NEXT STEPS WITH DATA

ST101 – DR. ARIC LABARR

ANALYSIS OF VARIANCE

NEXT STEPS WITH DATA

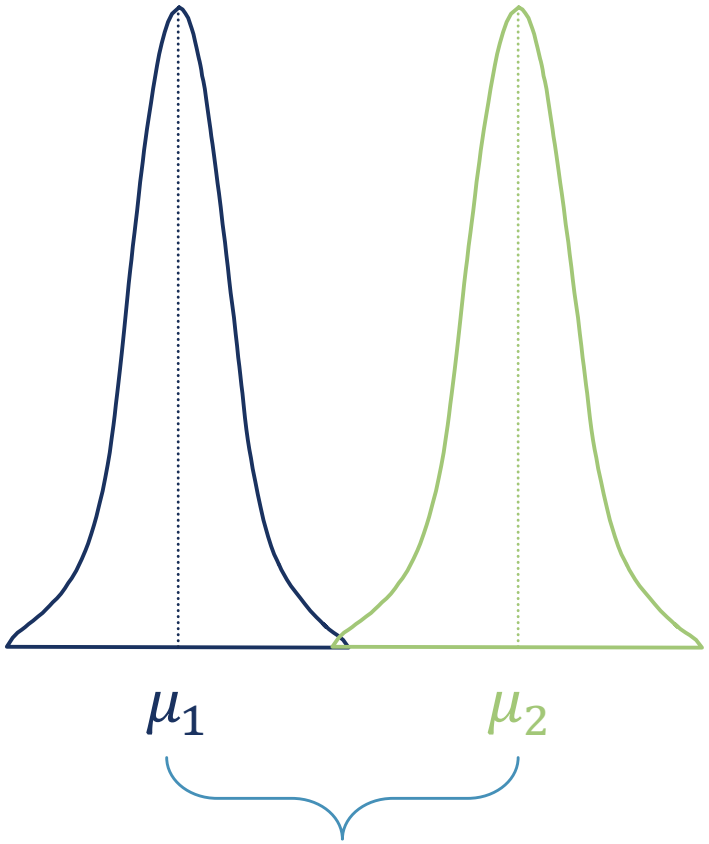
COMPARING TWO OR MORE AVERAGES

- We have studied hypothesis tests and confidence intervals that have focused on one population parameter – for example, one average compared to a number.
- However, sometimes we like to compare multiple parameters against each other, like comparing two or more averages.

COMPARING TWO OR MORE AVERAGES

- When comparing averages between two groups of data, we must think about how spread out the data is.
- When comparing many averages *statistically* we call this an **Analysis of Variance (ANOVA)**.
  - Need to account for the spread in the data when comparing means!

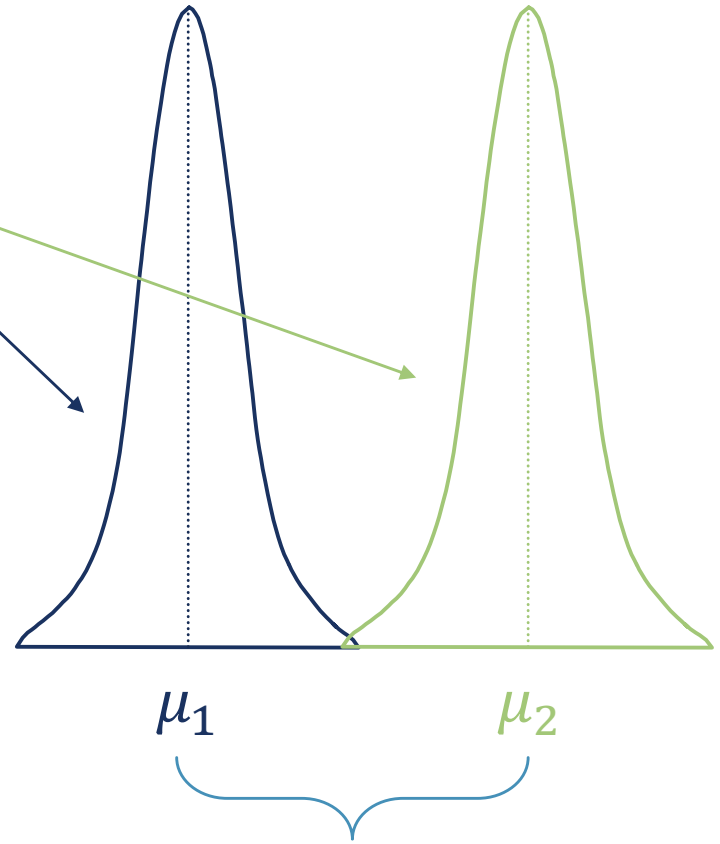
COMPARING TWO AVERAGES



How close are these values?

COMPARING TWO AVERAGES

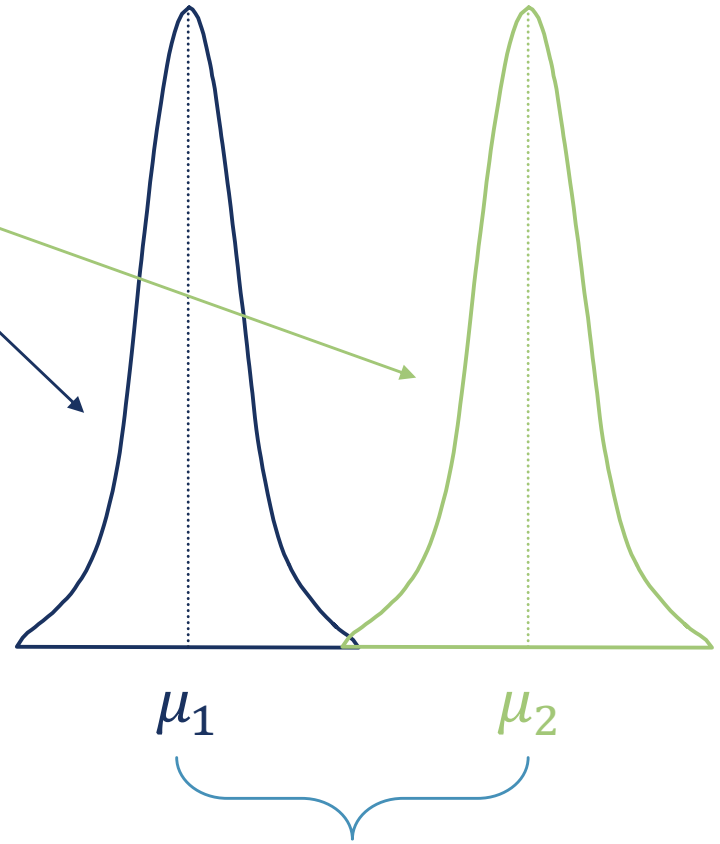
Spread of the two distributions is not overlapping by much.



How close are these values?

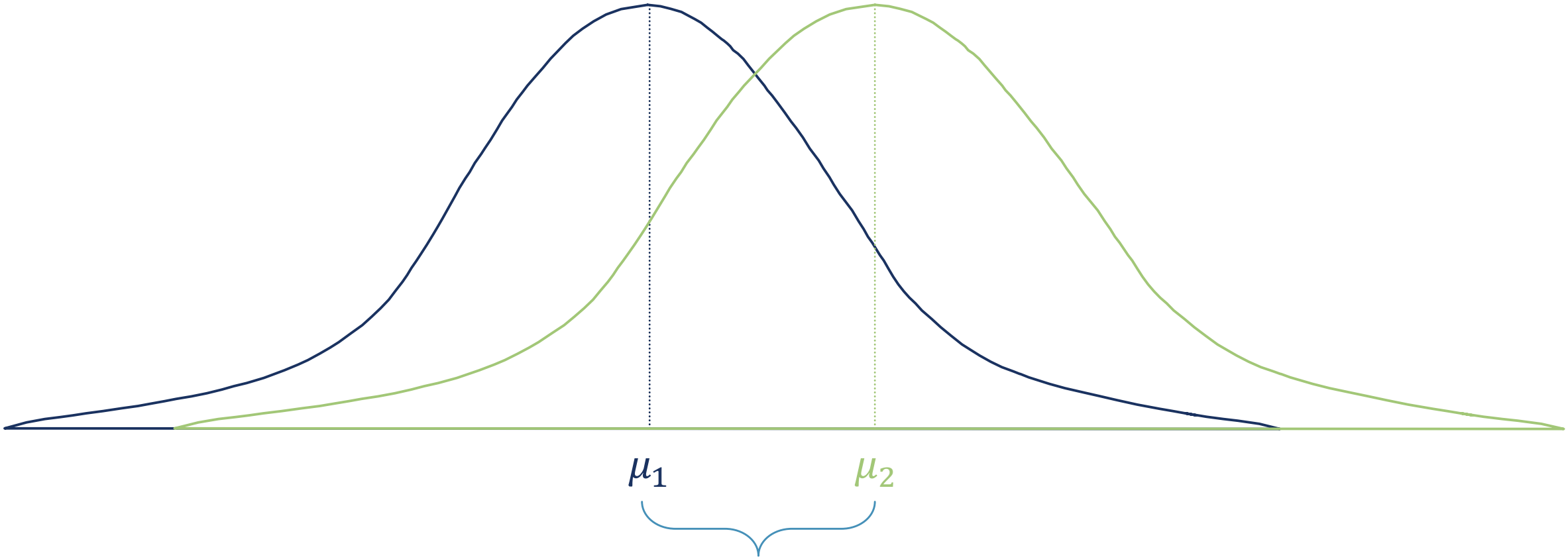
COMPARING TWO AVERAGES

Spread of the two distributions is not overlapping by much.

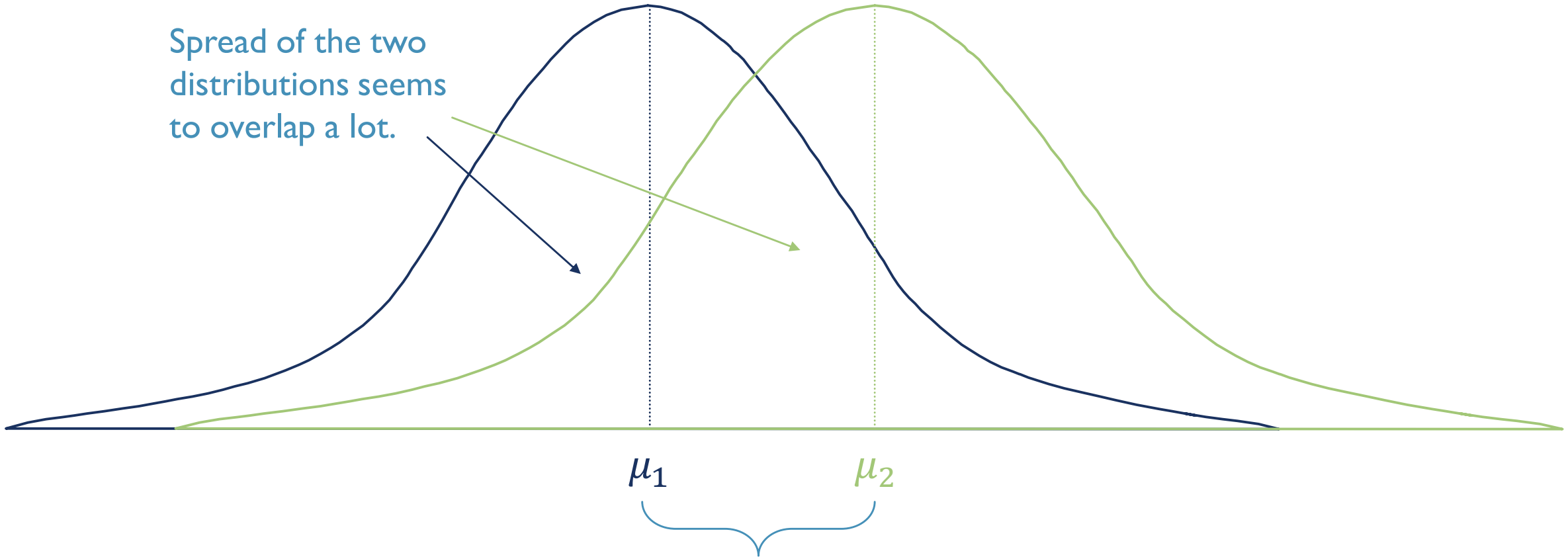


Don't appear to be too close!

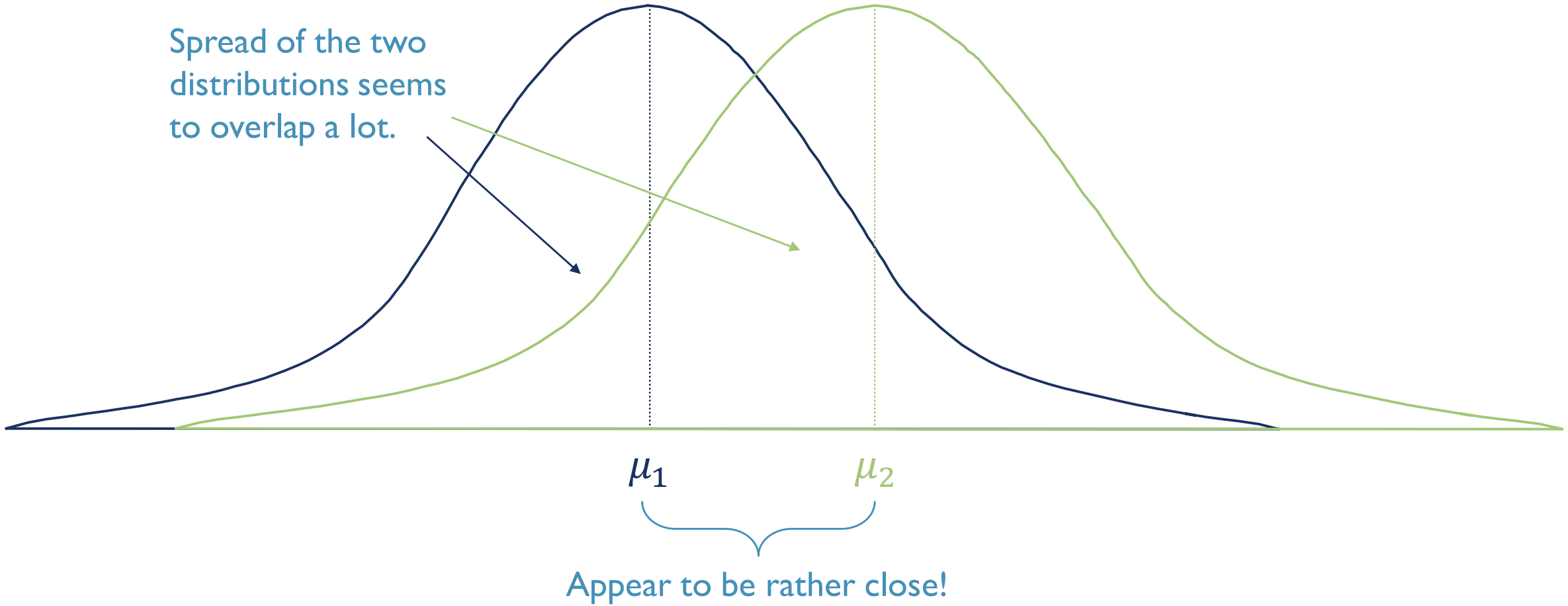
COMPARING TWO AVERAGES



COMPARING TWO AVERAGES



COMPARING TWO AVERAGES

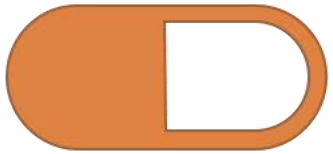


COMMON QUESTIONS ANOVA CAN HELP WITH

- Do accountants, on average, make more than teachers?

COMMON QUESTIONS ANOVA CAN HELP WITH

- Do people treated with one of the two new drugs have higher average T-cell counts than people in the control group?



