Activity #2: Variances; Chi-Square and F-distributions; hypothesis tests

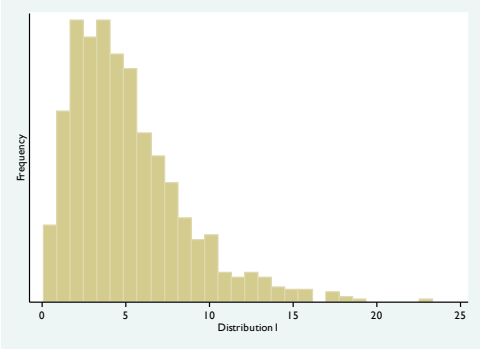
In a previous statistics course, you learned about the Central Limit Theorem:

> If we repeatedly sample ***n*** observations from a population with mean $\boldsymbol{\mu_x}$ and variance $\boldsymbol{\sigma_x^2}$ and calculate the mean, $\overline{X}$, of each sample, the <u>distribution of those sample means</u> will approach a normal distribution with a mean of $\mu_{\overline{X}} = \mu_x$ and standard error of $\sigma_{\overline{X}} = \sigma_x / \sqrt{n}$ as ***n*** increases.

Recall that the CLT applied if the population distribution was normal or if we the sample size was sufficiently large.

Go to the following site to see the CLT in action: http://onlinestatbook.com/stat_sim/sampling_dist/index.html. For this activity, we'll review two examples of the CLT.

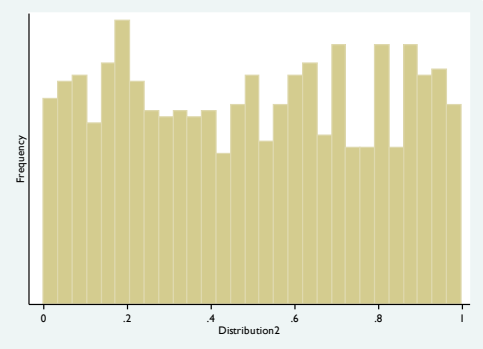
Suppose we have two population distributions. Describe the shape of each distribution.



Distribution of monthly incomes

$\mu = 5.119858$

$\sigma = 3.322086$

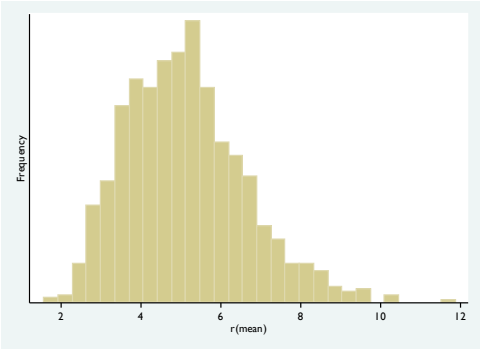


Distribution of arbitrarily chosen numbers

$\mu = 0.4977388$

$\sigma = 0.2930609$

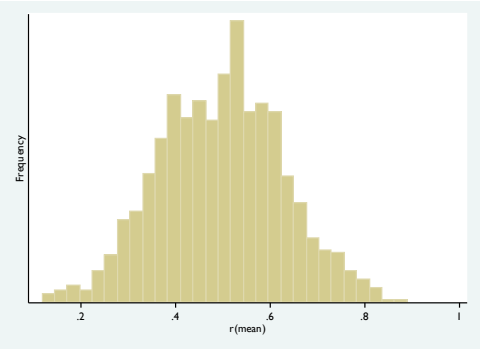
Let's take 5,000 random samples of size n=5 from each distribution. For each of those samples, we'll calculate an average. What should the graph of those 5,000 averages look like? Should the CLT apply to this situation? Let's see:



