Class 20: Chi-Square for Larger Tables (Text: Sections 9.1, 9.2. Section 9.3 is optional)

The Chi-Square Test

Null hypothesis: There is no association between variables.
Alternative hypothesis: There **is** an association between the variables, but we do not specify what.

We have a two-way table of **observed** counts and create a tables of **expected** counts:

$$\text{Expected cell count} = \frac{\text{Row total}\times\text{Column total}}{n}$$

Then we calculate the **Chi-Square** statistic with $df = (\#\text{ rows} - 1)(\#\text{ cols} - 1)$

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Notes on Using the Chi-Square Test

- Large values of χ^2 give significance.
- Small values of χ^2 give no significance.
- Rejecting null hypothesis does not tell us which way the interaction happens.
- Since χ^2 is always positive; there is only an upper tail. We never multiply the probabilities by 2.
- **Conditions**: We can use the Chi-Square test when
  - For tables larger than 2×2: Average of expected counts ≥ 5; all expected counts ≥ 1
  - For 2×2 tables: All expected counts ≥ 5

The example in the previous class could be done either chi-square test of by the z test. Now we do an example with a table that is larger than 2×2, which can only be done by chi-square.

Ex: Is the following drug effective?
In an experiment, 600 people were divided randomly into three groups and given a placebo, a single dose, or a double dose of a drug. The **observed** results were:

	Placebo	Single	Double	
Improve	50	57	78	185
Didn't	180	133	102	415
	230	190	180	600

Step 1:
Null hypothesis: There is no association between variables. That is, the drug has no effect of improvement.
Alternative hypothesis: There **is** an association between the variables. That is, the drug has an effect of improvement.

Step 2:
We calculate the table of **expected** counts:

	Placebo	Single	Double	
Improve	70.9	58.6	55.5	185
Didn't	159.1	131.4	124.5	415
	230	190	180	600

For example

$$\text{Expected cell count} = \frac{\text{Row total}\times\text{Column total}}{n} = \frac{185\cdot 230}{600} = 70.91667$$

Now we calculate the χ^2 statistic:

$$\chi^2 = \frac{(50-70.9)^2}{70.9} + \frac{(57-58.6)^2}{58.6} + \frac{(78-55.5)^2}{55.5} + \frac{(180-159.1)^2}{159.1} + \frac{(133-131.4)^2}{131.4} + \frac{(102-124.5)^2}{124.5} = 22.2 \quad \text{with} \quad \text{df} = (2-1)(3-1) = 2$$

Aside: On the test, you may have to write some of the terms of χ^2 and you will be expected to know its degrees of freedom and find its P-value, but you will not have to calculate the whole χ^2.
Step 3:
From Tables with df =2: We have $P(\chi^2 > 22.2) < 0.0005$
From Calculator:
$P(\chi^2 > 22.2) = \chi^2\text{cdf}(22.2, 50, 2) = 0.000015 = 0.0015\%$

Step 4: We reject the null hypothesis; there is an interaction.

Ex: Since the Chi-Squared test does not tell us how the interaction happens if we reject the null hypothesis, how do we know if the drug is effective? (Maybe it makes people worse?)

Find the conditional distributions:

Conditional Probability	**Placebo**	**Single**	**Double**
Improve	0.22	0.30	0.43
Didn't	0.78	0.70	0.57
	100	1.00	1.00

Thus, drug **is** doing better than placebo.

<u>Ex</u>: **Is a double dose of the drug significantly better than single dose? (Use Chi Squared.)**

Step1:

H_o: Proportion of patients improved independent of dose.
Dosage and improvement rate independent.
H_a: Proportion of patients improved **not** independent of dose.
Dosage and improvement **not** independent.

