Contingency (two-way) Tables **Wed., 11/1/2023**

GOAL: to know how to describe the variables in a table using marginal distributions and look for a relationship between variables in a table using conditional distributions.

Homework #13 (no quiz): due Friday, Nov. 3, 2023, at 11:59 p.m.

For this first section, we will be working entirely with categorical data. When we have one categorical variable, we find the percent in each category and show the distribution with either a pie chart or a bar chart. When we have two categorical variables, we put the data in a two-way table. We describe each variable separately, and then look for a relationship between variables. For example, if we were curious as to whether or not there was an association between age group and having cavities, we could come up with the following table:

	Cavities		
Age Group	Yes	No	Total
6-9	672	708	1380
10-12	1711	521	2232
13-14	1579	263	1842
15-16	1063	109	1172
Total	5025	1601	6626

These are data from Newburgh, NY, collected in 1954-1955. Fluoride was added to the drinking water beginning May 2, 1945, and they wondered if the children who had fluoride for a bigger part of their lives would have fewer cavities. (So the youngest age group had fluoride their entire lives, and the 15-16 year olds had fluoride for the last 9-10 years.)

Age group is the variable.

Cavities (yes/no) is the variable.

We first look at the variables one at a time: the Marginal distributions. We want to know, e.g., what percent of people (overall) have cavities. These are called the marginal distributions because we get the numbers we need from the margins: the Totals.

We often turn the numbers in which we're interested into percents. For example, to get the marginal distribution of Cavities (yes/no) in percents, we divide each total by the grand total (6626 children).

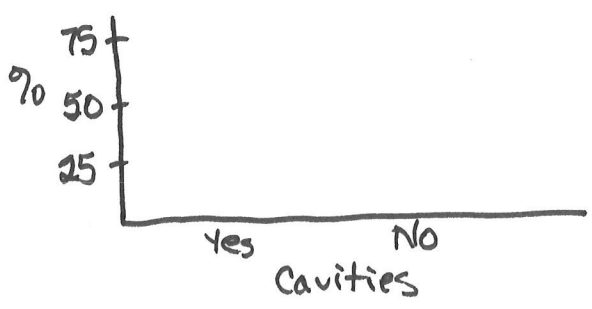
Marginal distribution of cavities:

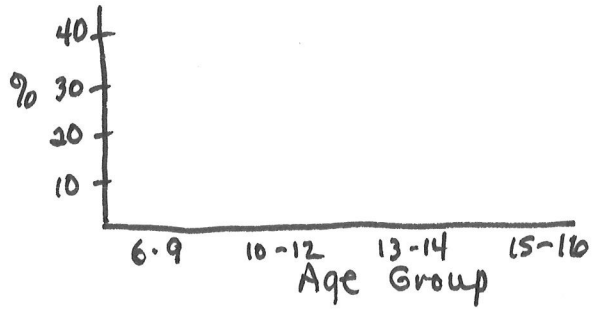
Cavities	Percent Total # with cavities/grand total /6626=
Yes	1601/6626=24.16%
No	

Marginal distribution of age group:

Age group	Percent Total # in an age group/grand total /6626=
6-9	2232/6626=33.69%
10-12	1842/6626=27.80%
13-14	1172/6626=17.69%
15-16	

Turn this into a bar graph or pie chart (the graphs for categorical data) to see this graphically.





To look for a relationship between age and cavities (instead of just determining what percent fall in a certain category) we need to look at the numbers in the table (rather than the margins).

e.g. What percent of children ages 6-9 have cavities?
Look at the ages 6-9 row only. There are 1380 children in this age group. Repeat what you did for the marginal distributions, but do it only for this row.

%children ages 6-9 with cavities= #children ages 6-9 with cavities / #children ages 6-9 =

%children ages 6-9 without cavities= #children ages 6-9 without cavities / #children ages 6-9 =

If we do this for every age group, it is called a conditional distribution. It is called conditional because we are looking at the distribution of a variable, conditional on limiting the distribution to some group of another variable. Basically, we are taking a table row by row or column by column and looking for differences in the percent that have a certain characteristic. (i.e. does the % with a certain characteristic CHANGE from group to group? If so, we think there's a relationship between those two variables.)

What is the conditional distribution of cavities, given that a person is in a certain age group? i.e. For each age group separately, what percent have cavities? What percent does not have cavities?

Cavities	Ages 6-9	Ages 10-12	Ages 13-14	Ages 15-16
Yes	672/1380=48.7%	1711/2232=76.7%	1579/1842=85.7%	1063/1172=90.7%
No	708/1380=51.3%	521/2232=23.3%	263/1842=14.3%	109/1172=9.3%

If the distribution of one variable is the same for each level of the other, the variables are **independent** (have no association). In this case, if the percent with cavities is the same for all age groups, there is no association between age group and having cavities. (i.e. No matter what your age is, you have the same probability of having cavities.)

What is happening to the percent with cavities as age increases? Staying the same / Increasing / Decreasing

So, based on the conditional distributions, does there appear to be a relationship between age group and the presence/absence of cavities, or are they independent?

There appears to be a relationship (they are dependent). The percent with cavities is not staying the same as age increases, but the percent with cavities increases with increasing age.

BUT—age is confounded with fluoride. Confounding happens when the effects of two variables can't be separated. In this case, we can't separate the effects of age and fluoride, since older kids had less fluoride during their lifetime, and the youngest kids had fluoride their entire lives. i.e., younger=more fluoride, and older = less fluoride. So, is the increase in the percent with cavities due to age (you'd expect older kids to have more cavities because they've had their teeth longer) or due to fluoride (does ↑fluoride ↓cavities?), or is it due to both, or something else entirely (e.g a new toothpaste introduced during the older kids' lifetime?). We can't tell, since the effects of age and fluoride are intertwined.

You can also compute the conditional distribution of age group, given whether or not there are cavities.

i.e. Among the 5,025 with cavities, what percent falls in each age group? (*672 / 5,025 = 13.4% are 6-9, . . .*) Among the 1,601 who do not have cavities, what percent falls in each age group? (*708 / 1,601 = 44.2% are 6-9, . . .*)

Look at both conditional distributions and determine which helps you to most clearly see the relationship (if any) between the two categorical variables. In general, if you think one variable may be causing the other, check the distribution of the second variable for each group of the first (possibly causal) variable.

One more example: smoking status and LDL (from the Healthy Women Study in Pittsburgh, PA)

--Our two categorical variables are smoker (Y/N) and LDL (below the median / above the median). (LDL is Low Density Lipids, a quantitative variable, so I split it at the median to make it categorical. Low LDL is healthier.)

Contingency table results:

Rows: LDL
Columns: Smoker

Cell format
Count
(Row percent)
(Column percent)

Smoker?

LDL

	Yes	No	Total
Below Median	62	211	273
	(22.71%)	(77.29%)	(100%)
	(38.75%)	(55.38%)	(50.46%)
Above Median	98	170	268
	(36.57%)	(63.43%)	(100%)
	(61.25%)	(44.62%)	(49.54%)
Total	160	381	541
	(29.57%)	(70.43%)	(100%)
	(100%)	(100%)	(100%)

Chi-Square test:

Statistic	**DF**	**Value**	**P-value**
Chi-square	1	12.466928	0.0004

How to read that table:

1. All you really need to know is that the top number in each box is the actual count (e.g. there's a grand total of 541 women. There are 62 women who smoked and had low LDL). The rest of the numbers can be easily computed.
2. The second number in each box is the Row percent. That gives you the percent of that row group that falls in that category. In this case, our rows are "Below median" and "Above median," so the row percent gives the percent of each LDL group that are smokers or non-smokers. E.g. We have 273 women with LDL ≤ median, and of those 273, 62 are smokers. 62/273 = 22.71%
3. The third number in each box is the Column percent. This is just like row percents, only for the columns. E.g. We have 160 smokers. Among the smokers, 98/160 = 61.25% have high LDL.

Let's try some:

1. How many women were non-smokers?

2. How many smokers had high LDL?

3. What percent of the women, overall, were smokers?

4. What percent of the women smoked and had low LDL?

5. Of the women who smoked, what percent had low LDL?

   Of the women who did not smoke, what percent had low LDL?

6. If those conditional percents are similar, LDL is independent of smoking. If they are different, smoking and LDL are associated. Do you think smoking and LDL are associated?

7. Can we say that smoking CAUSES high LDL?

Simpson's paradox: The effect of lurking variables can strongly affect the relationship between two variables. Simpson's paradox is a situation in which an association that holds for all of several groups reverses direction when the data are combined to form a single group. This is because combining dissimilar groups loses important information, so it can give misleading results and lead to unfair comparisons.

e.g. Graduate admissions at Berkeley in the 1970's. Berkeley was concerned about getting sued for sex discrimination.

