

## L12: Sections 7.1 and 7.2

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

# Hypothesis testing

- Often scientists wish to test a hypothesis, or statement of fact.
- **Example 7.2.1** scientists wish to test the hypothesis that norepinephrine (NE) levels are different in rats exposed to toluene (glue) and those that aren't.
- The scientist designs a controlled experiment in which  $n_1 = 6$  rats are exposed to toluene and  $n_2 = 5$  rats are not. NE levels are measured in the rats brains.
- The scientists wish to show that the population NE levels are different among rats exposed and non-exposed to toluene.
- This is encapsulated as  $H_A : \mu_1 \neq \mu_2$ , the *mean* NE levels differ across exposed and non-exposed.

# Hypothesis testing

- A hypothesis test is a proof by contradiction.
- We assume the null  $H_0 : \mu_1 = \mu_2$  then the data shows us something that is absurd, casting doubt on what we assumed, namely  $H_0 : \mu_1 = \mu_2$ .
- We thus have to conclude the opposite,  $H_A : \mu_1 \neq \mu_2$ .
- The **null hypothesis** is *what we are trying to disprove*,  $H_0 : \mu_1 = \mu_2$ .
- The **alternative hypothesis** is *what we're trying to show is true*,  $H_A : \mu_1 \neq \mu_2$ .

# Hypothesis testing

- Recall from Chapter 6, that if data are normal in both populations then

$$t_s = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{SE_{\bar{Y}_1 - \bar{Y}_2}}$$

has a  $t$  distribution with  $df$  given by the Satterthwaite-Welch formula.

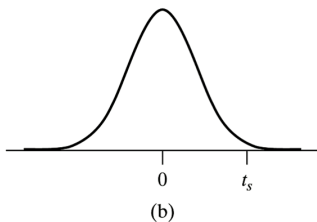
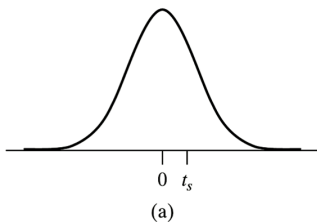
- In the hypothesis test, we are *assuming*  $\mu_1 - \mu_2 = 0$ , so

$$t_s = \frac{\bar{Y}_1 - \bar{Y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}}$$

has a  $t$  distribution, which is centered at zero.

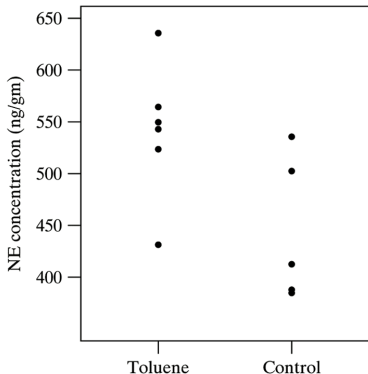
- If  $t_s$  is really far away from zero in either direction, we have evidence that  $H_0 : \mu_1 = \mu_2$  is *not true*.
- $|t_s|$  measures how far apart  $\bar{Y}_1$  and  $\bar{Y}_2$  are, i.e. how many  $SE_{\bar{Y}_1 - \bar{Y}_2}$ 's apart.

## $t$ test schematic



(a) Data compatible with  $H_0$  (so no evidence toward  $H_A$ ), (b) data not compatible with  $H_0$  (in favor of  $H_A$ ).

## Example 7.2.1 Parallel dotplots NE concentration



**Figure 7.2.1** Parallel dotplots of NE concentration

Normality okay? Does there appear to be a mean difference?

## Example 7.2.2 NE concentration (ng/gm)

	Toluene (Group 1)	Control (Group 2)
	543	535
	523	385
	431	502
	635	412
	564	387
	549	
$n$	6	5
$\bar{y}$	540.8	444.2
$s$	66.1	69.6
SE	27	31

$$t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}} = \frac{540.8 - 444.2}{\sqrt{\frac{66.1^2}{6} + \frac{69.6^2}{5}}} = 2.34.$$

$\bar{y}_1$  is 2.34 SE's away from  $\bar{y}_2$ . This is big, but how big?

