Lecture 7: Properties of Random Samples

1 Continued From Last Class

Theorem 1.1. *Let $X_1, X_2, ...X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$, then*

a) $\mathbb{E}\overline{X} = \mu$,

b) $Var\overline{X} = \frac{\sigma^2}{n}$,

c) $\mathbb{E}S^2 = \sigma^2$.

Proof. Part **(a)** of the theorem can be simply proved as follows :

$$\mathbb{E}\overline{X} = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}n\mathbb{E}X_1 = \mu. \tag{1}$$

A similar proof can be given for part**(b)** :

$$Var\overline{X} = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}nVarX_1 = \frac{\sigma^2}{n}. \tag{2}$$

From the definition of **sample variance** and using the equation,

$$(n-1)S^2 = \sum_{i\in[n]}(X_i - \overline{X})^2 = \sum_{i\in[n]} X_i^2 - n\overline{X}^2, \tag{3}$$

part **(c)** can be proved as follows:

$$\begin{aligned}
\mathbb{E}S^2 &= \mathbb{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]\right), \\
&= \frac{1}{n-1}(n\mathbb{E}X_1^2 - n\mathbb{E}\overline{X}^2), \\
&= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right), \\
&= \sigma^2.
\end{aligned} \tag{4}$$

□

Theorem 1.2. *Let $X_1, X_2, ...X_n$ be a random sample from a pmf or pdf $f(x|\theta)$, where,*

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)u_i\right)$$

is a member of an exponential family. Define statistics $T_1, T_2, ...T_k$ as,

$$T_i(X_1, X_2.....X_n) = \sum_{j=1}^{n} t_i(X_j),\ \ i = 1, 2....k.$$

If the set $\{w_1(\theta), w_2(\theta), ...w_k(\theta) : \theta \in \Theta\}$ contains an open subset of $\mathbb{R}^k$, then the distribution of $(T_1, ...T_k)$ is an exponential family of the form,

$$f_T(u_1, ..., u_k|\theta) = H(u_1, ...u_k)[c(\theta)]^n \exp\left(\sum_{i=1}^{k} w_i(\theta)u_i\right)$$

Example 1.3 (Sum of Bernoulli Random Variables). Let $X_1, X_2, ...X_n$ be random sample of size n from a Bernoulli distribution. Thus,

$$\begin{aligned} P(X_1, ...X_n|p) &= Bern(p), \\ &= P(X_1|p) = p^{X_1}(1-p)^{1-X_1}, \\ &= (1-p)exp\left(log\left[\frac{p}{1-p}X_1\right]\right). \end{aligned} \tag{5}$$

Comparing with the exponential family equation above, we get $h(X_1) = 1$, $c(p) = 1 - p$ and $w_1(p) = log(\frac{p}{1-p})$.

2 Sampling from Normal distribution

Theorem 2.1. *Let $X_1, ...X_n$ be a random sample from a Normal distribution $\mathcal{N}(\mu, \sigma^2)$ and $\overline{X}$ and S^2 are sample mean and variance respectively. Then,*

a) $\overline{X}$ and S^2 are independent random variables.

b) $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

c) $\frac{(n-1)S^2}{n}$ has a chi-squared distribution with $(n-1)$ degrees of freedom.

Proof. a) Without any loss of generality, we can assume that $\mu = 0$ and $\sigma = 1$. It can be shown that if X_1 and X_2 be two independent random variables, then $U_1 = g_1(X_1)$ and $U_2 = g_2(X_2)$ are also independent random variables where g_1 and g_2 are functions of X_1 and X_2 respectively. Thus we aim to show that $\overline{X}$ and S^2 are functions of independent random vectors. We can write S^2 as a function of $(n-1)$ deviations as follows:

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \\
&= \frac{1}{n-1}\left((X_1-\overline{X})^2+\sum_{i=2}^{n}(X_i-\overline{X})^2\right) \\
&= \frac{1}{n-1}\left(\left[\sum_{i=2}^{n}(X_i-\overline{X})\right]^2+\sum_{i=2}^{n}(X_i-\overline{X})^2\right) \qquad (6)
\end{aligned}$$

