Data Visualization and Basic Statistical Testing

Kimberly Greco, MPH

View course video.

Boston Children's Hospital
Until every child is well

Course Overview

Course Objective

Provide a foundation in the basic statistical methods and principles necessary to understand, interpret, and communicate insights from data.

Course Structure

Lecture 1: Getting to Know Your Data: Types of Data and Descriptive Statistics

Lecture 2: Sampling Concepts and Comparing Two Means

Lecture 3: Linear Models and Correlation

Lecture 4: Comparing Proportions and Measures of Association

Lecture Outline

- ❑ **Analysis of Variance (ANOVA)**
  - *Problem with Multiple Comparisons*
  - *Comparing ≥ 3 population means → ANOVA*
- ❑ **Correlation**
  - *Linear relationship between two continuous variables*
- ❑ **Linear Regression**

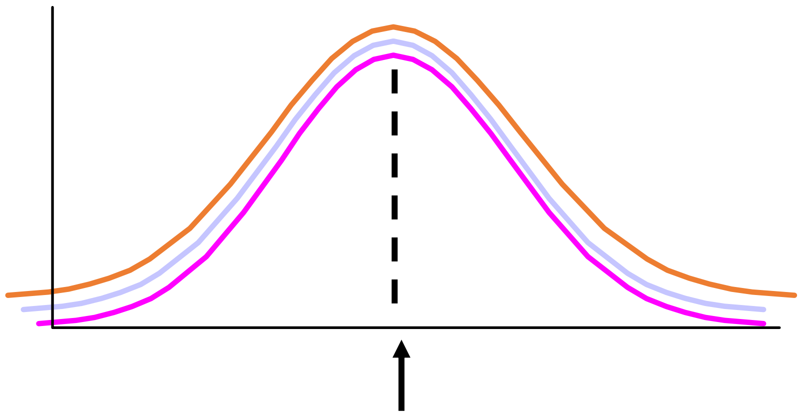
Example: Emergency room admissions by the time of month in 1999

Before Full Moon	During Full Moon	After Full Moon
6.4	12	11.4
7.1	13	10.3
6.5	14	15.8
8.1	12	11
8.6	16	11.1
9.4	11	5.8
11.5	13	9.2
9.5	16	7.9
5.4	19	7.7
11.7	13	11
10.8	20	10
9.6	14	12.1

Question: Is there a difference in the number of ER admissions based on moon cycle?

Example: No difference in mean admissions (μ) by moon cycle

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_c$$
$$H_1 : \text{ Not all } \mu_i \text{ are the same}$$



$$\mu_1 = \mu_2 = \mu_3$$

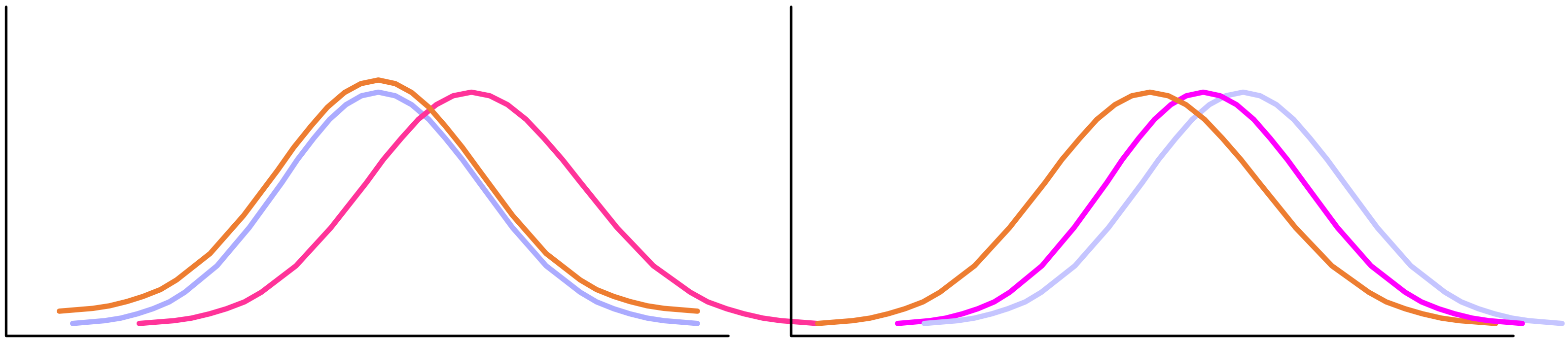
Scenario: Null Hypothesis is True

Example: Difference in mean admissions (μ) by moon cycle

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_c$$

$$H_1 : \text{ Not all } \mu_i \text{ are the same}$$

Scenario: Null Hypothesis is NOT True



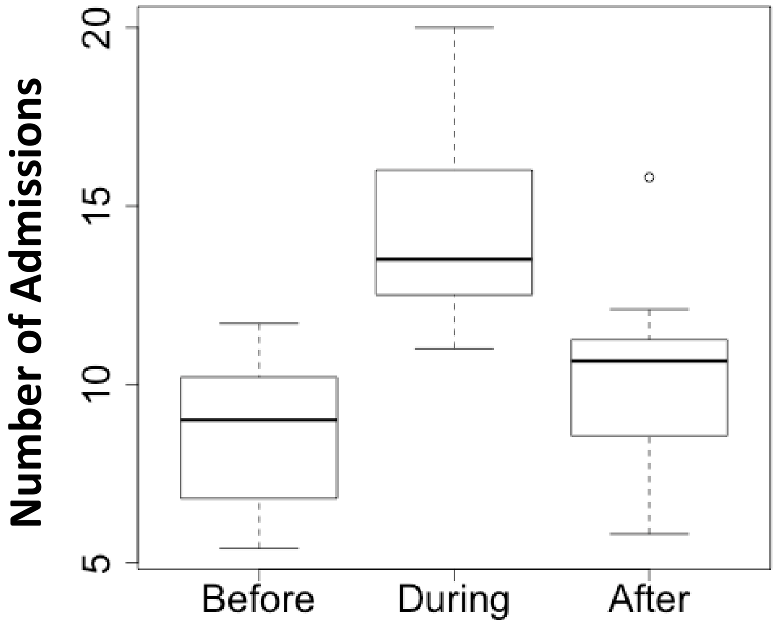
$$\mu_1 = \mu_2 \neq \mu_3$$

$$\mu_1 \neq \mu_2 \neq \mu_3$$

Example: Admissions Summary Statistics & Graph

	N	Mean	Standard Deviation	Standard Error of Mean
Before	12	8.717	2.0701	0.5976
During	12	14.42	2.811	0.8115
After	12	10.28	2.5295	0.7302
All Groups	36	11.14	3.434	0.5723

Variance of admissions = $(3.434)^2$ = 11.7924



***Note**: not a linear relationship between moon cycle and number of admissions*

We are comparing means…
So, can we use a t-test for this?

With three means, there are three possible comparisons:

- **Before** vs. **During** full moon
- **Before** vs. **After** full moon
- **During** vs. **After** full moon

We can use three pairwise t-tests to compare three means:

- T-test for mean **Before** vs. mean **During**
- T-test for mean **Before** vs. mean **After**
- T-test for mean **During** vs. mean **After**

However, problem arises when we do this….

Boston Children's Hospital
Until every child is well

The problem with using multiple t-tests...

More generally: The problem with multiple comparisons

Type I error is rejecting H_0 when H_0 is true (false positive)

- The probability of making a type I error is represented by the alpha level (α), which is the p-value below which you reject the null hypothesis
- Any time you reject H_0 because p-value < α, it's possible that you're wrong (i.e., H_0 is true and your significant result is due to chance)
- $\alpha = 0.05$ translates to a 5% chance of a false positive

		The Truth (Based on Entire Population)	
		Nothing Is There (H_0 Is True)	Something Is There (H_0 Is False)
Your Conclusion (Based on Your Sample)	I Don't See Anything (Nonsignificant)	Right!	Wrong (Type II Error)
	I See Something (Significant)	Wrong (Type I Error)	Right!

The problem with using multiple t-tests…

More generally: The problem with multiple comparisons

Each hypothesis test contains a type I error (α)

- So far we have used α=0.05 (i.e., 95% confidence interval)

Type I error for one comparison: $1 - (1 - \alpha) = 1 – (0.95) =$ **0.05**

Type I error for three comparisons: $1 - (1 - \alpha)^3 = 1 – (0.95)^3 =$ **0.14**

14% of the time we will reject H_0 (means are equal) in favor of H_1 (means are not equal) even when H_0 is true

- **14%** of the time we could draw the wrong conclusion – not **5%**!

Boston Children's Hospital
Until every child is well

The problem with using multiple t-tests…

More generally: The problem with multiple comparisons

What if we have five means and $\alpha = 0.05$?

- We need ten pairwise t-tests to compare five means

Type I error for <u>one</u> comparison: $1 - (1 - \alpha) = 1 – (0.95) =$ **0.05**
Type I error for <u>ten</u> comparisons: $1 - (1 - \alpha)^{10} = 1 – (0.95)^{10} =$ **0.40**

- **40%** of the time we could draw the wrong conclusion – not **5%**!

The problem with using multiple t-tests…

More generally: The problem with multiple comparisons

In general, if you have k comparisons:

$$\text{Total Type I error} = 1 - (1 - \alpha)^k$$

To avoid this issue with total type I error, we use the **analysis of variance (ANOVA)** method

What is Analysis of Variance (ANOVA)?

