Testing for Independence

1. Does ice cream preference affect performance on Math 10B midterms? A GSI surveys their 50 students on whether they like vanilla or chocolate more. Out of the students who did well, 15 prefer vanilla and 10 prefer chocolate; out of the remaining students, 11 prefer vanilla and the rest prefer chocolate.

   a) *What are the hypotheses you are testing?*
   The **null hypothesis** is that flavor preference and midterm performance are independent. The *alternate hypothesis* is that these are not independent.

   b) *Which test should you use here and why?*
   We will use a χ^2-test for independence, as we have discrete possible outcomes for our observations (e.g. prefers chocolate and did well), and our null hypothesis gives us some expectation on how these counts should be distributed.

   c) *What is the p-value of your experiment?*
   In order to get the p-value for the χ^2 test, we must calculate the corresponding χ^2 statistic. We can best keep track of our calculation with the following table.

	Vanilla		Chocolate		
	Observed	Expected	Observed	Expected	Total
Did Well	**15**		**10**		*25*
Did Not Well	**11**		**14**		*25*
Total	*26*		*24*		*50*

   Now we will use our null hypothesis to find the expected values. For instance, we see that $\frac{25}{50}$ of the students did well, and $\frac{26}{50}$ of students prefer vanilla. So by independence, we should have $\frac{25 \cdot 26}{50^2}$ of the students satisfying both of these. This means out of 50 students, we expect $\frac{25 \cdot 26}{50^2} \cdot 50 = 13$ students to have done well and like vanilla. So filling in the rest of the chart, we get:

	Vanilla		Chocolate		
	Observed	Expected	Observed	Expected	Total
Did Well	**15**	13	**10**	12	*25*
Did Not Well	**11**	13	**14**	12	*25*
Total	*26*		*24*		*50*

   Then we get our statistic by adding up $\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$ from each possible outcome. So we get the statistic $\chi^2 = \frac{4}{13} + \frac{4}{13} + \frac{4}{12} + \frac{4}{12} \approx 1.28$. We also need to know the degrees of freedom to look this up. In this case, we have two rows and two columns, so the degrees of freedom is $(2-1) \cdot (2-1) = 1$. Finally, we can see form the χ^2-chart that our p-value will be somewhere between 0.30 and 0.20.

   d) *What should you conclude?*
   Because our p-value is decently large, we would fail to reject the null hypothesis in this case. We conclude that there is no evidence that ice cream preference affects performance.

2. Darwin wonders if there is any relationship between the colors of sparrows he observes (white, brown, black, or red) and the shape of their beak (flat or curved). He records the features of 100 sparrows in the following chart. What conclusions should he draw about his hypothesis? (Hint: You will need to fix the shape of the table.)

Red Flat	Red Curved	Brown Flat	Brown Curved	Black Flat	Black Curved	White Flat	White Curved
10	15	13	7	11	19	6	19

The two factors we are testing the independence of in this case are coloration and beak shape. So our chart will look like: Again, we compute the expected value in each cell by looking at the row total times

	Flat Beak		Curved Beak		
	Observed	Expected	Observed	Expected	Total
Red	**10**	10	**15**	15	*25*
Brown	**13**	8	**7**	12	*20*
Black	**11**	12	**19**	18	*30*
White	**6**	10	**19**	15	*25*
Total	*40*		*60*		*100*

the column total divided by 100. And to get the p-value of our experiment, we can calculate the χ^2 statistic:

$$\chi^2 = \frac{0^2}{10} + \frac{0^2}{15} + \frac{5^2}{8} + \frac{5^2}{12} + \frac{1^2}{12} + \frac{1^2}{18} + \frac{4^2}{10} + \frac{4^2}{15} \approx 8.01$$

To look this up, we also need the degrees of freedom. We have 4 rows and 2 columns, so we get $(4-1)\cdot(2-1) = 3$ degrees of freedom. Then looking up 8.01 on the row corresponding to 3 degrees of freedom gives us something a bit less that 0.05. We get $p \approx 0.046$. This seems like enough evidence to reject the null hypothesis and claim there is some dependence between color and beak shape with significance level 5%. But if we had wanted to be more conservative in rejecting the null hypothesis, our p-value may not have been small enough.

Correlation & Best Fit Lines

3. Suppose you want to know whether performance on Quiz 1 is correlated with performance on Quiz 12. You randomly choose 5 students' quiz scores and get the following values.

Student	Quiz 1	Quiz 12
A	7	9
B	12	11
C	6	5
D	11	10
E	4	5

a) *Make a (rough) scatter plot of these scores. Do they seem to have a strong linear relationship? What would you guess the correlation coefficient (r) is approximately?* When you graph these, it seems like there is some sort of nearly linear relationship. In general, a higher score on Quiz 1 seems to yield a higher score on Quiz 12 for our sample. This would be confirmed by a *correlation coefficient* that was positive and close to 1. Since the data isn't a perfect line, we might guess $r \approx 0.9$.

b) *Calculate the sample standard deviation for Quiz 1 (σ_x). Do the same for Quiz 12 (σ_y).* In order to find the sample standard deviation, we first need to calculate the sample average ($\overline{x}$). In this case, we will get $\overline{x} = 8$. By coincidence, the average of the Quiz 12 scores also gives us $\overline{y} = 8$. To keep track of the calculation, the following chart is helpful. It is **very important** that you keep track

	$x_i - \overline{x}$	$y_i - \overline{y}$
A	-1	1
B	4	3
C	-2	-3
D	3	2
E	-4	-3

of whether each entry is positive or negative (for calculating covariance). Then we can get σ_x^2 by averaging the squares of the first column.