The last statement follows from the fact that $\sum_{i=1}^{n}(X_i-\overline{X})=0$. Hence, S^2 can be written as a function of only the $(n-1)$ deviations $(X_2-\overline{X}, X_3-\overline{X},\ldots,X_n-\overline{X})$. We can show that these random variables are independent of $\overline{X}$ and hence prove statement *(a)*. The joint pdf of the sample $X_1, X_2, \ldots, X_n$ is given by

$$f(x_1,\ldots,x_n)=\frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}x_i^2\right] \qquad -\infty < x_i < \infty,\ \forall\ i\in[n] \quad (7)$$

We make the following transformation,

$$\begin{aligned}
y_1 &= \overline{x}, \\
y_2 &= x_2-\overline{x}, \\
&\vdots \\
y_n &= x_n-\overline{x}. \qquad (8)
\end{aligned}$$

This linear transformation has a Jacobian of n and the distribution

$$\begin{aligned}
f(y_1,\ldots,y_n) &= \frac{n}{(2\pi)^{\frac{n}{2}}}\exp\left[-\frac{1}{2}(y_1-\sum_{i=2}^{n}y_i)^2\right]\exp\left[-\frac{1}{2}\sum_{i=2}^{n}(y_i+y_1)^2\right], \qquad -\infty<y_i<\infty, \\
&= \left(\frac{n}{2\pi}\right)^{1/2}\exp\left[\frac{-ny_1^2}{2}\right]\ \frac{n^{1/2}}{(2\pi)^{(n-1)/2}}\exp\left\{-\frac{1}{2}\left[\sum_{i=2}^{n}y_i^2+\left(\sum_{i=2}^{n}y_i\right)^2\right]\right\}. \qquad (9)
\end{aligned}$$

Hence, the joint pdf factors and thus the random variables $Y_1,\ldots,Y_n$ are independent.

b) Consider a random sample $X_1, \ldots, X_n$ obtained from $\mathcal{N}(\mu, \sigma^2)$. The moment generating function (mgf) of X_i, $i \in [n]$ is

$$M_{X_i}(t) = \exp(\mu t + \frac{\sigma^2 t^2}{2}). \tag{10}$$

Hence, for the variable $\frac{X_i}{n}$,the mgf is given by

$$M_{\frac{X_i}{n}}(t) = \exp(\mu \frac{t}{n} + \frac{\sigma^2 t^2}{2n^2}). \tag{11}$$

Now, or the sample mean $\overline{X} = \frac{(X_1+X_2+\cdots+X_n)}{n}$, the mgf is given by

$$\begin{aligned} M_{X_i}(t) &= \left[\exp(\mu \frac{t}{n} + \frac{\sigma^2 t^2}{2n^2})\right]^n, \\ &= \exp(n(\mu \frac{t}{n} + \frac{\sigma^2 t^2}{2n^2})), \\ &= \exp(\mu t + \frac{\sigma^2 t^2}{2n}). \end{aligned} \tag{12}$$

Because the mgf of a distribution is unique to that distribution, this mgf is from a Normal Distribution with mean μ and variance $\frac{\sigma^2}{n}$. Hence, $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. The chi-squared pdf is a special case of the gamma pdf and is given as,

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \qquad 0 < x < \infty. \tag{13}$$

Some properties of the chi squared distribution with p degrees of freedom are summarized in the following lemma.

Lemma 2.2. *Let χ_p^2 denote a chi squared random variable with p degrees of freedom, then,*

(a) If $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi_1^2$, i.e., the square of a standard normal random variable is a chi squared random variable.