- **Statistical test to compare 3 or more population means**
  - Continuous dependent variable & categorical independent variable(s)
  - Generalizes the t-test beyond two means

- **Hypotheses**

  $\mathbf{H_0}$: The population means of all groups are equal

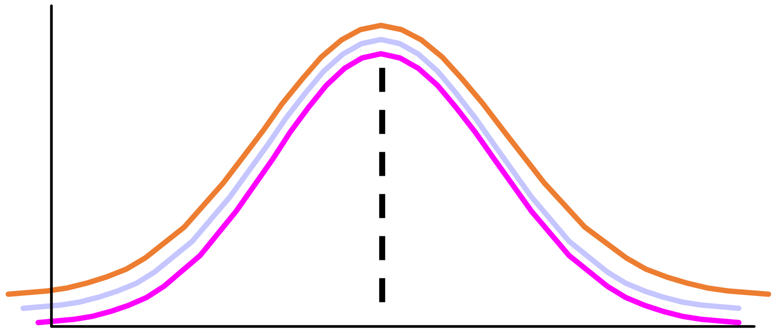
  $$(\mu_1 = \mu_2 = \ \dots = \mu_k)$$

  $\mathbf{H_1}$: <u>At least one</u> population mean differs from the others

ANOVA Assumptions

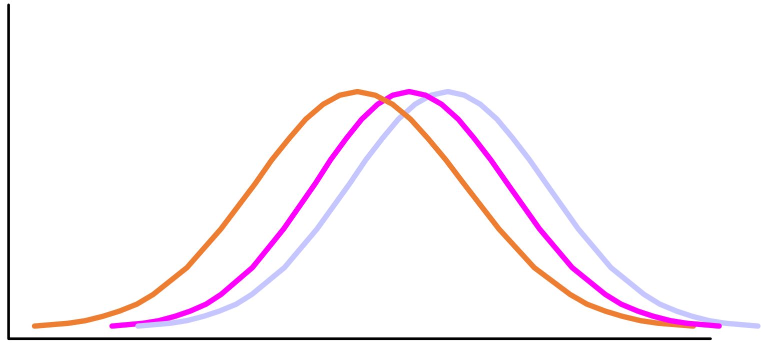
- Random samples are drawn from independent observations
- Underlying population variances are equal
- Underlying data are approximately normally distributed
- Use when data are quantitative
- Assume no shape to the relationship between dependent and independent variable (i.e., linear)

No difference in 3 means – variance equal



$$\mu_1 = \mu_2 = \mu_3$$

Difference in 3 means – variance equal



$$\mu_1 \neq \mu_2 \neq \mu_3$$

Analysis of Variance (ANOVA)

	N	Mean	Standard Deviation	Standard Error of Mean
Before	12	8.717	2.0701	0.5976
During	12	14.42	2.811	0.8115
After	12	10.28	2.5295	0.7302
All Groups	36	11.14	3.434	0.5723

Variance of admissions = $(3.434)^2$ = 11.7924

- ANOVA evaluates if independent variable(s) in a model (moon cycle) explain the **total variation** in the dependent variable (admissions)
- To get at this idea of **total variation**…
  - **Sample mean** of responses for each group (before, during, after)
  - **Grand mean** of all responses, irrespective of group

Analysis of Variance (ANOVA)

	N	Mean	Standard Deviation	Standard Error of Mean
Before	12	8.717	2.0701	0.5976
During	12	14.42	2.811	0.8115
After	12	10.28	2.5295	0.7302
All Groups	36	11.14	3.434	0.5723

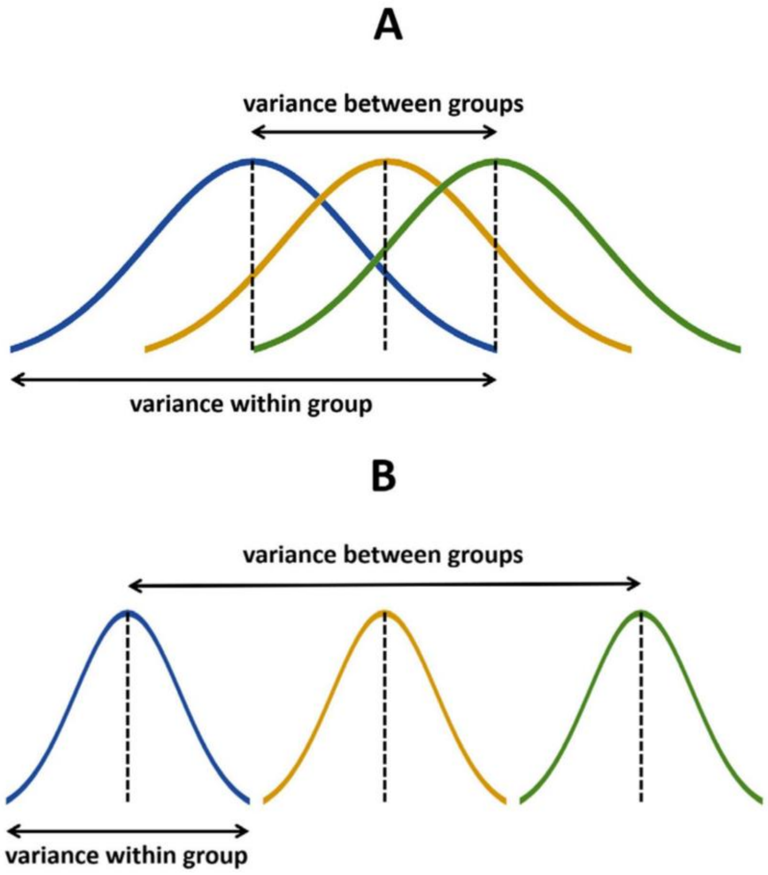
Variance of admissions = $(3.434)^2$ = 11.7924

- Viewed as one sample (rather than *k* samples from individual groups), we measure the total variability among observations (n=36)
- **Total variation** in the dependent variable is equal to:
  - Summing the squares of the differences between each observation (irrespective of group) and the grand mean
  - **sample variance * (n-1)**
  - Called "sum of squares total" (SST)

Total variation in admissions

SST = (11.7924)*(36-1) = 412.723

Partitioning the Variance



Total variation (SST) in the dependent variable has two sources:

1. **“Variation due to Model” → Variation due to independent variables**
   - Variance between groups
   - Calculated as the variance between each group mean and the grand mean
2. **“Variation due to Random Sampling” → Error variation**
   - Variance within groups
   - Calculated as the variance between each observation in a group and its group mean

Boston Children's Hospital
Until every child is well

Partitioning the Variance

Total Variation (SST) = **Variation Due to Model** + **Variation Due to Random Sampling**

Commonly referred to as:
- Sum of Squares Among
- **Sum of Squares Between**
- Sum of Squares Model
- Sum of Squares Explained
- Sum of Squares Treatment
- Among Groups Variation

Commonly referred to as:
- **Sum of Squares Within**
- Sum of Squares Error
- Sum of Squares Unexplained
- Within Groups Variation

Note: SPSS uses between group and within group terms in output

Example: Difference in mean admissions (μ_i) by moon cycle

Total Variation (SST)	=	Variation Due to Model	+	Variation Due to Random Sampling
SST = 412.723		SS Model = 208.287		SS Error = 204.436

To evaluate whether moon cycle explains the variation in admissions…

Step 1: Compute mean squares: **MS Model = SS Model / (k-1)**
MS Error = SS Error / (n-k)

k-1 df for MS model since it measures the variation of the k group means about the grand mean

n-k df for MS error since it measures the variation of the n observations about k group means

**May be helpful to think of mean squares as standard deviations*
**For the admissions example: n = number of observations = 36*
k = number of groups = 3

Boston Children's Hospital
Until every child is well™

Example: Difference in mean admissions (μ_i) by moon cycle

Total Variation (SST)	=	Variation Due to Model	+	Variation Due to Random Sampling
SST = 412.723		SS Model = 208.287		SS Error = 204.436

To evaluate whether moon cycle explains the variation in admissions…

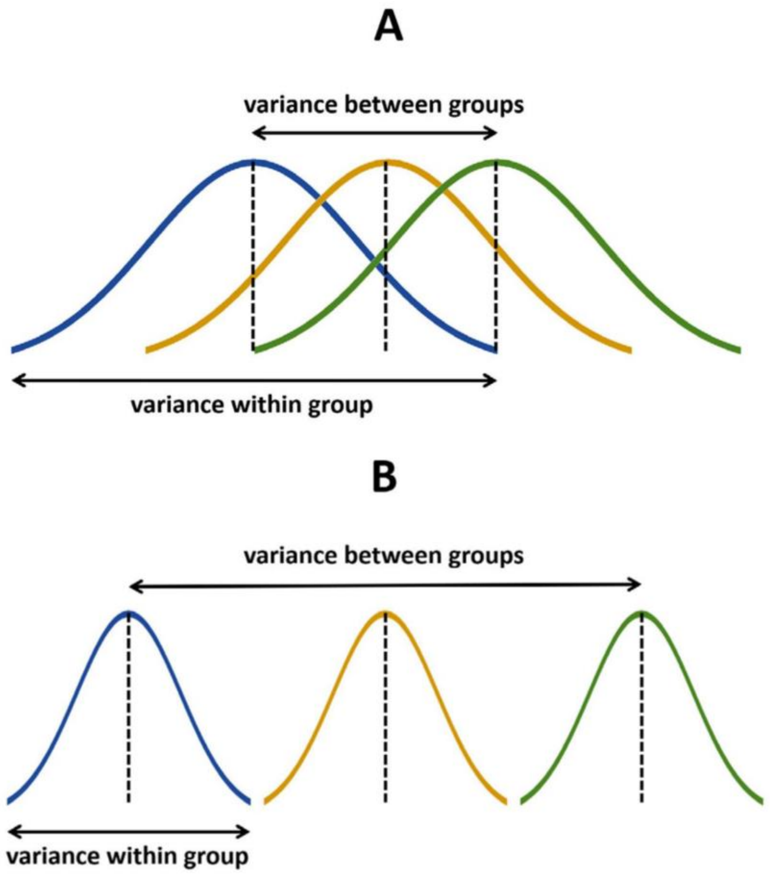
Step 1: Compute mean squares: **MS Model = SS Model / (k-1)**
MS Error = SS Error / (n-k)

