Stats 135, Fall 2019
Lecture 2, Friday, 8/30/2019

1 Review

Last time, we took the dichotomous case $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, and we have the sample mean:

$$\hat{p} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

as an (unbiased) estimate of the unknown parameter p.
We also found:

$$E(\hat{p}) = E(\overline{x}) = E(x) = p$$

and

$$\sigma_{\hat{p}}^2 = \text{SE of } (\hat{p})^2 = \text{var}(\overline{x}) = \frac{\overbrace{\text{var}(x)}^{p(1-p)}}{n}\left[\frac{N-n}{N-1}\right]$$

There are 2 estimates of SE of $(\hat{p})$: (1) a **conservative** estimate, and (2) a **bootstrap** estimate.

Topics Today:

- §7.3.3 : Confidence Intervals (CI) for $\mu = E(x)$ (or p in the dichotomous case).

  As an example in the dichotomous case, a 68% CI for p is

  $$\hat{p} \pm (\text{SE of } \hat{p})$$

  and a 95% CI for p is

  $$\hat{p} \pm 1.96\,(\text{SE of } \hat{p})\,.$$

  We can approximate SE of $\hat{p}$ to find a conservative or bootstrap confidence interval of p.
- §7.3.1 The expectation and variance of the sample mean.

2 Normal Approximation

First consider the normal approximation to the sampling distribution of $\overline{x}$:

Example: Consider a population of $N = 393$ hospitals. Let $x :=$ the number of patients discharged from the ith hospital, and let:

$$\mu = 814.6$$
$$\sigma = 590,$$

and a SRS (simple random sample) of $n = 50$ is taken. From our box (with μ, σ), we draw:

$$X_1, \ldots, X_{50}$$

and $\hat{\mu} = \overline{x}$. We want to find $P(|\overline{x} - \mu| > 100)$. The picture we have is a distribution centered at $\mu = 814.6$, and we have (from yesterday):

$$\sigma_{\overline{x}} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} = \frac{590}{\sqrt{50}}\sqrt{\frac{393-50}{393-1}} = 77.95$$

What's the chance of being at the tail? Take $\mu \pm \sigma_{\overline{x}}$.

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = \frac{\overbrace{\overline{x}}^{914.6} - 814.6}{77.95} = 1.28,$$

and this gives a standard normal curve, with bounds at -1.28 and 1.28.
The c.d.f. gives the tail end is

$$1 - \Phi(1.28)$$

The standard normal curve has c.d.f.

$$\Phi(z) = P(Z \leq z) = \texttt{pnorm(z)},$$

where in R we can find this using the command `pnorm(z)`.
And we have:

$$P\left(|\overline{x} - \mu| > 100\right) = 2\left(1 - \Phi(1.28)\right) = 2(.1) = .2$$

The question was posed in class what do we mean by a normal approximation? Lucas mentions that this curve is approximately normal, which follows from the Central Limit Theorem. So we just model it by the standard normal model. Given the mean and variance at the beginning, we obtain another SE $\sigma_{\overline{x}}$ which is written in terms of constants in our box. Usually we don't know these constants, and we would have to approximate this.

Example: Dichotomous Case In the hospital example, $n = 50$. Let p be the proportion of hospitals with fewer than 1000 discharges. We assign 0 if the hospital has more than or equal to 1000 discharges, and 1 if the hospital is less than 1000 discharges.
Suppose we know that $p = .65$. Find the tail-end probability:

$$P\left(|\hat{p} - p| > .13\right).$$

We use our finding from yesterday that:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}\left[\frac{N-n}{N-1}\right]} = \frac{.65 \cdot .35}{50}\left[\frac{N-n}{N-1}\right] = 0.063$$

$$P\left(|\hat{p} - p| > .13\right) = 2\left[1 - \Phi\left(\frac{.13}{.063}\right)\right] = 2\left(1 - \texttt{pnorm}(2.06)\right) = 0.039$$

3 Confidence Intervals (CI)

A Confidence Interval for a population parameter θ is a random interal, calculated from the sample that contains θ with some specified probability.
For $0 \leq \alpha \leq 1$, let $z(\alpha)$ be the number such that the area under the standard normal curve to the right of $z(\alpha)$ is α.
Then

$$P\left(-z\left(\frac{\alpha}{2}\right) \leq z < z\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha,$$

just by the definition of $z(\alpha/2)$. For example, if $\alpha = 0.05$, then

$$z\left(\frac{\alpha}{2}\right) = z\left(0.025\right) = \texttt{qnorm}(1 - .025) \approx 1.959964 = 1.96$$

and to find the point (z-value) such that this area is satisfied, we can use `qnorm` in R.
If $z = \frac{\overline{x}-\mu}{\sigma_{\overline{x}}}$, then $\overline{x}$ is approximately normal (so we normalize it). Then we take:

$$\begin{aligned} 1-\alpha &= P\left(\frac{\overline{x}-\mu}{\sigma_{\overline{x}}} \in \left[-z\left(\frac{\alpha}{2}\right), z\left(\frac{\alpha}{2}\right)\right]\right) \\ &= P\left(\overline{x}-\mu \in \left[-z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}}, z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}}\right]\right) \\ &= P\left(\mu-\overline{x} \in \left[-z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}}, z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}}\right]\right), \end{aligned}$$

since the interval is symmetric about zero, and this equals:

$$= P\left(\mu \in \left[\overline{x} - z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}}, \overline{x} + z\left(\frac{\alpha}{2}\right)\right]\right),$$

and we call this a $(1-\alpha)100\%$ confidence interval of μ.

