THE RANDOM SAMPLE

Defn: A collection of rvs $X_1, \dots, X_n$ which are mutually independent and which all have the same distribution is called a random sample.

Remarks:

- Such a collection of rvs arises from a data generating process in which the same experiment is conducted n times independently.

- "Random sample" often refers to a set of individuals/items drawn from a larger population such that each individual/item is equally likely to be drawn.

  In truth, selecting n individuals/items without replacement from a population of finite size does NOT result in a collection of mutually independent rvs, because every time an individual/item is drawn, the composition of the population changes, making the draws dependent.

  However, if the population is large, its change in composition due to sampling without replacement is so slight that we may treat draws without replacement as though they were independent.

- The common distribution of the rvs in a random sample is often called the population distribution, and quantities which describe the population distribution are often called population parameters.

- We often introduce a random sample by saying, "Let $X_1, \dots, X_n$ be independent identically distributed (iid) rvs with such-and-such a distribution."

  Or we say, "Let $X_1, \dots, X_n$ be independent copies of the rv X, where X has such-and-such a distribution."

- We generally wish to learn something from a random sample $X_1, \dots, X_n$ about certain population parameters.

- We learn about population parameters from sample statistics:

Defn: A sample statistic (just "a statistic" for short) is a function computed on the rvs of a random sample. The distribution of the statistic is called its sampling distribution.

The setup: Consider a random sample $X_1, \ldots, X_n$ from a population with cdf F_X and support on $\mathcal{X}$.

Let T_n be the rv defined as

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

for some function $\underbrace{T : \mathcal{X}^n \to \mathcal{Y}}$

Takes the n values in the random sample and returns a single number in a set $\mathcal{Y}$.

Here, T_n is a statistic, and the distribution of T_n is called its sampling distribution.

The statistic T_n may carry information about a certain population parameter.

Example: Let $X_1, \ldots, X_n$ be a random sample of times until failure of electronic components, where the failure times are assumed to have the Exponential (λ) distribution, but λ is not known.

It is of interest to learn the value of λ.

- The unknown λ is the population parameter of interest.

- The sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ is a sample statistic which carries information about λ.

What we can learn from a sample statistic about a population parameter has everything to do with the sampling distribution of the statistic.

In the following, we consider the sampling distributions of statistics which are sums of the rvs in the random sample and of statistics based on ordering the values in the random sample.

SAMPLE MEAN AND SAMPLE VARIANCE

We focus on two statistics which are based on sums of the rvs in a random sample:

1. $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$
2. $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$

Before trying to get the sampling distribution of $\bar{X}_n$ and S_n^2, we first focus only on $\mathbb{E}\bar{X}_n$, $\mathrm{Var}\,\bar{X}_n$, and $\mathbb{E}S_n^2$.

Theorem: Let $X_1, \dots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then

(a) $\mathbb{E}\bar{X}_n = \mu$ (Values of $\bar{X}_n$ are centered around μ)

(b) $\mathrm{Var}\,\bar{X}_n = \sigma^2/n$ (Variance of $\bar{X}_n$ diminishes as we increase the sample size)

(c) $\mathbb{E}S_n^2 = \sigma^2$ (Values of S_n^2 are centered around σ^2)

Proof: (a)

$$\mathbb{E}\bar{X}_n = \mathbb{E}\left[\frac{1}{n}(X_1 + \dots + X_n)\right] = \frac{1}{n}(\mathbb{E}X_1 + \dots + \mathbb{E}X_n) = \frac{1}{n} n\mu = \mu$$

(b)

$$\begin{aligned}
\mathrm{Var}\,\bar{X}_n &= \mathrm{Var}\left[\frac{1}{n}(X_1 + \dots + X_n)\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n} \mathrm{Var}\,X_i + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)\right] \\
&= \frac{1}{n^2}(n\sigma^2) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

