STAT 511

Lecture 11: Random Samples, Weak Law of Large Numbers and Central Limit Theorem

Devore: Section 5.3-5.5

Prof. Michael Levine

October 15, 2018

Definition of a Statistic

- A statistic is any quantity whose value can be calculated from sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result.
- A statistic is a random variable denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

- **Example** Consider a sample of $n = 3$ cars of a particular type; their fuel efficiencies may be $x_1 = 30.7$ mpg, $x_2 = 29.4$ mpg, $x_3 = 31.1$ mpg.
- It may also be $x_1 = 28.8$ mpg, $x_2 = 30.0$ mpg and $x_3 = 31.1$ mpg
- This implies that the value of the mean $\bar{X}$ is different in these cases. Clearly, $\bar{X}$ is a statistic. The first sample has the mean $\bar{X}_1 = 30.4$ mpg and the second one has $\bar{X}_2 \approx 30$ mpg

Statistic Examples

- A sample mean $\bar{X}$ of the sample $X_1, \ldots, X_n$ is a statistic; $\bar{x}$ is one of its possible values
- The value of the sample mean from any particular sample can be regarded as a *point estimate* of the population μ.
- Another example is the sample standard deviation S, while s is its computed value
- Yet another example is the difference between the sample means for two different populations $\bar{X} - \bar{Y}$

Sampling distribution

- Each statistic is a random variable and, as such, has its own distribution
- Consider two samples of size $n = 2$; if $X_1 = X_2 = 0$, $\bar{X} = 0$ with probability $P(X_1 = 0 \cap X_2 = 0)$
- On the other hand, if $X_1 = 1$ but $X_2 = 0$ or $X_1 = 0$ and $X_2 = 1$, we have $\bar{X} = 0.5$ with probability $P(X_1 = 1 \cap X_2 = 0) + P(X_1 = 0 \cap X_2 = 1)$
- This distribution is called the *sampling distribution* to emphasize its description of how the statistic varies in value across all possible sample

Random Sample

- The probability distribution of any statistic depends on the sampling method.
- Consider selecting a sample of size $n = 2$ from the population $1, 5, 10$. If the sampling is with replacement, it is possible that $X_1 = X_2$; then the sampling variance $S^2 = 0$ with a nonzero probability
- However, the sampling without replacement cannot produce $S^2 = 0$ and, therefore, $P(S^2 = 0) = 0$.

(Simple) Random Sample

- The RVs $X_1, \ldots, X_n$ are said to form a simple random sample of size n if
  - The X_is are independent RVs.
  - Every X_i has the same probability distribution.
- The usual way to describe these two conditions is to say that X_i's are *independent and identically distributed* or *iid*.

Example

- A certain brand of MP3 player comes in three configurations: a model with 2 GB of memory, costing 80, a 4 GB model priced at 100, and an 8 GB version with a price tag of 120
- 20% of all purchasers choose the 2 GB model, 30% choose the 4 GB model, and 50% choose the 8 GB model.
- The probability distribution of the cost X of a single randomly selected MP3 player purchase is given by

x	80	100	120
p(x)	0.2	0.3	0.5

- Here, $\mu 106$ and $\sigma^2 = 244$

Experiment

- On a particular day only two MP3 players are sold. Let $X_1 =$ the revenue from the first sale and $X_2 =$ the revenue from the second
- X_1 and X_2 are independent from the same tabled distribution above

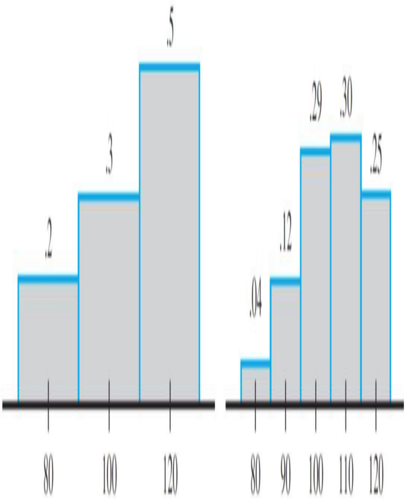
x_1	x_2	$p(x_1, x_2)$	$\bar{x}$	s^2
80	80	.04	80	0
80	100	.06	90	200
80	120	.10	100	800
100	80	.06	90	200
100	100	.09	100	0
100	120	.15	110	200
120	80	.10	100	800
120	100	.15	110	200
120	120	.25	120	0

Complete sampling distributions

$\bar{x}$	80	90	100	110	120
$p_{\bar{X}}(\bar{x})$	0.04	0.12	0.29	0.30	0.25

s^2	0	200	800
$p_{S^2}(s^2)$	0.38	0.42	0.20

Comparison of two histograms



Simulation Experiments

- This is usually employed when the direct derivation is too difficult
- The following characteristics must be specified
  1. The statistic of interest.
  2. The population distribution.
  3. The sample size n.
  4. The number of replications k.

