Topic 2 - Point Estimation

Estimation is the process of estimating certain parameters from the data set itself. The **sampling distribution** of a statistic is its probability distribution, given an assumed population distribution and a sampling scheme (e.g. random sampling). Sometimes one can determine it exactly, but often one might resort to simulation. One needs to know the properties of their sampling distributions, such as the mean and variance. For example, it is natural to expect that:

- Sample Mean ≈ Population Mean
- Sample Variance ≈ Population Variance

A **parameter** is a quantity that describes the population distribution (for example, the mean μ). The **parameter space** is the set of all possible values that a parameter might take. An **estimator** (or point estimator) is a statistic that is used to estimate a parameter. It refers specifically to the random variable version of the statistic. An **estimate** (or point estimate) is the observed value of the estimator for a given dataset. In other words, it is a realisation of the estimator.

1. Sample Mean

The sample mean is denoted by the formula:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The sample mean has the following important (expectation) properties):

- $\mathbb{E}(\bar{X}) = \mu$
- $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$

2. Sample Variance

The sample variance is denoted by the formula:

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

Note that the expectation of the sample variance is equal to the population variance (non-biased estimator).

3. Sample Proportion

For a discrete random variable, one might be interested in how often a particular value appears. Counting this gives the sample frequency. Let the population proportion be p = Pr(X = a). Then freq(a) ~ Bi(n, p). When this is divided by the sample size, one gets the sample proportion. This is often used as an estimator for the population proportion:

$$\hat{p} = \frac{\text{freq}(a)}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i = a)$$

More importantly, for large n, one can approximate this with a normal distribution:

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

Note that the sample pmf and the sample proportion are the same; both of them estimate the probability of a given event or set of events. The pmf is usually used when the interest is in many different events/values, and is written as a function. The proportion is usually used when only a single event is of interest (getting heads for a coin flip, a certain candidate winning an election, etc).

If the sample is drawn from a normal distribution, one can derive exact distributions for these statistics. The sample mean and sample variance are distributed as:

Sample mean:

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

Sample variance:

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

$$\mathbb{E}(S^2) = \sigma^2, \qquad \mathrm{var}(S^2) = \frac{2\sigma^4}{n-1}$$

Consider an estimator θ of $\boldsymbol{\theta}$. If $E(\theta) = \boldsymbol{\theta}$, then the estimator is said to be **unbiased**.
The **bias** of the estimator is $E(\theta)$ - $\boldsymbol{\theta}$. Note that unbiasedness is not preserved under transformations.

When choosing between estimators, evaluate and compare the sampling distributions of the estimators. Generally, prefer estimators that have smaller bias and smaller variance (and it can vary depending on the aim of the problem). Sometimes, one only knows asymptotic properties of estimators. Note that this approach to estimation is referred to as frequentist or classical inference.

Method of Moments

The idea is to make the population distribution resemble the empirical (data) distribution by equating theoretical moments with sample moments. Do this until one has enough equations, and then solve them. The general procedure (for *r* parameters) is:

1. $X_1, \ldots, X_n$ i.i.d. $f(x \mid \theta_1, \ldots, \theta_r)$.
2. kth moment is $E(X^k)$
3. kth sample moment is $M_k = \frac{1}{n}\sum X_i^k$
4. Set $E(X^k) = M_k$, for $k = 1, \ldots r$ and solve for $(\theta_1, \ldots, \theta_r)$

Note that one can use the variance instead of the second moment (which is sometimes more convenient). However, this is an intuitive approach to estimation and can work in situations where

other approaches are too difficult. Note that it is usually biased, usually not optimal (but may suffice). An example is shown below.

- Sampling from: $X \sim \text{Geom}(p)$
- The first moment:

$$E(X) = \sum_{x=1}^{\infty} xp(1-p)^{x-1} = \frac{1}{p}$$

- The MM estimator is obtained by solving

$$\bar{X} = \frac{1}{p}$$

which gives

$$\widetilde{p} = \frac{1}{\bar{X}}$$

Method of Maximum Likelihood

The main idea is to find the 'most likely' explanation for the data. More concretely, one needs to find parameter values that maximise the probability of the data. The general procedure to find the MLE, is to firstly have a random sample of iid variables. Hence, the likelihood function with *m* parameters and data is given by:

$$L(\theta_1, \ldots, \theta_m) = \prod_{i=1}^{n} f(x_i \mid \theta_1, \ldots, \theta_m)$$

If X is discrete, for *f* use the pmf. If X is continuous, for *f* use the pdf.

The **maximum likelihood estimates** (MLEs) or maximum likelihood estimators are values that maximise the likelihood function. It is often (but not always) useful to take logs and then differentiate and equate derivatives to zero to find MLE's. Sometimes this is too hard, but one can maximise numerically. An example is shown below.

Sampling (iid) from: $X \sim \text{Exp}(\lambda)$

$$f(x \mid \lambda) = \frac{1}{\lambda} e^{-x/\lambda}, \quad x > 0, \quad 0 < \lambda < \infty$$

$$L(\lambda) = \frac{1}{\lambda^n} \exp\left(\frac{-\sum_{i=1}^{n} x_i}{\lambda}\right)$$

$$\ln L(\lambda) = -n \ln(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} x_i$$

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = -\frac{n}{\lambda} + \frac{\sum x_i}{\lambda^2} = 0$$

This gives: $\hat{\lambda} = \bar{X}$

Also note that the MLE satisfies the **invariance property**. In other words, transformations don't affect the value of the MLE. The consequence is that MLEs are usually biased since expectations are not invariant under transformations. Some useful results is that the MLE is asymptotically unbiased, asymptotically optimal variance ('efficient') and is asymptotically normally distributed.

Topic 3 - Interval Estimation

Point estimates are usually only a starting point, but are insufficient to conclusively answer real questions of interest. The variance of the estimators calculated previously tells one a typical amount by which the estimate will vary from one sample to another, and thus (for an unbiased estimator) how close to the true parameter value it is likely to be.

The **standard error** of an estimate is the estimate standard deviation of the estimator. An **interval estimate** is of the form (est - error, est + error). It is more general and more useful than just reporting a standard error. For example, it can cope with skewed (asymmetric) sampling distributions. It is a pair of statistics defining an interval that aims to convey an estimate (of a parameter) with uncertainty.

The resulting interval estimate is called a **95% confidence interval** for the parameter, in which the interval has probability 0.95 of containing the parameter of interest. This interval estimator is a **random interval** and is calculable from the sample. The parameter is fixed and unknown. Before the sample is taken, the probability the random interval contains the parameter is 95%. After the sample is taken, there is a realised interval. It no longer has a probabilistic interpretation: it either contains the parameter or does not.

The following random interval contains μ with probability 1 - α (for known variance):

$$(\bar{X} - c\frac{\sigma}{\sqrt{n}}, \bar{X} + c\frac{\sigma}{\sqrt{n}})$$

The general technique for deriving a confidence interval is to start with an estimator, T, whose sampling distribution is known. Then follow these steps:

- Write the central probability interval based on its sampling distribution,
$$\Pr(\pi_{0.025} < T < \pi_{0.975}) = 0.95$$
- The endpoints will depend on the parameter, θ, so can write it as,
$$\Pr(a(\theta) < T < b(\theta)) = 0.95$$
- Invert it to get a random interval for the parameter,
$$\Pr\left(b^{-1}(T) < \theta < a^{-1}(T)\right) = 0.95$$
- Substitute observed value, t, to get an interval estimate,
$$\left(b^{-1}(t),\ a^{-1}(t)\right)$$