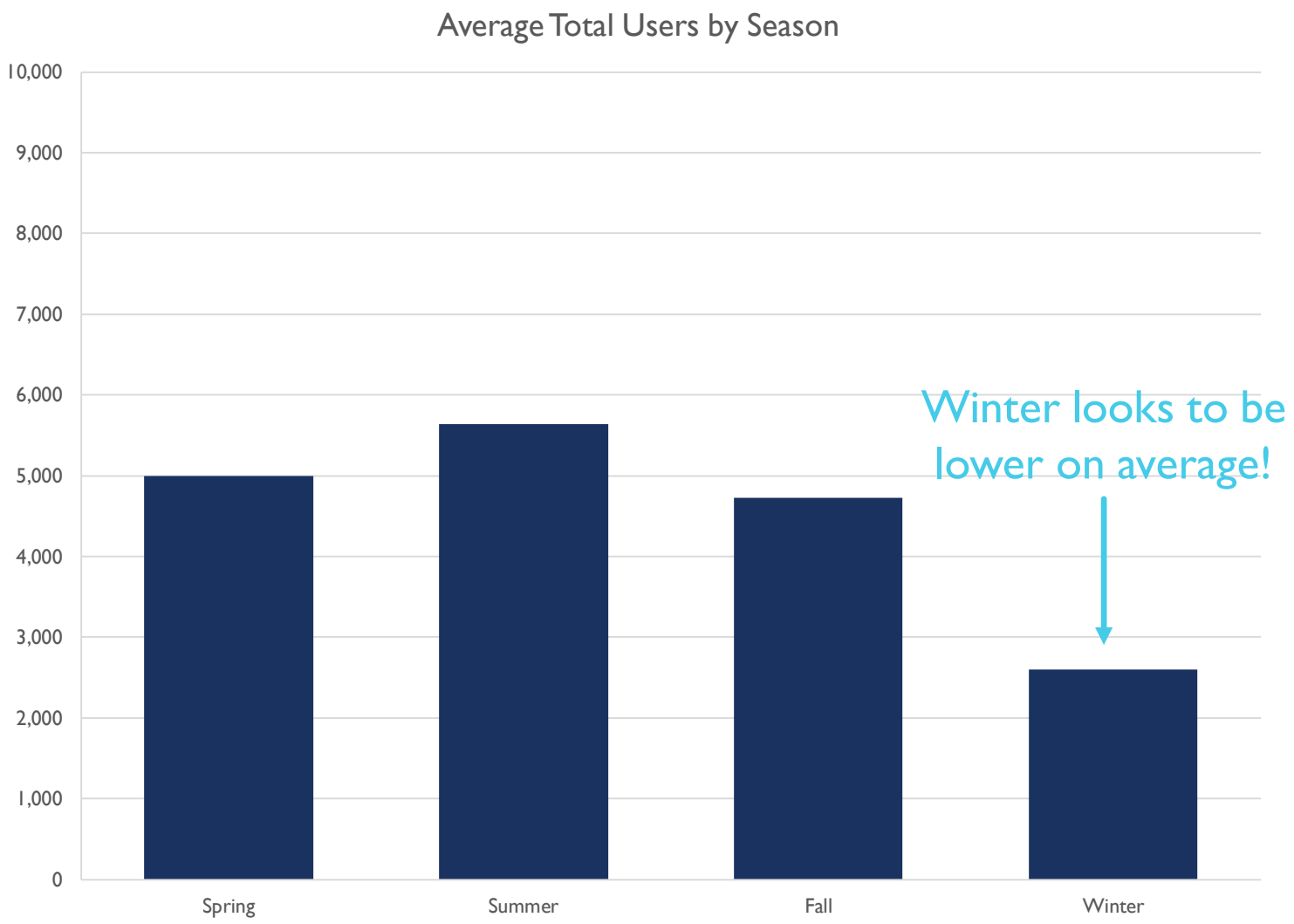
Treatment A

Treatment B

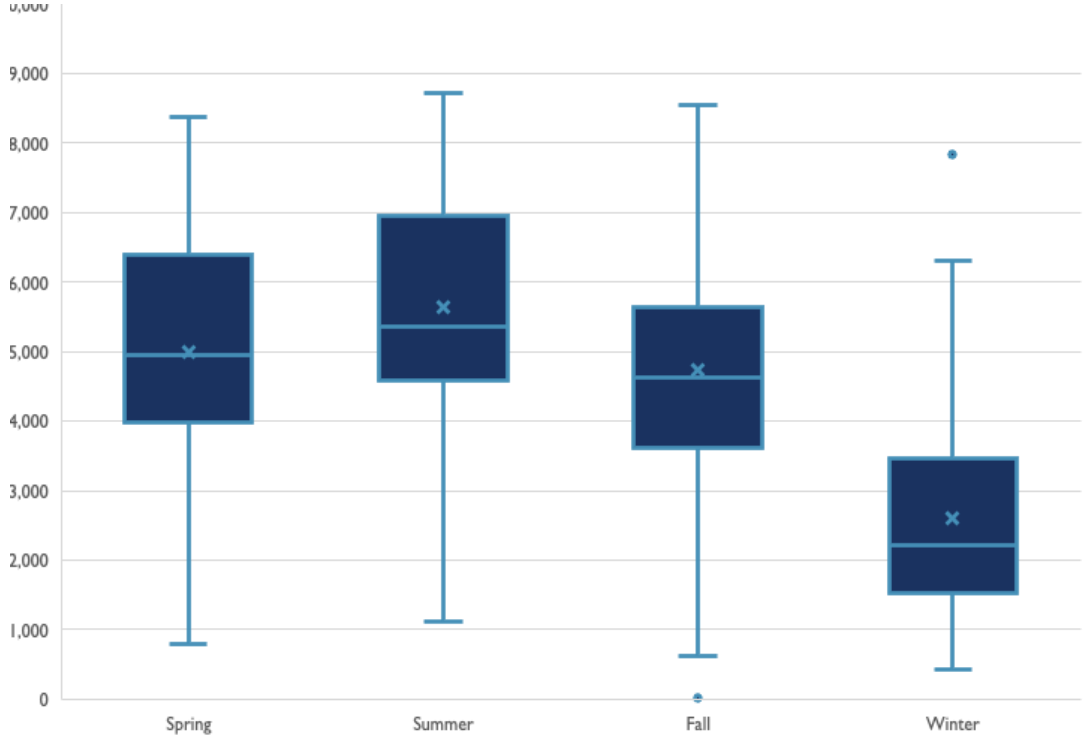
Placebo

COMMON QUESTIONS ANOVA CAN HELP WITH

- Do people spend different amounts depending on which type of credit card they have?



DOES WINTER HAVE LOWER AVERAGE TOTAL USERS?



DOES WINTER HAVE LOWER AVERAGE TOTAL USERS?

Look at how spread out the values of users in winter are!

ONE-WAY ANOVA

- **One-Way ANOVA** – design in which independent samples are obtained from 2 or more categories of a single explanatory variable, then **testing** whether these categories **have equal means**.
- For example:
  - Variable of interest is total users – looking at the average total users.
  - Explanatory variable is season – looking to see if average total users changes across season.
  - Number of categories is 4 – looking to see if the average total users changes across the 4 seasons.

ONE-WAY ANOVA HYPOTHESES

- **One-Way ANOVA** – design in which independent samples are obtained from k categories of a single explanatory variable, then **testing** whether these k categories **have equal means**.
- Null Hypothesis:

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

- Alternative Hypothesis:

$$H_a: \text{At least one mean different than another}$$

ONE-WAY ANOVA HYPOTHESES

- **One-Way ANOVA** – design in which independent samples are obtained from k categories of a single explanatory variable, then **testing** whether these k categories **have equal means**.
- Assumptions:
  1. Groups are Normally distributed
  2. Groups have equal variance / spread
  3. Independence of observations

ONE-WAY ANOVA HYPOTHESES

- **One-Way ANOVA** – design in which independent samples are obtained from k categories of a single explanatory variable, then **testing** whether these k categories **have equal means**.
- Assumptions:
  1. Groups are Normally distributed – total users across each season is Normally distributed
  2. Groups have equal variance / spread – total users across each season have equal variance
  3. Independence of observations – total users from each day don't depend on each other

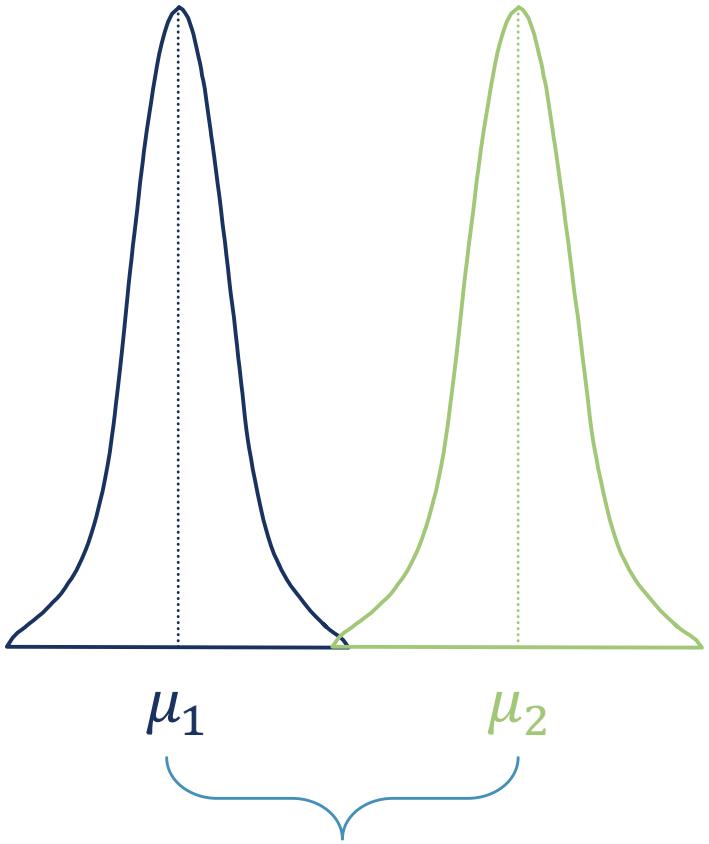
SOURCES OF VARIATION

- Variation in an ANOVA can come from two places – within a category and between different categories.
  - **Between-Sample Variability** – variability in the variable of interest that exists between categories of an explanatory variable.
  - **Within-Sample Variability** – variability in the variability of interest that exists within a category of an explanatory variable.

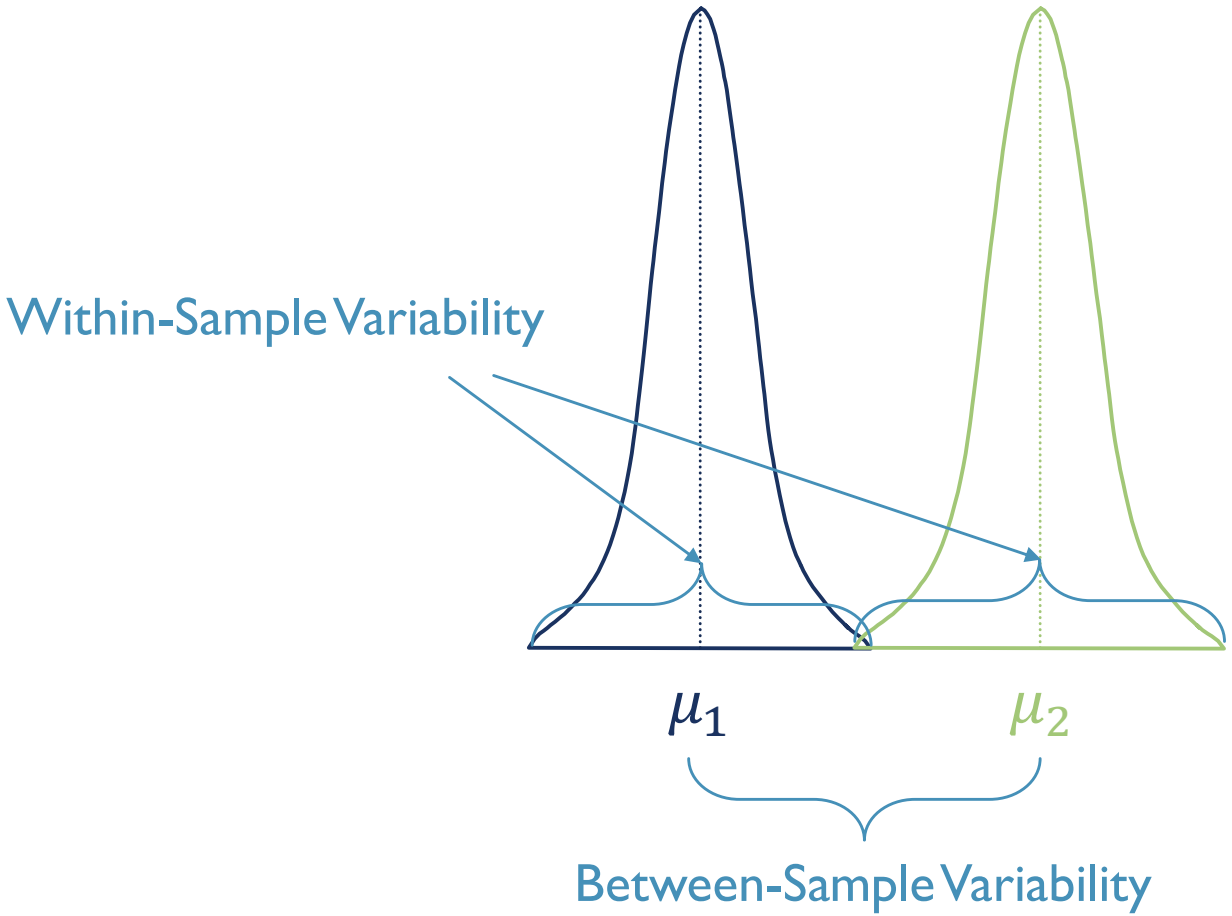
SOURCES OF VARIATION

- Variation in an ANOVA can come from two places – within a category and between different categories.
  - **Between-Sample Variability** – variability in the variable of interest that exists between categories of an explanatory variable. WHAT CATEGORIES CAN EXPLAIN
  - **Within-Sample Variability** – variability in the variability of interest that exists within a category of an explanatory variable. WHAT CATEGORIES CANNOT EXPLAIN

