5. Random Samples

Definition 5.1

A set of random variables $X_1, \ldots, X_n$ is called a Random Sample from a population if $X_1, \ldots, X_n$ are mutually independent and each X_i has the same cdf F.

- F describes the assumed distribution in the population.
- Corresponding to F is a pdf (or pmf) f.
- $X_1, \ldots, X_n$ are Independent and Identically Distributed (*iid*).

- Joint pdf (pmf) of the sample

$$f(x_1, \ldots, x_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta})$$

- Usually the parameter vector $\boldsymbol{\theta}$ is unknown.

- Aim is to make inference about $\boldsymbol{\theta}$ based on the observed sample $x_1, \ldots, x_n$.

- Inference is based on statistics.

Definition 5.2

Let $X_1, \ldots, X_n$ be a random sample from an infinite population and let $T(x_1, \ldots, x_n)$ be a function mapping the support of $X_1, \ldots, X_n$, $\mathcal{X}^n$ to $\mathbb{R}^m$ where $m \leqslant n$. Then the random variable (or vector)

$$Y = T(X_1, \ldots, X_n)$$

is called a **statistic**. *The distribution of the random variable Y is known as its* **sampling distribution**.

- Since Y is a function of $X_1, \ldots, X_n$ its sampling distribution is found from $f(x_1, \ldots, x_n \mid \boldsymbol{\theta})$.

Definition 5.3

If $X_1, \ldots, X_n$ is a random sample then the sample mean is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and the sample variance is

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

The positive square root, S, of the sample variance is called the sample standard deviation.

- Observed values of $X_1, \ldots, X_n$ are $x_1, \ldots, x_n$
- Observed values of $\overline{X}$ and S^2 are

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}.$$

Sums of Random Variables

Theorem 5.1

Let $X_1, \ldots, X_n$ be a sequence of random variables with finite means and variances and let $a_1, \ldots, a_n$ be real constants. Then

$$\begin{aligned}
\mathsf{E}\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i \mathsf{E}(X_i) \\
\mathsf{Var}\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i^2 \mathsf{Var}(X_i) + \sum_{j \neq i} a_i a_j \mathsf{cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} a_i^2 \mathsf{Var}(X_i) + 2 \sum_{i>j} a_i a_j \mathsf{cov}(X_i, X_j)
\end{aligned}$$

Corollary 5.1.1

Let $X_1, \ldots, X_n$ be a random sample from a distribution having finite mean, μ and finite variance, σ^2, and let $a_1, \ldots, a_n$ be real constants. Then

$$\begin{aligned} \mathsf{E}\left(\sum_{i=1}^{n} a_i X_i\right) &= \mu \sum_{i=1}^{n} a_i \\ \mathsf{Var}\left(\sum_{i=1}^{n} a_i X_i\right) &= \sigma^2 \sum_{i=1}^{n} a_i^2 \end{aligned}$$

Corollary 5.1.2

Let $X_1, \ldots, X_n$ be a random sample and let g be a function such that $Y = g(X_1)$ has finite mean and variance. Then

$$\begin{aligned} \mathsf{E}\left(\sum_{i=1}^{n} g(X_i)\right) &= n\mathsf{E}[g(X_1)] \\ \mathsf{Var}\left(\sum_{i=1}^{n} g(X_i)\right) &= n\mathsf{var}[g(X_1)] \end{aligned}$$

Convolutions

Theorem 5.2 (Bivariate Convolution)

If X and Y are independent random variables with pdfs f_X and f_Y respectively and $Z = X + Y$ then the pdf of Z is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z - w) dw.$$

Theorem 5.3 (General Convolution)

Let $X_1, \ldots, X_n$ be a sequence of independent random variables such that X_i has pdf f_{X_i} and let $Z = \sum X_i$. Then the pdf of Z is

$$f_Z(z) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[f_{X_1}(w_1) \prod_{i=2}^{n-1} f_{X_i}(w_i - w_{i-1}) \right. \\ \left. f_{X_n}(z - w_{n-1}) \right] dw_1 \cdots dw_{n-1}.$$

Sample Mean

Theorem 5.4

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and finite variance σ^2 and let $\overline{X}$ be the corresponding sample mean. Then

$$\mathsf{E}\left(\overline{X}\right) = \mu, \quad \textit{and} \quad \mathsf{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Theorem 5.5

Let $X_1, \ldots, X_n$ be a random sample from a population with moment generating function $M_X(t)$ then the sampling distribution of the sample mean $\overline{X}$ has moment generating function

$$M_{\overline{X}}(t) = \left[M_X\left(\frac{t}{n}\right)\right]^n$$

- A statistic that is used to estimate a population quantity (parameter) θ is called an **estimator**.

