Economics 142

Spring 2019

Today's agenda

1. quick overview of class

   - course requirements: problem sets; midterm; course project

   - course content

2. quick refresher on some basic stats

3. confidence intervals; minimum sample sizes

4. prep. for PS#1

Refresher on statistics (Appendix B, C of Wooldridge)

$Y_1, Y_2, ... Y_n$ random sample from a pop, mean μ, variance σ^2

$\overline{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$ is the sample mean; a "statistic" (a function of the sample that has no unknown parameters)

$s_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2$ is the sample variance (note $n-1$)

$E[\overline{Y}_n] = \mu$: the sample mean is "unbiased" for the pop. mean

$Var[\overline{Y}_n] = Var[\sum_{i=1}^{n} (Y_i/n)] = \sigma^2/n$.

$E[s_n^2] = \sigma^2$: the "d.f. corrected" sample var is unbiased for σ^2

convergence in probability: $Z_1, Z_2, ..$ is a sequence of r.v.'s *converges in probability* to b if for any $\varepsilon > 0$:

$$lim_{n \to \infty} P(|Z_n - b| < \varepsilon) = 1$$

Write as $plim\, Z_n = b$.

Three famous results: Markov inequality; Chebyshev inequality; WLLN

1. *Markov*: if X is r.v., with $P(X > 0) = 1$ then for any $t > 0$:

$$P(X \geq t) \leq \frac{E[X]}{t}.$$

Proof: $E[X] = \int_0^\infty xf(x)dx = \int_0^t xf(x)dx + \int_t^\infty xf(x)dx$

$\Rightarrow E[X] \geq \int_t^\infty xf(x)dx \geq t \int_t^\infty f(x)dx = tP(X \geq t)$.

Chebychev: If X is a random variable s.t. $Var[X]$ exists, then for any $t > 0$:

$$P(|X - E[X]| \geq t) \leq \frac{Var[X]}{t^2}.$$

Proof: consider r.v. $Y = (X - E[X])^2$. Note $E[Y] = Var[X]$. Using Markov $P(Y \geq \tau) \leq \frac{E[Y]}{\tau}$.

So, letting $\tau = t^2$,

$$P(Y \geq t^2) = P(|X - E[X]| \geq t) \leq \frac{E[Y]}{t^2} = \frac{Var[X]}{t^2}$$

WLLN. Suppose that $Y_1, Y_2, ...Y_n$ is a random sample from a pop with mean μ, variance σ^2(both finite). Then $plim \overline{Y}_n = \mu$.

Proof. Pick $\varepsilon > 0$. Applying Chebychev to $\overline{Y}_n$:

$$P(|\overline{Y}_n - \mu| \geq \varepsilon) \leq \frac{Var[\overline{Y}_n]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

$$\Rightarrow P(|\overline{Y}_n - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

So

$$lim_{n\to\infty} P(|\overline{Y}_n - \mu| < \varepsilon) = 1.$$

WLLN says that the distribution of the sample mean "collapses" to the point μ as the sample size gets bigger, (i.e., $plim(\overline{Y}_n - \mu) = 0$).

WLLN says $plim(\overline{Y}_n - \mu) = 0$. The *Central Limit Theorem (CLT)* says that the distribution of $\overline{Y}_n$ collapses to a *normal* at the rate $n^{1/2}$: if we consider the "scaled" r.v. $\sqrt{n}(\overline{Y}_n - \mu)/\sigma$, this has a normal distribution $N(0,1)$ as $n \to \infty$. A key idea of statistics is that for a given n we can step back from the limit and still be "approximately" OK.

