MA Advanced Econometrics: Asymptotics

Karl Whelan

School of Economics, UCD

Revised, February 15, 2011

Plan For Next Three Weeks

- My plan for the next three weeks is to cover the following topics:
  - Asymptotics: Properties of estimators in large samples.
  - Law of Large Numbers, Central Limit Theorem.
  - The Delta method.
  - Application of asymptotic results to least squares regression.
  - Extension to OLS estimation of AR(1).
  - Unit roots.
  - Spurious regressions and cointegration.
  - Finite-sample properties of OLS estimates of AR models.
  - Bootstrapping.
  - Testing for structural change.
- That's the plan: Let's see how far I get!

The Normal Regression Model

- Introductory econometrics often focuses on the regression model

$$Y = X\beta + e \tag{1}$$

where

  - Y is an $n \times 1$ vector of observations of data on the variable to be explained
  - X is an $n \times k$ matrix of data on k variables that are independent of all the error terms e_i
  - β is a $k \times 1$ vector of coefficients.
  - e is an $n \times 1$ vector of error terms that are independently and identically normally distributed with variance σ^2, i.e. $e \sim N\left(0, I_n \sigma^2\right)$.

- In this case, the OLS estimator $\hat{\beta} = (X'X)^{-1} X'Y$ has the property that

$$\hat{\beta} - \beta \sim N\left(0, \sigma^2 (X'X)^{-1}\right) \tag{2}$$

- This allows you to test hypothesis about the true cofficients β and construct confidence intervals.

Why Do We Need Asymptotics?

- However, there is often no good reason to assume that errors are normally distributed. Indeed, there is often no reason to have *a priori* knowledge of what is the underlying distribution of the error term.
- Without an arbitrary assumption about the distribution of the error term, we generally cannot derive the exact distribution of our estimator for the sample size used.
- However, it turns out there are powerful results that allow us to make statements about how the estimator behaves if our sample size becomes large.
- The two key results are:
  1. *The Law of Large Numbers* – this helps us construct estimators that are consistent, i.e. becomes less and less likely to be far away from the true value as sample sizes get bigger.
  2. *The Central Limit Theorem* – this tells us what happens to the distribution of estimators as sample sizes get bigger.

Convergence in Probability

- Consider a sequence of random variables z_n each based on a sample of n observations. We say that z_n **converges in probability** to z as $n \to \infty$ if for all $\delta > 0$
$$\lim_{n\to\infty} \text{Prob}\left(|z_n - z|) \le \delta\right) = 1 \tag{3}$$
- In other words, no matter how small δ is, the likelihood that z_n is no more than δ away from z gets closer and close to one as n gets larger.
- Convergence in probability is often indicated in one of the two ways: Plim $z_n = z$ or $z_n \xrightarrow{p} z$.
- For a sequence of matrices X_n, we say $X_n \xrightarrow{p} X$ if every element of X_n converges in probability to the corresponding element of X.
- The concept of covergence in probability is usually applied to assess estimators. If $\hat{\theta}_n$ is a finite sample estimator of a population moment θ then we could like for this estimator to converge in probability to the true value.
- We say an estimator $\hat{\theta}_n$ based on a sample of size n is **consistent** if $\hat{\theta} \xrightarrow{p} \theta$.

A Useful Result About Convergence in Probability

Continuous Mapping Theorem: Let $\{z_n\}$ denote a sequence of vectors of estimators based on a sample of size n such that $z_n \xrightarrow{p} c$ and let g be a continuous function at c that does not depend upon the sample size n. Then $g(z_n) \xrightarrow{p} g(c)$.

Examples: If $z_{1n} \xrightarrow{p} c_1$ and $z_{2n} \xrightarrow{p} c_2$, then $z_{1n} + z_{2n} \xrightarrow{p} c_1 + c_2$, $z_{1n}z_{2n} \xrightarrow{p} c_1 c_2$ and so on.

This result can also be applied to vectors and matrices of random variables. For example, let X_{1n} denote a $k \times k$ matrix of random variables calculated from samples of size n such that $X_{1n} \xrightarrow{p} C_1$ where C_1 is a nonsingular (invertible) matrix. Let X_{2n} denote a sequence of $k \times 1$ vectors such that $X_{2n} \xrightarrow{p} c_2$. Then

$$\text{Plim} X_{1in}^{-1} X_{2n} = \text{Plim}\{X_{1in}\}^{-1} \text{Plim} X_{2n} = C_1^{-1} c_2$$

Properties of the Sample Mean

- Many estimators in econometrics are constructed from sample means of various sorts.
- Consider a random sample $\{y_1, y_2, ..., y_n\}$ that are independently and identically distributed (i.i.d.) with mean μ and variance σ^2. The sample mean of this population, $\bar{y}_n$ is an unbiased estimator of μ:

$$\mathbb{E}\,\bar{y}_n = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,y_i = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \qquad (4)$$

- The variance of the sample mean can also be shown to be

$$\mathrm{Var}\left(\bar{y}_n\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu\right)\right)^2 \qquad (5)$$

$$= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu\right)\right)\left(\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \mu\right)\right) \qquad (6)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(y_i - \mu\right)\left(y_j - \mu\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n} \qquad (7)$$

