Inferences for Regression **Friday, Nov. 10, 2023**

GOAL: to know regression assumptions and how to do hypothesis testing for a slope, CI for a slope, CI for predictions, hypothesis testing for a correlation

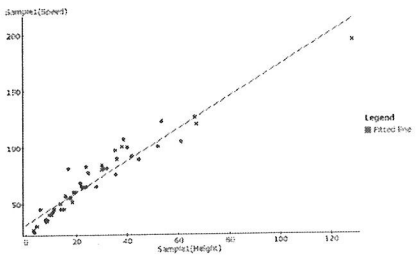
HW#17 and Quiz#17: on Moodle. Due Friday 11/17/2023 at 11:59 p.m.

Up until now, we used regression to *describe* the relationship between two quantitative variables. Now we want to think of the regression line we computed from our sample ($y=b_o+b_1x$) as an estimate of the true regression line in the population ($y = \beta_o + \beta_1 x$), just as $\bar{y}$ is an estimate of the true mean μ. So we'll learn how to compute CI and do hypothesis tests for the parameters β_o (population intercept) and β_1 (population slope). We'll also learn to compute confidence intervals for predictions we make, and to test hypotheses about a correlation coefficient.

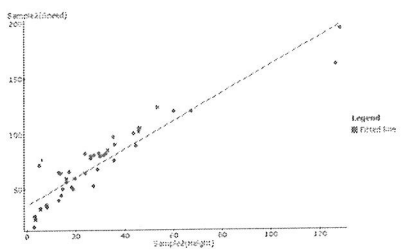
The Simple Linear Regression Model and Related Assumptions

Eg. There's a data set in Statcrunch that contains data on 408 roller coasters, including height (in meters) and speed (in mph). The regression line predicting speed from height using all available data is Speed = 29.6 + 1.54*Height. I took 3 random samples from the data, to see what sort of regression lines we'd get from the samples:

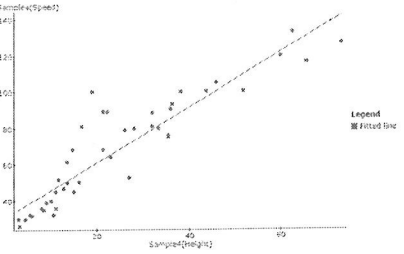
1) Speed = 31.2 + 1.43*Height



2) Speed = 35.4 + 1.27*Height



3) Speed = 30.5 + 1.53*Height

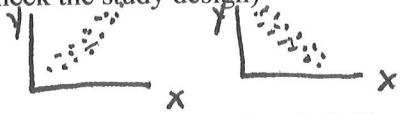


All of these samples are attempting to model the "true" linear relationship that may exist between Speed and Height in the **population** of roller coasters. The regression lines are all sample estimates of the "true" (but unknown) population relationship between Speed (y) and Height (x).

This is why we need inference: it helps us draw conclusions about the (unknown) truth in the population from our sample. We will use CI to estimate things ("What is the range in which we expect to find the true slope in the population?") and hypothesis testing to answer Y/N questions ("Is x useful in predicting y?").

Assumptions for linear regression:

1. SRS, with independent observations (check the study design)
2. The true relationship is linear.



3. The scatter of the points around the line is the same along the whole line (not close at one end and spread out at the other).

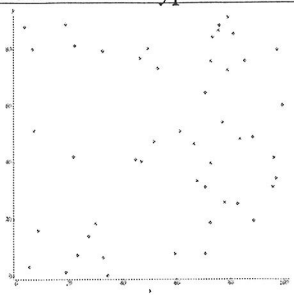
(should look like the graphs above, not

4. The distance of our data points to the regression line has a normal distribution (68% are within $\pm$ 1 s_e (s.e. of the line), 95% are within $\pm$ 2 s_e, . . .).

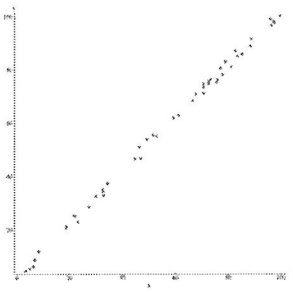
(residuals are the distance of data points to the line. The computer can calculate these & draw a histogram to see if it looks normal.)

The regression line in the population is: $y = \beta_o + \beta_1 x$. Unfortunately, of course, we generally don't have access to the entire population, so we don't know the true intercept and slope in the population (we don't know the parameters β_o and β_1). So we estimate the regression line in the population with the regression line in our sample: $y= b_o + b_1 x$. b_0 and b_1 are sample statistics (the intercept and slope in our sample), and we get them from our computer output. b_o and b_1 are called PARAMETER ESTIMATES, since they are . . .(wait for it . . . .) estimates of the parameters β_o and β_1 (true population intercept and slope). We also get our two measures of how well the regression line fits the data (R^2 and s_e) from our computer output.

The idea behind the hypotheses for hypothesis testing: For which sample does it help to know x when predicting y?



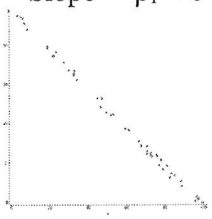
A



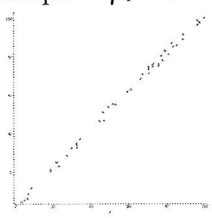
B

Our null hypothesis is that there is nothing going on in the population. In this case, our null hypothesis is that there is no linear relationship between x and y. How could we write that in terms of the slope?

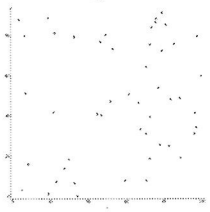
If this relationship is negative, Slope = $\beta_1 < 0$



and this relationship is positive, Slope = $\beta_1 > 0$



then we would write "no linear relationship between x and y" as:



Hypothesis testing: to test the usefulness of the model $y = \beta_o + \beta_1 x$ (is x useful/helpful in predicting y?):

1. Hypotheses:

H_0:

H_a:

2. Test statistic: The output will give you a t-statistic (df=n-2) to test these hypotheses. ($t = b_1/SE_{b1}$)

3. p-value: A few programs (e.g. StatCrunch) let you pick the H_a (<, >, ≠). Then the output will give you the p-value that matches your specified H_a. Most, however, do not. If you see output that does not mention an H_a, then the default H_a printed is two-sided (≠). If the H_a you are interested in is one-sided (<,>), then you will need to divide the p-value in half, since the p-value from a two-sided H_a (≠) is twice as big as the p-value from a one-sided H_a (<, >).

4. Decision rule: At the α level of significance, reject H_0 if $p \leq \alpha$, and accept H_0 if $p > \alpha$.

Confidence intervals for β_o and β_1:

Use the formulas $b_o \pm t * SE_{bo}$

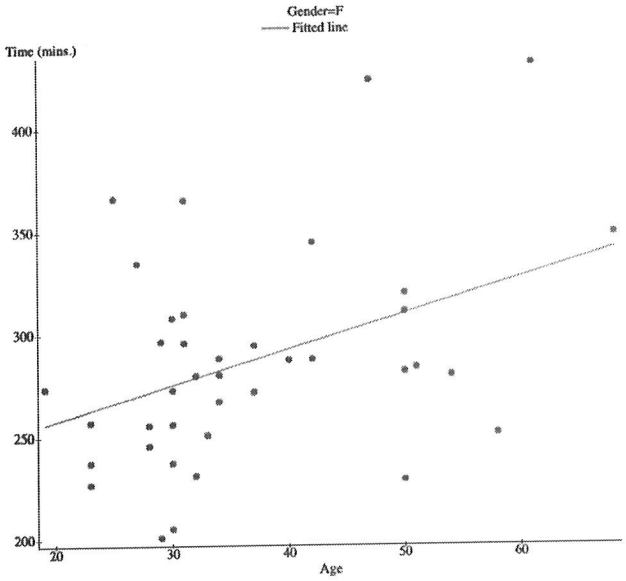
$b_1 \pm t * SE_{b1}$

We are generally only interested in a CI for the slope (β_1). To compute this CI, we get:

--b_1 and SE_{b1} from the output. b_1 is the estimate of the slope from our sample, and SE_{b1} tells us how reliable that sample estimate is (how much it might vary from sample to sample).

