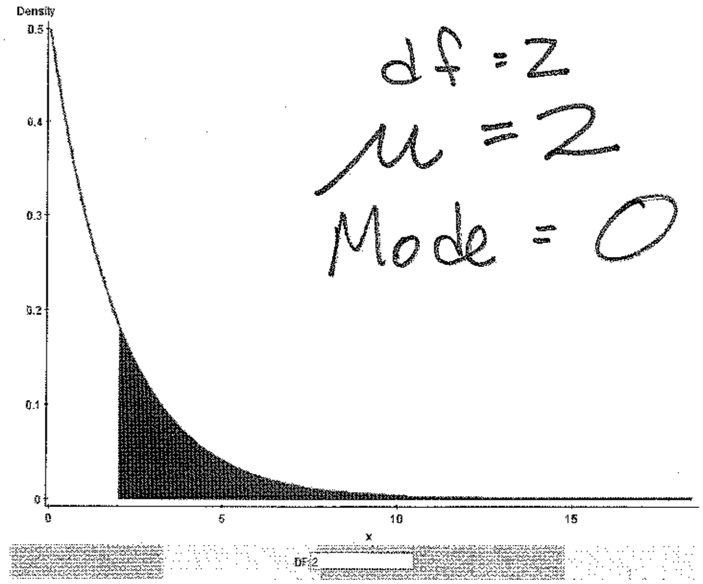
Chi-Square Techniques

- The chi-square model can be used to test many hypotheses and calculate confidence intervals that we have not encountered as of yet.

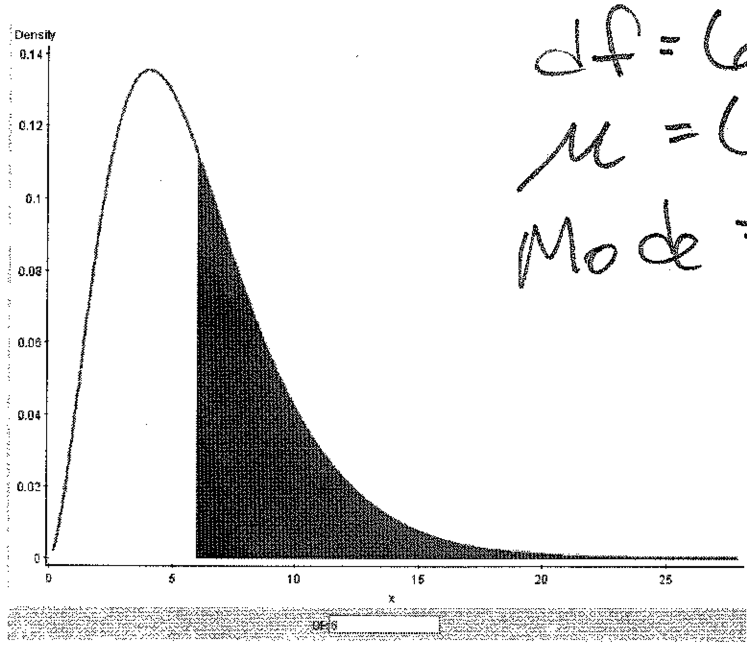
- The chi-square is a family of models. It has one parameter, degrees of freedom which is based on Sample size or number of categories or number of cells in a table.

- If degrees of freedom are small, the distribution is very skewed - right.