# P-values

- The  **$P$ -value** for a hypothesis test is the probability of the test statistic being at least as extreme as the observed test statistic, *assuming  $H_0$  is true*.
- P-value answers the question “how big is big?” for  $|t_s|$ .



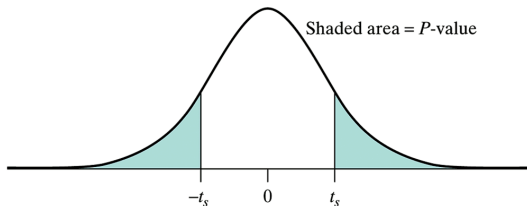
## P-value for two-sample problem

- For the two-sample problem, P-value is the probability of seeing *sample means*  $\bar{Y}_1$  and  $\bar{Y}_2$  *even further apart than what we saw, if*  $H_0 : \mu_1 = \mu_2$  *is true.*
- This is a standard probability calculation due to W.S. Gosset

$$\begin{aligned} \text{P-value} &= \Pr\{|\bar{Y}_1 - \bar{Y}_2| > |\bar{y}_1 - \bar{y}_2| | H_0 \text{ true}\} \\ &= \Pr\left\{\left|\frac{\bar{Y}_1 - \bar{Y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}}\right| > \left|\frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}}\right| \mid H_0 \text{ true}\right\} \\ &= \Pr\{|T_{df}| > |t_s|\} \end{aligned}$$

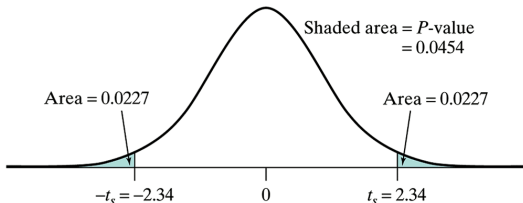
where  $T_{df}$  is a  $t$  random variable with  $df$  given by the Satterthwaite-Welch formula.

## Two-tailed $p$ -value for the $t$ test



- The P-value is computed in R using `t.test(data1,data2)`.
- It is the area in the tails of the  $t$  distribution with Satterthwaite-Welch  $df$ .

## Example 7.2.3 Two-tailed $P$ -value toluene data



```
> toluene=c(543,523,431,635,564,549)
> control=c(535,385,502,412,387)
> t.test(toluene,control)
```

Welch Two Sample t-test

```
data: toluene and control
t = 2.3447, df = 8.451, p-value = 0.04543
alternative hypothesis: true difference in means is not equal to 0
```

# What do we decide based on the P-value?

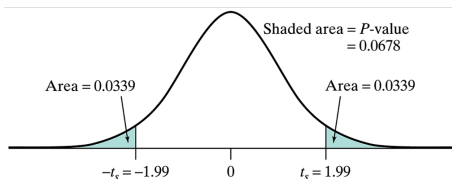
- $|t_s|$  is “big” when the P-value is “small.” But how small is small?
- We compare the P-value to a cutoff value denoted  $\alpha$ .
- $\alpha$  is called the **significance level** of the hypothesis test; it is almost always  $\alpha = 0.05$ .
- If P-value  $< \alpha$  then we **reject**  $H_0$  at significance level  $\alpha$ .
- If P-value  $> \alpha$  then we **accept**  $H_0$  at significance level  $\alpha$ .
- Some books demand that you say “do not reject  $H_0$ ” instead of “accept  $H_0$ .”
- $\alpha$  has an interpretation that we’ll talk about next time.

## Example 7.2.4 NE concentration data

	Toluene	Control
$n$	6	5
$\bar{y}$	540.8	444.2
$s$	66.1	69.6

We reject  $H_0 : \mu_1 = \mu_2$  at the 5% significance level because  $P\text{-value} = 0.0454 < 0.05 = \alpha$ . There is statistically significant evidence that the mean norepinephrine levels are different in toluene-exposed vs. non-exposed rats.

