14.30 Introduction to Statistical Methods in Economics
Spring 2009

14.30 Introduction to Statistical Methods in Economics
Lecture Notes 17

Konrad Menzel

April 16, 2009

1 The Central Limit Theorem

Remember that last week, we saw the DeMoivre-Laplace theorem for Binomial random variables, which essentially said that for large values of n, the *standardization* of the random variable $Y \sim B(n, p)$, $Z = \frac{Y - \mathbb{E}[Y]}{\sqrt{n \text{Var}(Y)}}$ follows approximately a standard normal distribution. Since a binomial is a sum of i.i.d. zero/one random variables X_i (counting the number of "trials" resulting in a "success"), we can think of $\frac{Y}{n}$ as the sample mean of $X_1, \dots, X_n$.
Therefore the DeMoivre-Laplace theorem is in fact also a result on the standardized mean of i.i.d. zero/one random variables. The Central Limit Theorem generalizes this to sample means of i.i.d. sequences from any other distribution with finite variance.

Theorem 1 ***(Central Limit Theorem)*** *Suppose $X_1, \dots, X_n$ is a random sample of size n from a given distribution with mean μ and variance $\sigma^2 < \infty$. Then for any fixed number x,*

$$\lim_{n \to \infty} P\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq x\right) = \Phi(x)$$

We say that $\sqrt{n}\bar{X}_n$ converges in distribution *(some people also say "converges in law") to a normal with mean μ and variance σ^2, or in symbols:*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

So how does the mean converge both to a constant μ (according to the Law of Large Numbers), and a random variable with variance one (according to the central limit theorem) at the same time? The crucial detail here is that for the central limit theorem, we blow the sample mean up by a factor of $\sqrt{n}$ which turns out to be exactly the right rate to keep the distribution from collapsing to a point (which happens for the Law of Large Numbers) or exploding to infinity.

Why is the standard normal distribution a plausible candidate for a limiting distribution of a sample mean to start with? Remember that we argued that the sum of two independent normal random variables again follows a normal distribution (though with a different variance, but since we only look at the *standardized* mean, this doesn't matter), i.e. the normal family of distributions is stable with respect to convolution (i.e. addition of independent random variables from the family). Note that this is not true for most other distributions (e.g. the uniform or the exponential).
Since the sample mean is a weighted sum of the individual observations, increasing the sample from n to $2n$, say, amounts to adding the mean of the sequence $X_{n+1}, \dots, X_{2n}$ to the first mean, and then dividing by 2. Therefore, if we postulated that even for large n, the distribution of $\bar{X}_n$ was not such that the sum

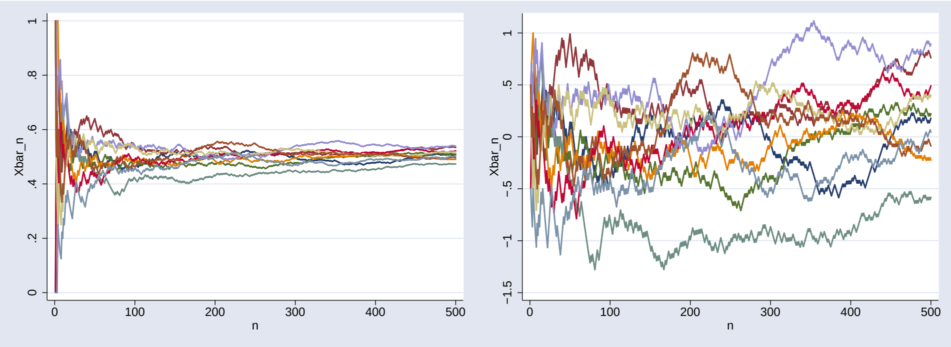


Figure 1: Number of heads in n coin tosses: sample mean $\bar{X}_n$ (left) and standardized sample mean $\sqrt{n}\bar{X}_n$ (right)

of two independent draws was in the same family of distributions, the distribution of the sample mean would still change a lot for arbitrarily large values of n, and can therefore not lead to a stable limit. This should motivate why it is plausible that the distribution of the mean approaches the normal distribution in the limit.