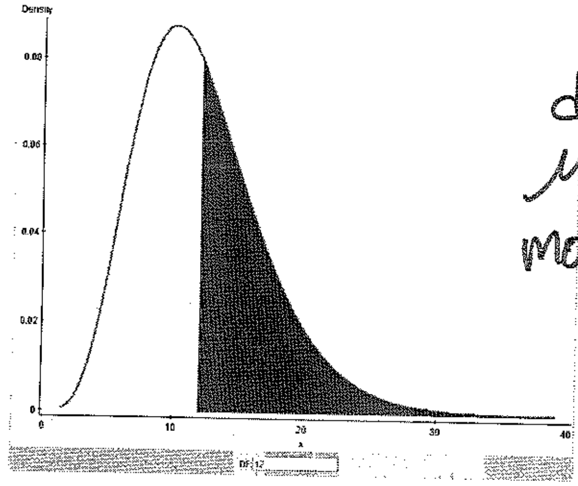


$df = 2$
$\mu = 2$
Mode = 0

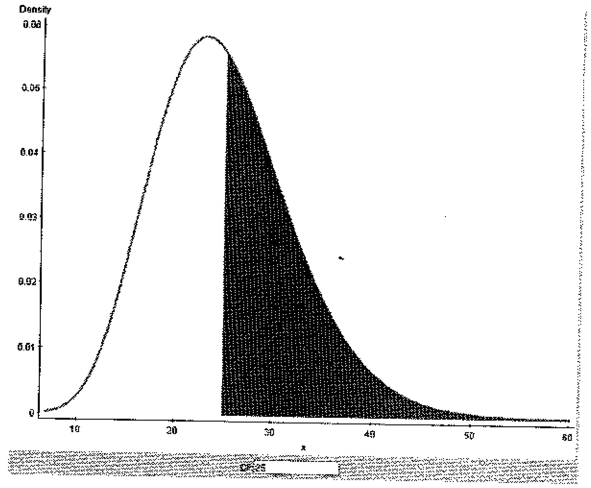
- As degrees of freedom increase, see what happens:



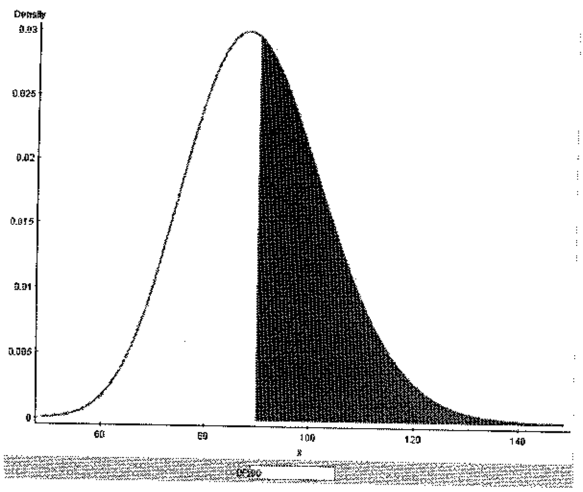
$df = 6$
$\mu = 6$
Mode = 4



df = 12
μ = 12
mode = 10



df = 25
μ = 25
mode = 23



df = 90
μ = 90
mode = 88

- For chi-square models, the mean is always equal to $\mu = df$.
- For chi-square models, the highest peak, or the Mode, is at $df - 2$.

In the limit (as $n \to \infty$) χ^2 model goes to a normal model.

Chi-Square Technique #1
Goodness-of-Fit Test

Example: Much debate has taken place among sports fans to determine which conference ranks supreme for NCAA college football. If we use ***Number of National Championships*** as our benchmark, is there evidence that certain conferences are superior?

Data were collected since 1950 and the number of football national championships were tallied up – we have actually have 80 champions for 61 years since during certain years past, multiple schools were crowned national champion (go figure!).

Here are the counts:

NCAA Conference	Big 12	SEC	Big 10	ACC	PAC 10	Other
Number of Titles	22	18	14	10	8	8

- If college football talent was equally spread among the conferences, we'd expect:

Each conference to have $\frac{80}{6} = 13.33$ championships.

- Do we have evidence that this <u>is not</u> the case? To start, write out the hypotheses to run the goodness-of-fit test:

H_0: Championships are equally-spread (uniformly) over the conferences.

H_A: Championships are <u>not</u> uniformly spread.

- Technically, we're testing this null hypothesis, though we never write it this way:

$$H_0: p_{BIG\,12} = p_{SEC} = p_{BIG\,10} = p_{ACC} = p_{PAC\,10} = p_{OTHER}$$

We cannot use a one-sample z-test for proportions because we are testing multiple proportions simultaneously. Thus, we must use a Goodness-of-Fit Test with the χ^2 statistic.

- If the null hypothesis were true, what would we expect the counts to be?

NCAA Conference	Big 12	SEC	Big 10	ACC	PAC 10	Other
Actual / Observed Number of Titles	22	18	14	10	8	8
Expected Number of Titles	13.33	13.33	13.33	13.33	13.33	13.33
Differences	+8.67	+4.67	+0.67	-3.33	-5.33	-5.33

- The χ^2 statistic looks at the differences between the observed counts and the expected counts.
- If the null hypothesis model fits our data well, the differences should be small.
- If the null hypothesis model **does not** fit our model well, the differences should be large and our test statistic will be large. This is evidence that the alternative hypothesis is plausible.

Goodness-of-Fit Test Statistic

$$\chi^2 = \sum \left[\frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}\right]$$ with df = number of categories − 1

Notes: Since positive and negative differences will always cancel out, we need to square the differences (which also puts emphasis on the large differences, sort of a penalty).

As we have more categories, the test statistic will always get bigger since we're adding up more things. Therefore, we scale the test statistic by dividing each squared difference by its expected count.