(b) If $X_1, X_2 \ldots, X_n$ are independent and $X_i \sim \chi_{p_i}^2$, then $\sum_{i=1}^n X_i \sim X_{\sum_{i=1}^n p_i}$. Thus, independent chi squared variables add to a chi squared variable and their degrees of freedom also add up.

c) To prove part **(c)**, first we prove the recursive relations for sample mean and variance. We know that, sample mean $\overline{X}_{n+1} = \frac{1}{n+1} \sum_{k=1}^{n+1} X_k$. We obtain the recursive relations for sample mean as follows,

$$\begin{aligned}\overline{X}_{n+1} &= \frac{1}{n+1}\sum_{k=1}^{n+1} X_k, \\ &= \frac{1}{n+1}[X_{n+1} + \sum_{k=1}^{n} X_k], \\ &= \frac{1}{n+1}[X_{n+1} + n\overline{X}_n].\end{aligned}$$

Hence the recursive relation for sample mean can be stated as,

$$\overline{X}_{n+1} = \frac{1}{n+1}[X_{n+1} + n\overline{X}_n]. \tag{14}$$

Now we will proceed to derive the recursive relationship for sample variance. For $n+1$, random samples, the sample variance can be stated as,

$$nS^2_{n+1} = \sum_{k=1}^{n+1}[X_k - \overline{X}_{n+1}]^2 \tag{15}$$

Using (14), we have,

$$\begin{aligned}nS^2_{n+1} &= \sum_{k=1}^{n+1}[X_k - \frac{1}{n+1}[X_{n+1} + n\overline{X}_n]]^2, \\ &= \sum_{k=1}^{n+1}[X_k - \frac{1}{n+1}[X_{n+1} + (n+1-1)\overline{X}_n]]^2, \\ &= \sum_{k=1}^{n+1}[X_k - \overline{X_n} - \frac{1}{n+1}[X_{n+1} - \overline{X}_n]]^2, \\ &= \sum_{k=1}^{n+1}[(X_k - \overline{X}_n)^2 + \frac{1}{(n+1)^2}[X_{n+1} - \overline{X}_n]^2 - 2\frac{1}{n+1}[X_{n+1} - \overline{X}_n][X_k - \overline{X}_n]].\end{aligned} \tag{16}$$

Since $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$, we have,

$$\begin{aligned}nS^2_{n+1} &= \sum_{k=1}^{n+1}(X_k - \overline{X_n})^2 + \frac{1}{n+1}[X_{n+1} - \overline{X_n}]^2 - 2\frac{1}{n+1}[X_{n+1} - \overline{X_n}]^2, \\ &= \sum_{k=1}^{n}(X_k - \overline{X_n})^2 + \left[1 - \frac{1}{n+1}\right][X_{n+1} - \overline{X_n}]^2, \\ &= \sum_{k=1}^{n}(X_k - \overline{X_n})^2 + \frac{n}{n+1}[X_{n+1} - \overline{X_n}]^2.\end{aligned} \tag{17}$$

Thus we have,

$$nS_{n+1}^2 = (n-1)S_n^2 + \frac{n}{n+1}[X_{n+1} - \overline{X}_n]^2. \tag{18}$$

Replacing n by $n-1$ in (18), we get a recursive relation for sample variance as,

$$(n-1)S_n^2 = (n-2)S_{n-1}^2 + \frac{n-1}{n}[X_n - \overline{X}_{n-1}]^2. \tag{19}$$

If we take $n = 2$ and use it in (19) and if we define $0 \times S_1^2 = 0$, then from (19), we have $S_2^2 = \frac{1}{2}(X_2 - X_1)^2$.Since the distribution of $\frac{1}{\sqrt{2}}(X_2 - X_1)$ is Gaussian with parameter (0,1), part (a) of lemma 2.2 shows that $S_2^2 \sim \chi_1^2$. Proceeding with induction, let us assume that for $n = k$, $(k-1)S_k^2 \sim \chi_{k-1}^2$.

So for $n = k+1$, we can write from (18),

$$kS_{k+1}^2 = (k-1)S_k^2 + \frac{k}{k+1}[X_{k+1} - \overline{X}_k]^2. \tag{20}$$

By inductive hypothesis, $(k-1)S_k^2 \sim \chi_{k-1}^2$, so if we can establish that $\frac{k}{k+1}\left[X_{k+1} - \overline{X}_k\right]^2 \sim \chi_1^2$ and is independent of S_k^2, then from part (b) of lemma 2.2, $kS_{k+1}^2 \sim \chi_k^2$ and the theorem will be proved.

