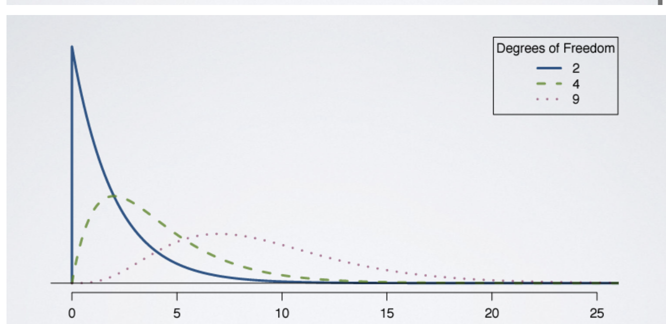
<table>
<tr><th colspan="2">Chi-Square test (Goodness of Fit and Test of Independence)</th></tr>
<tr><td>Chi Squared Test</td><td>Chi Square test is used when determining relationship between two categorical variables

Goodness of Fit test:
- Assess how well observed categorical data fits an expected or theoretical distribution
Test of Independence:
- Examine the association between two categorical variables to determine if they are independent or related</td></tr>
<tr><td>Chi Squared GOF Notation:</td><td>Chi-squared statistic
- Squared so that highly unusual differences between observed and expected will appear even more unusual
- χ^2 statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$ O : observed E : expected k : number of cells
Degrees of Freedom
- To determine if the chi-squared is considered unusually high or not, we describe its distribution
- Higher degree of freedom, closer to normal distribution
- ‣ degrees of freedom (df): influences the shape, center, and spread
χ^2 degrees of freedom for a goodness of fit test: $df = k - 1$ k : number of cells
- 

Condition: Cell based
Conditions for the chi-square test:
1. Independence: Sampled observations must be independent.
‣ random sample/assignment
‣ if sampling without replacement, n < 10% of population
‣ each case only contributes to one cell in the table
2. Sample size: Each particular scenario (i.e. cell) must have at least 5 expected cases.</td></tr>
<tr><td>Chi-Squared GOF Test Example:</td><td>Step 1: Identify the Hypothesis</td></tr>
</table>

H_0: The observed counts of jurors from various race/ethnicities follow the same ethnicity distribution in the population.

H_A: The observed counts of jurors from various ethnicities do not follow the same race/ethnicity distribution in the population.

Step 2: Calculate the expected count and compare with the actual (observed) distribution

ethnicity	white	black	nat. amer.	asian & PI	other	total
%in population	80.29%	12.06%	0.79%	2.92%	3.94%	100%
expected #	2007	302	20	73	98	2500
observed #	1920	347	19	84	130	2500

observed < expected

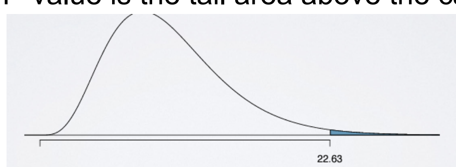
observed > expected

Step 3: FInd Chi-Squared and Df

$$\chi^2 = \frac{(1920-2007)^2}{2007} + \frac{(347-302)^2}{302} + \frac{(19-20)^2}{20} + \frac{(84-73)^2}{73} + \frac{(130-98)^2}{98} = 22.63$$

$$df = k - 1 = 5 - 1 = 4$$

Step 4: Find the p-value

- P-value is the tail area above the calculated test statistic
- 

- Can either use R or Chi-squared table

p-value

$\chi^2 = 22.63 \quad df = 4$

using R

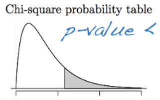
```
R
> pchisq(22.63, 4, lower.tail = FALSE)
[1] 0.0002
```

using the applet

http://bitly.com/dist_calc

using the table

Chi-square probability table



$p\text{-value} < 0.001$

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

-

Chi-Square Independence Test
(Two categorical variables with at least one variable with more than 2 levels)

	dating	cohabiting	married	total
obese	81	103	147	331
not obese	359	326	277	962
total	440	429	424	1293

Does there appear to be a relationship between weight and relationship status?

We need to be a bit more targeted in the ratio being calculated for

Step 1: Identify Hypothesis

H_0 (nothing going on): Weight and relationship status are independent. Obesity rates do not vary by relationship status.

H_A (something going on): Weight and relationship status are dependent. Obesity rates do vary by relationship status.

Step 2: Check the Conditions
1. Independence
2. Sample size

Step 3: Calculate the expected count

If in fact weight and relationship status are independent (i.e. if in fact H_0 is true) how many of the dating people would we expect to be obese? How many of the cohabiting and married?

- Look at the overall obesity rate in the sample and apply that for each relationship status
- $331 / 1293 = 0.256$

  dating: $440 \times 0.256 \approx 113$

  cohabiting: $429 \times 0.256 \approx 110$

  married: $424 \times 0.256 \approx 108$
-

Step 4: Find Chi-square and df
- Note that df is multiplication of the #row-1 and #column-1

Test the hypothesis that relationship status and obesity are associated at the 5% significance level.

	dating	cohabiting	married	total
obese	81 **(113)**	103 **(110)**	147 **(108)**	331
not obese	359 **(327)**	326 **(319)**	277 **(316)**	962
total	440	429	424	1293

$$\chi^2 = \frac{(81-113)^2}{113} + \frac{(103-110)^2}{110} + \frac{(147-108)^2}{108} + \frac{(359-327)^2}{327} + \frac{(326-319)^2}{319} + \frac{(277-316)^2}{316}$$

$$= 31.68$$

$$df = (2-1) \times (3-1) = 1 \times 2 = 2$$

Step 5: FInd p-value

- We cannot conclude that living with someone is making some people obese and that marry someone is making people even more obese

R

```
> pchisq(31.68, 2, lower.tail = FALSE)
[1] 1.320613e-07
```