Step2:

Observed	Single	Double	
Improve	57	78	135
Didn't	133	102	235
	190	180	370

Expected	Single	Double	
Improve	69.3	65.7	135
Didn't	120.7	114.4	235
	190	180	370

$$\chi^2 = \frac{(57-69.3)^2}{69.3} + \frac{(78-65.7)^2}{65.7} + \frac{(133-120.7)^2}{120.7} + \frac{(102-114.3)^2}{114.3} + \frac{(133-131.4)^2}{131.4}$$
$$= 7.09 \text{ with df } = (2-1)\times(2-1) = 1$$

Step 3:
From table, with $df = 1$, we have
$P(\chi^2 > 7.09)$ is between 0.005 and 0.01 (so between 0.5% and 1%)
From calculator
$P(\chi^2 > 7.09) = \chi^2\text{cdf}(7.09, 100, 1) = 0.0077 = 0.77\%$
Step 4:
We reject the hypothesis that there is no interaction. There is a significant difference in the distributions between the two different columns. But we don't yet know whether one dose is better than two.

Conditional probabilities show a double dose is more effective than a single dose. (The chi-square test has already told us that this difference is significant.)

Observed	Single	Double
Improve	0.30	0.43
Didn't	0.70	0.57
	1	1

<u>Ex</u>: **As more women work, the effect of childcare on infants has been investigated**. One study[1] looked at the relationship between infant-mother attachment and the time spent in childcare, with the following results:

	Low (0-3 hours/week)	*Moderate (4-19 hours/week)*	*High (20-54 hours/week)*
Anxious	24	35	5
Secure	11	10	8

(a) Does the data provide evidence of a difference in attachment patterns with the amount of time spent in childcare? Give the hypotheses, the test statistic, and the *p*-value with your conclusion. Use both the 5% and the 1% levels.
(b) Combine the moderate and high childcare groups together, and test again. What is your conclusion?

(a) We use the chi-square test because there are three samples. Null hypothesis: There is no association between attachment and time in day care. Alternate hypothesis: There is an association.
The expected matrix is as follows, where, for example $24.09 = 64 \cdot 35/93$, and so on.

Childcare	*Low (0-3 hours/week)*	*Moderate (4-19 hours/week)*	*High (20-54 hours/week)*	
Anxious	24.09	30.97	8.95	64
Secure	10.91	14.03	4.05	29
	35	45	13	93

Thus the statistic is the $\chi^2$with 2 degrees of freedom

$$\chi^2 = \frac{(24-24.09)^2}{24.09} + \frac{(35-30.97)^2}{30.97} + \frac{(5-8.95)^2}{8.95} + \frac{(11-10.91)^2}{10.91} + \frac{(10-14.03)^2}{14.03} + \frac{(8-4.05)^2}{4.05} = 7.27.$$

From the table, we see that the *p*-value is between 0.025 and 0.05. Using a calculator, we find $\chi^2\text{cdf}(7.27, 100, 2) = 0.0264 = 2.64\%$. Thus, we reject the null hypothesis at the 5% level, but not at the 1% level. We have some evidence, but not strong, that there is a significant interaction between childcare and infant-mother attachment.

(b) We use the same hypotheses. The new observed matrix is

	Low	*Moderate and High*	*Total*
Anxious	24	40	64
Secure	11	18	29
Total	35	58	93

The new expected matrix is

	Low	*Moderate and High*	*Total*
Anxious	24.09	39.91	64
Secure	10.91	18.09	29
Total	35	58	93

Then $\chi^2 = 0.00158$, and $\chi^2\text{cdf}(0.00158, 100, 1) = 0.968 = 98.6\%$, so we do not reject the null hypothesis at any level.

[1] By J. Jacobson and D. Willie, reported in *Statistical Record of Women Worldwide* (Galen Research, 1991) and *Introduction to Practice of Statistics*, 13-th ed, Mendenhall, Beaver and Beaver.

<u>Chi-Square Goodness of Fit Test</u> (Optional: Section 9.3):
<u>Example from UA Biology Major, 2006: Distribution of Creosote Bushes</u>
The question the student answered was whether the number of creosote bushes varied with setting.