Calculate the test statistic by hand:

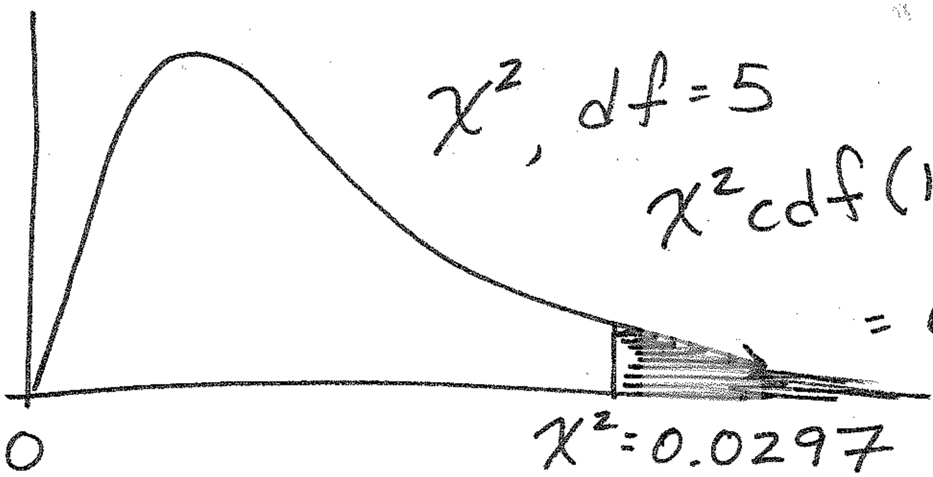
$$\chi^2 = \frac{8.67^2}{13.33} + \frac{4.67^2}{13.33} + \frac{0.67^2}{13.33} + \frac{(-3.33)^2}{13.33} +$$

$$\frac{(-5.33)^2}{13.33} + \frac{(-5.33)^2}{13.33} = 12.40$$

with $df = 6 - 1 = 5$

Chi-Square tests are always upper tail tests. We are looking for large differences from the null model to give a large test statistic. Always shade right.

Determine the P-value and make a decision and conclusion:



χ^2, $df = 5$

$\chi^2 cdf(12.40, E99, 5)$

$= 0.0297$

With the low P-Value, reject H_0.

We conclude NCAA championships are not uniformly spread across the six conferences.

Conditions and Assumptions for Chi-Square Tests

1. The data need to be counts.
2. The sample should be random, or at least unbiased. This helps to ensure the counts in the cells are independent of each other.
3. The expected number of observations in each cell must be at least 5.

Example: In 2010, there were 139 days in which the stock market went up (the S&P 500 Index). If day of the week was irrelevant, the number of up days should be proportional to the number of times the market was open on that day.

Here are the summarized data for the entire year (percentage of trading days are not equal due to holidays when the market is closed):

Day of Week	Monday	Tuesday	Wednesday	Thursday	Friday
Actual Number of "Up" Days	28	25	29	31	26
Percentage of Trading Days	19.0%	21.5%	20.7%	20.2%	18.6%
Expected Number of "Up" Days	26.41	29.885	28.773	28.078	25.854

Is there evidence that the distribution of up days differs from what we'd expect if day of the week was irrelevant. Run the goodness of fit test.

H_0: Day of week is irrelevant for stock market to have an "up" day

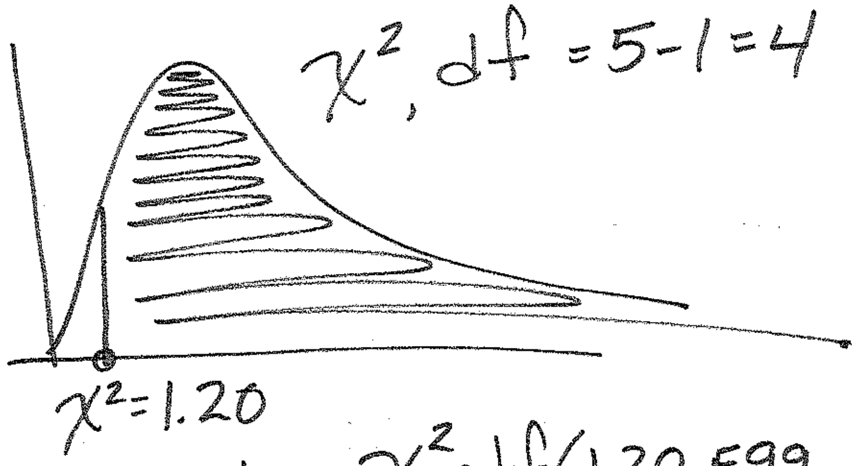
H_A: Day of week matters.

$$\chi^2 = \frac{(\text{obs} - \text{Exp})^2}{\text{Exp}} =$$

$$\frac{(28-26.41)^2}{26.41} + \cdots + \frac{(26-25.854)^2}{25.854} =$$

1.20

Get P-Value:



P-Value = χ^2cdf(1.20, E99, 4) = 0.878

Fail to Reject Ho, no evidence that day of the week matters when predicting a stock market "Up" day.