Example 1 *Suppose $X_1, \ldots, X_n$ are i.i.d. random variables where $X_i \sim U[0,1]$ is uniform, so the p.d.f. is*

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

We can now use the convolution formula from lecture 10 to compute the p.d.f. for the partial sums

$$S_k = X_1 + X_2 + \ldots + S_k$$

For $k = 2$, we get (need to be careful about integration limits)

$$\begin{aligned} f_{S_2}(s_2) &= \int_{-\infty}^{\infty} f_X(s_2 - w) f_X(w) dw = \int_{\max\{s_2-1,0\}}^{\min\{s_2,1\}} 1 dw \\ &= \min\{s_2, 1\} - \max\{s_2 - 1, 0\} = \begin{cases} s_2 & \text{if } 0 \leq s_2 \leq 1 \\ 2 - s_2 & \text{if } 1 \leq s_2 \leq 2 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Now, the next calculations become more tedious because we always have to keep track of the integration limits and the kink points in the density. After some calculations, I get for $k = 3$

$$f_{S_3}(s_3) = \int_{-\infty}^{\infty} f_{S_2}(s_3 - w) f_X(w) dw = \begin{cases} \frac{s_3^2}{2} & \text{if } 0 \leq s_3 \leq 1 \\ -\frac{3}{2} + 3s_3 - s_3^2 & \text{if } 1 \leq s_3 \leq 2 \\ \frac{9}{2} - 3s_3 + \frac{1}{2}s_3^2 & \text{if } 2 \leq s_3 \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

By the rule on expectations of sums of random variables,

$$\mathbb{E}[S_k] = k\mathbb{E}[X_1] = \frac{k}{2}$$

Also, since $X_1, X_2, \ldots, X_k$ are independent, we can use the rule on the variance of the sum

$$\begin{aligned}\text{Var}(S_k) &= \text{Var}(X_1 + X_2 + \ldots + X_k) = \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_k) \\ &= k\text{Var}(X_1) = k\int_0^1 \left(t - \frac{1}{2}\right)^2 dt = k\left(\frac{1}{3} - \frac{1}{2} + \frac{1}{4}\right) = \frac{k}{12}\end{aligned}$$

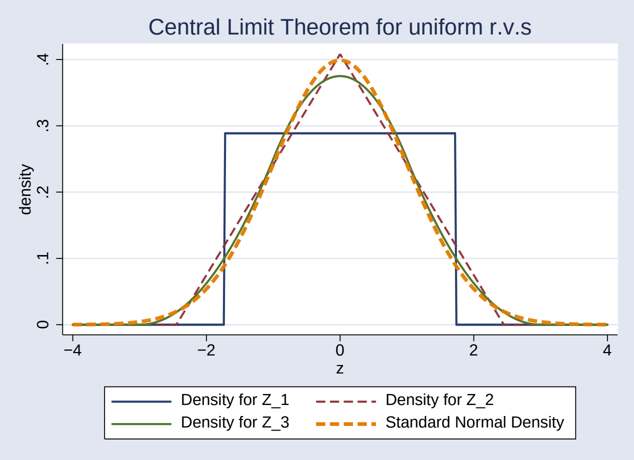
Therefore, the standardization Z_k of S_K is given by

$$Z_k = \frac{S_k - \frac{k}{2}}{\sqrt{\frac{k}{12}}}$$

We can therefore calculate the densities of the standardizations Z_1, Z_2, Z_3 using the change of variables formula (notice that the derivative is just equal to $\sqrt{\frac{12}{k}}$)

$$\begin{aligned}f_{Z_1}(z) &= \begin{cases} \frac{1}{\sqrt{12}} & \text{if } -\sqrt{3} \leq z \leq \sqrt{3} \\ 0 & \text{otherwise} \end{cases} \\ f_{Z_2}(z) &= \begin{cases} \frac{1}{\sqrt{6}} + \frac{z}{6} & \text{if } -\sqrt{6} \leq z \leq 0 \\ \frac{1}{\sqrt{6}} - \frac{z}{6} & \text{if } 0 \leq z \leq \sqrt{6} \\ 0 & \text{otherwise} \end{cases} \\ f_{Z_3}(z) &= \begin{cases} \frac{1}{4}\left(\frac{3}{2} + \frac{z}{2}\right)^2 & \text{if } -3 \leq z \leq -1 \\ \frac{3}{8} - \frac{z^2}{8} & \text{if } -1 \leq z \leq 1 \\ \frac{1}{2}\left(\frac{9}{8} - \frac{3}{4}z + \frac{z^2}{8}\right) & \text{if } 1 \leq z \leq 3 \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