Step 2: Compute F-statistic: **F = (MS Model) / (MS Error)**

**F-statistic is a measure of the variability between groups divided by a measure of the variability within groups*

Boston Children's Hospital
Until every child is well

F-Statistic



F = (MS Model) / (MS Error)
F = (MS Between) / (MS Within)

F is small → variability between groups is small relative to the variation within groups (there is probably no difference among these groups – do not reject the null hypothesis)

F is large → variability between groups is large relative to the variation within groups (there is probably a difference among these groups – reject the null hypothesis of equal means)

Example: Difference in mean admissions (μ_i) by moon cycle

Total Variation (SST)		Variation Due to Model		Variation Due to Random Sampling
SST = 412.723	=	SS Model = 208.287	+	SS Error = 204.436

To evaluate whether moon cycle explains the variation in admissions…

Step 1: Compute mean squares: **MS Model = SS Model / (k-1)**
MS Error = SS Error / (n-k)

Step 2: Compute F-statistic: **F = (MS Model) / (MS Error)**

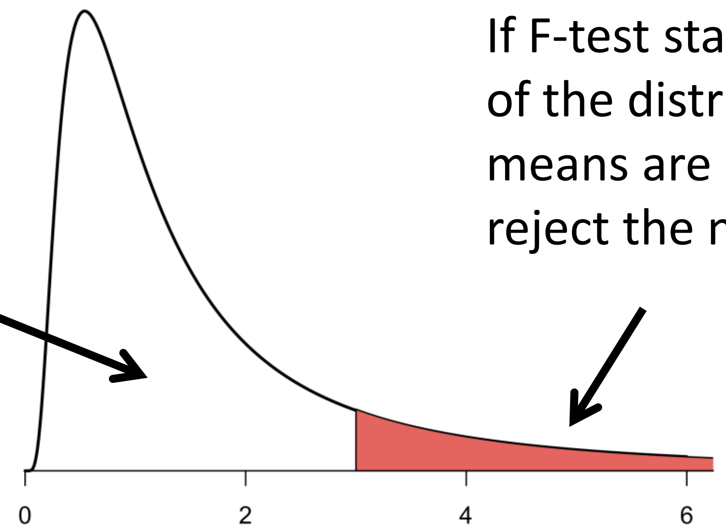
Step 3: Compare F-statistic to F-distribution

F-Distribution

The mathematical equation for the F-distribution below requires 2 values to define (denoted df_1 and df_2, where df = degrees of freedom):

- $df_1 = k-1$
- $df_2 = n-k$
- n = number of subjects and k = number of groups

If F-test statistic falls in this part of the distribution then do not conclude the means are different (i.e., do not reject the null hypothesis)



If F-test statistic falls in this part of the distribution then conclude means are different (i.e., do reject the null hypothesis)

Example: Difference in mean admissions (μ_i) by moon cycle

Total Variation (SST)		Variation Due to Model		Variation Due to Random Sampling
SST = 412.723	=	**SS Model = 208.287**	+	**SS Error = 204.436**

To evaluate whether moon cycle explains the variation in admissions…

Step 1: Compute mean squares:

MS Model = SS Model / (k-1)
MS Model = (208.287) / (3-1)
MS Model = 104.144

MS Error = SS Error / (n-k)
MS Error = (204.436) / (36-3)
MS Error = 6.195

Step 2: Compute F-statistic: F = (MS Model) / (MS Error) = 104.144 / 6.195 = **16.811 (p-value <0.0001)**

Step 3: Compare F-statistic to F-distribution with $df_1=2$, $df_2=33$

Boston Children's Hospital
Until every child is well

ANOVA
SPSS: Analyze > Compare Means > One-Way ANOVA

ANOVA Table in SPSS:

	Sum of Squares	df	Mean Square (MS)	F	Sig.
Between Groups	208.287	2	104.144	16.811	0.000
Within Groups	204.436	33	6.195		
Total	412.723	35			

Variation due to model (SS Model)

Variation due to random sampling (SS Error)

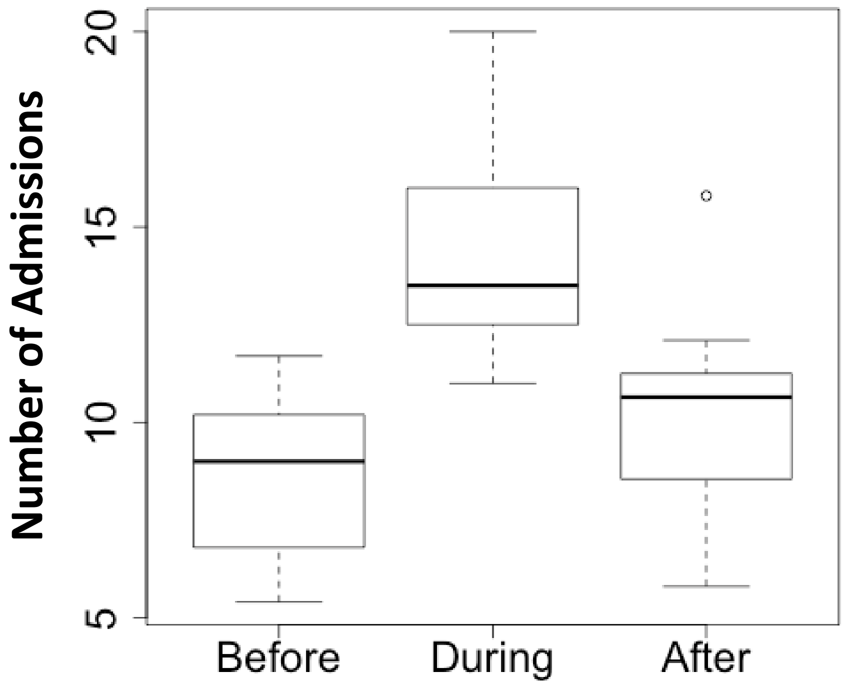
MS for Model

MS for Error

F-statistic

Conclusion: Data indicate that there is <u>at least one</u> difference in the mean admissions by moon cycle ($p<0.0001$) with mean number of admissions of 8.7, 14.4, and 10.3 for before, during, and after full moon, respectively.

Data shows means are different… but which ones?



- There are 3 possible comparisons of means:
  - **Before** vs. **During** full moon
  - **Before** vs. **After** full moon
  - **During** vs. **After** full moon
- Recall our hypotheses:

H_0 : The population means of all groups are equal

$$(\mu_{Before} = \mu_{During} = \mu_{After})$$

H_1 : **At least one** population mean differs

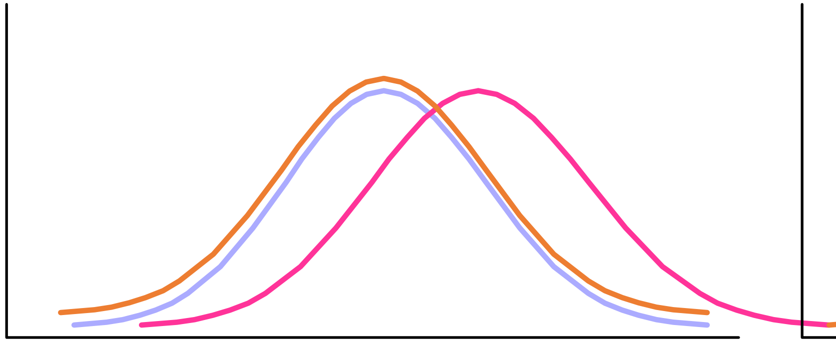
$$(NOT\ \mu_{Before} \neq \mu_{During} \neq \mu_{After})$$

Data shows means are different… but which ones?

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_c$$

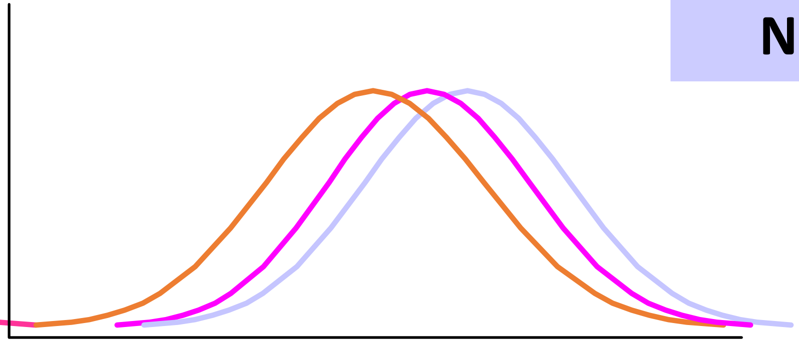
$$H_1 : \text{ Not all } \mu_i \text{ are the same}$$

Scenario: Null Hypothesis is NOT True



$$\mu_1 = \mu_2 \neq \mu_3$$

2 means are the same



$$\mu_1 \neq \mu_2 \neq \mu_3$$

All 3 means are different

Data shows means are different… but which ones?

Statistically significant F-test for ANOVA…

- Indicates that not all of the group means are equal
- Does NOT identify which particular differences between pairs of means are significant.

The role of post-hoc testing is to explore differences between multiple group means while controlling the experiment-wise error rate (usually $\alpha = 0.05$)

- Should only be performed after a statistically significant "global" F-test
- A few methods…
  - Comparing all groups against each other (all pairwise comparisons)
  - Comparing specific pairs of interest (specific pairwise comparison)
  - Comparing all treatment groups against one control group.