--t from Table T with n-2 df and the desired level of confidence.

e.g. When I turned 60 (in 2020), in an attempt to show that I'm not dead yet, I created a COVID-19K as part of my birthday celebration. That's minor league compared to people who can run a full marathon. The following graph and regression analysis uses age to predict marathon finishing time (in minutes) for these amazing women.



Simple linear regression results:
Where: Gender=F
Dependent Variable: Time (mins.)
Independent Variable: Age
Time (mins.) = 222.74 + 1.74 Age
Sample size: 39
R (correlation coefficient) = 0.39235019
R-sq = 0.15393867
Estimate of error standard deviation: 48.273143

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	222.74	25.8116		37	8.6295057	<0.0001
Slope	1.74	0.6702		37	2.5946203	0.0067

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	15687.681	15687.681	6.7320544	0.0135
Error	37	86220.963	2330.2963		
Total	38	101908.64			

Predicted values:

X value	Pred. Y	s.e.(Pred. y)	95% C.I. for mean	95% P.I. for new
60	327.1	17.4	(291.8, 362.3)	(223.1, 431.1)

Hypothesis testing: (Y/N question). Does finishing time increase with age?

H_o:

H_a:

t-statistic=

p-value =

decision (α=0.05) :

Compute a 95% CI for the slope (rate of change). i.e. Estimate the slope in the population. i.e. What is the range in which we expect to find the true slope of the population?

b_1 =

t(95% CI, 37 df) =

SE_{b1} =

$b_1 \pm t * SE_{b1}$

Interpret: We are 95% confident that the finishing time increases (on average) by between 0.38 minutes and 3.1 minutes for each additional year older a woman gets. So, though we don't know what the actual rate of change is for the population, we think that if we had access to the entire population of female runners, the slope of the group as a whole would probably be between 0.38 and 3.1.

Remember what affects the width of a CI. As we increase n, the width of the CI increases/decreases (i.e. we have more precision). As we increase the level of confidence, the width increases/decreases (so we're more confident of catching the true slope of the population).

Using the model to predict

We've been using a regression model to predict the response for one certain value of the predictor. The problem is that if we had a different sample, we'd have a slightly different regression model, and we'd make a different prediction. Therefore we are going to develop a confidence interval for our predictions. You've probably already seen CI for predictions that models make, though you didn't know that they were called CI. e.g. With COVID, when they gave forecasts for the number of cases and the number of deaths, you always saw a best guess and a range. This is exactly what we are doing with confidence intervals: the prediction based on our sample is our best guess, and the margin of error tells us how close we think that is to what will actually happen. CI: sample value ± margin of error. In COVID-19 news, they made their best prediction based on their model, but there was a margin of error around that best prediction, so they ended up with a range in which they expected to find what was going to happen.

We can use x to predict y for one particular individual where $x=x_o$, and we can use x to predict the mean of y for everyone who has $x=x_o$.

e.g. The regr. eqn. for finishing time from age is: finishing time (minutes) = 222.74 + 1.74*(age)

1. Predict the average finishing time of 60-year-old women who are runners (i.e. predict the mean finishing time of all women who run marathons where age=60).

2. Predict the finishing time of one 60-year-old (i.e. predict the finishing time of one particular woman where age=60).

3. Which number, if either, do you think may have the greater error (wider confidence interval)? Why?

Formulas for CI for predictions:

CI when predicting the mean of several observations when $x=x_o$:

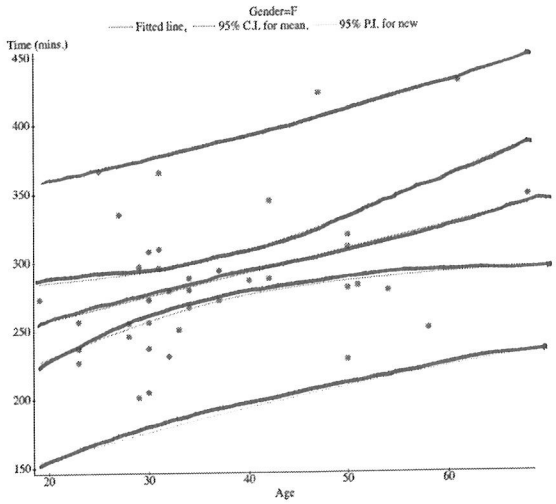
(Predicting μ_y from $x=x_0$)

$$\hat{y} \pm t^* s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

Prediction interval = PI = CI when pred. an indiv. response when $x=x_o$:

(Predicting y from $x=x_0$)

$$\hat{y} \pm t^* s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$



The width of a CI for a prediction depends on the same old stuff (level of confidence, sample size), but it adds two: what you are predicting (CI vs PI), and where you are predicting (i.e. is the predictor near the mean?).

1. CI. As our level of confidence decreases (i.e. 95% to 90%), the CI gets wider/narrower.
2. sample size. As our sample size increases, the CI gets wider/narrower.
3. CI vs PI. When predicting for an individual (e.g. height of one 25 year old man), the CI is wider/narrower than when predicting the mean of many individuals (e.g. mean height of all 25 year old men).
4. Predicting near the middle vs at the edges. When we are predicting in the middle of our predictor's range (e.g. if the mean age of men is 40, predicting the height of a 40 year old man) the CI is wider/narrower than if we are predicting at the edges (e.g. predicting the height of a 80 year old man).

You will never have to compute these confidence intervals for predictions. They are (in my opinion) the nastiest formulas all semester. But you need to know where to find them on the output, and which one to use. Look at the output on the previous page, under Predicted Values. It shows we are predicting the finishing time for a 60-year-old woman, and that our best guess at the finishing time (based on this sample) is 327.1 minutes. But we are 95% confident that an individual 60-year-old female runner will finish between 223.1 and 431.1 minutes, and we are 95% confident that the average finishing time for all 60-year-old female runners is between 291.8 and 362.3 minutes.

Hypothesis testing for the Correlation Coefficient (ρ)

ρ is the _____.

r is the _____.

We use _____ to estimate _____.

r and the slope are closely related: $\rho=0$ exactly when $\beta_1=0$. This makes sense, because there is no linear relationship between x and y when $\rho=0$, and when $\beta_1=0$.

When hypothesis testing for a correlation, the Y/N question we are asking is, "Is there a significant correlation?" i.e. "Is the correlation in the population equal to 0?" Since $\rho = 0$ exactly when $\beta_1 = 0$, hypothesis testing for a correlation is identical to hypothesis testing for the slope. So to test H_0: $\rho=0$, we will use exactly the same t-statistic and p-value we did for the slope (H_0: $\beta_1 = 0$). The t-statistic, p-value, and decision are identical for these two tests.

e.g. Does finishing time increase as age increases for women running a marathon?