Now let's check how this looks graphically:



The p.d.f. for standardized sums of uniform random variables looks very similar to the standard normal p.d.f. for a sum over as few as 3 independent draws - which is quite surprising since the uniform density itself doesn't look at all like that of a normal random variable.

While the last example is a little deceptive in that the normal approximation looks quite good for n as small as 3 (at least optically), with $n \to \infty$, we usually mean $n \approx 40$ or larger for the approximation to be reasonably accurate.

Summarizing, the Central Limit Theorem is particularly useful when we don't want to compute the true p.d.f. of the sample mean. There are two situations in which this happens

- We *can't* compute the actual p.d.f. because we don't know the exact distribution of the X_i's
- we *don't want to* compute the actual p.d.f. because the computations are too tedious - which is almost invariably true for the general convolution formula (see last example), but also for many discrete examples (see Binomial example from last lecture).

2 Estimation

So far in this class, we started by assuming that we knew the parameters of the distribution of a random variable - e.g. we knew that $X \sim P(\lambda)$ - and then calculated probabilities and other properties of random samples from that distribution. Now we are going to look at the reverse problem:
Assume that we have an i.i.d. sample of observations from a distribution with unknown parameters θ, how do we get a "reasonable" answer which value of θ in the family of distributions we are looking at may have generated the data.

Example 2 *If for a given coin we don't know the probability for heads in a single toss, we could toss it many times. Then we'd think that the fraction of heads, $\hat{p} = \frac{\sharp \text{ Heads}}{\sharp \text{ Tosses}}$ may be a "good guess" for the probability $P(\text{Heads})$ in a sense to be defined later.*

A *parameter* is a constant indexing a family of distributions given by the p.d.f.s $f(x|\theta)$, where we denote parameters generally as $\theta_1, \ldots, \theta_k$.

Example 3

- *for the binomial distribution,*

$$f_X(x|n,p) = \begin{cases} \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

the parameters are the number of trials n and the success rate p.

- *for the normal distribution,*

$$f_X(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

so that parameters are mean μ and standard deviation σ

- *the Poisson distribution has one parameter λ*

$$f_X(x|\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{for } x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Much of statistics is concerned with determining which member of a known family of distributions gives the correct probability distribution of an observed process or phenomenon. In symbols, we want to find the parameter value θ_0 such that $X \sim f(x|\theta_0)$. This is the problem of "estimating the parameters which characterize the distribution."

We'll always start off a random sample $X_1, \ldots, X_n$, and we'll always assume that

$$X \sim f(x|\theta_0) \text{ for unknown } \theta_0 \in \Theta$$

Definition 1 *An* estimator $\hat{\theta}$ *of* θ *is a statistic (i.e. a function of* $X_1, \dots, X_n$*),*

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$$

A *realization* $\hat{\theta}(x_1, \dots, x_n)$ of the estimator in a sample is called an *estimate* of θ.

Notice that, as a function of the random sample, the estimator is a proper random variable, so we will in general be interested in describing its distribution in terms of the p.d.f., and moments of its distribution.

Example 4 *Suppose,* $X \sim \text{Bernoulli}(\theta_0)$*, i.e.* X *is a zero/one random variable which takes the value 1 with probability* θ *and has p.d.f.*

$$f_X(x) = \begin{cases} \theta_0 & \text{if } x = 1 \\ 1 - \theta_0 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