Examining Residuals

- Whenever we reject the null, it's a good idea to investigate which cells are have the unusual counts.
- We must standardize the residuals; in other words, transform them to z-scores.
- Here is the formula to transform the residuals:

$$c = \frac{(\text{obs} - \text{Exp})}{\sqrt{\text{Exp}}} = \frac{\text{Difference}}{\sqrt{\text{Exp}}}$$

- When looking at the standardized residuals, recall the 68-95-99.7 Rule and remember what z-scores are unusual or rare. Look back to the categories and dive deeper into your collected data.

Example: Going back to the NCAA championships dataset, examine the residuals and make any relevant conclusions.

NCAA Conference	**Big 12**	**SEC**	**Big 10**	**ACC**	**PAC 10**	**Other**
Actual / Observed Number of Titles	**22**	**18**	**14**	**10**	**8**	**8**
Expected Number of Titles	13.33	13.33	13.33	13.33	13.33	13.33
Difference	8.67	4.67	0.67	-3.33	-5.33	-5.33
Standardized Residual	2.37	1.28	0.18	-0.91	-1.46	-1.46

$$c = \frac{8.67}{\sqrt{13.33}} = 2.37 \quad \text{etc...}$$

Since 1950, only the big 12 has an unusually high number of NCAA championships.

No other conference is much beyond or below the expected number of championships.

Chi-Square Test of Homogeneity

- Homogeneity means that "things are the same".
- We can ask if the distribution of counts for one variable is the same across many different categories.

Example: In May 2011, The Pew Research Center released a report summarizing political attitudes of many subgroups of the U.S. population. Respondents identified as Democrat were further classified as New Coalition Democrats, Hard-Pressed Democrats, and Solid Liberals.

One question asked respondents, "Do business corporations make too much profit?" The table summarizes the results:

	New Coalition Democrats	Hard-Pressed Democrats	Solid Liberals	Total
Yes	109	285	344	738
No	180	76	103	359
Total	289	361	447	1097

Write the hypotheses to test if the distribution of opinions are the same for the three classifications of democrats:

H_0: The proportion of yes/no is the same for all 3 Democrat classifications

H_A: There is at least one classification with a diffrent yes/no ratio.

To get a sense of the distributions, we must change to percentages:

	New Coalition Democrats	Hard-Pressed Democrats	Solid Liberals	Total
Yes	37.72%	78.95%	76.96%	67.27%
No	62.28%	21.05%	23.04%	32.73%
Total	100%	100%	100%	100%

We run the chi-square test for homogeneity much in the same way as we did for goodness-of-fit:

Test Statistic: $\chi^2 = \frac{(\text{obs} - \text{Exp})^2}{\text{Exp}}$

$$Df = (\text{Rows} - 1)(\text{Columns} - 1)$$

We need a table of the expected counts. In each cell goes:

$$\frac{\text{Row Total} \times \text{Column Total}}{\text{Total}}$$

Expected Counts

	New Coalition Democrats	Hard-Pressed Democrats	Solid Liberals	Total
Yes	194.42	242.86	300.72	
No	94.58	118.14	146.28	
Total				

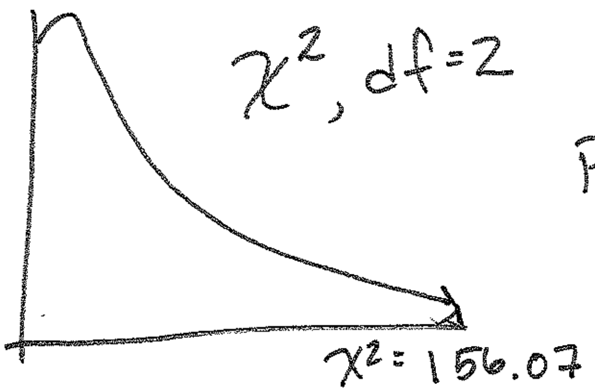
Calculate the test statistic by hand:

$$\chi^2 = \frac{(109 - 194.42)^2}{194.42} + \cdots + \frac{(103 - 146.28)^2}{146.28}$$

$$= 156.07$$

$$DF = (2-1)(3-1) = 2$$

Determine the P-value and make a conclusion:



P-Value ≈ 0

Reject H_0, conclude that at least one classification of Democrats feels differently about Businesses making too much profit.