- The observed value of an estimator is called the **estimate**.

Definition 5.4

A statistic $T(X_1, \ldots, X_n)$ is said to be an unbiased estimator of the parameter θ if, and only if,

$$\mathsf{E}_\theta \Big(T(X_1, \ldots, X_n) \Big) = \theta$$

for all possible values of θ.

Theorem 5.6

Let $X_1, \ldots, X_n$ be a random sample from a population with finite mean and variance μ and σ^2. Then $\overline{X}$ is an unbiased estimator of μ and S^2 is an unbiased estimator of σ^2.

- In the exponential family, we have the following important result.

Theorem 5.7

Suppose that $X_1, \ldots, X_n$ is a random sample from an exponential family with pdf or pmf given by

$$f(x \mid \boldsymbol{\theta}) \;=\; h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right)$$

and further suppose that the set $\{w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})\}$ contains an open subset in $\mathbb{R}^k$.

Define the statistics

$$T_i = \sum_{j=1}^{n} t_i(X_j) \qquad \text{for } i = 1, \ldots, k$$

then the joint distribution of $\boldsymbol{T} = (T_1, \ldots, T_k)$ is also an exponential family of the form

$$f_{\boldsymbol{T}}(t_1, \ldots, t_k \mid \boldsymbol{\theta}) \;=\; h_1(t_1, \ldots, t_k)\Big[c(\boldsymbol{\theta})\Big]^n \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i\right)$$

- The condition that $\{w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})\}$ contains an open subset in $\mathbb{R}^k$ eliminates curved exponential families.

- We shall see later that $\boldsymbol{T}$ contains the same information about $\boldsymbol{\theta}$ as the entire sample.

Normal Random Samples

Theorem 5.8

Let $X_1, \ldots, X_n$ be a sample from a $N(\mu, \sigma^2)$ population and let

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

be the sample mean and variance. Then

(i) *$\overline{X}$ and S^2 are independent.*

(ii) *$\overline{X} \sim N(\mu, \sigma^2/n)$.*

(iii) *$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.*

Definition 5.5

Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$ *is a sample from a population with cdf depending on some parameters* $\boldsymbol{\theta}$*. A quantity* $R(\boldsymbol{X}, \boldsymbol{\theta})$ *which is a function of the data and the parameters is called a* **pivot** *if the sampling distribution of* R *does not depend on the parameters* $\boldsymbol{\theta}$*.*

- If $X_1, \ldots, X_n$ are *iid* $\mathsf{N}(\mu, \sigma^2)$ then

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim \mathsf{N}(0, 1)$$

- Another pivot is

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S}$$

The Student's t distribution

Theorem 5.9

If Z and X are two independent random variables with $Z \sim N(0,1)$ and $X \sim \chi^2_\nu$ then the random variable

$$T = \frac{Z}{\sqrt{X/\nu}}$$

has pdf given by

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \text{for } t \in \mathbb{R}.$$

The distribution with this pdf is called the Student's t distribution with ν degrees of freedom.

- $\mathrm{E}(T^r)$ exists if, and only if, $r < \nu$.

- $\mathrm{E}(T) = 0 \quad$ for $\nu > 1$.

- $\mathrm{Var}(T) = \frac{\nu}{\nu-2} \quad$ for $\nu > 2$.

- Suppose that $X \sim t_1$ then

$$f_X(x) = \frac{1}{\pi(1+x^2)} \qquad -\infty < x < \infty.$$

  This is called the standard Cauchy distribution.

