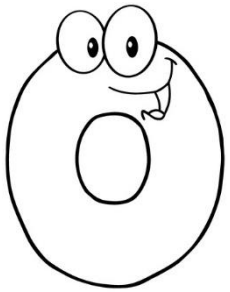
Learn SPSS in 45 Minutes

Requirement

- 0 requirement for statistics knowledge.
- 0 requirement for SPSS experience.
- Nick name of this tutorial: Zero to hero.
- You can download SPSS to try out a few days for free from IBM website. (Link will be provided.)



Target Audience

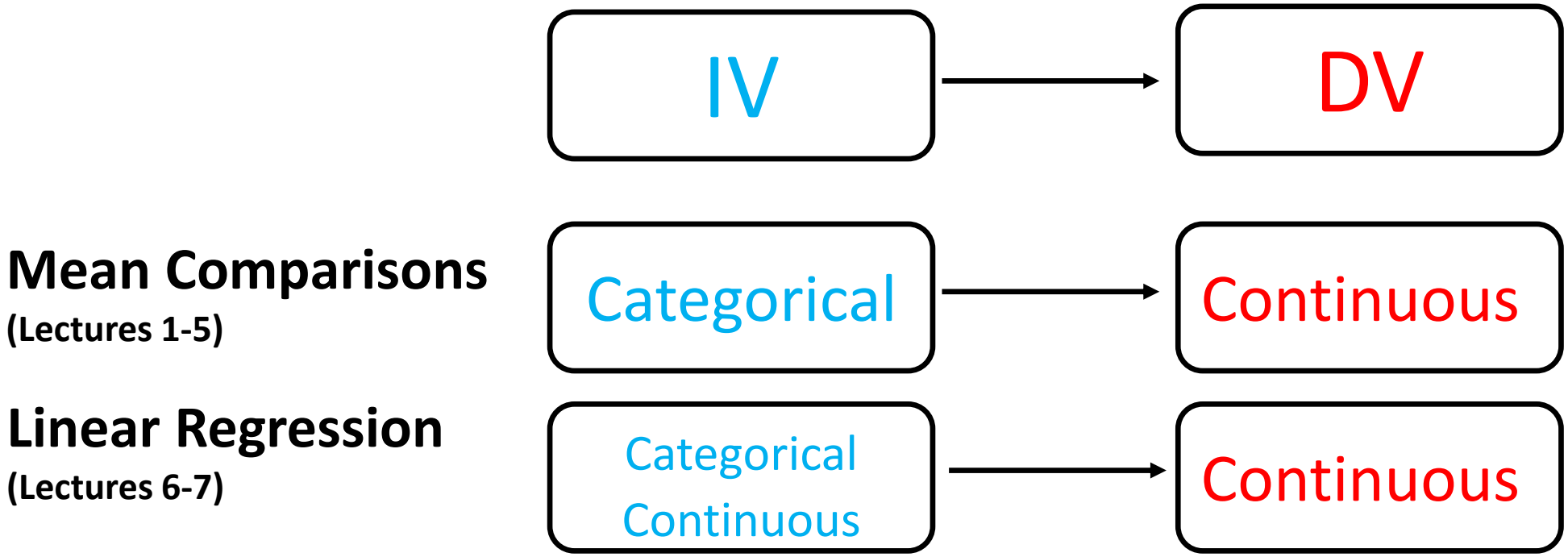
- Interested in data analytics (zero or limited experience).
- For users from both academia and industry.

Outcome of This Tutorial

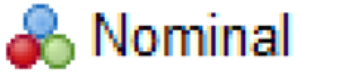
- Know how to do most common statistical analyses.
- Know how to interpret and report these analyses.
- Have an overall knowledge framework of data analytics.

Overview

- **Section 1: Mean Comparisons**
  - Lecture 1: One Sample t-test
  - Lecture 2: Paired sample t-test
  - Lecture 3: Independent t-test
  - Lecture 4: One-Way ANOVA
  - Lecture 5: Two-Way ANOVA
- **Section 2: Linear Regressions**
  - Lecture 6: Simple Linear Regression
  - Lecture 7: Multiple Linear Regression

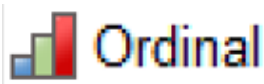


Data Type and Analysis Framework



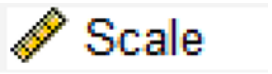
Nominal: Multiple distinct categories within a set.

- **NFL positions** (e.g., WR, LB, CB)
- **Gender** (e.g., male, female).



Ordinal: Multiple distinct categories within a set that have an inherent order.

- **Salary categories** (e.g., High Income jobs, Upper Middle Income jobs, Middle Income jobs, Entry Level Income jobs)
- **Language ability** (e.g., Beginner, Intermediate, Fluent)



Scale: numerical values.

- **Age** (e.g., 2, 18,80,99,102)
- **Income** (e.g., 10,000, 30,000, …..)