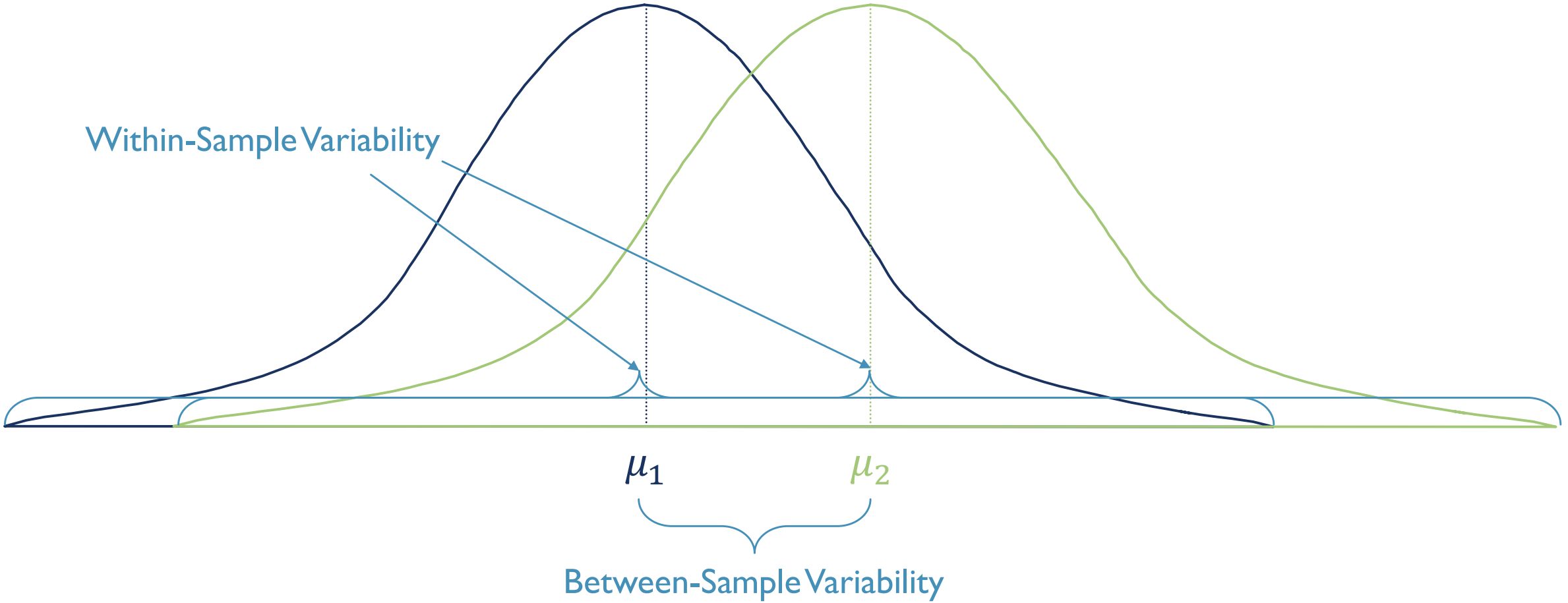
COMPARING TWO AVERAGES



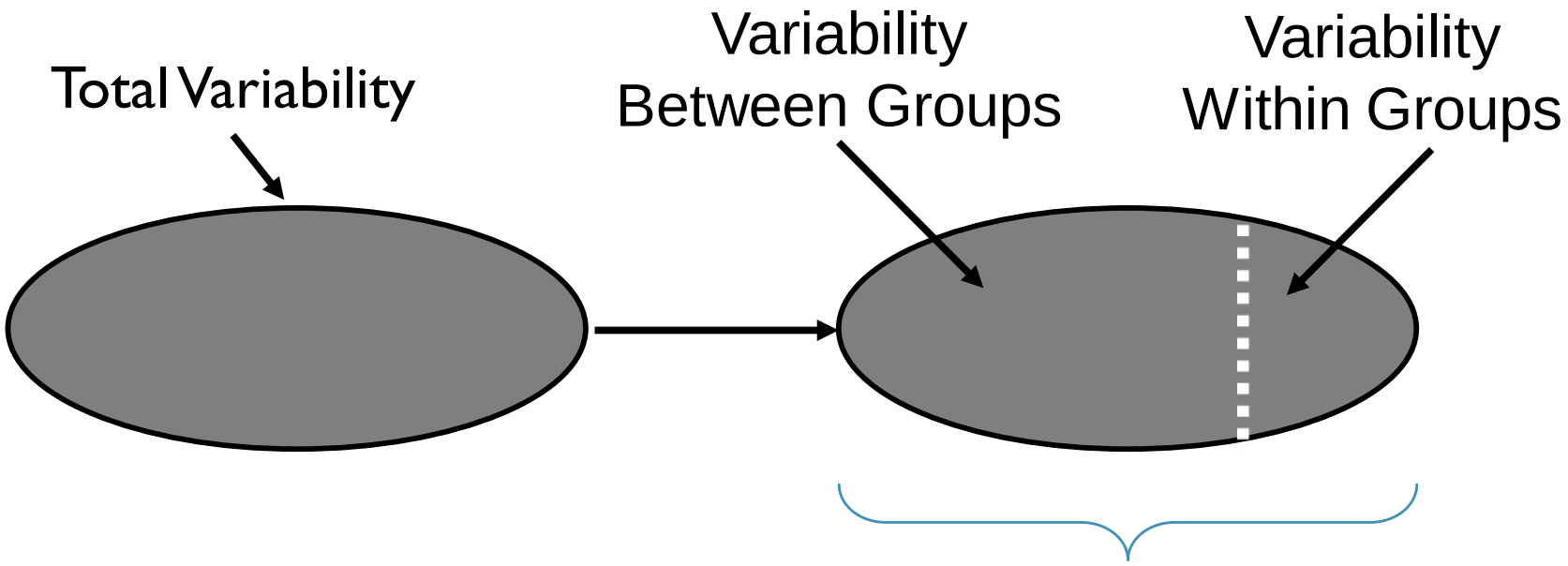
COMPARING TWO AVERAGES



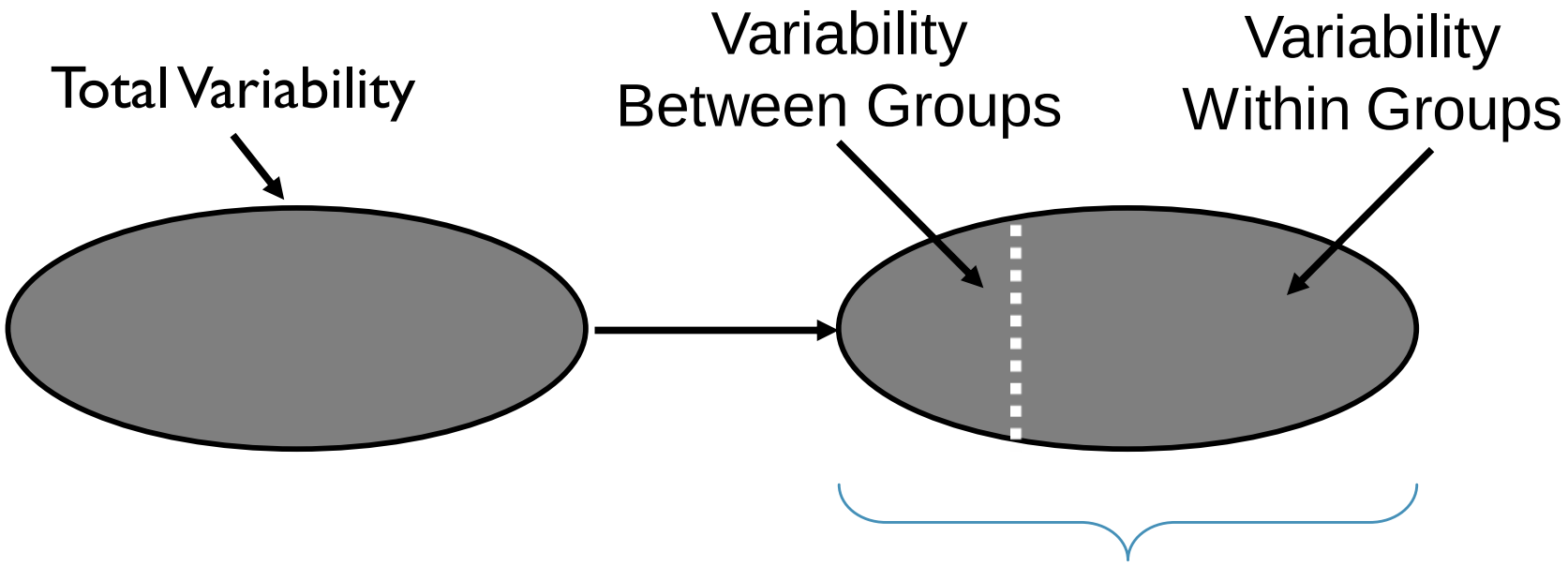
COMPARING TWO AVERAGES



SPLITTING VARIABILITY IN ANOVA

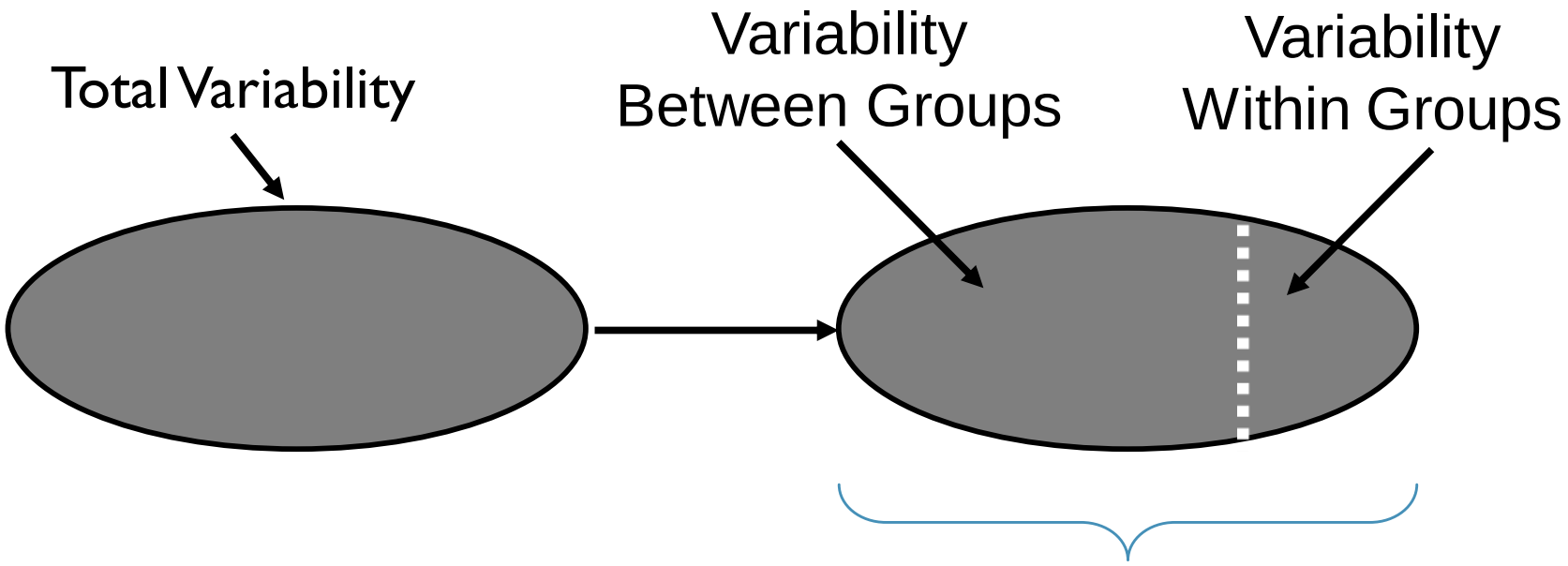


SPLITTING VARIABILITY IN ANOVA



If within group/sample variability is much bigger than the categories' averages aren't that different.

SPLITTING VARIABILITY IN ANOVA



If between group/sample variability is much bigger than the categories' averages are different.

BIKE DATA EXAMPLE

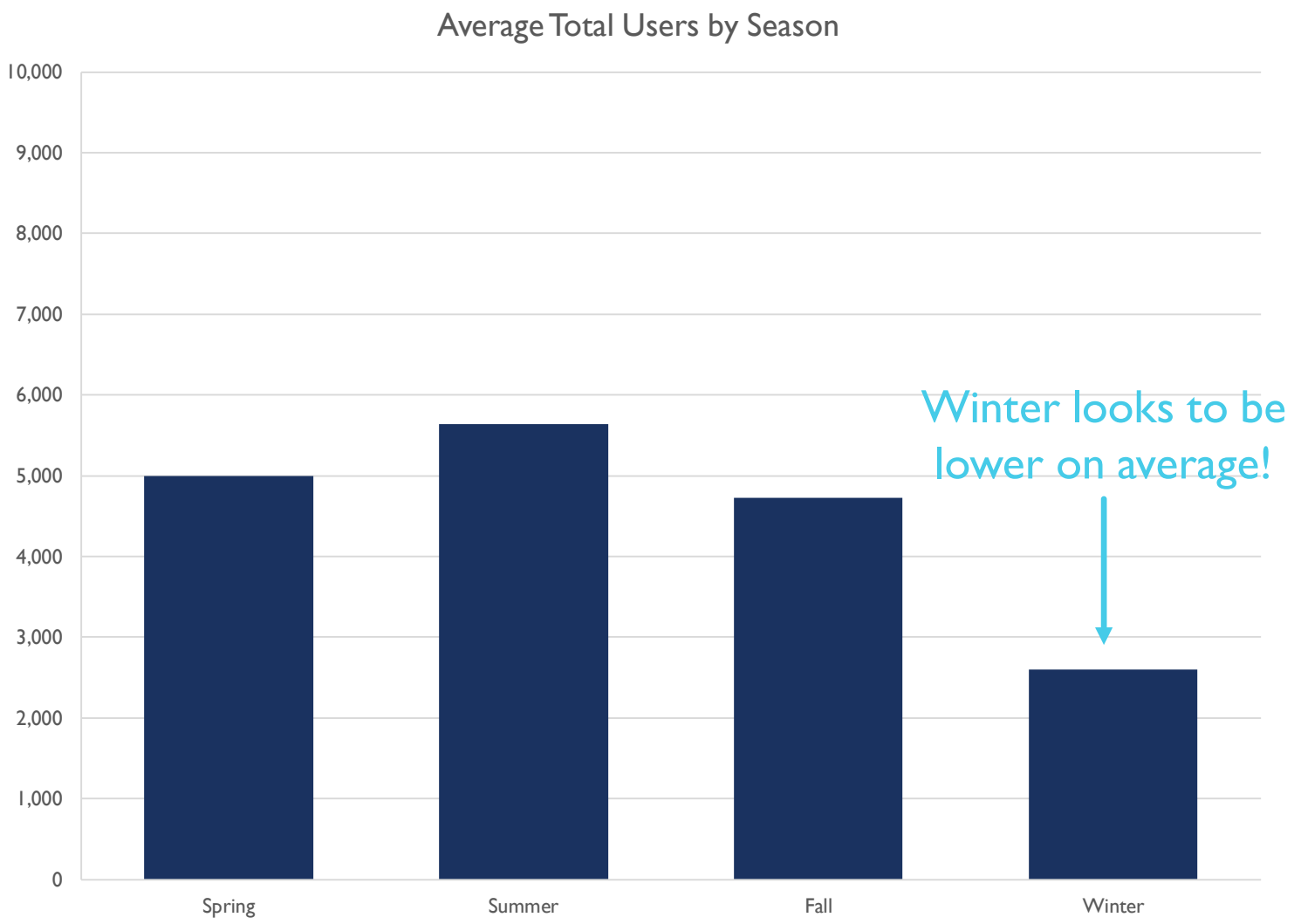
- Null Hypothesis:
  - Average total users is the same across seasons
  - $H_0: \mu_{spring} = \mu_{summer} = \mu_{fall} = \mu_{winter}$
- Test Statistic:
  - $F = 128.8$
- P-value:
  - $P(F \geq 144) < 0.0001$
- Decision:
  - REJECT NULL HYPOTHESIS → At least one of the seasons has a different average total number of users

SUMMARY

- One-Way ANOVA – design in which independent samples are obtained from 2 or more categories of a single explanatory variable, then testing whether these categories have equal means.
- Variation in an ANOVA can come from two places – within a category and between different categories.
  - Between-Sample Variability – variability in the variable of interest that exists between categories of an explanatory variable.
  - Within-Sample Variability – variability in the variability of interest that exists within a category of an explanatory variable.

MULTIPLE COMPARISONS

NEXT STEPS WITH DATA



DOES WINTER HAVE LOWER AVERAGE TOTAL USERS?

NEXT STEPS AFTER ANOVA

- If you reject the null hypothesis on the F-test what does that mean?

NEXT STEPS AFTER ANOVA

- If you reject the null hypothesis on the F-test what does that mean? **Evidence shows at least one category is different**.

- But which category?!?!?!?!

NEXT STEPS AFTER ANOVA

- If you reject the null hypothesis on the F-test what does that mean? **Evidence shows at least one category is different**.

- Once a difference is detected, must test each individual pair of categories to find where all the differences are – a process called **multiple comparisons** or **ad-hoc testing**.

