Supplementary Handout for Dis 9: Intro to Estimation

1 Motivation

- Last week, we talked about sampling distributions, which are distributions of some sample statistics of interest.
- Recall that our goal is to perform statistical inference: use sample statistic to draw conclusion on population parameter.
- We are finally going to connect the pieces:
  - The sample statistics of interest from last week are **point estimators**. A point estimator takes a best guess at the true value of an underlying population parameter.
  - Sometimes, one might instead want to estimate a range that's likely to include the true population parameter. The estimator that provides such a range is called an **interval estimator**.
- Sorting through some terminologies:
  - An estimator (point or interval) tries to estimate (a point value or a range of) the corresponding true population parameter.
  - An estimator follows a sampling distribution.
  - A population parameter follows a probability distribution.

2 Point Estimator

- Definition: a point estimator takes a (single) best guess at the true value of an underlying population parameter.
- Examples of point estimator:
  - The sample mean $\bar{X} = \frac{1}{n}\sum_i^n X_i$ used to estimate population mean μ_X.
  - The sample variance $s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})$ used to estimate population variance σ_X^2.
  - The sample covariance $s_{XY} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ used to estimate population covariance σ_{XY}.
- How do you evaluate if an estimator is "good"?

  $\Rightarrow$ use the following three criteria:

  1. **Unbiased**: an estimator is unbiased if

  $$E[\text{estimator}] = \text{population parameter}$$

  2. **Relatively efficient**: an estimator is relatively efficient if, compared to another estiamtor with the same amount of bias, it has lower variance. That is, if

  $$E[\text{estimator}_a] = \text{population parameter} \quad \text{and} \quad E[\text{estimator}_b] = \text{population parameter}$$

Then estimator$_a$ is relatively efficient if

$$V(\text{estimator}_a) < V(\text{estimator}_b)$$

3. **Consistent**: an estimator is consistent if, as $n \to \infty$, the following two hold:
   (a) **Asymptotically unbiased**: $E[\text{estimator}] \to$ population parameter, and
   (b) $V(\text{estimator}) \to 0$

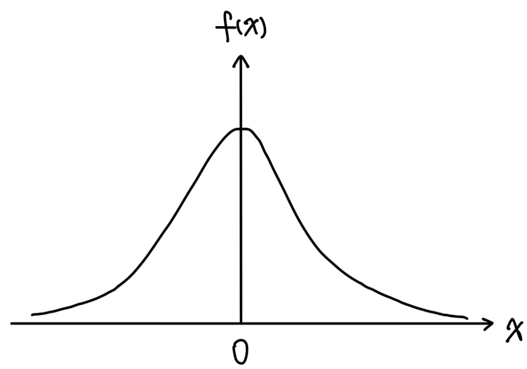
[Go to Exercise 1 and 2]

3 Interval Estimator: Confidence Interval

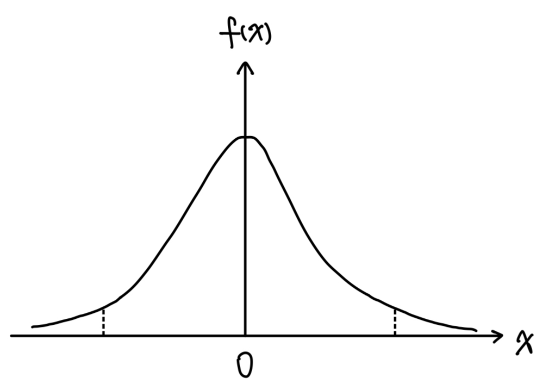
- Definition: an interval estimator estimates a range that's likely to include the true population parameter.
- One interval estimator that we often look at: **confidence interval**

3.1 Construct a confidence interval

- Think about a standard normal distribution:



- If we want to cover $(1 - \alpha)$ portion of this standard normal distribution, then



- We can think about this standard normal distribution as the sampling distribution of the mean, where the sample mean estimator has been standardized:

$$\begin{aligned}
& P\left(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\right) && = 1 - \alpha \\
\Leftrightarrow\quad & P\left(-Z_{\alpha/2} \leq \frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} \leq Z_{\alpha/2}\right) && = 1 - \alpha \\
\Leftrightarrow\quad & P\left(-Z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \leq \bar{X} - \mu_X \leq Z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right) && = 1 - \alpha \\
\Leftrightarrow\quad & P\left(\bar{X} - Z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \leq \mu_X \leq \bar{X} + Z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right) && = 1 - \alpha
\end{aligned}$$

  Thus, for $(1-\alpha)$ portion of area covered, the confidence interval constructed is

$$\left[\bar{X} - Z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right]$$

  We call $(1-\alpha)$ the **confidence level** for the above interval.

- What are some common confidence level and the associated Z score ($Z_{\alpha/2}$)?

Confidence level	α	$Z_{\alpha/2}$
90%	0.1	1.645
95%	0.05	1.96
99%	0.01	2.575

- **Interpretation**:

  Say that, for example, a 95% confidence interval of the mean of X, using a sample of size 70, is estimated to be $[4, 8]$. The following are some examples of correct interpretation of this confidence interval constructed.

  - **Correct version 1**: There's a 5% probability that the population mean of X lies outside of the confidence interval estimator. For this sample of size 70, we estimate the confidence interval to be $[4, 8]$.
  - **Correct version 2**: If random sample of size 70 were repeatedly selected, then in the long run, 95% of the confidence intervals formed would contain the true mean of X, which in this case is between 4 and 8.

[Go to Exercise 3]

3.2 Sample size needed given a already constructed confidence interval and confidence level

- We just saw that a confidence interval with $(1-\alpha)$ confidence level is constructed to be

$$\left[\bar{X} - Z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right]$$

In other words, the lower and upper bound of this confidence interval is calculated to be

$$\bar{X} \mp Z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}$$

- Say that instead, we want to specify how tight the confidence interval is. Usually, we do this by specifying a bound (B), which is the value that is subtracted from or added to the $\bar{X}$. That is, we want the lower and upper bound of a confidence interval to be calculated as

$$\bar{X} \mp B$$

- This implies that

$$B = Z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}$$

- Using this expression, we have chosen what B is. Often times, people also have in mind of what they want the confidence level to be (i.e. α is chosen), and σ_X is given. Thus, in order to set the bound as B, one can specify the sample size n:

$$\begin{aligned} B &= Z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}} \\ \sqrt{n} &= \frac{Z_{\alpha/2} \cdot \sigma_X}{B} \\ n &= \left(\frac{Z_{\alpha/2} \cdot \sigma_X}{B} \right)^2 \end{aligned}$$

- The sample size n obtained in this way is, more appropriately speaking, a lower bound, since a bigger n always shrinks the variance of the sample mean, meaning that the bound can be even tighter if needed.

  Thus, in order to achieve bound B under some α and σ_X, one needs sample size

$$n \geq \left(\frac{Z_{\alpha/2} \cdot \sigma_X}{B} \right)^2$$

[Go to Exercise 4]