H_o:

H_a:

t =

p-value =

decision:

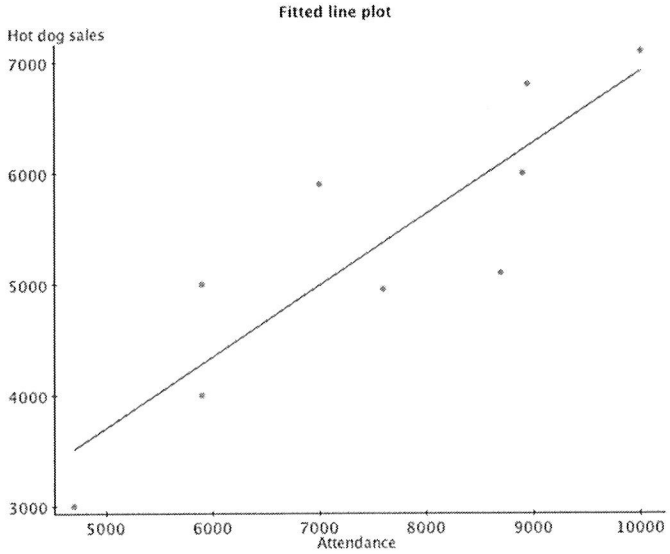
Comparing R^2, r (the correlation coefficient), and b_1 (the slope): what does each show?

R^2	r (corr)	b_1 (slope)

All of these have a value of_____at the same time, indicating that there is NO LINEAR RELATIONSHIP between x and y. We have just random scatter (or a curve), with a horizontal line, so knowing x tells us nothing about y.

Example from start to finish:

A scatter plot of the daily attendance and the number of hot dog sales for a sample of 9 games of a minor league baseball team suggest that the relationship between attendance and sales may be linear. The regression output obtained is as follows:



Simple linear regression results:
Dependent Variable: Hot dog sales
Independent Variable: Attendance
Hot dog sales = 476.89 + 0.64 Attendance
Sample size: 9
R (correlation coefficient) = 0.8734
R-sq = 0.7628491
Estimate of error standard deviation: 678.0368

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	476.89	1044.67		7	0.456	0.6619
Slope	0.64	0.136		7	4.745	0.001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	1.0351862E7	1.0351862E7	22.517073	0.0021
Error	7	3218137.5	459733.94		
Total	8	1.357E7			

Predicted values:

X value	Pred. Y	s.e.(Pred. y)	95% C.I. for mean	95% P.I. for new
7000	4983.999	236.64	(4424.45, 5543.55)	(3285.86, 6682.14)

1. What is the predictor (independent variable)?

2. What is the response (dependent variable)?

3. What does the scatter plot tell us about:

   Form

   Direction

   Strength

   Outliers

   Range of x's (attendance)

4. What are the assumptions of linear regression?
   --SRS, independent observations (check study design)
   --we have a line, not a curve (look at the scatter plot)
   --the points are scattered evenly along the length of the line (look at the scatter plot)
   --the distance of actual hot dog sales (the data points) to predicted hot dog sales (the line) has a normal distribution

5. Circle the regression equation.

6. What is the slope? What does it tell us?

7. Predict hot dog sales when attendance is 7000.

8. There are two confidence intervals for a prediction (95% CI, 95% PI). Which one would we use if we are the manager of the ballpark, and we want to predict hot dog sales for today's game with attendance of 7000?

   If attendance was 10,000, would the CI and PI be wider, narrower, or the same width as when attendance is 7000?

   If attendance was 12,000, should we make a prediction?

9. What is a 95% CI for the slope?

10. How do we interpret that CI?

    We are 95% confident that β_1 (the slope in the population of baseball games in this ballpark) is between 0.32 and 0.96. So, though we don't know the slope, since we don't have data from ALL the baseball games in this stadium, we think we know where to find that slope. We think that if we had all the data, we'd find that as attendance increases by 1 person, #hot dogs sold would increase by somewhere between 0.32 and 0.96 hotdogs.

    Do you think it would be reasonable for the manager to claim that they need 1 additional hotdog for each person who comes through the gate?

11. Do hot dog sales increase as attendance increases (α=0.05)?

12. What is the percent of the variation in hot dog sales that is explained by attendance? Do we have a good fit?

    What might explain the remaining 23.72% of what's going on in hotdog sales?

13. What does s_e (the standard error of the line) tell us? (Note that sy = s.d. of hot dogs sales=1250 hot dogs)

    s_e = _____. *That is a measure of spread around the* _____. *So the difference between actual hot dog sales and predicted hot dog sales has a normal distribution with a s.d. of 678 hotdogs. So* _____*% of the games have hot dog sales within* _____ *hot dogs of what we predicted, and* _____*% of the games have hot dog sales within* _____ *hot dogs of what we predicted.*

    Is this regression line a good fit, based on the s_e? s_e=678 hot dogs is a little more than half of the random overall variation in hot dog sales (s_y=1250 hot dogs), so it is a _____*fit.*

14. What is the correlation coefficient, and how is it computed?

15. Can we assume that this relationship between hot dog sales and attendance is also true for football games?

Analysis of Variance **Wed., Nov. 15, 2023**

GOAL: to know when to use ANOVA and to be able to interpret ANOVA output (including multiple/pairwise comparisons)

Homework#18 and Quiz#18: Due Monday (11/20/2023) at 11:59 p.m.

Where have we been since the last test, and where are we going now?

	Categorical	Quantitative
Categorical	Chi-square 2-sample z-test for proportions (if it's a 2x2 table)	Analysis of Variance 2-sample t-test (if 2 groups)
Quantitative		Correlation Regression

Recall, when we wanted to compare just two population means, we used the

Now we are going to expand this to compare two or more population means.

We use **Analysis of Variance** (ANOVA) to compare groups (categorical variable) on a quantitative variable. So, e.g. we could use ANOVA to answer the following questions:

1. Which academic department in the sciences gives out the lowest average grades?

2. Which kind of promotional campaign leads to greatest store income at Christmas time?

3. How does the type of career relate to the total cost in annual claims someone is likely to make on their health insurance?

The basic idea of ANOVA:
Analysis of Variance is well-named, because that is exactly what we are doing. We are comparing the variation (i.e. differences) between group means to the random variation of individuals within groups. If there is relatively more variation between groups than within groups, then we are going to conclude that at least one of the population means is significantly different from another.

Using graphics to get an idea of what is going on:

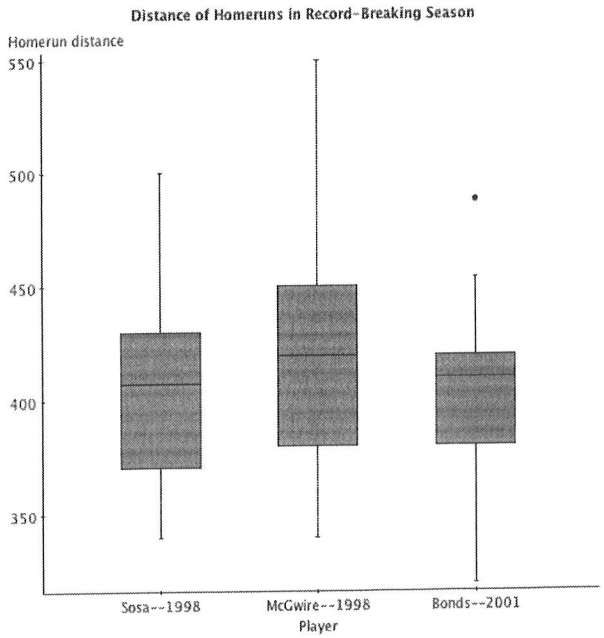
As with the two-sample t-test, we can use side-by-side boxplots:

ANOVA asks: What is the important story here?
--Are there big differences in homerun (HR) length within each batter, and minimal differences between medians?

OR

-Are there big differences between batter medians, relative to the random variation in HR length within each batter?

