Inference for Difference Between 2 Population Proportions

Chapter 10	Chapter 11
We are interested in whether 2 pop. means are different.	We are interested in whether 2 pop. proportions are different.
Example:	Example:
Mean Income in Bradford = \$38,010	2013 Data Entry Error Rate = 0.142
Mean Income in Kane = \$35,006	2014 Data Entry Error Rate = 0.107
difference = \$3,004	difference = 0.035
sample means	sample proportions

Are the differences due to sampling Error or are the significant differences that allow us to conclude that population parameters are different.

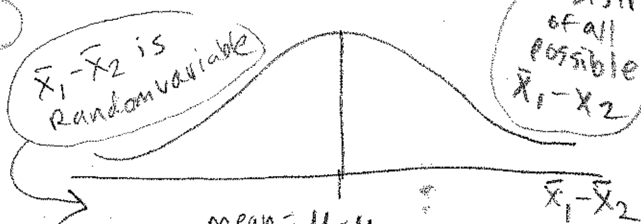
1. Confidence Interval for difference between 2 pop. proportions
2. Hypothesis Test for difference between 2 pop. proportions
3. Hypothesis test to check Equality of 2 or more pop. proportions using chi-square Test statistic
4. Test of Independence of 2 categorical variables using chi-square test statistic

---

Chapter 10	Chapter 11
Sampling Distribution of Xbar$_1$ − Xbar$_2$	Sampling Distribution of Pbar$_1$ − Pbar$_2$

Chapter 10

Sampling Distribution of Xbar$_1$ − Xbar$_2$



mean $= \mu_1 - \mu_2$

$= E(\text{Xbar}_1 - \text{Xbar}_2)$

$= \dfrac{\text{sum of all possible } \bar{x}_1-\bar{x}_2}{\text{count of all } \bar{x}_1-\bar{x}_2}$

Standard Error of Sampling Distribution of Xbar$_1$ − Xbar$_2$ $\Big\}$ SD $= \sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}$

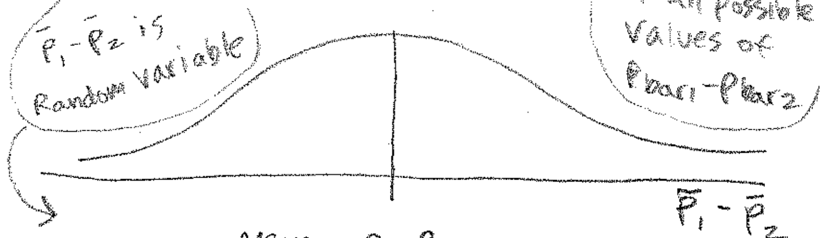
Correction Factor

$\sqrt{\dfrac{N-n}{N-1}}$

Needed when $\dfrac{n}{N} > 0.05$

Chapter 11

Sampling Distribution of Pbar$_1$ − Pbar$_2$



Mean $= p_1 - p_2$

$= E(\text{Pbar}_1 - \text{Pbar}_2)$

$= \dfrac{\text{sum of all possible } \bar{p}_1-\bar{p}_2}{\text{count of all } \bar{p}_1-\bar{p}_2}$

Standard Error/Deviation of sampling Distribution of Pbar$_1$ − Pbar$_2$ $\Big\}$ SD $= \sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\dfrac{p_1(1-p_1)}{n_1}+\dfrac{p_2(1-p_2)}{n_2}}$

In this chapter we wont Know population Proportions p_1 & p_2 so we will use $\bar{p}_1$ & $\bar{p}_2$ in formula to estimate $\sigma_{\bar{p}_1-\bar{p}_2}$

★ see Excel example to Prove these formulas work.

Remember Proportions from chapter 7:

Sample Proportion $= \bar{p} = \dfrac{x}{n}$ = Random variable

x = number of elements in sample that possess the characteristic of interest

x = categorical variable

n = sample size

Sampling Distribution of Pbar can be approximated by a normal Distribution whenever:

$$n * p \geq 5$$

$$n * (1-p) \geq 5$$

$$\left\{ \begin{array}{c} \text{Expected} \\ \text{value of} \\ \text{Pbar} \end{array} \right\} = E(\bar{p}) = p$$

mean of all possible $\bar{p}$

population proportion

$$\left\{ \begin{array}{c} \text{Standard} \\ \text{Error /} \\ \text{Deviation} \\ \text{of Sampling} \\ \text{Distribution} \\ \text{of Pbar} \end{array} \right\} = \sigma_{\bar{p}} = \sqrt{\frac{p*(1-p)}{n}} * \underbrace{\sqrt{\frac{N-n}{N-1}}}_{\text{Correction Factor needed when } \frac{n}{N} > 0.05}$$

① Confidence Interval for Difference Between 2 Population Proportions $P_1 - P_2$

* From Sample Data, we calculate a point estimate for $P_1 - P_2$, $\bar{P}_1 - \bar{P}_2$, and then add a margin of error to both sides to get an interval (lower & upper value) that will contain the population proportion difference $(P_1 - P_2)$ about 95 out of a 100 times.
* if we don't know P_1 or P_2, we use our sample proportions $\bar{P}_1$ and $\bar{P}_2$ in all of our calculations. (That is why we use word "Estimate" on SE calc.)
* we must take 2 Independent and random samples.
* we can use the Normal Distribution to estimate the sampling Distribution of $\bar{P}_1 - \bar{P}_2$ if 4 tests are passed:

① $n_1 * \bar{P}_1 \geq 5$ ③ $n_2 * \bar{P}_2 \geq 5$

② $n_1 * (1-\bar{P}_1) \geq 5$ ④ $n_2 * (1-\bar{P}_2) \geq 5$

$$\underbrace{\bar{P}_1 - \bar{P}_2}_{\text{point Estimate of } P_1 - P_2} \pm \overbrace{\underbrace{Z_{\alpha/2}}_{\text{\# of Standard Deviations}} * \underbrace{\sqrt{\frac{\bar{P}_1 * (1-\bar{P}_1)}{n_1} + \frac{\bar{P}_2 * (1-\bar{P}_2)}{n_2}}}_{\text{Estimate of Standard Deviation/ Standard Error for Sampling Distribution of } \bar{P}_1 - \bar{P}_2}}^{\text{Margin of Error}}$$

Variables Defined for Confidence Interval:

P_1 = population 1 proportion

P_2 = population 2 proportion

$\bar{P}_1$ = P_{bar1} = sample proportion from pop. 1

$\bar{P}_2$ = P_{bar2} = sample proportion from pop. 2

n_1 = sample size from population 1

n_2 = sample size from population 2

α = alpha = risk that the population proportion, $P_1 - P_2$, is not in interval

$1 - \alpha$ = Confidence Level / Coefficient = How sure we are that $P_1 - P_2$ is in interval

$\sqrt{\frac{N-n}{N-1}}$ = correction factor for Standard Error calculation if $\frac{n}{N} > 0.05$

② Hypothesis Test to check if there is <u>No Difference</u> Between 2 population proportions $P_1 - P_2$ (Z method, Hypothesized Difference = 0) (P.5)

* From Sample Data, we run our 5-step Hypothesis Test "No Difference" between 2 population proportions.