Examine the residuals, same formula as before:

	New Coalition Democrats	Hard-Pressed Democrats	Solid Liberals	Total
Yes	-6.13	2.70	2.496	
No	8.78	-3.88	-3.58	
Total				

$$C = \frac{(\text{obs} - \text{Exp})}{\sqrt{\text{Exp}}}$$

$$\text{E.G.} = \frac{(109 - 194.42)}{\sqrt{194.42}} = -6.13$$

Every standardized residual is very unusual to outright rare. Tons of evidence the 3 classifications of Democratics differ on this question

Example: In April 2009, Gallup published results from data collected from a large sample of adults in the 27 European Union member states. One of the questions asked was, "Which is the most practicable and realistic option for child care, taking into account the need to earn a living?"

	Male	Female
Both parents work full time	161	140
One works full time, other part time	259	308
One works full time, other stays home for kids	189	161
Both parents work part time	49	63
No opinion	42	28

Is there a difference in the male / female distribution of responses?
Run the test for homogeneity, showing all steps.

H_0: Males and Females feel the same ~~option~~ childcare options.
H_A: Males and Females feel differently.

STAT CRUNCH → Look at two pie charts.
STAT → Tables → Cont. → Summary
CHI-SQUARE FOR INDEP.

$\chi^2 = 12.49$
P-Value = 0.0141
Reject H_0, conclude men/women differ on what's most realistic for child care.

Residuals?

Example: The MPAA movie ratings are always controversial because at the present, moderate language or sexuality will garner an R rating but extreme gore and violence are sneaking through with PG-13 ratings.

Back in the mid 1980s when the PG-13 rating was invented, suppose the distribution of movies was G = 5.2%, PG = 18.3%, PG-13 = 9.9%, R = 63%, Other = 3.6%.

Run the appropriate test on StatCrunch to determine if the distribution is different today. Use the "***2010 Movie Revenue***" dataset.

This is a goodness-of-fit test.

H_0: Movies today follow the distribution from 1980s

H_A: Movies today follow a different distribution.

n = ~~[illegible]~~ 379 movies

Expected Counts

G = 0.052(~~[illegible]~~ 379) = ~~[illegible]~~ 19.708

PG = ~~[illegible]~~ 69.357

PG-13 = ~~[illegible]~~ 37.521

R = ~~[illegible]~~ 238.77

other = ~~[illegible]~~ 13.644

(All expecteds exceed 5)

Actual Counts

G = 9

PG = 64

PG13 = 113

R = 184

other = 9

GOODNESS-OF-FIT

DF = 4 $\chi^2 = 172.21$

P-Value < 0.0001

Reject Ho, conclude movie dist. of ratings differs now, compared to the 1980s.

Chi-Square Test For Independence

- The mechanics of the test are identical to that for the test for homogeneity.
- When given a contingency table with two categorical variables, we'd like to see if the distribution of one is contingent on the other.
- We can test for independence with a chi-square test statistic.

Example: In Math 127, we asked students for their living situation and whether or not they pay bills. We would think (we've been wrong before), that these two variable are related.

Contingency table results:
Rows: Living Situation
Columns: Bills

	I don't pay bills	I pay about half the bills	I pay all the bills	I pay less than half of the bills	I pay more than half of the bills	Total
Live with Parents	46	4	4	32	1	87
Own	2	1	3	3	0	9
Rent	1	5	4	4	1	15
Something Else	1	0	0	3	1	5
Total	50	10	11	42	3	116

Conditions Violated – We need to expect at least 5 in each cell. Expected counts are given by $\frac{\text{Row Total} \times \text{Column Total}}{\text{Total}}$. This will surely be violated.

Solution: We will combine categories and nix the "***Something Else***" category.

	I don't pay bills	I pay less than half of the bills	I pay more than half of the bills	Total
Live with Parents	46	32	9	87
Own or Rent	3	7	14	24
Total	49	39	23	111

Get the **expected counts**:

	I don't pay bills	I pay less than half of the bills	I pay at least half of the bills	Total
Live with Parents	38.41	30.58	18.03	87
Own or Rent	10.59	8.43	4.97	24
Total	49	39	23	111

The test statistic is the same as for the chi-square test for independence. We use StatCrunch:

Contingency table results:
Rows: var1
Columns: None

	I don't pay bills	Less than half	At least half	Total
Live with Parents	46	32	9	87
Own or Rent	3	7	14	24
Total	49	39	23	111

Chi-Square test for independence:

Measure	DF	Value	P-value
Chi-Square	2	28.162653	<0.0001

Warning: over 20% of cells have a expected count less than 5.
~~Chi-Square suspect.~~

?

Make a conclusion:

P-Value is tiny, Reject Ho.

There is evidence that living situation and bill situation are related.

