Central Limit Theorem

(Convergence of the sample mean's distribution to the normal distribution)

Let $X_1, X_2, \ldots, X_n$ be a random sample drawn from any distribution with a finite mean μ and variance σ^2. As $n \to \infty$, the distribution of:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

"converges" to the distribution $N(0,1)$. In other words,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1).$$

Note 1: What is $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$? Remember that we proved that $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \sigma^2/n$. That means we are taking the random variable $\bar{X}$, subtracting its mean, and dividing by its standard deviation. It's a z-score!

Note 2: "converge" means "convergence in distribution:"

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le z\right) = \Phi(z) \text{ for all } z.$$

Don't worry about this if you don't understand (it's beyond the scope of 15.075).

Note 3: CLT is really useful because it characterizes large samples from *any* distribution. As long as you have a lot of independent samples (from any distribution), then the distribution of the sample mean is approximately normal.

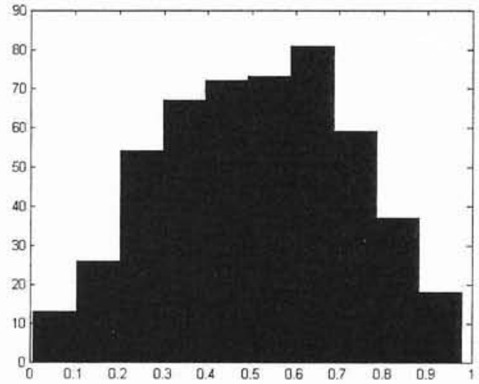
Let's demonstrate the CLT.
Pick n large. Draw n observations from $U[0,1]$ (or whatever distribution you like). Repeat 1000 times.

	x_1	x_2	x_3	x_n	$\bar{x} = \sum_{i=1}^n x_i$
$t = 1$	.21	.76	.57	.84	(.21+.76+...)/n
$\vdots$					
$\vdots$					
$\vdots$					
$t = 1000$					

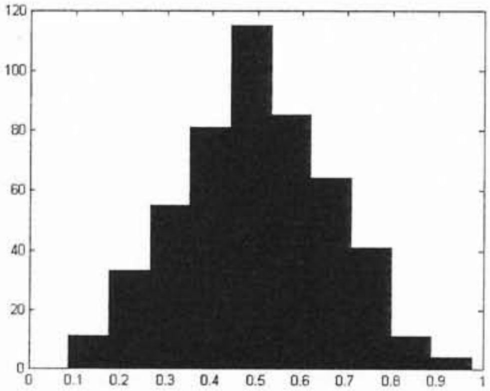
Then, histogram the values in the rightmost column and it looks normal.

CLT demo

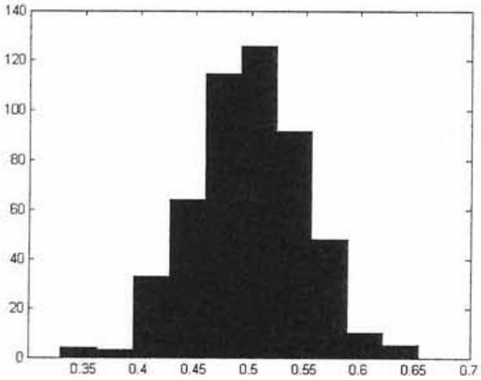
```
n=2;
myrand=rand(n,500);
mymeans=mean(myrand);
hist(mymeans)
```



n=2



n=3



n=30

Sampling Distribution of the Sample Variance - Chi-Square Distribution

From the central limit theorem (CLT), we know that the distribution of the sample mean is approximately normal. What about the sample variance?

Unfortunately there is no CLT analog for variance...
But there is an important special case, which is when $X_1, X_2, \ldots, X_n$ are from a *normal* distribution. (Recall that the CLT applies to arbitrary distributions.)
If this is true, the distribution of the sample variance is related to the *Chi-Square* (χ^2) *distribution.*

Let $Z_1, Z_2, \ldots, Z_\nu$ be $N(0,1)$ r.v.'s and let $X = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$. Then the pdf of X can be shown to be:

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} \quad \text{for } x \geq 0.$$

This is the χ^2 distribution with ν degrees of freedom (ν adjustable quantities). (Note: the χ^2 distribution is a special case of the Gamma distribution with parameters $\lambda = 1/2$ and $r = \nu/2$.)