$\{Z_n\}$, a sequence of r.v.'s, *converges in distribution to a r.v.* Z *with c.d.f.* $F(x)$ if

$$lim_{n \to \infty} P(Z_n \leq x) = F(x).$$

CLT: Let $Y_1, Y_2, ...Y_n$ be a random sample from a population with mean μ, variance σ^2. Then the "scaled" r.v. $\sqrt{n}(\overline{Y}_n - \mu)/\sigma$ converges in distribution to $N(0,1)$. That is, for any fixed x:

$$lim_{n\to\infty} P(\frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma} \leq x) = \Phi(x),$$

where $\Phi()$ is the standard normal c.d.f. This is often written as

$$\sqrt{n}(\overline{Y}_n - \mu)/\sigma \approx N(0,1)$$

$$\Rightarrow \overline{Y}_n \approx N(\mu, \sigma^2/n)$$

$$\overline{Y}_n \approx N(\mu, \sigma^2/n)$$

In fact, CLT remains true if, instead of scaling by σ, we scale by s_n (the *estimate* of σ):

$$\sqrt{n}(\overline{Y}_n - \mu)/s_n \approx N(0,1).$$

$$\Rightarrow \overline{Y}_n \approx N(\mu, s_n^2/n)$$

The quantity s_n^2/n is often called the "estimated sampling variance" of the mean, and $s_n/\sqrt{n}$ is often called the "estimated sampling error".

Sampling from a normal distribution.

CLT says that the sample mean is "asymptotically normal", regardless of the underlying distribution that the $Y's$ are drawn from (as long as μ and σ^2 are finite). Suppose that each Y_i is a draw from $N(\mu, \sigma^2)$. In this case,

$$\overline{Y}_n = \sum_{i=1}^{n} (Y_i/n).$$

Now, we know that if X and Z are independently distributed $X \sim N(\mu_x, \sigma_x^2)$ and $Z \sim N(\mu_z, \sigma_z^2)$ then

$$aX + bZ \sim N(a\mu_x + b\mu_z, a^2\sigma_x^2 + b^2\sigma_z^2)$$

Extending this result $\overline{Y}_n \sim N(\mu, \sigma^2/n)$ or $\sqrt{n}(\overline{Y}_n - \mu)/\sigma \sim N(0, 1)$. In this case, the distribution of $\overline{Y}_n$ is exact.

Aside on sampling from a normal distribution, continued....

Also, the distribution when we use s_n instead of σ (which is unknown) to scale is known to be a so-called "t-distribution":

$$\sqrt{n}(\overline{Y}_n - \mu)/s_n \sim t_{n-1}$$

where $t_{n-1}$is the t-distribution with $n-1$ degrees of freedom. For large n the t is very close to the standard normal. For smaller n the t distribution has fatter tails.

Confidence intervals.

Suppose $Z \sim N(0, 1)$. Then we know Z is symmetrically distributed around 0 with a "bell curve" distribution. Define $z_p > 0$ as the real number such that $\Phi(z_p) = 1 - p$ (for $p < .5$). This is the point such that $P(Z \geq z_p) = p$. We ask: what is the symmetric interval (around 0) such that a standard normal falls in the interval with probability $1 - \alpha$? This is the interval $(-z_{\alpha/2}, z_{\alpha/2})$. Why? Because the probability of falling *above* $z_{\alpha/2}$ is $\alpha/2$, and by symmetry the probability of falling *below* $-z_{\alpha/2}$ is also $\alpha/2$. So the probability of being outside the interval is α.

For $Z \sim N(0,1)$ $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$. Suppose we have obtained a random sample of some $Y's$ and formed the estimated mean and standard deviation. By the CLT $\sqrt{n}(\overline{Y}_n - \mu)/s_n \approx N(0,1)$, so (approximately):

$$P(-z_{\alpha/2} \leq \sqrt{n}(\overline{Y}_n - \mu)/s_n \leq z_{\alpha/2}) = 1 - \alpha$$

$$\Rightarrow P(-\frac{s_n z_{\alpha/2}}{\sqrt{n}} \leq \overline{Y}_n - \mu \leq \frac{s_n z_{\alpha/2}}{\sqrt{n}}) = 1 - \alpha$$

$$\Rightarrow P(-\overline{Y}_n - \frac{s_n z_{\alpha/2}}{\sqrt{n}} \leq -\mu \leq -\overline{Y}_n + \frac{s_n z_{\alpha/2}}{\sqrt{n}}) = 1 - \alpha$$

$$\Rightarrow P(\overline{Y}_n - \frac{s_n z_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \overline{Y}_n + \frac{s_n z_{\alpha/2}}{\sqrt{n}}) = 1 - \alpha$$

$$P(\overline{Y}_n - \frac{s_n z_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \overline{Y}_n + \frac{s_n z_{\alpha/2}}{\sqrt{n}}) = 1 - \alpha$$

This is interpreted as: if we kept repeating a sample of size n, $1-\alpha$ percent of the time the interval $\overline{Y}_n \pm \frac{s_n z_{\alpha/2}}{\sqrt{n}}$ would "capture" (or cover) the true mean μ. This is called the $(1-\alpha)$ "confidence interval".