* In this section we run H.T. for $P_1 - P_2$ using the <u>Z method</u>. Next section we will see how to check difference between 2 or more population proportions using Chi-square method. Chi-square can only do 2-tail Test.

* <u>Z method</u> Allows us to run 3 types of Tests:

① 1 tail, Lower	② 1 tail upper	③ 2 tail
$H_0: P_1 - P_2 >= 0$	$H_0: P_1 - P_2 <= 0$	$H_0: P_1 - P_2 = 0$
$H_a: P_1 - P_2 < 0$	$H_a: P_1 - P_2 > 0$	$H_a\ P_1 - P_2 <> 0$

* we must take 2 Independent & Random Samples

* We can use Normal Distribution to estimate the Sampling Distribution of $\bar{P}_1 - \bar{P}_2$ if 4 Tests are passed:

① $n_1 * \bar{P}_1 \geq 5$ ③ $n_2 * \bar{P}_2 \geq 5$

② $n_1 * (1 - \bar{P}_1) \geq 5$ ④ $n_2 * (1 - \bar{P}_2) \geq 5$

Formulas for Z method Hypothesis Test of $p_1 - p_2$

p.6

① If Ho is TRUE as an Equality:

$$p_1 - p_2 = 0$$

$$\downarrow$$

$$p_1 = p_2$$

population proportion becomes: $p_1 = p_2 = p$

② Estimate of Standard Error of $\bar{p}_1 - \bar{p}_2$ when $p_1 = p_2 = p$:

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1 * (1-p_1)}{n_1} + \frac{p_2 * (1-p_2)}{n_2}} = \sqrt{p * (1-p) * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

if $\rightarrow p_1 = p_2 = p \rightarrow$ then

Now we just need estimate of p

③ Pooled Estimator of p when $p_1 = p_2 = p$

{if $p_1 = p_2 = p$ we can just take weighted Average} → {Pooled Estimator of p}

$$= \bar{p} = \frac{n_1 * \bar{p}_1 + n_2 * \bar{p}_2}{n_1 + n_2} \leftarrow$$

weighted Ave. we learned in ch.3

④ Z Test statistic

$$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p} * (1-\bar{p}) * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

← Point Estimate of Difference

← SE

Variables Defined for Hypothesis Test of $P_1 - P_2$

(P.7)

P_1 = population 1 proportion

P_2 = population 2 proportion

$\bar{P}_1$ = P_{bar_1} = sample proportion from Pop. 1

$\bar{P}_2$ = P_{bar_2} = sample proportion from Pop. 2

P = Population proportion when $P_1 = P_2$

$\bar{P}$ = Estimate of P when we assume $P_1 = P_2 = P$.
called "Pooled Estimator of P" or weighted Average of $\bar{P}_1$ and $\bar{P}_2$.

n_1 = sample size taken from population 1

n_2 = sample size taken from population 2

α = Risk that we Reject H_0 when it is TRUE (Type 1 Error)

2013	2
No	No
No	No
No	No
No	No
No	No
Yes	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	Yes
No	No
No	No
Yes	No
No	Yes
No	Yes
No	No
No	No
No	Yes
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	Yes
No	No
No	No
No	Yes
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	Yes
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	Yes

An HMO wanted to check the accuracy of patient contact information data entry
It took independent random samples of records and rechecked the accuracy of the data.
No = no errors, data accurate
Yes = Yes, Errors were found
Is there a significant reduction in errors from 2013 to 2014? >>>>>>
p_1 = proportion for population 1
p_2 = proportion for population 2
p_{bar1} = sample proportion for simple random sample from population 1
p_{bar2} = sample proportion for simple random sample from population 2

Variables		Proportion 1	Proportion 2	Criteria
		2013	2014	
n	Count = Sample Size = n =	316	382	
	Count 'Yes' =	45	41	Yes
	Count 'No' =	271	341	No
Pbar	Yes Sample proportion = Pbar	0.142405063	0.107329843	
1-Pbar	No Sample proportion = 1-Pbar	0.857594937	0.892670157	

	Point Estimate of $p_1 - p_2$
Pbar1 - Pbar2 =	0.03507522

	Can we use Normal Curve?		Criteria
p*n>5?	45	41	5
(1-p)*n>5?	271	341	5

		Calculations for Confidence Interval		
Pbar*(1-Pbar)/n	Used For Confidence Interval	0.000386474	0.000250812	
	Alpha	0.05		Risk Pop. Proportion not in interval
	Alpha/2	0.025		
	Confidence Coefficient/Level	0.95		
	Upper Z	1.959963985		
	SE	0.025244527	= SQRT(p_{bar1}*(1-p_{bar1})/n_1 + p_{bar2}*(1-p_{bar2})/n_2)	Standard Error/Deviation for Sampling Distribution of Pabar1 - Pbar2
#SD*SD	Margin of Error	0.049478364		
	Interval Upper	0.084553584		
	Interval Lower	-0.014403144		

Conclusion: We are 95% sure that the population proportion for difference between 2013 errors and 2014 errors is between -0.0144 and 0.0846.
Because zero is in our interval, it seems reasonable to assume that there is not a significant reduction in errors

		Calculations for Hypothesis Testing	
	Hypothesized Difference	0	
Step 1:	$H_0 : p_1 - p_2$	<=	0
	$H_a : p_1 - p_2$	>	0
Step 2:	Alpha	0.05	
Step 3:	"Pooled" Estimate of p = p_{bar} =	0.123209169	(n_1*p_{bar1}+n_2*p_{bar2})/(n_1+n_2)
	SE =	0.024993208	SQRT(p_{bar}*(1-p_{bar})*(1/n_1 + 1/n_2))
	Z Test Statistic =	1.40339009	(p_{bar1} - p_{bar2})/SE
Step 4:	Critical Value (Hurdle)	1.644853627	
	p-value	0.080250273	

Because Question is structured this way:
Is there a significant reduction in errors from 2013 to 2014?
And sample 1 = 2013 and Sample 2 = 2014
The difference between (2013 - 2014) would have to be significantly positive
This means we set Alternative Hypothesis up as test on the high end like:
$p_1 - p_2 > 0$
We want to conclude Yes/No there a significant reduction in errors from 2013 to 2014 so we set up alternative this way

Because we are interested in "No Difference", p1 = p2 = p
If H0 is TRUE as an equality:
$p_1 - p_2 = 0$
$p_1 = p_2$
Pop. Proportion becomes: $p_1 = p_2 = p$
SE = SQRT
SE = SQRT(p_1*(1-p_1)/n_1 + p_2*(1-p_2)/n_2) becomes: SQRT(p*(1-p)*(1/n_1 + 1/n_2))
Where p will be estimated by "Pooled Estimator of p when p_1 = p2 = p"
p_{bar} = Weighted Average = Pooled Estimator of p when $p_1 = p_2 = p$ ==>> (n_1*p_{bar1}+n_2*p_{bar2})/(n_1+n_2)
z = (p_{bar1} - p_{bar2})/SQRT(p_{bar}*(1-p_{bar})*(1/n_1+1/n_2))

Step 5: Because our test statistics is not past our hurdle and the p-value is bigger than our alpha, we fail to reject the Null Hypothesis.
The evidence does not suggest that there has been a significant reduction in errors.
We do run a 5% risk that the null was not rejected even though the Alternative was TRUE.