MULTIPLE COMPARISONS PROBLEM

- You have a coin which lands on heads 50% of the time when flipped.
- What is the probability of flipping a head on your first flip?
- What is the probability of flipping a head on your second flip?
- What is the probability of flipping *at least* one head in two flips?

MULTIPLE COMPARISONS PROBLEM

- You have a coin which lands on heads 50% of the time when flipped.
- What is the probability of flipping a head on your first flip?
  - 50%
- What is the probability of flipping a head on your second flip?
  - 50%
- What is the probability of flipping *at least* one head in two flips?
  - 75%

MULTIPLE COMPARISONS PROBLEM

- You have a test which makes an error 5% of the time when performed.
- What is the probability of making an error on your first test?
- What is the probability of making an error on your second test?
- What is the probability of making *at least* one error in two tests?

MULTIPLE COMPARISONS PROBLEM

- You have a test which makes an error 5% of the time when performed.
- What is the probability of making an error on your first test?
  - 5%
- What is the probability of making an error on your second test?
  - 5%
- What is the probability of making *at least* one error in two tests?
  - 9.75%

MULTIPLE COMPARISONS PROBLEM

- You have a test which makes an error 5% of the time when performed.
- What is the probability of making an error on your first test?
  - 5%

Comparison-wise Error

- What is the probability of making an error on your second test?
  - 5%
- What is the probability of making *at least* one error in two tests?
  - 9.75%

MULTIPLE COMPARISONS PROBLEM

- You have a test which makes an error 5% of the time when performed.
- What is the probability of making an error on your first test?
  - 5%
- What is the probability of making an error on your second test?
  - 5%
- What is the probability of making *at least* one error in two tests?
  - 9.75%

Experiment-wise Error

TWO DIFFERENT TYPES OF ERROR

- **Comparison-wise error rate** is the error rate for each individual test or comparison.
- **Experiment-wise error rate** is the error rate across all comparisons – proportion of experiments/comparisons in which at least one error occurs.

- Tests and confidence intervals typically control for comparison-wise error rates, α, but ideally we want to control for experiment-wise error.

MULTIPLE COMPARISONS METHODS

Number of Groups Compared	Number of Comparisons	Experimentwise Error Rate (α=0.05)
2	1	.05
3	3	.14
4	6	.26
5	10	.40

Comparison-wise Error:
$\alpha = 0.05$

Experiement-wise Error:
$1 - (1 - \alpha)^{\#\ comparisons}$

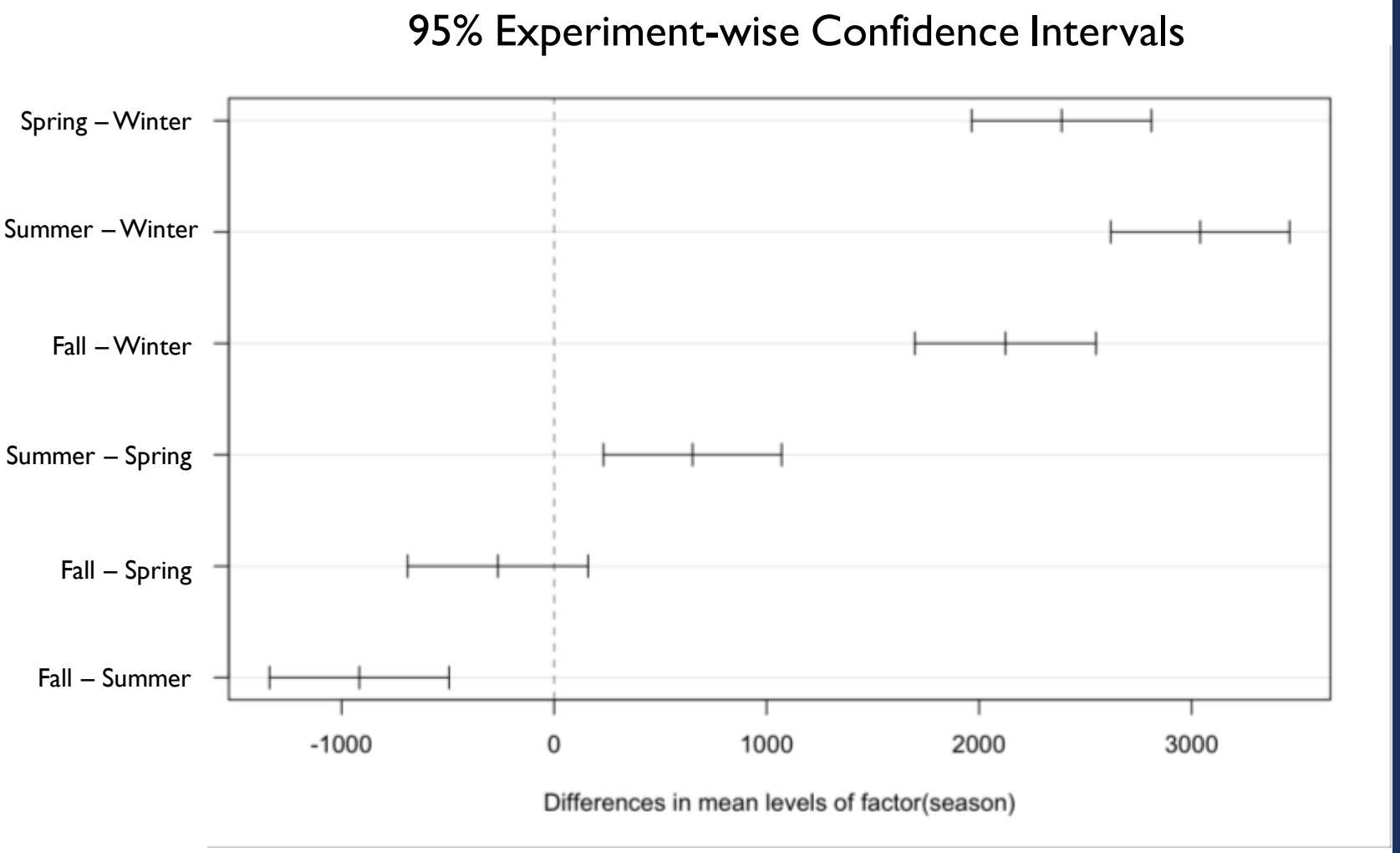
TUKEY-KRAMER TESTS

- **Tukey-Kramer tests** *statistically* compare averages between two groups while controlling for all pairwise comparisons that will be considered.
- Tukey-Kramer tests control for experiment-wise error.
  - Experiment-wise error rate equal to α when **all** pairwise comparisons are considered.
  - Can use hypothesis tests or confidence intervals.

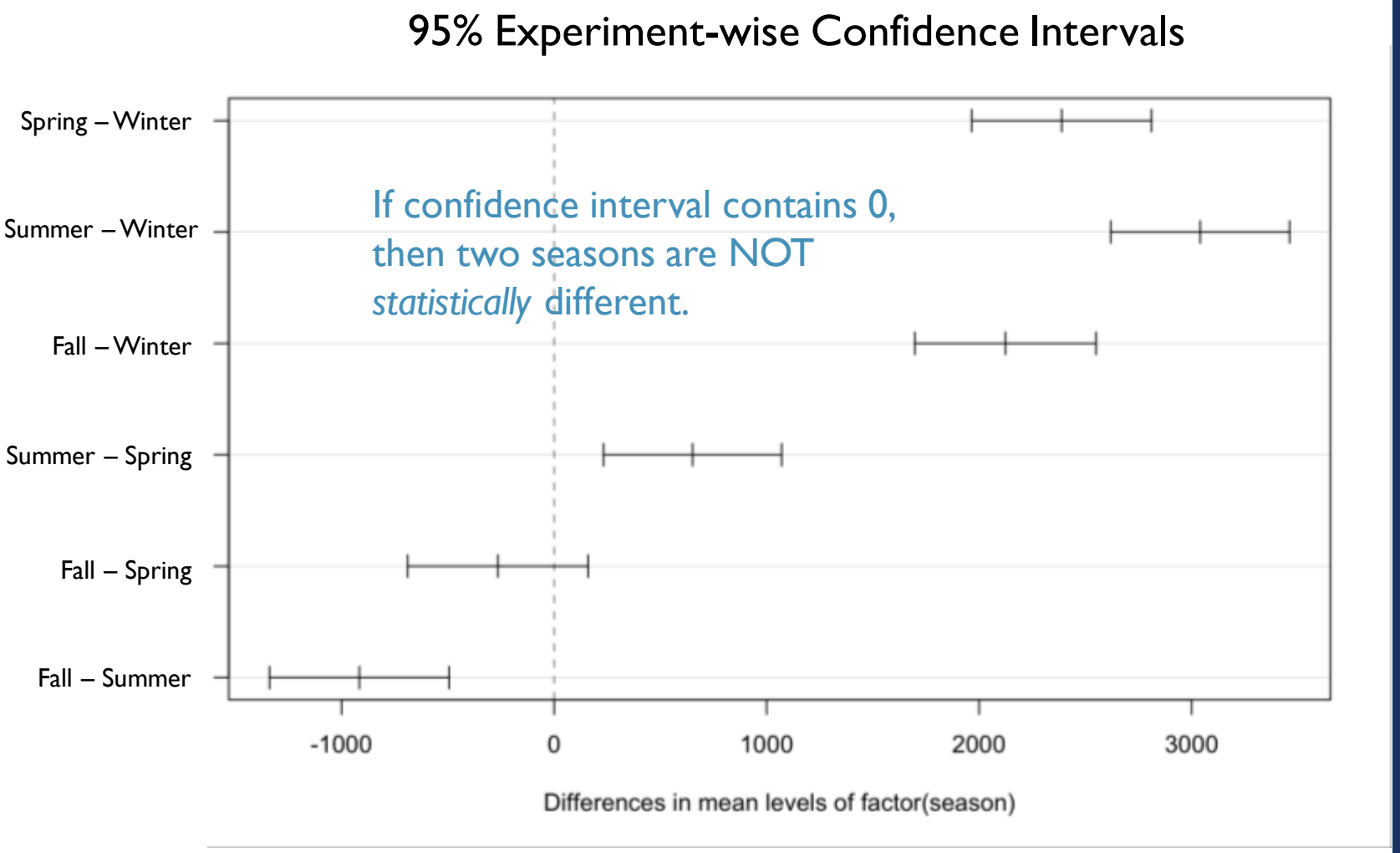
BIKE DATA EXAMPLE – ANOVA

- Null Hypothesis:
  - Average total users is the same across seasons
  - $H_0: \mu_{spring} = \mu_{summer} = \mu_{fall} = \mu_{winter}$
- Test Statistic:
  - $F = 128.8$
- P-value:
  - $P(F \geq 144) < 0.0001$
- Decision:
  - REJECT NULL HYPOTHESIS → At least one of the seasons has a different average total number of users

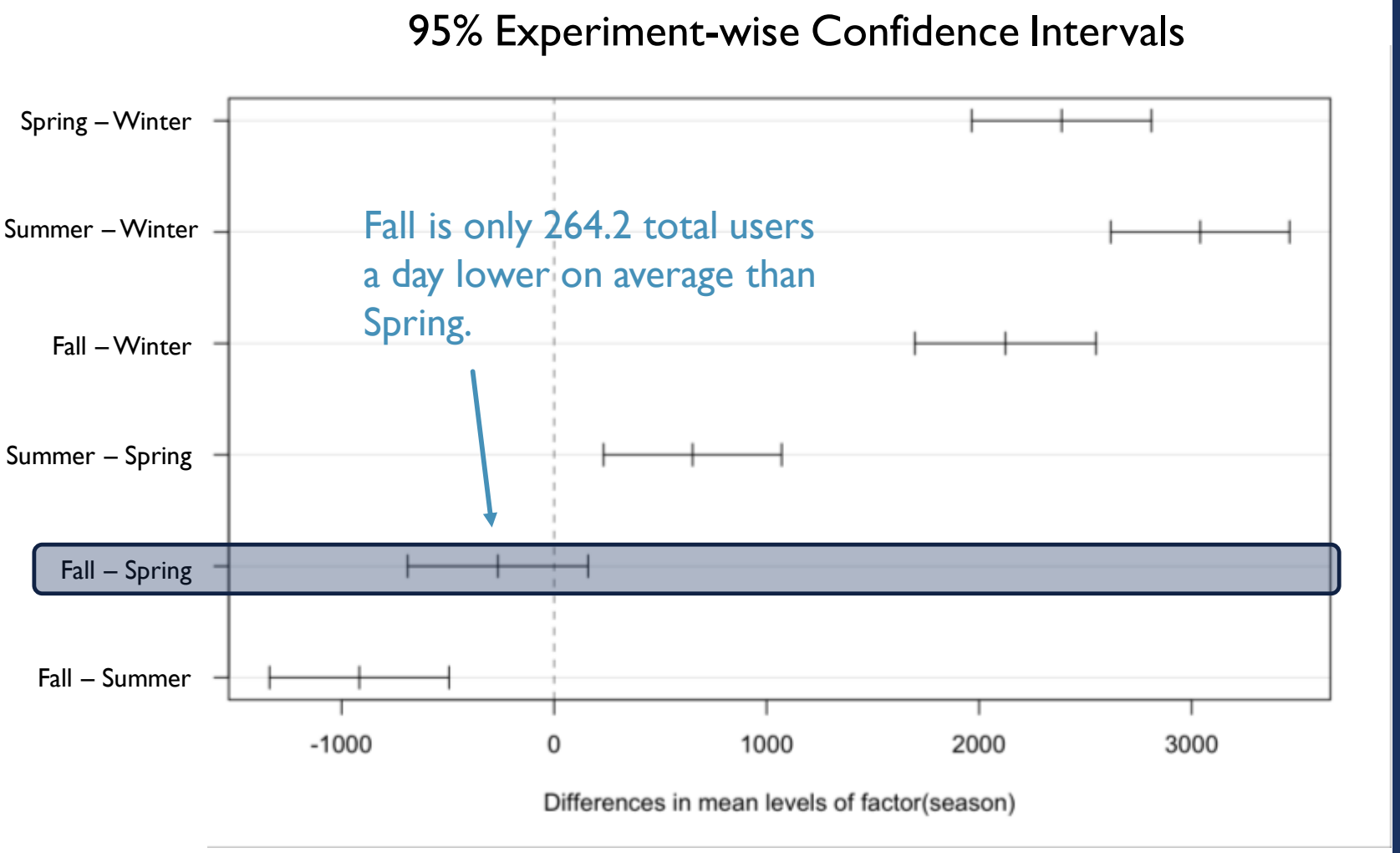
BIKE DATA EXAMPLE – MULTIPLE COMPARISONS



