Now, Two-Sample Tests (No need to memorize formulas in detail)

Two sample t-Test Example

You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	NYSE	**NASDAQ**
Number	**21**	**25**
Sample mean	**3.27**	**2.53**
Sample std dev	**1.30**	**1.16**

Assuming both populations are approximately normal with equal variances, is there a difference in mean yield ($\alpha = 0.05$)?

F Test: An Example

You are a financial analyst for a brokerage firm. You want to compare dividend yields between stocks listed on the NYSE & NASDAQ. You collect the following data:

	NYSE	NASDAQ
Number	**21**	**25**
Mean	**3.27**	**2.53**
Std dev	**1.30**	**1.16**

Is there a difference in the variances between the NYSE & NASDAQ at the $\alpha = 0.05$ level?

Paired Difference Test: *Pairwise t-test*

- Assume you send your salespeople to a "customer service" training workshop. Has the training made a difference in the number of complaints? You collect the following data:

Salesperson	Number of Complaints: Before (1)	After (2)	(2) - (1) Difference, D_i
C.B.	6	4	- 2
T.F.	20	6	-14
M.H.	3	2	- 1
R.K.	0	0	0
M.O.	4	0	- 4
			-21

Two-Sample Tests

Two-Sample Tests

- Population Means, <u>Independent Samples</u>
- Population Means, <u>Related Samples</u>
- Population Proportions (Skip this)
- <u>Population Variances</u>

Examples:

- Group 1 vs. Group 2
- Same group before vs. after treatment
- Proportion 1 vs. Proportion 2
- Variance 1 vs. Variance 2

Related Populations The Paired Difference Test

Related samples

Tests Means of 2 Related Populations

- Paired or matched samples
- Repeated measures (before/after)
- Use difference between paired values:

$$D_i = X_{1i} - X_{2i}$$

- Assumptions:
  - Both Populations Are Normally Distributed
  - Or, if not Normal, need large sample size
- Called Pairwise t-test

Related Populations
The Paired Difference Test

(continued)

Related samples

The i^{th} paired difference is D_i , where

$$D_i = X_{1i} - X_{2i}$$

The point estimate for the paired difference population mean μ_D is $\bar{D}$:

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

The sample standard deviation is S_D

$$S_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{n-1}}$$

n is the number of pairs in the paired sample

The Paired Difference Test: Finding t_{STAT}

Paired samples

- The test statistic for μ_D is:

$$t_{STAT} = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}}$$

- Where t_{STAT} has $n - 1$ d.f.

The Paired Difference Confidence Interval

Paired samples

The confidence interval for μ_D is

$$\overline{D} \pm t_{\alpha/2} \frac{S_D}{\sqrt{n}}$$

where $S_D = \sqrt{\frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}}$

Paired Difference Test: Example

- Assume you send your salespeople to a "customer service" training workshop. Has the training made a difference in the number of complaints? You collect the following data:

Salesperson	Number of Complaints: Before (1)	After (2)	(2) - (1) Difference, D_i
C.B.	6	4	- 2
T.F.	20	6	-14
M.H.	3	2	- 1
R.K.	0	0	0
M.O.	4	0	- 4
			-21

$$\bar{D} = \frac{\sum D_i}{n}$$

$$= -4.2$$

$$S_D = \sqrt{\frac{\sum (D_i - \bar{D})^2}{n-1}}$$

$$= 5.67$$

P-value? Let us find it.

- Has the training made a difference in the number of complaints (at the 0.01 level)?

Paired Difference Test: Solution

- Has the training made a difference in the number of complaints (at the 0.01 level)?

H_0: $\mu_D = 0$
H_1: $\mu_D \neq 0$

$\alpha = .01$ $\quad \bar{D} = -4.2$

d.f. = n - 1 = 4

Decision: Do not reject H_0

Test Statistic:

$$t_{STAT} = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} = \frac{-4.2 - 0}{5.67/\sqrt{5}} = \boxed{-1.66}$$

Conclusion: There is not a significant change in the number of complaints.

JMP Demo

- The data (Therm.jmp) contains temperature measurements on 20 people. Temperature is measured using two types of thermometers: oral and tympanic (ear).
- You want to determine whether the two types of thermometers produce equal temperature readings.
- The small p-value (Prob > |t|) indicates that this difference is statistically significant, and not due to chance.