*The set-up here is slightly different from the previous example as the data has only **one** observed row. This alters how we calculate the degrees of freedom, but nothing else. **This is called a Chi-Square Goodness of Fit Test.***

Areas of the same size were marked out in four regions. To do this, the pace length was found by walking a known 100 ft and dividing this 100 ft by the number of paces needed to complete the 100 ft.

$$\text{Pace length} = \frac{100 \text{ ft}}{45.5 \text{ paces}} = 2.20 \text{ ft/pace}$$

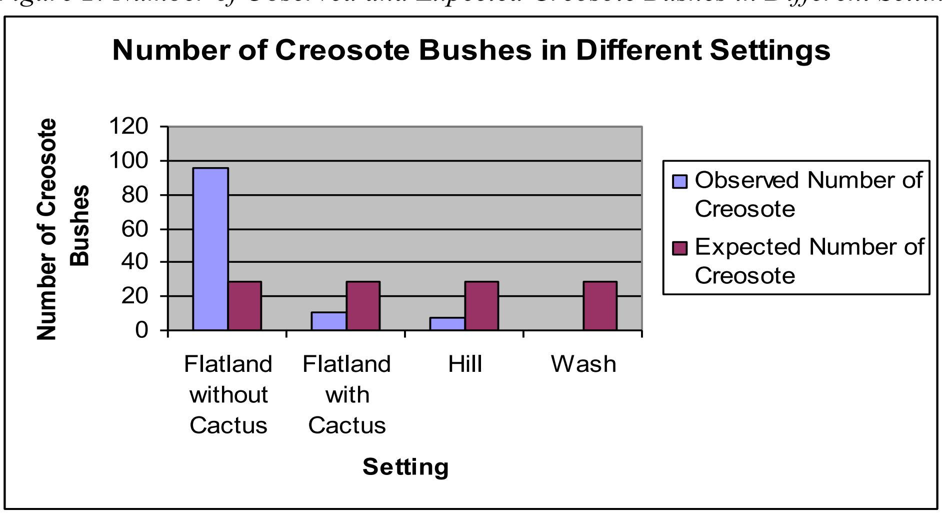
This pace length was then used to measure an area of 95 ft x 95 ft in each region. The numbers of creosote bushes were counted in each one.
Step 1:
Null hypothesis: There is no association between variables. That is, the number of creosote bushes is not affected by region.
Alternative hypothesis: There **is** an association between the variables. That is, the number of creosote bushes is affected by region.
Step 2:
Figure 1: Number of Creosote Bushes Observed and Expected in Each Setting

	Flatland	**Flatland without Cactus**	**Hill**	**Wash**
Number of Creosote Bushes Observed	96	11	7	0
Number of Creosote Bushes Expected	28.5	28.5	28.5	28.5

The expected number of creosote bushes found in each setting was found by as follows:

$$\text{Expected Number of Creosote Bushes} = \frac{\text{Total Number of Creosote Bushes Observed}}{\text{Number of Settings}}$$
$$= \frac{(96 + 11 + 7 + 0)\text{bushes}}{4 \text{ settings}} = 28.5 \text{ creosote bushes per setting}$$

Graphically, we can compare the observed and expected counts:
Figure 2: Number of Observed and Expected Creosote Bushes in Different Settings



$$\chi^2 = \frac{(96 - 28.5)^2}{28.5} + \frac{(11 - 28.5)^2}{28.5} + \frac{(7 - 28.5)^2}{28.5} + \frac{(0 - 28.5)^2}{28.5} = 186.83$$

For a Goodness of Fit Test:

$$\text{Degrees of Freedom} = (\text{Number of Variables} - 1)$$

In this case, Degrees of Freedom $= (\text{Number of Variables} - 1) = 3$

Step 3:

From table, with $df = 31$, we have
$P(\chi^2 > 186.83)$ is less that $0.0005 = 0.05\%$. The P-value is approximately 0.
From calculator
$P(\chi^2 > 186.83) = \chi^2\text{cdf}(186{,}83, 10020, 1) = 3 \cdot 10^{-40}$. Tiny!

Step 4:

We reject the null hypothesis. There is strong evidence that the number of creosote bushes depends on the setting.