Distribution of averages from n=5

$\mu = 5.121062$

$\sigma = 1.485658$

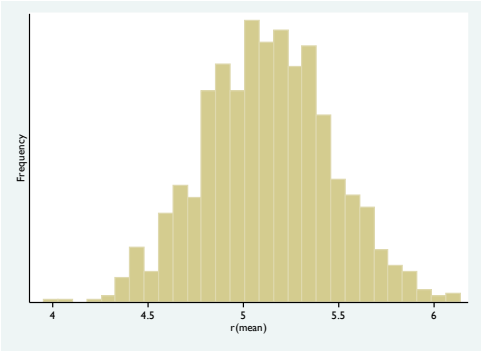


Distribution of averages from n=5

$\mu = 0.4970913$

$\sigma = 0.1313342$

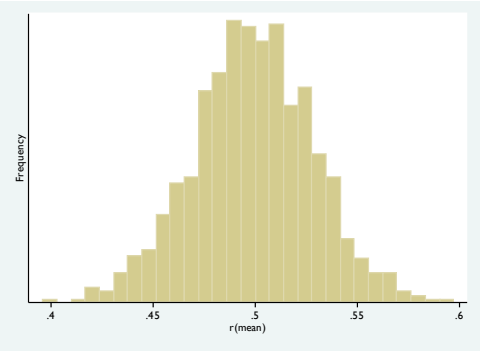
Those distributions do not appear to be approximately normal, so the sample size of n=5 was not sufficiently large. What would happen if we repeatedly took samples of size n=100? Would the distribution of those sample averages approximate a normal distribution? What would the mean and standard error of the sample distributions be?



Distribution of averages from n=100

$\mu = 5.129322$

$\sigma = 0.3349493$



Distribution of averages from n=100

$\mu = 0.4992945$

$\sigma = 0.0293165$

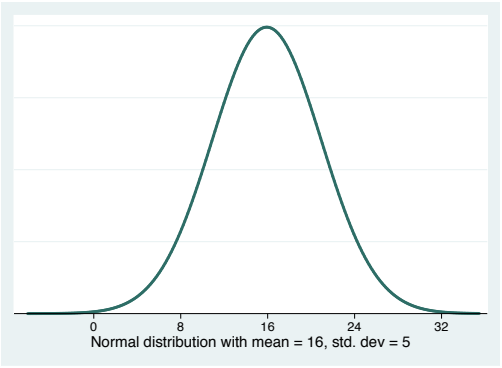
Those distributions do appear to be approximately normal. To check, I took each of these sampling distributions and ran a test called the *Shapiro-Wilk W Test for Normality.* I won't tell you any details about this test other than the fact that it tests a null hypothesis that the data come from a normal distribution.

Running the tests yielded a p-value of 0.97 for the distribution on the left and a p-value of 0.91 for the distribution on the right.

1) How do you interpret these p-values? Do they convince us that the two sampling distributions are normal distributions?

So the CLT tells us what we can expect from sampling distributions of the mean. But what about variances (or standard deviations)? If we repeatedly take samples and calculate the variance of each sample, what will the distribution of those sample variances look like? What will be the mean and standard error of those sampling distributions? If we know the answers to these questions, we could construct confidence intervals for variances and use hypothesis tests to compare variances between two groups.

To get us started, suppose we have a population that follows a normal distribution with $\boldsymbol{\mu_x = 16}$ and variance $\boldsymbol{\sigma_x = 5.}$



Suppose we randomly select 5 observations from this distribution. Notice that while we couldn't predict exactly which 5 values we'd choose, we would be able to predict the values are more likely to come from the center of our distribution. Suppose we take the 5 values that we choose at random and convert them to z-scores. Arbitrarily, we then decide to square those z-scores and find their sum.

2) What is a z-score? How do we convert a number into a z-score?

Now suppose we repeat this process an infinite number of times. We repeatedly sample 5 observations, convert the values to z-scores, square those z-scores, and add them up to get a single number. Can we predict the shape of the distribution of the sum of those squared z-scores?

To get us started, let's pretend we're going to do this process by hand. Suppose we select 5 observations at random and get:

Random sample #1

Value	z	z^2
11.0	$z = \frac{11-16}{5} = -1$	1
15.5	$z = \frac{15.5-16}{5} = -.02$	0.0004
16.0		
16.4		
21.0		
	$\sum_{i=1}^{n} z_i^2 =$	

Random sample #2

Value	z	z^2
1.0	$z = \frac{1-16}{5} = -3$	9
6.0	$z = \frac{6-16}{5} = -2$	4
8.5		
26.0		
31.0		
	$\sum_{i=1}^{n} z_i^2 =$	

3) Which of those random samples are we more likely to get from our population distribution? Use that logic to sketch the distribution of these "sum of squared z-scores" we would expect to get if we repeated this process an infinite number of times. Describe the shape of this distribution.

4) Suppose we repeatedly took random samples of size n = 20 and calculated the sum of the squared z-scores. What would you predict that distribution to look like? Why?