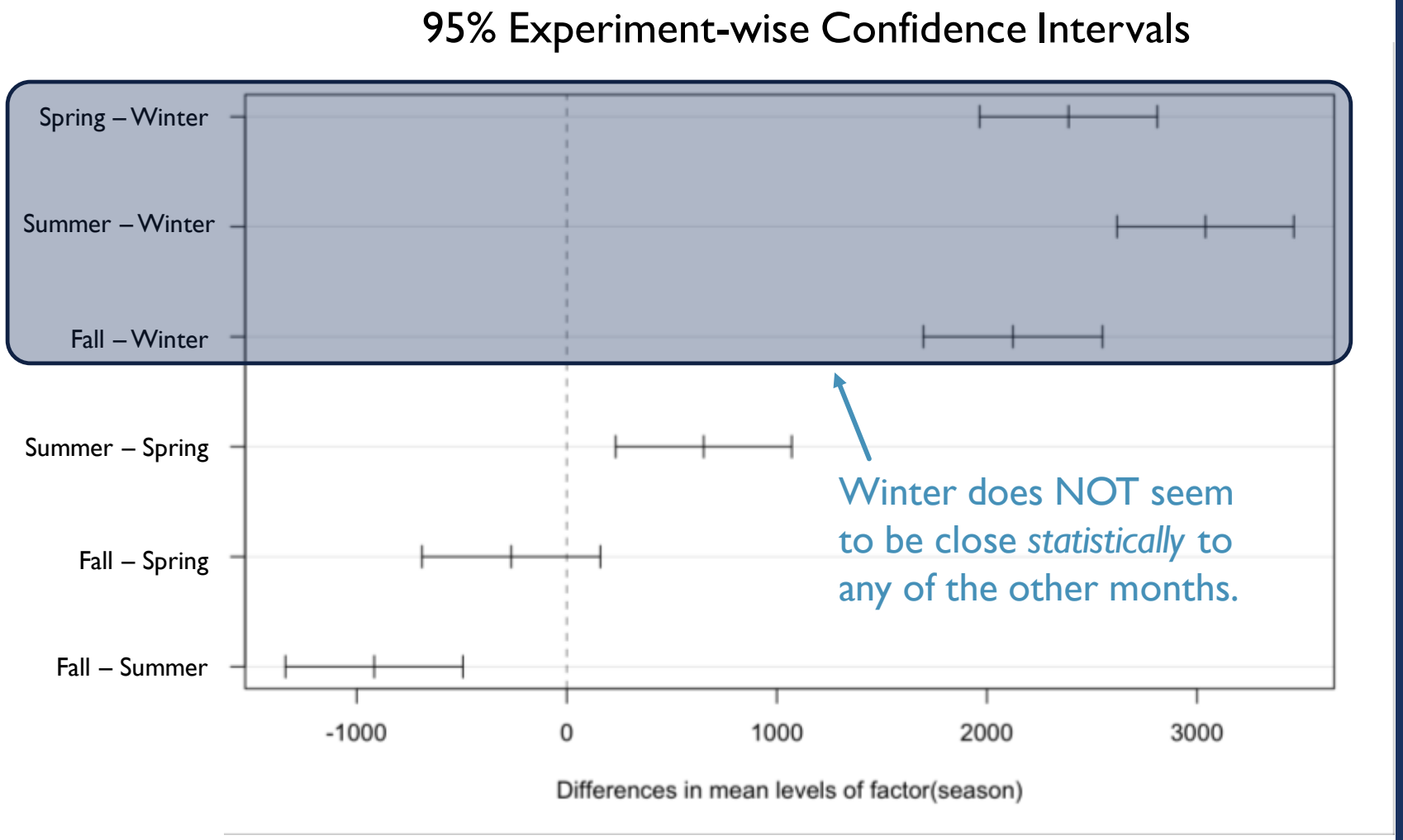
BIKE DATA EXAMPLE – MULTIPLE COMPARISONS



BIKE DATA EXAMPLE – MULTIPLE COMPARISONS



BIKE DATA EXAMPLE – MULTIPLE COMPARISONS



SUMMARY

- If you reject the null hypothesis on the F-test that means there is evidence that shows at least one category is different.
- Once a difference is detected, must test each individual pair of categories to find where all the differences are – a process called multiple comparisons or ad-hoc testing.
- In the process of testing many individual pairs, errors (comparison-wise) are bound to happen if not controlled across an entire experiment (experiment-wise).

LINEAR REGRESSION

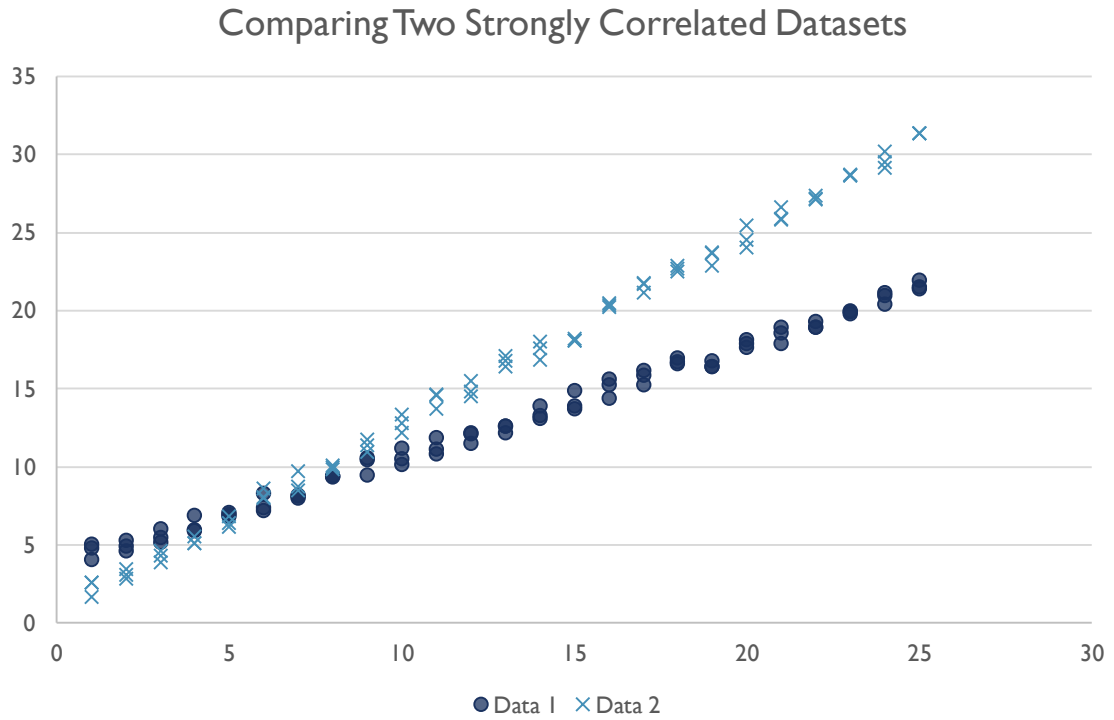
NEXT STEPS WITH DATA

REVIEW OF CORRELATION

- The Pearson correlation coefficient, r, is a measure of strength of the *linear* relationship between two variables.
  - Negative values imply a negative *linear* relationship.
  - Positive values imply a positive *linear* relationship.
  - Values near 0 implies no real *linear* relationship.

CORRELATION IS NOT EVERYTHING

- Correlation is a measure of strength of a linear relationship but does not say what the linear relationship is.
- Plot has two sets of data with exact same correlation of 0.99.
- However, the relationship is different between the two.



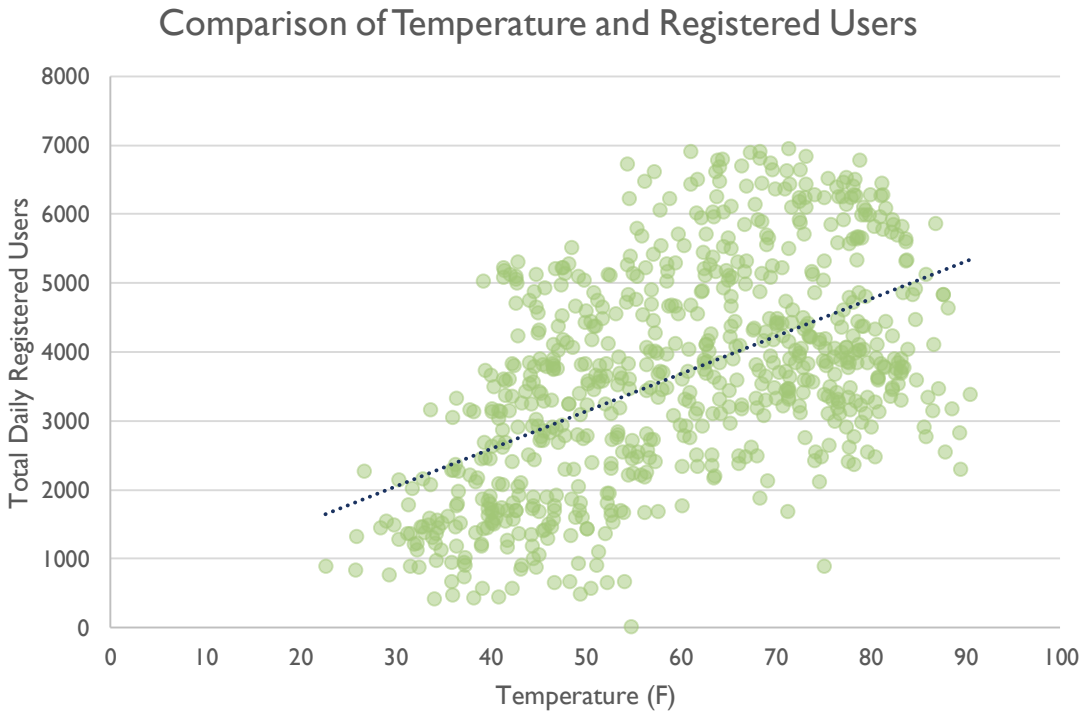
REGRESSION MODELING

- Many people across industries devote research funding to discover how variables are related (modeling).
- The simplest graphical technique to relate two quantitative variables is through a straight-line relationship – called the **simple linear regression** (**SLR**) **model**.
- Most models are more extensive and complicated than SLR models, but SLR models form a good foundation.

BIKE DATA EXAMPLE

- What if you wanted to predict the number of registered users based on the temperature outside?
- What is the best guess line for the following?

$$Predicted\ Users = \beta_0 + \beta_1 \times Temp$$

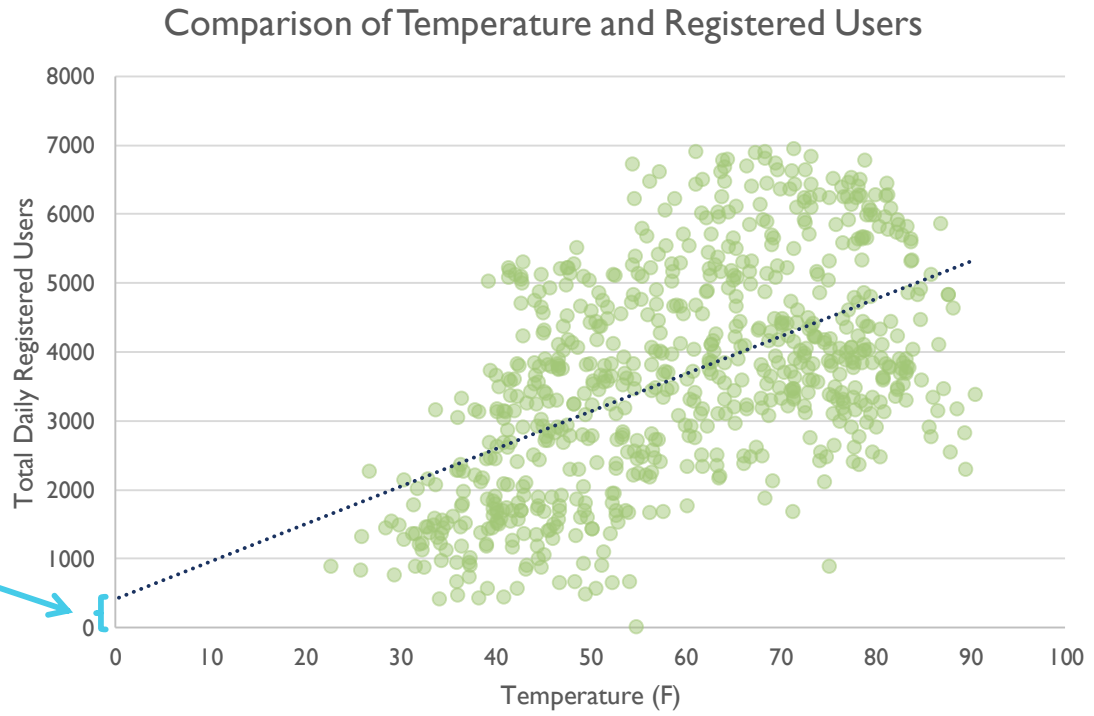


BIKE DATA EXAMPLE

- Simple Linear Regression:

$$Predicted\ Users = \beta_0 + \beta_1 \times Temp$$

Intercept

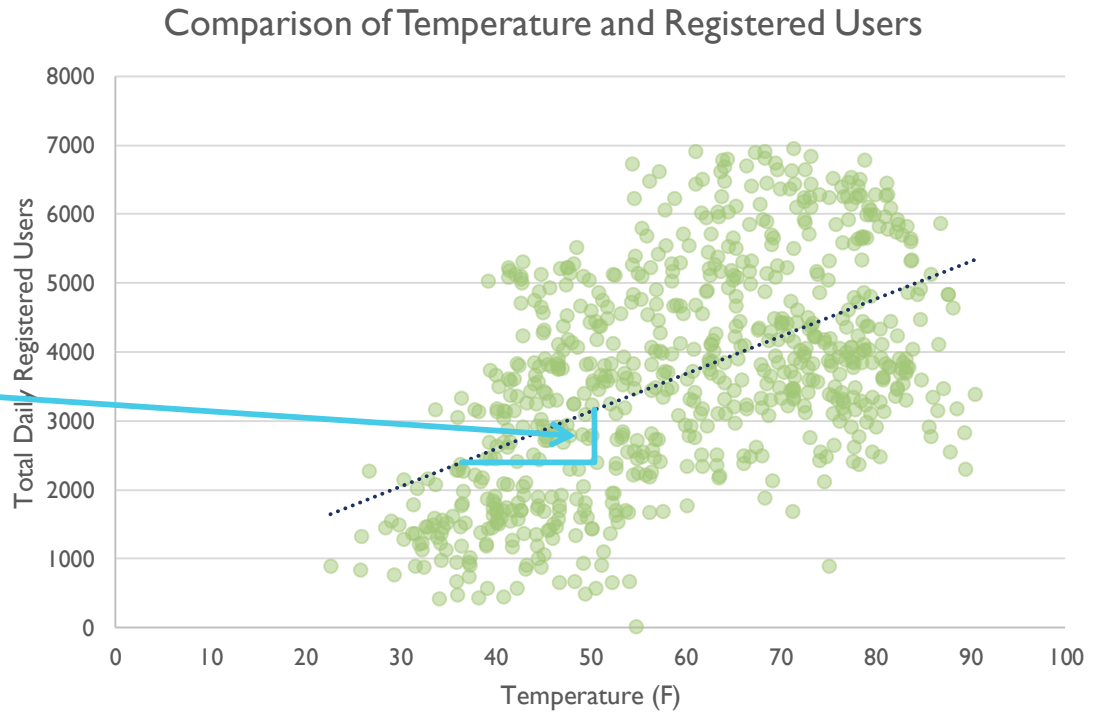


BIKE DATA EXAMPLE

- Simple Linear Regression:

$$Predicted\ Users = \beta_0 + \beta_1 \times Temp$$

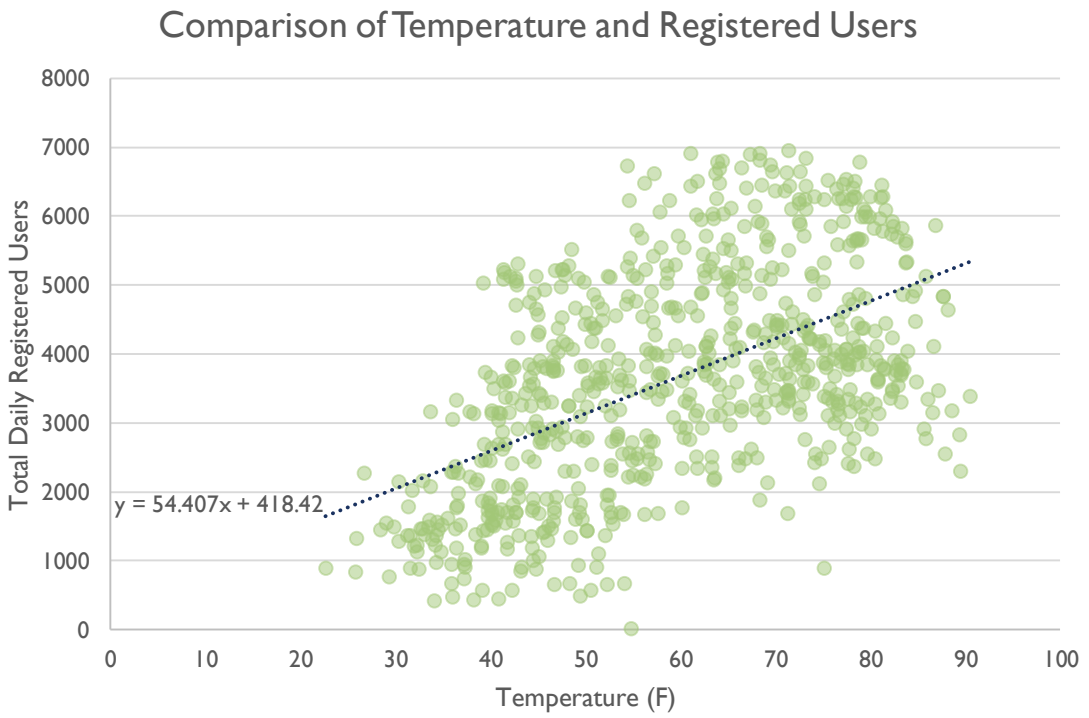
Slope



BIKE DATA EXAMPLE

- What if you wanted to predict the number of registered users based on the temperature outside?
- What is the best guess line for the following?