Generalised Chebyshev's Inequality

Let X be a random variable with a finite value for $\mathbb{E}|X|^r$. Then

$$\text{Prob}\{|X-c|>\delta\} \leq \frac{\mathbb{E}|X-c|^r}{\delta^r} \tag{8}$$

Proof: Let the probability density function of X be denoted by f_X. Let S denote the set of all x such that $|X-c|>\delta$ and $\tilde{S}$ be its complement, i.e. the set of all x such that $|X-c|\leq\delta$. Then one can calculate the following expectation

$$\begin{aligned}
\mathbb{E}|X-c|^r &= \int |X-c|^r f_X(x)dx \\
&= \int_S |X-c|^r f_X(x)dx + \int_{\tilde{S}} |X-c|^r f_X(x)dx \\
&\geq \int_S |X-c|^r f_X(x)dx \\
&\geq \int_S \delta^r f_X(x)dx \\
&= \delta^r \text{Prob}\{|X-c|>\delta\}
\end{aligned}$$

Weak Law of Large Numbers

- We can use Chebyshev's inequality to prove that sample means are consistent for the case where sample observations are independently and identically distributed with finite variance.
- Using $r = 2$ the inequality becomes

$$\text{Prob}\{|X - c| > \delta\} \leq \frac{\mathbb{E}\,|X - c|^2}{\delta^2} \tag{9}$$

- Now apply to the case of a sample mean

$$\text{Prob}\{|\bar{y}_n - \mu| > \delta\} \leq \frac{\mathbb{E}\,|\bar{y}_n - \mu|^2}{\delta^2} = \frac{\sigma^2}{n\delta^2} \tag{10}$$

- The probability bound on the right-hand side gets smaller as $n \to \infty$.
- This means that $\bar{y}_n \xrightarrow{p} \mu$: Sample means converge in probability to population means. This is known as the **Weak Law of Large Numbers**.
- Note, however, the strong assumptions that we made to get this result: i.i.d y_i observations with finite variance. The i.i.d part can be changed and LLN results still obtained. But, as we shall see, the finite variance assumption is harder to dispense with.

Convergence in Distribution

- The WLLN is a useful result but it just tells us that certain estimators get closer to the population values as samples get larger. It doesn't tell us about the shape of the distribution of the estimator, which is required for confidence intervals.
- Suppose z_n is a statistic constructed from a sample of size n with a cumulative distribution function given by $F_n(u) = \text{Prob}\{z_n \leq u\}$. We say that z_n converges in distribution to a random variable z with distribution F if for all u at which $F(u)$ is continuous, $F_n(u) \to F(u)$ as $n \to \infty$.
- Convergence in distribution is denoted as $z_n \xrightarrow{d} z$. Effectively, it just means that the graph of z's CDF looks more and more like that of z as $n \to \infty$.
- A key result about convergence in distributions is the **Central Limit Theorem**: If a series $\{y_1, y_2, ..., y_n\}$ is i.i.d with mean μ and finite variance σ^2 then as $n \to \infty$
$$\sqrt{n}\left(\bar{y}_n - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right). \tag{11}$$
- This result provides a justification for the usual hypothesis tests and confidence intervals for μ, even when the errors are not normally distributed.

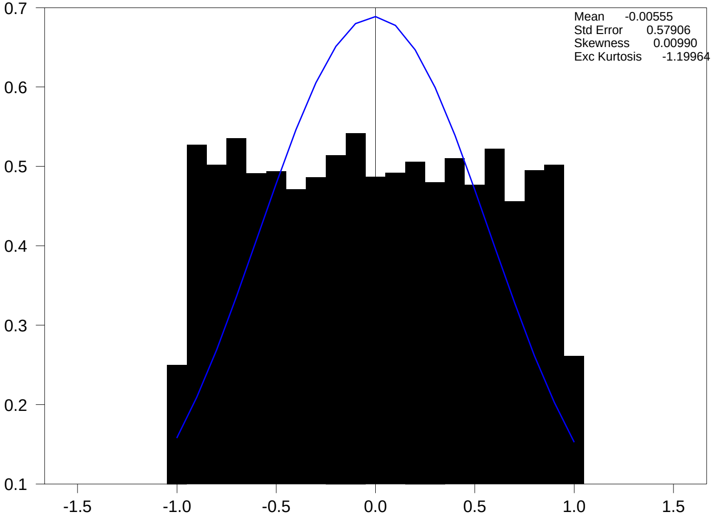
Comments on the Central Limit Theorem

- The CLT is one of the most profound results in mathematics or statistics.
- Think about what it says for a moment: Take averages of samples of size n from a distribution—*any* distribution—and once n gets large, the distribution of the sample-size adjusted average $\sqrt{n}\left(\bar{y}_n - \mu\right)$ takes the form of a Normal distribution.
- How large does n need to get for the distribution of the adjusted sample average to be approximately normal? It depends.
- In the i.i.d case, $n = 30$ is often used as a conservative rule of thumb, meaning that by samples of that size, the distribution is usually close to normal. In practice, the convergence is quicker in many cases.
- But what is the distribution of $\bar{y}_n$? One might be tempted to multiply across by $\sqrt{n}$ and subtract μ and assume $\bar{y}_n \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$. But as $n \to \infty$ the variance of this distribution goes to zero and it collapses on μ so there is no distribution.
- That said $N\left(\mu, \frac{\sigma^2}{n}\right)$ will tend to work well for large values of n and this is often called the asymptotic distribution, denoted $\bar{y}_n \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$