Multiple Comparison Post-Hoc Methods

Several procedures (partial list):

Duncan	**Most Liberal** (i.e., higher type I error than expected)
Dunnett	↓
Tukey's honest square difference (Tukey)	↓
Sidak	↓
Bonferroni	**Most Conservative** (i.e., lower type I error than expected)
Scheffe	

Note: Least square difference (LSD) is included in SPSS but does not provide adjustment for multiple comparisons, so not listed here

SPSS: Anatomy of Multiple Comparisons Table

Dependent variable: Admissions

	(I) fullmoon	(J) fullmoon	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	Before	During	-5.70000*	1.01612	.000
		After	-1.55833	1.01612	.289
	During	Before	5.70000*	1.01612	.000
		After	4.14167*	1.01612	.001
	After	Before	1.55833	1.01612	.289
		During	-4.14167*	1.01612	.001
Bonferroni	Before	During	-5.70000*	1.01612	.000
		After	-1.55833	1.01612	.404
	During	Before	5.70000*	1.01612	.000
		After	4.14167*	1.01612	.001
	After	Before	1.55833	1.01612	.404
		During	-4.14167*	1.01612	.001

Groups evaluated in each pairwise comparison (one comparison per row)

Difference in means between groups
Note: mean differences significant at the 0.05 level are denoted with "*"

Standard error for difference in means
Note: SE = 1.01612 for all comparisons is equal because n=12 in each group.

P-value for difference in means

Boston Children's Hospital
Until every child is well

SPSS: Anatomy of Multiple Comparisons Table

Dependent variable: Admissions

	(I) fullmoon	(J) fullmoon	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	Before	During	-5.70000*	1.01612	.000
		After	-1.55833	1.01612	.289
	During	Before	5.70000*	1.01612	.000
		After	4.14167*	1.01612	.001
	After	Before	1.55833	1.01612	.289
		During	-4.14167*	1.01612	.001
Bonferroni	Before	During	-5.70000*	1.01612	.000
		After	-1.55833	1.01612	.404
	During	Before	5.70000*	1.01612	.000
		After	4.14167*	1.01612	.001
	After	Before	1.55833	1.01612	.404
		During	-4.14167*	1.01612	.001

Groups evaluated in each pairwise comparison (one comparison per row)

Difference in means between groups
Note: mean differences significant at the 0.05 level are denoted with "*"

Standard error for difference in means
Note: SE = 1.01612 for all comparisons is equal because n=12 in each group

P-value for difference in means

Boston Children's Hospital
Until every child is well

SPSS: Anatomy of Multiple Comparisons Table

Dependent variable: Admissions

	(I) fullmoon	(J) fullmoon	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	Before	During	-5.70000*	1.01612	.000
		After	-1.55833	1.01612	.289
	During	Before	5.70000*	1.01612	.000
		After	4.14167*	1.01612	.001
	After	Before	1.55833	1.01612	.289
		During	-4.14167*	1.01612	.001
Bonferroni	Before	During	-5.70000*	1.01612	.000
		After	-1.55833	1.01612	.404
	During	Before	5.70000*	1.01612	.000
		After	4.14167*	1.01612	.001
	After	Before	1.55833	1.01612	.404
		During	-4.14167*	1.01612	.001

Groups evaluated in each pairwise comparison (one comparison per row)

Difference in means between groups
Note: mean differences significant at the 0.05 level are denoted with "*"

Standard error for difference in means
Note: SE = 1.01612 for all comparisons is equal because n=12 in each group

P-value for difference in means

Boston Children's Hospital
Until every child is well

SPSS: Anatomy of Multiple Comparisons Table

Dependent variable: Admissions

	(I) fullmoon	(J) fullmoon	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	Before	During	-5.70000*	1.01612	.000
		After	-1.55833	1.01612	.289
	During	Before	5.70000*	1.01612	.000
		After	4.14167*	1.01612	.001
	After	Before	1.55833	1.01612	.289
		During	-4.14167*	1.01612	.001
Bonferroni	Before	During	-5.70000*	1.01612	.000
		After	-1.55833	1.01612	.404
	During	Before	5.70000*	1.01612	.000
		After	4.14167*	1.01612	.001
	After	Before	1.55833	1.01612	.404
		During	-4.14167*	1.01612	.001

Groups evaluated in each pairwise comparison (one comparison per row)

Difference in means between groups
Note: mean differences significant at the 0.05 level are denoted with "*"

Standard error for difference in means
Note: SE = 1.01612 for all comparisons is equal because n=12 in each group.

P-value for difference in means

SPSS: Anatomy of Multiple Comparisons Table

Dependent variable: Admissions

	(I) fullmoon	(J) fullmoon	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	Before	During	-5.70000*	1.01612	.000
		After	-1.55833	1.01612	.289
	During	Before	5.70000*	1.01612	.000
		After	4.14167*	1.01612	.001
	After	Before	1.55833	1.01612	.289
		During	-4.14167*	1.01612	.001
Bonferroni	Before	During	-5.70000*	1.01612	.000
		After	-1.55833	1.01612	.404
	During	Before	5.70000*	1.01612	.000
		After	4.14167*	1.01612	.001
	After	Before	1.55833	1.01612	.404
		During	-4.14167*	1.01612	.001

There is a significant difference in mean admissions between…

- Before vs. during a full moon (diff=5.70, $p<0.0001$)
- During vs. after a full moon (diff =4.14, $p=0.001$)

No difference in mean admissions for Before vs. After (diff=1.56, $p=0.289$)

ANOVA with more than one independent variable

ANOVA with one independent variable: One-way ANOVA

Example: Dependent variable=admissions
Independent variable 1=moon cycle

ANOVA with two independent variables: Two-way ANOVA

Example: Dependent variable=admissions
Independent variable 1=moon cycle
Independent variable 2=Friday (yes/no)

With 2 or more independent variables….use another procedure in SPSS called the ***General Linear Model (GLM)***

General Linear Model (1 Independent Variable)
SPSS: Analyze > General Linear Model > Univariate

GLM Table in SPSS:

Tests of Between-Subjects Effects					
Dependent Variable: Admission					
Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	208.287[a]	2	104.144	16.811	.000
Intercept	4464.467	1	4464.467	720.654	.000
fullmoon	208.287	2	104.144	16.811	.000
Error	204.436	33	6.195		
Total	4877.190	36			
Corrected Total	412.723	35			
a. R Squared = .505 (Adjusted R Squared = .475)					

In GLM, output labeled differently:

Between groups = **fullmoon**

Within groups = **Error**

Total = **Corrected Total**

General Linear Model (2 Independent Variable)
SPSS: Analyze > General Linear Model > Univariate

GLM Table in SPSS:

Tests of Between-Subjects Effects					
Dependent Variable:Admission					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	214.018[a]	5	42.804	6.462	.000
fullmoon	204.162	2	102.081	15.412	.000
Friday	4.371	1	4.371	.660	.423
fullmoon * Friday	1.159	2	.580	.088	.916
Error	198.705	30	6.624		
Corrected Total	412.723	35			
a. R Squared = .519 (Adjusted R Squared = .438)					

- Add independent variable Friday as a ”main effect” into the model.
- Add interaction between fullmoon and Friday into the model (Does the relationship between fullmoon and admissions depend on Friday?)

Total variation (SST) is the same as with 1 independent variable

No interaction between fullmoon and Friday (p=0.916) and no effect of Friday (p=0.423)

Boston Children's Hospital
Until every child is well

Questions?

Correlation

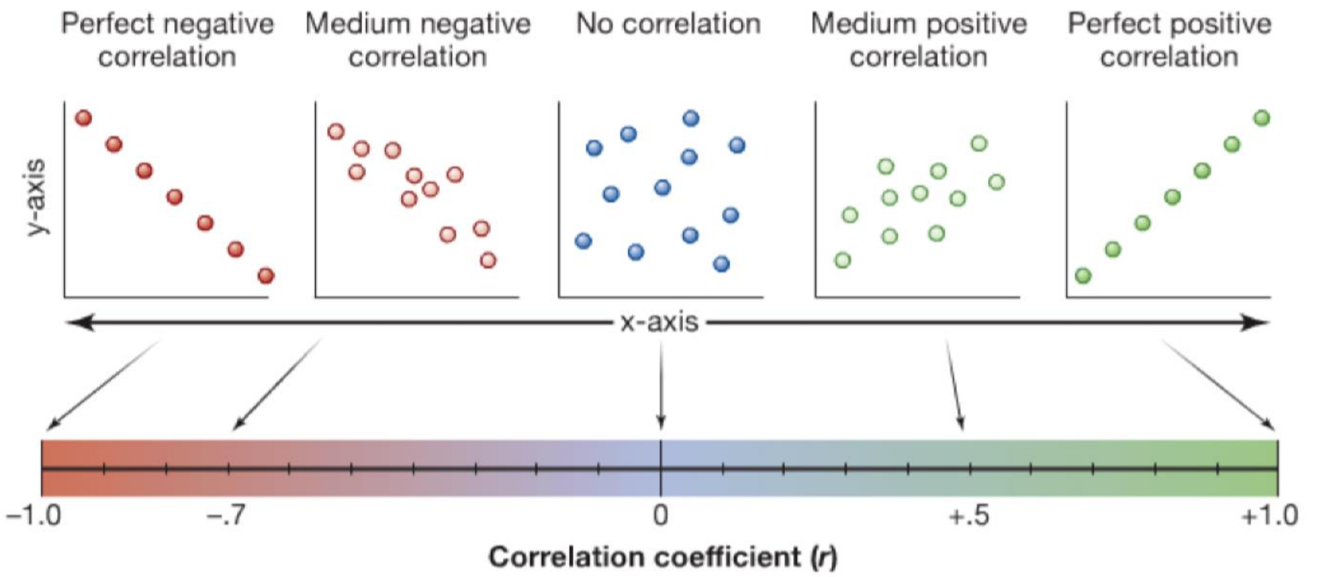
So far we have assumed that the independent variable is categorical (2 or more groups)… What if the independent variable is continuous?