$$Predicted\ Users = 418.42 + 54.4 \times Temp$$

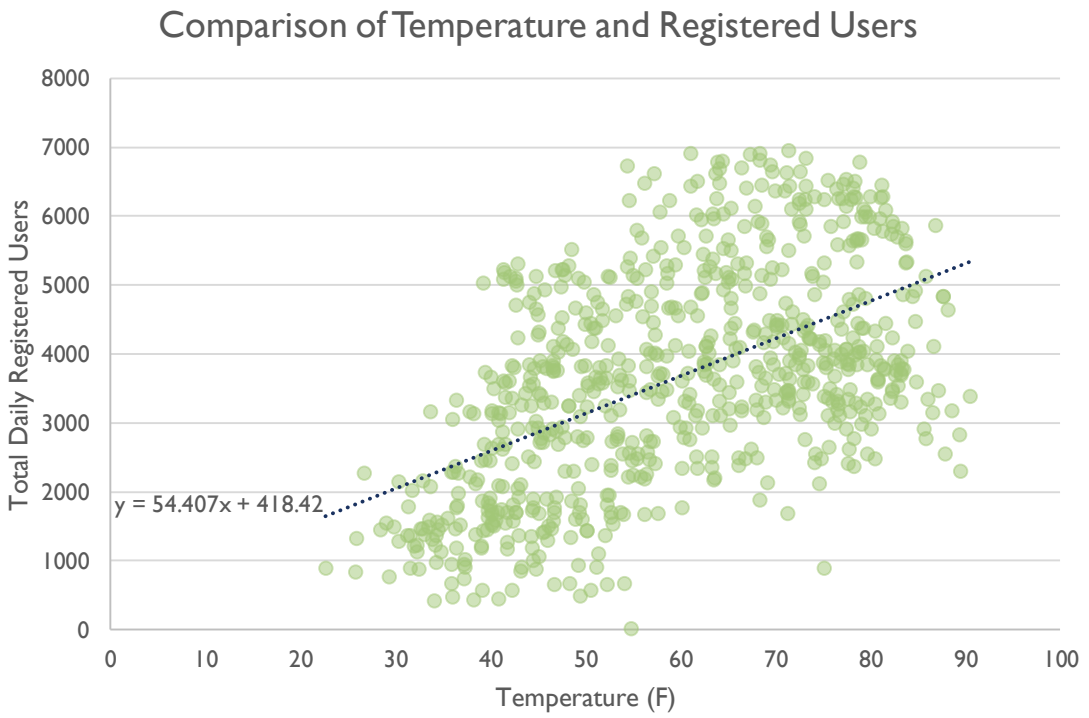


BIKE DATA EXAMPLE

- What if you wanted to predict the number of registered users based on the temperature outside?
- What is the ***best guess line*** for the following?

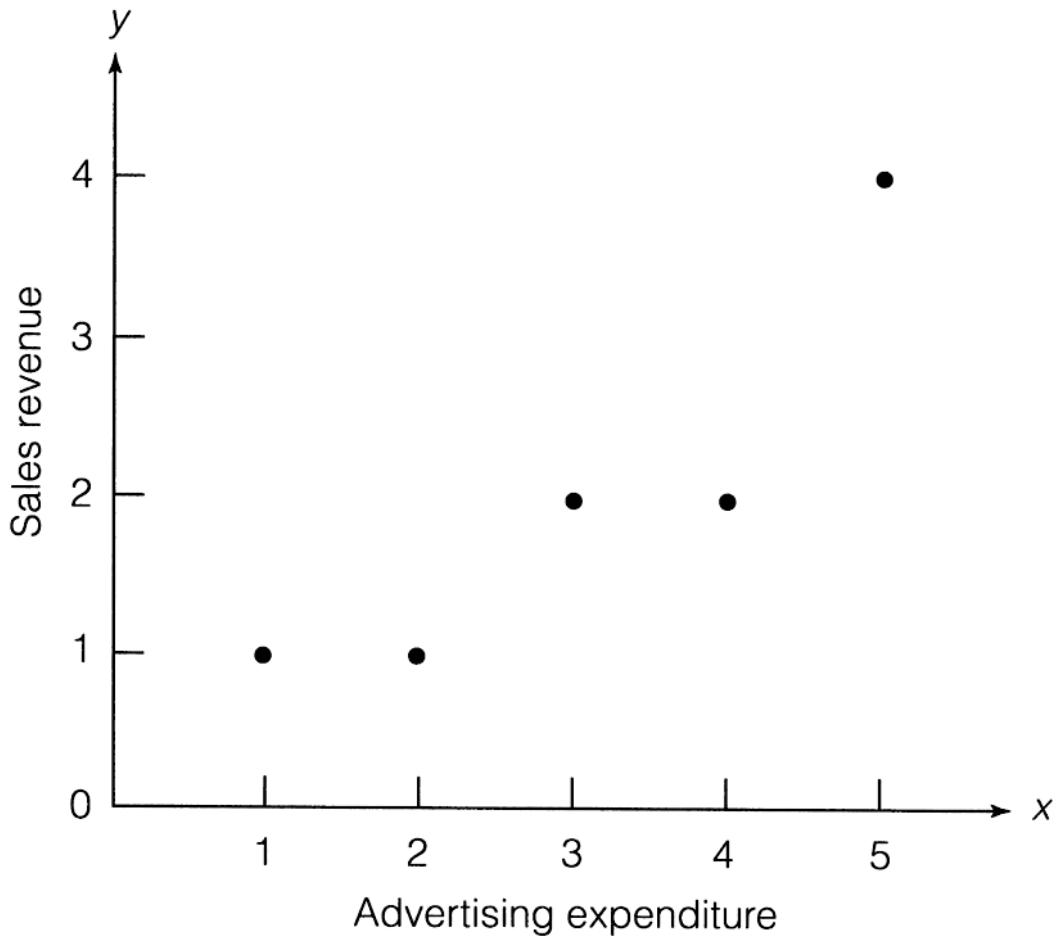
$$Predicted\ Users = 418.42 + 54.4 \times Temp$$

How do we determine this?



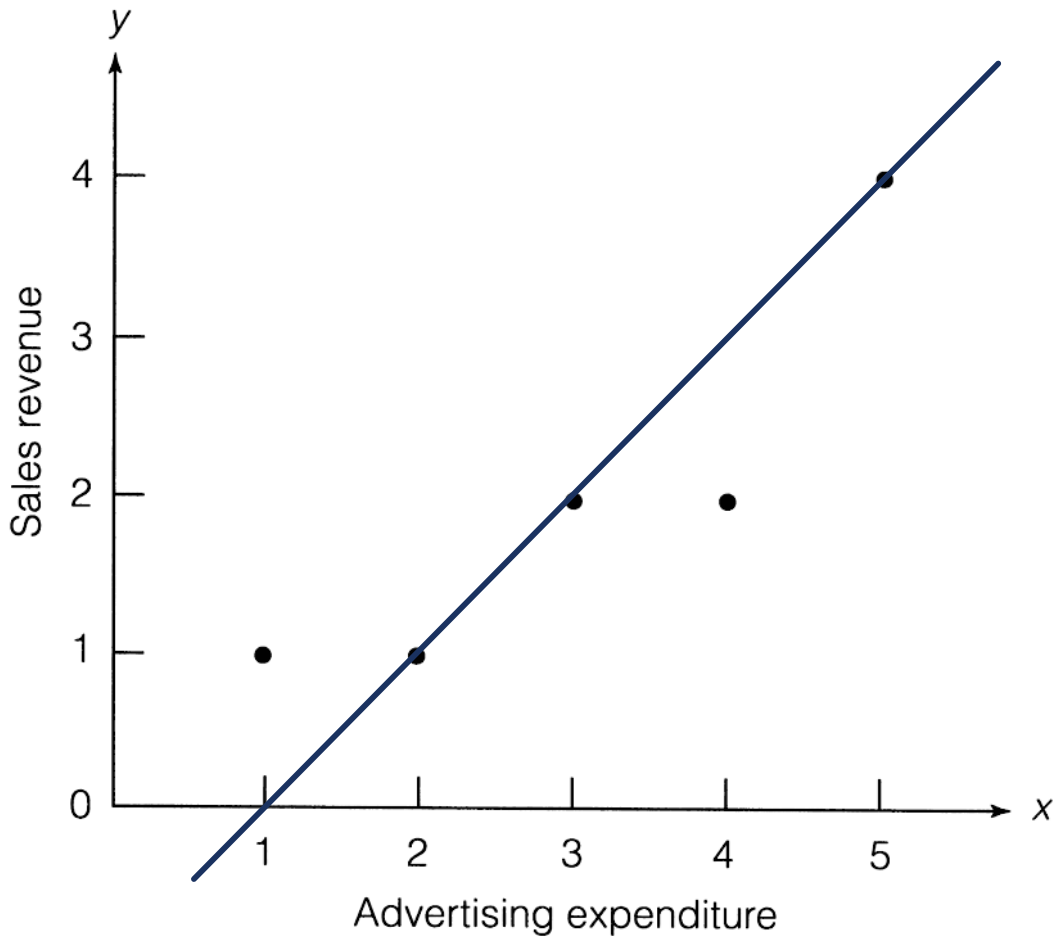
SIMPLE EXAMPLE

- Predicting sales revenue (thousands of $) with advertising expenditure (hundreds of $).
- What is the "best" line through these 5 data points?



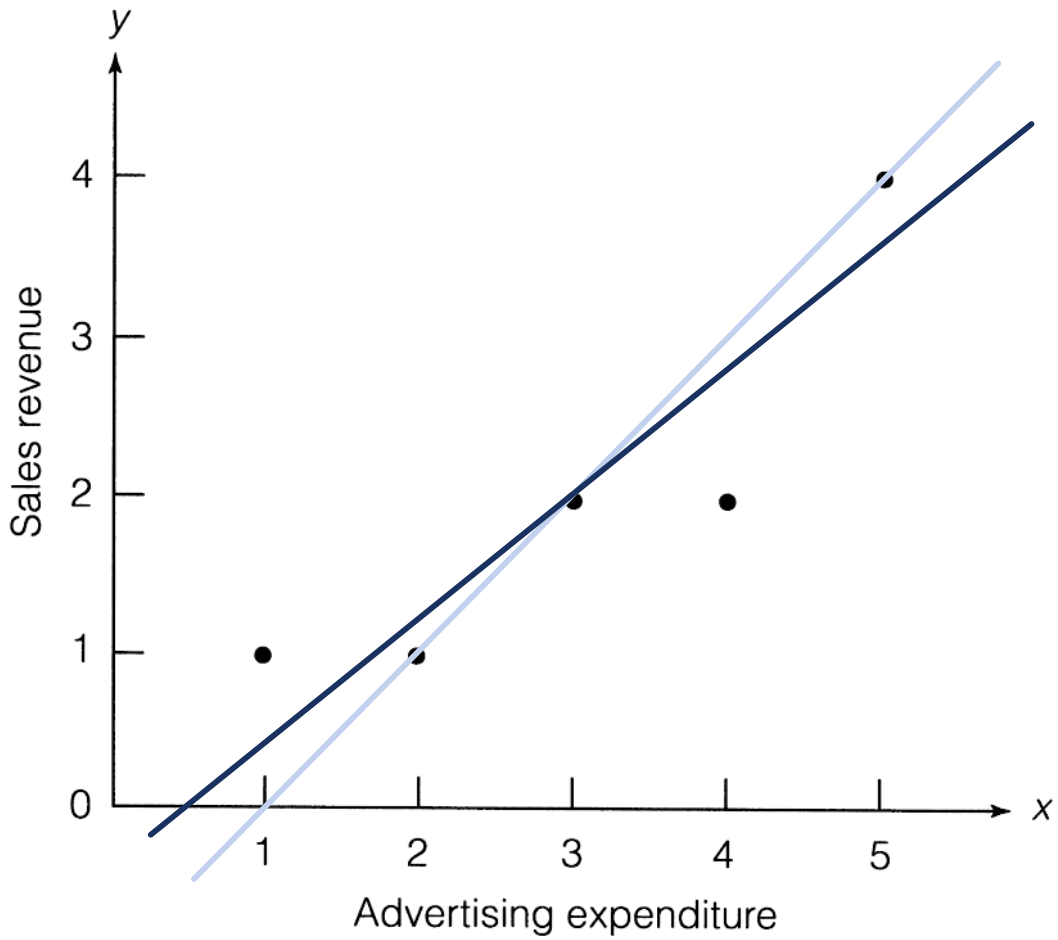
SIMPLE EXAMPLE

- Predicting sales revenue (thousands of $) with advertising expenditure (hundreds of $).
- What is the "best" line through these 5 data points?



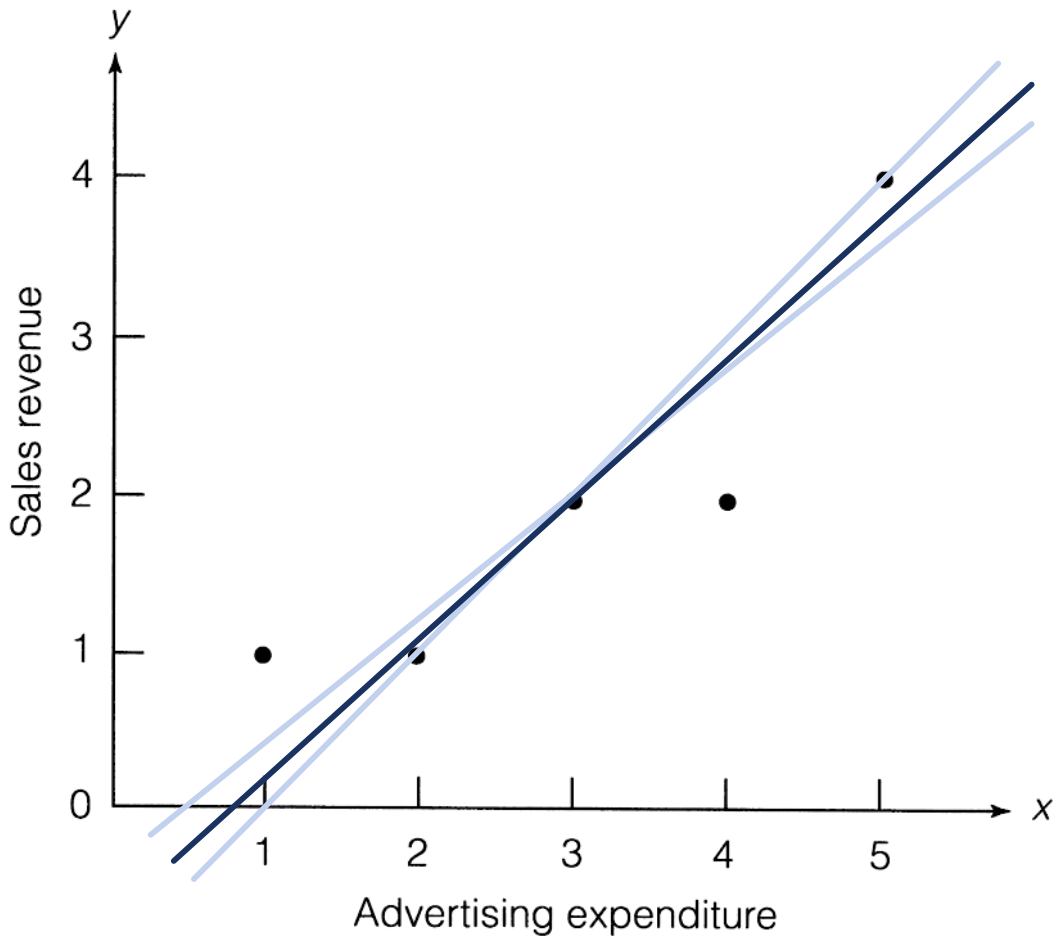
SIMPLE EXAMPLE

- Predicting sales revenue (thousands of $) with advertising expenditure (hundreds of $).
- What is the "best" line through these 5 data points?