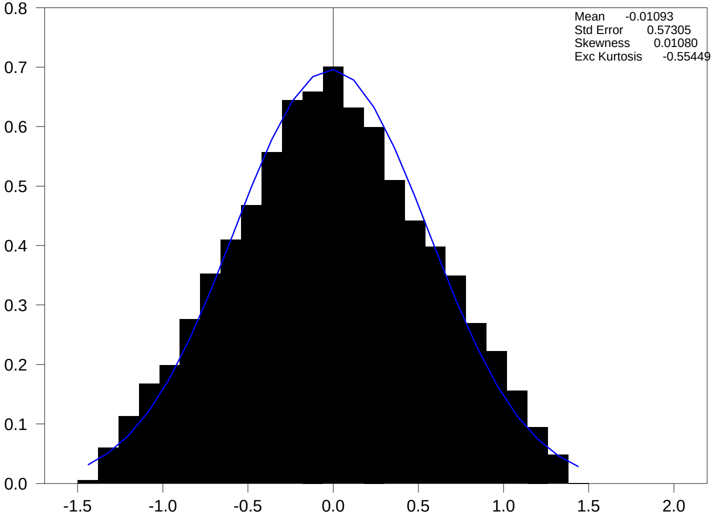
An Example

- Consider the example of taking random draws from a uniform distribution on $[-1, 1]$, in which case the population mean is zero and standard deviation is 0.577.
- The next few graphs shows the results of computer simulations based on 10,000 replications, in which we take sample averages and then chart the adjusted sample average $\sqrt{n}\bar{y}$.
- The amount of data points used to construct the sample averages increases with each chart. The blue line on the chart is the normal distribution corresponding to the mean and variance of the distribution in the chart.
- The graphs also show figures for the skewness (slantedness or asymmetry) and excess kurtosis (presence of fat tails) of the distribution. Both of these values should be zero for a normal distribution.
- We see that for this distribution, the CLT works very well for samples much smaller than $n = 30$.

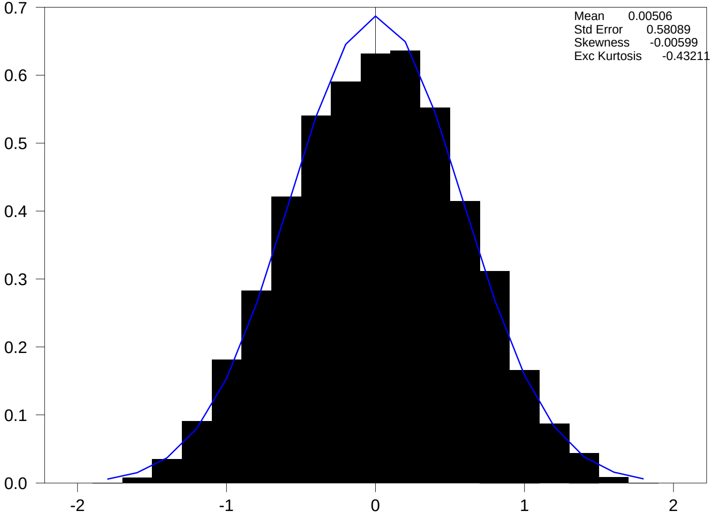
Uniform Distribution



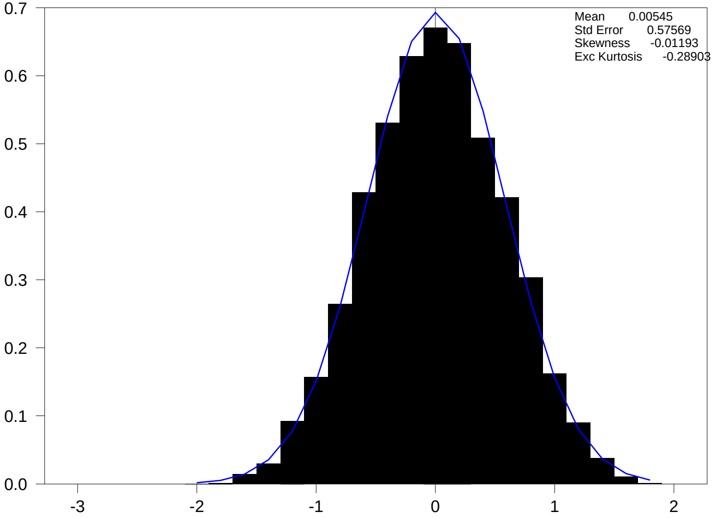
Adjusted Averages from a Uniform Distribution: $n = 2$