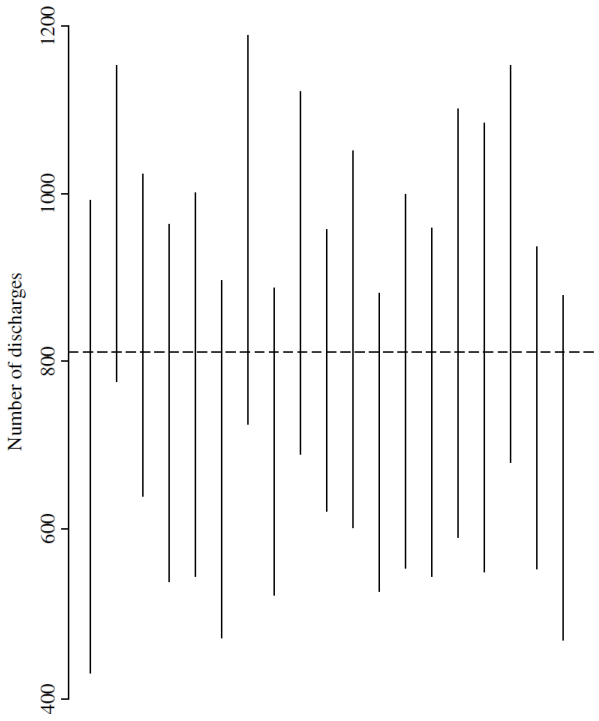


FIGURE **7.4** Vertical lines are 20 approximate 95% confidence intervals for μ. The horizontal line is the true value of μ.

The confidence interval is a variable. There is a 95% (95 out of 100) chance that our interval contains the true parameter μ. The width of all these are the same, but the center is different.

3.1 Example

Consider $N = 393$ hospitals, and let x_i be the number of patients discharged from the ith hospital. Take $\mu = 814.6$ and $\sigma = 590$. Take a simple random

sample (SRS) of $n = 50$. We showed:

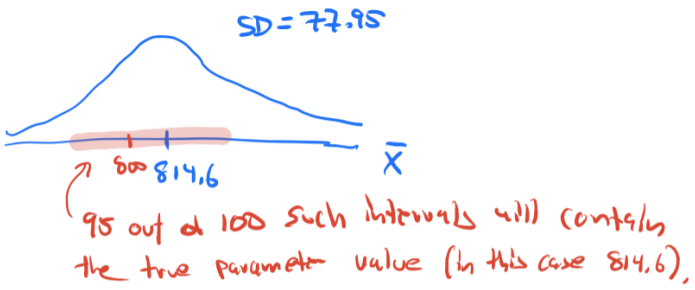
$$\sigma_{\overline{x}} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} = 77.95$$

Suppose $\overline{x} = 800$. Find the 95% CI for μ and interpret the result.

Solution. $\alpha = 0.05$, so take:

$$\overline{x} \pm z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}} = 800 \pm 1.96(77.95) = [647.2, 952.8],$$

and we can draw a picture of our distribution centered around 814.6, and that our confidence interval centered at 800 **contains** 814.6.



To interpret results, we can say that 95 out of 100 such intervals will contain 814.6. □

4 §7.3.1 Expectation, Variance of the Sample Mean

Definition: Unbiased Estimator -

We say that an estimator $\hat{p}$ of p is unbiased if $E(\hat{p}) = p$.

Example: If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, then $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of p since:

$$E(\overline{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot \sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n} \cdot np = p$$

Let $X_1, \ldots, X_n$ be iid with $E(x) = \mu$ and $\text{var}(x) = \sigma^2$. Show that

$$E\left(\overline{x}^2\right) = \frac{\sigma^2}{n} + \mu^2.$$

To see this, notice:

$$\text{var}(y) = E(y^2) - E(y)^2,$$

so

$$E\left(\overline{x}^2\right) = \text{var}(\overline{x}) + E\left(\overline{x}^2\right),$$

and because this is i.i.d., we have:

$$\text{var}(\overline{x}) = \text{var}\left(\frac{1}{n}\sum_{X_i}\right) = \frac{1}{n^2}n \cdot \text{var}(x) = \frac{1}{n}\sigma^2$$

We will use this in the the next lecture to prove a theorem that the sample variance is an unbiased estimator of the true population variance r^2.

Lecture ends here.