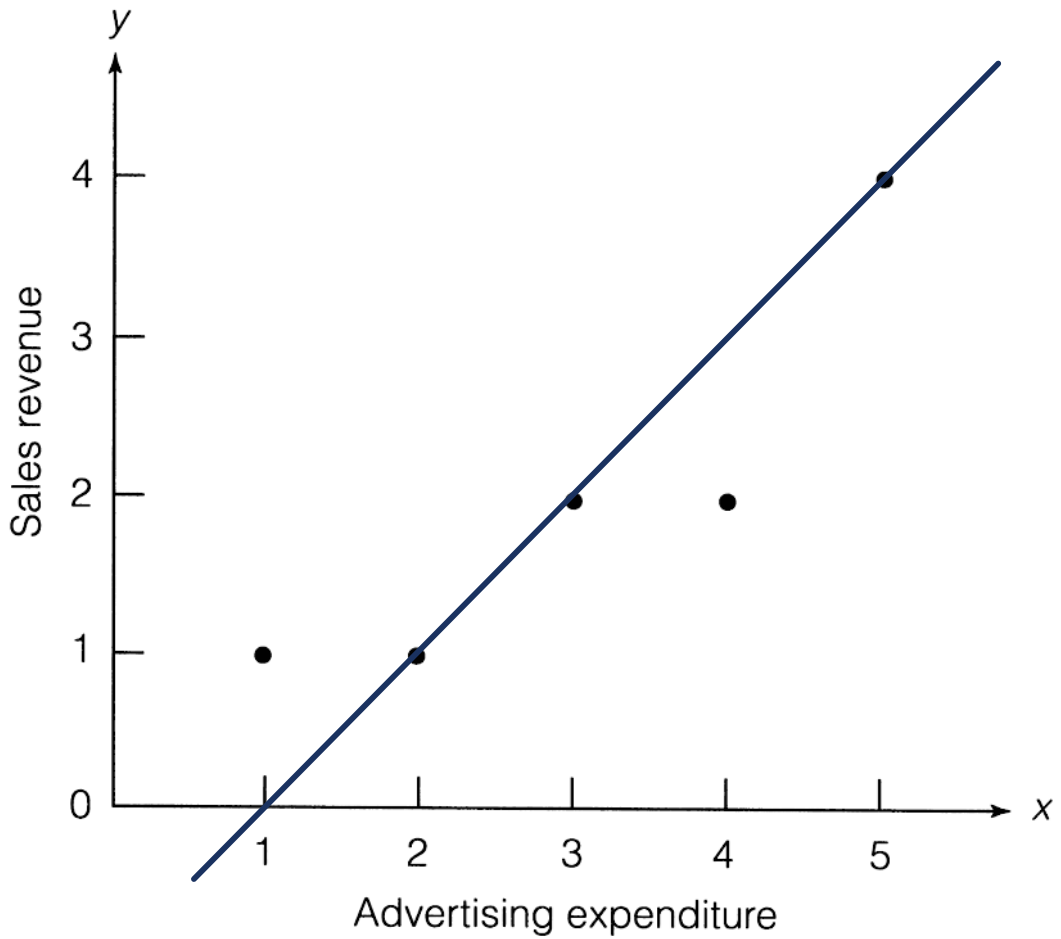
SIMPLE EXAMPLE

- Predicting sales revenue (thousands of $) with advertising expenditure (hundreds of $).
- What is the "best" line through these 5 data points?



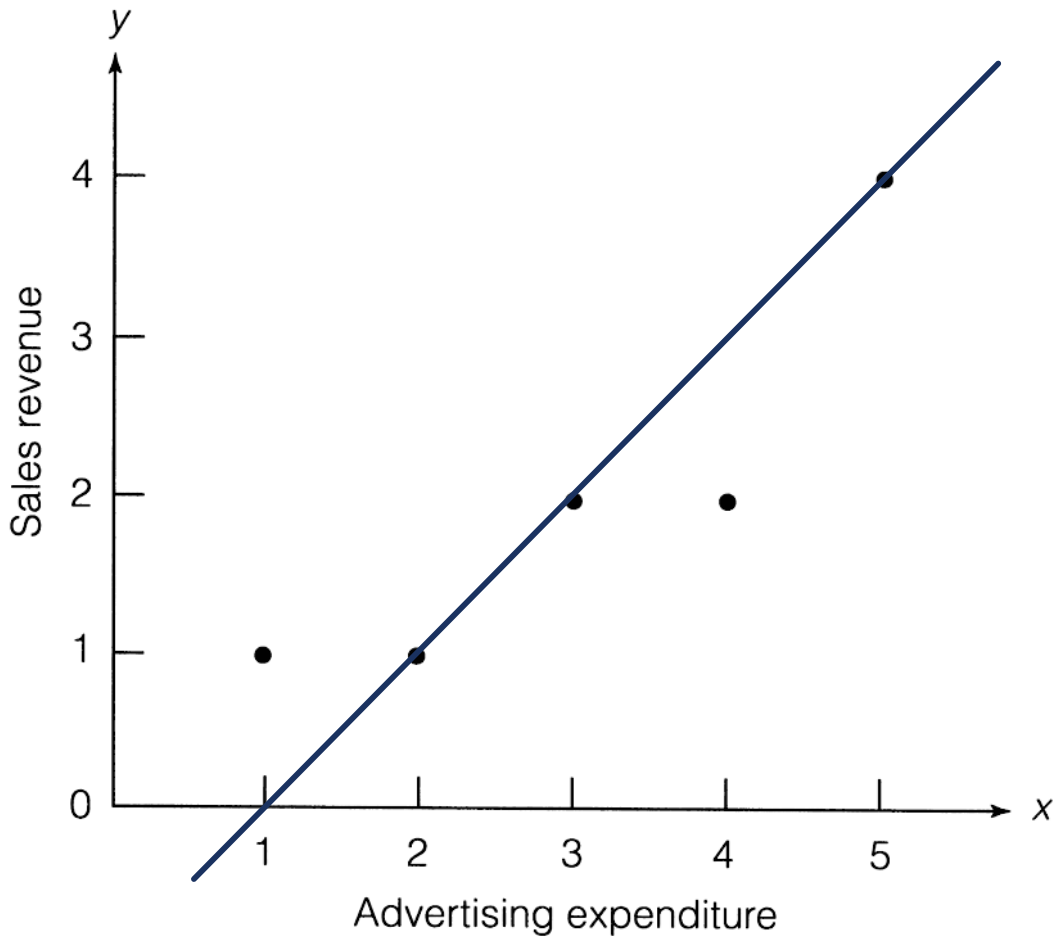
SIMPLE EXAMPLE

- Let's pick one line and work through how we would approach this.



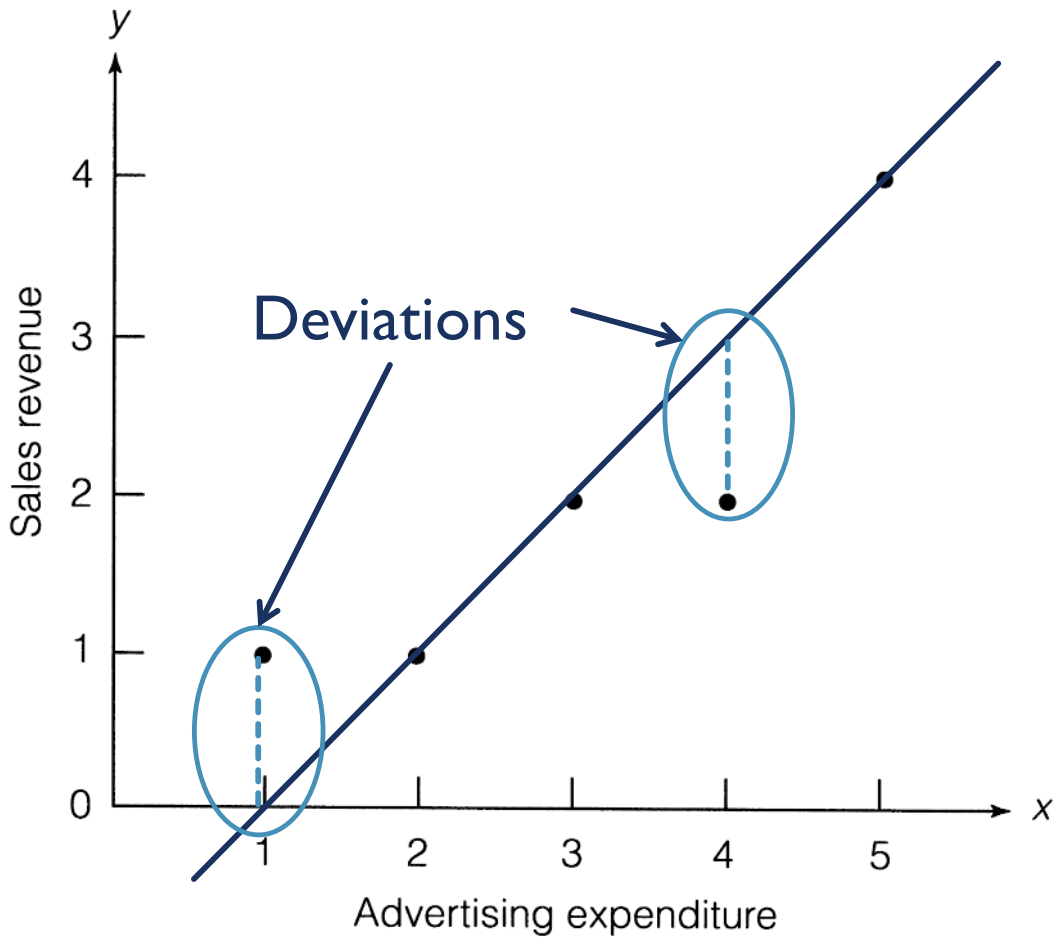
SIMPLE EXAMPLE

- How "wrong" were you at each point?



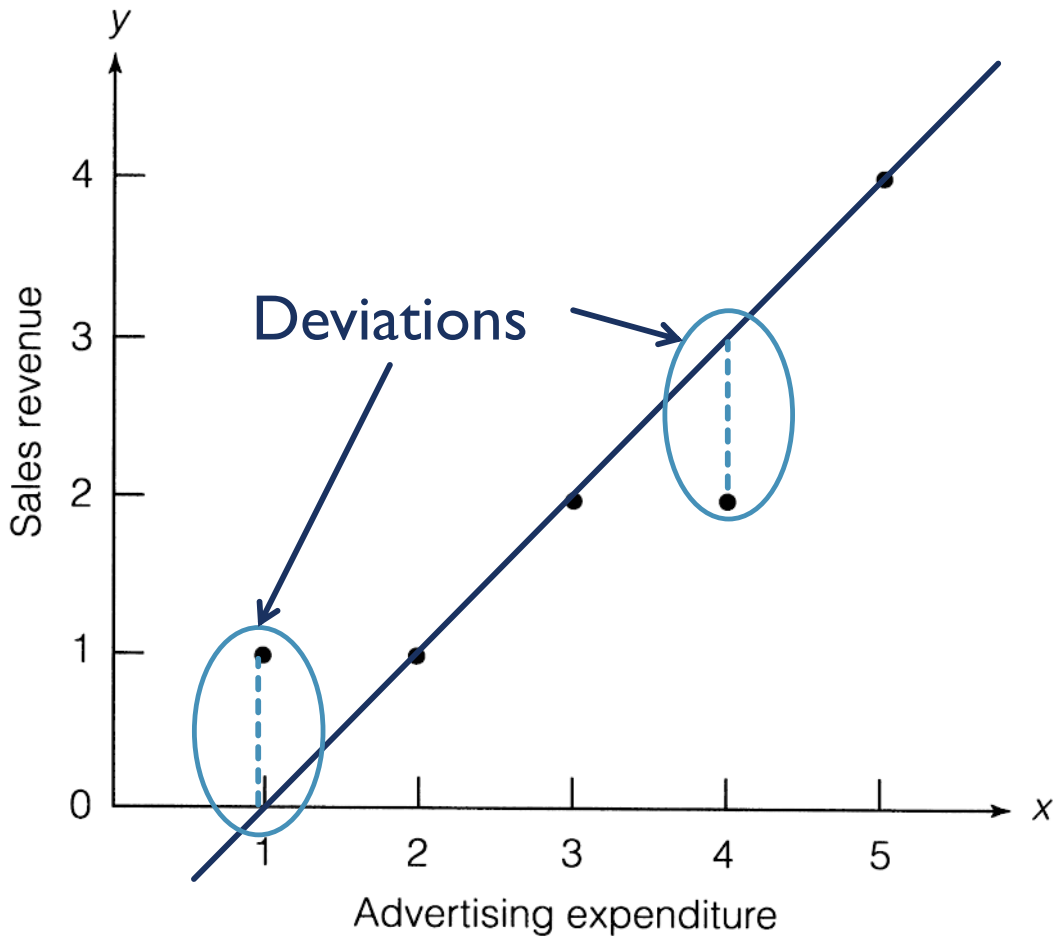
SIMPLE EXAMPLE

- How "wrong" were you at each point?
- Look at the **vertical deviations** from the data point to the line – called **residuals**.