The correlation coefficient (ρ) measures the strength of association between two variables

- **Pearson's correlation coefficient "r"** is the most commonly used correlation coefficient
- Quantifies the linear relationship between two continuous variables

Pearson's Correlation Coefficient "r"

Correlation coefficient ranges from -1 to 1 and shows **magnitude (strong, medium, weak)** and **direction (positive, negative)** of association



Pearson's Correlation Coefficient "r"

Person's correlation coefficient is calculated as the **covariance of the two variables** (a measure of how the variables change together) divided by the **product of their standard deviations**:

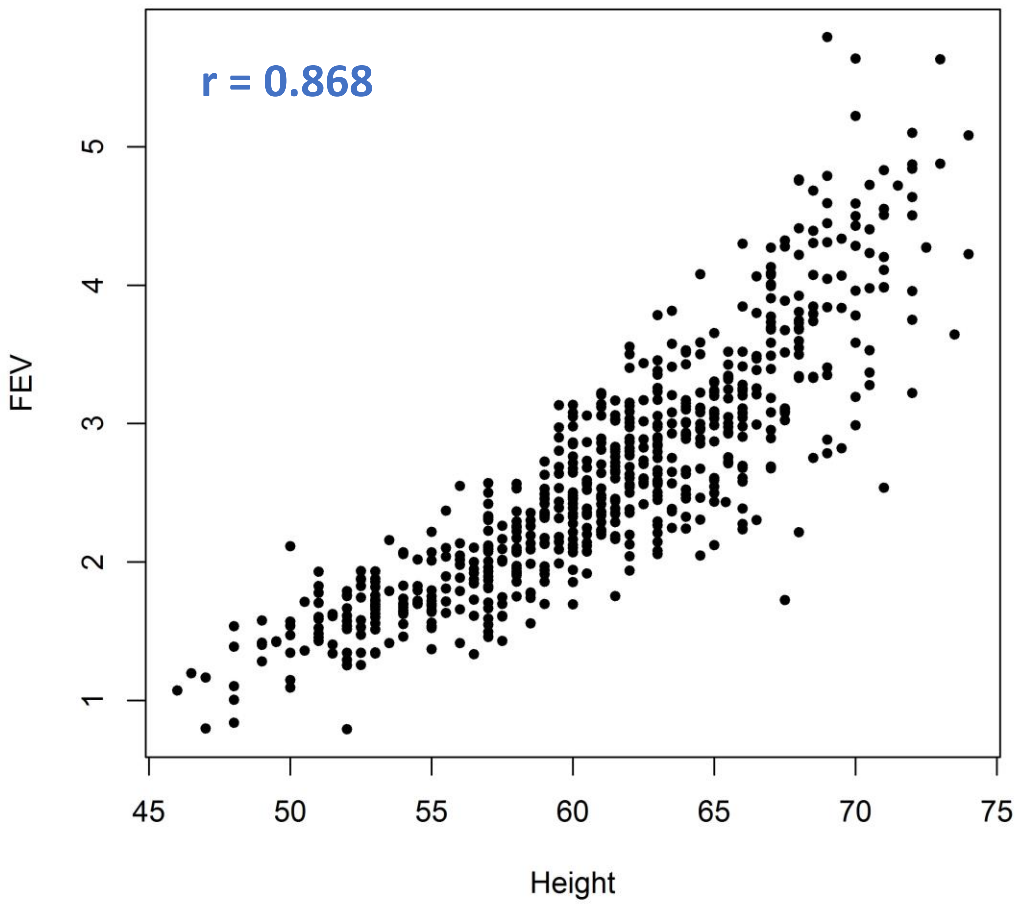
$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Example: FEV & Height

- **Sample**: 654 children ages 3 to 19 who were seen in the Childhood Respiratory Disease Study in East Boston
- **Objective**: Evaluate the linear relationship between FEV and height

Boston Children's Hospital
Until every child is well

Example: FEV & Height



Step 1: Hypotheses

$\mathbf{H_0}: \rho = 0$ vs. $\mathbf{H_1}: \rho \neq 0$

Step 2: Test Statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

r = sample correlation

Step 3: Compare test statistic to a t-distribution with n-2 degrees of freedom

Pearson Correlation

SPSS: Analyze > Correlate > Bivariate > Coefficients = Pearson

Correlations

		FEV	Hgt
FEV	Pearson Correlation	1	.868**
	Sig. (2-tailed)		.000
	N	654	654
Hgt	Pearson Correlation	.868**	1
	Sig. (2-tailed)	.000	
	N	654	654

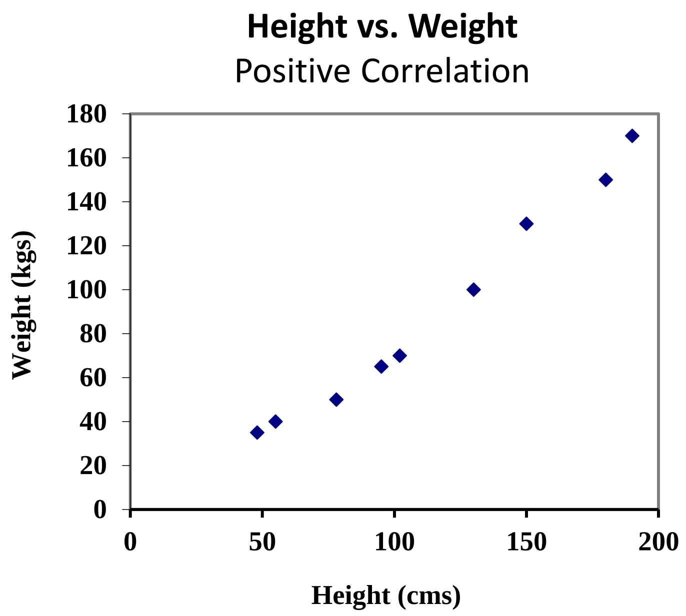
**. Correlation is significant at the 0.01 level (2-tailed).

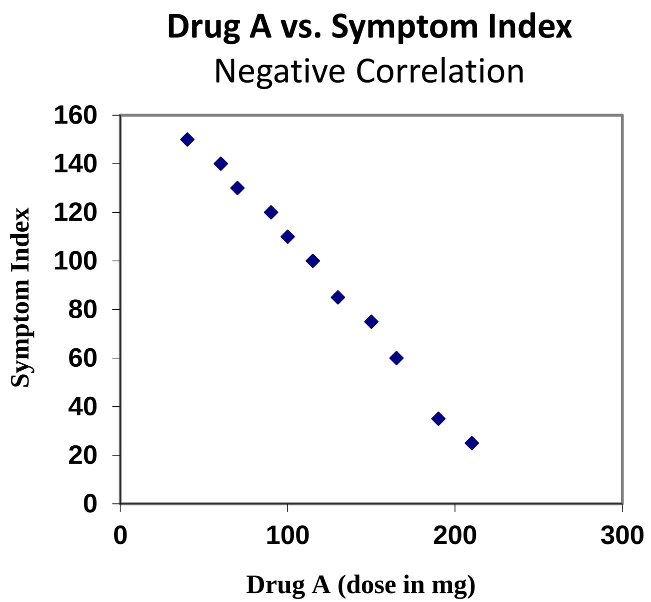
Pearson Correlation "r"

P-Value

Conclusion: There is a strong, positive correlation between FEV and height (Pearson Correlation r = 0.868; p<0.0001)

Why do we need linear regression?

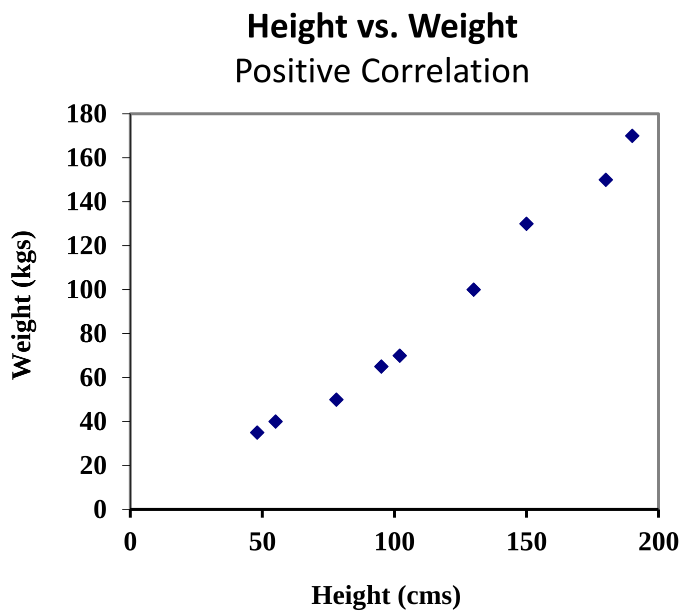


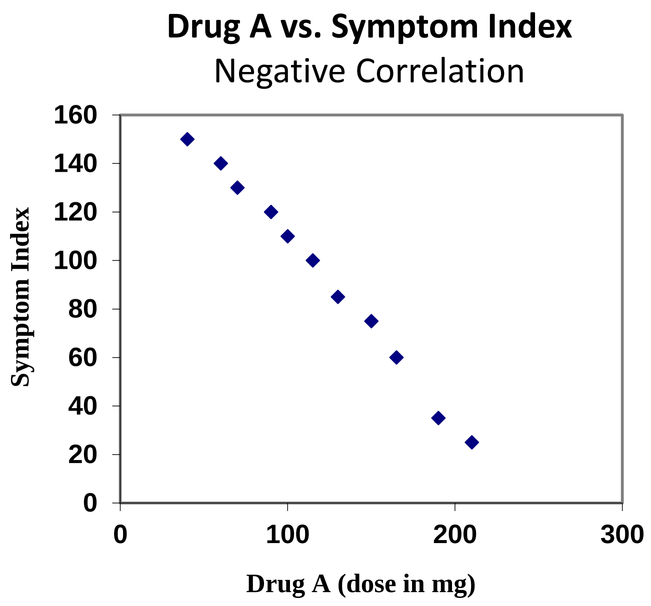


Correlation

- Useful measure to summarize the relationship (magnitude & direction) between two variables
- Describes the extent to which two variables move together
  - Weight increases with height
  - Symptoms decrease with drug A dose

Why do we need linear regression?