Chi-Square Tests – Reference Page

Step 1: Write down the correct set of hypotheses, based on the context of the problem. Write the hypotheses in words and in context.

Goodness-of-fit: H_0: The distribution of observed counts follows the specified distribution
Test for Homogeneity: H_0: The distribution of observed counts is the same across categories
Test for Independence: H_0: The distribution of one variable is independent of the other variable

Step 2: Check the conditions:

a. The data must be counts.
b. Each cell must **<u>expect</u>** to have 5 observations.
c. The data were collected randomly or at least in an unbiased manner

Step 3: Convert the data into the test statistic:

$$\chi^2 = \sum \frac{(\text{Observerved} - \text{Expected})^2}{\text{Expected}}$$

Goodness-of-fit: Expected counts are based on the expected distribution defined in H_0. Degrees of freedom are number of categories – 1.

Other Two Tests: Expected counts are Row Total X Column Total / Total. Degrees of freedom are $(\text{Rows}-1)(\text{Columns}-1)$

Step 4: Determine the *P*-value by shading under the chi-square model, always to the right. Use χ^2cdf(left, right, df) on the TI.

NOTE: Steps 3 and 4 can be performed directly on StatCrunch using:

Goodness-of-fit: Type in Observed and Expected counts into two columns. Use Stat→Goodness of Fit.

Other Two Tests: Type in the Observed counts into a table, with row labels in column 1 and column labels in columns 2, 3, …. Use Stat→Tables→Contingency→With Summary.

Step 5: If the *P*-value is low (rule of thumb, under 0.05, but even lower is better), reject the null hypothesis. This means we do have compelling evidence in favor of the alternative hypothesis. Write a concluding remark in context.

Wrap Up Chi-Square Tests for Categorical Variables

- When running a test for independence, do not interpret a small *P*-value as evidence of causation.
- We should also not say that one variable depends on the other variable. There easily and likely will be lurking variables we must consider.
- When rejecting the null hypothesis for chi-square tests, remember that there are many scenarios for this to happen. Inspect the residuals to garner an understanding of exactly which categories or counts are inconsistent with the null.

Example: For each, identify the test and write the correct hypotheses.

a. The manager at Klondike Kate's wonders if customers on Tuesday 1/2 Price Nacho Day have the same preferences for the six nacho plates as the customers who purchase nacho plates during the rest of the week. He'll compare the distributions of nachos chosen on Tuesday versus the rest of the week. ~~Independence~~ Homogeneity.

H_0: Tuesday and Rest of the Week customers have the same preferences

H_A: They don't

b. A student wants to find out if political ideology (conservative, liberal, or moderate) is related to choice of major at Towson. She randomly selects 354 Towson students and texts the question to garner responses. Independence

H_0: Political ideology and major are indep.

H_A: They're not

c. In Cecil County, 89.7% are white, 6.7% are black, 1.2% are Asian, and 2.4% are listed as "Other". Does the college population differ?

Goodness-of-Fit

H_0: Cecil College follows the racial breakdown of the county

H_A: It doesn't

Tests and Intervals for a Population Variance / Standard Deviation

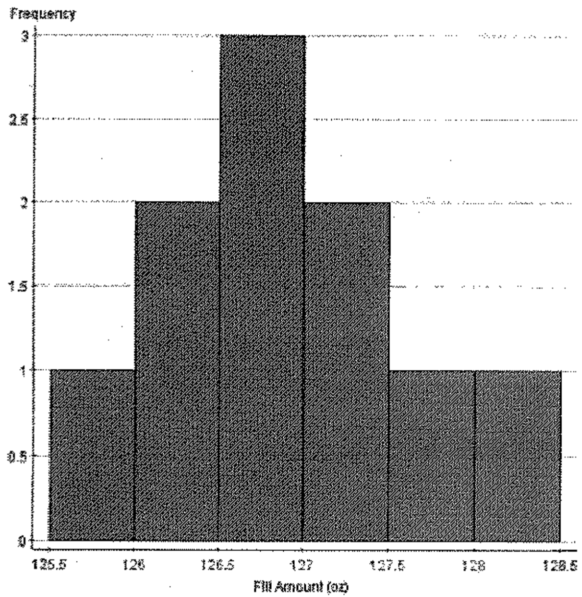
- We have mastered tests for means, but we have never thought about testing measures of variability. That's next, and these tests require the χ^2 model.

Example: The manufacturer of house paint fills each 1-gallon can with 127 fluid ounces of product. Regulatory bodies require that the standard deviation of fill amounts be less than 0.5 ounces, and they check the manufacturer's fill equipment every other month. at $\alpha = 5\%$

This month's measurements are as follows:

126.82	**126.94**	**126.28**	**126.48**	**126.92**	**127.01**	**127.91**	**127.38**	**125.96**	**128.06**

Do we have evidence that the population mean differs from 127 oz.? Do we have evidence that the population standard deviation exceeds 0.5 oz.?