Data Type and Analysis Framework

Categorical Variables:

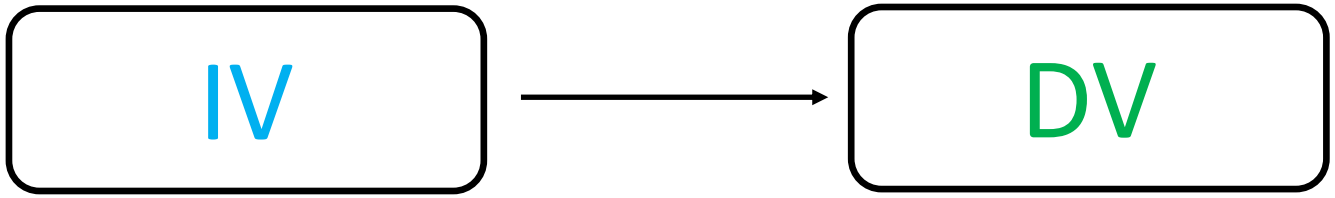
- **Nominal:** Multiple distinct categories within a set.
- **Ordinal:** Multiple distinct categories within a set that have an inherent order.

Continuous Variables:

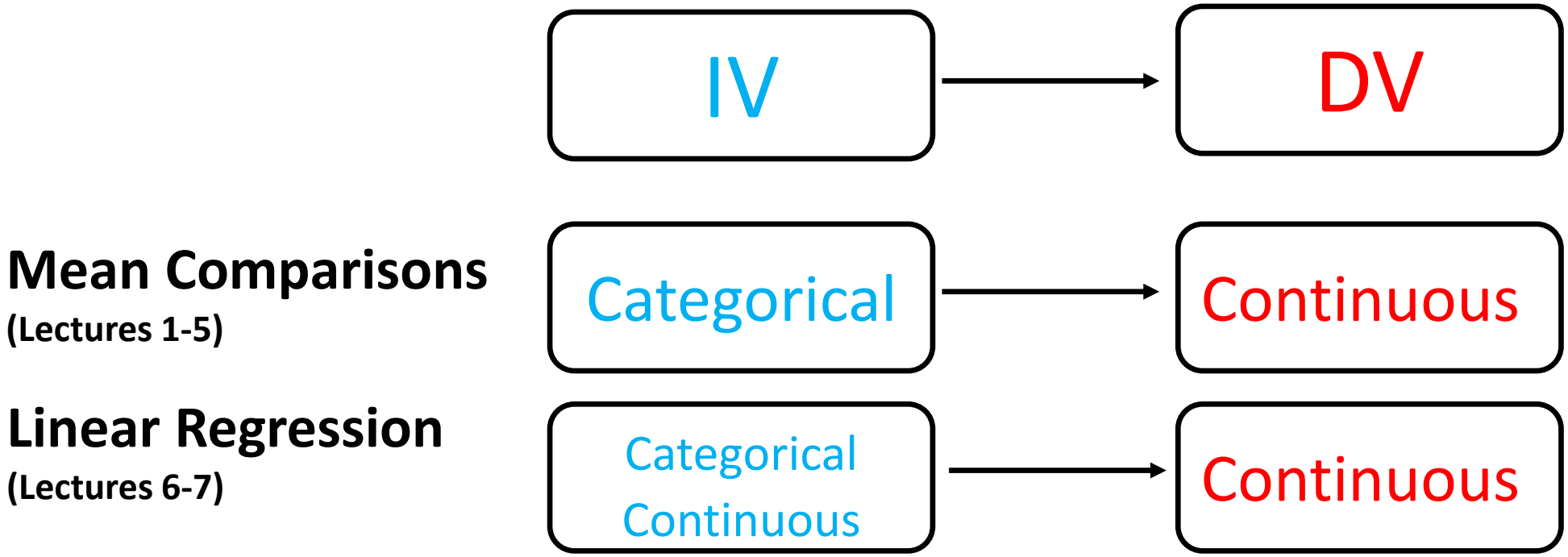
Scale: numerical values.