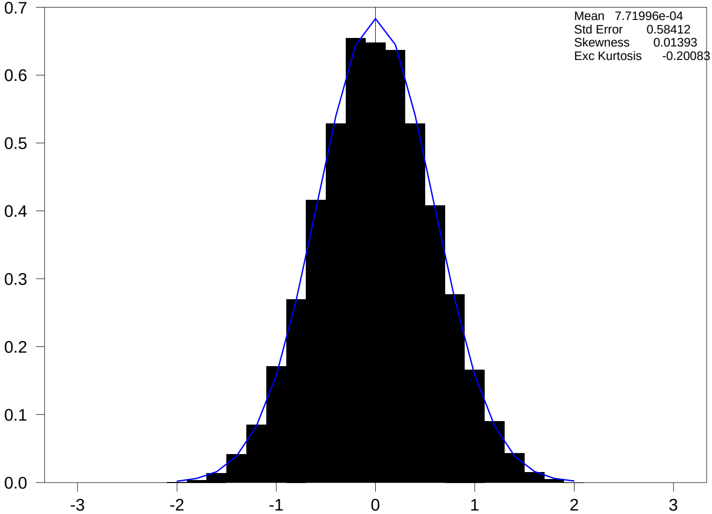
Adjusted Averages from a Uniform Distribution: $n = 3$



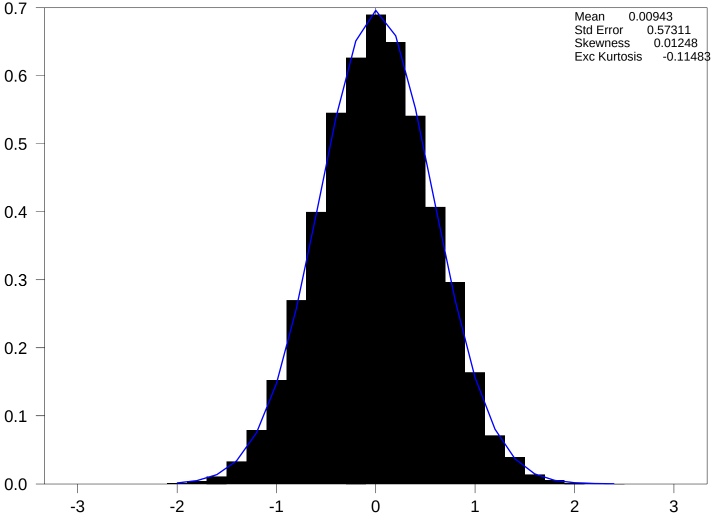
Adjusted Averages from a Uniform Distribution: $n = 4$



Adjusted Averages from a Uniform Distribution: $n = 6$



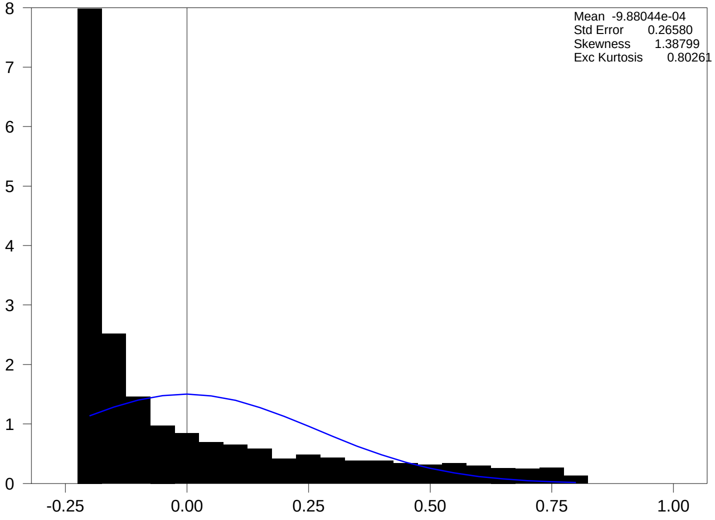
Adjusted Averages from a Uniform Distribution: $n = 10$



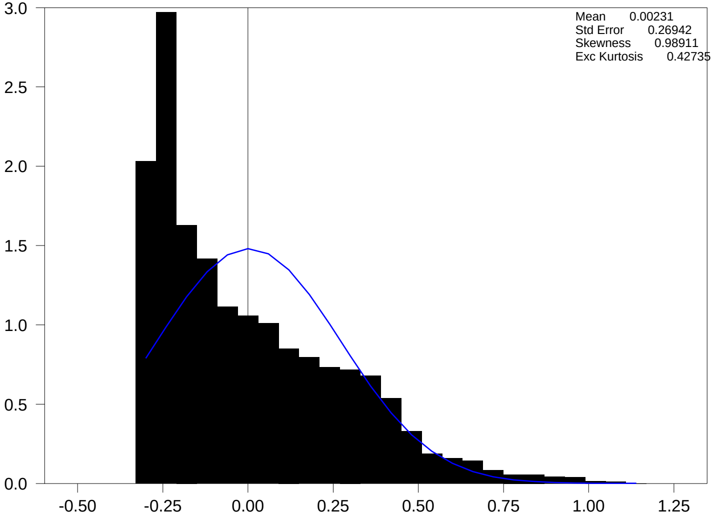
A More Complex Example

- In the uniform distribution case, the distribution of sample averages becomes approximately normal by about $n = 8$.
- But this isn't always the case. Consider the following example: Define the variable $Z_t = X_t^4$ where X_t is uniformly distribution on $[-1, 1]$.
- The distribution of Z_t is highly skewed distribution with no negative values, a mean of 0.2 and a maximum possible value of 1.
- The graph on the next page shows that the distribution of $\sqrt{n}\left(\bar{Z}_t - 0.2\right)$ looks like.
- In this example, the distribution of the adjusted sample averages are also very skewed for small values of n. However, once n gets to 10, the distribution looks a bit more bell-shaped and the Normal approximation works pretty well by the time we get to $n = 30$.
- But only in very large samples over over 1000 do we see the skewness fall to low levels.