	Males	Females	
Program	Percent Admitted	Percent Admitted	Overall % Admitted
A	511/825 = 61.9%	89/108 = 82.4%	600 /933 = 64.3%
B	352/560 = 62.9%	17/25 = 68.0%	369 /585 = 63.1%
C	137/407 = 33.7%	132/375 = 35.2%	269 /782 = 34.4%
D	22/373 = 5.9%	24/341 = 7.0%	46 / 714 = 6.4%
Total	1022/2165 = 47.2%	262/849 = 30.9%	1284/3014 = 42.6%

Overall, the percent of males admitted was:

Overall, the percent of females admitted was:

So, should Berkeley be sued for discriminating against females?

Now let's compare males and females on the percent admitted to each program. Which sex had the higher percent admitted for:

Program A	Males	Females
Program B	Males	Females
Program C	Males	Females
Program D	Males	Females

So:

Which sex had a higher admittance rate overall?	Males	Females
Which sex had a higher admittance rate for each program?	Males	Females

This is Simpson's Paradox: we thought Berkeley was discriminating against females, since it appeared that male applicants were more likely to be accepted. But once we took into account the lurking variable of Program, our conclusions completely flipped, and we realized that female applicants were more likely to be accepted into each program (so maybe males are being discriminated against!).

How can this reversal happen?

Now it's time to look at the percent of applicants admitted to each program overall. Almost 2/3 of applicants to Programs A and B were admitted (64.3% and 63.1%). Only about 1/3 of applicants were admitted to Program C (34.4%). Notably, only 46/714 = 6.4% of applicants were admitted to Program D.

Now look to see where males and females tended to apply.

Among the 2,165 males:		Among the 849 females:	
	A 825/2165 = 38.1%		A 108/849 = 12.7%
	B 560/2165 = 25.9%		B 42/849 = 5.0%
	C 407/2165 = 18.8%		C 375/849 = 44.2%
	D 373/2165 = 17.2%		D 341/849 = 40.2%

So the females were much more likely to apply to Program C (where only 34.4% of all applicants were admitted) and Program D (where only 6.4% of applicants were admitted). The males were more likely to apply to Programs A and B, where almost two-thirds of applicants were admitted.

So a lower percent of females were admitted to Berkeley, not because of discrimination, but because they were much more likely to apply to the programs that had low admittance rates.

This shows the danger of aggregating (combining tables by adding their corresponding cells). Keep your thinking caps on! Any observed association should always be examined critically for possible lurking variables.

The Chi-square (χ^2) test

Wed., Nov. 1, 2023

GOAL: to know when to use the chi-square test (both Goodness of Fit chi-square and regular chi-square), to understand the ideas behind the chi-square test, and to be able to interpret chi-square output.

Homework and Quiz #14: due Wednesday, 11/8/2023 at 11:59 p.m.

The chi-square test is ONLY used with categorical data, and we are interested in the count in each category or combination of categories. We use it when:

1. Goodness of fit test (one categorical variable, and we wonder if the percent in each category matches some preconceived idea of the percent in each category) e.g. does a package of M&M's have 13% brown, 14% yellow, 13% red, . . . as the company claims, or are percentages of each color significantly different than that?

2. Regular old chi-square (two categorical variables, and we wonder if there's an association between the two categorical variables / if the percent in each group of one of the categorical variables changes depending on the other variable.
   a. e.g. do two or more groups have a similar distribution on another categorical variable. e.g. do three treatment groups (placebo, herbal remedy, pill) have the same outcomes (depression returned/no depression)?
   b. e.g to see if there is a relationship between two categorical variables from one group. e.g. is there a relationship between handedness (left/right) and type of major (STEM / music or art / humanities / business) among college students?

For which of the following could we use a chi-square?

1. Is there a relationship between smoking (Y/N) and personality (extrovert/introvert)?

2. Are children more likely to be born on certain days of the week, or are approximately 1/7 of babies born each day?

3. Is there a difference between Hope and Calvin in average IQ?

4. Is there an association between weight and blood pressure?

5. Is there a difference between the genders in their favorite breed of dog?

For all of these chi-square tests, we assume:

1) SRS
2) Independent observations
3) at least 5 individuals expected in every cell (we're checking the count we would have expected if Ho is true, not the count we actually saw in our sample).

Let's start with the Goodness of fit test (looking at one categorical variable at a time)

We'll use the same 4 step hypothesis testing procedure (hypotheses, test statistic, p-value, and decision), only now we'll be using a chi-square statistic as our test statistic.

Our four step hypothesis test procedure is as follows for the goodness of fit test:

1) hypotheses.
   H_o: we specify the percent we expect to see in each group if there is goodness of fit (what the percentages in each group would be if they matched our comparison group/theory).
   H_a: not so (the percent in each group is somehow different than what we said in H_o)

2) χ^2 (computed by the computer)
3) P-value (determined by the computer)
4) Decision (if the p-value $\leq \alpha$, we reject H_o)

For example: (note: Michigan stopped collecting race/ethnicity data on March 7, 2021)
As of March 2, 2021, there were 15,398 deaths from COVID-19 among the 4 main ethnic groups in Michigan: Asian (177), Black (3,609), Latino (502), and White (11,110). In Michigan, including only those four ethnicities, 3.3% of residents are Asian, 14.3% are Black, 3.7% are Latino, and 78.7% are White.

We have one categorical variable (ethnicity), with 4 groups (Asian, Black, Latino, and White), and we wonder if there is goodness of fit between the percent of deaths that come from each ethnicity, and the percent of each ethnicity in the state's population. i.e. Do we see the same percent of each ethnicity among the deaths as we do in the population?

We need the Goodness of Fit chi-square test.

We have the observed counts (of 15,398 deaths: 177 Asian, 3,609 Black, 502 Latino, and 11,110 White).

Now we need the expected counts. We know 3.3% of residents are Asian, 14.3% are Black, 3.7% are Latino, and 78.7% are White. If deaths reflected the population, 3.3% of the 15,398 deaths would be Asian, 14.3% Black, 3.7% Latino, and 78.7% White.

3.3% of 15,398 = 0.033 x 15,398 = 508.1 Asian deaths expected, if the ethnic pattern of deaths follows the ethnic pattern in the population.

14.3% of 15,398 = 0.143 x 15,398 = 2,201.9 Black deaths expected, if . . .

3.7% of 15,398 = 0.037 x 15,398 = 569.7 Latino deaths expected, if . . .

78.7% of 15,398 = 0.787 x 15,398 = 12,118.2 White deaths expected, if . . . .

Checking assumptions: All of those expected counts are greater than 5. This is not a SRS, since it's all the deaths so far, but we will treat this sample as a representative sample of the population of Michiganders. (This may be a questionable assumption.) We will also assume that all the deaths are independent, which again may be questionable if family members with similar health challenges infected each other. But we'll boldly go on.

So far we have:	Obs. count of deaths	% of deaths	Known % in population	Exp. count of deaths
Asian	177	1.1%	3.3%	508.1
Black	3,609	23.4%	14.3%	2,201.9
Latino	502	3.3%	3.7%	569.7
White	11,110	72.2%	78.7%	12,118.2

Let's do some hypothesis testing using our 4-step method, to see if this is a statistically significant difference.

1. Hypotheses:

H_o: p_{Asian}= p_{Black} = p_{Latino} = p_{White} =

(each ethnicity makes up the same % of the deaths as they do in the population)

H_a: not so

(we don't offer a specific alternative. We just say that if we reject Ho, we conclude that the percent of deaths from each ethncity don't match—somehow—the percent of each ethnicity in the population)

2. Test statistic: Remember? Our z-statistics and t-statistics were always measures of how far our sample data fell from the null hypothesis.

e.g. $t = (\bar{y} - \mu_o) / (s/\sqrt{n})$ How far is our sample mean ($\bar{y}$) from our hypothesized mean (μ_o)?

$z = (\hat{p} - p_o) / \sqrt{(p_o*(1- p_o)/n)}$ How far is our sample % ($\hat{p}$) from our hypothesized % (p_o)?

The chi-square statistic is no different. It's a measure of distance: how far is our sample from what we hypothesized about the population?

The chi-square test statistic is: $\chi^2 = \sum(o-e)^2/e$, where o=observed count in a cell (what we see in our sample), and e=expected count in a cell if H_o is true (what we expected to see if Ho is true), with df=#categories – 1.

$\chi^2 = \sum(o-e)^2/e = (177 - 508.1)^2/508.1 + (3,609 - 2,201.9)^2/2,201.9 + (502 - 569.7)^2/569.7 + (11,110 - 12,118.2)^2/12,118.2 = ??$

$= 216 + 899 + 8 + 84 = \mathbf{1,207}$

3. p-value. We'll let Statcrunch find the p-value (the χ^2 table looks just like the T-table, so we could have done this by hand very easily):

$\chi^2 = 1,207$ df=3 p-value < 0.0001

4. Decision: No matter what our level of significance (α) is, our **p-value $\leq \alpha$, so we reject the null hypothesis** of goodness of fit and conclude that the percent of deaths for each race does not match the percent of each race in the population. By comparing the observed and expected counts, and/or by seeing where the big contributors to the χ^2 statistic lie, it's very easy to see where the differences lie (and, as it should, this has very much caught the attention of state and national leaders):

So, to summarize, H_o spells out the percent we expect to see in each group if there's goodness of fit, and H_a just says that the percentages are somehow different from that ("not so"). The computer will calculate the chi-square statistic and p-value for us. We will look at the p-value and decide whether or not to reject H_o. If we decide to reject Ho, we'll look at the number we observed in each category, and the number we would have expected to see in each category if H_o was true, and see where the goodness of fit broke down.