It's the first one, so ANOVA would say there is no significant difference in average home run length between these 3 batters.



e.g. Let's try one: We want to determine which of two treatments is better at healing blisters. In addition, we want to see if either is better than a placebo. Therefore, we will randomly assign 24 patients to the three treatments (8 patients each), and measure the number of days it took for their blisters to heal. (categ var = trt. quant var = #days to healing)

Analysis of Variance results:
Data stored in separate columns.
Column means

Column	n	Mean	Std. Dev.	Std. Error
Placebo	8	10.25	2.1213	0.75
Trt. A	8	7.25	1.6962	0.59009683
Trt. B	8	8.875	1.4577	0.5153882

ANOVA table

Source	df	SS	MS	F-Stat	P-value
Treatments	2	36.083332	18.041666	5.7514234	0.0102
Error	21	65.875	3.1369047		
Total	23	101.958336			

Tukey 95% Simultaneous Confidence Intervals

Placebo subtracted from

	Difference	Lower	Upper	P-value
Trt. A	-3	-5.23213	-0.76787007	0.0075
Trt. B	-1.375	-3.60713	0.85712993	0.2877

Trt. A subtracted from

	Difference	Lower	Upper	P-value
Trt. B	1.625	-0.60712993	3.85713	0.1828

1. Look at the descriptive statistics.
   a. Which treatment took the longest, on average, to heal blisters?

   b. One of our assumptions is that the s.d. are equal. Our sample s.d. will rarely be exactly equal, so we need to check if they are close enough to be considered equal. Are any of them more than twice another? (i.e. if the biggest / smallest > 2, then they are not close enough)

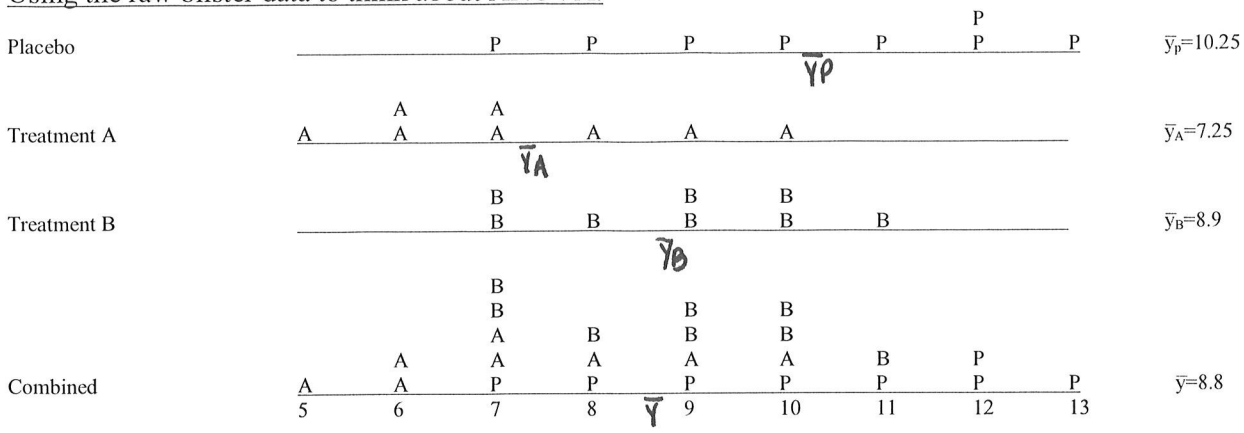
2. Hypothesis testing (look at the ANOVA table)
   H_o: $\mu_{Trt\ A} = \mu_{Trt\ B} = \mu_{Placebo}$ (i.e. the mean days to healing is the same for the 3 treatments)
   H_a: not so (at least one of the treatments take a different amount of time, on average)

   F = 5.75 df=(2,21) p-value= 0.0102 p-value $\leq$ 0.05, so reject H_o.

3. So we think at least one treatment differs from another in the average days to healing. Now we need pairwise comparisons (also called multiple comparisons), where we compare each treatment to every other treatment, and we look for differences. (look at the Tukey 95% Simultaneous Confidence Intervals)
   Placebo vs Trt. A Difference: Difference in sample means = 3 days
   Lower, Upper: 95% CI for the difference in population means (-5.23, -0.77)
   We first look for a 0 in the CI. If there's a 0, there's no significant difference in population means. There is a significant difference here, and we know from the sample means that Placebo > Trt. A. *So, we are 95% confident that Treatment A, on average, takes somewhere between 0.77 and 5.23 fewer days to heal blisters than the placebo.*
   P-value: If p-value $\leq \alpha$, there's a significant difference between Trt. A and Placebo
   Placebo vs Trt. B (p-value = 0.2877>α, so no signif. diff in means between Placebo and Trt. B)
   Trt. A vs Trt. B (p-value = 0.1828>α, so no signif. diff in means between Trt. A and Trt. B)

So the only significant difference in mean #days to healing is between Treatment A and the Placebo.

Using the raw blister data to think about ANOVA:



What do we see in this data?

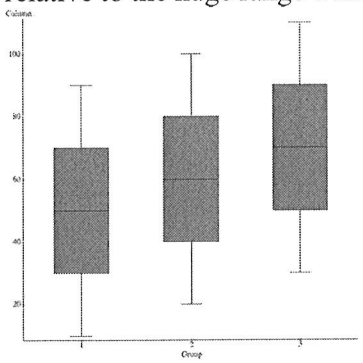
1. There are differences between the means of the groups (i.e. people who got Treatment A tended to heal faster, and people who got the placebo tended to heal more slowly).

2. There are differences within the groups (i.e. not all the people who got Treatment A were magically healed at exactly 7.25 days, but there was random variation within the group).

The whole point of ANOVA is to figure out which is the important story. Our null hypothesis is that there is no difference, on average, between the groups. Our test statistic for ANOVA is the **F-statistic**. Like other test statistics, it is comparing our sample data to our null hypothesis. In this case, a big F statistic means that there is a big difference between at least two of the population means, relative to the random variation with the groups, instead of the "no diff. in means" we claimed in H_0.

Picturing this:

Here we have little difference between medians, relative to the huge range within each group.



Analysis of Variance results:

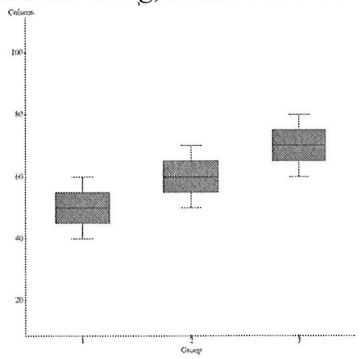
Responses: Column
Factors: Group
Response statistics by factor

Group	n	Mean	Std. Dev.
1	5	50	31.62
2	5	60	31.62
3	5	70	31.62

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Group	2	1000	500	0.5	0.6186
Error	12	12000	1000		
Total	14	13000			

Here we have the same medians, but now the difference in medians is big, relative to the small range within grps.



Analysis of Variance results:

Responses: Column
Factors: Group
Response statistics by factor

Group	n	Mean	Std. Dev.
1	5	50	7.91
2	5	60	7.91
3	5	70	7.91

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Group	2	1000	500	8	0.0062
Error	12	750	62.5		
Total	14	1750			

So we drew different conclusions about a significant difference between means, even though the means are 50, 60, and 70 in both cases, because the first one has a lot of random variation in each group, and the second does not.

NOTEThis is precisely why researchers must report BOTH means and standard deviations (or medians and IQRs). Without measures of both center and spread, we really have no way of knowing if there is an important difference between groups.

Assumptions for One-Way Analysis of Variance, and how to check the assumptions:

1. Assumption: Subjects are chosen (a) via simple random samples (SRS)
   (b) from independent groups.

   Check it: Look at the design of the study.

