Sampling & Statistical Inference

Random Sample

A set of items selected from a parent population is a random sample if:

- the probability that any item in the population is included in the sample is proportional to its frequency in the parent population and

- the inclusion/exclusion of any item in the sample operates independently of the inclusion/exclusion of any other item.

Random Sample Notation

- A random sample is made up of (iid) random variables and so they are denoted by capital X's.
- We will use the shorthand notation $\underline{X}$ to denote a random sample, that is, $\underline{X} = (X_1, X_2, \ldots, X_n)$.
- An observed sample will be denoted by $\underline{x} = (x_1, x_2, \ldots, x_n)$.
- The population distribution will be specified by a density (or probability function) denoted by $f(x;\theta)$, where θ denotes the parameter(s) of the distribution such as mean (denoted by μ) and variance (denoted by σ^2).

Sum of Independent R.V.

- $E[X_1+X_2+....+X_n] = E[X_1]+E[X_2]+....+E[X_n]$

- $var[X_1+X_2+....+X_n] = var[X_1]+var[X_2]+....+var[X_n]$

Definition of a statistic

- A statistic is a function of $\underline{X}$ only and does not involve any unknown parameters.

- Examples:
  Sample Mean, $\bar{X} = (1/n) * \Sigma X_i$
  Sample Variance, $S^2 = (1/(n-1)) * \Sigma (X_i - \bar{X})^2$
- Non-Example:
  $(1/n) * \Sigma (X_i - \mu)^2$ is not a statistic, unless μ is known.

- A statistic can be generally denoted by $g(\underline{X})$. Since a statistic is a function of random variables, it will be a random variable itself and will have a distribution, its sampling distribution.

The Sample Mean

- Suppose that we have n independent and identically distributed random variables, X_i , i = 1, 2,..., n, each with mean μ and variance σ^2.

- Sample Mean, $\bar{X} = (1/n) * \Sigma X_i$

- $E[\bar{X}] = \mu$
- $var[\bar{X}] = \sigma^2/n$
- $sd[\bar{X}] = \sigma/\sqrt{n}$

- For large n, $\bar{X} \sim N(\mu, \sigma^2/n)$ OR $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$
- This result is known as the Central Limit Theorem(CLT) or the z result.

Normal Approximations Using CLT

- Bin(n, p) ≈ N(np, np(1-p)), if np>5 and n(1-p) > 5
- Poi(λ) ≈ N(λ, λ), if λ is large
- $\chi_n^2 \approx$ N(n, 2n), if n is large

The Sample Variance

- Suppose that we have n independent and identically distributed random variables, X_i , i = 1, 2,..., n, each with mean μ and variance σ^2.

- Sample Variance, $S^2 = (1/(n-1)) * \Sigma (X_i - \bar{X})^2$

- $E[S^2] = \sigma^2$
- $var[S^2]$ -> Dependent on population distribution

- The sampling distribution of S^2 when sampling from a normal population, with mean μ and variance σ^2, is: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

The t result

- In most cases, σ^2 is not known and so the z result cannot be used in such cases. We use the t result in such cases.

$$\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}$$

- This is the t result or the t sampling distribution.
- This result is valid for samples from normal distribution only.
- The t distribution is symmetrical about zero.

The F result

- If independent random samples of size n_1 and n_2 respectively are taken from normal populations with variances σ_1^2 and σ_2^2 , then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1} \quad \text{OR} \quad \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \sim F_{n_2-1,n_1-1}$$

- This result is valid for samples from normal distribution only.

Point Estimation

Method of Moments

- **The basic principle is to equate population moments (i.e the means, variances, etc of the theoretical model) to corresponding sample moments (i.e the means, variances, etc of the sample data observed) and solve for the parameter(s).**
- **One parameter case:**

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Two parameter case:**

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad E(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Method of Moments Example

- A random sample from a Exp(λ) distribution is as follows:

  14.84, 0.19, 11.75, 1.18, 2.44, 0.53

  Calculate the method of moments estimate for λ.

- $\bar{x} = (14.84+0.19+11.75+1.18+2.44+0.53)/6 = 5.155$
  $E[X] = 1/\lambda$
  According to method of moments: $E[X] = \bar{x} \Rightarrow 1/\lambda = 5.155 \Rightarrow \lambda = 0.1940$

Method of Maximum Likelihood

- The method of maximum likelihood is widely regarded as the best general method of finding estimators.
- Likelihood of a sample $x_1, x_2, \ldots, x_n$ from a population with density or probability function f (x; θ) is given by:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

- Differentiating the likelihood or log likelihood with respect to the parameter and setting the derivative to zero gives the maximum likelihood estimator (MLE) for the parameter (denoted by $\hat{\theta}$).

Method of Maximum Likelihood Example

Given a random sample of size n $(ie\ x_1, \ldots, x_n)$ from the exponential population with density $f(x) = \lambda e^{-\lambda x}$, $x > 0$, the MLE, $\hat{\lambda}$, is found as follows:

$$L(\lambda) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}$$

$$\therefore \log L(\lambda) = n \log \lambda - \lambda \sum x_i$$

$$\frac{\partial}{\partial \lambda} \log L(\lambda) = \frac{n}{\lambda} - \sum x_i$$

equating to zero:

$$\frac{n}{\lambda} - \sum x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$$

$$\therefore \text{MLE is } \hat{\lambda} = \frac{1}{\bar{X}}$$

Properties of Estimators

- Let us take a random sample $\underline{X} = (X_1, X_2, \ldots, X_n)$ from a distribution with an unknown parameter θ and $g(\underline{X})$ is an estimator of θ.
- Bias = $E[g(\underline{X})] - \theta$
- If bias of an estimator is zero, it is said to be an unbiased estimator.
- Important: S^2 is an unbiased estimator of σ^2. Hence, we take (n-1) instead of n.
- Mean Squared Error, $MSE(g(\underline{X})) = E[(g(\underline{X}) - \theta)^2] = bias^2 + variance$
- Estimator with a lower MSE is said to be more efficient.