Looking at data two categorical variables at a time

So far, we merely described (based on a sample) the relationship between two categorical variables. We first put the data into a table (called a two-way table). We then looked at the marginal distributions, which is the distribution of each variable ignoring the other. We next examined the conditional distributions. The conditional distributions are key to determining if there is a significant association between two categorical variables, and what this association looks like.

The conditional distributions, recall, are the distribution of one variable for each level of the other variable. If these distributions are the same for each level of the second variable, we would conclude there was no association (they are independent). If these distributions differ for the different levels of the second variable, we might conclude that there is an association.

When the conditional distribution suggests there may be an association, we want to see if that association is *statistically significant* (greater/stronger than what we would expect to see just by chance, i.e. sampling variation). To do this, we will do some hypothesis testing.

Hypothesis testing will be the same four step procedure as before. For our hypotheses, we're making a claim about what is true in the population, and H_o is claiming that there is nothing going on (no association, no difference, ...). H_a is saying there is something going on (an association, a difference, ...). The computer will find the chi-square statistic and p-value for us, and then we'll make a decision about the null hypothesis based on our p-value (remember, a low p-value means that it'd be very unlikely to see a sample like ours if H_o is true, so we think H_o is probably not true).

Hypothesis testing when you have two or more categorical variables:

1. General hypotheses for the chi-square test:

   H_o:

   H_a:

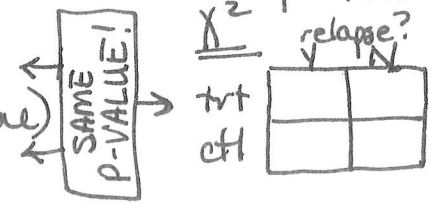
2. Look at the value of the Pearson chi-square statistic (our test statistic).
3. Find the p-value.
4. Draw a conclusion. As before, at the α level of significance, reject H_o if $p \leq \alpha$, and do not reject H_o if $p > \alpha$. (Note that if we reject H_o, we do not know the nature of the relationship. Look at the conditional percents to get an idea.)

Note that for a 2x2 table, you can use either the chi-square test or the 2 sample z-test for proportions (H_o: $p_1=p_2$). It is generally better, in this case, to use the 2-sample z, because your alternate hypothesis can have a direction (<, >) so the p-value will be cut in half, and you can compute a confidence interval to get an idea of the size of the difference in the percents. e.g. comparing trt. group to ctl. on relapse from cancer.

2-sample z-test

H_o: $p_{TRT} = p_{CTL}$ (% that relapse is the same)

H_a: $p_{TRT} \neq p_{CTL}$



H_o: no assoc. b/t which group you are in & whether or not you relapse

H_a: assoc.

Examples: A paper in 2021 by Jin, et al, classified COVID-19 patients by sex (M/F) and severity of illness (mild, severe, critical). Since these are two categorical variables, we may use a chi-square test to see if there's a relationship between sex and COVID-19 severity (i.e. if there's a difference between males and females in the severity of cases). A 2x3 table follows:

Contingency table results:
Rows: Sex
Columns: Severity

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

COVID severity

Sex

	Mild	Severe	Critical	Total
Male	5	6	12	23
	(21.74%)	(26.09%)	(52.17%)	(100%)
	(38.46%)	(42.86%)	(80%)	(54.76%)
	(7.12)	(7.67)	(8.21)	
Female	8	8	3	19
	(42.11%)	(42.11%)	(15.79%)	(100%)
	(61.54%)	(57.14%)	(20%)	(45.24%)
	(5.88)	(6.33)	(6.79)	
Total	13	14	15	42
	(30.95%)	(33.33%)	(35.71%)	(100%)
	(100%)	(100%)	(100%)	(100%)

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	2	6.0519627	0.0485

1. Do we have two categorical variables?

2. We start by describing the sample with marginal percents. Sex: Male Female Severity: Mild Severe Critical

3. If there is no relationship between sex and illness severity, we'd expect the same percent of males and females to have each level of severity. What percent would we expect both sexes to have of each level of severity?
   Mild Severe Critical

4. What do we actually see? (Look for a relationship between sex and severity with conditional percentages.)
   Among males, what percent had a mild case? Among females, % with a mild case?

   Severe case? Severe case?

   Critical case? Critical case?

5. Does there appear to be a relationship? Is there a difference between males and females in COVID severity?

6. Before we do a chi-square test, we need to check if the expected counts are big enough to do a chi-square test. Are they?

7. Hypothesis testing

   H_o:

   H_a:

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

   Is there a statistically significant relationship between sex and COVID severity (α=0.05)? i.e. Is there a significant difference between males and females in the percent with each level of severity?

e.g. Is alcohol consumption associated with cancer? (Allen N et al. Moderate Alcohol Intake and Cancer Incidence in Women. *J Natl Cancer Inst*, 2009;101: 296-305) (Within one population, are these two categorical variables independent?) This is a prospective observational study.

Contingency table results:
Rows: Outcome
Columns: Alcohol consumption

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

	Non-drinker	1-2 drinks/week	3-6 drinks/week	7-14 drinks/week	>14 drinks/week	Total
Cancer	17416 (25.32%) (5.677%) 16479	19307 (28.07%) (5.198%) 19954	15183 (22.08%) (5.166%) 15787	12838 (18.67%) (5.329%) 12940	4031 (5.861%) (5.99%) 3615	68775 (100.00%) (5.372%)
No cancer	289344 (23.88%) (94.32%) 290281	352146 (29.07%) (94.8%) 351499	278708 (23%) (94.83%) 278104	228056 (18.82%) (94.67%) 227954	63261 (5.222%) (94.01%) 63677	1211515 (100.00%) (94.63%)
Total	306760 (23.96%) (100.00%)	371453 (29.01%) (100.00%)	293891 (22.96%) (100.00%)	240894 (18.82%) (100.00%)	67292 (5.256%) (100.00%)	1280290 (100.00%) (100.00%)

Statistic	DF	Value	P-value
Chi-square	4	154.44698	<0.0001

1. Are the expected counts $\geq$ 5?
2. Marginal percents:
   What % of people, overall, had cancer?

   What % of people, overall were non-drinkers? Heavy drinkers?
3. Conditional percents:
   Among non-drinkers, what % had cancer?

   Among heavy drinkers, what % had cancer?

   Which level of drinking had the lowest rate of cancer?
4. Is there a statistically significant association between level of alcohol use and getting cancer?

   H_o:

   H_a:

   Circle the χ^2 statistic, df, and p-value.
   Conclusion (α=0.05):

5. There's not much difference between the alcohol groups in the percent with cancer. Why is it statistically significant?

Note: This article took a lot of heat because:
1) they said that no level of alcohol was safe

2) they said that based on this study, alcohol causes cancer.

How to find expected values for the chi-square test (2 categorical variables):

Just as before, the test statistic is: $\chi^2 = \sum(o-e)^2/e$, where o=observed count in a cell, and e=expected count in a cell if H_o is true [but now df = (#rows – 1) x (#columns – 1)].

So we need to find the expected values, both for computing the chi-square AND checking our assumption that all the expected values are at least 5.

The null hypothesis is that there is no association between the two variables, or no difference between the groups in the distribution of the second variable. This means for the COVID example:

Males and females are _____ when it comes to COVID severity (each sex has the same distribution of mild, severe, and critical cases.)

Therefore, to compute the expected count, we would take the overall percent that had each level of severity, and apply that percent to the number of males and females.

Overall, 30.95% had mild cases. So if H_o was true, we'd expect 30.95% of the males to have mild cases, and 30.95% of the females to have mild cases.

So, if H_o is true, we'd expect 0.3095*23 = 7.12 males with mild cases in a sample of this size, and 0.3095*19 = 5.88 females to have mild cases in a sample like this. We'd do the same thing for severe and critical cases.

So, in general, to find the expected values, we . . . *find the _____ (i.e. marginal %) in each*

category/group of one of the variables, and multiply that by the _____ of each group of the other variable. This gives us how many we'd expect in the various boxes if the groups of one variable were exactly the same relative to the other variable.

Another example: Pew Research Center, Aug. 23, 2012: gender and political party.

Note that overall, 53.75% of the sample is female. If H_o of no association is true, we'd expect 53.75% of Democrats to be female, 53.75% of Republicans to be female, and 53.75% of Independents to be female.

