Notation (not a big deal): X = capital "chi" (also capital "ex"), χ = lower "chi"

- A ____________________________ gets a capital letter: Z, X, X_i, $\bar{X}$, T, X^2
- A ______________ gets lower-case: $z = 1.96$, $z_{\alpha/2}$, x, x_i, $\bar{x}$, $t = -8.87$, $t_{n-1,\alpha/2}$, $\chi^2 = 12.6$
- Distributions follow the culture. e.g. $N(\mu, \sigma^2)$, t_{n-1}, χ^2_ν

6.5 The Chi-Square Test

The chi-square test is for H_0: "The frequency distribution of ______________ events observed in a sample is ______________ with a particular distribution" against H_1: "Not H_0". We consider three of its forms: the tests for goodness-of-fit, independence, and homogeneity.

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 ____.

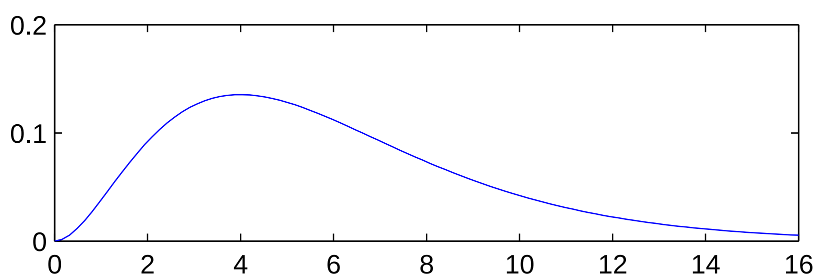
If expected counts are all at least _______, and under a suitable H_0, then X^2 fits a χ^2 distribution.

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:



- Table A.5 (p. 525) 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 (§6.3), $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$, 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_1: ____________________.

e.g. The Nice family gives trick-or-treaters a scoop of _____________ M&Ms. The Naughty family gives _____________ M&Ms. 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%
Margaret, Monica, Andrew, Mary, & 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_1: "They did not".

Expected Counts

Let $k =$ #category values = ______. If n is the sample size and p_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_1: "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_1: "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	______

The *marginal* distribution of education level (circle) is found by summing over ____________________, and the marginal distribution of smoking status (box) is found by summing over ____________________.

Test H_0: "Education and smoking ______________________" vs. H_1: "There's ______________________ between education and smoking".

Expected Counts

Under H_0, $P(\text{Primary} \cap \text{Nonsmoker}) =$ _________________________________, 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 $E_{ij} = \dfrac{O_{i.}O_{.j}}{O_{..}} = \dfrac{\text{(row total)(column total)}}{\text{table total}}$.
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:

	Smoking status			
Education	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:

The Chi-Square Test For Independence Doesn't Say What The Relationship Is

We rejected H_0, concluding that there *is* a relationship between education and smoking. What is the relationship? The chi-square test ________________________.

We can get some insight by comparing conditional distributions of the dependent variable for the separate values of the independent variable. Here are these distributions (as percentages), along with the marginal distribution of smoking status for comparison. What relationship can we see?

	Smoking status			
Education	Nonsmoker	Former	Moderate	Heavy
Primary	29.9%	28.9%	21.9%	19.3%
Secondary	26.6%	30.9%	19.4%	23.0%
University	39.8%	21.1%	27.1%	12.0%
Overall	31.8%	27.2%	22.7%	18.3%

The Chi-Square Test For Homogeneity

The *chi-square test for homogeneity* checks whether different populations have a common distribution with respect to a categorical variable. It uses ____________ expected counts, test statistic, and degrees of freedom as the test for independence.

e.g. (p. 246 #5) Here are data from a study linking exposure to beryllium to disease. Test H_0: "the proportions in disease categories are the same across exposure levels" vs. H_1: "they differ".

	Exposure (Years)		
	< 1	1 to < 5	≥ 5
Diseased	10	8	23
Sensitized	9	19	11
Normal	70	136	206

R gives $\chi^2 =$ ____________ (see the R guide for §6.5). Since $\nu =$ ____________, the P-value is $P(X_4^2 > 10.83) =$ __________________, and we conclude ...