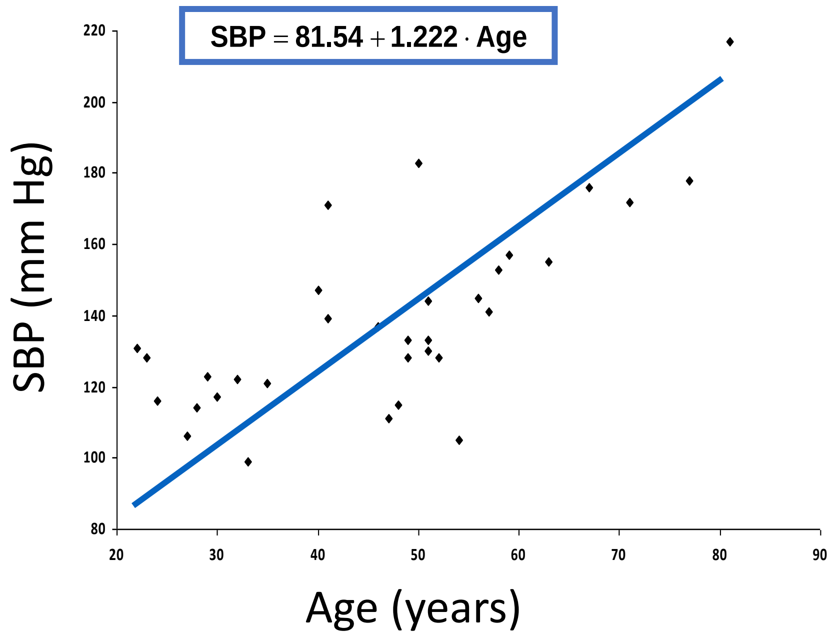


Linear Regression

- Provides additional information on the magnitude of the relationship
- Measures the impact of 1 unit change in independent variable on dependent variable
  - A 1 unit increase in height results in a 5 unit increase in weight
  - A 1 unit decrease in dose results in a 5 unit decrease in symptom index

Boston Children's Hospital
Until every child is well

How does linear regression work?
Example: Age vs. Systolic Blood Pressure (SBP)



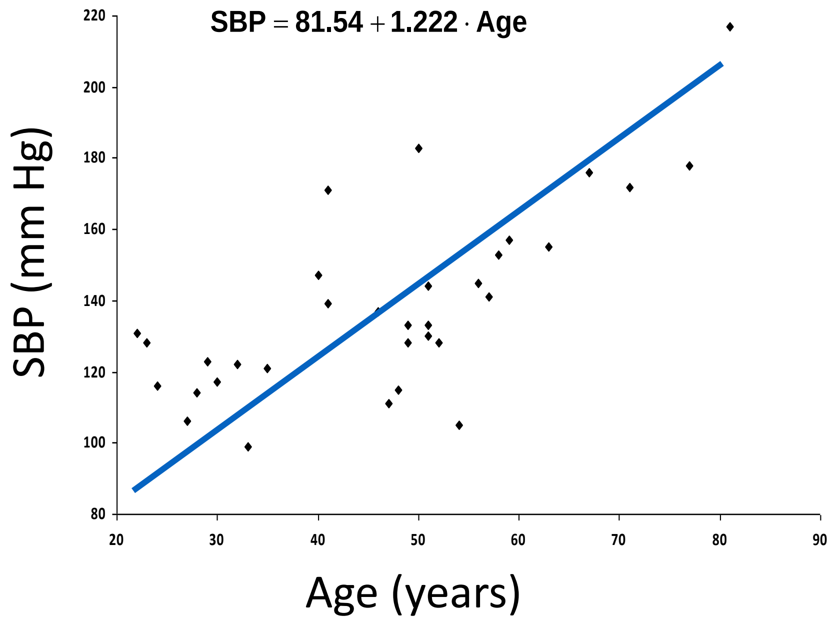
Adapted from Colton T. Statistics in Medicine. Boston: Little Brown, 1974

Diamonds represent the individual observations (N=33)

Equation for the blue "line of best fit" outputs the predicted SBP value

- **Intercept**: SBP value if age = 0 is 81.54
  - Denoted $\alpha = 81.54$
- **Slope**: Average change in SBP per 1 year change in age is 1.222
  - Denoted $\beta_1 = 1.222$

How does linear regression work?
Example: Age vs. Systolic Blood Pressure (SBP)



Adapted from Colton T. Statistics in Medicine. Boston: Little Brown, 1974

The vertical deviation from each diamond to the line represents the difference in the observed and predicted SBP values

- The sum of these squared deviations measures "goodness of fit" (how well the line fits the data)
- The smaller the deviation, the closer the points are to the predicted line

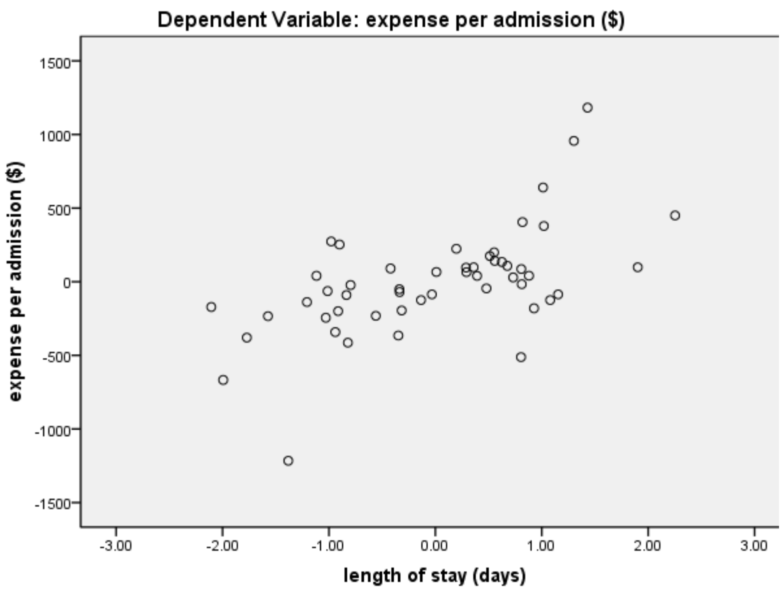
Our goal is to find the α and β_1 that give the minimum value for the sum of squared deviations (smallest error)

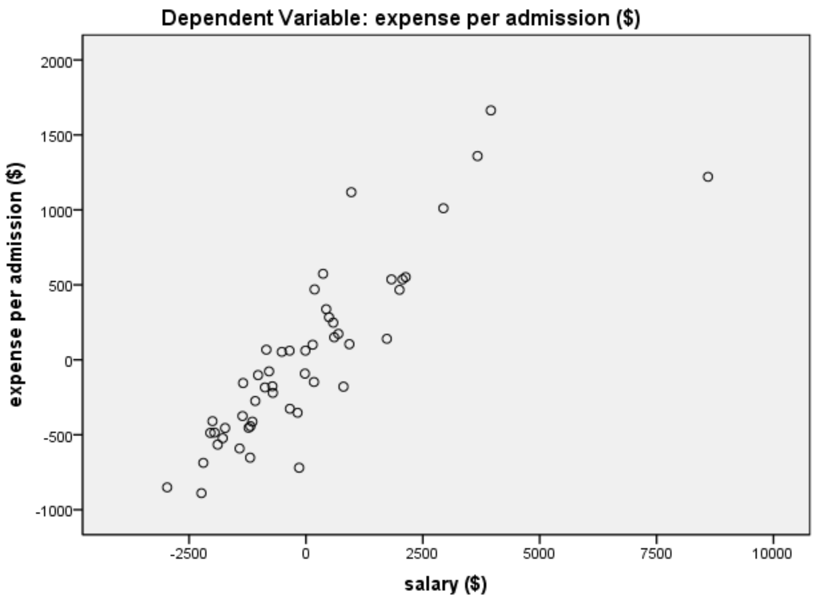
- Called least squares method

Effect of age on SPB addressed by testing whether slope (β_1) is different from zero using a t-test.

Boston Children's Hospital
Until every child is well

Example: Predicting hospital expenses from length of stay and salary level





Similar to ANOVA, with more than 1 variable:

- First, determine whether both variables explain the relationship
- Second, determine variables that are important to the outcome

Example: Predicting hospital expenses from length of stay and salary level

First, let's look at correlation:
Both length of stay and salary are significantly correlated with hospital expenses.