Fact proved in book:

If $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma)$ r.v.'s, then

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{n-1}.$$

That is, the sample variance times a constant $\frac{(n-1)}{\sigma^2}$ has a χ^2_{n-1} distribution.

Technical Note: we lost a degree of freedom when we used the sample mean rather than the true mean. In other words, fixing $n-1$ quantities completely determines s^2, since:

$$s^2 := \frac{1}{n-1} \sum_i (x_i - \bar{x})^2.$$

Let's simulate a χ^2_{n-1} distribution for $n = 3$. Draw 3 samples from $N(0,1)$. Repeat 1000 times.

	z_1	z_2	z_3	$\sum_{i=1}^3 z_i^2$
$t = 1$	-0.3	-1.1	0.2	1.34
$\vdots$				
$\vdots$				
$\vdots$				
$t = 1000$				

Then, histogram the values in the rightmost column.

For the chi-square distribution, it turns out that the mean and variance are:

$$\begin{aligned}\mathbf{E}(\chi^2_\nu) &= \nu\\ \mathrm{Var}(\chi^2_\nu) &= 2\nu.\end{aligned}$$

We can use this to get the mean and variance of S^2:

$$\begin{aligned}\mathbf{E}(S^2) &= \mathbf{E}\left(\frac{\sigma^2\chi^2_{n-1}}{n-1}\right) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2,\\ \mathrm{Var}(S^2) &= \mathrm{Var}\left(\frac{\sigma^2\chi^2_{n-1}}{n-1}\right) = \frac{\sigma^4}{(n-1)^2}\mathrm{Var}(\chi^2_{n-1}) = \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{n-1}.\end{aligned}$$

So we can well estimate S^2 when n is large, since $\mathrm{Var}(S)$ is small when n is large.

Remember, the χ^2 distribution characterizes normal r.v. with known variance. You need to know σ! Look below, you can't get the distribution for S^2 unless you know σ.

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2) \quad \rightarrow \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Student's t-Distribution

Let $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$. William Sealy Gosset aka "Student" (1876-1937) was looking at the distribution of:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Contrast T with $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ which we know is $N(0,1)$.
So why was Student looking at this?
Because he had a small sample, he didn't know the variance of the distribution and couldn't estimate it well, and he wanted to determine how far $\bar{x}$ was from μ. We are in the case of:

- $N(0,1)$ r.v.'s

- comparing $\bar{X}$ to μ

- unknown variance σ^2

- small sample size (otherwise we can estimate σ^2 very well by s^2.)

Rewrite

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\frac{\sqrt{S^2}}{\sqrt{n}} \frac{1}{\sigma/\sqrt{n}}} = \frac{Z}{\sqrt{S^2/\sigma^2}}.$$

The numerator Z is $N(0,1)$, and the denominator is sort of the square root of a chi-square, because remember $S^2(n-1)/\sigma^2 \sim \chi^2_{n-1}$.
Note that when n is large, $S^2/\sigma^2 \to 1$ so the T-distribution $\to N(0,1)$.

Student showed that the pdf of T is:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\nu/2\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} \qquad -\infty < t < \infty$$

Snedecor's F-distribution

The F-distribution is usually used for comparing variances from two separate sources. Consider 2 independent random samples $X_1, X_2, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$. Define S_1^2 and S_2^2 as the sample variances. Recall:

$$\frac{S_1^2(n_1-1)}{\sigma_1^2} \sim \chi^2_{n_1-1} \qquad \text{and} \qquad \frac{S_2^2(n_2-1)}{\sigma_2^2} \sim \chi^2_{n_2-1}.$$

The F-distribution considers the ratio:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \frac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)}.$$

When $\sigma_1^2 = \sigma_2^2$, the left hand side reduces to S_1^2/S_2^2.
We want to know the distribution of this! Speaking more generally, let $U \sim \chi^2_{\nu_1}$ and $V \sim \chi^2_{\nu_2}$.
Then $W = \frac{U/\nu_1}{V/\nu_2}$ has an F-distribution, $W \sim F_{\nu_1,\nu_2}$.
The pdf of W is:

$$f(w) = \frac{\Gamma\left((\nu_1+\nu_2)/2\right)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} w^{\nu_1/2-1} \left(1 + \frac{\nu_1}{\nu_2} w\right)^{-(\nu_1+\nu_2)/2} \qquad \text{for } \ w \geq 0.$$

