Packet 6: Chi-Square Test for Independence Textbook pages: 18 – 24; 624 – 631

After completing this material, you should be able to:

- calculate the expected count for any cell of a contingency table.
- find marginal distributions for a contingency table and use those to construct a side-by-side bar graph.
- compute chi-square contributions for any cell of a contingency table.
- conduct the chis-square test (with the aid of StatCrunch output) to determine if two categorical variables are related.

The legalization of medicinal marijuana has been a hotly contested subject. A survey conducted in April 2015 was undertaken to investigate whether a relationship exists between feelings on the legalization of medicinal marijuana (for/against) and political party (Republican/Democrat/Independent).

During this study, a total of 500 individuals were surveyed. For each American adult, what variables were recorded? Are these variables categorical or quantitative?

- Political Party (R/D/I)
- Opinion about legalization

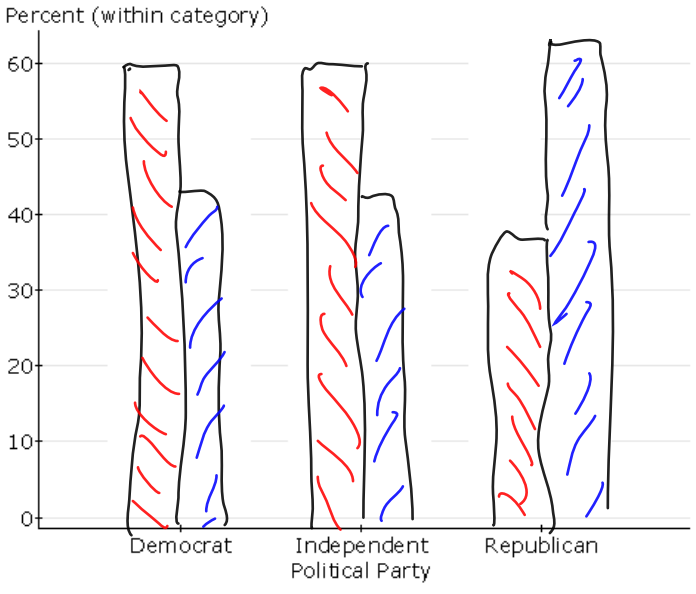
Because two variables were recorded for each individual, we need a way to organize this data. A **contingency table** categorizes counts on two (or more) categorical variables – in other words, this table summarizes the number of individuals in all possible combinations of categories. We can summarize the responses to the survey in the table below:

		Legalization of Marijuana		row sums
		Supports	**Does not support**	
Political Party	**Democrat**	116	84	200
	Republican	74	126	200
	Independent	59	41	100
col sums		249	251	500 = n

Instead of looking at the counts, let's split the table into *marginal distributions* – that is, what percentage of each political party surveyed gave each of the two responses?

	Support	Oppose
Dems	116/200 = 58%	42%
Reps	74/200 = 37%	63%
Ind	59/100 = 59%	41%

gave % between two?



In order to determine if the differences are *significant*, we need to conduct a hypothesis test. **One can never simply examine sample data and draw some conclusion about the population – we need to conduct a hypothesis test in order to determine if the results are significant**.

What is the goal of the **chi-square test for independence**? To determine whether two categorical variables are independent or not.

If we wanted to conduct this test on the legalization of marijuana data, what **hypotheses** would be tested?

H_0: Political Party & support for medical marijuana are unrelated (hence independent)

H_a: Political Party & support are related. ($\neq$ alternative)

In order to conduct a hypothesis test, we need some quantity to compare the observed counts from the survey to. This is referred to as the **expected count** (in other words, what should we have observed if the null hypothesis were true). Fill in the table below with the expected counts.

from data

Observed Counts

Political Party	Legalization of Marijuana: Supports	Legalization of Marijuana: Does not support
Democrat	116	84
Republican	74	126
Independent	59	41
	249	251

500

$\frac{249}{500} = .498$ support legalization

Expected Counts

Political Party	Legalization of Marijuana: Supports	Legalization of Marijuana: Does not support	
Democrat	$\frac{249}{500} \cdot 200 = 99.6$	100.4	200
Republican	99.6	100.4	200
Independent	$\frac{249}{500} \cdot 100 = 49.8$	50.2	100
	249	251	

What do you notice when the observed (top table) and expected counts (bottom table) are compared?