$$\sigma_x = \sqrt{\frac{1}{5}\left[(-1)^2 + 4^2 + (-2)^2 + 3^2 + (-4)^2\right]} = \sqrt{\frac{46}{5}} \approx 3.03$$

A similar calculation will give us that $\sigma_y = \sqrt{\frac{32}{5}} \approx 2.53$.

c) *Calculate the sample covariance for the two variables.* To get the sample covaraince, we average the **product** of each row.

$$\text{cov}(x, y) = \frac{1}{5}\left[(-1)\cdot 1 + 4\cdot 3 + (-2)\cdot(-3) + 3\cdot 2 + (-4)\cdot(-3)\right] = 7$$

d) *What is the correlation coefficient equal to? What does this tell us about the data?* Finally we can simply calculate the correlation coefficient with the formula:

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \approx \frac{7}{3.03 \cdot 2.53} \approx 0.913$$

This confirms what we thought with our initial picture. There is a fairly strong *positive* linear relationship between scores on Quiz 1 and scores on Quiz 12.

4. Using the same data as above, we now want to make the best prediction for how someone who got an 9 on Quiz 1 would've scored on Quiz 12. Do this by first finding the *line of best fit* ($y = ax + b$) using the basic formulas for the MLE of a and b (use what you calculated above).

There are many different formulas for calculating the *slope* (a) of the line of best fit. But the easiest one to remember (that also gives a good way to think about this line) is:

$$a = r\frac{\sigma_y}{\sigma_x}.$$

Why is this easy to remember? The slope of a line is generally "rise over run", i.e. how much do we change in the y-direction per change in the x-direction. The sample standard deviation of y and x roughly capture this idea of how much we are varying in the y-direction based on the variation in the x-direction, so the slope should be similar to $\frac{\sigma_y}{\sigma_x}$. However, since our data is not a perfect line, we have to multiply by r, which tells us how close we are to being a perfect line. For our example,

$$a \approx 0.913 \cdot \frac{2.53}{3.03} \approx 0.762$$

Finally, we can calculate the y-intercept in our line of best fit by noting that $\overline{y} = a\overline{x} + b$ and solving for b, since we know the other 3 variables. Here we get $8 = 0.762 \cdot 8 + b$, so $b \approx 1.901$. So our line of best fit is $y = 0.762x + 1.901$. So if someone get 9 on Quiz 1, we would guess a score of about 8.76 on Quiz 12. Our large r-value gives us a pretty high confidence that this is an accurate guess.

5. You are interested in whether two variables (x and y) are correlated, but due to budget constraints, you can only collect two data points. You get the data points $(1, 8)$ and $(3, 4)$.

   a) *What do you guess the correlation coefficient will be?* Since we only have two points, there is a line that goes perfectly through both of these. So r will equal either +1 or -1, since this line is decreasing, we expect to get $r = -1$.

   b) *Calculate r and explain what this means about your data.* We can make the same chart as above to calculate σ_x, σ_y, and $\text{cov}(x, y)$. Note $\overline{x} = 2$ and $\overline{y} = 6$. Then we get $\sigma_x = \sqrt{\frac{1}{2}(1+1)} = 1$, and

	$x_i - \overline{x}$	$y_i - \overline{y}$
A	-1	2
B	1	-2

   $\sigma_y = \sqrt{\frac{1}{2}(4+4)} = 2$. Finally $\text{cov}(x, y) = \frac{1}{2}((-2) + (-2)) = -2$. So we get $r = \frac{-2}{1 \cdot 2} = -1$, exactly as expected.

   c) *Find the line of best fit.* We could use a variety of techniques to find the line of best fit, since it will just be the line that passes between our two points. But using our general formula $a = r\frac{\sigma_y}{\sigma_x} = -1 \cdot \frac{2}{1} = -2$ and $b = \overline{y} - a\overline{x} = 6 - (-2) \cdot 2 = 10$. So the line is just $y = -2x + 10$.

6. Is there a relationship between the amount of antibody A and antibody B in a sick patient? You take antibody A and B counts per milliliter from 4 patients (in reality you will have a much, much larger sample size).

Patient	Antibody A	Antibody B
A	120	100
B	95	110
C	115	130
D	110	80

   a) *What do you guess the correlation coefficient will be?* If you make a scatter plot of the data, and try to draw a line of best fit, it seems almost flat. This generally corresponds to an r value of somewhere around 0.

b) *What is the correlation coefficient?* Note that $\overline{x} = 110$ and $\overline{y} = 105$. Then we can make the small chart. So we can calculate:

	$x_i - \overline{x}$	$y_i - \overline{y}$
A	10	-5
B	-15	5
C	5	25
D	0	-25

$$\sigma_x = \sqrt{\frac{1}{4}(100 + 225 + 25 + 0)} = \sqrt{\frac{350}{4}} \approx 9.35$$
$$\sigma_y = \sqrt{\frac{1}{4}(25 + 25 + 625 + 625)} = \sqrt{\frac{1300}{4}} \approx 18.03$$
$$\text{cov}(x, y) = \frac{1}{4}(-50 - 65 + 125 + 0) = \frac{10}{4} = 2.5$$
$$r = \frac{2.5}{9.35 \cdot 18.03} \approx 0.015$$

c) *What will our line of best fit look like? If someone has an Antibody A count of 100, should we feel confident in guessing their Antibody B count?* Our r-value is extremely close to 0 as expected. While we can still calculate the line of best fit to get $y = 0.0286x + 101.854$. However, with an r value so close to 0, we should generally expect Antibody A and Antibody B to be not correlated, so we shouldn't use this line to try to make predictions.