The vector $(X_{k+1}, \overline{X_k})$ is independent of S_k^2, so is any function of this vector. Furthermore, $(X_{k+1} - \overline{X}_k)$ is a normally distributed random variable with mean 0 and variance,

$$Var(X_{k+1} - \overline{X_k}) = \frac{k+1}{k}.$$

and therefore $\frac{k}{k+1}\left[X_{k+1} - \overline{X}_k\right]^2 \sim \chi_1^2$. This completes our proof of the theorem.

$\square$

3 Order Statistics

Definition 3.1. The order statistics of a random sample $X_1, X_2, \ldots X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, X_{(2)}, \ldots X_{(n)}$.

The order statistics are random variables satisfying $X_{(1)} \leq \cdots \leq X_{(n)}$. In particular,

$$
\begin{aligned}
X_{(1)} &= \min_{1\le i\le n} X_i, \\
X_{(2)} &= \text{second smallest } X_i, \left(\min_{1\le i\le n, X_i\neq X_{(1)}} X_i\right) \\
&\vdots \\
X_{(n)} &= \max_{1\le i\le n} X_i.
\end{aligned}
\tag{21}
$$

Theorem 3.2. *Let f_X be the probability density function associated with the population, then the joint density of order statistics can be written as,*

$$
f_{X_{(1)},X_{(2)},\dots X_{(n)}}(x_1,x_2,\dots x_n) = \begin{cases} n!\prod_{i=1}^{n} f_X(x_i), & \textit{if } \; x_1 < x_2 \dots < x_n, \\ 0, & \textit{otherwise.} \end{cases}
\tag{22}
$$

Remark 1. The term $n!$ comes into this formula, because for any set of values $x_1, x_2 \dots x_n$, there are $n!$ equally likely assignments for these values to $X_1, X_2, \dots X_n$ that all yields the same values of the order statistics.

Definition 3.3. The *sample range*, $R = X_{(n)} - X_{(1)}$ is the distance between the smallest and the largest observations. It is a measure of the dispersion of the sample and should reflect the dispersion in the population.

Definition 3.4. The *sample median*, which we will denote by M, is a number such that approximately one half of the observations are less than M and one half are greater. In terms of order statistics, M can be defined as,

$$
M = \begin{cases} X_{(n+1)/2} & \text{if } n \text{ is odd}, \\ (X_{n/2} + X_{(n/2)+1})/2, & \text{if } n \text{ is even}. \end{cases}
\tag{23}
$$

Definition 3.5. For any number p between 0 and 1, the $(100p)th$ percentile is the observation such that approximately np of the observations are less than this observation and $n(1-p)$ are greater than it. As a special case, for $p = .5$, we have the $50th$ sample percentile, which is nothing but the sample median.

Theorem 3.6. *Let $X_1, X_2, \dots X_n$ be a random sample from a discrete distribution with pmf $f_X(x_i) = p_i$ where $x_1 < x_2 \dots$ are the possible values of X in ascending* *order. We define,*

$$\begin{aligned}
&P_0 = 0, \\
&P_1 = p_1, \\
&P_2 = p_1 + p_2, \\
&\vdots \\
&P_i = p_1 + p_2 \ldots + p_i, \\
&\vdots
\end{aligned} \tag{24}$$

Let $X_{(1)}, X_{(2)}, \ldots X_{(n)}$ be the order statistics from the sample. Then,

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}, \tag{25}$$

and

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} [P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}]. \tag{26}$$

Proof. First we fix i. Let Y be a random variable which counts the number of $X_1, X_2 \ldots, X_n$ which are less than of equal to x_i. For each of $X_1, X_2 \ldots, X_n$, we denote the event $\{X_j \leq x_i\}$ as success and the event $\{X_j > x_i\}$ as failure. So Y can be regarded as the number of successes in n trials. Since $X_1, X_2 \ldots, X_n$ are identically distributed, the probability of success for each trial is a same value, which is P_i. We can write P_i as,

$$P_i = P[X_j \leq x_i]. \tag{27}$$

The success or failure of the j^{th} trial is independent of the outcome of any other trial, since X_j is independent of other X_i's. Thus we can write $Y \sim Bin(n, P_i)$. The event $\{X_j \leq x_i\}$ is equivalent to the event $Y \geq j$; that is, atleast j of the sample values are less than or equal to x_i. Since Y follows a Binomial distribution, we can write,

$$P(Y \geq j) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}. \tag{28}$$

As $P(Y \geq j) = P(X_{(j)} \leq x_i)$, we can write,

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}. \tag{29}$$

This completes the proof of (25). For the proof of (26), we note that,

$$P(X_{(j)} = x_i) = P(X_{(j)} \leq x_i) - P(X_{(j)} \leq x_{i-1}).$$

Hence, we can write using (29),

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} [P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}]. \tag{30}$$