SPSS: Analyze > Correlate > Bivariate > Coefficients = Pearson

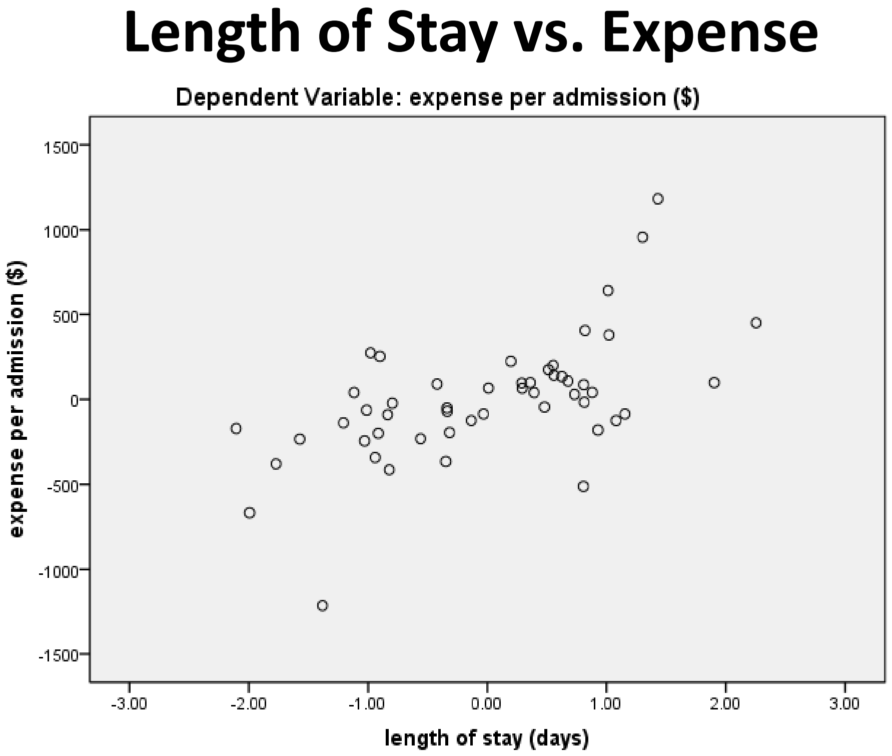
Correlations

		expense per admission ($)	length of stay (days)	salary ($)
length of stay (days)	Pearson Correlation	.322*	1	-.046
	Sig. (2-tailed)	.021		.748
	N	51	51	51
salary ($)	Pearson Correlation	.794**	-.046	1
	Sig. (2-tailed)	.000	.748	
	N	51	51	51

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

Boston Children's Hospital
Until every child is well

Example: Predicting hospital expenses from length of stay and salary level



Next, let's fit the regression model including only length of stay:

SPSS: Analyze > Regression > Linear

Output Tables:

- Variables entered/removed
- Model summary
- ANOVA
- Coefficients
- Residual statistics

Example: Predicting hospital expenses from length of stay and salary level

SPSS: Analyze > Regression > Linear

Model Summary[b]

Model	R	R-Square	Adjusted R-Square	SE of the Estimate
1	0.322[a]	0.104	0.085	577.589

a. Predictors: (Constant), length of stay (days)
b. Dependent variable: expense per admission ($)

Simple Linear Regression:

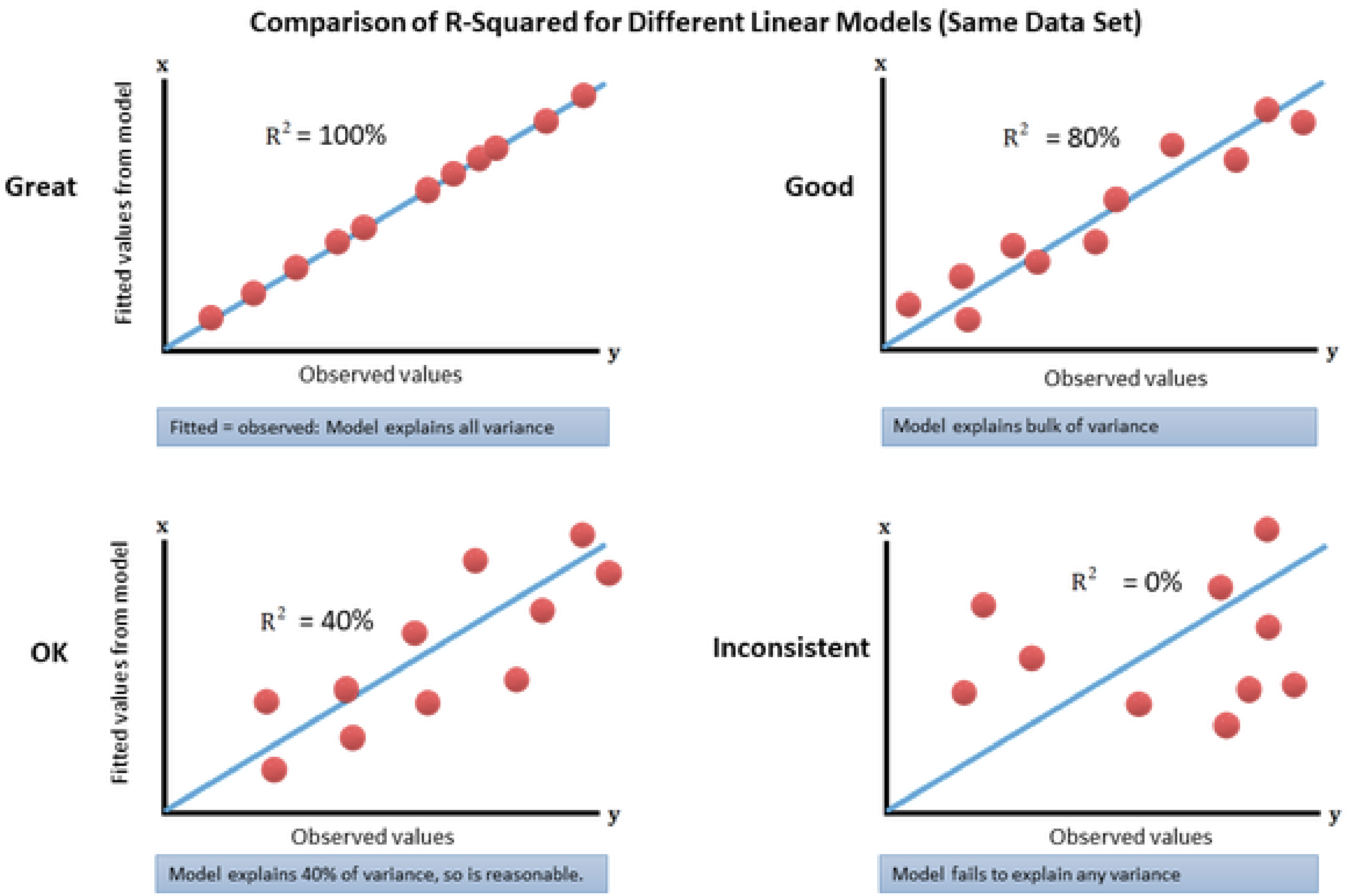
R = Pearson correlation of length of stay and expense = 0.322

R^2 (R-square) = $(0.322)^2$ = 0.104

R^2 → "Goodness of Fit" Measure

- **R^2 reflects how well your data fits a regression line**
  - Formally defined as the proportion of the variance for a dependent variable (i.e., hospital expense) that is explained by the independent variables (i.e., LOS) in a regression model
  - The better the model fits the data (i.e., the closer observations are to the best-fit line), the smaller the variance and the higher the R^2
- **Ranges between 0 (0%) and 1 (100%)**
- **Often expressed as percentage, rather than decimal**

R^2 → "Goodness of Fit" Measure



Boston Children's Hospital
Until every child is well

Example: Predicting hospital expenses from length of stay and salary level

SPSS: Analyze > Regression > Linear

ANOVA Table[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	0.18908	1	0.18908	5.668	.021[a]
	Residual	1.635	49	0.03336		
	Total	1.824	50			

a. Predictors: (Constant), length of stay (days)

b. Dependent Variable: expense per admission ($)

Indicates that the predictors in the model (in this case, LOS) significantly explain the variation in the data (p=0.021).

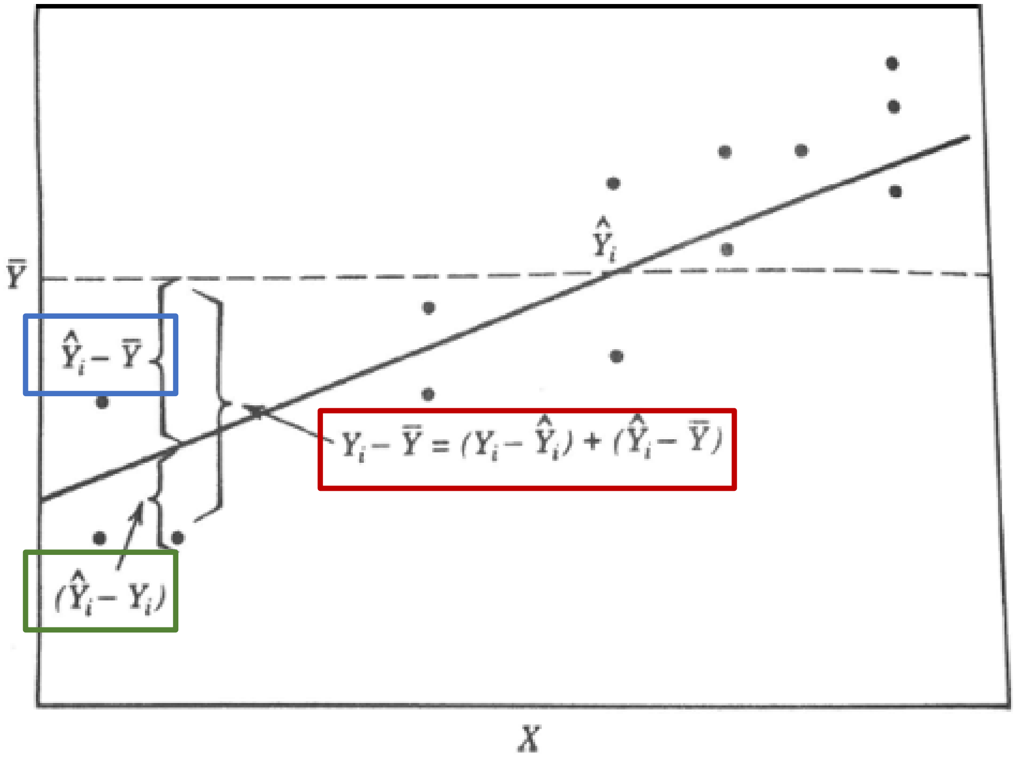
Regression vs. Residual Sum of Squares

In regression analysis, there are three main types of sum of squares:

