12 Chi-squared (χ^2) Tests for Goodness-of-fit and Independence

The chi-squared tests are for H_0: "The frequency distribution of ____________ events observed in a sample is ____________ with a particular distribution" against H_A: "Not H_0". We consider two of its forms: the test for goodness-of-fit of counts for one categorical variable to a distribution and the test for independence of two categorical variables.

Each uses a *chi-square* statistic of the form

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

This is a measure of ____.

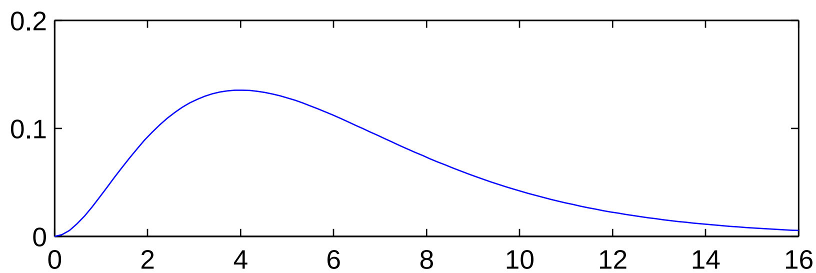
If expected counts are all at least ______, and under a suitable H_0, then X^2 fits a χ^2 distribution.

The Chi-Square Distributions

(Background: if $Z_1, \cdots, Z_\nu$ are independent, $N(0,1)$ random variables, then $X^2 = \sum_{i=1}^{\nu} Z_i^2 \sim \chi^2_\nu$.)

A χ^2 distribution is specified by its degrees of freedom, ν. Here are some of its properties:

- $X^2 \geq 0$ (it's a measure of distance)
- $X^2 = 0 \implies$ observed and expected counts are __________________
- Large $X^2 \implies$ observed counts aren't ______________________________
- Each χ^2_ν density function is skewed ____________
- e.g. Here's χ^2_6:



- The χ^2 table gives, in row ___ and column ___, the point $\chi^2_{\nu,\alpha}$ with area α to its right.

  e.g. $\chi^2_{6,.05} =$ ______ (draw)

The Chi-Square Test For Goodness-of-Fit

Recall the z-test for a population proportion, $H_0 : \pi = \pi_0$ vs. $H_A : \pi \neq \pi_0$, for which an outcome takes one of ______ values, success or failure. The *chi-square test for goodness-of-fit* generalizes to the case of an outcome taking any of _____________ values of a categorical variable, testing H_0: "These categorical data came from the specified distribution" vs. H_A: ___________________.

e.g. The Nice family gives trick-or-treaters a scoop of _____________ M&Ms. The Naughty family gives _____________ M&Ms. Anna, Teresa, Margaret, Monica, Andrew, Mary, and Philip return from trick-or-treating, and their father says, "Where did you get the M&Ms?" They know they visited only one of the Nice and Naughty homes, but can't remember which one. Their father says, "Throw away the M&Ms." The children ________. Their mother (a _____________) says, "Let's figure out their source." She investigates and finds these color distributions:

	Brown	Yellow	Green	Red	Total
Nice supply	20%	25%	40%	15%	100%
Naughty supply	50%	20%	10%	20%	100%
Anna, ..., & Philip (sample)	12	15	17	6	$n =$ _____

From which family did the kids get their M&Ms?

Test H_0: "The kids got M&Ms from the Nice family" vs. H_A: "They did not".

Expected Counts

Let $k =$ #category values = ______. If n is the sample size and π_i is the expected proportion in category i under H_0, the *expected count* of each type is $E_i =$ ______. The test statistic is $X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$, whose value for the M&Ms is $\chi^2 =$

Under H_0, $X^2 \sim \chi^2_\nu$, where $\nu = k - 1 =$ ______. The P-value is $P(X_3^2 >$ ______$) =$ __________.

Conclusion:

Next, test H_0: "The kids got M&Ms from the Naughty family" vs. H_A: "They did not". Here $\chi^2 =$

The P-value is $P(X_3^2 >$ ______$) =$ _________________________.

Conclusion:

The Chi-Square Test for Independence

The *chi-square test for independence* tests H_0: "Categorical variables A and B are independent" against H_A: "There is ____________________ between A and B".

e.g. Here is a *contingency table* of ____________________ that relates the education level and smoking status of a SRS of 459 French men. Are education and smoking related?

	Smoking status				
Education	Nonsmoker	Former	Moderate	Heavy	Total
Primary	56	54	41	36	______
Secondary	37	43	27	32	139
University	53	28	36	16	133
Total	______	125	104	84	______

Test H_0: "Education and smoking ______________________" vs. H_A: "There's ______________________ between education and smoking".

Expected Counts

Under H_0, P(Primary and Nonsmoker) = ________________________________, so the expected count in the Primary / Nonsmoker cell is

More generally, let

- O_{ij} = _____________ count in row i and column j
- $O_{i.}$ = ______ i total, $O_{.j}$ = _____________ j total
- $O_{..}$ = _____________ total
- $I = \#$______, $J = \#$_____________

Then, under H_0, the *expected cell count* in row i and column j is $E_{ij} = \frac{O_{i.}O_{.j}}{O_{..}} = \frac{\text{(row total)(column total)}}{\text{table total}}$. Here are the 12 expected counts:

	Smoking status				
Education	Nonsmoker	Former	Moderate	Heavy	Total
Primary	______	50.9	42.4	34.2	187
Secondary	44.2	37.9	31.5	25.4	139
University	42.3	36.2	30.1	______	133
Total	146	125	104	84	459

The chi-square statistic is $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$. For the smokers, its value χ^2 has 12 terms:

Education	Smoking status Nonsmoker	Former	Moderate	Heavy
Primary	______	.19	.04	.09
Secondary	1.2	.7	.6	1.7
University	2.7	1.9	1.1	______

The table sum is $\chi^2 = 13.3$. The required degrees of freedom is $\nu =$ (#rows - 1)(#columns - 1) = ________________________, and the P-value is $P(X_6^2 > 13.3) =$ ________________________.

Conclusion:

R for χ^2 tests

```
rm(list=ls()) # Remove all variables to start with a clean slate.

# Test goodness-of-fit of kids' sample of M&Ms to Nice distribution.
kids.sample = c(12,15,17,6)
Nice.population = c(.20, .25, .40, .15)
chisq.test(x=kids.sample, p=Nice.population)

# Make comparative bar plots.
colors = c("Brown", "Yellow", "Green", "Red")
layout(matrix(data=1:2, nrow=2, ncol=1)) # Allow two graphs in one plot.
barplot(height=kids.sample, names.arg=colors, main="M&M's sample")
barplot(height=Nice.population, names.arg=colors, main="Nice population")
layout(1) # Return to one graph per plot.

# Do it again for the Naughty population.
Naughty.population = c(.50, .20, .10, .20)

layout(matrix(data=1:2, nrow=2, ncol=1)) # Allow two graphs in one plot.
barplot(height=kids.sample, names.arg=colors, main="M&M's sample")
barplot(height=Naughty.population, names.arg=colors, main="Naughty population")
layout(1) # Return to one graph per plot.

chisq.test(x=kids.sample, p=Naughty.population)

# Test independence of education and smoking.

# matrix(data, nrow, ncol, byrow=FALSE) fills an nrow by ncol matrix,
# by column, from the vector data.
(French.men = matrix(data = c(56,37,53, 54,43,28, 41,27,36, 36,32,16),
   nrow=3, ncol=4, byrow=FALSE))
chisq.test(French.men)
```