STAT 234 Lecture 21
Analysis of Two-Sample Data
Section 10.1-10.2

Yibi Huang
Department of Statistics
University of Chicago

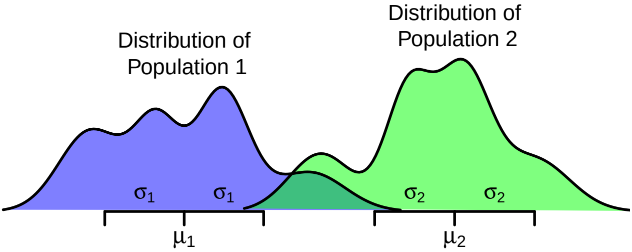
Two Sample Problems (1)

- E.g., is the air more polluted in Chicago or in LA?
- E.g., Do smokers or nonsmokers suffer more from depression?
- E.g., Does the mean response for the treatment group differ from that for the control group?

Two Sample Problems (2)

Population 1

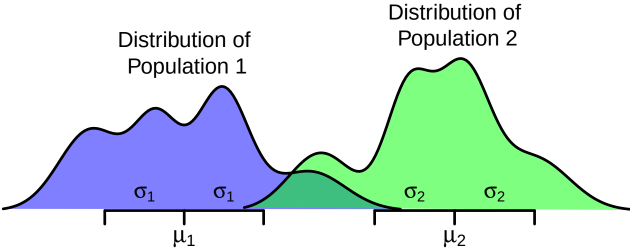
Population 2



Two Sample Problems (2)

Population 1

Population 2

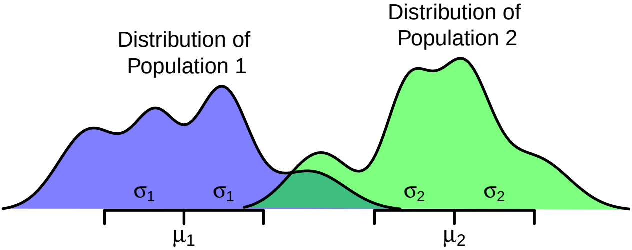


Population distributions may be normal or not normal or of the different shape.

Two Sample Problems (2)

Population 1

Population 2

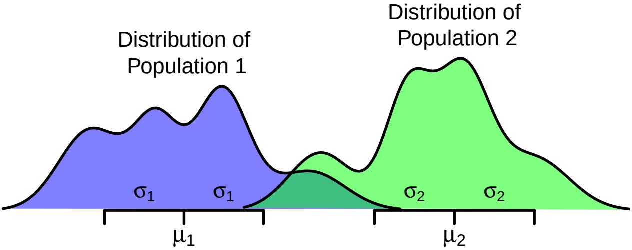


Population SDs σ_1 and σ_2 may not be equal.

Two Sample Problems (2)

Population 1

Population 2



Goal: difference in population means $\mu_1 - \mu_2$.

Two Sample Data

$$\text{Population 1} \quad \longrightarrow \text{random sample } X_1, X_2, \ldots, X_m$$
$$\text{Population 2} \quad \longrightarrow \text{random sample } Y_1, Y_2, \ldots\ldots\ldots, Y_n$$

- Observations in one group are **independent** of those in the other group
- the two samples can be of different sizes m and n

Two Sample Problems (3)

A natural estimate of $\mu_1 - \mu_2$ is the difference of the two sample means $\overline{X} - \overline{Y}$.

How close is $\overline{X} - \overline{Y}$ to $\mu_1 - \mu_2$?

