Note: For a CI for p, it is possible for a calculated bound to be above 1, or less than zero. This typically happens when $\tilde{p}$ is close to zero or one.
However, we would want to report the interval as being cut off at 0, or 1, rather than listing an impossible bound.

✳ Goodness of Fit

If a categorical Y can take on more than two categories, or if we are interested in both categories at once, we may want to know if "The true proportions are roughly equal to my hypothesized proportions"?

We have a hypothesis test for this situation.

Step 1) Suppose we have k categories, and $\sum_{i=1}^{k} p_i = 1$.
H_0: $P_r\{\text{Cat } 1\} = p_1$, $P_r\{\text{Cat } 2\} = p_2$ ⋯ $P_r\{\text{Cat } k\} = p_k$
H_A: At least two of the categories differs from H_0
Here, $p_i = P_r\{\text{subject is in cat } i\} = \text{true hypothesized proportion}$.

Step 2) Let O_i = observed # of subjects in cat i
Let e_i = expected # of subjects in cat i if H_0 were true $= np_i$

Our test-statistic is

$$\chi^2_s = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \quad \text{with d.f.} = k-1$$

This is a chi-squared distribution, which is roughly a normal distribution square.
Its shape is similar to an F, but it has only one d.f.

Note: We do not round e_i to a whole #

Step 3) The p-value is (two-sided by default): $Pr\{\chi^2 > \chi^2_s\}$. There is a Chi-Squared table, which we use in a very similar way to a t-table.

Step 4) If p-value $< \alpha$, reject H_0
If p-value $\geq \alpha$, fail to reject H_0

Ex: A hospital manager want to test if the proportion of ER patients during four phases of the moon are equal. A random sample of 316 showed:

Cycle	New	1st Quarter	Full	3rd Quarter	Total
Count	85	66	97	68	316

a) State H_0, H_A

Since the claim is that the proportions are equal,
H_0: $Pr\{\text{New}\} = Pr\{\text{1st}\} = Pr\{\text{Full}\} = Pr\{\text{3rd}\} = 1/4$
H_A: At least two proportions are not $1/4$.

b) Calculate the test-statistic

					Total
O_i	85	66	97	68	316
e_i	$316(1/4) = 79$	$316(1/4) = 79$	$316(1/4) = 79$	$316(1/4) = 79$	316

$$\chi^2_s = \sum_{i=1}^{4} \frac{(O_i - e_i)^2}{e_i} = \frac{(85-79)^2}{79} + \frac{(66-79)^2}{79} + \frac{(97-79)^2}{79} + \frac{(68-79)^2}{79}$$
$$= 0.4557 + 2.1392 + 4.1013 + 1.5316 = 8.2278$$

c) Calculate the p-value

At d.f. $= k - 1 = 4 - 1 = 3$, using the χ^2 table at row 3
7.81 is in column 0.05 ($Pr\{\chi^2 > 7.813\} = 0.05$), and
9.84 is in column 0.02 ($Pr\{\chi^2 > 9.84\} = 0.02$).
Thus, $0.02 <$ p-value < 0.05

d) State your decision and interpret your conclusion in terms of the problem if $\alpha = 0.05$

Decision: Since p-value $< 0.05 = \alpha$, reject H_0

Interpretation: We cannot conclude that the proportion of patients at the ER is equal for four phases of the moon.

e) Which category differed the most from what we expected if H_0 was true?

The full moon category, since its contribution to χ^2s is the largest.

Ex: A researcher believes that a certain plant has purple blooms 55% of the time, red blooms 25% of the time, and pink 20% of the time. A random sample of 143 showed

	Purple	Red	Pink	Total
Count	63	37	43	143

a) Test the claim using $\alpha = 0.05$, showing all four steps.

Step 1) H_0: $Pr\{\text{Purple}\} = 0.55$, $Pr\{\text{Red}\} = .25$, $Pr\{\text{Pink}\} = .2$

H_1: At least two proposed proportions do not match

Step 2)

	Purple	Red	Pink	Total
O_i	63	37	43	143
e_i	$(143)(.55) = 78.63$	$(143)(.25) = 35.75$	$(143)(.20) = 28.6$	143

$$\chi^2_s = \frac{(63-78.63)^2}{78.63} + \frac{(37-35.75)^2}{35.75} + \frac{(43-28.6)^2}{28.6}$$

$$= 3.114 + 0.0437 + 7.250 = 10.408$$

Step 3) d.f. $= k - 1 = 3 - 1 = 2$

P-value $= Pr\{\chi^2 > 10.408\}$.

At row 2, 9.21 is in column 0.01, 13.82 is in column 0.001.

Thus, $0.001 <$ p-value < 0.01

Step 4) Since p-value $< \alpha$, reject H_0 and conclude that we do not have enough evidence to support the researchers claim about the percentage of flower blooms.

b) Which category was the most different from the researchers claim?