Advantage of z method:
Can do two-tail test or one tail test.
Disadvantage of z method:
Can only compare two proportions (could compare more but would have build up of Type 1 error)

P.8

2013	2014
No	No
No	No
No	No
No	No
No	No
Yes	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	Yes
No	No
No	No
Yes	No
No	Yes
No	Yes
No	No
No	No
No	Yes
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	Yes
No	No
No	No
No	Yes
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	Yes
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	Yes

An HMO wanted to check the accuracy of patient contact information data entry
It took independent random samples of records and rechecked the accuracy of the data.
No = no errors, data accurate
Yes = Yes, Errors were found
Is there a significant reduction in errors from 2013 to 2014?
p_1 = proportion for population 1
p_2 = proportion for population 2
p_{bar1} = sample proportion for simple random sample from population 1
p_{bar2} = sample proportion for simple random sample from population 2

>>>>>>

Variables		Proportion 1	Proportion 2	Criteria
		2013	2014	
n	Count = Sample Size = n =	=COUNTA(A2:A317)	=COUNTA(B2:B383)	
	Count 'Yes' =	=COUNTIFS($A$2:$A$317,I14)	=COUNTIFS($B$2:$B$383,I14)	Yes
	Count 'No' =	=COUNTIFS($A$2:$A$317,I15)	=COUNTIFS($B$2:$B$383,I15)	No
Pbar	Yes Sample proportion = Pbar	=F14/F$13	=G14/G$13	
1-Pbar	No Sample proportion = 1-Pbar	=F15/F$13	=G15/G$13	

	Point Estimate of $p_1 - p_2$
Pbar1 - Pbar2 =	=F16-G16

	Can we use Normal Curve?		Criteria
p*n>5?	=F16*F$13	=G16*G$13	5
(1-p)*n>5?	=F17*F$13	=G17*G$13	5

		Calculations for Confidence Interval	
Pbar*(1-Pbar)/n	Used For Confidence Interval	=F16*F17/F13	=G16*G17/G13
	Alpha	0.05	
	Alpha/2	=F28/2	
	Confidence Coefficient/Level	=1-F28	
	Upper Z	=NORM.S.INV(1-F29)	
	SE	=SQRT(SUM(F27:G27))	$= SQRT(p_{bar1}*(1-p_{bar1})/n_1 + p_{bar2}*(1-p_{bar2})/n_2)$
#SD*SD	Margin of Error	=F31*F32	
	Interval Upper	=F20+F33	
	Interval Lower	=F20-F33	

Risk Pop. Proportion not in interval

Standard Error/Deviation for Sampling Distribution of Pabar1 - Pbar2

Conclusion: #VALUE!

Because zero is in our interval, it seems reasonable to assume that there is not a significant reduction in errors

		Calculations for Hypothesis Testing		
	Hypothesized Difference	0		
Step 1:	$H_0 : p_1 - p_2$	<=	0	
	$H_a : p_1 - p_2$	>	0	
Step 2:	Alpha	0.05		
Step 3:	"Pooled" Estimate of $p = p_{bar}$ =	=SUMPRODUCT(F13:G13,F16:G16)/SUM(F13:G13)	$(n_1*p_{bar1}+n_2*p_{bar2})/(n_1+n_2)$	
	SE =	=SQRT(F48*(1-F48)*(1/F13+1/G13))	$SQRT(p_{bar}*(1-p_{bar})*(1/n_1 + 1/n_2))$	
	Z Test Statistic =	=F20/F49	$(p_{bar1} - p_{bar2})/SE$	
Step 4:	Critical Value (Hurdle)	=NORM.S.INV(1-F46)		
	p-value	=1-NORM.S.DIST(F50,1)		

Step 5: Because our test statistics is not past our hurdle and the p-value is bigger than
our alpha, we fail to reject the Null Hypothesis.
The evidence does not suggest that there has been a significant reduction in errors.
We do run a 5% risk that the null was not rejected even though the Alternative was TRUE.

Because Question is structured this way:
Is there a significant reduction in errors from 2013 to 2014?
And sample 1 = 2013 and Sample 2 = 2014
The difference between (2013 - 2014) would have to be significantly positive
This means we set Alternative Hypothesis up as test on the high end like:
$p_1 - p_2 > 0$
We want to conclude Yes/No there a significant reduction in errors from 2013 to 2014 so we set up alternative this way

Because we are interested in "No Difference", p1 = p2 = p
If H0 is TRUE as an equality:
$p_1 - p_2 = 0$
$p_1 = p_2$
Pop. Proportion becomes: $p_1 = p_2 = p$
SE = SQRT
$SE = SQRT(p_1*(1-p_1)/n_1 + p_2*(1-p_2)/n_2)$ becomes: $SQRT(p*(1-p)*(1/n_1 + 1/n_2))$
Where p will be estimated by "Pooled Estimator of p when $p_1 = p2 = p$"
p_{bar} = Weighted Average = Pooled Estimator of p when $p_1 = p_2 = p$ ==>> $(n_1*p_{bar1}+n_2*p_{bar2})/(n_1+n_2)$
$z = (p_{bar1} - p_{bar2})/SQRT(p_{bar}*(1-p_{bar})*(1/n_1+1/n_2))$

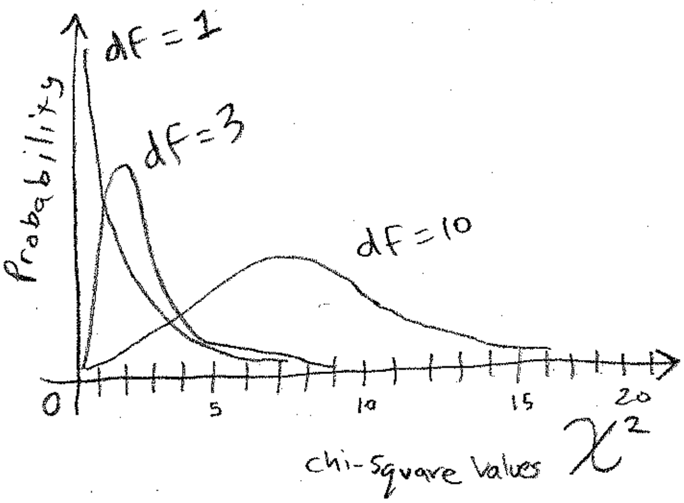
Advantage of z method:
Can do two-tail test or one tail test.
Disadvantage of z method:
Can only compare two proportions (could compare more but would have build up of Type 1 error)

P. 9

(3) Hypothesis Test to check Equality of 2 or more Population Proportions using Chi-square

χ^2 New Distribution

Chi-square ("Kie" like "pie") Distribution



(1) Family of chi-square Distributions based on Degrees of Freedom.
where:
K = # of populations
r = # of rows or # of categorical variables

For 2 categorical variables	$df = K - 1$
For 3 or more categorical variables	$df = (r-1)*(K-1)$

*called Multi Nomial population

(2) Continuous Distribution where Area = Probability. All Area = 1

(3) postive skew, but starts to approach Normal at df = 10

(4) Because of calculations, χ^2 never negative. It is always an upper test.

