

## Chapter 5 — Estimation

### 1. Distributions of Functions of RVs

- Two RVs are said to be **independent** if the realization of one of them does not change the probability distribution of the other, and vice versa. If two RVs are not independent, then they are **dependent**.
- Some rules of expectation and variance follow:
  - (a)  $E(c) = c$ .
  - (b)  $E(c * X) = c * E(X)$ .
  - (c)  $E(X + c) = E(X) + c$ .
  - (d)  $E(X + Y) = E(X) + E(Y)$ .
  - (e)  $VAR(c) = 0$ .
  - (f)  $VAR(c * X) = c^2 VAR(X)$ .
  - (g)  $VAR(X + c) = VAR(X)$ .
  - (h) If  $X$  and  $Y$  are independent,  $VAR(X + Y) = VAR(X) + VAR(Y)$ .
- A sample of size  $n$  from a population is called a **simple random sample** if every possible sample of size  $n$  is equally likely to be drawn.
- We say a sample is drawn **with replacement** if an element is replaced to the population before the next element is drawn. There is a chance the same element could be drawn more than once. Otherwise we say the sample is drawn **without replacement**, and every element can be drawn at most once.
- A collection of RVs  $X_1, X_2, \dots, X_n$  are said to be **independent and identically distributed**, or **iid**, if the following things are true:
  - They are all independent from one another. That is, the realization of any one of them does not change the probability distribution of any other one.
  - They all have exactly the same probability distribution.

### 2. Estimation

- Sample mean:  $\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$

- Sample variance of  $X$ :  $\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$
- Sample standard deviation of  $X$ :  $\hat{\sigma} = S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$
- The formula that describes how data from a sample would be used to compute a guess about a population parameter is called an **estimator**, or a **statistic**. The numerical value computed once the data is collected is called an **estimate**. An estimator is an RV, and an estimate is a realization of that RV.
- The **bias** in an estimator  $\hat{\theta}$  is defined as:

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

If the the bias is equal to zero, the estimator  $\hat{\theta}$  is called **unbiased** for  $\theta$ . All other things being equal, smaller bias is better.

- The variance of an estimator  $\hat{\theta}$  is defined as  $\text{VAR}(\hat{\theta})$ . All other things being equal, smaller variance is better. The square root of the variance is usually called the standard deviation or SD. However, when we are talking about estimating a parameter, we instead use the term **standard error** or **SE**, to remind us that this is the amount of error in estimation. Thus the square root of the variance of an estimator will be denoted  $SE(\hat{\theta})$ .
- The **mean squared error**, or **MSE**, of an estimator  $\hat{\theta}$  is defined as:

$$\text{MSE}(\hat{\theta}) = \text{VAR}(\hat{\theta}) + \text{bias}(\hat{\theta})^2.$$

All other things being equal, smaller MSE is better.

- $E(\bar{X}) = E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{\mu + \mu + \dots + \mu}{n} = \mu.$
- $\text{VAR}(\bar{X}) = \text{VAR}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{\sigma^2 + \sigma^2 + \dots + \sigma^2}{n^2} = \frac{\sigma^2}{n}.$
- $SE(\bar{X}) = \sqrt{\text{VAR}(\bar{X})} = \frac{\sigma}{\sqrt{n}}.$
- Estimated standard error of  $\bar{X}$ :  $\widehat{SE}(\bar{X}) = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{S}{\sqrt{n}}.$

3. A **normal quantile-quantile plot** or **normal QQ plot** can be used to evaluate normality. If the data appears to be drawn from a normally distributed population, the points in the plot will usually fall on a roughly straight line.
4. The Central Limit Theorem can be stated as follows. Let  $X_1, X_2, \dots, X_n$  be a collection of iid RVs with  $E(X_i) = \mu$  and  $VAR(X_i) = \sigma^2$ . For large enough  $n$ , the distribution of  $\bar{X}$  will be approximately normal with  $E(\bar{X}) = \mu$  and  $VAR(\bar{X}) = \frac{\sigma^2}{n}$ . That is,  $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ . The required size for  $n$  depends on the nature of the true distribution of  $X_i$ . The closer the distribution of  $X_i$  is to normal, the smaller  $n$  is required for the approximation to be good. Usually about  $n = 30$  is sufficient.