## Example 7.2.5 Two-week height of control &amp; ancy plants



```
> control=c(10.0,13.2,19.8,19.3,21.2,13.9,20.3,9.6)
> ancy=c(13.2,19.5,11.0,5.8,12.8,7.1,7.7)
> t.test(control,ancy)
```

```
t = 1.9939, df = 12.783, p-value = 0.06795
alternative hypothesis: true difference in means is not equal to 0
```

As  $P\text{-value} = 0.068 > 0.05 = \alpha$ , we accept  $H_0 : \mu_1 = \mu_2$  at the 5% level. There is no evidence that the mean heights are different in control vs. ancy populations.

# Assumptions behind two-sample $t$ test

- If the sample sizes are large enough ( $n_1 > 30$  and  $n_2 > 30$ , say) the  $t$ -test works okay because of the central limit theorem.
- If the sample sizes are small, data from each population *needs to be normal* for the procedure to work okay. If not, we can't trust the  $t$ -test P-value.
- When sample sizes are small and data are not normal, there is an alternative method to compute the P-value that *doesn't assume anything* about the population shapes.
- This approach is called a randomized test or **permutation test**, and uses *resampling methods*.
- Resampling methods are a powerful approach to statistics and include permutation tests and bootstrapping.

## Example 7.1.1

- An exercise science researcher studied the trunk flexion flexibility (cm) of  $n_1 = 4$  women in an aerobics class, and  $n_2 = 3$  women who were dancers.
- Among aerobics class participants we have 38, 45, 58, and 64 cm.
- Among dancers we have 48, 59, and 61 cm.
- Is there a difference in  $\mu_1 =$  population mean stretching of aerobics vs.  $\mu_2 =$  population mean stretching of dancers?
- If there *truly is no difference*, then all 7 observations *came from the same population distribution*.
- If there *truly is no difference*, then all arrangements of  $n_1 = 3$  observations and  $n_2 = 4$  observations are *equally likely*.



Permutation test of  $H_0 : \mu_1 = \mu_2$ 

- Compute the observed test statistic  $d_s = \bar{y}_1 - \bar{y}_2$ .
- Consider all possible arrangements of  $n_1 = 4$  and  $n_2 = 3$  and compute the mean differences  $d$  from these. The histogram from this is called the **permutation density**. This is the distribution of the test statistic assuming  $H_0 : \mu_1 = \mu_2$  is true.
- The P-value for  $H_0 : \mu_1 = \mu_2$  vs.  $H_A : \mu_1 \neq \mu_2$  is the proportion of  $|d|$ 's bigger than  $|d_s|$ .

## Permutation samples

**Table 7.1.2**

Sample 1 ("aerobics")	Sample 2 ("dance")	Mean of sample 1	Mean of sample 2	Difference in means
38 45 58 64	48 59 61	51.25	56.00	<b>-4.75</b>
38 45 58 48	64 59 61	47.25	61.33	<b>-14.08</b>
38 45 58 59	64 48 61	50.00	57.67	<b>-7.67</b>
38 45 58 61	64 48 59	50.50	57.00	<b>-6.50</b>
38 45 64 48	58 59 61	48.75	59.33	<b>-10.58</b>
38 45 64 59	58 48 61	51.50	55.67	<b>-4.17</b>
38 45 64 61	58 48 59	52.00	55.00	<b>-3.00</b>
38 45 48 59	58 64 61	47.50	61.00	<b>-13.50</b>
38 45 48 61	58 64 59	48.00	60.33	<b>-12.33</b>
38 45 59 61	58 64 48	50.75	56.67	<b>-5.92</b>
38 58 64 48	45 59 61	52.00	55.00	-3.00
38 58 64 59	45 48 61	54.75	51.33	3.42
38 58 64 61	45 48 59	55.25	50.67	4.58
38 58 48 59	45 64 61	50.75	56.67	<b>-5.92</b>
38 58 48 61	45 64 59	51.25	56.00	<b>-4.75</b>

*(Continues on next page)*

10  $|d|$ 's are bigger than  $|d_s| = 4.75$  out of the first 15 possible combinations.