2. Assumption: Within each group, the response variable (quantitative var) is normally distributed.

   Check it: Create histograms. Note that the ANOVA is reasonably robust against skewness, just like the t-test, if we have no outliers and a total sample size of at least 40 (or smaller if the groups have similar distributions)

3. Assumption: The population standard deviations are the same for all groups.

   Check it: (a) Is the largest standard deviation divided by the smallest standard deviation ≤ 2? If yes, the s.d. are close enough to be considered equal because none of them are more than twice another.
   (b) If the group sample sizes are the same, then you don't have to worry about this.
   (c) There are also statistical tests (e.g. Levene's test) to check whether they are the same. A p-value > 0.05 would suggest that we do not reject the null hypothesis that they are equal.

Testing hypotheses in a one-way ANOVA

Hypotheses: Ho:

(The population means are the same.) (not $\bar{y}_1 = \bar{y}_2 = \bar{y}_3 = \bar{y}_4$)

Ha:

(At least one population mean is different from another. not: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$)

Test statistic: F = a measure of the difference between the means / a measure of spread within the groups

p-value: (df=#groups - 1, total N - #groups) *(Note: you need both d.f.!)*

Decision Rule: At the α level of significance, reject H_0 if $p \leq \alpha$. Do not reject H_0 if $p > \alpha$.

If $p>\alpha$, we are done, because we've concluded that the population group means are probably the same. If $p \leq \alpha$, we conclude that at least two population means are not equal, but we don't know which ones. Therefore we follow up a significant F statistic with **pairwise comparisons** of the means (sometimes called **multiple comparisons**), to see which are significantly different from each other.

We could do a lot of two-sample t-tests to compare each pair of means, but then we'd make Mr. Bonferroni unhappy if we weren't lowering α. Fortunately, several people have figured out ways to compare all the means pairwise for us, adjusting α at the same time. Tukey and Scheffe are the most common.

So, we do pairwise comparisons such as Tukey, Scheffe, or Bonferroni:

1) only if we need to (i.e. if we have statistically significant results, so we reject the null hypothesis of equal means and think that at least one mean differs from another)
2) to compare each group with every other group to see which means differ significantly
3) to lower α, because we're doing several comparisons of means, thereby making Mr. Bonferroni happy.

Thinking a little more about the F statistic:

From above: F = a measure of the difference between group means / a measure of spread within the groups

If at least two of the **sample means** are far from each other, we will have a big/small F-statistic. If we have a big F-statistic, that suggests that our sample data is far from our null hypothesis of equal means, so our p-value will be big/small. If our p-value is small, we will reject/fail to reject Ho.

If we have a BIG difference between **individuals** in the groups, we will have a big/small F-statistic. If we have a small F-statistic, that suggests that our sample is close to our null hypothesis of equal means, so our p-value will be big/small. If our p-value is big, we will reject/fail to reject Ho.

More StatCrunch output:

e.g. Some analysts have suggested that Spain left the World Cup earlier than expected in 2014 because they tended to be older than the other teams. It made me wonder about the average ages of the various positions. Descriptive statistics and an ANOVA table from a random sample of 2014 World Cup players follow:

Analysis of Variance results:
Responses: Age
Factors: Position

Response statistics by factor

Position	n	Mean	Std. Dev.	Std. Error
Defender	44	27.113636	3.9190406	0.59081761
Forward	28	26	3.3665016	0.63620901
Goalkeeper	26	29.923077	3.8979284	0.76444666
Midfielder	42	27.428571	4.0251821	0.62109908

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Position	3	222.42917	74.143057	5.0152382	0.0025
Error	136	2010.5637	14.783557		
Total	139	2232.9929			

Tukey 95% Simultaneous Confidence Intervals:

Defender subtracted from

	Difference	Lower	Upper	P-value
Forward	-1.1136364	-3.5313458	1.304073	0.6291
Goalkeeper	2.8094406	0.33555723	5.2833239	0.0192
Midfielder	0.31493506	-1.8425185	2.4723886	0.9813

Forward subtracted from

	Difference	Lower	Upper	P-value
Goalkeeper	3.9230769	1.1992821	6.6468718	0.0015
Midfielder	1.4285714	-1.0114215	3.8685643	0.4267

Goalkeeper subtracted from

	Difference	Lower	Upper	P-value
Midfielder	-2.4945055	-4.9901708	0.0011598	0.0502

1. Do we have one categorical variable and one quantitative variable?

2. Describe the means.

3. Let's check the assumption of equal standard deviations. Are they equal? How do you know?

4. What are the hypotheses we are testing?

5. What are the parameters we are comparing? (i.e. What are μ_{df}, μ_{fw}, μ_{gk}, μ_{mf}?)

6. Circle the F-statistic, df, and p-value for testing these hypotheses.
7. What do you conclude (α=.05)?

8. Which means differ significantly from each other (if any)? How did you decide?

9. What is the difference in mean age between these samples of goalkeepers and defenders?

10. If we were to compute the mean age of all goalkeepers and the mean age of all defenders, in what range do you expect to find the difference in means?

11. How confident are we that (10) contains the true difference in population means?

Matching. Put the letter of the correct statistical test on the lines below.

a. Chi-square goodness of fit
b. Chi-square
c. Correlation
d. Regression
e. Analysis of Variance

A boy I know loves watching all of Calvin's sports, including basketball (both men's and women's). He's got some questions, and let's tell him what kind of analysis would be most appropriate for answering them:

_____He wondered if guards, forwards, and centers scored the same number of points, on average.

_____He wondered if the amount of expensive sport drink consumed was related to the number of points scored in a game.

_____He wondered if he could predict the number of points scored in a game by the number of minutes practiced that week.

_____He wondered how the number of fans at the games changed as the temperature outside changed.

_____He wondered if shots were evenly distributed between lay-ups, mid-range shots, and 3-pointers.

_____He wondered whether there was an association between shooting hand (left/right) and whether or not he made a basket.

_____He wondered whether men and women differ in the kinds of shots they attempt.

_____He wondered, when looking at what fans wear to games, if 1/2 of the fans wear specifically Calvin clothes, 1/3 wear a Calvin color without the word "Calvin", and 1/6 wear something else.

_____He wondered if there was an association between the number of fouls a person got in a season and their score on an anger management test.

_____He divided the teams into three height groups, and wondered if the height groups played the same average number of minutes.

True/False, Multiple choice, short answer:

1. What is the definition of p-value?
   The p-value is the probability, when the null hypothesis is true, of obtaining a sample as extreme or more extreme than the one we have.

2. True/False. We reject the null hypothesis when $p \leq \alpha$.

3. True/False. Our results are statistically significant when $p \leq \alpha$.

4. True/False. All expected counts need to be 5 or more before the chi-square is valid.

5. Do marginal or conditional distributions better help us understand the relationship between two categorical variables?
   -conditional. The marginal distribution simply gives us the % in each group for one variable at a time.

e.g. Is there other life in the universe? (Gallup poll)
Rows: Year
Columns: Other life in the universe?