This completes our proof. Here, for the case $i = 1$, $P(X_{(j)} = x_i) = P(X_{(j)} \leq x_i)$. The definition of $P_0 = 0$,takes care of this situation. □

Theorem 3.7. *Let $X_1, X_2, \ldots X_n$ denote the order statistics of a random sample, $X_1, X_2, \ldots X_n$ with cdf $F_x(x)$ and pdf $f_X(x)$. Then the pdf of of X_j is,*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) F_X(x)^{j-1} [1 - F_X(x)]^{n-j}. \tag{31}$$

Proof. We will first find the *cdf* of $X_{(j)}$ and then will differentiate it to get the *pdf*. As in theorem 3.6, let Y be a random variable which counts the number of $X_1, X_2, \ldots X_n$ which are less than or equal to x. Then, if we consider the event $X_j \leq x$ as success, then following the approach for the proof of 3.6, we can write that $Y \sim Bin(n, F_X(x))$. It is to be noted that although $X_1, X_2, \ldots X_n$ are continuous random variables, Y is discrete.

Hence, we have,

$$P(Y \geq j) = \sum_{k=j}^{n} \binom{n}{k} F_X(x)^k (1 - F_X(x))^{n-k}. \tag{32}$$

Since $P(Y \geq j) = P(X_j \leq x_i) = F_{X(j)}(x)$, we will differentiate (32) to obtain the *pdf* of $X_{(j)}$. Thus,

$$f_{X_{(j)}}(x) = \frac{d(F_{X_{(j)}}(x))}{dx}.$$

After differentiating the above expression, it can be written as,

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

$$= \binom{n}{j} j F_X(x)^{j-1} (1 - F_X(x))^{n-j} f_X(x) + \sum_{k=j+1}^{n} \binom{n}{k} k F_X(x)^{k-1} (1 - F_X(x))^{n-k} f_X(x),$$

$$- \sum_{k=j}^{n-1} \binom{n}{k} F_X(x)^k (n-k)(1 - F_X(x))^{n-k-1} f_X(x),$$

$$= \frac{n!}{(j-1)!(n-j)!} f_X(x) F_X(x)^{j-1} [1 - F_X(x)]^{n-j}$$
$$+ \sum_{p=j}^{n-1} \binom{n}{p+1} (p+1) F_X(x)^p (1 - F_X(x))^{n-p-1} f_X(x)$$
$$- \sum_{k=j}^{n-1} \binom{n}{k} F_X(x)^k (n-k) (1 - F_X(x))^{n-k-1} f_X(x).$$

The 1^{st} equality was obtained from the fact that the second term under the summation will be zero when $n = k$ and the 2^{nd} equality followed, when we make the transformation $p = k - 1$. Thus,

$$\begin{aligned} f_{X_{(j)}}(x) = {} & \frac{n!}{(j-1)!(n-j)!} f_X(x) F_X(x)^{j-1} [1 - F_X(x)]^{n-j} \\ & + \sum_{p=j}^{n-1} \binom{n}{p+1} (p+1) F_X(x)^p (1 - F_X(x))^{n-p-1} f_X(x) \\ & - \sum_{k=j}^{n-1} \binom{n}{k} F_X(x)^k (n-k) (1 - F_X(x))^{n-k-1} f_X(x). \end{aligned} \tag{33}$$

Now we utilize the following results,

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

and

$$\binom{n}{k} \times (n-k) = \frac{n!}{(n-k-1)!k!}.$$

Using these above 2 results, we can write (33) as,

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) F_X(x)^{j-1} [1 - F_X(x)]^{n-j}. \tag{34}$$

This completes our proof of the theorem. $\square$