Two Sample Problems (4)

Recall

$$\mathrm{E}(\overline{X}) = \mu_1, \quad \mathrm{E}(\overline{Y}) = \mu_2, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma_1^2}{m}, \quad \mathrm{Var}(\overline{Y}) = \frac{\sigma_2^2}{n}.$$

Observe $\overline{X} - \overline{Y}$ is an **unbiased estimate** of $\mu_1 - \mu_2$ because

$$\mathsf{E}(\overline{X} - \overline{Y}) = \mathsf{E}(\overline{X}) - \mathsf{E}(\overline{Y}) = \mu_1 - \mu_2.$$

Furthermore, since the two samples are *independent*, $\overline{X}$ and $\overline{Y}$ are independent, we have

$$\mathrm{Var}(\overline{X} - \overline{Y}) = \mathrm{Var}(\overline{X}) - 2\underbrace{\mathrm{Cov}(\overline{X}, \overline{Y})}_{=0 \text{ by indep.}} + \mathrm{Var}(\overline{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

Thus the **standard error** of $\overline{X} - \overline{Y}$ is

$$\mathsf{SD}(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Two-Sample z-Test w/ ***Known*** σ_1 & σ_2

For testing H_0: $\mu_1 - \mu_2 = \Delta_0$, the z-statistic is

$$z\text{-stat} = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

We reject $\mathrm{H}_0 : \mu = \mu_0$ at the significance level α if

- z-stat $> z_\alpha$ for H_A: $\mu_1 - \mu_2 > \Delta_0$
- z-stat $< -z_\alpha$ for H_A: $\mu_1 - \mu_2 < \Delta_0$
- $|z\text{-stat}| > z_{\alpha/2}$ for H_A: $\mu_1 - \mu_2 \neq \Delta_0$

Two-Sample z-CI w/ *Known* σ_1 & σ_2

A $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ is given by

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Two-Sample t-Statistic w/ Unknown σ_1 & σ_2

Of course, σ_1^2 and σ_2^2 are often unknown, We thus replace them with the sample variances s_1^2 and s_2^2.

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} \quad \text{where} \quad \begin{aligned} s_1^2 &= \frac{\sum_{i=1}^{m}(X_i - \overline{X})^2}{m-1} \\ s_2^2 &= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1} \end{aligned}$$

- Unfortunately, the two-sample t-statistic does NOT have a t-distribution
- Fortunately, it can be approximated by a t-distribution with a certain degrees of freedom.

See the next slide for the approximation

Approximate Distribution of the Two-Sample t-Statistic

The two-sample t-statistic has an **approximate t_ν distribution**.

For the degrees of freedom ν we have two formulas:

- software formula:

$$\nu = \frac{(w_1 + w_2)^2}{w_1^2/(m-1) + w_2^2/(n-1)}, \quad \text{where} \quad \begin{aligned} w_1 &= s_1^2/m, \\ w_2 &= s_2^2/n. \end{aligned}$$

- simple formula: $\boxed{\nu = \min(m-1, n-1)}$

Comparison of the two formulas:

- The software formula is more accurate. It gives larger d.f. and yields shorter CIs and smaller P-value
- The simple formula is conservative. I.e., it yields wider CIs and larger P-values than the actual P-value
- In the exam, it is fine **just using the simple formula**.

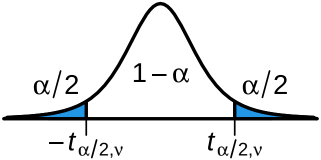
Confidence Intervals for $\mu_1 - \mu_2$

A $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ is given by

$$(\overline{X} - \overline{Y}) \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where $t_{\alpha/2,\nu}$ is the value of the t distribution with ν degrees of freedom such that

density curve of t_ν



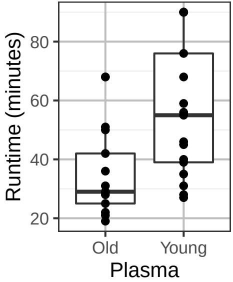
which can be found in R using the `qt()` command.

```
qt(alpha/2, df, lower.tail=F)
```

Example: Young Blood Helps Old Brains?