These positively skewed distributions of the sum of squared z-scores are called ***Chi-Square Distributions.***

If $X = Z_1^2 + Z_2^2 + \ldots + Z_v^2$, the distribution of X would then be: $\chi_v^2 = \dfrac{x^{\left(\frac{v}{2}-1\right)} e^{\left(\frac{-x}{2}\right)}}{2^{\left(\frac{v}{2}\right)} \Gamma\left(\frac{v}{2}\right)}$ with **v = n-1** degrees of freedom.

To sketch this distribution, note $E(\chi^2) = v$, $Var(\chi^2) = 2v$, and the mode = v-2.

Why are we learning about the chi-square distribution? What possible use could it have? To answer that, let's go through a simple derivation. Explain what is happening at each step:

$$\chi^2 = \sum_{i=1}^{n} z_i^2 = \sum \left(\frac{x_i - \mu}{\sigma} \right)^2 = \frac{\sum (x_i - \mu)^2}{\sigma^2} = \frac{(n-1)\dfrac{\sum (x_i - \mu)^2}{(n-1)}}{\sigma^2} = \frac{(n-1)s_x^2}{\sigma^2}$$

We can then rearrange terms to see: $\dfrac{\sigma^2 \chi^2}{n-1} = s_x^2$

We just showed that the chi-square distribution is directly related to the sample variance.
Using this fact, along with information about the expected value of a chi-square distribution, we can derive:

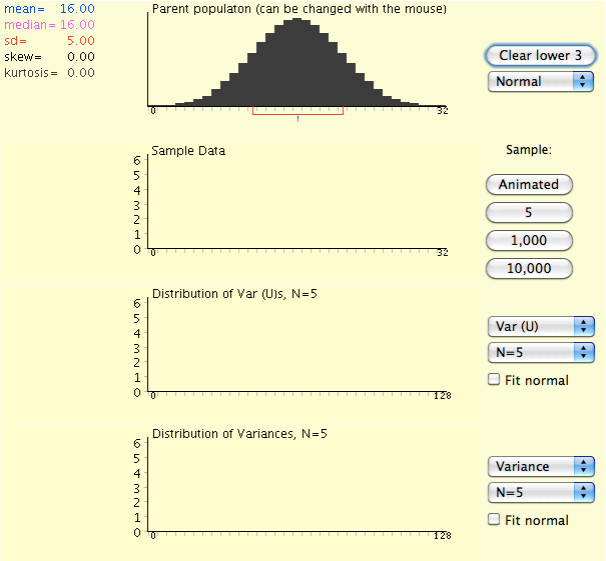
$$E\left[s_x^2\right] = E\left[\frac{\sigma^2 \chi_{n-1}^2}{n-1}\right] = \frac{\sigma^2}{n-1} E\left[\chi_{n-1}^2\right] = \frac{\sigma^2}{n-1}(n-1) = \sigma^2$$

5) What is the importance of what was just derived?

6) Write out the formula used to calculate the standard deviation deviation for a set of data? What does the standard deviation represent? How did the formula differ between a population and a sample standard deviation? How do we convert standard deviations to variances?

Let's simulate the sampling distribution of a sample variance looks like.
Go to http://onlinestatbook.com/stat_sim/sampling_dist/index.html and click ***BEGIN*** to bring up the simulation window.

7) On this simulation, let's begin with a normal distribution for our population (at the top). Using a sample size of **N=5**, let's see what happens if we calculate the sample variance (unbiased, VAR(U) statistic) and population variance of our samples.



Click **ANIMATED** and watch as 5 observations are randomly selected from our population distribution. The sample and population variances are then calculated for these 5 observations.

Which will be larger: the sample variance or the population variance?

Note that our population has a standard deviation of 5 and, therefore, a variance of 25. Go ahead and click ***10,000*** and see what happens when we take 10,000 samples of size ***n=5*** and calculate the variance for each sample.

8) Describe the shape of the sampling distributions of the sample and population variances. Record the mean for each distribution in the table below:

Sampling Distribution	Mean	Shape
Var(U): Sample variance		
Population Variance		

Remember that an estimate is unbiased if its expected value equals the parameter it is estimating. According to our simulation, which statistic (the sample variance or the population variance) is unbiased?

9) Click **CLEAR LOWER 3** to reset the simulation. This time, let's try a sample size of ***n=25***. Take 10,000 samples and comment on the shape of the distribution of the sample variances. Are these distributions normal in shape?