There are tables in the appendix of the book which solve:

$$\begin{array}{rlcl} P(\chi^2_\nu > & \chi^2_{\nu,\alpha}) & = & \alpha \\ & \uparrow & & \uparrow \end{array}$$

$$\begin{array}{rlcl} P(T_\nu > & t_{\nu,\alpha}) & = & \alpha \\ & \uparrow & & \uparrow \end{array}$$

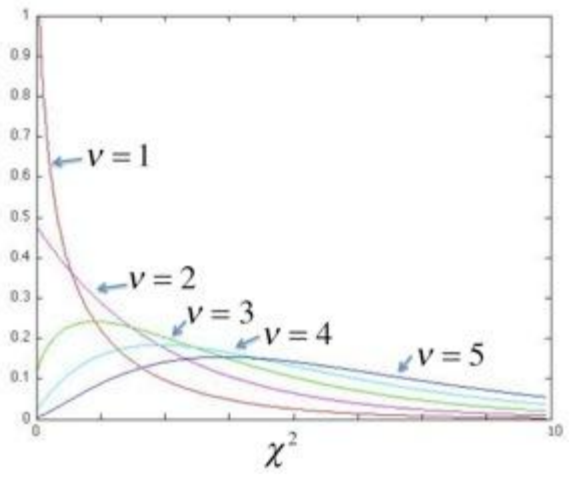
$$\begin{array}{rlcl} P(F_{\nu_1,\nu_2} > & f_{\nu_1,\nu_2,\alpha}) & = & \alpha \\ & \uparrow & & \uparrow \end{array}$$

Note that because F is a ratio,

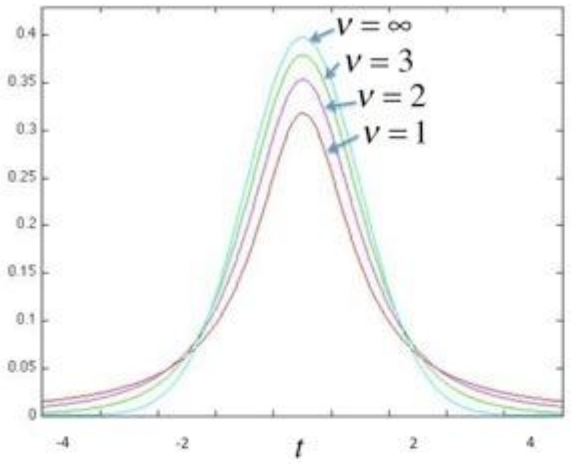
$$F_{\nu_1,\nu_2} = \frac{1}{F_{\nu_2,\nu_1}}$$

which you might need to use in order to look up the F-scores in a table in the book. Actually, you will need to know:

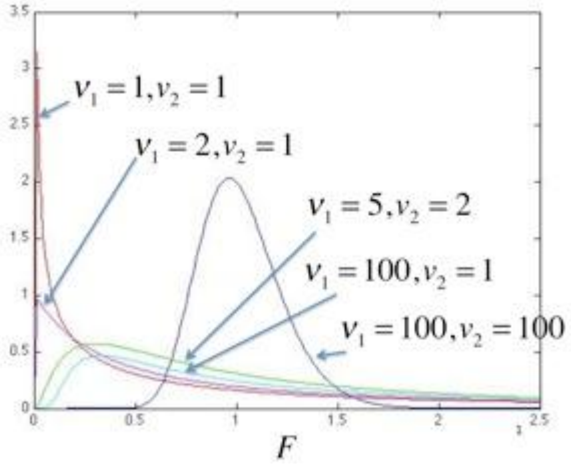
$$f_{\nu_1,\nu_2,1-\alpha} = \frac{1}{f_{\nu_2,\nu_1,\alpha}}.$$



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
$\nu = 1$
$\nu = 2$
$\nu = 3$
$\nu = 4$
$\nu = 5$
χ^2



0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-2
2
4
$\nu = \infty$
$\nu = 3$
$\nu = 2$
$\nu = 1$
t



3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
$\nu_1 = 1, \nu_2 = 1$
$\nu_1 = 2, \nu_2 = 1$
$\nu_1 = 5, \nu_2 = 2$
$\nu_1 = 100, \nu_2 = 1$
$\nu_1 = 100, \nu_2 = 100$
F

15.075J / ESD.07J Statistical Thinking and Data Analysis
Fall 2011