The Pink category, since it contributed the most to χ^2's (with 7.250).

* Assumptions for Goodness of Fit

1) A random sample was taken
2) $e_i \geq 5$ for all groups (we expect at least 5 for all groups)

Two Categorical Variables

When we have two categorical variables, the question becomes "Does one variable effect the other?" In other words, "Are the variables independent?"

Recall that if two events are independent,

i) $Pr\{A \mid B\} = Pr\{A\} = Pr\{A \mid B^c\}$

ii) $Pr\{A \text{ and } B\} = Pr\{A\} Pr\{B\}$

The main idea is that we create a hypothetical sample that is perfectly independent, and compare that to the sample we have.

Ex: We want to see if the flu shot was effective for a particular year, and select 162 students randomly:

		F: Got Flu	F^c: Did Not	
S =	Got Shot	23	55	78
S^c =	Did Not	40	44	84
		63	99	162

The overall probability of getting the flu is $\hat{Pr}\{F\} = \frac{63}{162} = 0.388$

If the 78 people who got the shot had the same prob. of getting the flu as the 84 who did not (i.e. getting the flu and getting the flu shot is indep) we would expect

$78(.388) = 78\left(\frac{63}{162}\right) = 30.333$ people in the S group to get the flu,

and

$84(.388) = 84\left(\frac{63}{162}\right) = 32.666$ people from the S^c group to get the flu.

In other words, if $Pr\{F \mid S\} = Pr\{F\} = Pr\{F \mid S^c\}$, out of the 78 who got the shot, 30.333 should get the flu, and out of the 84 who did not get the shot, 32.66 should get the flu.

We can generilize the above concepts into formulas for
e_{ij} = expected count if the categories were indep.

* χ^2 Test for Independence

Step 1) H_0: The two categorical variables are independent.
H_A: The two categorical variables are dependent.
or $H_0: Pr\{A|B\} = Pr\{A|B^c\}$ vs $H_A: Pr\{A|B\} \neq Pr\{A|B^c\}$

Step 2) O_{ij} = observed value for i^{th} row, j^{th} column
Let r_i = row total for row i
c_j = column total for column j
e_{ij} = expected count for row i, column j if H_0 true
$= (r_i)(c_j)/n = r_i\left(\frac{c_j}{n}\right)$
I = number of categories for A, J = # of categories for B

$$\chi^2_s = \sum_{\text{all } ij} \frac{(O_{ij} - e_{ij})^2}{e_{ij}} \qquad d.f. = (I-1)(J-1)$$

Step 3) P-value = $Pr\{\chi^2 > \chi^2_s\}$ (for two sided)

Step 4) If p-value $< \alpha$, reject H_0
If p-value $\geq \alpha$, fail to reject H_0.

Ex: Flu shot example continued. Let $\alpha = 0.01$
Step 1) $H_0: Pr\{F|S\} = Pr\{F|S^c\}$
$H_A: Pr\{F|S\} \neq Pr\{F|S^c\}$
or H_0: Getting the flu is indep. of getting flu shot.
H_A: Getting the flu is depen. of getting flu shot.

Step 2)

O_{ij}	F	F^c		e_{ij}		
S	23	55	$78 = r_1$	$\frac{(78)(63)}{162} = 30.333$	$\frac{(78)(99)}{162} = 47.666$ or $78 - 30.333$	78
S^c	40	44	$84 = r_2$			
	$63 = c_1$	$99 = c_2$	$162 = n$	$\frac{(84)(63)}{162} = 32.666$ or $63 - 30.333$	$\frac{44(99)}{162} = 51.333$ or $84 - 32.666$	84
				63	99	162

Thus, we have a hypothetical sample of 162 that is perfectly indep, and we can compare it to our actual sample.

$$\chi^2_{obs} = \frac{(23-30.333)^2}{30.333} + \frac{(55-47.666)^2}{47.666} + \frac{(40-32.666)^2}{32.666} + \frac{(44-51.333)^2}{51.333}$$

$$= 1.773 + 1.128 + 1.647 + 1.048 = 5.5953$$

$$d.f = (2-1)(2-1) = 1$$

Step 3) p-val = $Pr\{\chi^2 > 5.5953\}$ at d.f = 1

At row 1, 5.41 is in column 0.02, 6.63 is in column 0.01

$0.01 < p\text{-value} < 0.02$

Step 4) Since p-value $> \alpha$, fail to reject H_0. We cannot conclude the prob of getting the flu differs for those who got the flu shot vs. those who did not.

Interpretation of p-value: If in reality getting the flu was indep. of getting a flu shot, we would observe our data or more extreme (more dependent) with prob between 0.01 and 0.02.

* Assumptions for χ^2 test for indep

1) A random sample was taken
2) $e_{ij} \geq 5$ for all rows, all columns.