## Permutation samples cont'd

**Table 7.1.2 (Continued)**

Sample 1 ("aerobics")	Sample 2 ("dance")	Mean of sample 1	Mean of sample 2	Difference in means
38 58 59 61	45 64 48	54.00	52.33	1.67
38 64 48 59	45 58 61	52.25	54.67	-2.42
38 64 48 61	45 58 59	52.75	54.00	-1.25
38 64 59 61	45 58 48	55.50	50.33	<b>5.17</b>
38 48 59 61	45 58 64	51.50	55.67	-4.17
45 58 64 48	38 59 61	53.75	52.67	1.08
45 58 64 59	38 48 61	56.50	49.00	<b>7.50</b>
45 58 64 61	38 48 59	57.00	48.33	<b>8.67</b>
45 58 48 59	38 64 61	52.50	54.33	-1.83
45 58 48 61	38 64 59	53.00	53.67	-0.67
45 58 59 61	38 64 48	55.75	50.00	<b>5.75</b>
45 64 48 59	38 58 61	54.00	52.33	1.67
45 64 48 61	38 58 59	54.50	51.67	2.83
45 64 59 61	38 58 48	57.25	48.00	<b>9.25</b>
45 48 59 61	38 58 64	53.25	53.33	-0.08
58 64 48 59	38 45 61	57.25	48.00	<b>9.25</b>
58 64 48 61	38 45 59	57.75	47.33	<b>10.42</b>
58 64 59 61	38 45 48	60.50	43.67	<b>16.83</b>
58 48 59 61	38 45 64	56.50	49.00	<b>7.50</b>
64 48 59 61	38 45 58	58.00	47.00	<b>11.00</b>

10  $|d|$ 's are bigger than  $|d_s| = 4.75$  out of the last 20 possible combinations.

## Final P-value

There are 20  $|d|$ 's bigger than  $|d_5|$  out of 35 possible combinations.  
The P-value is  $\frac{20}{35} = 0.57$ .

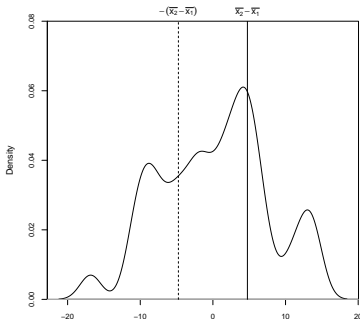
We accept  $H_0 : \mu_1 = \mu_2$  at the 5% level because P-value  $= 0.54 > 0.05 = \alpha$ . There is no evidence that the mean trunk flexion is different between dancers and aerobics participants.

## DAAG package in R

- Many people have written specialized R code to carry out non-standard analyses.
- The DAAG (Data Analysis And Graphics data and functions) package has a function to do a two-sample permutation test as described in the book.
- The function works the same as `t.test`, e.g. `twot.permutation(data1,data2)`.
- You will get a P-value and an estimate of the permutation density (based on an approximation).
- In R click Packages, then Install package(s) . . . , then pick a mirror (a place to download from – any will work), then scroll down until you find DAAG and install it.
- After installing it, you need to load it. Under Packages pick Load package . . . , then pick out DAAG.

## Trunk flexion in R

```
> aerobics=c(38,45,58,64)
> dance=c(48,59,61)
> twot.permutation(aerobics,dance)
[1] 0.572
```



The P-value is the area outside of  $-d_s$  and  $d_s$ .

# Comparing permutation test to t-test

```
> control=c(10.0,13.2,19.8,19.3,21.2,13.9,20.3,9.6)
> ancy=c(13.2,19.5,11.0,5.8,12.8,7.1,7.7)
> twot.permutation(control,ancy)
[1] 0.08
> t.test(control,ancy)
t = 1.9939, df = 12.783, p-value = 0.06795

> aerobics=c(38,45,58,64)
> dance=c(48,59,61)
> twot.permutation(aerobics,dance)
[1] 0.572
> t.test(aerobics,dance)
t = -0.6615, df = 4.86, p-value = 0.5384
```