χ^2 "Chi-square" Test statistic

* using sample Data, χ^2 test statistic can be used to test whether **2** or more population proportions are equal.

* χ^2 can be used to check Equality of 2 pop. proportions, but only for a 2-tail Test.

* To calculate the χ^2 test statistic and p-value for χ^2 we will compare "Observed Frequencies" for our proportions to the calculated "Expected Frequencies" for our proportions in a multiple step process that uses "Cross Tab" tables (Remember: chapter 2) also know as "Contingency tables".

* The χ^2 test statistic calculations for "testing 2 or more pop. proportions" in this section is similar to the calculations for a "Test of Independence" in the next section. The diffence is that "Test of Independence" has only 1 population, whereas "Testing 2 or more pop. proportions" has multiple Populations.

* If you are checking Equality of 2 pop. proportions & you do Z method & χ^2 method, then: $(Z \text{ test statistic})^2 = \chi^2$ test statistic

Variables for χ^2 Hypothesis Test of Equality of 2 or more population proportions

p_1 = population 1 proportion

p_2 = population 2 proportion

$\vdots$

p_K = population K proportion

α = Alpha = Risk that we Reject Ho when it is TRUE

K = # of populations

r = # of rows in "Cross Tab" table or # of categorical variables.

$\bar{p}_1$ = p_{bar_1} = sample proportion from pop. 1

$\bar{p}_2$ = p_{bar_2} = sample proportion from pop. 2

$\vdots$

$\bar{p}_K$ = p_{bar_K} = sample proportion from pop. K

df = Degrees of Freedom

f_{ij} = observed Frequency for Row i, column j in "Cross Tab" table

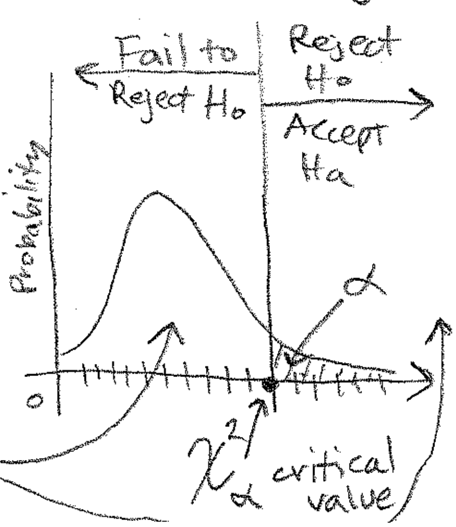
e_{ij} = Expected Frequency for Row i, column j

χ^2 = calculated Test statistic

χ^2_α = critical value (Hurdle)

p-value = $P(\geq \chi^2)$

Is calculated Test statistic χ^2?



Example of Hypothesis Test for Equality of 2 or more Population Proportions

Question: Executive of HMO wanted to check the accuracy of patient contact information. A sample of patient records was taken for each of the years: 2011, 2012, 2013, 2014. Are error rates the same for each year?

observed frequencies

"Cross Tab" table with Frequencies

Errors / Years	2011 (Population 1)	2012 (Population 2)	2013 (Population 3)	2014 (Population 4)	Total
Yes	39	43	45	41	168
No	304	223	271	341	1139
Total	343	266	316	382	1307

Sample sizes

sample from pop. 1 — sample from pop. 2 — sample from pop. 3 — sample from pop. 4

We are interested in P(yes)

	2011 $\bar{p}_1$	2012 $\bar{p}_2$	2013 $\bar{p}_3$	2014 $\bar{p}_4$	overall P(yes)
Yes	$\frac{39}{343} = 0.114$	$\frac{43}{266} = 0.162$	$\frac{45}{316} = 0.142$	$\frac{41}{382} = 0.107$	$\frac{168}{1307} = 0.129$

* If each $\bar{p}$ (sample proportion) was equal to 0.129, then there would be zero difference (All Equal)

* Remember, some difference (sampling error) is acceptable. We need to check for <u>significant</u> Difference.

The essence of how we will check to see if there is a <u>significant</u> Difference in the proportions is by comparing:

Observed Frequencies and Expected Frequencies

All observed Frequencies

Errors / Years	2011	2012	2013	2014	Total
yes	39	43	45	41	168
No	304	223	271	341	1139
Total	343	266	316	382	1307

Sample sizes

sum of all sample sizes

* Remember, overall $P(\text{yes}) = \frac{168}{1307} = 0.129$

Observed Frequency of "Yes" in 2013 = 45

Sample size in 2013 = 316

2013 $\bar{P}_3 = \frac{45}{316} = 0.142$

* if P(yes in 2013) were to equal 0.129 then:

$\left\{\begin{matrix}\text{sample}\\ \text{size}\\ 2013\end{matrix}\right\} * 0.129 = 316 * 0.129 = 40.6$

"Expected Frequency" if all proportions Equal!!!

check $\frac{40.6}{316} = 0.129$ ✓

5 steps for Hypothesis Test to check Equality of 2 or more population proportions

Step 1 Null & Alternative Hypothesis:

$H_0: p_1 = p_2 = \cdots = p_K$

H_a: Not all population proportions are equal

1 or more pop. proportions differ from other pop. proportions

Step 2a Alpha α = Risk that we reject H_0 when it was TRUE.

Step 2b Select Random Samples from each of the populations & create "Cross Tab" table with observed Frequencies.

Example:

"observed Frequencies" $= f_{ij}$

i = Row Number = categorical variable Number

j = column Number = Population number

		Pop. 1 ($j=1$)	Pop. 2 ($j=2$)	Pop. 3 ($j=3$)	Pop. 4 ($j=4$)	
	Errors / Years	2011	2012	2013	2014	Total
$r=1$	Yes	39	43	45	41	168 ← overall yes total
$r=2$	No	304	223	271	341	1139 ← overall No total
sample sizes →	Total	343	266	316	382	1307 ← Sum of all sample sizes

$f_{ij} = f_{13} = 45$ = observed Frequency Row 1, column 3

Step 3

Assume H_0 TRUE and compute Expected Frequencies e_{ij}

P. 15

e_{ij} = "Expected Frequencies"

i = Row Number = categorical variable Number

j = Column Number = population Number

$$e_{ij} = \frac{\text{Row } i \text{ Total}}{\text{Sum of all sample sizes}} * \left(\text{Column } j \text{ Total}\right)$$

Observed Frequencies →

		$j=1$	$j=2$	$j=3$	$j=4$	Row Totals
	Errors/Years	2011	2012	2013	2014	Total
$i=1$	Yes	39	43	45	41	168
$i=2$	No	304	223	271	341	1139
Column Totals →	Total	343	266	316	382	1307

$$e_{ij} = e_{13} = \frac{168}{1307} * 316 = 0.129 * 316 = 40.6$$

Grand over Probability of getting "Yes" P(yes)