*We have _____ Democrats. 53.75% of 4686 = 0.5375*4686 = **2519** female Democrats expected in a sample of this size if there is no association between sex and political party.*

Contingency table results:
Rows: Gender
Columns: Political party affiliation

Alternatively, overall _____% are Democrats. If H_o of no association is true, we'd expect 36.35% of Females to be Democrats and 36.35% of Males to be Democrats.
*We have _____ Females. 36.35% of 6929 Females=**2519 female Dems expected**.*

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

	Democrat	Independent	Republican	Total
Female	2887 (41.67%) (61.61%) 2519	2093 (30.21%) (46.99%) 2394	1949 (28.13%) (51.95%) 2017	6929 (100.00%) (53.75%)
Male	1799 (30.17%) (38.39%) 2167	2361 (39.59%) (53.01%) 2060	1803 (30.24%) (48.05%) 1735	5963 (100.00%) (46.25%)
Total	4686 (36.35%) (100.00%)	4454 (34.55%) (100.00%)	3752 (29.1%) (100.00%)	12892 (100.00%) (100.00%)

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	2	203.17804	<0.0001

One more example, from start to finish: The 2015 survey of more than 7800 randomly selected Millennials (born after 1982) from 29 countries, all with a college degree and employed full-time, examined world regions, gender, opinions, and goals. (n=300 from most countries). Is there a difference between people in developed and emerging markets in the kind of company they would like to work for if they were to leave their current job?

Contingency table results:
Rows: Markets
Columns: Desired employment if left current job

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

	Large Co.	Medium Co.	Start New	Small Co.	Self-employed	Total
Developed	1194 (36.84%) (34.76%) (1474.94)	1092 (33.69%) (63.97%) (732.96)	375 (11.57%) (27.94%) (576.24)	375 (11.57%) (63.03%) (255.48)	205 (6.33%) (43.71%) (201.38)	3241 (100%) (42.94%)
Emerging	2241 (52.03%) (65.24%) (1960.06)	615 (14.28%) (36.03%) (974.04)	967 (22.45%) (72.06%) (765.76)	220 (5.11%) (36.97%) (339.52)	264 (6.13%) (56.29%) (267.62)	4307 (100%) (57.06%)
Total	3435 (45.51%) (100%)	1707 (22.62%) (100%)	1342 (17.78%) (100%)	595 (7.88%) (100%)	469 (6.21%) (100%)	7548 (100%) (100%)

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	4	623.25248	<0.0001

1. How many people in developed markets wanted to work for a large company?

2. What is the expected number of people in emerging markets who wanted to be self-employed?

3. This is the expected number, assuming what?

4. Marginal distribution: What % of respondents, overall, came from developed markets?

   What % of respondents, overall, wanted to work for a large company?

5. Conditional distribution: What % of respondents from developed markets wanted to work for a big company?

   What % of respondents from emerging markets wanted to work for a big company?

6. Which distribution (marginal or conditional) helps us decide if there is a relationship?

7. Based on those percents, does there appear to be a difference between Millennials from developed and emerging markets in the percent who want to work in these settings?

8. If there was no difference/relationship, what % of respondents from developed markets would we expect to want to work for a large company?

9. May we do a chi-square? (Hint: look at the expected values)

10. What are the null and alternate hypotheses for this question?

    Ho:

    Ha:

11. Circle the chi-square statistic, df, and p-value,
12. What do you conclude (alpha=.01)?

Scatterplots and Correlation Mon., 11/6/2023

GOAL: to know how to create and interpret a scatter plot, and to know the properties and purpose of correlations.

Quiz on Correlations (Quiz #15): Due Friday 11/10/2023 at 11:59 p.m.

We use regression and/or correlation when we have two **quantitative** variables, and we want to see if there is an **association** between these two variables. Two variables are associated if certain values of one variable tend to occur with certain values of the other variable.

e.g. high values of height tend to occur with high values of weight.
e.g. high values of practice (in sports or music) tend to occur with low values of mistakes.

In some cases, we hope merely to describe the relationship (e.g. SAT math and verbal scores).

For this, we will use _____.

In other cases, we hope to either 1) predict one variable from the other (predict someone's blood pressure from the #hours of exercise they tend to get in a week) or 2) see the rate of change (how does blood pressure change as the #hours of exercise increases?).

For these, we will use _____.

For regression, a response variable measures an outcome of a study. An explanatory variable explains or causes changes in the response variable. Response variables are often called dependent variables, and explanatory variables are often called independent variables.

big Caution:

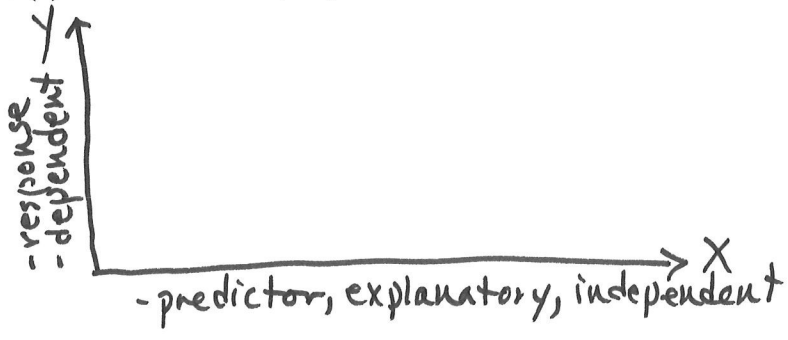
e.g. There is a strong correlation between shoe size and reading ability in elementary schools. If parents can stretch their kids' feet out, will they become better readers?

e.g. There is a strong correlation between the number of firemen at a fire and the amount of damage done. Are these firemen out there doing as much damage as possible?

In this section, we are going to start with a graphical display of the relationship between two quantitative variables (scatter plot) and then move on to a numerical summary of that relationship (correlation). In the next section, we will use a mathematical model to describe a regular pattern (regression analysis), and we'll also look for deviations from the pattern.

1. **Graphical display**=Scatter plots: portray the relationship between two quantitative variables measured on the same individuals.

-one dot per individual
-explanatory variable (x) on horizontal axis, response variable on vertical axis (y)



Once you've graphed the data, describe the **form, strength, and direction** of the relationship, look for **outliers,** and notice the **range** of the x's (predictor).

(1) Form= Do we have a line or a curve?

(2) Strength= Is the relationship strong or weak?

(3) Direction=

--Positive association: High values of one variable tend to occur with high values of the other variable, and low values of one variable tend to occur with low values of the other variable.

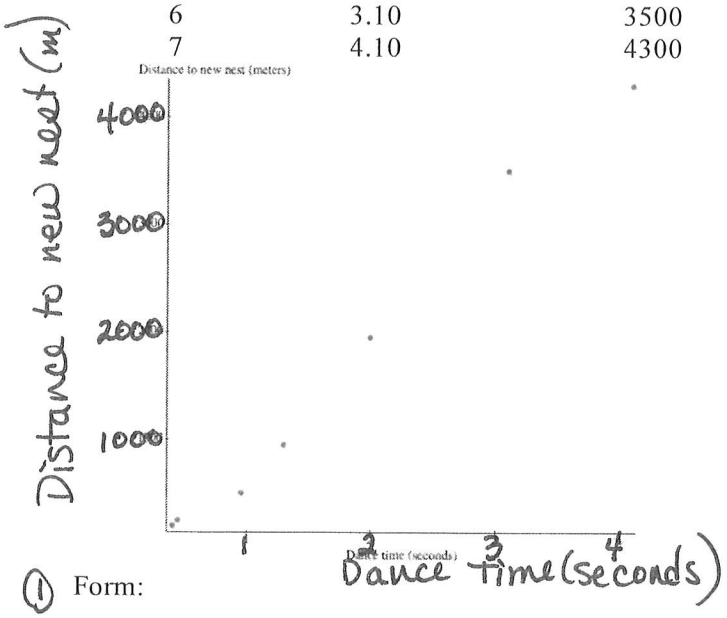
--Negative association: High values of one variable tend to occur with low values of the other variable.

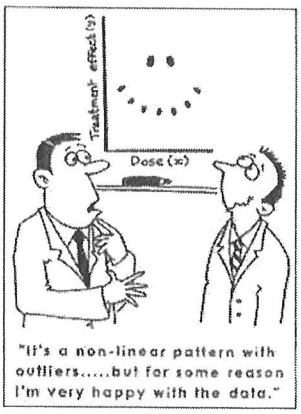
(4) Outliers=Data points that don't fit the pattern.

(5) Range of the x's (the predictor): What is the minimum and maximum value for our predictor (x)?

e.g. In a study of honeybees, Seeley (2010) observed that scout bees do a "waggle dance" to help communicate the distance to a new nest site to bees back in the original nest. The table below shows the distance to the new site (in meters) and duration of the dance (in seconds) recorded for seven different scout honeybees.