Total sum of squares
- Reflects total variation in the sample

Regression sum of squares ($\widehat{Y_i} - \bar{Y}$)
- Reflects how well a regression model represents the modeled data
- Higher regression sum of squares (i.e., larger difference between predicted and mean values) indicates that the model does not fit the data well

Residual sum of squares ($\widehat{Y_i} - Y_i$)
- Reflects variation in the dependent variable that cannot be explained by the model (measuring error)
- Higher residual sum of squares (i.e, larger difference between predicted and observed values) indicates that the model poorly explains the data



$\widehat{Y}_i$ The predicted value estimated by the regression line
$\bar{Y}$ The mean value of the sample
Y_i The observed value

Boston Children's Hospital
Until every child is well

Example: Predicting hospital expenses from length of stay and salary level

SPSS: Analyze > Regression > Linear

Coefficients Table[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1281.959	608.104		2.108	.040
	length of stay (days)	191.563	80.465	.322	2.381	.021

a. Dependent Variable: expense per admission ($)

T-Test Statistic

= Unstandardized B / SE

= 191.563 / 80.465 = **2.381**

Indicates that slope for length of stay is significantly different from 0 (p=0.021).

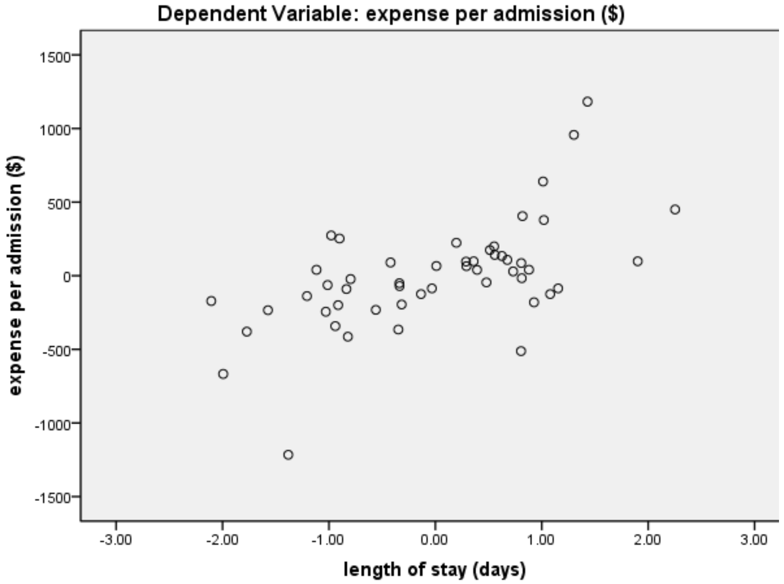
For every 1-day increase in length of stay, hospital expenses increase by $191.56.

Note: With 1 independent variable F test (ANOVA Table) and t-test results (Coefficient Table) are equal. Specifically, F test statistic = 5.688 is equal to the square of the t-test statistic = $t^2 = (2.381)^2$

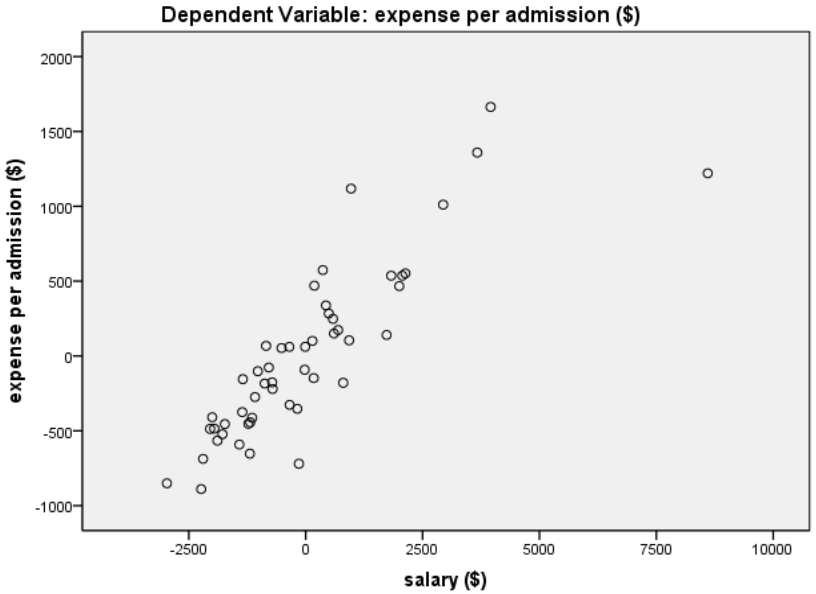
Boston Children's Hospital
Until every child is well

Example: Predicting hospital expenses from length of stay and salary level

Length of Stay vs. Expense



Salary vs. Expense



Example: Predicting hospital expenses from length of stay and salary level

SPSS: Analyze > Regression > Linear

ANOVA Table[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.384	2	.69205	75.554	.000[a]
	Residual	.4396	48	.00916		
	Total	1.824	50			

a. Predictors: (Constant), salary ($), length of stay (days)
b. Dependent Variable: expense per admission ($)

Indicates that the predictors in the model significantly explain the variation in the data ($p<0.0001$)

Example: Predicting hospital expenses from length of stay and salary level

SPSS: Analyze > Regression > Linear

Coefficients Table[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-2582.736	464.770		-5.557	.000
	length of stay (days)	213.797	42.208	.359	5.065	.000
	salary ($)	.249	.022	.810	11.422	.000

a. Dependent Variable: expense per admission ($)

Indicates that slope for both **length of stay** and **salary** are significantly different from zero ($p<0.0001$)

Example: Predicting hospital expenses from length of stay and salary level

Prediction Model:

Expense = $\alpha + \beta_1$*(length of stay) + β_2*(salary)

Expense = -2582.736 + 213.797*(length of stay) + 0.249*(salary)

Note:

- *Least squares method is used to estimate* α, β_1, β_2
- *With 2 or more independent variables, coefficients* (β_1, β_2) *are called partial regression coefficients*

Example: Predicting hospital expenses from length of stay and salary level

Prediction Model:

Expense = α + β_1*(length of stay) + β_2*(salary)

Expense = -2582.736 + 213.797*(length of stay) + 0.249*(salary)

Interpretation:

β_1 is the amount expense changes on average with 1 unit increase in length of stay at a fixed value of salary (i.e., controlling for salary)

β_2 is the amount expense changes on average with 1 unit increase in salary at a fixed value of length of stay (i.e., controlling for length of stay)

Boston Children's Hospital
Until every child is well™

Example: Predicting hospital expenses from length of stay and salary level

SPSS: Analyze > Regression > Linear

Model Summary[b]

Model	R	R-Square	Adjusted R-Square	SE of the Estimate
1	0.871[a]	0.759	0.749	302.649

a. Predictors: (Constant), salary ($), length of stay (days)
b. Dependent variable: expense per admission ($)

Multiple Linear Regression:

R^2 (R-square) = (Sum of squares regression) / (Sum of squares total)

= 1.384 / 1.824 = 0.759

What is the adjusted R-square?

Adjusted R^2

- Takes into account number of predictors in model
- Define: N=number of observations

  p=number of predictors
- Calculate as:

$$R^2_{adj} = 1 - (1-R^2)\,(N-1) \,/\, (N-p)$$

$$= 1 - (1-0.759)*(50) \,/\, (50-2)$$

$$= 0.749$$

- R^2_{adj} will always be smaller than R^2

Interpretation

"Length of stay and salary significantly explain the variation in hospital expenses (F-test statistic = 75.55, p<0.0001). The estimated coefficient for length of stay was positive indicating that expenses increase by approximately $214 for an additional day (SE = 42). The estimated coefficient for salary was also positive indicating that expenses increase by approximately $0.25 for every $1 increase in salary (SE=0.022). The adjusted R-square for this model is 0.749."

Model Diagnostics

- **Residual Plots in SPSS**
  - Check to ensure normally distributed
  - Independent of one another
  - Similar in terms of variance

- **Unusual Observations: Need to determine reason for them and have a strong justification for exclusion**
  - **Outliers** (extreme residuals) → Data points that diverge from the overall pattern
  - **Influential Observations** (extreme predicted values) → Influence the slope of regression line

Boston Children's Hospital
Until every child is well™

Linear Regression Summary:

- **F-test** → used to determine overall significance of relationship
- **Coefficients, SE and 95% CI** → used to describe effect of each independent variable on outcome
- **R^2 and R^2_{adj}** → provide estimate of strength of relationship
- **Model Diagnostics** → check model assumptions and identify outliers that could bias the estimates (residual plots, etc.)

Best Model?

- If models have the **same** number of independent variables…
  - Choose model with highest value of R^2_{adj}
  - This gives 'maximum value' per independent variable
  - This model will also have the highest value of R^2 and F
- If models have a **different** number of independent variables…
  - Highest value of R^2_{adj} (more independent variables)
  - Highest value of F (fewer independent variables)
- **Clinical Relevance!**

Next Class

- We interpret regression coefficients for continuous predictors as slopes… what about **categorical predictors**?
- We've spent the last two classes discussing methods for continuous outcomes… what about **categorical outcomes**?

Questions?

kimberly.greco@childrens.harvard.edu