	Yes	No	Total
1996	385	150	535
1999	326	209	535
Total	711	359	1070

marginal
Year < 1996 / 1999

other life < Yes / No

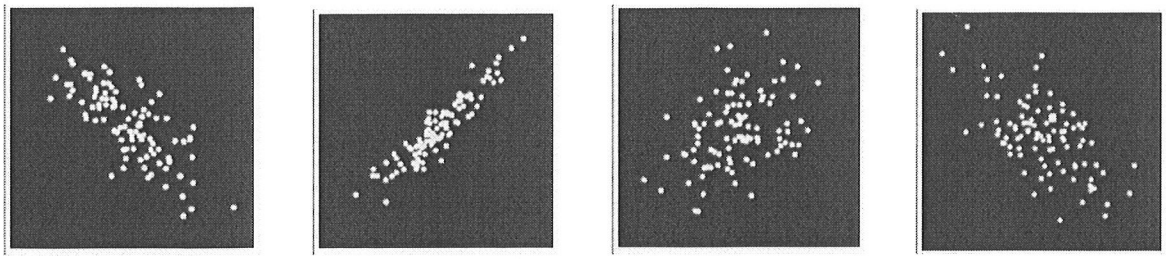
conditional
In 1996:

In 1999:

6. Matching. Which correlation coefficient goes with which scatter plot?

_____ 0.93 _____ 0.37 _____ -0.60 _____ -0.75

A B C D

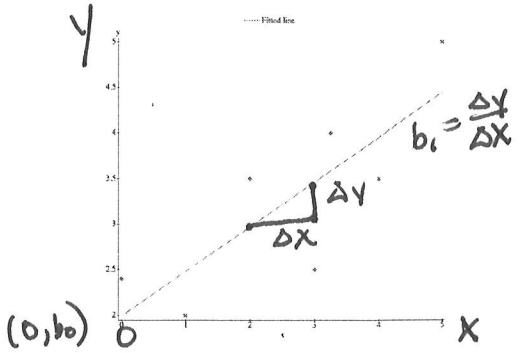


Reminders about corr. coeff: used when we want to know the nature of the relationship between two quantitative variables (i.e. is it +/-, strong/weak?); only works with a linear relationship; ranges from -1 to 1; has no units; affected by outliers—which can make the correlation stronger or weaker, depending on where the outlier is; the sign (+/-) tells the direction (positive/negative); the number tells the strength (close to 0 → weak, close to 1 or - 1 → strong).

7. What is the form of a regression equation?

$y=b_0+b_1x$

y=response
x=predictor
b_o=intercept = the point where the line hits the y-axis.
i.e. the value of y when x=0
b_1=slope = the rate of change.
i.e. as x increases by 1, y changes by the slope.



8. In regression, what name do we give the percent of variation in Y (response variable) explained by X (predictor)?

9. What are two measures of fit for a regression line?

-R^2: (definition above). Gives us an idea of how good or bad our predictions are likely to be.

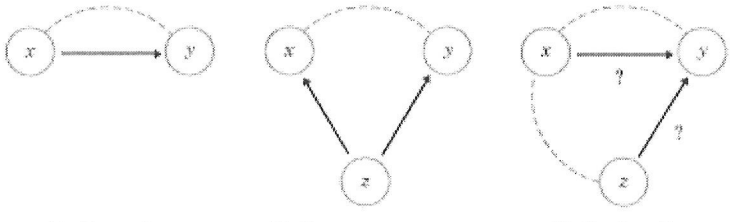
-the s_e (s.e. of the line): a measure of how far the data points (what we observed) fall from the regression line (what we predicted). The distance the points fall from the line has a normal distribution, so we can use the 68-95-99.7 rule to find the ranges in which the data falls. 68% of the data we observed falls within 1 s_e of what we predicted, 95% falls within 2 s_e, and 99.7% of the data points fall within 3 s_e.

10. What is the relationship between the correlation coefficient and R^2?

11. What is <u>extrapolation</u>?
-predicting outside our range of x's.

12. True/False. The correlation coefficient is sensitive to outliers.

13. What are some possible explanations, other than one variable causing the other, for a significant association between two variables?



(a) Causation (b) Common response (c) Confounding

14. What do we look for to see if the linear regression model is appropriate?
    --do we really have a line?
    --is the scatter around the line even along the entire length of the line?
    --are there outliers?
    --does the distance of points to the line have a normal distribution?

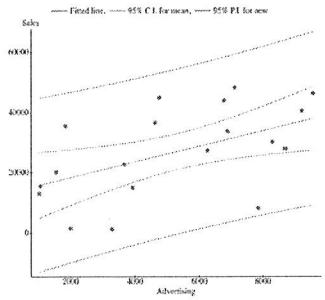
15. True/False. The null hypothesis H_0: $\beta_1 = 0$ is equivalent to the null hypothesis H_0: $\rho = 0$.

16. A study suggested that high school GPA was positively associated with the number of meals per week that families ate together. A 95% confidence interval for the slope of the regression line was (0.06, 0.16).

    a. How should we interpret such an interval in context?

    b. Based on this interval, is the linear association between GPA and family togetherness significant? Why or why not?

17. You are using a regression equation to predict sales from the amount of money spent on advertising. The confidence interval for a prediction is narrower when (select all that are correct):
    a. you are predicting sales from the average amount spent on advertising.
    b. the confidence interval for a prediction is constant, so this is a bad question.
    c. You are predicting the average sales over a whole year for a given level of advertising rather than the individual sales for one day.



18. True/False. Whether or not the ANOVA F-statistic is significant, we compare each pair of means using multiple comparisons.

19. True/False. An analysis of variance looks at the ratio of the variance between groups to the variance within groups to determine if the group means differ significantly.

20. If we reject the ANOVA null hypothesis, why do we follow up with pairwise comparisons?
    -if we reject the null hypothesis, all we "know" (there's always a chance we've made a Type I error because our sample is flawed) is that at least one mean is different from another. We use pairwise comparisons to look at the means pairwise (two at a time), to see which means differ from each other, and which means we think are the same. In addition, Tukey or Scheffe will lower α for us, since we're doing many comparisons.

21. What assumption does every inferential statistical test have in common?

22. Give an example of the null and alternate hypotheses for a chi-square goodness of fit test.

23. Give an example of the null and alternate hypotheses for a chi-square test.

24. Give an example of the null and alternate hypotheses for a regression coefficient.

25. Give an example of the null and alternate hypotheses for a correlation coefficient.

26. Give an example of the null and alternate hypotheses for analysis of variance.

27. We wonder, among 2-child families, if ¼ have 2 girls, ¼ have 2 boys, and ½ have one child of each gender (which is what you'd expect if boys and girls are equally likely). The National Health Interview Survey (NHIS) examined 4288 two-child families and found the following results:

Children	count	percent
2 girls	952	22.2
2 boys	1133	26.4
1 of each	2203	51.4

a. What are the appropriate hypotheses?

b. What do you conclude (α=.05)?

$\chi^2 = 18.53$, df=2, p<.0001

28. The following table shows the Myers-Briggs personality preference and professions for a random sample of 2408 people in the listed professions.

Contingency table results:

Rows: Personality
Columns: Profession

Cell format
Count
(Row percent)
(Column percent)
(Expected count)

	Clergy	Lawyer	M.D.	Total
Extrovert	308	112	667	1087
	(28.33%)	(10.3%)	(61.36%)	(100%)
	(57.68%)	(41.33%)	(41.61%)	(45.14%)
	(241.05)	(122.33)	(723.61)	
Introvert	226	159	936	1321
	(17.11%)	(12.04%)	(70.86%)	(100%)
	(42.32%)	(58.67%)	(58.39%)	(54.86%)
	(292.95)	(148.67)	(879.39)	
Total	534	271	1603	2408
	(22.18%)	(11.25%)	(66.57%)	(100%)
	(100%)	(100%)	(100%)	(100%)

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	2	43.556144	<0.0001

a. May we use the chi-square in this situation? Why or why not?

b. What is the actual number of clergy who were extroverted?

c. What is the expected number of clergy who were extroverted, if H_o is true?

d. What percent of people in the sample were extroverts?

e. If there is no relationship between profession and personality type, what percent of clergy would you expect to be extroverts?

f. If there is no relationship between profession and personality type, what percent of lawyers would you expect to be extroverts?

g. Among clergy, what percent were extroverts?