Honeybee	Dance time (secs)	Distance to new nest (meters)
1	0.40	200
2	0.45	250
3	0.95	500
4	1.30	950
5	2.00	1950
6	3.10	3500
7	4.10	4300





(1) Form:

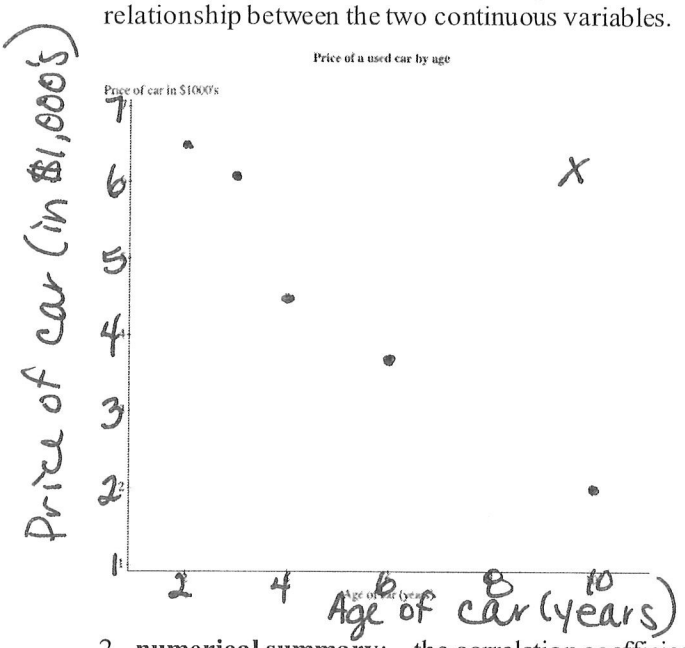
(2) Strength:

(3) Direction:

(4) Outliers?

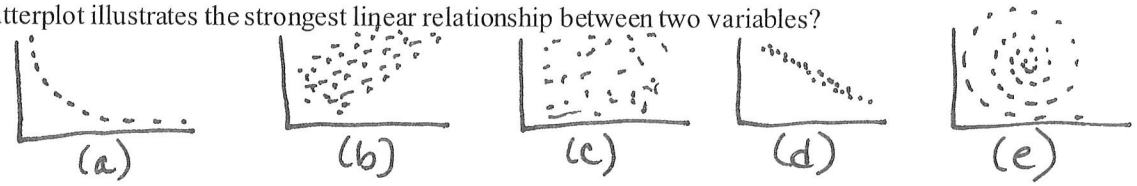
(5) Range of x's:

If you have more than one group, plot points in different colors, or with different letters, to see if group affects the relationship between the two continuous variables.



2. **numerical summary**: the correlation coefficient, r

Which scatterplot illustrates the strongest linear relationship between two variables?



A numerical measure of the direction and strength of the linear relationship between two quantitative variables is the correlation coefficient:

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

x_i = indiv. value of x
$\bar{x}$ = mean of x
s_x = sd of x
n = sample size

What does $(y-\bar{y})/sd$ look like? _____ That is the number of s.d. away from the mean; it has NO UNITS. Therefore you can change units in a correlation without changing the correlation. This is a very desirable characteristic. e.g. If we want to know the nature of the relationship between the weight of a car and the gas mileage of a car, we don't want the number describing that relationship to change depending on whether you are in Canada (kg and km/liter) or the U.S. (pounds and mpg).

(Do not despair. You will never have to compute r! We'll let the computer do this for us. ☺)

Properties of r:

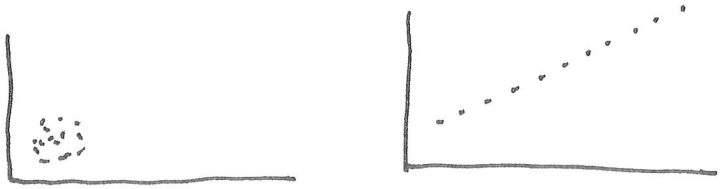
1.

2.

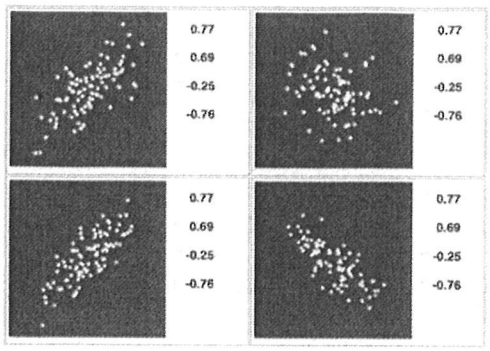
3.

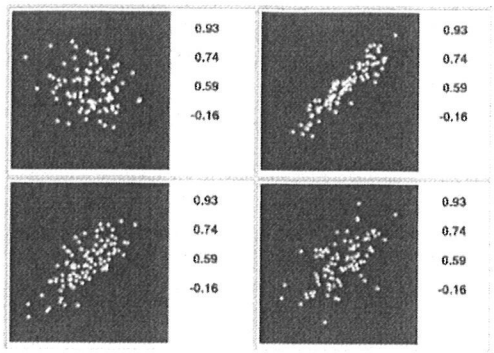
4.

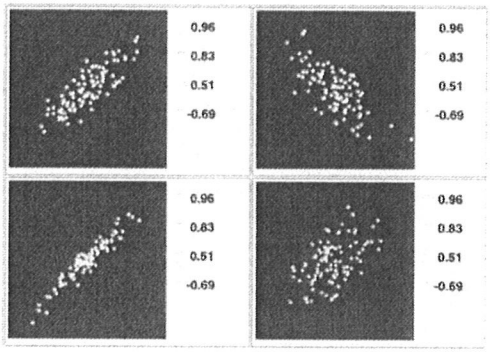
5.

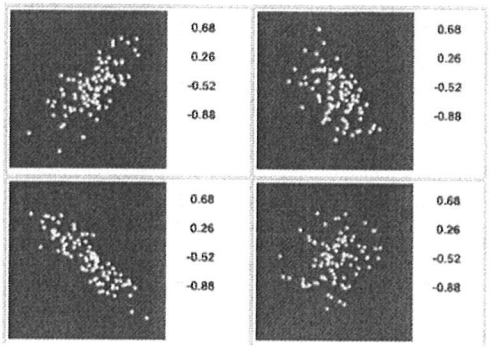


Matching correlations with scatter plots: http://istics.net/Correlations/

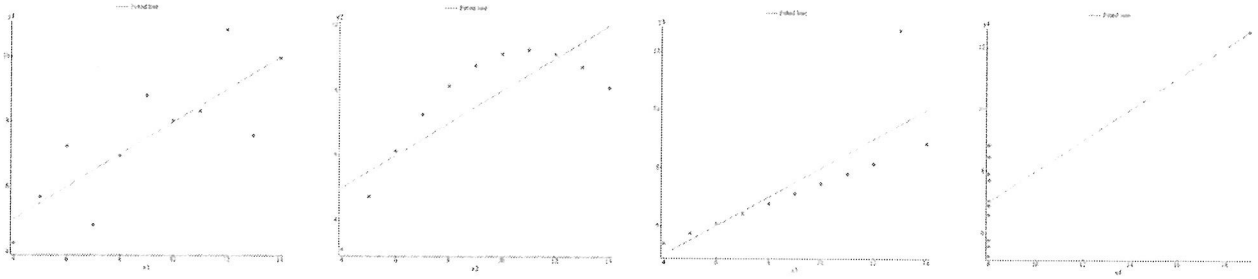








Anscombe's quartet shows the importance of _____.



r=0.816 r=0.816 r=0.816 r=0.816

A partial review:

1. To see the relationship between two quantitative variables graphically, we use a:
   a. Box plot b. histogram c. bar chart d. scatter plot

2. Correlations are used to describe the relationship between two quantitative / categorical variables. If we have a linear / quadratic relationship, the correlation will tell us the _____ and _____ of that relationship.

3. The _____ tells us the direction.

4. The _____ tells us the strength.

   Sort the following correlations from weakest to strongest, crossing out any that can't be a correlation:
   -1.7 .7 .3 -.1 1.25 -.9

5. Correlations are affected by outliers. Outliers make the correlation:
   a. Weaker b. Stronger c. Either weaker or stronger, depending on where the outlier is.

Linear Regression

Mon., 11/6/2023

GOALS: To understand slope and intercept. To be able to make predictions using a regression equation and know where predictions are likely to be best. To be able to sketch a line. To know the measures of fit for a regression line. To know when graphs suggest trouble for our linear regression model. To know possible explanations for an association between two variables.

Homework/Quiz #16: Due Monday, 11/13/2023 at 11:59 p.m.

For correlations, we do not talk about an explanatory and response variable because we are just looking for the association between two quantitative variables. In regression, we do talk about explanatory and response variables because we are using one quantitative variable to predict another, or we want to see the rate of change.

Correlations measure the direction and strength of the linear relationship between two quantitative variables. If we want to go beyond that, and to draw a line to graphically show the relationship more specifically, we need linear regression. A regression line describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x.

Matching: What would we use for:

a. chi-square b. correlation c. regression

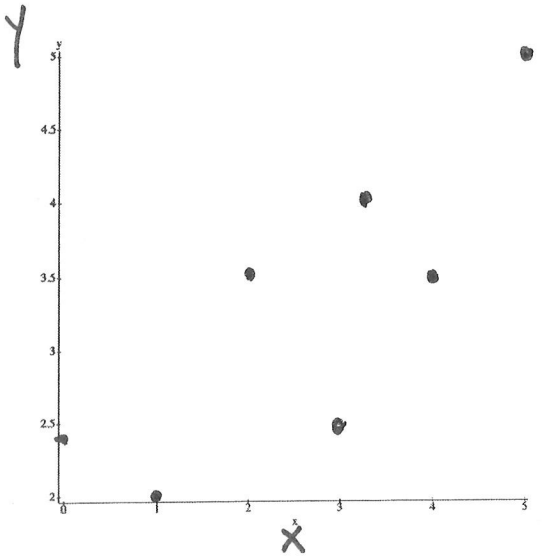
1. _____Is there a relationship between the number of my tulips that bunnies eat and how much I like bunnies that week (on a scale of 1-100)?
2. _____Can I predict the number of weeds in my garden from the number of minutes I work on teaching that week?
3. _____Is there a relationship between whether or not I see squirrels in my backyard and whether or not my dog is outside?
4. _____How does the number of minutes that I work in my garden change as the temperature decreases?

