Summary of Lecture 6

Terminology:

- A **parameter** of a model refers to an unknown quantity we are interested in.
- An **estimator** is a function of the data that is used to infer the value of an unknown parameter in a statistical model. An estimator is a random variable because it is a function of the data, which are random variables.
- An **estimate** is the observed outcome of the estimator for a specific data set. In other words, an estimate is a realization of the random variable (the estimator).
- The **sampling distribution** of an estimator is a probability distribution that describes all the possible outcomes of the estimator we might see if we could "repeat" our sample over and over again.
- The **standard error** is the special name given to the standard deviation of the sampling distribution of an estimator.

We use the sample mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$, to estimate the unknown mean of an i.i.d. model $\mu = E[X_i]$.

The single **most important idea** of this lecture is that the sample mean $\bar{x}$ is a random variable before you see observations. This means that is has its own sampling distribution, which is different from the distribution of the samples. However, the two are related. We use the distribution of the sample mean to construct confidence intervals and perform hypothesis testing. Sometimes, the distribution of the sample mean is not known and we need to approximate it using the central limit theorem.

Suppose that the model for observations $X_1, \ldots, X_n$ is i.i.d. $N(\mu, \sigma^2)$. Here the mean μ and variance σ^2 are not known. The expected value of the sample mean is the same as the expected value of the observations, $E[\bar{x}] = \mu$, therefore, we say that the sample mean is an unbiased estimator. The variance of the sample mean is

$$\text{Var}[\bar{x}] = \frac{\sigma^2}{n}.$$

The average has a normal distribution,

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

We see that the sample mean is becoming a better estimator of μ as we get more samples. Since we do not know what the variance σ^2 is, we will use the sample variance s_x^2 as the best guess. The standard error for $\bar{x}$ is defined as

$$S.E.(\bar{x}) = \frac{s_x}{\sqrt{n}}.$$

The standardized quantity

$$\frac{\bar{x} - \mu}{s_x/\sqrt{n}} \approx N(0,1)$$

has approximately normal distribution based on the central limit theorem. We will use this distribution to construct confidence intervals for μ.

We compute $100(1-\alpha)\%$ confidence interval for μ as

$$\bar{x} \pm z_{1-\alpha/2} \cdot \frac{s_x}{\sqrt{n}}$$

where z_α is a number such that $P(Z < z_\alpha) = \alpha$ and $Z \sim N(0,1)$. For example, an approximate 95% confidence interval for μ is $\bar{x} \pm 1.96 s_x/\sqrt{n}$. A confidence interval captures our uncertainty about the mean μ. The interpretation of a $100(1-\alpha)\%$ confidence interval is that it will contain the true mean $(1-\alpha)$ fraction of times we constructed the interval.

Estimation and confidence intervals for success proportion under an i.i.d. Bernoulli model are constructed in the same way. In an i.i.d. Bernoulli model, we use proportion of successes in a sample, $\hat{p}$, as an estimator of p. The estimator $\hat{p}$, is a random variable before we see data. This estimator, $\hat{p}$, is an unbiased estimator, that is, $E[\hat{p}] = p$. Furthermore $\text{Var}(\hat{p}) = p(1-p)/n$. Its approximate distribution is given by the central limit theorem as

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right).$$

The standard error is $S.E.(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$. We will use the following standardized quantity

$$\frac{\hat{p} - p}{S.E.(\hat{p})} \approx N(0,1)$$

to construct confidence intervals as before. For example, we construct a 95% confidence interval for the true success probability as $\hat{p} \pm 1.96 \cdot S.E.(\hat{p})$.