Data Type and Analysis Framework

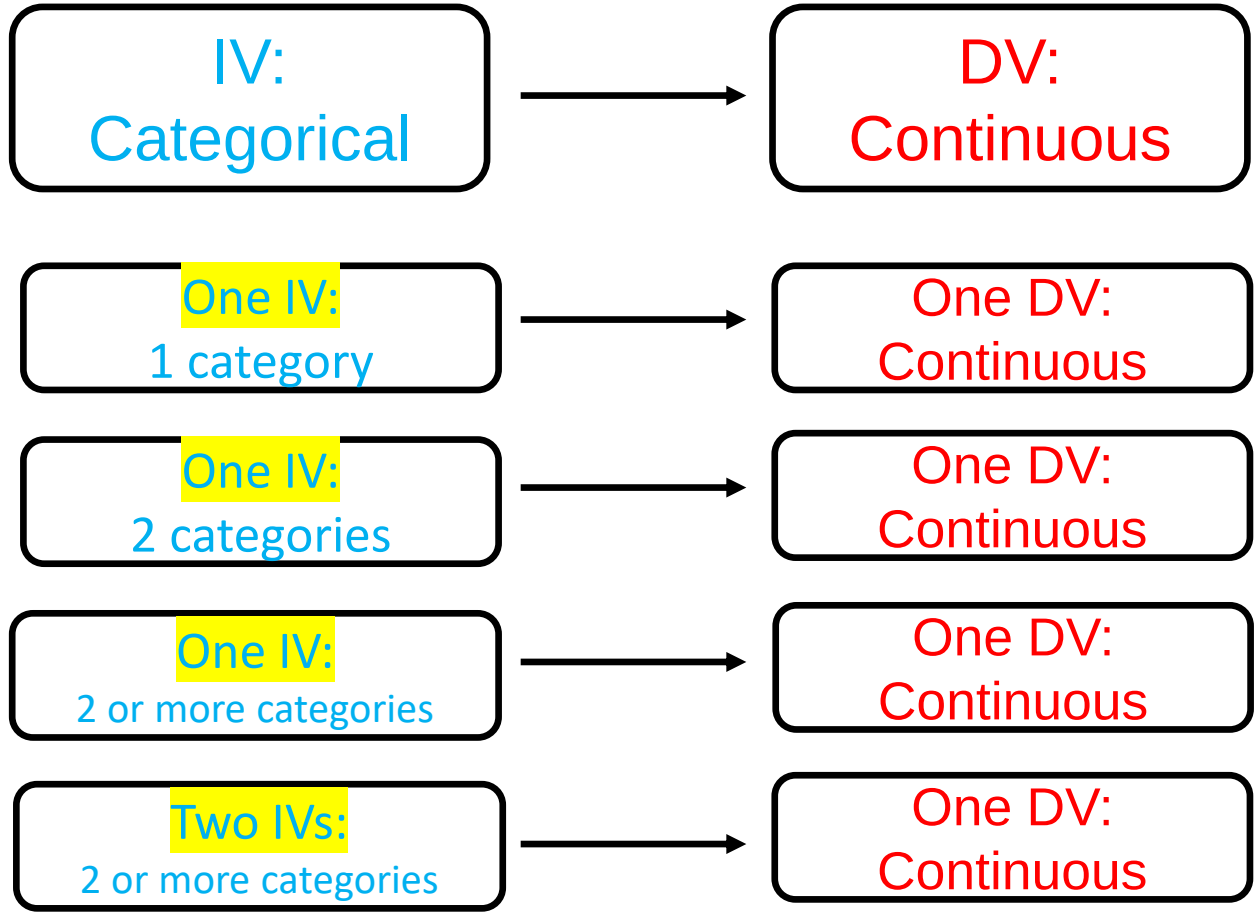
- Most data analytics test relationship between IV (Independent Variable) and DV (Dependent Variable).



- Depending on different data types of IV and DV, you need to use different types of test.



Section 1: Mean Comparisons (Lectures 1-5)



- One Sample t-test
- Paired sample t-test
- Independent sample t-test
- One-Way ANOVA
- Two-Way ANOVA

Notes:
(1) This PDF is part of a YouTube tutorial (https://youtu.be/Vda7kobpLLU). This PDF is for individual, personal usage only.
(2) The author accepts no responsibility for the topicality, correctness, completeness or quality of the information provided.

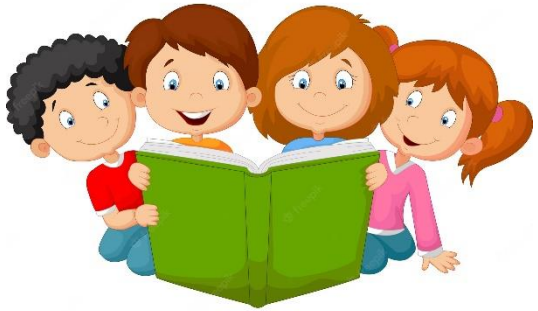
Overview

- **Section 1: Mean Comparisons**
  - Lecture 1: One Sample t-test
  - Lecture 2: Paired sample t-test
  - Lecture 3: Independent t-test
  - Lecture 4: One-Way ANOVA
  - Lecture 5: Two-Way ANOVA
- **Section 2: Linear Regressions**
  - Lecture 6: Simple Linear Regression
  - Lecture 8: Multiple Linear Regression

Lecture #1: One Sample t-test

One sample t-test is to compare the mean of a sample to a fixed value.

- For instance, compare whether the average score of students' reading is different from 45.



- **H0:** The mean of reading scores is not different from 45.
- **H1:** The mean of reading scores is different from 45.

Lecture #1: One Sample t-test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
read	200	52.23	10.253	.725