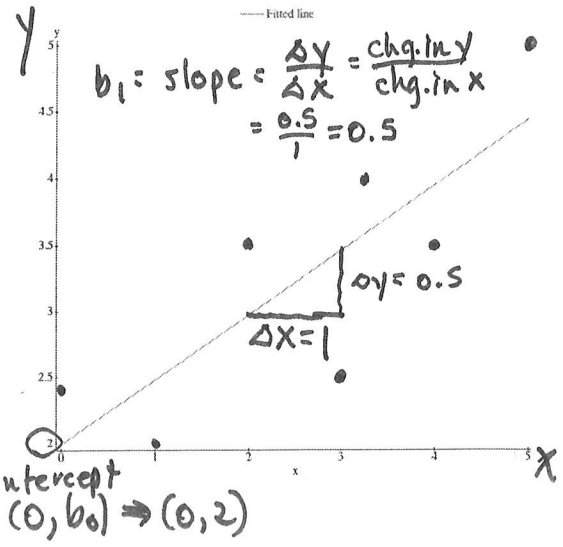
The least squares regression line of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible. (How's that for a mouthful?!?! It's easier to show:

Review from high school (y=mx+b) ??

In statistics, we generally use: $y = b_0 + b_1x$

y=response variable
x=explanatory variable
b_0=intercept (where the line hits the y-axis. i.e. the value of y when x is 0)
b_1=slope (the rate of change. As x increases by 1, y will change by b_1)

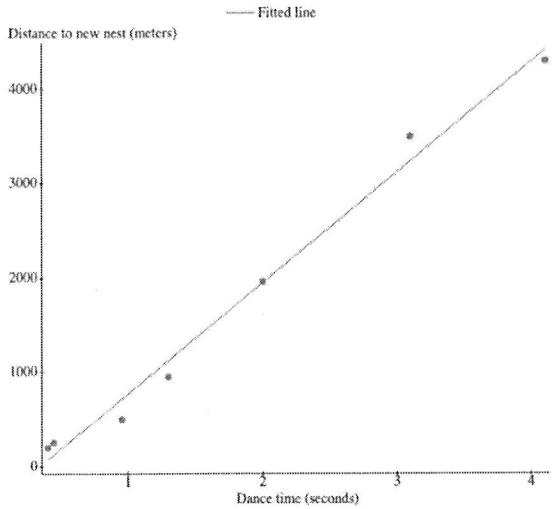




Finding the slope and intercept of a regression equation

There are formulas for computing slope and intercept, but we will be getting them from output. We will never be computing them. You need to know where to find them and how to interpret them.

Back to the honeybee waggle dance (using how long the dance lasts to estimate the distance to the new nest)



Simple linear regression results:
Dependent Variable: Distance to new nest (meters)
Independent Variable: Dance time (seconds)
Distance to new nest (meters) = -399.05265 + 1174.2576 Dance time (seconds)
Sample size: 7
R (correlation coefficient) = 0.99419419
R-sq = 0.98842209
Estimate of error standard deviation: 194.92409

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	-399.05265	124.10346	> 0	5	-3.2154836	0.9882
Slope	1174.2576	56.83584	≠ 0	5	20.660513	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	16218594	16218594	426.85679	<0.0001
Error	5	189977	37995.4		
Total	6	16408571			

1. Intercept: What is the intercept?

   How do we interpret it?

2. Slope: What is the slope?

   How do we interpret it?

Making predictions. To predict the value of the response variable (y) from a given value of the predictor variable (x), we substitute that particular value of x into the regression equation and determine the predicted value of y.

The best predictions occur for

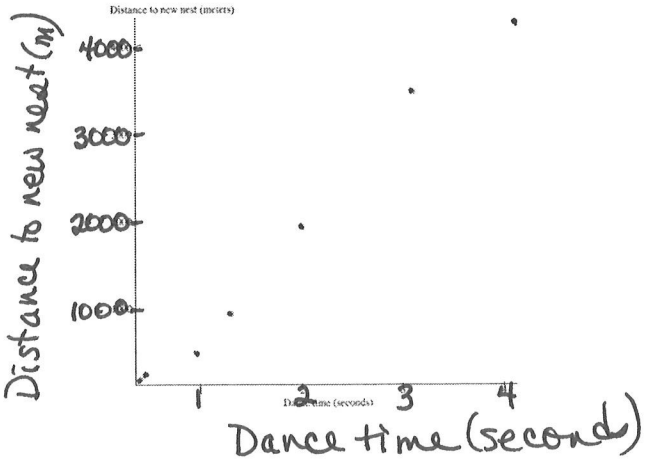
e.g. Predict the distance to the new nest for various times of the waggle dance.

distance (when the dance lasts 1 second) =

distance (when the dance lasts 4 seconds) =

distance (when the dance lasts 0.25 second) =

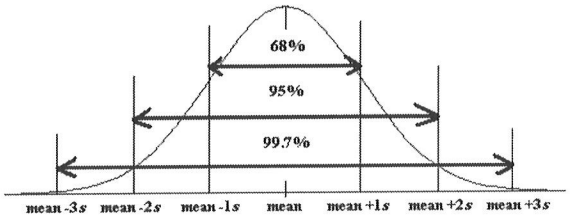
Sketching the line Make two predictions, plot them, and connect them. e.g. honeybee data



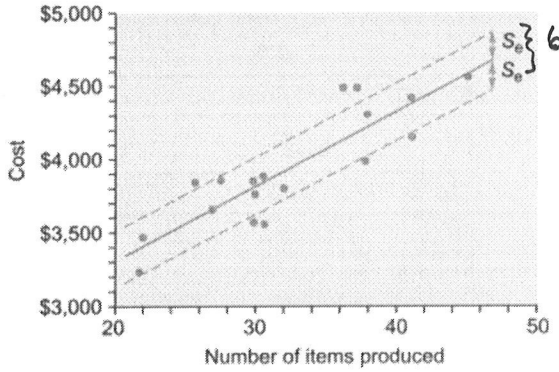
How well does the regression line fit the data?

Two measures: **standard error of the line** (s_e) and $\mathbf{R^2}$. The standard error of the line (s_e) is the same idea as standard deviation (a measure of spread). In fact, the s_e is the s.d. of the distances of the points to the line. It is a measure of the scatter, or dispersion, of our data around the regression line. $s_e = \sum(y_i -)^2/(n-2)$ y_i = actual observed y, $\hat{y}$ = predicted y (line)

The distance the points fall from the line (i.e. how far off the actual observation is from the predicted value) has a normal distribution. Remember the 68-95-99.7 rule?



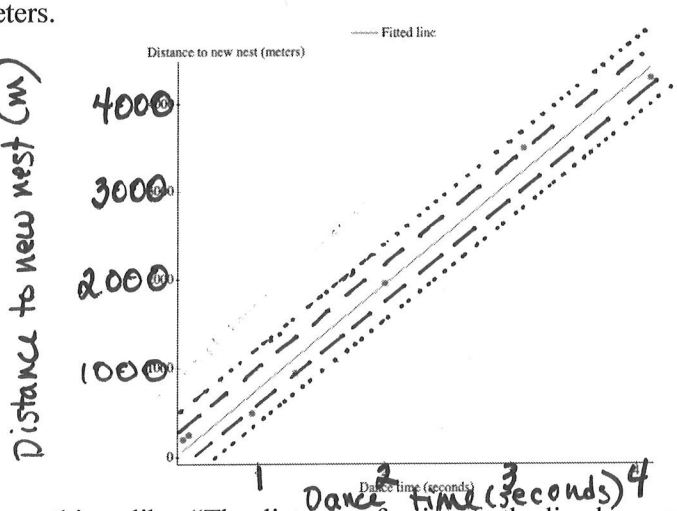
Now we're doing the same thing, only we're doing it around the regression line instead of the mean.



68% of the sample data falls within ±1 s_e (±1 s.e. of the line)

So 68% of the data points fall within 1 s_e, 95% fall within 2 s_e, and 99.7% of the data points fall within 3 s_e.

The s_e is called the ESTIMATE OF ERROR STANDARD DEVIATION on Statcrunch output. For the honeybee data, the s_e = 195 meters.



To interpret this, we say things like, "The distance of points to the line has a normal distribution with s.d.=195 meters. So 68% of the actual distances (from old nest to new nest) in our sample are within 195 m of what we predicted, and 95% of the distances in our sample are within 390 m of what we predicted." Or "The actual distances to the new nests in our sample differed from what we predicted, and the amount we were off had a s.d. of 195 m."