Sample size for year 2013

expected value = Expected count for "Yes" in 2013 if Proportion same for each year

Completed Table of Epected Frequencies:

Expected Frequencies →

Errors/Years	2011	2012	2013	2014	Total
Yes	44.1	34.2	40.6	49.1	168
No	298.9	231.8	275.4	332.9	1139
Total	343	266	316	382	1307

Step 4a — check to see if all $e_{ij} >= 5$

Expected Frequencies:

$\boxed{\text{TRUE All } e_{ij} >= 5}$

Errors/Years	2011	2012	2013	2014
yes	44.1	34.2	40.6	49.1
NO	298.9	231.8	275.4	332.9

Step 4b — Because All $e_{ij} >= 5$, we can calculate:

{Chi-Square Test Statistic}

$$\chi^2 = \underbrace{\sum_i \sum_j}_{} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

Summation notation for adding whole Table with i Rows & j columns.

* Example for Row i column j:

$$\frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \frac{(f_{13} - e_{13})^2}{e_{13}} = \frac{(45 - 40.62)^2}{40.62} = 0.4727$$

Full Table of calculation Answers for χ^2:

Errors	years 2011	2012	2013	2014
yes	0.587	2.269	0.4727	1.337
NO	0.087	0.3347	0.0697	0.1971

Adding all these numbers $= \chi^2 = 5.35$

calculated Test statistic

Step $\boxed{4c}$ Remaining Calculations:

Alpha $= \alpha = 0.05$

k = # of populations = 4 years

$df = 4 - 1 = 3$

or if we use $(r-1)*(k-1)$
same answer: $(2-1)*(4-1) = 1*3 = 3$

χ^2_α = critical value = CHISQ.INV.RT(Alpha, df)

= CHISQ.INV.RT(0.05, 3) = 7.815

$\left\{\begin{matrix}\text{P-value} \\ \text{Method 1}\end{matrix}\right\}$ = CHISQ.TEST(actual_range, expected_range)

= CHISQ.TEST(Table of observed Frequencies, Table of Expected Frequencies)

= 0.1476

$\left\{\begin{matrix}\text{P-value} \\ \text{Method 2}\end{matrix}\right\}$ = CHISQ.DIST.RT(x, df)

calculated χ^2 test statistic = 5.35   3

= CHISQ.DIST.RT(5.35, 3) = 0.1476

Step $\boxed{5}$ Rules same as with previous Hypothesis Tests:

P-value $\le \alpha$, then Reject Ho & Accept Ha

critical value $\ge$ Calculated Test statistic, Reject.

Conclusion:

Statistical evidence does not suggest a significant difference between pop. proportions.

$\boxed{\underset{\text{Pvalue}}{0.1476} \ge \underset{\text{Alpha}}{0.05}}$

$\boxed{\underset{\text{calculated } \chi^2}{5.35} \le \underset{\text{critical v. } \chi^2}{7.815}}$

2011	2012	2013	2014
Yes	Yes	No	No
No	No	No	No
No	No	No	No
No	No	No	No
Yes	No	No	No
Yes	No	Yes	No
No	No	No	No
No	No	No	No
Yes	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	Yes
No	No	No	No
No	No	No	No
Yes	No	Yes	No
No	No	No	Yes
No	No	No	Yes
Yes	No	No	No
No	No	No	No
No	No	No	Yes
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
Yes	No	No	No
No	No	No	Yes
No	No	No	No
No	No	No	No
No	No	No	Yes
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
Yes	No	No	Yes

Question: Executive at HMO wanted to check Patient Record Data accuracy. An independent and random sample of medical records was taken for each of the four years.

Step 1:

$H_0 : p_1 = p_2 = p_3 = p_4$

H_a : Not all proportions are equal?

Chi-Square method can only do a two-tail test

Because of the way the calculations are set up, it will always be a test on the upper end (test statistic is always >= 0)

Step 2:

Alpha	0.05

I = row

j = column

Step 3:

Observed Frequencies = f_{ij} :

Errors in Form/Year	2011	2012	2013	2014	Total
Yes	39	43	45	41	**168**
No	304	223	271	341	**1139**
Total	**343**	**266**	**316**	**382**	**1307**

Sample Proportions of Error Rate:

Errors in Form/Year	2011	2012	2013	2014
Yes	0.113702624	0.161654135	0.142405063	0.107329843
No	0.886297376	0.838345865	0.857594937	0.892670157
Total	**1**	**1**	**1**	**1**

Expected Frequencies:

Errors in Form/Year	2011	2012	2013	2014
Yes	44.08875287	34.19127774	40.61820964	49.10175976
No	298.9112471	231.8087223	275.3817904	332.8982402
Total	**343**	**266**	**316**	**382**

Overall Proportions
0.128538638
0.871461362

i = row

j = column

Expected Frequency = e

e_{ij} = (Row i Total)/(Sum of all Sample Sizes)*(Column j Total)

Deviations Squared:

Errors in Form/Year	2011	2012	2013	2014
Yes	25.89540576	77.59358794	19.20008676	65.63851113
No	25.89540576	77.59358794	19.20008676	65.63851113

Deviations Squared/Expected Frequencies: Final Calculations to add to get Chi-Square

Errors in Form/Year	2011	2012	2013	2014
Yes	0.587347205	2.269397141	0.47269653	1.33678531
No	0.086632424	0.334731097	0.069721701	0.197172899

((f_{ij} - eij)^2/e_{ij}

Step 4:

Test Statistic Chi-Square = χ^2 =	5.354484306	5.354484306	5.354484306
k = count populations =	4		
df	3		
p-value =	0.147605801		
p-value =	0.147605801		
Chi Critical Value	7.814727903		

SUM of ALL ((f_{ij} - eij)^2/e_{ij}

Step 5:

Because p-value is bigger than alpha and the test statistic is not past our critical value, we fail to reject H0.
It is reasonable to assume that there is not a difference between the population proportions.
The statistical evidence suggests that there is not a significant difference between the population proportions, that is the error rate for each one of the years is not significantly different.
There is a 5% risk that we have concluded that there is a difference, when in fact there is a difference.

P.18

2011	2012	2013	2014
Yes	Yes	No	No
No	No	No	No
No	No	No	No
No	No	No	No
Yes	No	No	No
Yes	No	Yes	No
No	No	No	No
No	No	No	No
Yes	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	Yes
No	No	No	No
No	No	No	No
Yes	No	Yes	No
No	No	No	Yes
No	No	No	Yes
Yes	No	No	No
No	No	No	No
No	No	No	Yes
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
Yes	No	No	No
No	No	No	Yes
No	No	No	No
No	No	No	No
No	No	No	Yes
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
No	No	No	No
Yes	No	No	Yes

Question: Executive at HMO wanted to check Patient Record Data accuracy. An independent and random sample of medical records was taken for each of the four years.