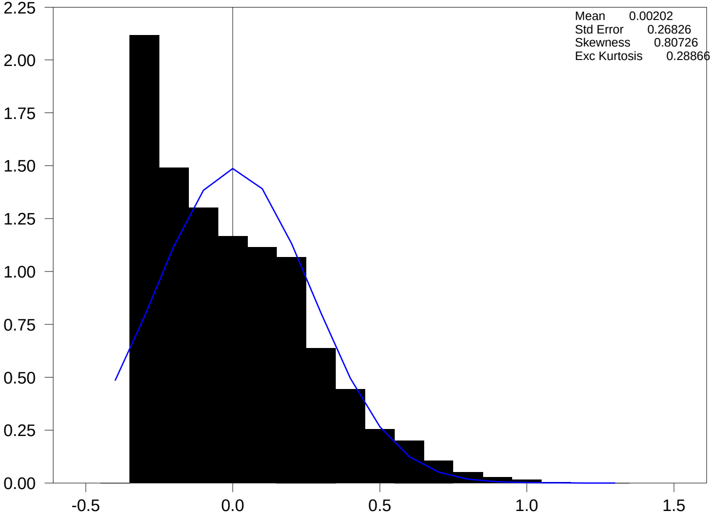
Distribution of Uniform Variable To The Fourth Power