(c)

$$
\begin{aligned}
\mathbb{E}\, S_n^2 &= \mathbb{E}\, \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \\
&= \mathbb{E}\, \frac{1}{n-1} \sum_{i=1}^{n} (X_i^2 - 2X_i \bar{X}_n + \bar{X}_n^2) \\
&= \mathbb{E}\, \frac{1}{n-1} \Big[\sum_{i=1}^{n} X_i^2 - 2 \underbrace{\sum_{i=1}^{n} X_i}_{= n\bar{X}_n} \bar{X}_n + n\bar{X}_n^2 \Big] \\
&= \mathbb{E}\, \frac{1}{n-1} \Big[\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2 \Big] \\
&= \frac{1}{n-1} \Big[n\, \mathbb{E}X_1^2 - n\, \mathbb{E}\bar{X}_n^2 \Big] \\
&= \frac{1}{n-1} \Big[n(\sigma^2 + \mu^2) - n\Big(\frac{\sigma^2}{n} + \mu^2\Big) \Big] \\
&= \frac{(n-1)\sigma^2}{(n-1)} \\
&= \sigma^2 .
\end{aligned}
$$

$\mathrm{Var}\, X_1 = \mathbb{E}X_1^2 - (\mathbb{E}X_1)^2$

$\iff \mathbb{E}X_1^2 = \mathrm{Var}\, X_1 + (\mathbb{E}X_1)^2$

Getting the sampling distribution of $\bar{X}_n$:

We can get the mgf of $\bar{X}_n$ in terms of the mgf of the population distribution. We have

$$
M_{\bar{X}_n}(t) = M_{\frac{1}{n}(X_1 + \dots + X_n)}(t) = M_{X_1 + \dots + X_n}(t/n) = \prod_{i=1}^{n} M_{X_i}(t/n) = \left[M_{X_1}(t/n) \right]^n
$$

Example: Let $X_1, \dots, X_n$ be a random sample from the Gamma(α, β) distribution. Then

$$
M_{X_1}(t) = (1 - \beta t)^{-\alpha},
$$

giving

$$
M_{\bar{X}_n}(t) = (1 - \beta(t/n))^{-n\alpha},
$$

so that $\bar{X}_n \sim \text{Gamma}(n\alpha, \beta/n)$.

Example: We have already used this method to show that if $X_1, \dots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$, then $\bar{X}_n \sim \text{Normal}(\mu, \sigma^2/n)$.

It is difficult to get the distribution of S_n^2, unless the random sample comes from a Normal population. We will come back to this later.

ORDER STATISTICS

Many interesting statistics are functions of the ordered values of a random sample $X_1, \dots, X_n$.

Define the following notation:

$$X_{(1)} = \text{the least of } X_1, \dots, X_n$$

$$X_{(2)} = \text{the next-to-least of } X_1, \dots, X_n$$

$$\vdots$$

$$X_{(n)} = \text{the greatest of } X_1, \dots, X_n,$$

assuming there are no ties among $X_1, \dots, X_n$. Then

$$X_{(1)} < X_{(2)} < \dots < X_{(n)}$$

are called the order statistics of the random sample.

Equipped with the notation of order statistics, we can use them to express familiar sample statistics:

The range: $R_n = X_{(n)} - X_{(1)}$

The midrange: $M_n = (X_{(n)} + X_{(1)})/2$

The median: $\hat{\xi}_{0.5} = \begin{cases} \left(X_{(n/2)} + X_{(\frac{n+2}{2})}\right)/2, & \text{if } n \text{ even} \\ X_{(\frac{n+1}{2})}, & \text{if } n \text{ odd} \end{cases}$

Once we have studied order statistics we can derive the distributions of sample statistics based on ordering the values in the random sample.

First we derive the marginal pdf of the k^{th} order statistic $X_{(k)}$, for any $k = 1, \dots, n$.

Theorem: Let $X_{(1)}, \dots, X_{(n)}$ be the order statistics of a random sample from a population with cdf F_X and pdf f_X. Then the pdf of $X_{(k)}$ is given by

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

for $k = 1, \dots, n$.