One-Sample Test

Test Value = 45

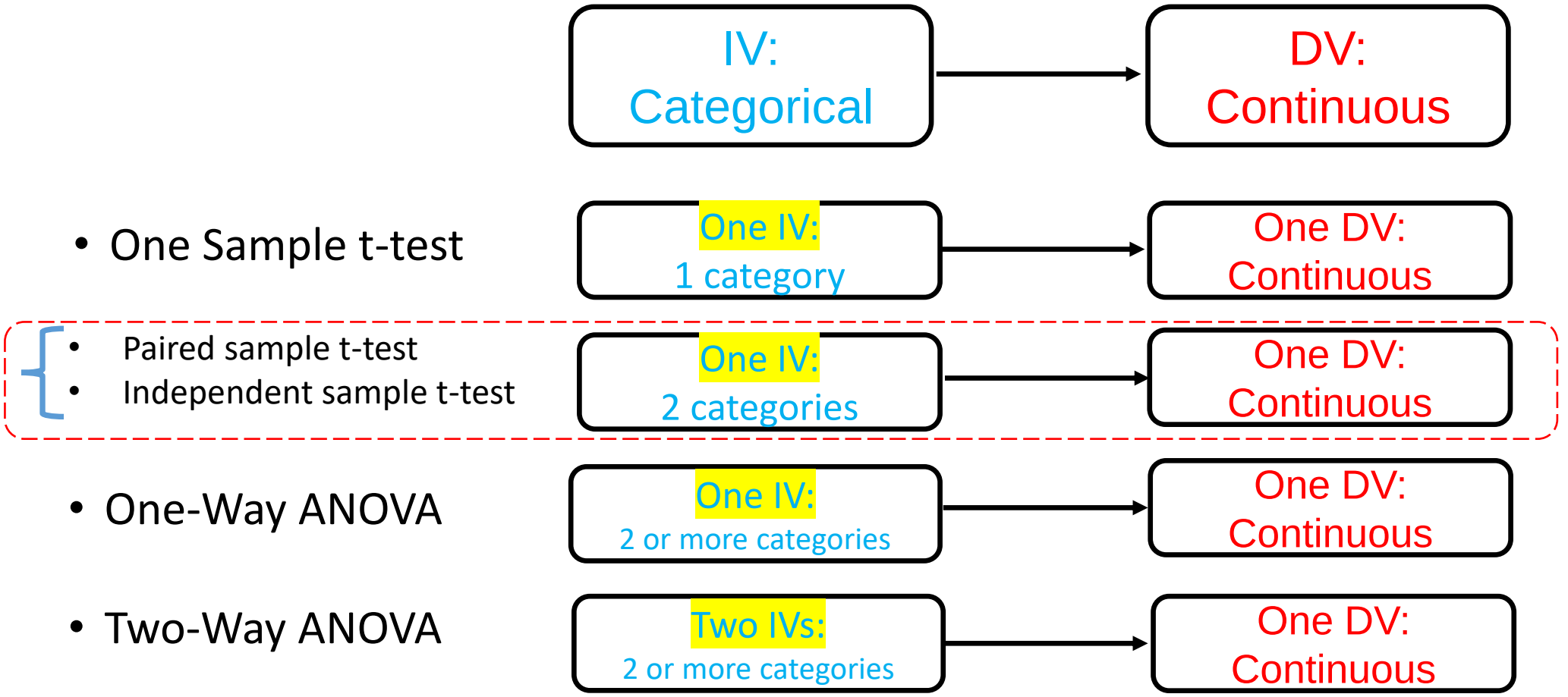
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
read	9.973	199	<.001	7.230	5.80	8.66

- Mean of reading is 52.23 and the p-value is smaller than 0.001. Thus, we reject null hypothesis.
- We conclude that reading scores is significantly different from 45.

Overview

- **Section 1: Mean Comparisons**
  - Lecture 1: One Sample t-test
  - Lecture 2: Paired sample t-test
  - Lecture 3: Independent t-test
  - Lecture 4: One-Way ANOVA
  - Lecture 5: Two-Way ANOVA
- **Section 2: Linear Regressions**
  - Lecture 6: Simple Linear Regression
  - Lecture 8: Multiple Linear Regression

Mean Comparisons (Lectures 1-5)



Notes:

(1) This PDF is part of a YouTube tutorial (https://youtu.be/Vda7kobpLLU). This PDF is for individual, personal usage only.

(2) The author accepts no responsibility for the topicality, correctness, completeness or quality of the information provided.

Lecture #2: Paired Sample t-test

- Paired t-test is used when Category 1 (data 1) and Category 2 (data 2) are from the same group of people or objects.

Example 1: Exam 1 Vs. Exam 2

Example 2: Reading scores Vs. Writing scores

Notes:
(1) This PDF is part of a YouTube tutorial (https://youtu.be/Vda7kobpLLU). This PDF is for individual, personal usage only.
(2) The author accepts no responsibility for the topicality, correctness, completeness or quality of the information provided.

Lecture #2: Paired Sample t-test

Reading scores Vs. Writing scores

- **H0:** Writing and Reading are not different.
- **H1:** Writing and Reading are different.

Lecture #2: Paired Sample t-test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	read	52.23	200	10.253	.725
	write	52.78	200	9.479	.670

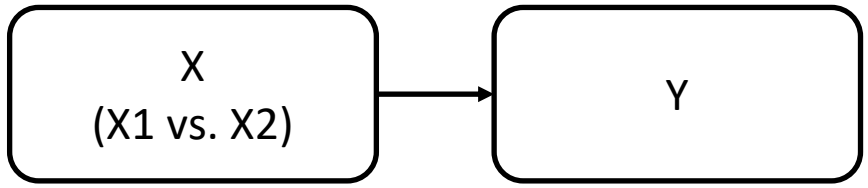
Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	read - write	-.545	8.887	.628	-1.784	.694	-.867	199	.387

- $M_{Reading}$ = 52.23 and $M_{writing}$ = 52.78. P-value of 0.387 is greater than 0.05. We failed to reject the null hypothesis.
- We conclude that writing and reading scores are not significantly different.