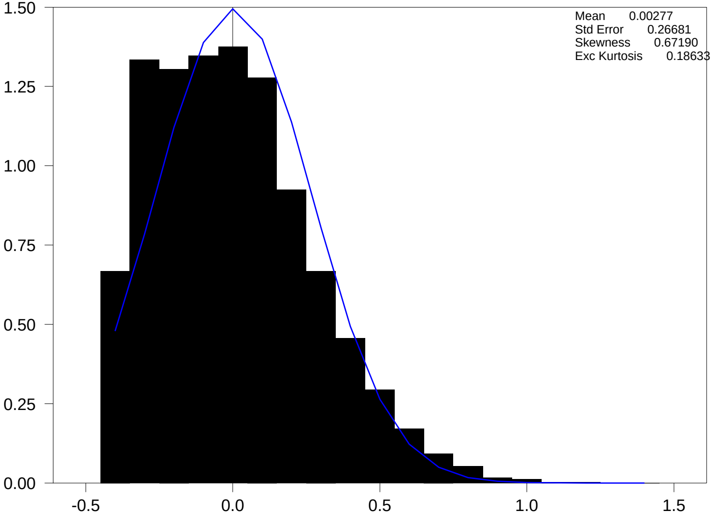
Adjusted Averages: $n = 2$



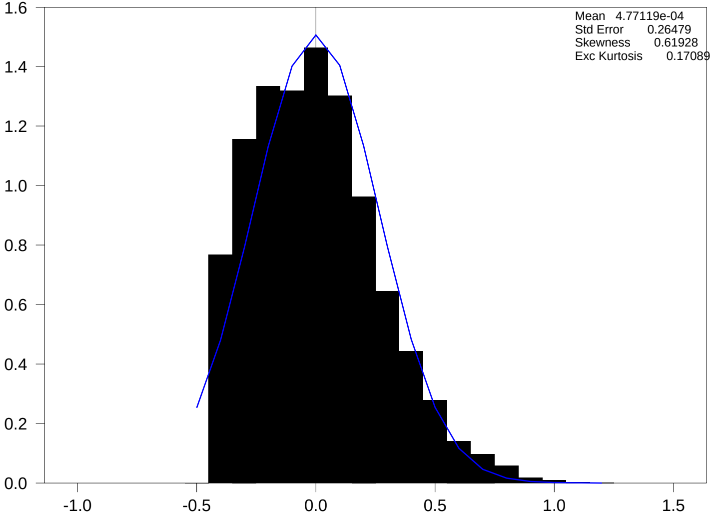
Adjusted Averages: $n = 3$



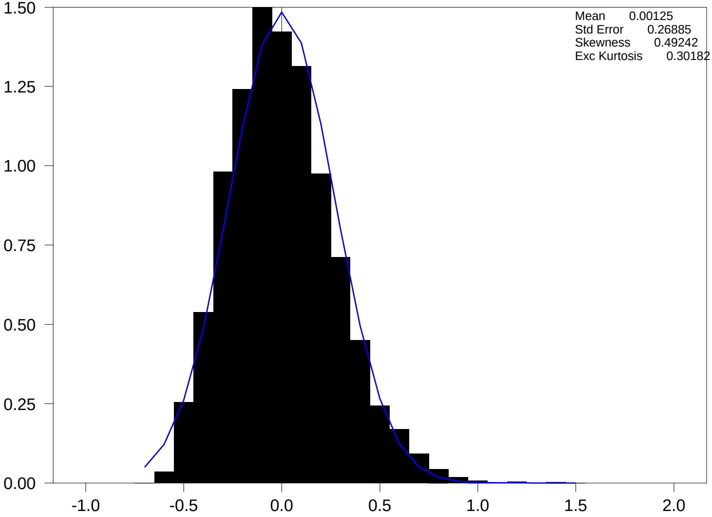
Adjusted Averages: $n = 4$



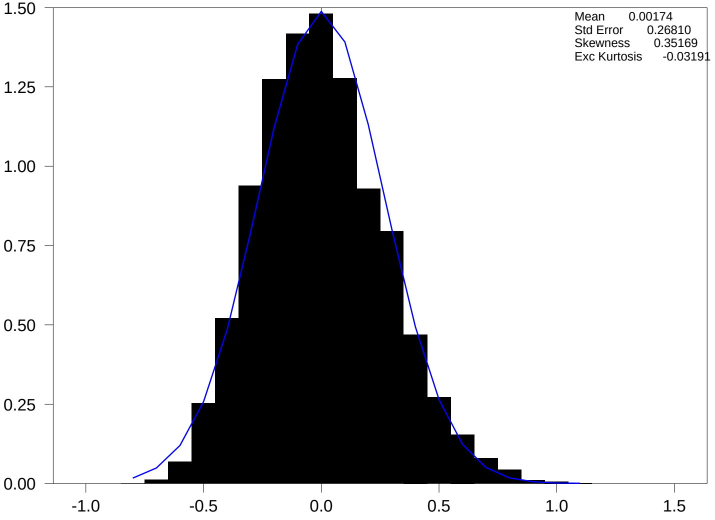
Adjusted Averages: $n = 5$



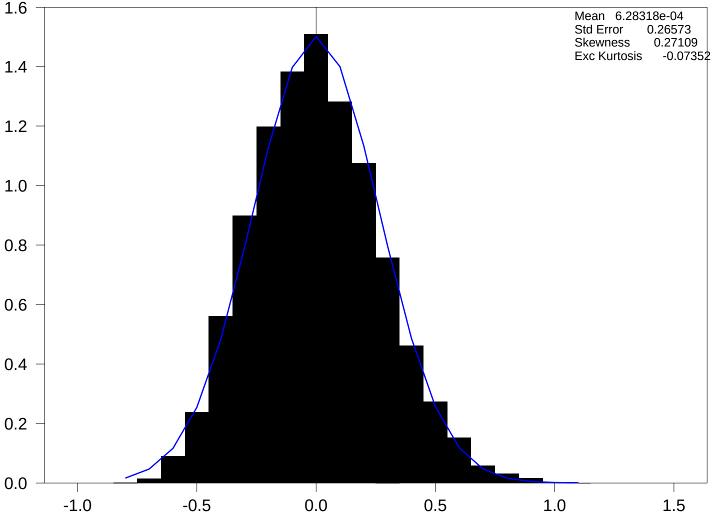
Adjusted Averages: $n = 10$



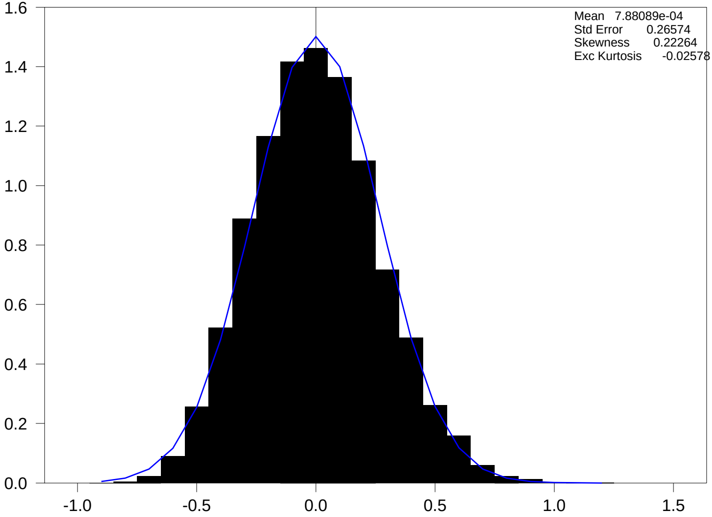
Adjusted Averages: $n = 15$



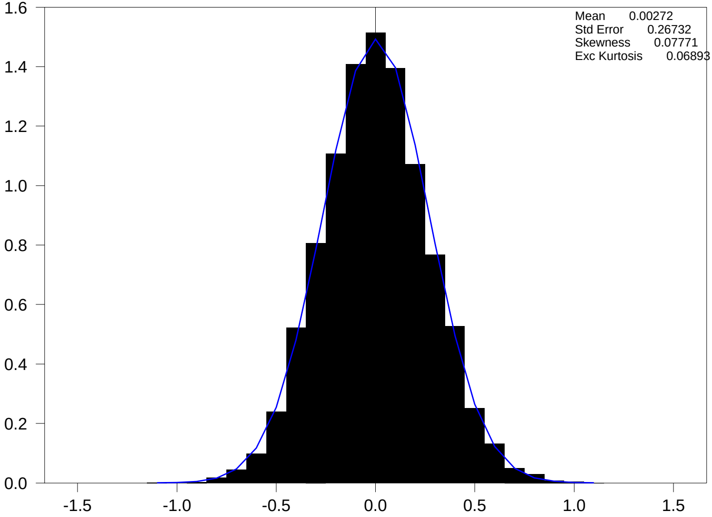
Adjusted Averages: $n = 20$



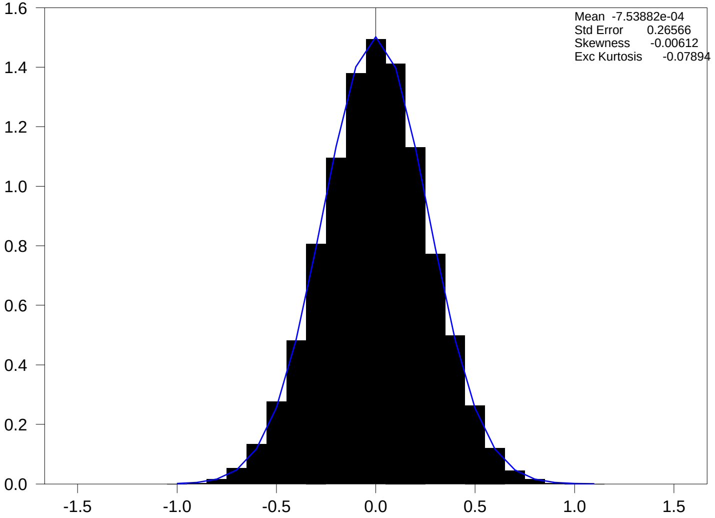
Adjusted Averages: $n = 30$



Adjusted Averages: $n = 1000$



Adjusted Averages: $n = 5000$



Computer Simulation Methods

- These graphs were made using computer simulations. These methods are incredibly helpful for understanding statistical distributions. In this example, the simulations illustrate how analytical results based on asymptotic theorems often work well. However, we will also see examples where asymptotic approximations don't work well and, in these cases, computer simulations are often used to provide a better idea as what the finite-sample distribution of estimators look like.
- The RATS program used to make the uniform distribution charts is repeated on the next page. How to interpret the program?
  - `do ssize=1,10` tells RATS to execute the commands that follow (taking averages for samples of size `ssize`) for a value of `ssize=1` and then when that's finished to repeat the calculations for `ssize=2` and so on up to `ssize=10`. The loop is completed with `end do ssize`.
  - The `do k=1,10000` loop gets us 10,000 sample averages that can be drawn as a histogram to give us a good idea of the true distribution of the various sample means.