10) Click **CLEAR LOWER 3** to reset the simulation. This time, let's try a different population distribution. Change the population to a **UNIFORM** distribution and take 10,000 samples of size **n=5.** Why are the distributions positively skewed?

11) Sketch two chi-square distributions below: one with 2 degrees of freedom and one with 25 degrees of freedom. Label the mean and standard deviation of these distributions.

12) In a previous statistics course, you may have briefly discussed *degrees of freedom*. One working definition of degrees of freedom is: *The number of independent scores used in the estimate minus the number of parameters estimated en route to the estimation of the parameter of interest.*

When we're dealing with variances, we're using the sample variance as an estimate of the population variance. Let's look at the formula for our estimate and see if we can figure out the degrees of freedom:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{n-1}$$

If we have a sample of size ***n,*** how many "independent scores" do we use to calculate our sample variance?

How many parameters did we need to estimate in our formula?

So for a sample of size ***n,*** how many degrees of freedom would a chi-square distribution have?

13)You should have a chi-square distribution table in front of you. Suppose we have a chi-square distribution with 8 degrees of freedom. Sketch this distribution below and label the mean.

Find the values **a** and **b** such that $P\left(\chi_8^2 < a\right) = 0.025$ and $P\left(\chi_8^2 > b\right) = 0.025$

Label and shade in these areas. How much area under the chi-square distribution is between ***a*** and ***b***?

10) Let's generalize what we just did:

$$0.95 = P(2.18 < \chi^2_8 < 17.53)$$

$$0.95 = P(\chi^2_{8,0.025} < \chi^2_8 < \chi^2_{8,0.975})$$ *Just rewriting our values of 2.18 and 17.53 as chi-squares*

$$100(1-\alpha) = P\left(\chi^2_{n-1,\frac{\alpha}{2}} < \chi^2_8 < \chi^2_{n-1,1-\frac{\alpha}{2}}\right)$$ *Rewriting our 0.95 as having an alpha error of 0.05*

$$100(1-\alpha) = P\left(\chi^2_{n-1,\frac{\alpha}{2}} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}\right)$$ *Rewriting chi-square using what we derived earlier.*

$$100(1-\alpha) = P\left(\frac{1}{\chi^2_{n-1,\frac{\alpha}{2}}} < \frac{\sigma^2}{(n-1)s^2} < \frac{1}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right)$$ *Taking the reciprocal of everything in the parentheses*

$$100(1-\alpha) = P\left(\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right)$$ *Multiplying to get the variance by itself*

We just derived the confidence interval for a population variance. A 95% confidence interval for σ would be:

$$s\sqrt{\frac{n-1}{\chi^2_{n-1,\ 0.025}}} < \sigma < s\sqrt{\frac{n-1}{\chi^2_{n-1,\ 0.975}}}$$

Let's practice using this confidence interval.

11) The PGA establishes strict rules for the golf balls used in its tournaments. In 1942, the PGA established a rule stating that golf balls must have an initial velocity between 243.75 and 256.25 feet per second (when measured at sea-level in 70° weather). Using the current process and materials, your company has manufactured golf balls whose initial velocity is normally distributed with a standard deviation of 4.3 ft/sec. Using an experimental manufacturing process, you sample 16 golf balls and calculate their initial velocity (mean = 250; standard deviation = 2.4). Does the experimental process have significantly more or less variation than the current process? (Use α=0.05)

12) The inventor of another experimental golf ball manufacturing process advertises a precision level of σ=0.8 ft/sec for the initial velocity of golf balls created by his process. You sample 25 balls and find a standard deviation of 1.1 ft/sec. Does his process meet his advertised claims? Use α=0.01.

13) **Complete this exercise and turn it in at the beginning of our next class session**:

Diabetic patients monitor their blood sugar levels with a home glucose monitor that analyzes a drop of blood from a finger stick. Although the monitor gives precise results in a laboratory, the results are too variable when it is used by patients. A new monitor is developed to improve the precision of the assay results under home use. Home testing on the new monitor is done by 25 persons using drops from a sample having a glucose concentration of 118 mg/dl. If σ < 10 mg/dl, then the precision of the new device under home use is better than the current monitor. The readings from the 25 tests are as follows:

125	123	117	123	115
112	128	118	124	111
116	109	125	120	113
123	112	118	121	118
122	115	105	118	131

The following statistics were calculated from the 25 tests: $\overline{X} = 118.5$ and $s = 6.2$.

Construct a 90% confidence interval to determine if the new home monitor is better than the previous monitor.