Lecture #3: Independent sample t-test

- Paired sample t-test vs. independent sample t-test:



Paired sample t-test: X1 and X2 are from the same group of respondents.

Independent sample t-test: X1 and X2 are from two different groups of respondents.

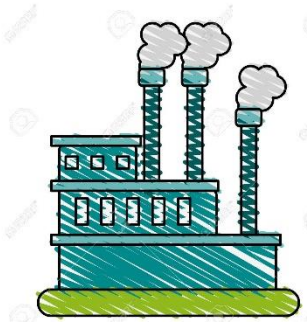
Lecture #3: Independent sample t-test

- Independent (sample) t-test is also called two sample t-test.
  - It deals with situations where Category 1 (data 1, X1) and Category 2 (data 2, X2) are from two different groups of people or objects.

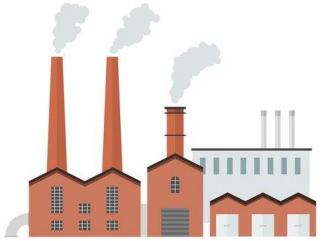
Example 1:
reading scores between males vs. females

Vs.

Example 2:
Product quality ratings between factory A vs. B



Vs.



Notes:
(1) This PDF is part of a YouTube tutorial (https://youtu.be/Vda7kobpLLU). This PDF is for individual, personal usage only.
(2) The author accepts no responsibility for the topicality, correctness, completeness or quality of the information provided.

Lecture #3: Independent sample t-test

Writing score between males vs. females

Vs.

- **H0:** Men and women do not differ in writing scores.
- **H1:** Men and women do differ in writing scores.

Lecture #3: Independent sample t-test

Group Statistics

	Gender_dummy	N	Mean	Std. Deviation	Std. Error Mean
write	male	91	50.12	10.305	1.080
	female	109	54.99	8.134	.779

Independent Samples Test

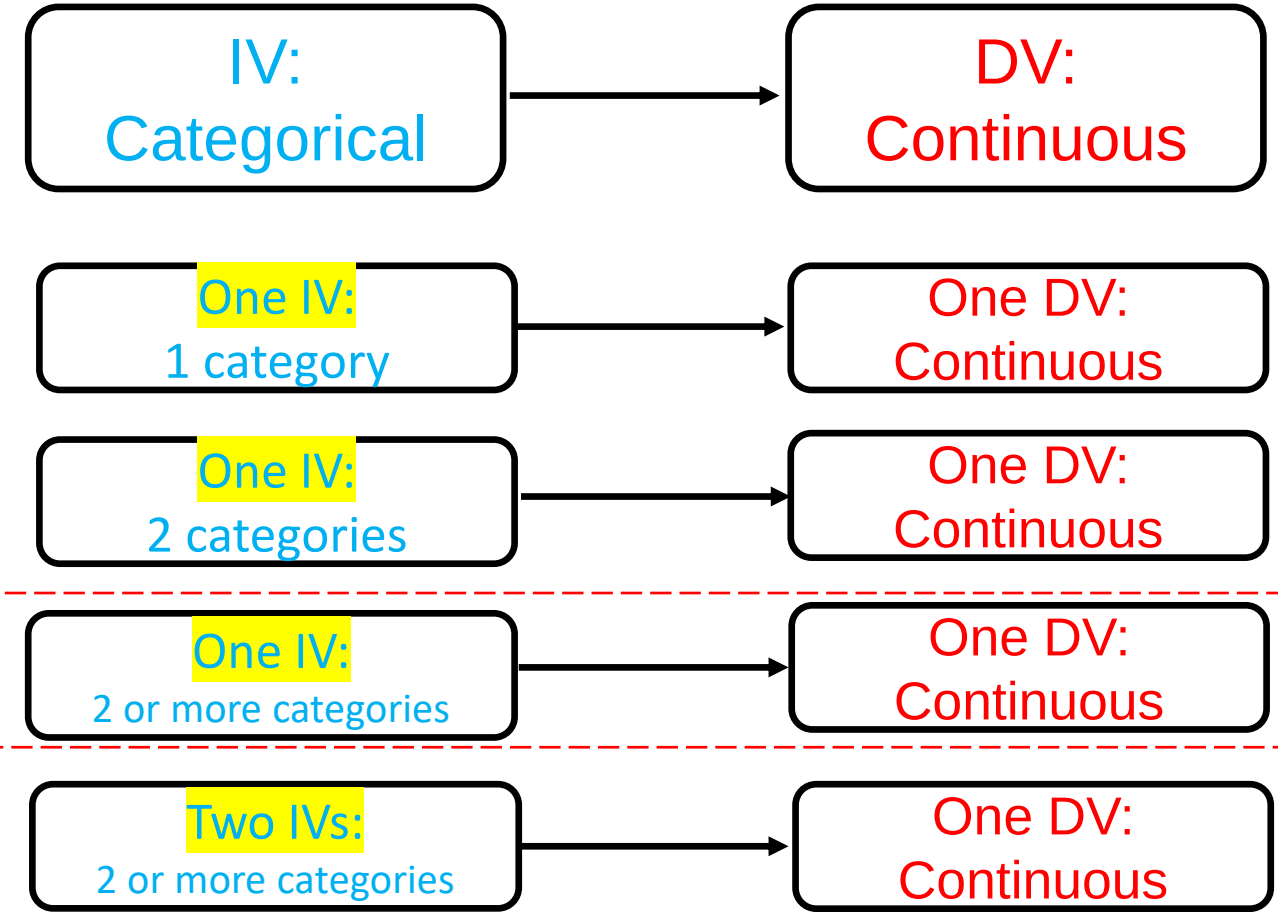
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
write	Equal variances assumed	11.133	.001	-3.734	198	<.001	-4.870	1.304	-7.442	-2.298
	Equal variances not assumed			-3.656	169.707	<.001	-4.870	1.332	-7.499	-2.241

