5.2 Interval Estimation

The next type of estimate we look at is the interval estimate. Instead of trying to guess the exact value of p, we can guess a region within which we are reasonably confident p lies. Usually this region will be an interval, hence the name "Interval Estimation". The motivation behind the setup is that we may know very little about the distribution of interest (from which we are sampling), but the normal distribution is already very well characterized. We know a lot about the normal distribution, and the CLT tells us that almost all random variables can be connected to the normal distribution by taking sample means. Since sample means are approximately normal, we can use the known pdf of a normal distribution to compute probabilities that the sample means land in certain intervals on the real line. From here, we can construct intervals that are highly probable to contain the true parameters.

5.2.1 Applying the CLT

Without knowing anything about the underlying distribution of a sequence of random variables $\{X_i\}$, the CLT gives a statement about the sample means. In particular, if Y is a $N(0,1)$ random variable, and $\{X_i\}$ are distributed iid with mean μ and variance σ^2, then by Corollary 4.17,

$$P\Big(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \in A\Big) \approx P(Y \in A).$$

In particular, if we want an interval in which Y lands with probability 0.95, we look online or in a book for a z table, which will tell us that for a $N(0,1)$ random variable Y,

$$P(Y \in (-1.96, 1.96)) = P(-1.96 \leq Y \leq 1.96) = 0.95.$$

Since $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is nearly $N(0,1)$ distributed, we can roughly say that

$$P\Big(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\Big) = 0.95.$$

From the above statement we can make statements about experiments in order to quantify confidence. We will use the above expression in the next section to construct a confidence interval, and also in the section after that to describe hypothesis testing.

5.2.2 Confidence Intervals

Suppose that during the presidential election, we were interested in the proportion p of the population that preferred Hillary Clinton to Donald Trump. It wouldn't

be feasible to call every single person in the country and write down who they prefer. Instead, we can take a bunch of samples, $X_1, \ldots, X_n$ where

$$X_i = \begin{cases} 1 & \text{if person } i \text{ prefers Hillary} \\ 0 & \text{otherwise.} \end{cases}$$

What is the distribution on each X_i? In this description of voters, each X_i will equal 1 with probability p and 0 with probability $1-p$, so we can think of each voter as a coin flip with bias p (the proportion supporting Hillary). Then the sample mean, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, is the proportion of our sample that prefers Hillary. Note that $E\bar{X} = p$, since each $EX_i = 1 \cdot p + 0 \cdot (1-p) = p$. Then by the CLT,

$$\frac{\bar{X} - p}{\sigma/\sqrt{n}} \sim N(0,1).$$

Since we don't know the true value of σ, we estimate it using the sample variance, defined

$$S^2 \doteq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

This is a consistent estimator for σ^2, so for large n, the probability that it differs greatly from the true variance σ^2 is small. Supposing our sample is large, we can replace σ in our expression with $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$. Since $\frac{\bar{X}-p}{S/\sqrt{n}}$ is approximately $N(0,1)$ distributed, we have

$$P\Big(-1.96 \leq \frac{\bar{X} - p}{S/\sqrt{n}} \leq 1.96\Big) = 0.95.$$

Rearranging the expression for p, we have

$$\begin{aligned} P\Big(-1.96 \cdot \frac{S}{\sqrt{n}} \leq \bar{X} - p \leq 1.96 \cdot \frac{S}{\sqrt{n}}\Big) &= 0.95 \\ \Rightarrow P\Big(-1.96 \cdot \frac{S}{\sqrt{n}} - \bar{X} \leq -p \leq 1.96 \cdot \frac{S}{\sqrt{n}} - \bar{X}\Big) &= 0.95 \\ \Rightarrow P\Big(1.96 \cdot \frac{S}{\sqrt{n}} + \bar{X} \geq p \geq \bar{X} - 1.96 \cdot \frac{S}{\sqrt{n}}\Big) &= 0.95 \end{aligned}$$

Even though we do not know the true value for p, we can conclude from the above expression that with probability 0.95, p is contained in the interval

$$\Big(\bar{X} - 1.96 \cdot \frac{S}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{S}{\sqrt{n}}\Big).$$

This is called a 95% confidence interval for the parameter p. This approximation works well for large values of n, but a rule of thumb is to make sure $n > 30$ before using the approximation.

On the website, there is a confidence interval visualization. Try selecting the Uniform distribution to sample from. Choosing a sample size of $n = 30$ will cause batches of 30 samples to be picked, their sample means computed, and their resulting confidence intervals displayed on the right. Depending on the confidence level picked (the above example uses $\alpha = 0.05$, so $1 - \alpha = 0.95$), the generated confidence intervals will contain the true mean μ with probability $1 - \alpha$.