  - `set x 1 ssize = %uniform[1,1]` uses the RATS random number generator to get draws from the required uniform distribution.

RATS Program to Illustrate the Central Limit Theorem

```
allocate 10000

*** TAKING AVERAGES FROM A UNIFORM DISTRIBUTION

do ssize=2,10

do k=1,10000

set x 1 ssize = (ssize**0.5)*%uniform(-1,1)

stats(noprint) x 1 ssize
comp xmean(k) = %mean

end do k

@histogram(distrib=normal, stats) xmean 1 10000

end do ssize
```

Multivariate Version of the CLT

- These examples described what happened when taking averages of a single variable. Sometimes, we will want to use a vector that describes multiple averages of different variables.
- In this case, the Central Limit Theorem easily generalises.
- **Multivariate Central Limit Theorem**: Consider a sequence of $n \times k$ matrices of random variables, y_n $(n = 1, 2, 3...)$ each with k different columns each made of up n i.i.d. observations with finite variances and covariances. Then the sequence of $1 \times k$ vectors of sample means $\bar{y}_n$ has the property that

$$\sqrt{n} \left(\bar{y}_n - \mu \right) \xrightarrow{d} N \left(0, V \right). \tag{12}$$

where

$$\mu = \mathbb{E}\, y$$

and

$$V = \mathbb{E} \left(\left(y - \mu \right) \left(y - \mu \right)' \right)$$

Useful Results for Convergence in Distribution

Continuous Mapping Theorem: Let $\{X_n\}$ denote a sequence of vectors of k different estimators based on a sample of size n such that $X_n \xrightarrow{d} z$ (i.e. each of the k estimators converges in probability to the corresponding slots in the c vector) and let $g : \mathbb{R}^k \rightarrow \mathbb{R}^m$ be a continuous function at c that does not depend upon the sample size n. Then $g(X_n) \xrightarrow{d} g(z)$.