Theorem 5.10

Suppose that $T_1, T_2, \ldots$ *is a sequence of random variables such that* $T_\nu \sim t_\nu$ *and* $Z \sim N(0,1)$. *Then*

$$T_\nu \xrightarrow{d} Z \text{ as } \nu \to \infty$$

Theorem 5.11

Suppose that $X_1, \ldots, X_n$ *is a random sample from a Normal*(μ, σ^2) *population and that* $\overline{X}$ *and* S^2 *are the sample mean and sample variance. Then*

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t_{n-1}$$

Snedecor's F Distribution

Definition 5.6

A random variable Y is said to have and F distribution with p numerator degrees of freedom and q denominator degrees of freedom if, and only if, its pdf is given by

$$f_Y(y) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2}\frac{y^{(p/2)-1}}{[1+(p/q)y]^{(p+q)/2}} \text{ for } 0 < y < \infty.$$

Theorem 5.12

Suppose that X_1 has a Chi-squared(p) distribution, X_2 has a Chi-squared(q) distribution and X_1 and X_2 are independent. Then the random variable

$$Y = \frac{X_1/p}{X_2/q}$$

has an F distribution with p and q degrees of freedom.

Theorem 5.13

Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent random samples from normal populations with parameters (μ_X, σ_X^2) and (μ_Y, σ_Y^2) respectively. Let $\overline{X}$ and $\overline{Y}$ be the sample means and S_X^2 and S_Y^2 be the sample variances. Then

(i)
$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}$$

Now suppose that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ and define the **pooled variance estimate**

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

Then we also have

(ii)
$$\frac{(n+m-2)S_p^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

(iii)
$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

Order Statistics

Definition 5.7

Let $X_1, \ldots, X_n$ be a random sample then the **order statistics** *of the sample are denoted $X_{(r)}, \quad r = 1, \ldots, n$ where*

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}.$$

Theorem 5.14

Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf F_X. Then the cdf of the sample maximum, $X_{(n)}$, is

$$F_{X_{(n)}}(x) = [F_X(x)]^n .$$

and that for the minimum, $X_{(1)}$, is

$$F_{X_{(1)}}(x) = 1 - [1 - F_X(x)]^n .$$

Theorem 5.15

Let $X_1, \ldots, X_n$ be a random sample from a discrete distribution on the values $x_1 < x_2 < \cdots$. Let the common probability mass function of the random variables be $\mathsf{P}(X = x_i) = p_i$ with corresponding cdf

$$\mathsf{P}(X \leqslant x_i) = P_i = \sum_{k=1}^{i} p_k$$

and let us define $P_0 = 0$.

If $X_{(r)}$ is the r^{th} order statistic of the sample then

$$\mathsf{P}(X_{(r)} \leqslant x_i) = \sum_{k=r}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

$$\text{and} \quad \mathsf{P}(X_{(r)} = x_i) = \sum_{k=r}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k} \right]$$

Theorem 5.16

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with pdf f_X and cdf $F_X(x)$ and let $X_{(r)}$ be the r^{th} order statistic. Then the pdf of $X_{(r)}$ is

$$f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!} f_X(x) \left[F_X(x)\right]^{r-1} \left[1 - F_X(x)\right]^{n-r}.$$

Theorem 5.17

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with pdf f_X and cdf $F_X(x)$ and let $X_{(r)}$ and $X_{(s)}$ be two order statistics with $r < s$. Then the joint pdf of $X_{(r)}$ and $X_{(s)}$ is

$$\begin{aligned} f_{X_{(r)},X_{(s)}}(u,v) &= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} f_X(u) f_X(v) \\ &\quad \times [F_X(u)]^{r-1} [F_X(v) - F_X(u)]^{s-r-1} [1 - F_X(v)]^{n-s} \end{aligned}$$

for $-\infty < u < v < \infty$.

Theorem 5.18

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with pdf f_X and let $X_{(1)}, \ldots X_{(n)}$ be the order statistics. Then the joint pdf of all of the order statistics is

$$f_{X_{(1)},\ldots X_{(n)}}(x_1, \ldots, x_n) = n! f_X(x_1) \cdots f_X(x_n)$$

for $-\infty < x_1 < \cdots < x_n < \infty$.