Proof: We begin by finding the cdf $F_{X_{(k)}}$ of $X_{(k)}$ and then we take its derivative.

For any x, we have

$$F_{X_{(k)}}(x) = P_{X_{(k)}}\left(X_{(k)} \le x\right)$$

$$= P_{(X_1,\dots,X_n)}\left(\text{At least } k \text{ of } X_1,\dots,X_n \text{ are less than or equal to } x\right)$$

$Y = \#\{X_1,\dots,X_n \text{ less than or equal to } x\}$

$$= \sum_{j=k}^{n}\binom{n}{j}\left[F_X(x)\right]^{j}\left[1 - F_X(x)\right]^{n-j}$$

$\sim \text{Binomial}\left(n, F_X(x)\right)$,

$\Rightarrow$ find $P_Y(Y \ge k)$

We now take the derivative with respect to x:

$$\frac{d}{dx}F_{X_{(k)}}(x) = \frac{\partial}{\partial x}\sum_{j=k}^{n}\binom{n}{j}\left[F_X(x)\right]^{j}\left[1 - F_X(x)\right]^{n-j}$$

$$= \sum_{j=k}^{n}\binom{n}{j}\Big(j\left[F_X(x)\right]^{j-1}\left[1 - F_X(x)\right]^{n-j} f_X(x)$$

$$- \left[F_X(x)\right]^{j}(n-j)\left[1 - F_X(x)\right]^{n-j-1} f_X(x)\Big)$$

Split up first summation

j=k term

$$= \binom{n}{k} k\left[F_X(x)\right]^{k-1}\left[1 - F_X(x)\right]^{n-k} f_X(x)$$

$$+ \sum_{j=k+1}^{n}\binom{n}{j} j\left[F_X(x)\right]^{j-1}\left[1 - F_X(x)\right]^{n-j} f_X(x)$$

We can remove the j=n case, because this term is zero

$$- \sum_{j=k}^{n-1}\binom{n}{j}(n-j)\left[F_X(x)\right]^{j}\left[1 - F_X(x)\right]^{n-j-1} f_X(x)$$

$$= \frac{n!}{(k-1)!\,(n-k)!}\left[F_X(x)\right]^{k-1}\left[1-F_X(x)\right]^{n-k} f_X(x)$$

change index and change summand accordingly

$$\underbrace{\begin{aligned} &+ \sum_{j=k}^{n-1} \binom{n}{j+1}(j+1)\left[F_X(x)\right]^{j}\left[1-F_X(x)\right]^{n-j-1} f_X(x) \\ &- \sum_{j=k}^{n-1} \binom{n}{j}(n-j)\left[F_X(x)\right]^{j}\left[1-F_X(x)\right]^{n-j-1} f_X(x) \end{aligned}}_{=0}$$

Noting that

$$\binom{n}{j+1}(j+1) = \frac{n!}{j!\,(n-j-1)!} = \binom{n}{j}(n-j),$$

we see that the last two terms cancel out, giving the result.

For the minimum and maximum of a random sample, we have the following, coming directly from the theorem.

Corollary: Let $X_1, \dots, X_n$ be a random sample from a population with cdf F_X and pdf f_X. Then

1. $\min\{X_1, \dots, X_n\} = X_{(1)}$ has cdf and pdf given by

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

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

2. $\max\{X_1, \dots, X_n\} = X_{(n)}$ has cdf and pdf given by

$$F_{X_{(n)}}(x) = \left[F_X(x)\right]^n$$

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

Example: Let $X_1, \dots, X_n$ be iid Uniform $(0,1)$ rvs. Find the density of the k^{th} order statistic $X_{(k)}$.