Slutsky's Theorem: This tells us we can mix together variables that converge in distribution with other variables that converge in probability to a particular value. If $z_n \xrightarrow{d} z$ and $c_n \xrightarrow{p} c$ then

1. $z_n + c_n \xrightarrow{d} z + c$
2. $c_n z_n \xrightarrow{d} cz$
3. $\frac{z_n}{c_n} \xrightarrow{d} \frac{z}{c}$

Test Procedures Based on the CLT

- So you find out that for the i.i.d. sequence y_i with mean μ and finite variance σ^2
$$\sqrt{n}\left(\frac{\bar{y}_n-\mu}{\sigma}\right)\xrightarrow{d}N\left(0,1\right). \tag{13}$$
- Perhaps you might imagine using this as a test for figuring out the value for μ. However, if you don't know μ, then presumably you don't know σ^2 either. It turns out though, that you can use any consistent estimate of σ^2 instead.
- The WLLN and Continous Mapping Theorem for Plims tells us that sample second moment is a consistent estimator:
$$\hat{\sigma}_n=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}_n\right)^2}\xrightarrow{p}\sigma\Rightarrow\frac{\sigma}{\hat{\sigma}}\xrightarrow{p}1 \tag{14}$$
- The results from the last slide now tell us that the standardised mean constructed using the sample standard deviation also converges to the standard normal distribution:
$$\sqrt{n}\left(\frac{\bar{y}_n-\mu}{\hat{\sigma}}\right)=\sqrt{n}\left(\frac{\bar{y}_n-\mu}{\sigma}\right)\frac{\sigma}{\hat{\sigma}}\xrightarrow{d}N\left(0,1\right). \tag{15}$$

The Delta Method

- Consider the case where we know that an estimator θ_n behaves such that

$$\sqrt{n}\left(\theta_n - \theta\right) \xrightarrow{d} N\left(0, \sigma^2\right). \tag{16}$$

- Now suppose we want to know the asymptotic distribution of a nonlinear transformation $g\left(\theta_n\right)$. What can we tell about its asymptotic distribution?
- The mean value theorem from calculus tells us that there exists a value θ_n^* between θ_n and θ such that

$$g\left(\theta_n\right) = g\left(\theta\right) + g'\left(\theta_n^*\right)\left(\theta_n - \theta\right) \tag{17}$$

- This means that

$$\sqrt{n}\left(g\left(\theta_n\right) - g\left(\theta\right)\right) = g'\left(\theta_n^*\right)\sqrt{n}\left(\theta_n - \theta\right) \tag{18}$$

- We know the asymptotic distribution of $\sqrt{n}\left(\theta_n - \theta\right)$ and $g'\left(\theta_n^*\right) \xrightarrow{p} g'\left(\theta\right)$. This means

$$\sqrt{n}\left(g\left(\theta_n\right) - g\left(\theta\right)\right) \xrightarrow{d} \left(g'\left(\theta\right)\right)\left(\sqrt{n}\left(\theta_n - \theta\right)\right) \sim N\left(0, \sigma^2\left(g'\left(\theta\right)\right)^2\right) \tag{19}$$

Example of the Delta Method

- What is the asymptotic distribution of the log of the mean of a random sample of i.i.d. observations?
- $g(x) = \log x \Rightarrow g'(x) = \frac{1}{x}$.
- If we take i.i.d. observations $\{y_i\}$ from a sample of size n and calculate the average $\bar{y}_n$, we know that

$$\sqrt{n}\left(\bar{y}_n - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right). \tag{20}$$

- From the Delta method, we can calculate that

$$\sqrt{n}\left(\log \bar{y}_n - \log \mu\right) \xrightarrow{d} N\left(0, \left(\frac{\sigma}{\mu}\right)^2\right). \tag{21}$$

- This is because in this case $g'(\theta)^2 = g'(\mu)^2 = \left(\frac{1}{\mu}\right)^2$

Multivariate Delta Method

- Consider the case where we know that a sequence of $1 \times k$ vector θ_n of estimators behaves such that

$$\sqrt{n}\left(\theta_n - \theta\right) \xrightarrow{d} N\left(0, V\right). \tag{22}$$

where

$$V = \mathbb{E}\left(\left(y - \mu\right)\left(y - \mu\right)'\right) \tag{23}$$

- Now suppose we want to know the asymptotic distribution of some nonlinear transformation $g\left(\theta_n\right)$ where $g : \mathbb{R}^k \to \mathbb{R}^m$ is a differentiable function. Then

$$\sqrt{n}\left(g\left(\theta_n\right) - g\left(\theta\right)\right) \xrightarrow{d} N\left(0, G'VG\right) \tag{24}$$

where G is a matrix of derivatives of g evaluated at θ:

$$G = \frac{\partial}{\partial\theta} g\left(\theta\right)' \tag{25}$$

Example of the Delta Method

- Consider the case where we know that

$$\sqrt{n}\left((\theta_{1n},\theta_{2n})-(\theta_1,\theta_2)\right)\xrightarrow{d}N\left(0,\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}\right). \tag{26}$$

- We can test hypotheses about the ratio $\frac{\theta_1}{\theta_2}$ because the Delta method tells us that

$$\sqrt{n}\left(\frac{\theta_{1n}}{\theta_{2n}}-\frac{\theta_1}{\theta_2}\right)\xrightarrow{d}N(0,\omega) \tag{27}$$

where

$$\omega = \begin{pmatrix}\frac{1}{\theta_2} & \frac{-\theta_1}{\theta_2^2}\end{pmatrix}\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}\begin{pmatrix}\frac{1}{\theta_2}\\ \frac{-\theta_1}{\theta_2^2}\end{pmatrix} \tag{28}$$

$$= \frac{\sigma_1^2}{\theta_2^2}+\left(\frac{\theta_1}{\theta_2^2}\right)^2\sigma_2^2 \tag{29}$$