Step 1:
$H_0 : p_1 = p_2 = p_3 = p_4$
H_a : Not all proportions are equal?

Chi-Square method can only do a two-tail test (that is: H0 uses "Equal" and Alternative uses "Not Equal")
Because of the way the calculations are set up, it will always be a test on the upper end (test statistic is always >= 0)

Step 2: Alpha 0.05

i = row
j = column

Step 3: Observed Frequencies = f_{ij} :

Errors in Form/Year	2011	2012	2013	2014	Total
Yes	=COUNTIFS(A:A,$G11)	=COUNTIFS(B:B,$G11)	=COUNTIFS(C:C,$G11)	=COUNTIFS(D:D,$G11)	**=SUM(H11:K11)**
No	=COUNTIFS(A:A,$G12)	=COUNTIFS(B:B,$G12)	=COUNTIFS(C:C,$G12)	=COUNTIFS(D:D,$G12)	**=SUM(H12:K12)**
Total	**=SUM(H11:H12)**	**=SUM(I11:I12)**	**=SUM(J11:J12)**	**=SUM(K11:K12)**	**=SUM(H13:K13)**

324
18

Sample Proportions of Error Rate:

Errors in Form/Year	2011	2012	2013	2014
Yes	=H11/H$13	=I11/I$13	=J11/J$13	=K11/K$13
No	=H12/H$13	=I12/I$13	=J12/J$13	=K12/K$13
Total	**=SUM(H17:H18)**	**=SUM(I17:I18)**	**=SUM(J17:J18)**	**=SUM(K17:K18)**

Expected Frequencies:

Errors in Form/Year	2011	2012	2013	2014
Yes	=$M23*H$13	=$M23*I$13	=$M23*J$13	=$M23*K$13
No	=$M24*H$13	=$M24*I$13	=$M24*J$13	=$M24*K$13
Total	**=SUM(H23:H24)**	**=SUM(I23:I24)**	**=SUM(J23:J24)**	**=SUM(K23:K24)**

Overall Proportions
=L11/$L$13
=L12/$L$13

i = row
j = column
Expected Frequency = e
e_{ij} = (Row i Total)*(Column j Total)/(Overall Total)

Deviations Squared:

Errors in Form/Year	2011	2012	2013	2014
Yes	=(H11-H23)^2	=(I11-I23)^2	=(J11-J23)^2	=(K11-K23)^2
No	=(H12-H24)^2	=(I12-I24)^2	=(J12-J24)^2	=(K12-K24)^2

summation notation

Deviations Squared/Expected Frequencies:

Errors in Form/Year	2011	2012	2013	2014
Yes	=H29/H23	=I29/I23	=J29/J23	=K29/K23
No	=H30/H24	=I30/I24	=J30/J24	=K30/K24

$((f_{ij} - eij)^2/e_{ij}$

Step 4:

Test Statistic Chi-Square = χ^2 =	=SUM(H34:K35)	=SUMPRODUCT((H11:K12-H23:K24)^2/H23:K24)	=SUMPRODUCT((H11:K12-L11:L12/L13*H13:K13)^2/(L11:L12/L13*H13:K13))
k = count populations =	=COUNT(H10:K10)		
df	=H39-1		
p-value =	=CHISQ.DIST.RT(H38,H40)		
p-value =	=CHISQ.TEST(H11:K12,H23:K24)		
Chi Critical Value	=CHISQ.INV.RT(H7,H40)		

SUM of ALL $((f_{ij} - eij)^2/e_{ij}$

Step 5:

Because p-value is bigger than alpha and the test statistic is not past our critical value, we fail to reject H0.
It is reasonable to assume that there is not a difference between the population proportions.
The statistical evidence suggests that there is not a significant difference between the population proportions, that is the error rate at the different offices is not significantly different.
There is a 5% risk that we have concluded that there is a difference, when in fact there is a difference.

p. 19

(4) Test of Independence of 2 Categorical Variables using χ^2 chi-squre ("Kie" like Pie) (P. 20)

---

* Using Sample Data, we test if 2 categorical Variables sampled from 1 population are Independent (not dependent or associated).

---

* Example:

① Researchers want ~~to~~ determine if there is a difference in "Hiring Plans over the Next year" based on the "[redacted]" Private Firm or Public Firm. (2 categorical variables)

② Human Resource Executives were surveyed and asked about their Hiring Plans over next year. The survey looked like this:

categorical variables

① What are your hiring plans over next year? select one:

Hiring ____

Not Hiring ____

Lay off workers ____

② What sort of firm do you manage? select one:

Private ____

Public ____

one possible survey result might look like this

① What are your hiring plans over next year? select one:

Hiring X

Not Hiring ____

Lay off workers ____

② What sort of firm do you manage? select one:

Private ____

Public X

(3) From the random sample our survey results would look like:

2 categorical variables from 1 population of executives

(p. 21)

Fields →

Records →

Hiring Plans over Next Year	Firm Type
Hiring	Public
Hiring	Private
Not Hiring	Public
Lay off workers	Private
⋮	⋮

(4) If objective of study is to determine if there is a difference in Hiring over Next Year Based on Firm Type when we cross tabulate to get Observed Frequencies we must:

must put "Response variable" (Row variable) "Hiring Plans over Next Year"

must put Explanatory Variable (Column Variable) "Firm Type"

		Firm Type		
		Private	Public	Total
Hiring Plans over Next Year	Hiring	40	32	72
	Not Hiring	16	32	48
	Lay off Employees	16	46	62
	Total	72	110	182

Sample size from 1 population

(5) If we look at observed frequencies & proportions for just the "Response variable": P.22

Hiring Plans over Next Year	observed Frequencies	overall sample proportion
Hiring	72	40%
Not Hiring	48	26%
Lay off Employees	62	34%
Total	182	100%

we can think of checking Independence of 2 categorical variables this way:

"Do the proportions for the "response variable" significantly differ when a second "explanatory variable" is added?"

2nd variable added

Hiring Plans over Next Year		Firm Type		
		Private	Public	Total
	Hiring	56%	29%	40%
	Not Hiring	22%	29%	26%
	Lay off workers	22%	42%	34%
	Total	100%	100%	100%

* is there a significant change?
* if there was zero change, they would all be the same as overall sample proportion.

(6) ways we can ask Independence Question: (P.23)

* Is the Response variable "Hiring Plans over next year" independent of the Explanatory variable "Firm Type"?

* Is the Response variable "Hiring Plans over Next Year" dependent on the Explanatory variable "Firm Type"?

* Is there an association between the Response variable "Hiring Plans over Next Year" and the Explanatory variable "Firm Type"?

(7) our 2 possible conclusions:

$\boxed{1}$ If 2 categorical variables are Independent:

* "Hiring Plans over Next Year" will not depend on "Firm Type" and the proportions for Hiring, Not Hiring and Laying off workers will be same for Public & Private Firms.

$\boxed{2}$ If 2 categorical variables are <u>NOT</u> Independent:

* Proportions for Hiring, Not Hiring & Lay off workers should differ by Firm.
* After our chi-squre test we will have evidence that "Hiring Plans over Next Year" variable is associated with or dependent on "Firm Type"!
* We can gain more insights by comparing Sample proportions and making charts.