Difference Between Two Means

Population means, independent samples *

- σ_1 and σ_2 unknown, assumed equal
- σ_1 and σ_2 unknown, not assumed equal

Goal: Test hypothesis or form a confidence interval for the difference between two population means, $\mu_1 - \mu_2$

The point estimate for the difference is

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

Difference Between Two Means: Independent Samples

Population means, independent samples *

- Different data sources
  - Unrelated
  - Independent
    - Sample selected from one population has no effect on the sample selected from the other population

σ_1 and σ_2 unknown, assumed equal → Use S_p to estimate unknown σ. Use a **Pooled-Variance t test.**

σ_1 and σ_2 unknown, not assumed equal → Use S_1 and S_2 to estimate unknown σ_1 and σ_2. Use a **Separate-variance t test**

Hypothesis tests for μ_1 - μ_2 with σ_1 and σ_2 unknown and assumed equal

Population means, independent samples

σ_1 and σ_2 unknown, assumed equal *

σ_1 and σ_2 unknown, not assumed equal

Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed, *or* both sample sizes are at least 30
- Population variances are unknown but assumed equal

Hypothesis tests for μ_1 - μ_2 with σ_1 and σ_2 unknown and assumed equal

(continued)

Population means, independent samples

- σ_1 and σ_2 unknown, assumed equal *

- σ_1 and σ_2 unknown, not assumed equal

- The pooled variance is:

$$S_p^2 = \frac{(n_1 - 1)S_1^{\,2} + (n_2 - 1)S_2^{\,2}}{(n_1 - 1) + (n_2 - 1)}$$

- The test statistic is:

$$t_{STAT} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- Where t_{STAT} has d.f. = $(n_1 + n_2 - 2)$

Confidence interval for μ_1 - μ_2 with σ_1 and σ_2 unknown and assumed equal

Population means, independent samples

- σ_1 and σ_2 unknown, assumed equal *
- σ_1 and σ_2 unknown, not assumed equal

The confidence interval for $\mu_1 - \mu_2$ is:

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where $t_{\alpha/2}$ has d.f. = $n_1 + n_2 - 2$

“Tests for Variances”: F Test

You are a financial analyst for a brokerage firm. You want to compare dividend yields between stocks listed on the NYSE & NASDAQ. You collect the following data:

	NYSE	NASDAQ
Number	**21**	**25**
Mean	**3.27**	**2.53**
Std dev	**1.30**	**1.16**

Is there a difference in the variances between the NYSE & NASDAQ at the $\alpha = 0.05$ level?

BE MINDFUL…

- You need to check whether the normal assumption holds before you can use the *F*-test.

Hypothesis Tests for Variances

Tests for Two Population Variances *

F test statistic

Hypotheses	F_{STAT}
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	S_1^2 / S_2^2
$H_0: \sigma_1^2 \leq \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	

Where:

S_1^2 = Variance of sample 1 (the larger sample variance)

n_1 = sample size of sample 1

S_2^2 = Variance of sample 2 (the smaller sample variance)

n_2 = sample size of sample 2

n_1 –1 = numerator degrees of freedom

n_2 – 1 = denominator degrees of freedom

F test is useful…

- In the sense of previous example: one has genuine interest in comparing variances.
  - For instance, in financial investment, variance of return indicates market/portfolio's level of *risk*.
- On the other hand, recall in two sample t-test, one should check whether two variances are equal or not.
- F-test can provide info on that.
- <u>Basis of ANOVA (Analysis of Variance)</u>

The F Distribution (**F**isher, by G. Snedecor)

- <u>There are two degrees of freedom required</u>: numerator and denominator

- When $\boxed{F_{STAT} = \dfrac{S_1^2}{S_2^2}}$ $df_1 = n_1 - 1$; $df_2 = n_2 - 1$

F-test example

- Important Step: Check the normality assumption in both samples
- Big Class.jmp: Are the variances of height (or weight) different between Male and Female?
- (Select ‘Fit Y by X’, and select ‘Unequal Variances.’)
- Since the p-value from the 2-sided F-Test is large, you cannot conclude that the variances are unequal.

One more example

- NHANES.jmp example: large data set related to patients' profile and their health status
- BMI (Body Mass Index) vs. Male/Female
- Two sample t-test
  - Equal variance?

Pooled-Variance t Test Example

You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	NYSE	NASDAQ
Number	**21**	**25**
Sample mean	**3.27**	**2.53**
Sample std dev	**1.30**	**1.16**

Assuming both populations are approximately normal with equal variances, is there a difference in mean yield (α = 0.05)?

Pooled-Variance t Test Example: Calculating the Test Statistic

(continued)

H0: $\mu_1 - \mu_2 = 0$ i.e. ($\mu_1 = \mu_2$)
H1: $\mu_1 - \mu_2 \neq 0$ i.e. ($\mu_1 \neq \mu_2$)

The test statistic is:

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\left(3.27 - 2.53\right) - 0}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = \boxed{2.040}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(21-1)1.30^2 + (25-1)1.16^2}{(21-1) + (25-1)} = 1.5021$$

P-value? Sketch the area.

Hypothesis tests for $\mu_1 - \mu_2$ with σ_1 and σ_2 unknown, not assumed equal

Population means, independent samples

- σ_1 and σ_2 unknown, assumed equal
- σ_1 and σ_2 unknown, not assumed equal *

Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed or both sample sizes are at least 30
- Population variances are unknown and cannot be assumed to be equal

Hypothesis tests for μ_1 - μ_2 with σ_1 and σ_2 unknown and not assumed equal

(continued)

Population means, independent samples

σ_1 and σ_2 unknown, assumed equal

σ_1 and σ_2 unknown, not assumed equal *

The test statistic is:

$$t_{STAT} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

t_{STAT} has d.f. v, which has a very complex expression.