For a good fit, the s_e should be small / big. Note that *small* and *big* are relative to what you are measuring. e.g. if you are looking at the weights of animals, a s_e of 5 pounds would be very small for elephants and very big for meerkats. *You can compare the s_e (i.e. spread around the line) to the sd of the dependent variable (s_y) (i.e. spread around the mean) to get an idea of whether it's small or big (relatively speaking).* If s_e is close to the s_y (in this case, the s_y of the distance to the new nests), it's a big s_e, and your regression is not useful, because the regression hasn't done anything to reduce the random variation of y. In this case, the s_y of the distances to the new nest is 1654 m, so our s_e (195 m) is a huge reduction in variability, and this regression line is very useful.

To get an idea of how well our model can predict (another measure of fit), we use **R^2**. R^2 is usually expressed as a percent: the percent of the variation in the values of y that is explained by x. You can find R^2 on the Statcrunch output right above the s_e.

R^2 is the correlation coefficient (r) squared, so it ranges from _____ to _____ or _____ to _____.

When predicting the distance to a new nest, if we knew nothing about how long the waggle dance lasted, we'd do best by predicting the overall mean distance to the new nest (1664 m). R^2 gives us an idea of how much better we can do by predicting the distance using the waggle dance. An R^2 of 0 suggests we won't do any better. An R^2 of 100% means that we can predict exactly, with no error, when we use the waggle dance as a predictor.

For a good fit, R^2 should be small / big. What's small or big depends somewhat on the situation (perhaps biology is looking for a higher R^2 than psychology is.) But a general rule of thumb is that an R^2 of 0-19% is a very bad fit (i.e. very little of what's going on in y is explained by x), 20-39% is a pretty bad fit, 40-59% is a moderate fit, 60-79% is a pretty good fit, and 80%-100% is a very good fit.

e.g. Honey bee R^2=0.988. This means that 98.8% of the variability in distance to the new nest is explained by how long the waggle dance lasts. Only 1.2% of the variation in distance is due to unknown sources (e.g. weather? how drunk the honey bee is on flower fumes?). With an R^2=98.8%, our predictions are likely to be VERY close to the actual distance to the new nest. When the R^2 of a regression is low, you want to consider trying a different predictor, or adding a predictor (multiple regression), because otherwise our predictions could be miles off from what actually happens.

Let's say that we want to predict the distance to the new nest from how much pollen a honey bee collected, and that the R^2 from that regression is 2%.

Timing of waggle dance / pollen collected is the better predictor because R^2 is lower / higher.

There are many cases where newspaper and journal articles don't mention R^2, but this is what they are giving you, so you should know the definition of R^2.

e.g. Stealing Signs, by Peter Keating, trying to predict won-lost records in major league baseball (MLB). "Differences in team salaries accounted for 58% of the variation in winning percentages. All the rest combined (injuries, career seasons, . . .) explained just 42% of the spread in wins and losses."

R^2 = (so over half of what's going on in won-lost records is explained by salaries!)

e.g. The Distance from the Goal, by Glenn Kessler of the *Washington Post*, 4/17/2016, looking at the wage gap between men and women. "48% of the wage gap could be explained by a variety of factors largely in a women's control, such as choice of occupation and industry. The level of experience helped explain another 14% of the wage gap. The rest is 'unexplained' and could be the result of discrimination."

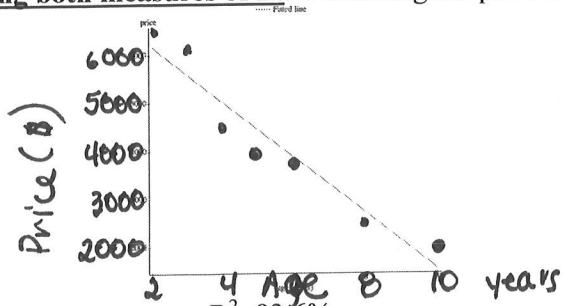
R^2 = unexplained variation =

e.g. Ahmad, et al, in the *J. of the Am. Heart Assoc.*, 9/17/19, reported that the poverty rate in a county "statistically explained ≈30% of the observed between-county variation in heart failure mortality rates."

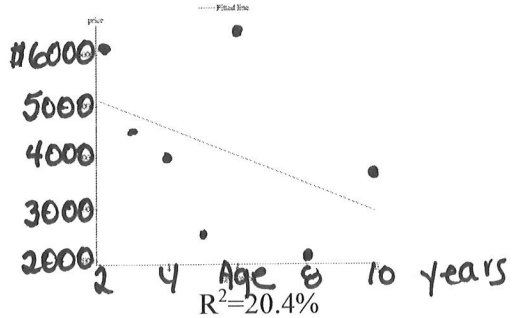
R^2 =

Given how complicated life is, a predictor can still be useful in learning about the world, even if the R^2 is not particularly high. These R^2 don't show a great fit, but are still really valuable information to have.

Showing both measures of fit. Predicting the price of a car from the age of a car (2 possible scenarios):

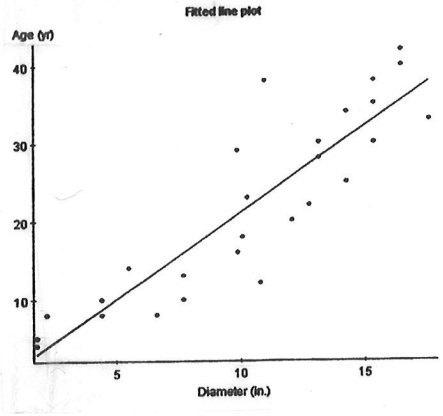


R^2=93.6%
s_e=$467.60 (s.e. of the line)
s_y=$1,683.70 (s.d. of price)



R^2=20.4%
s_e=$1,645.90
s_y=$1,683.70

e.g. Predicting the age of a tree (years) from the diameter of a tree (inches). (Both are quantitative.)



Simple linear regression results:
Dependent Variable: Age (yr)
Independent Variable: Diameter (in.)
Age (yr) = -0.974424 + 2.205518 Diameter (in.)
Sample size: 27
R (correlation coefficient) = 0.8882
R-sq = 0.7889163
Estimate of error standard deviation: 5.5764275

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	-0.974424	2.604332	≠0	25	-0.37415507	0.7114
Slope	2.205518	0.22816683	>0	25	9.666251	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	2905.5493	2905.5493	93.43641	<0.0001
Error	25	777.4136	31.096542		
Total	26	3682.963			

1. Circle the regression equation.

2. What is the intercept, and what does it mean?

3. What is the slope, and what does it tell us (very specifically)?

4. Predict the age of a tree that has diameter 10 inches.

5. Are we extrapolating?

6. What are our two measures of fit for the regression?

7. What is the s_e? What does it tell us? (FYI, s_y (the s.d. of tree age) is 11.7 years.)

s_e=5.58 years. This is a measure of spread around the line. Specifically, the difference between the actual tree ages in our sample and the tree age we predicted (i.e. the regression line) has a normal distribution with a s.d. of 5.58 years. Remember the 68-95-99.7 rule? This means 68% of our sample trees fall within ±5.58 years of the regression line, and 95% of our sample trees fall within 11.16 years of the line, and 99.7% of our sample trees . . . .

Also, s_e (along with R^2) tells us if a regression line fits this data well. In this case, (s_e=5.58 years), the s_e is about _____ of the general overall spread of tree age around the mean tree age (s_y=11.7 years). Cutting the random variation in half is good, so diameter is a useful predictor of tree age.

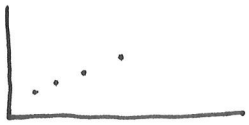
8. What percent of the variation in age is explained by diameter? Is this a good fit? (i.e. Is diameter a good predictor?)

9. What is the correlation coefficient? How could we compute this?

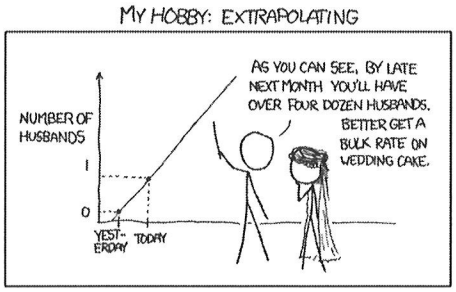
Cautions about correlation and regression:
1. Extrapolation.
2. Do graphs show anything that suggests trouble (non-linearity, unequal spread, outliers)?
3. Association vs causation.