Test for means
No evidence the mean fill amount different than 127 oz.

Summary statistics:

Column	n	Mean	Std. Dev.	Median	Min	Max	Q1	Q3
Fill Amount (oz)	10	126.98	0.666	126.93	125.96	128.055	126.48	127.38

Hypothesis test results:
μ : mean of Variable
$H_0 : \mu = 127$
$H_A : \mu \neq 127$

Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
Fill Amount (oz)	126.977425	0.21048333	9	-0.10724389	0.9169

Definition: If the population standard deviation is σ^2, then the population variance is S^2.

Write the hypotheses to test if the population variance exceeds $\sigma^2 = (0.5)^2 = 0.25$.

$H_0: \sigma^2 = 0.25$

$H_A: \sigma^2 > 0.25$ (test for excess variation)

The test statistic is:

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma^2}$$

The degrees of freedom are $n - 1$.

- Notice that if our observed variance S^2 is much larger than hypothesized, σ^2, the test statistic will be large.
- Notice we run the tests with variances, but can easily translate back to standard deviations.
- We can run "less than" and "not equals to" tests for a population variance using the same test statistic.
- It is crucially important that the data are plausibly normal. Check with graphs.

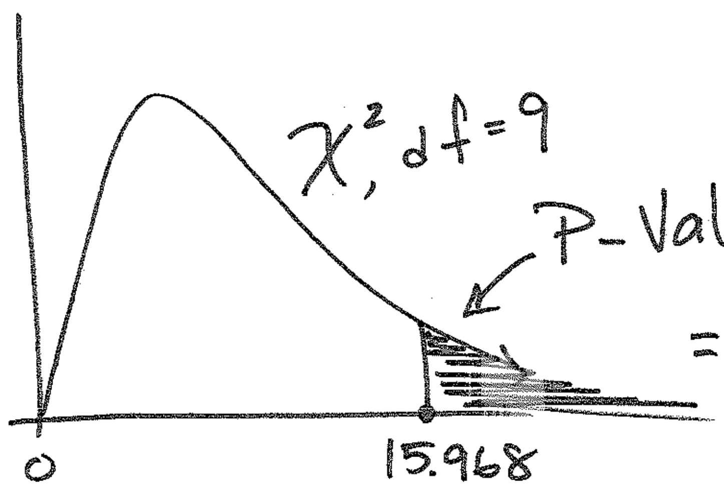
Calculate the value of the test statistic for the paint can example:

$$\chi^2 = \frac{(10-1)(0.666)^2}{(0.5)^2} = 15.968$$

$S = 0.666$ $df = 10 - 1 = 9$

$\sigma = 0.500$

Determine the *P*-value:



P-Value = χ^2cdf(15.968, E99, 9)

= 0.0676

Make a decision and conclusion in the context of the problem:

At $\alpha = 0.05$, fail to reject Ho.
There is no evidence that the standard deviation fill amount exceeds 0.5 ounces.
Company passes bimonthly exam.

Example: In the past, diameters of the ball-bearings produced by a certain manufacturer had a variance of 0.00156 mm. To improve quality, they bought a more expensive piece of machinery, and would like to test if there is less variation in the ball-bearings.

From 101 measurements, the variance of the ball-bearings was 0.00094. At the 1% level of significance, was the new machine worth the cost in reducing variation in the product?

$H_0: \sigma^2 = 0.00156$ mm

$H_A: \sigma^2 < 0.00156$ mm

Data:

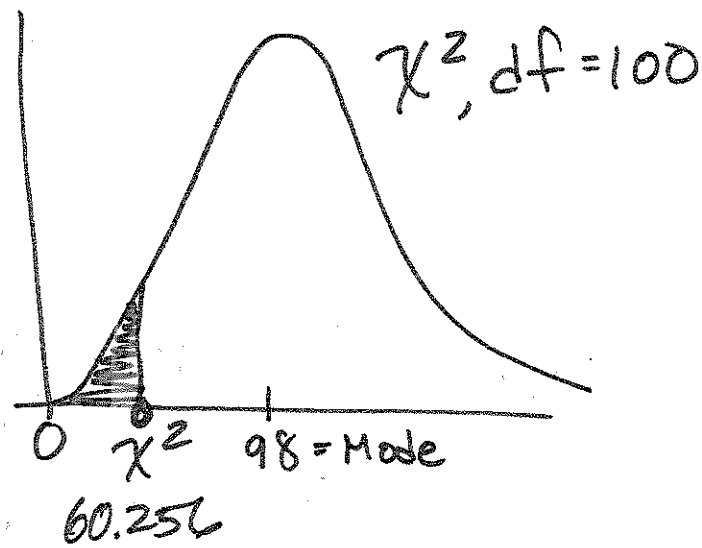
$s^2 = 0.00094$

$\alpha = 0.01$

Must assume normal-looking data.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(10\cancel{0}-1)(0.00094)}{(0.00156)} = 60.256$$

P-Value = $\chi^2\text{cdf}(0, 60.256, 100) = 0.00057$



At $\alpha = 0.01$, Reject Ho.