Several studies[1] on mice indicate that young blood help old brains. Old mice were randomly assigned to receive blood plasma either from a young mouse or another old mouse, and then tested on treadmill. The maximum treadmill runtime in minutes for 17 mice receiving young blood and 13 mice receiving old blood are

Blood	Runtime (minutes)	Mean	SD
Young	27 28 31 35 39 40 45 46 55 56 59 68 76 90 90 90 90	56.76	23.22
Old	19 21 22 25 28 29 29 31 36 42 50 51 68	34.69	14.37



[1] Sanders, L., "Young blood proven good for old brain," *Science News*, 185(11), May 31, 2014

Example: CI for the Young Blood Effect

Using the simple df $= \min(17-1, 13-1) = 12$, the critical value $t_{0.05/2,12} \approx 2.179$ for 95% CI can be found in R as follows

```r
qt(0.05/2, df=12, lower.tail=F)
## [1] 2.178813
```

α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
ν 12	1.356	1.782	2.179	2.681	3.055	3.930	4.318

The 95% CI for $\mu_Y - \mu_O$ (Young – Old) is hence

$$\overline{X}_Y - \overline{X}_O \pm t_{0.05/2,12}\sqrt{\frac{s_Y^2}{m} + \frac{s_O^2}{n}} \approx 56.76 - 34.69 \pm 2.179\sqrt{\frac{23.22^2}{17} + \frac{14.37^2}{13}}$$

$$\approx 22.07 \pm 15.03 = (7.04, 37.10)$$

With 95% confidence, the maximum treadmill runtime of old mice receiving plasma from a young mouse is 7.04 to 37.10 minutes longer on average than those who received plasma from a old mouse.

Example: CI for the Young Blood Effect

If we use the software formula for the df,

$$w_1 = \frac{s_Y^2}{m} \approx \frac{23.22^2}{17} \approx 31.71, \quad w_2 = \frac{s_O^2}{n} \approx \frac{14.37^2}{13} \approx 15.88$$

$$df = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{m-1} + \frac{w_2^2}{n-1}} \approx \frac{(31.71 + 15.88)^2}{\frac{31.71^2}{17-1} + \frac{15.88^2}{13-1}} \approx 27.007.$$

The critical value for 95% CI is $t_{0.05/2,27} \approx 2.052$.

```
qt(0.05/2, df=27.007, lower.tail=F)
## [1] 2.051806
```

α		0.1	0.05	0.025	0.01	0.005	0.001	0.0005
ν	27	1.314	1.703	2.052	2.473	2.771	3.421	3.690

The 95% CI for $\mu_Y - \mu_O$ becomes

$$\overline{X}_Y - \overline{X}_O \pm t_{0.05/2,27.007}\sqrt{\frac{s_Y^2}{m} + \frac{s_O^2}{n}} \approx 56.76 - 34.69 \pm 2.052\sqrt{\frac{23.22^2}{17} + \frac{14.37^2}{13}}$$

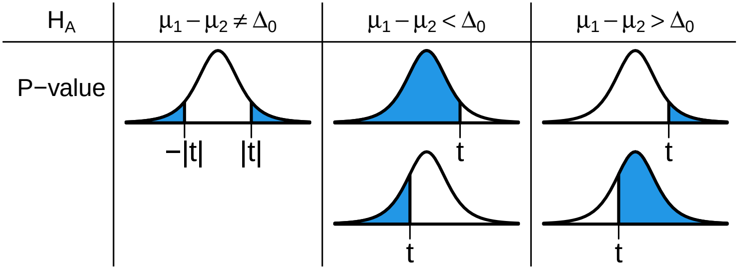
$$\approx 22.07 \pm 14.16 = (7.91, 36.23)$$

Hypothesis Tests for $\mu_1 - \mu_2$

To test H_0: $\mu_1 - \mu_2 = \Delta_0$, the two-sample t-statistic is

$$t = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{\sqrt{s_1^2/m + s_2^2/n}} \sim \text{approx. } t_\nu$$

where the df is $\nu = \min(m-1, n-1)$, or the one given by the software formula, and the P-value is computed as follows depending on H_A.



The bell curve above is the t-curve with ν degrees of freedom.