(8) Chi-square Test will involve same Formulas as last section. We will compare "observed Frequencies" to Expected Frequencies" in a "cross tabulated table" sometimes called a "contigency Table":

(P. 24)

* Assume H_0 TRUE & compute Expected Frequencies

$$e_{ij} = \frac{\text{Row } i \text{ Total}}{\text{sample size}} * \left(\text{column } j \text{ total}\right)$$

* IF ALL $e_{ij} \geq 5$ then:

$$\left\{\begin{array}{c}\text{Chi-square}\\ \text{test}\\ \text{statistic}\end{array}\right\} = \chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

④ chi-square Test for Independence of 2 categorical variables (contigency Table Test)

P. 25

① Null & Alternative:

H_0: The Response Variable (Row Variable) is independent of the Explanatory Variable (column variable).

H_a: The Response Variable (Row variable) is NOT independent of the Explanatory Variable (column variable).

② select alpha. select Random sample from one population and collect data for both variables for every element in sample. Explanatory variable is the column variable & the Response variable is the row variable for a cross tabulated Frequency table w/ observed frequencies $= f_{ij}$, where i = Row and j = column.

③ Assume Ho TRUE & compute Expected Frequencies:

$$e_{ij} = \frac{\text{Row } i \text{ Total}}{\text{sample size}} * (\text{column } j \text{ Total})$$

④ if all $e_{ij} >= 5$, then compute chi-square Test Statistic:

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

⑤ {Rejection Rules}

Reject Ho & Accept Ha if p-value $\leq \alpha$

Reject Ho & Accept Ha if Test Statistic $\chi^2 \geq$ critical value χ^2

$df = (r-1)*(c-1)$, where r = # Rows, c = # columns

⑥ IF variables are NOT independent, investigate association by calculating Pbars & creating charT.

Firm Type	Hiring Plans [illegible]he Next Year
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Lay Off Workers
Private	Lay Off Workers
Private	Lay Off Workers
Private	Lay Off Workers
Private	Lay Off Workers
Private	Lay Off Workers

Survey asked human resource executives about their firms hiring plans over the next year
The survey offered three nominal "Hiring Over Next Year" categories: Hiring, Not hiring, Laying off employees
The survey also ask whether the firm was public or private. We will call this nominal variable: "Firm Type"
At the 0.05 alpha risk level, is the "Hiring Over Next Year" Nominal Variable independent of the "Firm Type" Nominal Variable?
Categorical Variables, or more specifically, Nominal Variables are:
"Firm Type" will be the "Explanatory Variable" = "Column Variable"
"Hiring Over Next Year" is the "Response Variable" = "Row Variable"
Different ways to ask Question about Independence:
Is the "Hiring Over Next Year" variable Independent of the "Firm Type" variable?
Does the "Firm Type" variable influence/affect/bias the "Hiring Over Next Year" variable?
Is the "Hiring Over Next Year" variable dependent on the "Firm Type" variable?
Conduct a "Test of Independence" or "Contingency Table Test" to determine if the "Hiring Over Next Year" variable is independent of the "Firm Type" variable.
Create "Cross Tab" or "Contigency Tables" for the "Observed Frequencies" and the "Expected Frequencies"

Variables:
Firm Type
Hiring Plans Over The Next Year
Key: 2 variables from 1 population

Step 1: H0 : the "Hiring Over Next Year" variable is Independent of the "Firm Type" variable.
Ha : the "Hiring Over Next Year" variable is NOT Independent of the "Firm Type" variable.

Step 2: Alpha = 0.05

Step 3: Observed Frequencies:

Hiring Over Next Year/Firm Type	Private	Public	Total
Hiring	40	32	72
Not Hiring	16	32	48
Lay Off Workers	16	46	62
Total	72	110	182

Pbar Proportions:	pbar1	pbar2	Overall pbar
Hiring Over Next Year/Firm Type	Private	Public	Total
Hiring	56%	29%	40%
Not Hiring	22%	29%	26%
Lay Off Workers	22%	42%	34%
Total	100%	100%	100%

Expected Frequencies:

Hiring Over Next Year/Firm Type	Private	Public	Total
Hiring	28.48352	43.51648352	72
Not Hiring	18.98901	29.01098901	48
Lay Off Workers	24.52747	37.47252747	62
Total	72	110	182

Step 4: All eij >= 5? TRUE <<== Must use Ctrl + Shift + Enter to put "array formula" into cell.

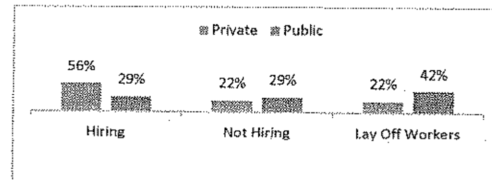
Calculations for Chi-Square Test Statistic:

Hiring Over Next Year/Firm Type	Private	Public
Hiring	4.656356	3.047796648
Not Hiring	0.470492	0.307958708
Lay Off Workers	2.964749	

Chi Test Statistic = χ^2 =	11.44735	13.38791499
k = # pops. = # Columns = c =	2	
# Rows = r =	3	
df	2	df = (r-1)*(c-1)
p-value	0.001238	0.003267676
Chi-Square Critical Value =	5.991465	

Step 5: Because the p-value is much smaller than our alpha, and because the calculated test statistic is much bigger than our critical value, we reject H0 and accept Ha.
The sample evidence suggests that the "Hiring Over Next Year" variable is NOT independent of the "Firm Type" variable.
It is reasonable to assume that the proportions for Hiring Over Next Year Over are different for Private and Public Firms.
It is reasonable to assume that the "Hiring Over Next Year" variable is dependent on the "Firm Type" variable.
It is reasonable to assume that the "Firm Type" variable influences/affects the "Hiring Over Next Year" variable.
We do run a 5% risk of reject H0 when in fact it was TRUE.

Step 6:
It looks like Private and Public plans are going in opposite directions.
Add Employees is: 56% for Private Firms, but 29% for Public Firms.
Lay Off Employees is: 22% for Private Firms, but 42% for Public Firms



P.26

Firm Type	Hiring Plans Over Next Year
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Not hiring
Private	Lay Off Workers
Private	Lay Off Workers
Private	Lay Off Workers
Private	Lay Off Workers
Private	Lay Off Workers
Private	Lay Off Workers

Survey asked human resource executives about their firms hiring plans over the next year
The survey offered three nominal "Hiring Over Next Year" categories: Hiring, Not hiring, Laying off employees
The survey also ask whether the firm was public or private. We will call this nominal variable: "Firm Type"
At the 0.05 alpha risk level, is the "Hiring Over Next Year" Nominal Variable independent of the "Firm Type" Nominal Variable?
Categorical Variables, or more specifically, Nominal Variables are:
"Firm Type" will be the "Explanatory Variable" = "Column Variable"
"Hiring Over Next Year" is the "Response Variable" = "Row Variable"
Different ways to ask Question about Independence:
Is the "Hiring Over Next Year" variable Independent of the "Firm Type" variable?
Does the "Firm Type" variable influence/affect/bias the "Hiring Over Next Year" variable?
Is the "Hiring Over Next Year" variable dependent on the "Firm Type" variable?
Conduct a "Test of Independence" or "Contingency Table Test" to determine if the "Hiring Over Next Year" variable is independent of the "Firm Type" variable.
Create "Cross Tab" or "Contigency Tables" for the "Observed Frequencies" and the "Expected Frequencies"

Variables:
Firm Type
Hiring Plans Over The Next Year
Key: 2 variables from 1 population

Step 1: H0 : the "Hiring Over Next Year" variable is Independent of the "Firm Type" variable.
Ha : the "Hiring Over Next Year" variable is NOT Independent of the "Firm Type" variable.