Note that for $\alpha = 0.05$, $z_{\alpha/2} = z_{.025} = 1.96 \approx 2$. So we can say that

$$P(\overline{Y}_n - 2\frac{s_n}{\sqrt{n}} \leq \mu \leq \overline{Y}_n + 2\frac{s_n}{\sqrt{n}}) \approx 0.95$$

The term $\frac{s_n}{\sqrt{n}}$ is the "estimated sampling error" of $\overline{Y}_n$. So if we repeated sampling, 95% of the time the $\overline{Y}_n \pm 2\frac{s_n}{\sqrt{n}}$ confidence interval would contain the true mean.

Using these ideas.

Suppose we have to draw a sample of a binary random variable (e.g., the fraction of people who vote Democrat; or the fraction of mortgages in a "mortgage-backed security" portfolio that were improperly underwritten). Let p represent the true probability of the event of interest being "true": so Y_i is a Bernoulli r.v. with mean p and variance $p(1-p)$. For a sample of size n we estimate the mean of the $Y's$, which is the fraction of "1's" we get. For simplicity call this $\overline{p}_n$. Note that $E[\overline{p}_n] = p$. Also, in this case our estimate of the variance term is

$$s_n^2 = \frac{1}{n-1}\sum_i (Y_i - \overline{p}_n)^2 = \frac{1}{n-1}\sum (Y_i^2 - 2\overline{p}_n Y_i + \overline{p}_n^2) = \frac{n}{n-1}\overline{p}_n(1-\overline{p}_n).$$

In the Bernoulli case, people often divide the sum by n so the estimate is $s_n = \sqrt{p_n(1-p_n)}$.

How big a sample do we need?

"Margin of Error" – one way people decide the sample size is to set a "margin of error" with a given level of confidence. The margin of error is 1/2 of the width of a $(1-\alpha)$ confidence interval. For a 95% confidence interval (the "industry standard"), $z_{\alpha/2} = 1.96$. Thus the width of half of the CI is

$$m = \frac{s_n z_{\alpha/2}}{\sqrt{n}}.$$

Now we don't know p so we don't know s_n : but a "worst case" is $p_n = 0.5$ which implies that $s_n = \sqrt{0.5(1-0.5)} = 0.5$. So the "worst case" for m with a sample size of n is

$$m = \frac{0.5 z_{\alpha/2}}{\sqrt{n}}.$$

If we choose a margin of error m, we need a sample size of

$$n = \left(\frac{z_{\alpha/2}}{2m}\right)^2$$

A standard setting is $m = 0.05$, which with a 95% confidence needs $n \approx 400$. Note that if we use a 95% confidence level then $z_{\alpha/2} \simeq 2$ and

$$n \simeq \left(\frac{1}{m}\right)^2$$

"Minimum Detectable Effect" – another way to choose a sample size is to ask what deviation from a given value would you like to be able to "reliably detect". If the default "null hypothesis" is $p = p^0$, we might want to be able to say that if we obtain a point estimate of $p_n = p'$ then it will be "significantly different from p^0. Assuming a null of $p = p^0$, we would "reject the null" with an estimate of p' at the α level of significance (under a 2-tailed test) if

$$\frac{p' - p^0}{s_n/\sqrt{n}} > z_{\alpha/2}.$$

Again, using $s_n = 0.5$ as the "worst case" scenario, the sample size we need satisfies:

$$\sqrt{n} > \frac{z_{\alpha/2}}{2(p' - p^0)}.$$

For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$, and we need (approximately):

$$n > \frac{1}{(p' - p^0)^2}.$$

If for example you want $p' - p^0 = 0.05$, you'll need $n = 400$. Notice that this is the same thing as having a 0.05 margin of error.