Lecture 23, Mar 9, 2023

Point Estimates

- So far we've worked with point estimates in our sampling
- We have IID measurements $X_1, \cdots, X_n$ with realizations $x_1, \cdots, x_n$
- In general we write θ is the true parameter, $\hat{\theta}$ is the observed value, and $\hat{\Theta}$ is the statistic
  - e.g. $\theta = \mu$ is the true parameter (true mean), $\hat{\theta} = \bar{x}$ is the observed value (observed mean), $\hat{\Theta} = \bar{X}$ is the statistic (sample mean RV)
- In general we want to estimate θ from $\hat{\theta}$

Definition

$\hat{\Theta}$ is an *unbiased estimator* if $E[\hat{\Theta}] = \theta$, that is, the expectation of the statistic is the true mean

Out of all unbiased estimators, the most *efficient* estimator has the lowest variance

- e.g. with the mean, $E[X_i] = \mu$ for all i, so any of the individual estimates is an unbiased estimator; however $\bar{X}$ has lower variance (σ^2/n vs. σ^2), so $\bar{X}$ is the most efficient estimator of the sample mean

Interval Estimates

- Instead of estimating an exact value, interval estimates give an interval $\theta_L \leq \theta \leq \theta_U$
  - The most well known example are confidence intervals
  - This gives us a sense of how good our estimate is
- θ_L, θ_U should be the realization of some sampling statistic based on the data

Definition

A *confidence interval* is of the form

$$P(\Theta_L \leq \theta \leq \Theta_U) = 1 - \alpha$$

where Θ_L, Θ_U are statistics

- e.g. a 95% confidence interval has $\alpha = 0.05$
- To calculate confidence intervals of the mean we can use the CLT
  - $Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
  - By the CLT the distribution of Z approaches $n(z; 0, 1)$
  - Recall the CDF is $\displaystyle\int_{-\infty}^{x} n(z; 0, 1)\,\mathrm{d}z$
  - Let $\beta < 0.5$, define z_β such that $\Phi(-z_\beta) = \beta \implies z_\beta = -\Phi^{-1}(\beta)$, that is, the area under the normal PDF above $x = \beta$ is equal to α
    - By symmetry $1 - \Phi(z_\beta) = \beta$
  - $1 - \alpha = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$
    - $z_{\alpha/2}$ has α area above it, and α area below $-z_{\alpha/2}$ by our previous definition
  - $1 - \alpha = P\left(-z_{\alpha/2} \leq \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}\right) = P\left(\bar{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)$
  - This gives us $\Theta_L = \bar{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}, \Theta_U = \bar{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

Definition

Given data $x_1, \cdots, x_n$, let

$$\theta_L = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$\theta_U = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2} = -\Phi^{-1}\left(\frac{\alpha}{2}\right)$, such that $\alpha/2$ of the normal distribution is below $-z_{\alpha/2}$, then $[\theta_L, \theta_U]$ is the *observed* (or realized) confidence interval for the mean for some confidence $1-\alpha$

- Example: $n = 20, \bar{x} = 4$, we know $\sigma = 2$, we want a 95% confidence interval
  - $\alpha = 0.05 \implies z_{\alpha/2} = -\Phi^{-1}(0.025) = 1.96 \approx 2$
  - $0.95 = P\left(\bar{X} - 2\frac{2}{\sqrt{20}} \leq \mu \leq \bar{X} - 2\frac{2}{\sqrt{20}}\right) = P(\bar{X} - 0.88 \leq \mu \leq \bar{X} + 0.88)$
  - Therefore our realized confidence interval is $[4-0.88, 4+0.88] = [3.12, 4.88]$
- This does **not** mean that there is a 95% chance that the true mean falls within $[3.12, 4.88]$ (this statement is not mathematically valid since the true mean is not a random variable)
  - It means that if we did this experiment a large number of times, each time collecting 20 samples, 95% of the time the realization $[\bar{X} - 0.88, \bar{X} + 0.88]$ contains the true mean
  - When a confidence interval is reported as a pair of numbers, as $[3.12, 4.88]$, it is only a particular realization of the confidence interval for that particular experiment