- $M_{Female} = 54.99$ vs. $M_{Male} = 50.12$, t(169.71) = -3.656, p-value < 0.001, and thus we reject the null hypothesis.
- Thus, male and female students significantly differ in writing scores.

Overview

- **Section 1: Mean Comparisons**
  - Lecture 1: One Sample t-test
  - Lecture 2: Paired sample t-test
  - Lecture 3: Independent t-test
  - Lecture 4: One-Way ANOVA
  - Lecture 5: Two-Way ANOVA
- **Section 2: Linear Regressions**
  - Lecture 6: Simple Linear Regression
  - Lecture 7: Multiple Linear Regression

Mean Comparisons (Lectures 1-5)



Notes:

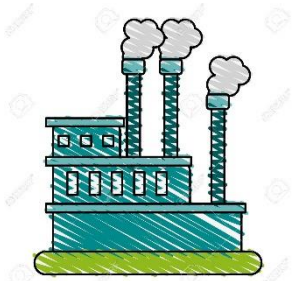
(1) This PDF is part of a YouTube tutorial (https://youtu.be/Vda7kobpLLU). This PDF is for individual, personal usage only.

(2) The author accepts no responsibility for the topicality, correctness, completeness or quality of the information provided.

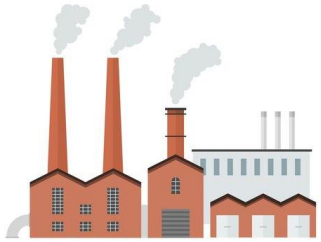
Lecture #4: One-Way ANOVA

- One-way ANOVA ("Analysis of Variance") compares the means of 2 or more groups to determine whether the means are significantly different.

Example 1:
Product quality between
factory A vs. B vs. C



Vs.



Vs.

Example 2:
Sales between
NYC vs. SF vs. Seattle

Vs.

Vs.

Notes:
(1) This PDF is part of a YouTube tutorial (https://youtu.be/Vda7kobpLLU). This PDF is for individual, personal usage only.

Lecture #4: One-Way ANOVA

Writing scores between
3 different programs

- Vocation
- General
- Academic

- **Ho:** $M_{\text{vocation}} = M_{\text{General}} = M_{\text{Academic}}$
- **H1:** At least one pair of M_{vocation} M_{General} M_{Academic} is not equal.

Lecture #4: One-Way ANOVA

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
vocation	50	46.76	9.319	1.318	44.11	49.41	31	67
general	45	51.33	9.398	1.401	48.51	54.16	31	67
academic	105	56.26	7.943	.775	54.72	57.79	33	67
Total	200	52.78	9.479	.670	51.45	54.10	31	67

ANOVA

write

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3175.698	2	1587.849	21.275	<.001
Within Groups	14703.177	197	74.635		
Total	17878.875	199			

- M_{vocation} = 46.76, M_{General} = 51.33, M_{Academic} =56.26. $F(2, 197) = 21.28$, $p < 0.001$. Thus, we reject the null hypothesis.
- At least one pair of M_{vocation} M_{General} M_{Academic} is not equal.

Lecture #4: One-Way ANOVA

Multiple Comparisons

Dependent Variable: write

Tukey HSD

(I) prog_numeric	(J) prog_numeric	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
vocation	general	-4.573*	1.775	.029	-8.77	-.38
	academic	-9.497*	1.484	<.001	-13.00	-5.99
general	vocation	4.573*	1.775	.029	.38	8.77
	academic	-4.924*	1.539	.005	-8.56	-1.29
academic	vocation	9.497*	1.484	<.001	5.99	13.00
	general	4.924*	1.539	.005	1.29	8.56

*. The mean difference is significant at the 0.05 level.

- We can see that all pairs have a p-value that is smaller than 0.05.
- That means all 3 categories are significantly different from one another.

