Lecture Note 7 *

Random Sample

MIT 14.30 Spring 2006

Herman Bennett

17 Definitions

17.1 Random Sample

Let $X_1, ..., X_n$ be mutually independent RVs such that $f_{X_i}(x) = f_{X_j}(x)\ \forall\ i \neq j$. Denote $f_{X_i}(x) = f(x)$. Then, the collection $X_1, ..., X_n$ is called a random sample of size n from the population $f(x)$.

Examples:

– Rolling a die n times.

– Selecting 10 MIT students and measuring their height.

- Sampling with and without replacement: Sampling from a large population ("nearly independent").

- Alternatively, this collection (or sampling), $X_1, ..., X_n$, is also called independent and identically distributed random variables with pmf/pdf $f(x)$, or *iid* sample for short.

- Note that the difference between X and x still holds (we continue to deal with random variables).

*Caution: These notes are not necessarily self-explanatory notes. They are to be used as a complement to (and not as a substitute for) the lectures.

17.2 Statistic

Let the RVs $X_1, X_2, ..., X_n$ be a random sample of size n from the population $f(x)$. Then, any real-valued function $T = r(X_1, X_2, ..., X_n)$ is called a statistic.

• Remember that $X_1, X_2, ..., X_n$ are RVs, and therefore T is a RV too, which can take any real value t with pmf/pdf $f_T(t)$.

17.3 Sample Mean

The sample mean, denoted by $\bar{X}_n$, is a statistic defined as the arithmetic average of the values in a random sample of size n.

$$\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{52}$$

17.4 Sample Variance

The sample variance, denoted by S_n^2, is a statistic defined as:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{53}$$

The sample standard deviation is the statistic defined by $S_n = \sqrt{S_n^2}$. [1]

• Remember, the observed value of the statistic is denoted by lowercase letters. So, $\bar{x}, s^2,$ and s denote observed values of the RVs $\bar{X}, S^2,$ and S.

[1]The sample variance and the sample standard deviation are sometimes denoted by $\hat{\sigma}^2$ and $\hat{\sigma}$, respectively.

18 Important Properties of the Sample Mean Distribution and the Sample Variance Distribution

18.1 Mean and Variance of $\bar{X}$ and S^2

Let $X_1, ..., X_n$ be a random sample of size n from a population $f(x)$ with mean μ (finite) and variance σ^2 (finite). Then,

$$E(\bar{X}) = \mu, \qquad E(S^2) = \sigma^2, \qquad Var(\bar{X}) = \frac{\sigma^2}{n}, \qquad \text{and} \qquad Var_{n\to\infty}(S^2) \to 0. \tag{54}$$

- Standard Error: $\sqrt{Var(\bar{X})}$

Example 18.1. Show the first 3 statements of (54).

18.2 The Special Case of a Random Sample from a Normal Population

Let $X_1, ..., X_n$ be a random sample of size n from a $N(\mu, \sigma^2)$ population. Then,

a. $\bar{X}$ and S^2 are independent random variables. (55)

b. $\bar{X}$ has a $N(\mu, \sigma^2/n)$ distribution. (56)

c. $\dfrac{(n-1)S^2}{\sigma^2}$ has a $\chi^2_{(n-1)}$ distribution. (57)

Example 18.2. Show (56).

18.3 Limiting Results ($n \to \infty$)

These concepts are extensively used in econometrics.

18.3.1 (Weak) Law of Large Numbers

Let $X_1, ..., X_n$ be independent and identically distributed (*iid*) random variables with $E(X_i) = \mu$ (finite) and $\text{Var}(X_i) = \sigma^2$ (finite). Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1 \ . \tag{58}$$

This condition is denoted,

$$\bar{X}_n \xrightarrow{p} \mu \qquad (\bar{X}_n \text{ converges in probability to } \mu.) \tag{59}$$

Example 18.3. Prove (58) using Chebyshev's inequality. Note that $S^2 \xrightarrow{p} \sigma^2$ can be proved in a similar way.

18.3.2 Central Limit Theorem (CLT)

Let $X_1, ..., X_n$ be independent and identically distributed (*iid*) random variables with $E(X_i) = \mu$ (finite) and $\text{Var}(X_i) = \sigma^2$ (finite). Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for any value $x \in (-\infty, \infty)$,

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = \Phi(x) \tag{60}$$

Where $\Phi(\ )$ is the cdf of a standard normal.

In words...From (56) we know that if the X_is are normally distributed, the sample mean statistic, $\bar{X}_n$, will also be normally distributed. (60) says that if $n \to \infty$, the function of the sample mean statistic, $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$, will be normally distributed **regardless** of the distribution of the X_is.

In practice(1)...If n is sufficiently large, we can assume the distribution of a function of $\bar{X}_n$, $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$, without knowing the underlining distribution of the random sample $f_{X_i}(x)$. [Very powerful result!]

In practice(2)...Define $Z = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$. If n is sufficiently large, then

$$F_Z\left(\frac{\sqrt{n}(\bar{x}_n - \mu)}{\sigma}\right) \approx \Phi\left(\frac{\sqrt{n}(\bar{x}_n - \mu)}{\sigma}\right) \tag{61}$$

$$\Downarrow$$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \overset{a}{\sim} N(0,1) \quad \text{or} \quad \bar{X}_n \overset{a}{\sim} N(\mu, \sigma^2/n) \qquad (a: \text{ for approximately}) \tag{62}$$

...**regardless** of the pmf/pdf $f_{X_i}(x)$!

- The larger the value of n is, the better the approximation. But, how much is "sufficiently large"? No straight forward rule. It will depend on the underlying distribution $f_{X_i}(x)$. The less bell-shaped $f_{X_i}(x)$ is, the large the n required. Having said this, some authors suggest the following rule of thumb: $n \geq 30$.

- Magnifying glass (see simulations).

Example 18.4. An astronomer is interested in measuring the distance from his observatory to a distant star (in light years). Due to changing atmospheric conditions and measuring errors, each time a measurement is made it will not yield the exact distance. As a result, the astronomer plans to make several measures and then use the average as his estimated distance. He believes that measurement values are *iid* with mean d (the actual distance) and variance 4 (light years). How many measurements does he need to perform to be reasonably sure that his estimated distance is accurate to within ± 0.5 light years?