally distributed variable. Was it OK to do a t-test here? Why or why not?

	Disaster-Related PTSD, % (No./Total No.)	PTSD, Mean (SD)	No PTSD, Mean (SD)	Significance
<b>Demographics</b>				
Age, y		42.5 (11.9)	46.7 (15.2)	$t = 3.78, P < .001$
No. of injuries		5.1 (4.2)	3.1 (2.4)	$t = 4.08, P < .001$

## Interactive Case Study: Impaired Driving

The Department of Motor Vehicles wishes to compare impairment of drivers while texting to impairment after drinking two beers. Suppose we use a design with two independent groups. How do we do a sample size calculation?

Stata: `sampsi mean1 mean2, power( ) r( ) sd1( ) sd2( )`

Value in `r( )` should be  $\frac{n_2}{n_1}$ . Use `r(1)` if you want groups of equal size. `r(2)` calls for twice as many subjects in group 2 as in group 1

## Interactive Case Study: Impaired Driving

A recent paper in *Traffic Injury Prevention* showed that impairment was similar among drivers who were texting and drivers who were eating. Suppose you want a 4 group design: drinking, texting, eating, and control. Next time we'll discuss how to analyze responses across more than 2 groups with ANOVA!

## Exercise: Serum DHEA-S and Meditation

Glaser et al. presented the following results from a study of transcendental meditation (TM) and serum DHEA-S (a hormone secreted by the adrenal gland).

**Table 1:** Serum DHEA-S Concentrations ( $\pm$  SEM) in Women for Comparison and TM Groups

Age Group	Comparison Group			TM Group	
	N	DHEA-S ( $\mu\text{g}/\text{dl}$ )	N	DHEA-S ( $\mu\text{g}/\text{dl}$ )	% Elevation in TM Group
45-49	51	$88 \pm 12$	30	$117 \pm 11$	34

Define  $H_0$  and  $H_A$  for a hypothesis test that there is no difference in serum DHEA-S levels between the two groups. Carry out the test and interpret the results.

## Reading for Next Time

- ▶ Pagano and Gauvreau, Chapter 12
- ▶ Gelman, “Of Beauty, Sex, and Power”