Step 2: Alpha = 0.05

Step 3: Observed Frequencies:

Hiring Over Next Year/Firm Type	Private	Public	Total
Hiring	=COUNTIFS($A$2:$A$183,F$22,$B$2:$B$183,$E23)	=COUNTIFS($A$2:$A$183,G$22,$B$2:$B$183,$E23)	=SUM(F23:G23)
Not Hiring	=COUNTIFS($A$2:$A$183,F$22,$B$2:$B$183,$E24)	=COUNTIFS($A$2:$A$183,G$22,$B$2:$B$183,$E24)	=SUM(F24:G24)
Lay Off Workers	=COUNTIFS($A$2:$A$183,F$22,$B$2:$B$183,$E25)	=COUNTIFS($A$2:$A$183,G$22,$B$2:$B$183,$E25)	=SUM(F25:G25)
Total	=SUM(F23:F25)	=SUM(G23:G25)	=SUM(F26:G26)

Pbar Proportions:	pbar1	pbar2	Overall pbar
Hiring Over Next Year/Firm Type	Private	Public	Total
Hiring	=F23/F$26	=G23/G$26	=H23/H$26
Not Hiring	=F24/F$26	=G24/G$26	=H24/H$26
Lay Off Workers	=F25/F$26	=G25/G$26	=H25/H$26
Total	=F26/F$26	=G26/G$26	=H26/H$26

Expected Frequencies:

Hiring Over Next Year/Firm Type	Private	Public	Total
Hiring	=$H23/$H$26*F$26	=$H23/$H$26*G$26	=SUM(F38:G38)
Not Hiring	=$H24/$H$26*F$26	=$H24/$H$26*G$26	=SUM(F39:G39)
Lay Off Workers	=$H25/$H$26*F$26	=$H25/$H$26*G$26	=SUM(F40:G40)
Total	=SUM(F38:F40)	=SUM(G38:G40)	=SUM(F41:G41)

Step 4: All eij >= 5? {=AND(F38:G40>=5)} <<== Must use Ctrl + Shift + Enter to put "array formula" into cell.

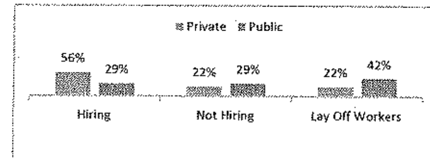
Calculations for Chi-Square Test Statistic:

Hiring Over Next Year/Firm Type	Private	Public
Hiring	=(F23-F38)^2/F38	=(G23-G38)^2/G38
Not Hiring	=(F24-F39)^2/F39	=(G24-G39)^2/G39
Lay Off Workers	=(F25-F40)^2/F40	#N/A

Chi Test Statistic = χ^2 =	=SUM(F47:G49)	=SUMPRODUCT((F23:G25-H23:H25/H26*F26:G26)^2/(H23:H25/H26*F26:G26))
k = # pops. = # Columns = c =	=COLUMNS(F22:G22)	
# Rows = r =	=ROWS(E23:E25)	
df	=(F52-1)*(F53-1)	df = (r-1)*(c-1)
p-value	=CHISQ.TEST(F23:G25,F38:G40)	=CHISQ.DIST.RT(F51,F54)
Chi-Square Critical Value =	=CHISQ.INV.RT(F19,F54)	

Step 5: Because the p-value is much smaller than our alpha, and because the calculated test statistic is much bigger than our critical value, we reject H0 and accept Ha.
The sample evidence suggests that the "Hiring Over Next Year" variable is NOT independent of the "Firm Type" variable.
It is reasonable to assume that the proportions for Hiring Over Next Year Over are different for Private and Public Firms.
It is reasonable to assume that the "Hiring Over Next Year" variable is dependent on the "Firm Type" variable.
It is reasonable to assume that the "Firm Type" variable influences/affects the "Hiring Over Next Year" variable.
We do run a 5% risk of reject H0 when in fact it was TRUE.

Step 6: It looks like Private and Public plans are going in opposite directions.
Add Employees is: 56% for Private Firms, but 29% for Public Firms.
Lay Off Employees is: 22% for Private Firms, but 42% for Public Firms



P. 27

5) chi-square Distribution & chi-square test statistic can be used in 3 situations: (P.28)

1) Hypothesis Test for 2 or more Population Proportions

used to test whether 2 or more pop. proportions are equal.

Same

Expected Frequencies under Assumption H_0 is TRUE $= e_{ij} = \frac{\text{Row } i \text{ Total}}{\text{Sum of all sample sizes}} * (\text{Column } j \text{ Total})$

chi-square Test Statistic if All $e_{ij} \geq 5$ $= \chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$

α = alpha
k = # populations or c = # columns
$df = (r-1) * (k-1)$ → Also: $(r-1)*(c-1)$

3 or more row categorical variables (multinomial pop.)

only 2 Rows: $df = k-1$

Critical value for each combination of 2 proportions - done if test says there is a difference $CV_{ij} = \sqrt{\chi^2_\alpha} * \sqrt{\frac{\bar{p}_i(1-\bar{p}_i)}{n_i} + \frac{\bar{p}_j(1-\bar{p}_j)}{n_j}}$

Same

2) Test of Independence "Contigency Table Test"

used to test if 2 categorical variables sampled from 1 population are independent. Does a set of proportions from the response variable (Row variable) significantly differ when a second explanatory variable (column variable) is added? $df = (r-1)*(c-1)$ $e_{ij} = \frac{\text{Row } i \text{ Total}}{\text{Sample size}} * (\text{Column } j \text{ total})$

Not covered in this class

3) Goodness of Fit Test (for Nominal Data)

used to determine whether a random nominal variable has a specific Probability Distribution. Check to see if observed frequencies for a multinomial probability distribution are equal to the expected frequencies.

chi-square test statistic if ALL $e_i \geq 5$ $= \chi^2 = \sum_{i=1}^{K} \frac{(f_i - e_i)^2}{e_i}$

f_i = observed frequency for category i
e_i = expected frequency for category i
K = # categories
$df = K-1$

Independence Defined:

① Probability Theory (Chapter 4):

If occurance of 1 event does not affect the probability of the other event.

② Probability Distributions (chapter 5 on...)

2 Random Variables are Independant if the occurance of one does not affect the probability Distribution of the other