How do we estimate θ_0*?*
Could use sample mean

$$\hat{\theta}(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

e.g. from 5 Bernoulli trials $1, 0, 0, 1, 1$

$$\hat{\theta}(1, 0, 0, 1, 1) = \frac{3}{5}$$

Since the estimator $\hat{\theta}$ *for a sample of 5 observations is a random variable, we can derive its p.d.f.: recall that for* $S_5 \equiv \left(\sum_{i=1}^{5} X_i\right)$*,* $S_5 \sim B(5, \theta_0)$*. Applying the methods for finding p.d.f.s of functions of discrete random variables to* $\hat{\theta} = \frac{S_5}{5}$*, we get*

$$f_{\hat{\theta}}(t) = \begin{cases} \begin{pmatrix} 5 \\ 5t \end{pmatrix} \theta_0^{5t} (1 - \theta_0)^{5(1-t)} & \text{if } t \in \left\{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1\right\} \\ 0 & \text{otherwise} \end{cases}$$

In particular, the distribution of the estimator depends on the true probability θ_0 *- which can be anywhere in the interval [0,1] - but can only take 6 different discrete values.*

Example 5 *If* $X \sim U[0, \theta]$ *is uniform over an interval depending on the parameter, the p.d.f. is*

$$f_X(x) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

How could we estimate θ*? Could use e.g.*

$$\begin{aligned} \hat{\theta}_1 &= \max\{X_1, \dots, X_n\} \\ \hat{\theta}_2 &= 2\bar{X}_n \end{aligned}$$

Say, we sampled three observations from the distribution, $0.2, 0.6, 0.4$*. Then* $\hat{\theta}_1 = 0.6$ *and* $\hat{\theta}_2 = 0.8$*, so the two estimators give different answers on the same parameter. How should we choose among those different estimators? - We'll get back to this in a moment.*

- How do you come up with these functions $\hat{\theta}(X_1, \dots, X_n)$?
- How can we determine whether these estimators are reasonable?
- How should we choose between two or more estimators for the same parameter?

3 General Properties of Estimators

We will denote the expectation of X *under the parameter* θ - i.e. the expectation of X if the true parameter is equal to θ - by

$$\mathbb{E}_\theta[X] = \int_{-\infty}^{\infty} x f_X(x|\theta)dx$$

Similarly, I'll write the variance under the parameter θ as

$$\mathrm{Var}_\theta(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}_\theta[X])^2 f_X(x|\theta)dx$$

The *bias* of an estimator is the difference between its expectation and the true parameter,

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}_{\theta_0}[\hat{\theta}] - \theta_0$$

Of course, we'd like an estimator to get the parameter right on average, so that ideally, the bias should be zero.

Definition 2 *An estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ is* unbiased *for θ if*

$$\mathbb{E}_{\theta_0}[\hat{\theta}] = \theta_0$$

for all values of θ_0.

Example 6 *Suppose $X_1, \dots, X_n$ is an i.i.d. sample from a $N(\mu, \sigma^2)$ distribution. We already saw last week that the expectation of the sample mean*

$$\mathbb{E}_{\mu,\sigma^2}[\bar{X}_n] = \mathbb{E}_{\mu,\sigma^2}[X] = \mu$$

for any value of μ, so that $\bar{X}_n$ is an unbiased estimator for the mean μ of a normal distribution.

Example 7 *Suppose we want to estimate the variance parameter σ^2 for $X \sim N(\mu, \sigma^2)$ with unknown mean μ from an i.i.d. random sample $X_1, \dots, X_n$. Since $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$, an intuitively appealing estimator would be*

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

(where we substituted the sample mean for the actual expectation). What is the expectation of this estimator if the true parameters of the distribution are (μ_0, σ_0^2)?
Recall that $\mathbb{E}[X^2] = \mathbb{E}[X]^2 + \mathrm{Var}(X)$, so that

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2 - \bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mu^2 + \sigma^2) - \left(\mu^2 + \frac{1}{n}\sigma^2\right) \\
&= \mu^2 + \sigma^2 - \mu^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2
\end{aligned}$$

Therefore $\hat{\sigma}^2$ is not *an unbiased estimator for σ^2, but we can easily construct an unbiased estimator $\tilde{\sigma}^2$*

$$\tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Where does this bias come from? Broadly speaking, the reason is that inside the square we are replacing μ with a (noisy) estimate $\hat{\mu} = \bar{X}_n$. You can check on your own that if μ_0 was known, the estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2$ would be unbiased for σ.
Having to estimate the mean uses up one "degree of freedom" in the data - e.g. if we only had a sample of one observation, the estimated mean would be equal to that observation, and the "naive" estimator of the variance would give us $\hat{\sigma}^2 = 0$, which is clearly not the right answer.