Quite a difference! If they were similar it would support independence - but we're not seeing that!

The **test statistic** for the chi-square test for independence compares the observed and expected counts. Its formula is the following:

$$\chi^2 = \sum \frac{(obs - exp)^2}{exp}$$

Sigma – Sum over all cells

Formula Alert!! This formula will be given on the formula sheet.

Let's look at how this test statistic is calculated by going back to the marijuana example:

$$\chi^2 = \frac{(116 - 99.6)^2}{99.6} + \ldots + \frac{(41 - 50.2)^2}{50.2} = 21.87$$

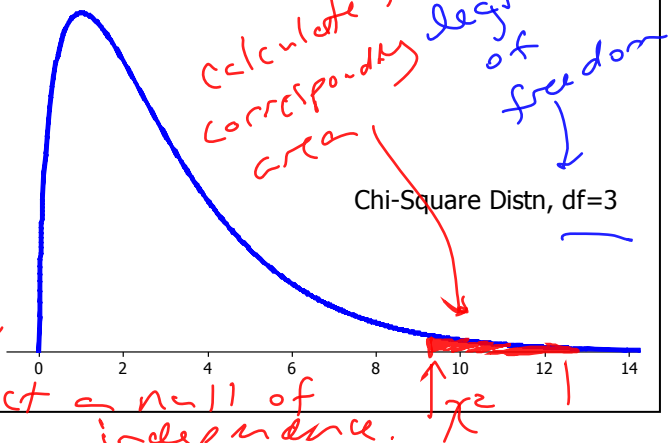
not an integer but not a problem!

What is the Chi-Square distribution? Different than a normal, for sure!

In statistical inference, there are several common distributions used for inference. In addition to the normal distribution (which we have already used), the chi-square distribution is also a common distribution used for inference.

How is the chi-square distribution used to find a probability?

Compute χ^2; look to the area to the right of χ^2 for the prob of a χ^2 that extreme or more so. Compare to α & decide if we reject a null of independence.



In general, we won't calculate the chi-square test statistic (the calculation can be tedious) or the p-value associated with the test. Instead, we will rely on StatCrunch output for our calculations. Let's look at the StatCrunch output for the legalization of marijuana example:

	Supports	Does not support	Total
Democrat	116 (99.6) (2.7)	84 (100.4) (2.68)	200
Republican	74 (99.6) (6.58)	126 (100.4) (6.53)	200
Other	59 (49.8) (1.7)	41 (50.2) (1.69)	100
Total	249	251	500

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	2	21.87235	<0.0001

observed

expected

contribution to χ^2 from that cell.

probability to the right of χ^2.

Complete the appropriate hypothesis test using a significance level of 0.05 to determine if political party and support of legalization of marijuana are related, or independent.

(1) H_0: party & opinion are independent; H_a: no, they're related.

(2) $\alpha = .05$; reject H_0 if prob. (p-value) $< \alpha$.

(3) T.S. $\chi^2_{3\,df} = 21.87$; p-value $< .0001 < \alpha$

(4) Because the p-value (prob of χ^2 that extreme or more so) is so ridiculously small ($\ll \alpha$), we reject the null of independence in favor of dependence: Republicans favor legalization less than others.

Example: All new drugs must go through a drug study before being approved by the FDA. A drug study typically includes clinical trials whereby participants are randomized to receive different dosages as well as a placebo. To control as many factors as possible, it is best to assign participants randomly across the treatments. A recent study for a new drug consisted of two dosages (10mg, 20mg) and a placebo. Those who designed the study would like to know if the dosage assigned was related to the participants' gender. The responses are summarized in the StatCrunch output below:

	10mg	20mg	Placebo	Total
Female	54 57.33 0.1938	56 55.33 0.008	60 57.33 0.124	170
Male (obs / exp / ·)	32 28.67 0.3868	27 27.67 0.0161	26 28.67 0.2481	85
Total	86	83	86	255 (n)

Compute the **expected** number of females receiving the placebo. What does this quantity mean?

$\frac{86}{255} \cdot 170 = 57.33$ of those

Relative frequency receiving placebo times the number of females.

$\frac{86 \cdot 170}{255} = \frac{\text{row sum} \cdot \text{col sum}}{n}$

Find the marginal distribution for each gender by filling in the tables below.