Example: Test for the Young Blood Effect

To test H$_0$: $\mu_Y - \mu_O = 0$ v.s. H$_a$: $\mu_Y - \mu_O \neq 0$, the t-statistic is

$$t = \frac{\overline{X}_Y - \overline{X}_O}{\sqrt{\frac{s_Y^2}{m} + \frac{s_O^2}{n}}} = \frac{56.76 - 34.69}{\sqrt{\frac{23.22^2}{17} + \frac{14.37^2}{13}}} = \frac{22.07}{6.899} \approx 3.199$$

df $= 13 - 1 = 12$ (simple) or 27.007 (software). The two-sided P-value can be found in R to be ≈ 0.0076 or 0.0035

```
2*pt(3.199, df=12, lower.tail=F)
## [1] 0.007646717
2*pt(3.199, df=27.007, lower.tail=F)
## [1] 0.003507634
```

	α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
ν	12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
	27	1.314	1.703	2.052	2.473	2.771	3.421	3.690

The difference is significant at 1% level.

The maximum treadmill runtime of old mice receiving young blood is significantly longer on average than those receiving old blood.

Analysis of Two Sample Data Assuming Equal Population SD's

What if $\sigma_1 = \sigma_2$?

So far we have assumed that $\sigma_1 \neq \sigma_2$. What if we have reasons to believe $\sigma_1 = \sigma_2 = \sigma$ albeit σ is unknown?

When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, both s_1^2 and s_2^2 are unbiased estimates of σ^2. We can combine s_1^2 and s_2^2 to get a better estimate for σ^2, the so-called **pooled sample variances**

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

Observe that s_p^2 is a weighted average of s_1^2 and s_2^2, and it gives more weights to the sample with larger size.

Moreover, as $s^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$, we can see that

$$s_p^2 = \frac{\sum_i (X_i - \overline{X})^2 + \sum_i (Y_i - \overline{Y})^2}{m+n-2}$$

is simply an "average" of the *squared deviations from the corresponding means*, though the divider is $m+n-2$ not $m+n$.

The Pooled Two-Sample t-Statistic Asumming Equal SDs

The two-sample t-statistic then becomes

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{m} + \frac{s_p^2}{n}}} = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

which is specifically called the **pooled two-sample t-statistic**.

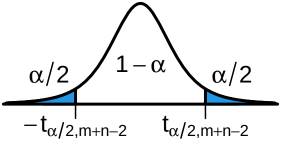
- It has an **exact** t-distribution with $m + n - 2$ **degrees of freedom** when the two populations are normal.
- It is approximately $t_{(m+n-2)}$ for non-normal population w/ equal SDs as long as the sample size m, n is not too small.
- The degrees of freedom, $m + n - 2$ is greater than the df of two-sample t -statistic when $\sigma_1 \neq \sigma_2$ (both software formula or the simple formula)

Two Sample Problems w/ Equal but Unknown SD's

A $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ is

$$(\overline{X} - \overline{Y}) \pm t_{\alpha/2, m+n-2}\, s_p \sqrt{\frac{1}{m} + \frac{1}{n}}$$

where $t_{\alpha/2,m+n-2}$ is the value of the t distribution with df $= m+n-2$ such that



which can be found in R using the `qt()` command.

```
qt(alpha/2, df=m+n-2, lower.tail=F)
```

To test $H_0 : \mu_1 - \mu_2 = \Delta_0$, we use the pooled 2-sample t-statistic

$$t = \frac{\overline{X} - \overline{Y} - \Delta_0}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2} \qquad \text{under } H_0$$

Young Blood Example Assuming Equal SD's — 95% CI

Assuming $\sigma_1 = \sigma_2$, the pooled SD is

$$s_p = \sqrt{\frac{(17-1)23.22^2 + (13-1)14.37^2}{17+13-2}} \approx 19.915$$

with df $= m + n - 2 = 17 + 13 - 2 = 28$. The critical value $t_{0.05/2,28} \approx 2.048$ for 95% CI is found in R below.

```
qt(0.05/2, df=28, lower.tail=F)
## [1] 2.048407
```

So the 95% CI for $\mu_Y - \mu_O$ (Young – Old) is

$$\begin{aligned}\overline{X}_Y - \overline{X}_O \pm t_{0.05/2,28}\, s_p \sqrt{\frac{1}{m} + \frac{1}{n}} &= 56.76 - 34.69 \pm 2.048 \times 19.915 \times \sqrt{\frac{1}{17} + \frac{1}{13}} \\ &\approx 22.07 \pm 15.03 = (7.04, 37.10)\end{aligned}$$

Observe the CI is shorter when assuming equal SDs for the greater df.