Conclude this new machine has less variation, as desired.

Confidence Intervals for Variances and Standard Deviations

- A bit of algebra gives us a confidence interval formula for σ^2 and σ:

CI for the Variance

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}, \quad df = n-1$$

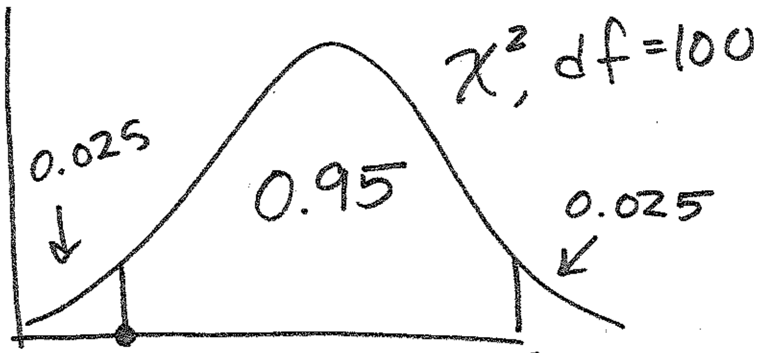
$\alpha/2$ in upper tail $\qquad$ $1-\alpha/2$ in upper tail

CI for the Standard Deviation

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}}, \quad df = n-1$$

Example: For the ball-bearings example, calculate a 95% confidence interval for the true variance for all ball-bearings made by this machine. Recall the sample variance was 0.00094 coming from 101 observations.

$s^2 = 0.00094$



$\chi^2_{0.975} = 74.222 \qquad \chi^2_{0.025} = 129.561$

$$\frac{(101-1)(0.00094)}{129.561} < \sigma^2 < \frac{(101-1)(0.00094)}{74.222}$$

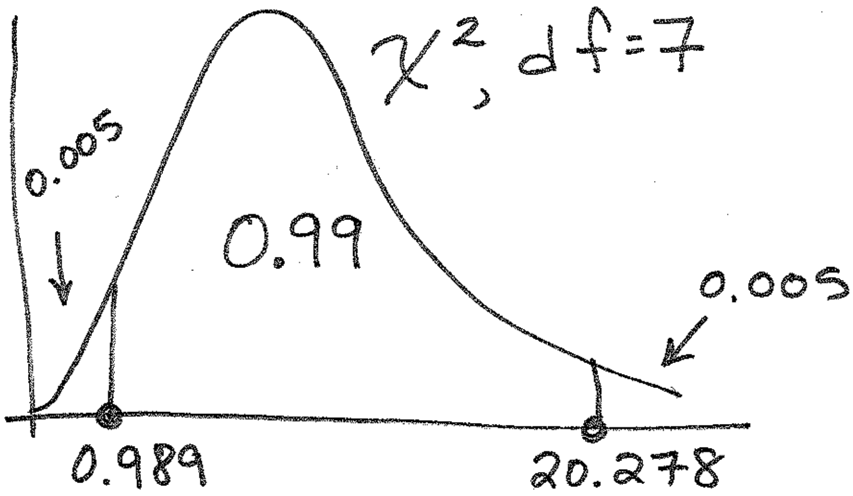
$$0.000726 < \sigma^2 < 0.00127$$

Example: A candy machine dispenses "Hot Tamales" for a quarter. The machine has been giving wildly different numbers of candies on each purchase. Create a 99% confidence interval for the true standard deviation, assuming the following data came from a normal distribution:

22	21	22	24	20	19	21	22

$\bar{y} = 21.375$, $s = 1.506$, $s^2 = 2.268$

$df = 8 - 1 = 7$



99% CI for σ:

$$\sqrt{\frac{(8-1)\,2.268}{20.278}} < \sigma < \sqrt{\frac{(8-1)\,2.268}{0.989}}$$

$$\sqrt{0.7829} < \sigma < \sqrt{16.0526}$$

$$0.885 < \sigma < 4.007$$

99% confident the true pop. SD. lies inside this interval.