Overview

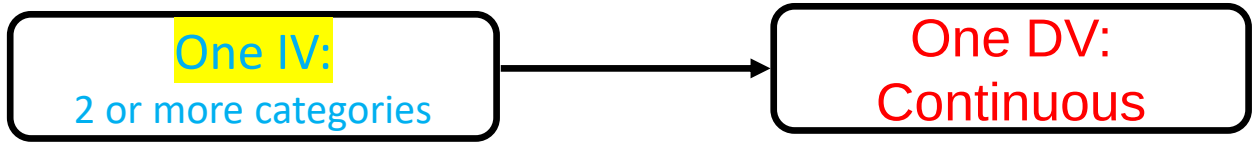
- **Section 1: Mean Comparisons**
  - Lecture 1: One Sample t-test
  - Lecture 2: Paired sample t-test
  - Lecture 3: Independent t-test
  - Lecture 4: One-Way ANOVA
  - Lecture 5: Two-Way ANOVA
- **Section 2: Linear Regressions**
  - Lecture 6: Simple Linear Regression
  - Lecture 7: Multiple Linear Regression

Mean Comparisons (Lectures 1-5)

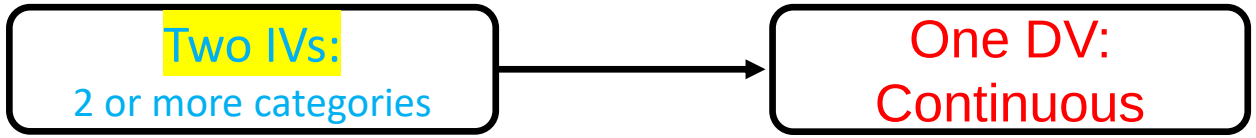
- One Sample t-test

- Paired sample t-test
- Independent sample t-test

- One-Way ANOVA



- Two-Way ANOVA



Lecture #5: Two-Way ANOVA

- It tests whether the effect of X_1 on Y is dependent on X_2.

- Brand = X_1
- City = X_2
- Sales = Y

- Sales of Brand 1 and Brand 2 are different.
  - For this, we say there is an effect of X_1 on Y.
- Brand 1 and Brand 2 sales difference in City 1 is much bigger than City 2.
  - For this, we say the effect of X_1 on Y is dependent on X_2.

Lecture #5: Two-Way ANOVA

- H0: The effect of Brand (Brand 1 vs. Brand 2) on Sales does not depend on the effect of City (a.k.a. There is no interaction effect).

- H1: The effect of Brand (Brand 1 vs. Brand 2) on Sales does depend on the effect of City (a.k.a. There is interaction effect).

Lecture #5: Two-Way ANOVA

Tests of Between-Subjects Effects

Dependent Variable: sales

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6365.800[a]	3	2121.933	6.462	.005
Intercept	5848.200	1	5848.200	17.810	<.001
City	2121.800	1	2121.800	6.462	.022
Brand	2163.200	1	2163.200	6.588	.021
City * Brand	2080.800	1	2080.800	6.337	.023
Error	5254.000	16	328.375		
Total	17468.000	20			
Corrected Total	11619.800	19			

a. R Squared = .548 (Adjusted R Squared = .463)

- The interaction effect is $F(1, 19) = 6.34$, $p = 0.023$. Thus, the interaction of Brand × City is significant.

Lecture #5: Two-Way ANOVA

Descriptive Statistics

Dependent Variable: sales

City	Brand	Mean	Std. Deviation	N
City1	brand1	48.00	35.637	5
	brand2	6.80	4.382	5
	Total	27.40	32.319	10
City2	brand1	7.00	3.391	5
	brand2	6.60	3.578	5
	Total	6.80	3.293	10
Total	brand1	27.50	32.195	10
	brand2	6.70	3.773	10
	Total	17.10	24.730	20