We have

$$f_X(x) = \mathbb{1}(0 < x < 1) \quad \text{and} \quad F_X(x) = \begin{cases} 1 & x \geq 1 \\ x & 0 < x < 1 \\ 0 & x \leq 0, \end{cases}$$

so that

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{n-k} \mathbb{1}(0<x<1)$$

$$= \frac{\Gamma(k+(n-k+1))}{\Gamma(k)\Gamma(n-k+1)} x^{k-1} (1-x)^{(n-k+1)-1} \mathbb{1}(0<x<1),$$

which we recognize as the Beta$(k, n-k+1)$ distribution.

Moreover, we have (from what we know of the Beta dist)

$$\mathbb{E} X_{(k)} = \frac{k}{k+(n-k+1)} = \frac{k}{n+1}$$

$$\text{Var } X_{(k)} = \frac{k(n-k+1)}{[k+(n-k+1)]^2[k+(n-k+1)+1]} = \frac{k(n-k+1)}{(n+1)^2(n+2)}.$$

<u>**Example:**</u> Let $X_1, \dots, X_n$ be independent rvs with the same distribution as X, where X has the Exponential (λ) distribution.

Find the pdf of $X_{(n)} = \max\{X_1, \dots, X_n\}$.

We have

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda} \mathbb{1}(x>0) \quad \text{and} \quad F_X(x) = \begin{cases} 1 - e^{-x/\lambda}, & x > 0 \\ 0, & x \leq 0, \end{cases}$$

so that the cdf and pdf of $X_{(n)}$ are given by

$$F_{X_{(n)}}(x) = \begin{cases} \left(1 - e^{-x/\lambda}\right)^n, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$f_{X_{(n)}}(x) = n\left(1 - e^{-x/\lambda}\right)^{n-1} \frac{1}{\lambda} e^{-x/\lambda} \mathbb{1}(x>0).$$

We also have an expression for the joint density of two order statistics, which is needed to get, for example, the density of the range of a random sample.

We present the expression in the following theorem, but we omit its proof, as it is considerably more complicated than that of the previous theorem.

Theorem: Let $X_{(1)}, \dots, X_{(n)}$ be the order statistics of a random sample from a population with cdf F_X and pdf f_X. Then for $1 \le j < k \le n$, the joint pdf of $X_{(j)}$ and $X_{(k)}$ is given by

$$f_{(X_{(j)}, X_{(k)})}(u,v) = \frac{n!}{(j-1)!\,(k-1-j)!\,(n-k)!} f_X(u) f_X(v)$$

$$* \underbrace{[F_X(u)]^{j-1}}_{\substack{j-1 \text{ values} \\ < X_{(j)}}} \underbrace{[F_X(v) - F_X(u)]^{k-1-j}}_{\substack{k-1-j \text{ values } < X_{(k)} \\ \text{and } > X_{(j)}}} \underbrace{[1 - F_X(v)]^{n-k}}_{\substack{n-k \text{ values} \\ > X_{(k)}}},$$

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

We can simplify the expression above if we are interested in the joint density of the minimum and maximum, i.e. in $X_{(1)}$ and $X_{(n)}$.

Corollary: Let $X_{(1)}, \dots, X_{(n)}$ be the order statistics of a random sample from a population with cdf F_X and pdf f_X. Then the joint pdf of $X_{(1)}$ and $X_{(n)}$ is given by

$$f_{(X_{(1)}, X_{(n)})}(u,v) = n(n-1) f_X(u) f_X(v) [F_X(v) - F_X(u)]^{n-2},$$

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

Example: Let $X_1, \ldots, X_n$ be a random sample from a population with the Uniform$(0, \theta)$ distribution.

(i) Find the joint pdf of $X_{(1)}$ and $X_{(n)}$.