The greater the df, the smaller the critical value $t_{\alpha/2,df}$.

Young Blood Example: Hyp Test Assuming Equal SD's

For testing $H_0 : \mu_Y - \mu_O = 0$ v.s. $H_a : \mu_Y - \mu_O \neq 0$, assuming $\sigma_1 = \sigma_2$ the pooled t-statistic is

$$t = \frac{\overline{X}_Y - \overline{X}_O}{s_p \sqrt{1/m + 1/n}} = \frac{56.76 - 34.69}{19.915 \sqrt{1/17 + 1/13}} = \frac{22.07}{7.337} \approx 3.008$$

The df is $m + n - 2 = 17 + 13 - 2 = 28$.

The 2-sided P-value can be found in R to be ≈ 0.0055 or using table to be between 0.01 and 0.002.

```
2*pt(3.008, df=28, lower.tail=F)
## [1] 0.00550726
```

α	0.1	0.05	0.025	0.01	*0.005*	*0.001*	0.0005
ν 28	1.313	1.701	2.048	2.467	*2.763*	*3.408*	3.674

The pooled t-test gives smaller P-value and the result appears more significant.

Two-Sample Tests/CIs in R

```
Young = c(27,28,31,35,39,40,45,46,55,56,59,68,76,90,90,90,90)
Old = c(19,21,22,25,28,29,29,31,36,42,50,51,68)
```

By default, the R command `t.test()` does NOT assume $\sigma_1 = \sigma_2$.

```
t.test(Young, Old, conf.level=0.95)
##
##  Welch Two Sample t-test
##
## data:  Young and Old
## t = 3.1997, df = 27.006, p-value = 0.003502
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.918414 36.226383
## sample estimates:
## mean of x mean of y
##  56.76471  34.69231
```

Note R uses the software formula to compute the df $= 27.006$.

Two-Sample Tests/CIs in R

One can force σ_1, σ_2 to be equal by adding `var.equal = T`.

```
t.test(Young, Old, conf.level = 0.95, var.equal = T)
##
##  Two Sample t-test
##
## data:  Young and Old
## t = 3.0086, df = 28, p-value = 0.005499
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.044474 37.100323
## sample estimates:
## mean of x mean of y
##  56.76471  34.69231
```

Which Two-Sample Tests/CIs to Use?

We have introduced two different two-sample tests/CIs:

- the one assuming $\sigma_1 = \sigma_2$ used the **pooled SD**.
- the one w/o assuming $\sigma_1 = \sigma_2$ is called **Welch's method**.

Though in many cases, the two methods agree in the conclusion, but they can provide different answers when:

- the sample SDs are very different, and
- the sizes of the groups are also very different

So which method should I use?

- When σ_1 and σ_2 are indeed equal, the method based on pooled SD is more powerful
- However, it is usually hard to check whether $\sigma_1 = \sigma_2$. So it's safer to use Welch's method.

Robustness of Two-Sample t-Procedures (1)

Even when the populations are not normal, the two-sample statistics

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

can be well-approximated by t-distributions, as long as *the sample sizes are not too small*.

This is the so-called **robustness** of the two-sample t-procedures.

Robustness of Two-Sample t-Procedures (2)

- The t-approximation is generally good if $m + n$ is not too small (both ≥ 15), the data are not strongly skewed, and there are no outliers.
  - Check histograms or side-by-side boxplots of the data
- With $m + n$ sufficiently large (say both ≥ 30), the approximation is good even when the data are clearly skewed.
- Given a fixed sum of the sample sizes $m + n$ the t-approximation works the best when the sample sizes are equal $m = n$
  - In planning a two-sample study, choose equal sample sizes if you can