Here we go:

1. **Extrapolation**. Predicting outside the range of x's. This is dangerous because



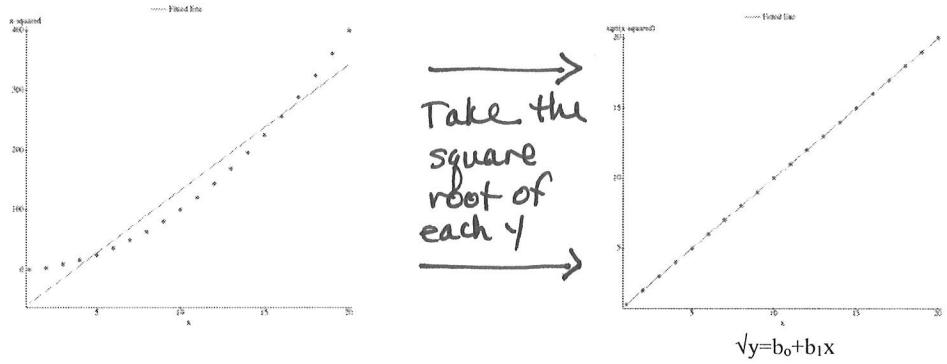
e.g. Using data on the day before your wedding & your wedding day to predict the future #spouses. (XKCD)



2. **Do graphs show anything that suggests trouble for our linear regression model?**

   a. Is the relationship between the predictor and response variables _____ ?

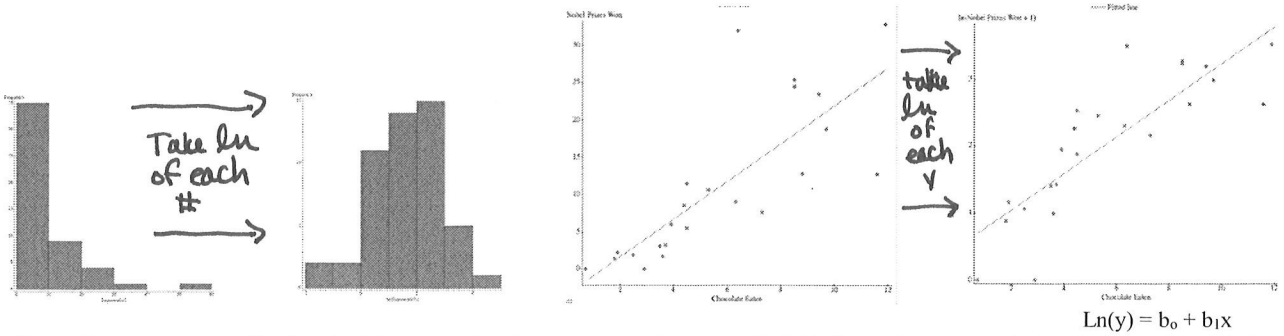
   If you have a curve instead of a line, there are two possible solutions: Add a quadratic term to the model (e.g. $y=bo+b_1x+b_2x^2$), or transform the data. (e.g. $\sqrt{y}$, ln(y))



(If you transform the data (e.g. $\sqrt{y}$), then remember that when you are making predictions, you are predicting $\sqrt{y}$ instead of y, and you would have to square the prediction to get the correct prediction.)

   b. Is the spread around the line the same over the whole line, or is there an increase as x increases? Possible solution: transform the data. ln(y) (natural log of y)

This is often found when Y is right skewed (e.g. cholesterol), because the predictions are not as precise as X increases (since a right skewed Y will have a progressively wider spread as X increases). Again, try transforming: take the square root or natural log (ln) of y, since if you have a right skewed distribution, taking the ln of each number will pull in the long right tail and make it more normal.



Note, if you are predicting ln(y) instead of y, you need to exp(ln(y)) in order to get it back into the original units.

3. **Are there outliers?**

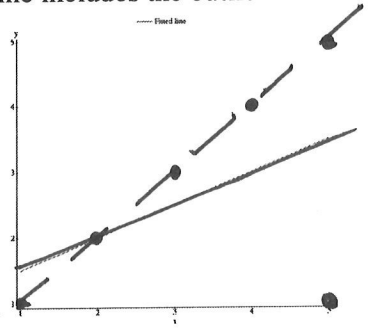
--Remember, correlations are affected by outliers. Outliers make the correlation:

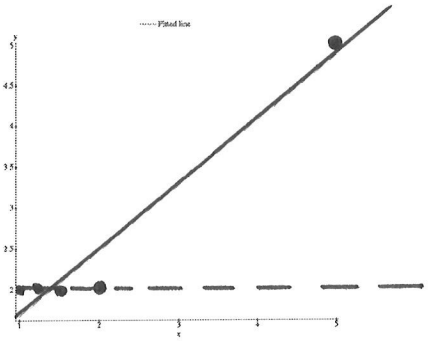
a. Weaker b. Stronger c. Either weaker or stronger, depending on where the outlier is.

--Regression lines are also affected by outliers. Given that the computer is trying to minimize the distance between points and the line, what effect will outliers have on regression?

a. Push the line away b. Pull the line towards themselves c. Could be either one.

Which line includes the outlier?





Association vs Causation

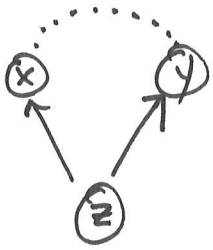
We often want to say that a change in one variable CAUSES a change in another variable. When can we say this with reasonable certainty? When do we use caution?

(Broken lines show an association. Arrows show a cause-and-effect link. We observe x and y, but z is a lurking variable.)

Causation means that a change in one variable causes a change in another. i.e. X and Y are associated, AND a change in X causes a change in Y.

e.g. ↑running speed ⇒ ↑heart rate
e.g. ↑radiation ⇒ ↑cancer

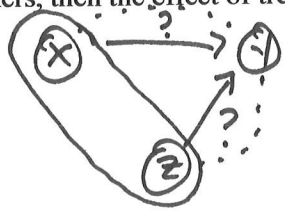
Common response means that there is a lurking variable (z) that causes both x and y, so that though x and y are in fact related, x is not causing y



e.g. ↑shoe size ⇒ ↑reading scores among elementary school children???

e.g. ↓sleep ⇒ ↑BP ??? Is there a lurking variable?

Confounding means that the effects of two variables (x and z) are mixed up together, so we cannot determine which of them (or perhaps both of them) is causing y. This often occurs when groups differ in important demographics at the beginning of a study (e.g. If the group getting Trt A has a lot of smokers, and the group getting Trt B. has fewer smokers, then the effect of treatment is mixed up with the effect of smoking.)

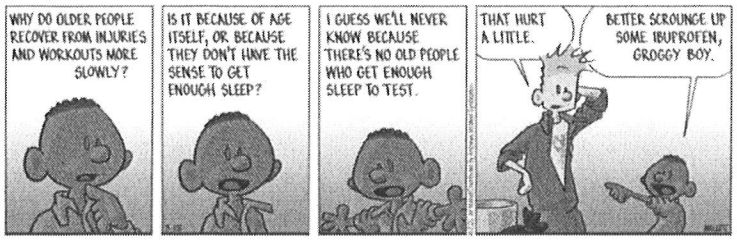


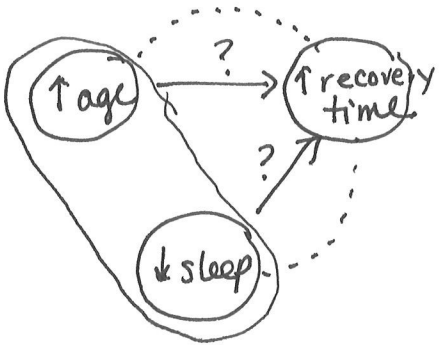
e.g. ↓fluoride ⇒ ↑cavities ??? ↑age ⇒ ↑cavities???

e.g. ↑chocolate eaten ⇒ ↑Nobel prizes??? Maybe countries that eat more chocolate have other characteristics that lead to more Nobel prizes??

Confounding in the comics: ☺

Frazz. By Jeff Mallett. March 18, 2020





So, when can we say that something causes something else?

1. A well-designed experiment, controlling for all the possible lurking variables. This is only possible in the lab. When you're working with people, you generally can't do this. (We can't, e.g., randomly assign babies at Butterworth Hospital to be smokers or non-smokers and see what happens to them.) What you can do is try to control for lurking variables (make your groups as similar as possible).

2. When we can't do an experiment, we look for (using smoking and cancer as an example):
   --a strong association

   --a consistent association

   --higher doses are associated with stronger responses

   --the alleged cause precedes the effect in time

   --the alleged cause is plausible

Researchers and organizations spend a great deal of time and effort trying to decide what's simply correlation and what's causation. The World Health Organization (WHO) has various groupings. Group 1 (smoking, processed meats, exposure to solar radiation) is listed as carcinogenic to humans on the basis of epidemiological studies. Group 2A (steroids, red meat, alcohol, pesticide DDT, . . .) is probably carcinogenic, but they have not ruled out "chance, bias, or confounding".

Chance: sheer bad _____ that their sample doesn't reflect the truth in the population

Bias: they did something _____ so their sample doesn't reflect the truth in the population

Confounding: the effect of these possible carcinogens is _____ with another variable.

In summary: General Cautions about Correlation and Regression

1. Correlation only measures linear / quadratic association.

2. Extrapolation (predicting outside your range of x's / y's) is dangerous.

3. Correlations and regressions are not resistant (they are greatly affected by _____).

4. Association does not imply _____. (e.g. lurking variables may make a correlation or regression misleading.)

5. Correlations based on a limited range of data may be too small.