We have

$$f_X(x) = \frac{1}{\theta} \mathbb{1}(0 < x < \theta) \quad \text{and} \quad F_X(x) = \begin{cases} 1 , & x \geq \theta \\ x/\theta , & 0 < x < \theta \\ 0 , & x \leq 0 , \end{cases}$$

so that

$$f_{(X_{(1)}, X_{(n)})}(u,v) = n(n-1) \frac{1}{\theta^2} \left(\frac{v}{\theta} - \frac{u}{\theta} \right)^{n-2} \mathbb{1}(0 < u < v < \theta)$$

(ii) Find the joint pdf of the rv pair (R, M) where

$$R = X_{(n)} - X_{(1)} \quad \text{(The range)}$$
$$M = X_{(n)} .$$

We have $(R, M) \in \{ (r,m) : 0 < r < \theta , \; r < m < \theta \}$.

We get the inverse transformation

$$\begin{aligned} r &= v - u \; (= g_1(u,v)) \\ m &= v \; (= g_2(u,v)) \end{aligned} \iff \begin{aligned} u &= m - r \; (= g_1^{-1}(r,m)) \\ v &= m \; (= g_2^{-1}(r,m)), \end{aligned}$$

giving the Jacobian

$$J(r,t) = \begin{vmatrix} \frac{\partial}{\partial r} m-r & \frac{\partial}{\partial m} m-r \\ \frac{\partial}{\partial r} m & \frac{\partial}{\partial m} m \end{vmatrix} = \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = -1 .$$

So we have

$$f_{(R,M)}(r,m) = n(n-1) \frac{1}{\theta^2} \left(\frac{m}{\theta} - \frac{m-r}{\theta} \right)^{n-2} |-1| \, \mathbb{1}(0 < r < \theta , \; r < m < \theta)$$

$$= \frac{n(n-1)}{\theta^2}\left(\frac{r}{\theta}\right)^{n-2} \mathbb{1}(0<r<\theta,\ r<m<\theta).$$

(iii) Find the pdf of $R = X_{(n)} - X_{(1)}$.

Integrate joint pdf over m

$$f_R(r) = \int_{-\infty}^{\infty} \frac{n(n-1)}{\theta^2}\left(\frac{r}{\theta}\right)^{n-2} \mathbb{1}(0<r<\theta,\ r<m<\theta)\, dm$$

$$= \int_r^{\theta} \frac{n(n-1)}{\theta^2}\left(\frac{r}{\theta}\right)^{n-2} dm\ \mathbb{1}(0<r<\theta)$$

$$= \frac{n(n-1)}{\theta^2}\left(\frac{r}{\theta}\right)^{n-2}(\theta-r)\ \mathbb{1}(0<r<\theta)$$

$$= \frac{n(n-1)}{\theta}\left(\frac{r}{\theta}\right)^{n-2}\left(1-\frac{r}{\theta}\right) \mathbb{1}(0<r<\theta).$$

* Check to see whether $\int_{-\infty}^{\infty} f_R(r)\,dr = 1$:

$$\int_{-\infty}^{\infty} f_R(r)\,dr = \int_{-\infty}^{\infty} \frac{n(n-1)}{\theta}\left(\frac{r}{\theta}\right)^{n-2}\left(1-\frac{r}{\theta}\right) \mathbb{1}(0<r<\theta)\, dr$$

$$= \int_0^{\theta} \frac{n(n-1)}{\theta}\left(\frac{r}{\theta}\right)^{n-2}\left(1-\frac{r}{\theta}\right) dr$$

$s = \frac{r}{\theta}$, $r = \theta s$, $dr = \theta\, ds$, $\Rightarrow s \in (0,1)$

$$= \int_0^1 \frac{n(n-1)}{\theta} s^{n-2}(1-s)\,\theta\, ds$$

$n(n-1) = \frac{n!}{(n-2)!\,2!} = \frac{\Gamma(n-1+2)}{\Gamma(n-1)\Gamma(2)}$

$$= \underbrace{\int_0^1 \frac{\Gamma(n-1+2)}{\Gamma(n-1)\Gamma(2)} s^{(n-1)-1}(1-s)^{2-1}\, ds}_{\text{Integral over Beta}(n-1,\,2)\text{ pdf}}$$

$$= 1.$$