#### 5. Confidence Intervals

- The interpretation for a confidence interval constructed for a population parameter  $\theta$ , is that if you had theoretically taken many samples from the population, and created a different interval for each sample,  $100(1 - \alpha)\%$  of them would cover the true value of  $\theta$ . This is usually shortened to saying we have  $100(1 - \alpha)\%$  **confidence** that the interval covers  $\theta$ .
- When using  $\bar{X}$  to estimate  $\mu$ , if the  $X_i$  are normal and  $\sigma$  is known, or  $n$  is large enough for the CLT to work, then a  $100(1 - \alpha)\%$  CI for  $\mu$  is given by:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- When using  $\bar{X}$  to estimate  $\mu$ , if the  $X_i$  are normal,  $\sigma$  is unknown, and the sample size is small, then a  $100(1 - \alpha)\%$  CI for  $\mu$  is given by:

$$\bar{X} \pm t_{(n-1, \alpha/2)} \frac{S}{\sqrt{n}}.$$

- The general form for a CI often looks like:

$$\text{estimate} \pm \text{multiplier} * \text{estimated SE(estimator)}$$

- When intending to create a  $100(1 - \alpha)\%$  CI for  $\mu$ , assuming normality and a large sample size, the  $n$  required to achieve a half-width of no larger than  $H$  is given by:

$$n = \frac{(z_{\alpha/2}^2)(\sigma^2)}{H^2}.$$

## 6. Bootstrap Methods

- When the sample does not look like it was drawn from a normal population, and the sample size is too small to use the CLT to approximate the sampling distribution of  $t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$ , the **bootstrap** can be used to approximate the distribution of  $t$ . The steps are as follows:
  - (1) Compute the estimate of the sample mean from the data sampled,  $\bar{x}$ .
  - (2) Draw a simple random sample, with replacement, of size  $n$ , from the sample data. Call these observations  $x_1^*, x_2^*, \dots, x_n^*$ . Often this means that the same data point will be repeated twice in the resampling.
  - (3) Compute the mean and sd of the resampled data. Call these things  $\bar{x}^*$  and  $s^*$ .
  - (4) Compute the statistic  $\hat{t} = \frac{\bar{x}^* - \bar{x}}{\frac{s^*}{\sqrt{n}}}$ .
  - (5) Repeat steps 2-4 a large number of times, and compute  $\hat{t}$  from each one. This is an approximation to the sampling distribution of  $t$ .
- Using the bootstrap, a  $100(1 - \alpha)\%$  CI for  $\mu$  based on the approximate sampling distribution of  $t$  is given by:

$$(\bar{x} - \hat{t}_{(\alpha/2)} \frac{s}{\sqrt{n}}, \bar{x} - \hat{t}_{(1-\alpha/2)} \frac{s}{\sqrt{n}}),$$

where  $\hat{t}_{(\alpha/2)}$  and  $\hat{t}_{(1-\alpha/2)}$  are the  $\alpha/2$  and  $1 - \alpha/2$  critical values of the approximate sampling distribution.

## 7. Estimation of a Population Proportion

- If a sample can be considered a collection of iid RVs  $Y_i$  where the outcome of each is either zero or one, then we define the sample proportion:

$$\text{Sample proportion: } \hat{\pi} = P = \frac{\sum_{i=1}^n Y_i}{n}.$$

- $E(P) = \pi$ ,  $VAR(P) = \frac{\pi(1-\pi)}{n}$ ,  $SE(P) = \sqrt{\frac{\pi(1-\pi)}{n}}$ .
- So long as  $n\pi > 5$  and  $n(1 - \pi) > 5$ , the approximate distribution of  $P$  is:

$$P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right).$$

- So long as  $n\pi > 5$  and  $n(1 - \pi) > 5$ , an approximate  $100(1 - \alpha)\%$  CI for  $\pi$  would be of the form:

$$P \pm z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}.$$