Unbiasedness may not be the only thing we care about, since the estimator being equal to the true parameter *on average* doesn't mean that in for a given sample, the *estimate* is actually going to be close to the true parameter.

Definition 3 *For a sample $X_1, \dots, X_n$, we say that $\hat{\theta}$ is a* consistent *estimator for θ if as we increase n, the estimator* converges in probability *to θ_0, i.e. for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} P_{\theta_0}\left(|\hat{\theta}(X_1, \dots, X_n) - \theta_0| < \varepsilon\right) = 1$$

for all values of θ_0.

In words, in a sufficiently large sample, a consistent estimator will be within a small distance from the true parameter with high probability. Notice that unbiasedness and consistency are two very different concepts which overlap, but neither implies the other:

Example 8 *Back to one of our estimators for the uniform distribution, $X \sim U[0, \theta_0]$. If we look at*

$$\hat{\theta}_1 = \max\{X_1, \dots, X_n\}$$

we can easily see that $\hat{\theta}_1$ is not unbiased *for θ, because due to the nature of the uniform distribution, all possible values of X_i are less than θ_0. Therefore, no matter how large n is, $P(\max\{X_1, \dots, X_n\} < \theta_0) = 1$. Therefore the expectation $\mathbb{E}_{\theta_0}[\hat{\theta}_1] < \theta_0$. However, $\hat{\theta}_1$ is* consistent *for θ_0: We can easily see that for a single observation X from the uniform, the c.d.f. is $F_X(x) = \frac{x}{\theta_0}$. Since $Y_n := \max\{X_1, \dots, X_n\}$ is the nth order statistic of the sample, we get from our previous discussion that for $0 \le y \le 1$, $F_{Y_n}(y) = (F_X(y))^n = \left(\frac{y}{\theta_0}\right)^n$. Since $\hat{\theta}_1 < \theta_0$ with probability 1, we can therefore calculate for any sample size n and any $\varepsilon > 0$*

$$P(|\hat{\theta}_1 - \theta_0| > \varepsilon) = P(Y_n < \theta_0 - \varepsilon) = \left(\frac{\theta_0 - \varepsilon}{\theta_0}\right)^n \equiv p^n$$

where $p := \frac{\theta_0 - \varepsilon}{\theta_0} < 1$ since $\varepsilon > 0$. Therefore, the probability of a deviation from θ_0 by more than ε vanishes as we increase n, and $\hat{\theta}_1$ is therefore consistent.

Example 9 *By the Law of Large Numbers, the sample mean converges in probability to $\mathbb{E}[X] = \mu$. Therefore, for an i.i.d. sample $X_1, \dots, X_n$ of $N(\mu, \sigma^2)$ random variables, the sample mean is a* consistent *estimator for μ.*
Alternatively, let's look at an "unreasonable" estimator $\tilde{\theta}(X_1, \dots, X_n) = X_n$. Then

$$\mathbb{E}[\tilde{\theta}(X_1, \dots, X_n)] = \mathbb{E}[X_n] = \mu$$

so this estimator is unbiased. However, for any sample size n, the distribution of the estimator is the same as that for $X_i \sim N(\mu, \sigma)$, so e.g. for $\varepsilon = \sigma_0$, the probability

$$\begin{aligned} P(|\tilde{\theta}(X_1, \ldots, X_n) - \mu_0| < \sigma_0) &= P(\mu_0 - \sigma_0 \leq X_n \leq \mu_0 + \sigma_0) \\ &= P\left(-1 < \frac{X_n - \mu_0}{\sigma_0} < 1\right) = P(-1 < Z < 1) \\ &= \Phi(1) - \Phi(-1) \approx 0.6825 << 1 \end{aligned}$$

for all n, where the standardization $Z := \frac{X_n - \mu_0}{\sigma_0}$ follows a $N(0,1)$ distribution. By this argument, $\tilde{\theta}$ is unbiased, but not *consistent.*