	Dosage		
	10mg	**20mg**	**Placebo**
Female	$\frac{54}{170} = 31.76\%$	$\frac{56}{170} = 32.94\%$	$\frac{60}{170} = 35.29\%$

	Dosage		
	10mg	**20mg**	**Placebo**
Male	37.65%	31.76%	30.59%

Based on these distributions, do you believe gender is somehow related to dosage? Explain.

There are differences, but not that big; could just be random effects.

Compute the **chi-square contribution** for male participants who were given 10mg of the drug.

$$\frac{(32-28.67)^2}{28.67} = .3868$$

Using the StatCrunch output below, conduct the appropriate test to determine if there is a relationship between gender and the dosage received. Use a significance level of 0.01.

$\alpha = 0.01$

Chi-Square test for independence:

Measure	DF	Value	P-value
Chi-Square	2	0.9775848	0.6134

(1) H_0: independence of gender & dosage
H_a: they're related,

(2) $\alpha = .01$; reject H_0 in favor of H_a if p-value $< \alpha = .01$

(3) TS, $\chi^2_{2 \text{ df}} = .9776$ with p-value = .6134
So p-value = prob $> \alpha$.

(4) Since the p-value is $> \alpha$, we fail to reject a null of independence.

What assumptions must be satisfied for the chi-square test of independence to be valid?

(1) The data are from a random sample

(2) The expected counts is > 5 or more in each cell.

Example: A sample of 1000 traffic crashes occurring in either Kentucky or Ohio was selected from the National Highway Safety Traffic Administration database. For each crash, it was noted whether or not alcohol was involved in the accident. A reporter has questioned whether there is a relationship between alcohol involvement and the state in which the accident occurred. The information gathered is summarized in the StatCrunch output provided.

Contingency table results:

	No	Yes	Total
Kentucky	289 (0.21)	85 (0.82)	374
Ohio	505 (0.13)	121 (0.49)	626
Total	794	206	1000

Statistic	DF	Value	P-value
Chi-square	1	1.65	0.1986

— Compute the number of accidents one would expect to involve alcohol in KY if there is no relationship.

Yes: $\frac{206 \cdot 374}{1000} = 77.04$

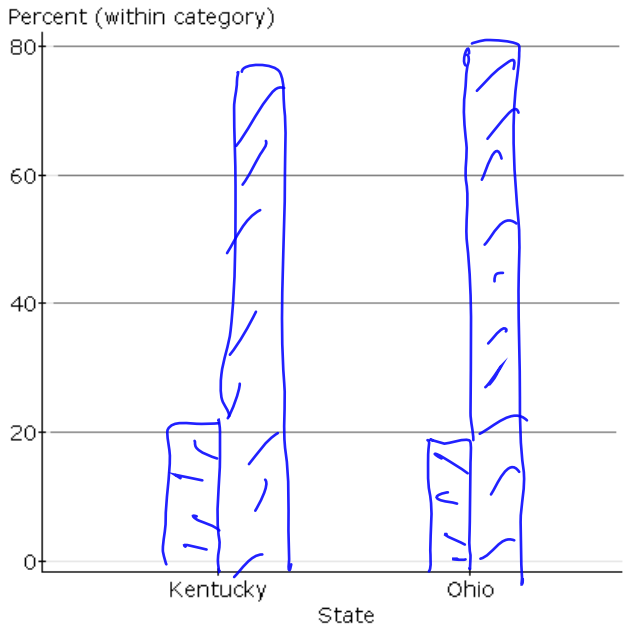
No: $\frac{794 \cdot 206}{1000} = \frac{296.96}{374}$

.21 + .82 + .13 + .49 =

— Fill in the tables below with the marginal distributions for each state. Then create a side-by-side bar graph comparing the percentages for the two states.

	Alcohol – yes	Alcohol – no
KY	$\frac{85}{374} = 22.73\%$	77.27 %

	Alcohol – yes	Alcohol – no
OH	19.33%	80.67 %



— Conduct the appropriate test to address the conjecture made by the reporter. Use a significance level of 0.05.

① H_0: independence of state & alcohol related crashes
H_a: they are related.

② $\alpha = .05$; reject independence if p-value $< \alpha$

③ T.S. $\chi^2_{1=df} = 1.65$; p-value $= 0.1986 > \alpha$; so we

④ Fail to reject a null of independence - seems like state & alcohol-related crashes is independent.