Among lawyers, what percent were extroverts?

Among M.D.'s, what percent were extroverts?

Do you think there is a relationship between profession and personality type?

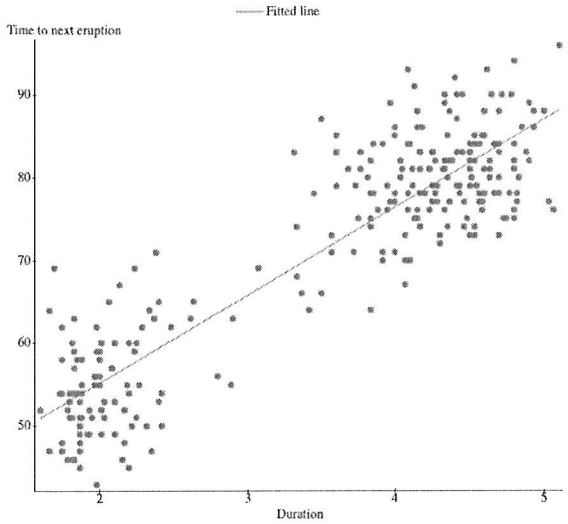
h. What are the null and alternate hypotheses for this problem?

i. Circle the chi-square statistic, df, and p-value.

j. What exactly does the p-value mean in this situation?

k. Is profession independent of personality type (α=.05)? Why do you think so?

29. Old Faithful, located in Yellowstone National Park, may be the world's most famous geyser. The duration (in minutes) of an Old Faithful eruptions can be used to predict the times (in minutes) until the next eruption. A scatter plot and regression output follow:



Simple linear regression results:
Dependent Variable: Time to next eruption
Independent Variable: Duration
Time to next eruption = 33.987808 + 10.611776 Duration
Sample size: 270
R (correlation coefficient) = 0.89606971
R-sq = 0.80294093
Estimate of error standard deviation: 6.0035495

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	33.987808	1.1812171	≠0	268	28.773548	<0.0001
Slope	10.611776	0.3211272	>0	268	33.045398	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	39358.466	39358.466	1091.9983	<0.0001
Error	268	9659.4187	36.042607		
Total	269	49017.885			

Predicted values:

X value	Pred. Y	s.e.(Pred. y)	95% C.I. for mean	95% P.I. for new
3	65.823136	0.39882599	(65.037905, 66.608367)	(53.976963, 77.669309)

a. Which variable is the explanatory variable? Which is the response?

b. What information does the scatter plot give about the relationship between duration of eruption and time until the next eruption?

c. What is the slope, and what does it mean?

d. Does time to next eruption increase significantly with increased duration of eruption (at α=.05)? How do you know?

e. Give a 95% confidence interval for the slope. How do we interpret this CI?

f. What percent of the variability in time to next eruption is explained by duration? Is eruption duration a good predictor of time to next eruption?

g. What is the correlation coefficient? How is this computed?

h. What is the standard error of the regression line (s_e)? What does this number mean?

i. Predict the time to the next eruption if the duration of the most recent eruption was 3 minutes.

j. On the output, under "predicted values for new observations", are two intervals for predicted time when the duration is 3 minutes. Which interval is for the prediction of the average of several eruptions? Which interval is for the prediction of a single observation? How do you know?

k. What affects the width of the CI when we are making prediction?

30. To detect the presence of harmful insects in farm fields, researchers can put up boards covered with a sticky material and examine the insects trapped on the boards. Which colors attract insects best? Experimenters placed 8 boards of each of four colors at random locations in a field of oats and measured the number of cereal leaf beetles trapped.

a. Describe the means.

b. Can we consider the standard deviations to be equal? Why or why not?

c. What other assumptions are we making?

d. Write the hypotheses being tested.

e. Circle the F-statistic, the df, and the p-value.

f. Can you conclude that, for at least one color, the mean number of beetles trapped is different (α=.01)?

g. Which means differ significantly from each other according to Tukey?

h. What is the difference in sample means between Green and Blue? What about population means?

Analysis of Variance results:
Data stored in separate columns.
Column means

Column	n	Mean	Std. Dev.	Std. Error
Blue	8	14.875	4.53	1.597
Green	8	31.125	5.32	1.884
White	8	16.125	3.18	1.125
Yellow	8	47.125	5.74	2.030

ANOVA table

Source	df	SS	MS	F-Stat	P-value
Treatments	3	5495.375	1831.7916	79.705	<0.0001
Error	28	643.5	22.982143		
Total	31	6138.875			

Tukey 95% Simultaneous Confidence Intervals
Blue subtracted from

	Difference	Lower	Upper	P-value
Green	16.25	9.705485	22.794516	<0.0001
White	1.25	-5.294515	7.794515	0.9532
Yellow	32.25	25.705484	38.794514	<0.0001

Green subtracted from

	Difference	Lower	Upper	P-value
White	-15	-21.544516	-8.455485	<0.0001
Yellow	16	9.455485	22.544516	<0.0001

White subtracted from

	Difference	Lower	Upper	P-value
Yellow	31	24.455484	37.544514	<0.0001

Wrapping up

Approaching a journal article

1. **What led up to the study?**
   a. What is the goal of the study?
   b. Who paid for it? (Do they have a financial interest in the results?)
   c. Who carried it out? (Are they reputable?)

2. **Who are the study participants?**
   a. How did the researchers pick their participating people, animals, or inanimate objects? Is the sample they are using a good one? Did they choose the sample at random, or assign subjects to treatments at random? Is the sample representative of the population they hope to learn something about? (no undercoverage, ...)
   b. Let's say it again, because this is so often where things fall completely apart for a study. How did the researchers pick their participating people, animals, or inanimate objects? Is the sample they are using a good one? i.e. Is it representative of the population they hope to learn something about?
   c. What is the sample size?
   d. If this involves humans, what was the rate of nonresponse?

3. **Carrying out the study:**
   a. What variables are they studying? Are there any problems with defining or measuring the variables of interest? (e.g. they hoped to measure 3 levels of smiling but couldn't distinguish between "no smile" and "closed mouth smile")
   b. Are there confounding variables that could make conclusions difficult or impossible (e.g. is the group getting the treatment different from the group getting the placebo in important ways, such as age, health status, . . .)?
   c. Consider the setting and wording of any survey. Is response bias being introduced?

4. **Statistical analysis**
   a. Do they give descriptive statistics and graphs, so you have a clear idea of what is going on? Do they have measure of center AND spread for quantitative variables?
   b. Did they choose appropriate analyses for their variable types (categorical/quantitative) and hypotheses? (e.g. regression to predict one quantitative variable from another)
   c. Did they report actual p-values, or at least several levels of significance (e.g. $p \leq 0.05$, $p \leq 0.01$, $p \leq 0.001$, etc.) so you have a ballpark idea of the actual p-value?
   d. If they are doing many hypothesis tests, did they choose a lower alpha, so they aren't getting as many statistically significant results just by chance (making too many Type I errors)?

5. **Conclusions**
   a. Are they confusing correlation with causation?
   b. Are their conclusions justified, based on #2, #3, and #4 above? Do their conclusions make sense? Do they have practical significance?
   c. Did they achieve the goals of the study?

Analyzing a data set

1. **Is it worth doing?**
   a. Is it important?
   b. How were the data obtained? Are they high quality and worth analyzing?

2. **Look at the data.**
   a. List each variable and decide if it is categorical or quantitative
   b. Describe the variables with graphs and numbers.
      1. Categorical: bar graph, pie chart, counts, %
      2. Quantitative: Histogram (does it have a normal distribution?), mean&sd or 5#summary
      3. Look for problems in the data set. i.e. Look for outliers, incorrect values, missing data. Correct as many problems as you can. Then redo the descriptive statistics.