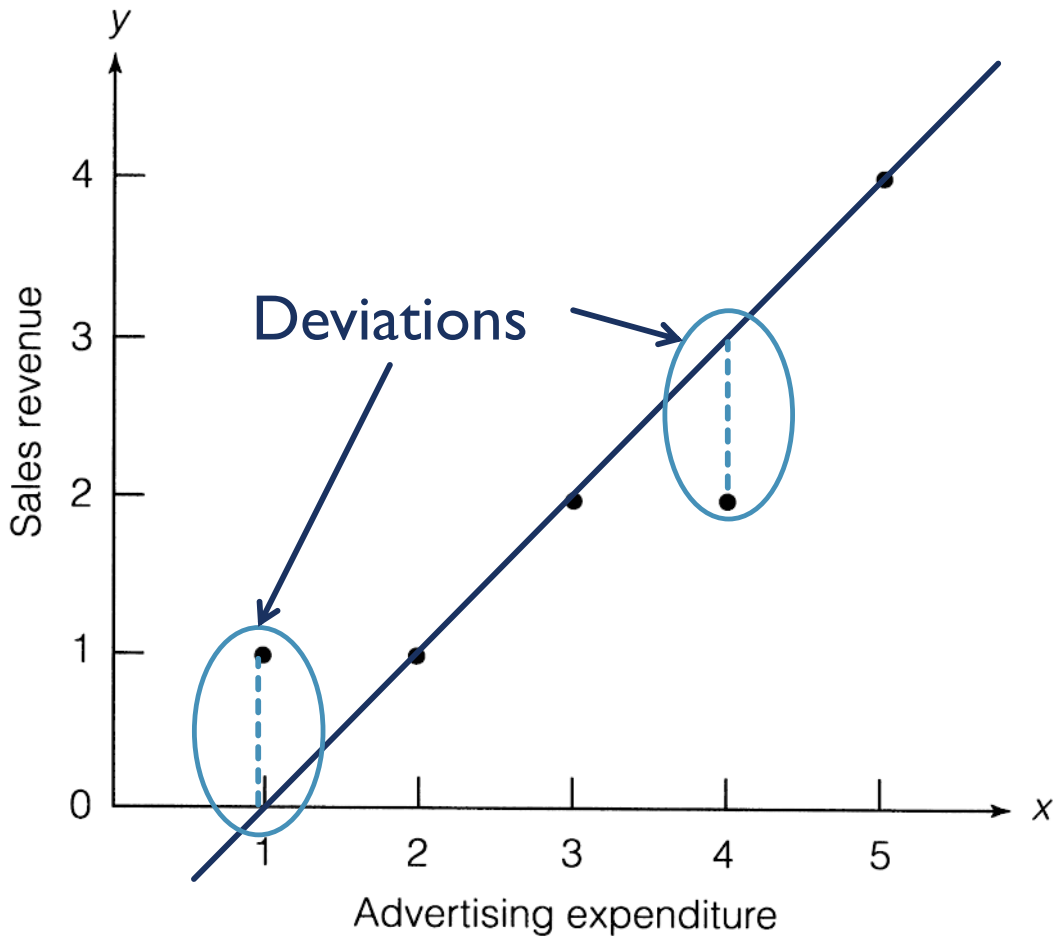
SIMPLE EXAMPLE

- How "wrong" were you at each point?
- Look at the **vertical deviations** from the data point to the line – called **residuals**.



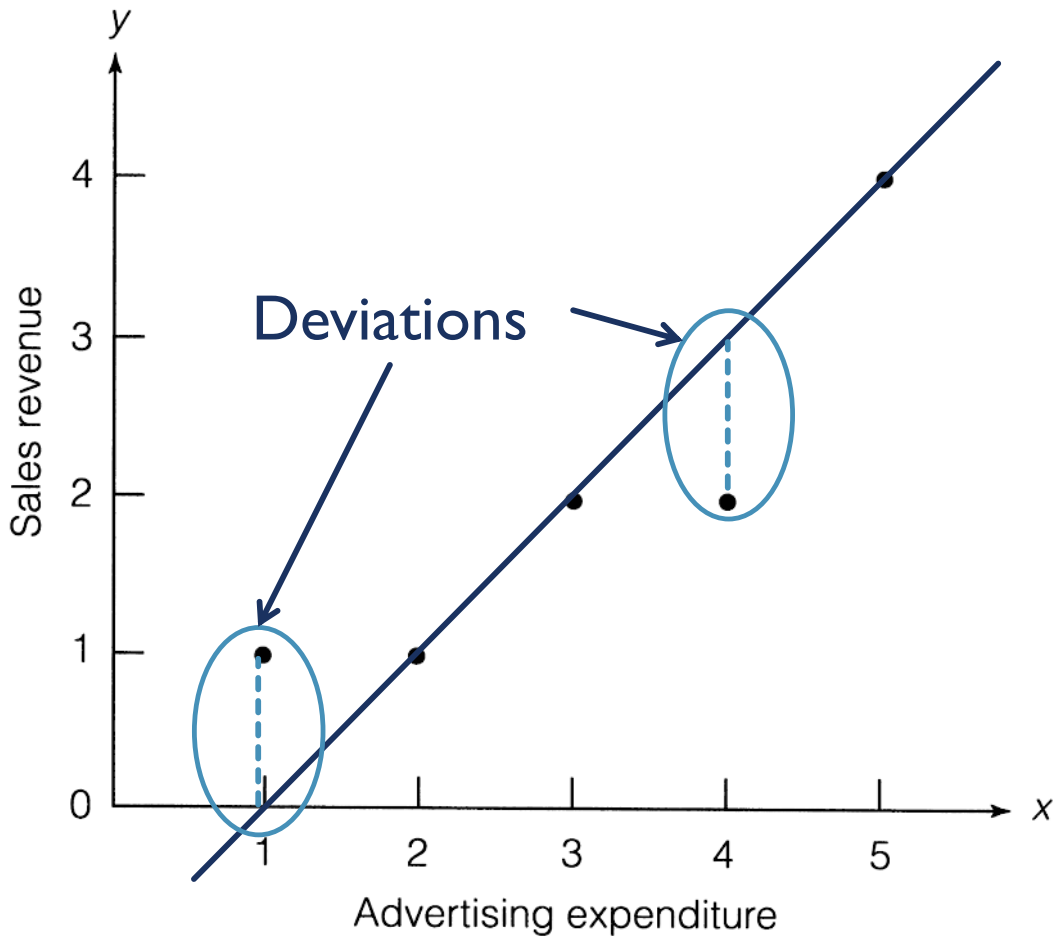
SIMPLE EXAMPLE

- How "wrong" were you at each point?
- Look at the **vertical deviations** from the data point to the line – called **residuals**.
- Can sum up all the deviations to calculate "total" error.



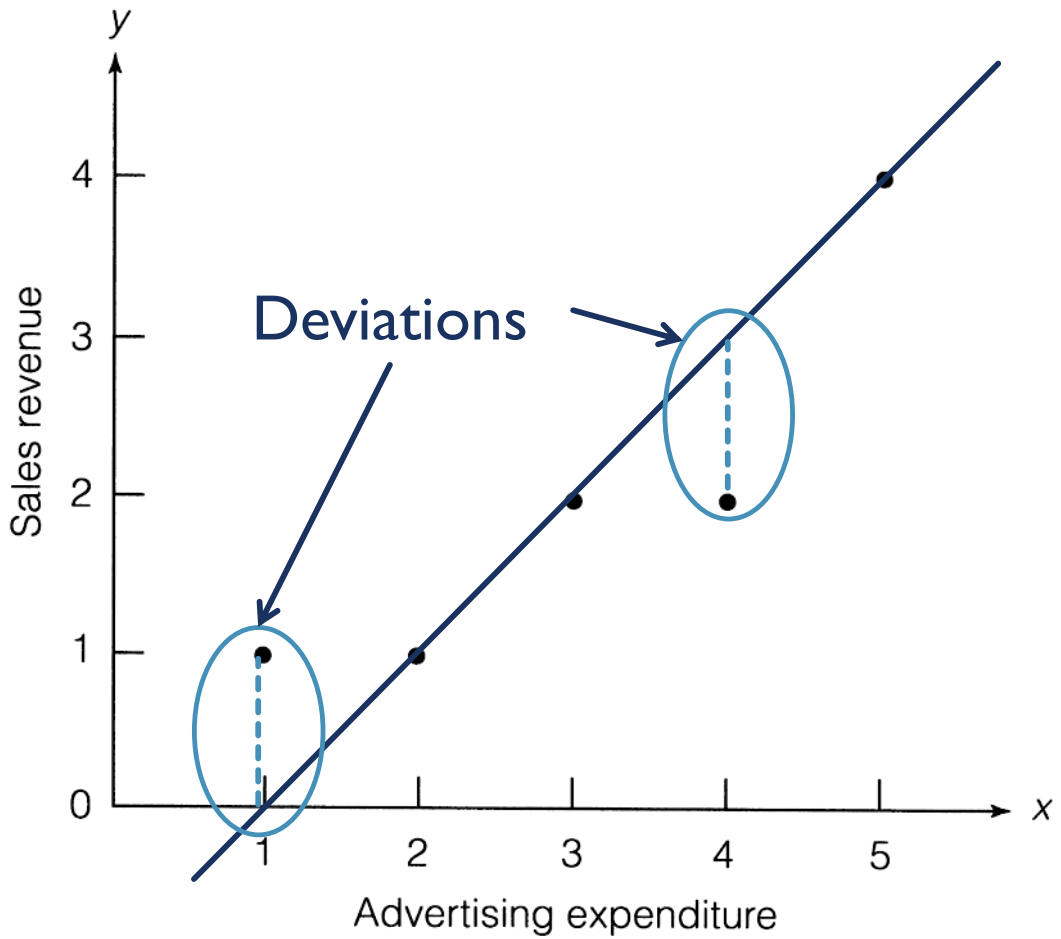
SIMPLE EXAMPLE

- How "wrong" were you at each point?
- Look at the **vertical deviations** from the data point to the line – called **residuals**.
- Can sum up all the deviations to calculate "total" error.
- These errors have both positive and negative values so they would cancel each other out if we just added them.



SIMPLE EXAMPLE

- How "wrong" were you at each point?
- Look at the **vertical deviations** from the data point to the line – called **residuals**.
- Can sum up all the deviations to calculate "total" error.
- Summing the squared errors (error^2) removes the effect of the direction of the error.

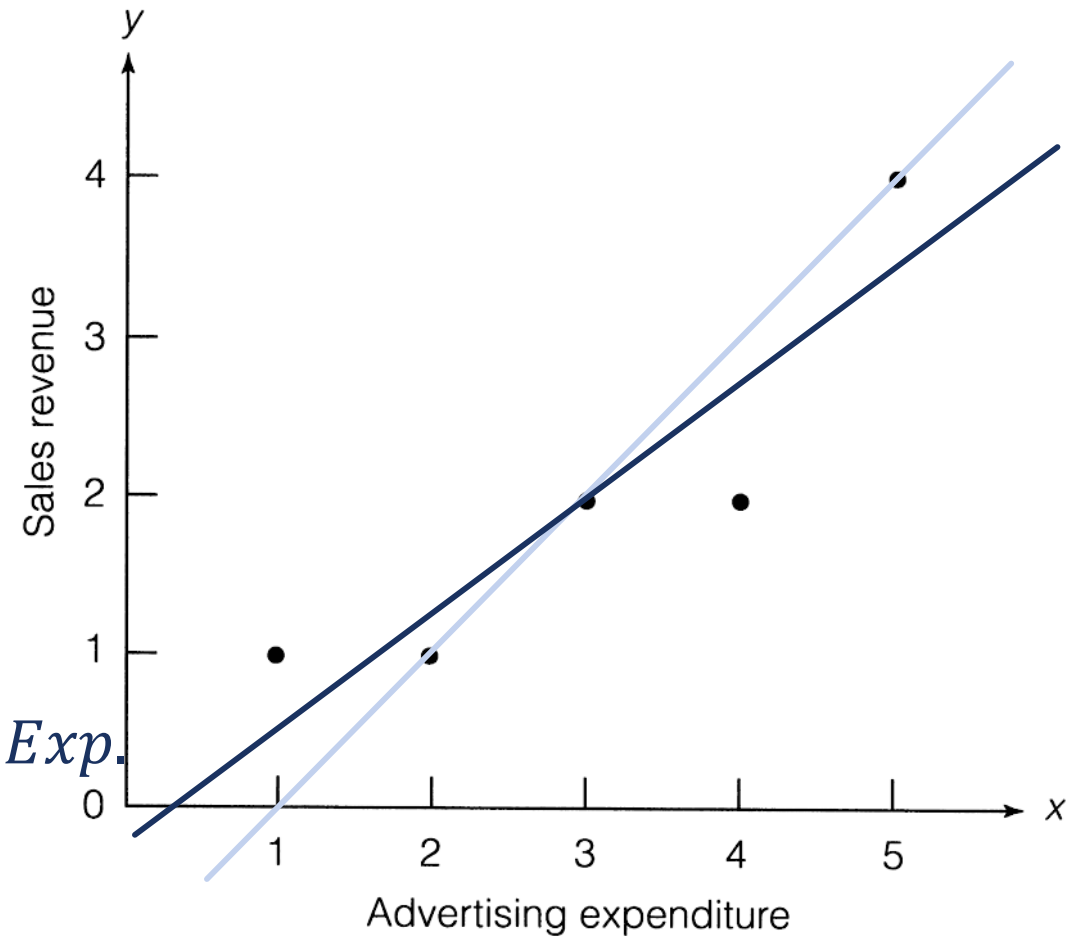


LEAST SQUARES REGRESSION

- It can be shown that there is only **one** line for which the sum of the squared errors is minimized.
- This line is called the **line of best fit** or the **least squares regression line**.

SIMPLE EXAMPLE

- The line of best fit for the 5 data points in the scatterplot is shown in the darker line.
- Not the original line we used for our prediction.
- Computers can easily and quickly calculate this best line for us.



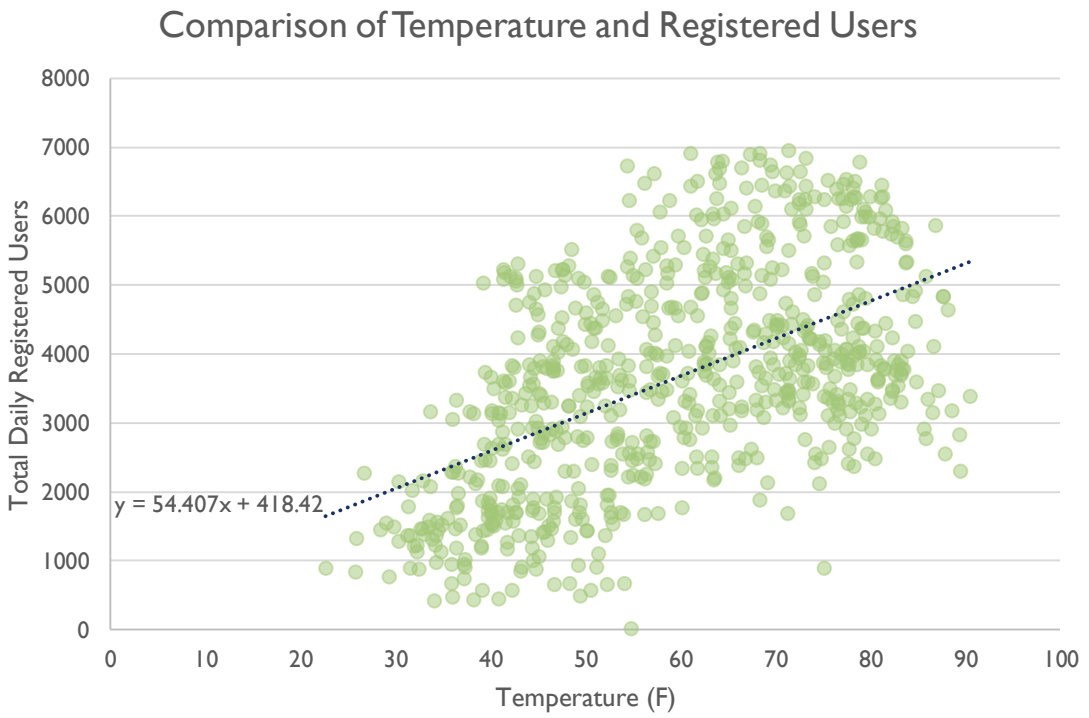
$$Predicted\ Sales\ Revenue = -0.1 + 0.7 \times Adv.Exp.$$

BIKE DATA EXAMPLE

- What if you wanted to predict the number of registered users based on the temperature outside?
- The ***best fit line*** for this relationship is:

$$Predicted\ Users = 418.42 + 54.4 \times Temp$$

- This line is the closest line to each point simultaneously in terms of squared vertical distances.



SUMMARY

- The simplest graphical technique to relate two quantitative variables is through a straight-line relationship – called the simple linear regression (SLR) model.
- Look at the vertical deviations from the data point to the line – called residuals.
- It can be shown that there is only one line for which the sum of the squared errors is minimized – called the line of best fit or the least squares regression line.