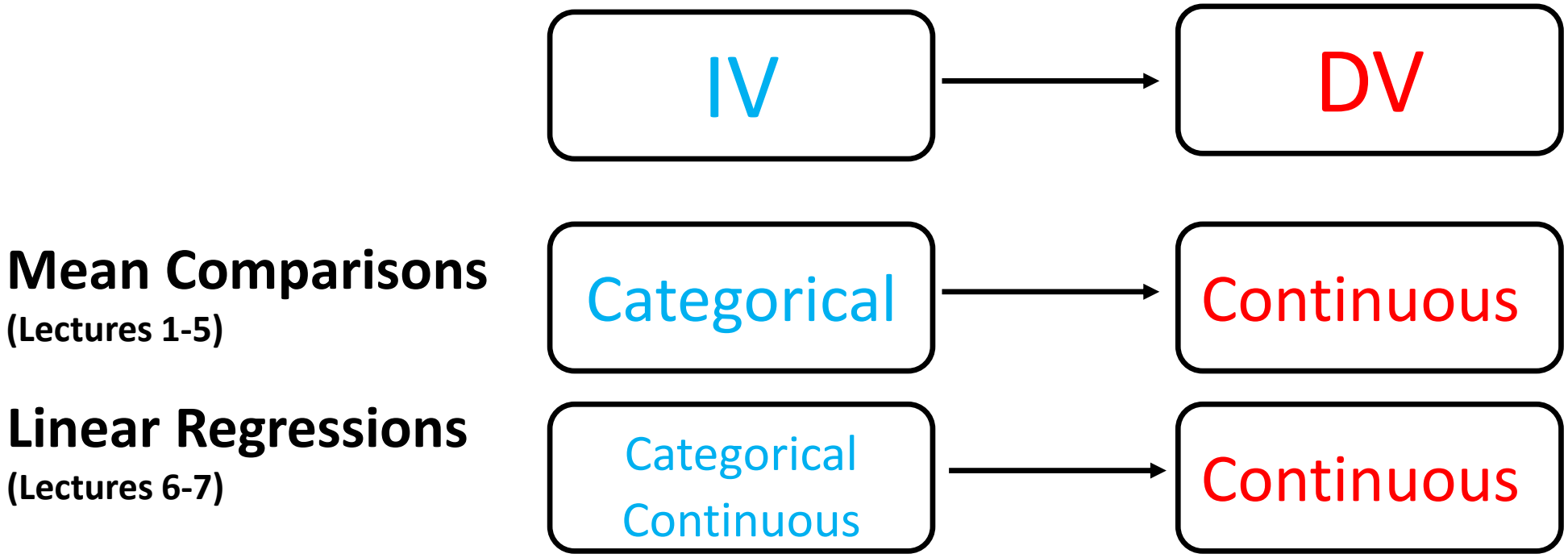
Lecture #5: Two-Way ANOVA

	City 1	City 2
Brand 1	48.0	7.0
Brand 2	6.8	6.6
Difference between Brand 1 and Brand 2	48.0-6.8=41.2	7.0-6.6=0.4

- For City 1, the sales difference between Brand 1 and Brand 2 is 41.2. For City 2, the difference is 0.4.
- Thus, the effect of Brand (Brand 1 vs. Brand 2) on Sales does depend on the effect of City (a.k.a. There is interaction effect).

Overview

- **Section 1: Mean Comparisons**
  - Lecture 1: One Sample t-test
  - Lecture 2: Paired sample t-test
  - Lecture 3: Independent t-test
  - Lecture 4: One-Way ANOVA
  - Lecture 5: Two-Way ANOVA
- **Section 2: Linear Regressions**
  - Lecture 6: Simple Linear Regression
  - Lecture 7: Multiple Linear Regression



Notes:
(1) This PDF is part of a YouTube tutorial (https://youtu.be/Vda7kobpLLU). This PDF is for individual, personal usage only.
(2) The author accepts no responsibility for the topicality, correctness, completeness or quality of the information provided.

Lecture #6: Simple Linear Regression

- Simple linear regression estimates the linear relationship between one independent variable and one dependent variable.

- We can test whether students' writing scores can be used to predict reading scores.

$$\textit{Read} = b_0 + b_1\,\textit{Write}$$

- **H0:** $b_1 = 0$.
- **H1:** $b_1 \neq 0$.

Lecture #6: Simple Linear Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	18.162	3.307		5.492	<.001
	write	.646	.062	.597	10.465	<.001

a. Dependent Variable: read

- b_1 =0.646, and the p-value for "write" is < 0.001, which suggests we should reject the null hypothesis.
- Thus, writing scores can be use as a predictor for reading scores.

$$Read = b_0 + b_1\ Write = 18.16 + 0.65\ Write$$

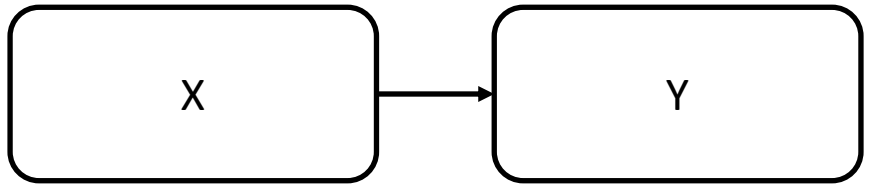
It means that, for writing scores to increase 1 unit, reading scores will increase 0.65 unit.

Overview

- **Section 1: Mean Comparisons**
  - Lecture 1: One Sample t-test
  - Lecture 2: Paired sample t-test
  - Lecture 3: Independent t-test
  - Lecture 4: One-Way ANOVA
  - Lecture 5: Two-Way ANOVA
- **Section 2: Linear Regressions**
  - Lecture 6: Simple Linear Regression
  - Lecture 7: Multiple Linear Regression

Lecture #7: Multiple Linear Regression

- Multiple linear regression estimates the linear relationship between multiple independent variables and one dependent variable.



Simple linear regression: One X

Multiple linear regression: Multiple X

Lecture #7: Multiple Linear Regression

- We would like to test whether students' writing scores and math scores can be used to predict reading scores.

$$\textit{Read} = b_0 + b_1\,\textit{Write} + b_2\,\textit{Math}$$

- **H0:** $b_1 = 0$.
- **H1:** $b_1 \neq 0$.

- **H0:** $b_2 = 0$.
- **H1:** $b_2 \neq 0$.

Lecture #7: Multiple Linear Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	7.542	3.268		2.308	.022
	write	.328	.070	.304	4.720	<.001
	math	.520	.070	.475	7.382	<.001

a. Dependent Variable: read

- The p-values for both "write" and "math" are < 0.001, and thus we reject both null hypotheses.

$$\textit{Read} = b_0 + b_1\,\textit{Write} + b_2\,\textit{Math} = 7.54 + 0.33\,\textit{Write} + 0.52\,\textit{Math}$$

- For writing scores to increase 1 unit, reading scores will increase 0.33 unit.
- For math scores to increase 1 unit, reading scores will increase 0.52 unit.