3. **Are there any relationships between the explanatory and response variables?**
   a. For each pair of variables, decide on the appropriate analysis, hypotheses, etc.

	Categorical	Quantitative
Categorical	2 sample z-test for proportions chi-square	2 sample t-test ANOVA Side-by-side boxplots
Quantitative		correlation regression scatter plot

   b. Check your assumptions before carrying out the test.

4. **Are there relationship between other variables (e.g. demographic variables) and either the explanatory or response variables?**
   These are possible confounders, and you may want to include them in:

   multiple regression: predicting a quantitative variable from several predictors
   2-way ANOVA: like the ANOVA we did, only with two categorical variables
   Analysis of Covariance (ANCOVA): ANOVA, while adjusting for an extra quant. variable

5. **Reporting on your results** (also look at the section on Approaching a Journal Article)
   a. Carefully describe how your sample and data were obtained.
   b. Describe your data with graphs and numbers (for quantitative variables, include measures of center and spread)
   c. Include the actual p-values of the statistical tests you did.
   d. Choose an alpha lower than 0.05 as your cut-off if you are doing many hypothesis tests.
   e. When writing up your conclusions:
      1. Be careful to only infer your results to the correct population.
      2. Be careful not to confuse association with causation.
      3. Be careful to look for alternate explanations (possible confounders)

e.g. A study on hospital-acquired infections (the sample was randomly selected hospitals):

1. Which variables are categorical?
   Avg. length of stay
   Avg age of the patients
   Infection risk (estimated probability of getting an infection while at the hospital)
   #beds
   Medical school affiliation (Y/N)
   Region of the country (NE, NC, S, W)
   Average #patients
   #nurses
2. What graph would you use to show the distribution of the categorical variables?
   a. bar chart b. scatter plot c. side-by-side boxplots d. histogram
3. What graph would you use to show the distribution of the quantitative variables?
   a. bar chart b. scatter plot c. side-by-side boxplots d. histogram
4. What graph and analysis would you use to look for an association between the average length of stay and infection risk?
   a. Scatter plot / correlation b. side-by-side box plots / ANOVA c. chi-square
5. What graph and analysis would you use to predict the infection risk from the avg. length of stay?
   a. Scatter plot / regression b. side-by-side box plots / ANOVA c. chi-square
6. What analysis would you use to look for a relationship between whether or not there is a medical school affiliation and the region of the country?
   a. correlation b. regression c. ANOVA d. 2-sample t-test e. chi-square
7. What graph and analysis would you use to compare the regions of the country on avg. infection risk?
   a. Scatter plot / regression b. side-by-side box plots / ANOVA c. chi-square

Big Data

"Big data promises to be transformative." (Big Data Imperatives, p. 11).
"Data are becoming the new raw material of business. Economic input is almost equivalent to capital and labor." The Economist (2010)
"Information will be the 21st century oil." The Gartner Company (2010)

Sources:

1. Big Data Imperatives (by Mohanty, Jagadeesh, Srivatsa. Apress, 2013)
2. Big Data: Related Technologies, Challenges, and Future Prospects (by Chen, Mao, Zhang, Leung. Springer Briefs in Computer Science, 2014.)
3. The big dangers of 'big data' (by Konstantin Kakaes, CNN, Feb. 2, 2015)
4. Math is racist: How data is driving inequality. (by Aimee Rawlins, 2016).
5. Weapons of Math Destruction (by Cathy O'Neil, Broadway Books, 2016).

What is Big Data? 3, 4, or 5 V's.

1.

--Businesses have left paper behind and become digital
--People contribute ENORMOUS amounts of data (Facebook, YouTube, tweets, blogs, searching the internet, shopping, MyFitnessPal, ...)
--Medical data (medical tests, keeping track of BP at home, doctor's visits, . . .)
--Environmental data (thousands of sensors recording temperature, pollution, . . .)
--Location recording devices (GPS, smart phones, cameras, ...)

2.

--Many kinds of data (traditional data, webpages, video, audio, tweets/texts)

3.

--data are being generated at enormous speed. In fact, they are being generated so quickly that at times companies can't even afford to store them, but have to analyze the data as they fly by.

4.

--High value, but you need to look through a lot of data to get the value.

5.

--Is this data trustworthy?

A couple of ways Big Data has been used in the past:

1. Amazon. "Recommendations for you." Dynamic pricing.

2. Facebook. –social experiments
   --figuring out things about users & selling targeted ads (Cambridge Analytica, businesses)
   --propagating similar but more extreme stories to get more clicks

3. Google --massive information collector, for good and ill.
   --good: figured out the 2009 flu pandemic because of 45 search terms.

4. Elections --politicians are targeting voters with the ad most likely to get the desired result (vote, donate, . .)

Potential for the future: The McKinsey Global Institute predicted (in *Big data: The next frontier for innovation, competition, and productivity,* by Manyika, et al, 2011) that Big Data could

--save the US medical industry over 300 billion dollars (8% of US healthcare expenditures) by figuring out the best treatment for a given patient and determining public health policies that will have the greatest effect.

--could increase retailers' profits by more than 60% by figuring out who buys things, when, and why

--save the governments of Europe over 100 billion euros through improved efficiency

--help in company logistics by aiding in route optimization and increasing efficiency

--be critical in aiding financial services, utilities, and media

In short, there's not much of the world that will be unaffected by Big Data.

Cautions:

1. Bad use of data can be worse than no data at all.

2. Falsified/incorrect data

3. What data won't tell us (at times we need informed judgments that use not only data, but also our values, the context, ...).

4. Big data can propagate racism&sexism, and discourage movement in economic class. "Any algorithm can – and often does – simply reproduce the biases inherent in its creator, in the data it's using, or in society at large." (Leigh Alexander, 2016) Therefore, although Big Data attempts to be objective, it does not always succeed. By using someone's name, zip code, education, credit score, and grammar, people are profiled. Big Data can propagate bias instead of remove it.

Rawlins say, "Denied a job because of a personality test? Too bad-the algorithm said you wouldn't be a good fit. Charged a higher rate for a loan? Well, people in your zip code tend to be riskier borrowers. Received a harsher prison sentence? Here's the thing: Your friends and family have criminal records, too, so you're likely to be a repeat offender. . . . "

O'Neill says that these algorithms are opaque, unregulated, and unaccountable, but they have enormous influence.

--mostly hurts the poor (higher mortgage and insurance rates, predatory for-profit college ads, scheduling software at restaurants, ...)

--but you are not immune!

--You voluntarily give loads of data through your phone, internet searches, websites visited, apps we use, ... Many companies collect that data and create scores for potential landlords, stores, insurance rates, ...

--Are you looking for a job? Hirevue uses a computer or cellphone to analyze facial movements, word choice, intonation, . . . to come up with an employability score.

So, Big Data has tremendous potential and is already transforming the world. But it needs to be constantly monitored and re-evaluated. We need to make sure that the goal is helping people, not taking advantage of them.

The analyses of Big Data require complex algorithms, so Big Data is at the intersection of statistics and computer science.

Which statistics does it use?

1. correlation analysis looks for relationships between variables. They are often not interested in causation at all—they are just looking for correlations.
2. Multiple regression analysis to predict one quantitative variable from several predictors
3. Cluster analysis groups individuals together using several variables (e.g. who shops at certain stores and why?)
4. Factor analysis groups variables together (a group of similar variables is called a factor).
5. Experiments (called A/B testing or bucket testing). E.g. Facebook, Amazon

More and more decisions in business, education, health care, and government are being driven by data—and having some understanding of graphs, descriptive statistics, and the analyses we've talked about in this course will only help you!