Example

- Consider the platelet volume distribution in individuals with no known heart problems. It is commonly assumed to be normal; particular research publication assumes $\mu = 0.25$ and $\sigma = 0.75$.
- Four experiments are performed, 500 replications each
- In the first experiment, 500 samples of $n = 5$ observations were generated; in the other three sample sizes were $n = 10$, $n = 20$ and $n = 30$, respectively

Distribution of sample mean

- Let $X_1, \dots X_n$ be a random sample from a distribution with mean value μ and standard deviation σ. Then
  1. $E(\bar{X}) = \mu_{\bar{X}} = \mu$
  2. $V(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$
  3. $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- In addition to the above, for the sample total $T = X_1 + X_2 + \dots + X_n$ we have $E\,T = n\mu$, $V(T) = n\sigma^2$ and $\sigma_T = \sqrt{n}\sigma$

Example

- Consider a notched tensile fatigue test on a titanium specimen.
- The expected number of cycles to first acoustic emission (indicates crack initiation) is $\mu = 28,000$. The standard deviation of the number of cycles is $\sigma = 5,000$.
- Let $X_1, \dots, X_{25}$ be a random sample; each X_i is the number of cycles on a different randomly selected specimen
- Then, $E(\bar{X}) = \mu = 28,000$ and the expected total number of cycles for all 25 specimens is $E\,T = n\mu = 700,000$.
- The standard deviations are

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5,000}{\sqrt{25}} = 1000$$

and

$$\sigma_T = \sqrt{n}\sigma = \sqrt{25}(5,000) = 25,000$$

Normal Population Distribution Case

- Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean value μ and standard deviation σ. Then for any n, $\bar{X}$is normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.
- Note that this is true no matter what n is. It need not go to infinity.

Linear combination of the random variables

- Given a collection of n random variables $X_1, \ldots, X_n$ and constants $a_1, \ldots, a_n$, the RV

$$Y = \sum_{i=1}^{n} a_i X_i$$

  is called a **linear combination** of X_i's
- $\bar{X}$ is a special case with $a_1 = \ldots = a_n = \frac{1}{n}$ while the total T is another special case with $a_1 = \ldots = a_n = 1$

Properties of linear combinations of random variables

- Let X_1, X_2,...,X_n be random variables with means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$ respectively.
  1. $E \sum_{i=1}^{n} a_i X_i = \sum_{i=1}^{n} a_i \mu_i$
  2. If $X_1, \ldots, X_n$ are independent, $V(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$

Example

- A gas station sells regular, extra and super gasoline. The prices are 3.00, 3.20 and 3.40 per gallon. Let X_1, X_2, X_3 be the amounts purchased on a particular day (in gallons).
- Let X_1, X_2, X_3 be independent with $\mu_1 = 1000$, $\mu_2 = 500$, $\mu_3 = 300$, $\sigma_1 = 100$, $\sigma_2 = 80$ and $\sigma_3 = 50$.
- The revenue from sales is $Y = 3X_1 + 3.2X_2 + 3.4X_3$
- The average revenue is

$$E\,Y = 3\mu_1 + 3.2\mu_2 + 3.4\mu_3 = 5620$$

- The variation in revenue is

$$\sigma_Y = \sqrt{9\sigma_1^2 + (3.2)^2\sigma_2^2 + (3.4)^2\sigma_3^2} = \sqrt{184,436} = 429.46$$

Example

- The time n it takes a rat of a certain subspecies to reach the end of the maze is normal with mean $\mu = 1.5$ min and $\sigma = .35$ min.
- If we have measurements for 5 rats $X_1, \ldots, X_n$, what is the probability that the total time $T = X_1 + \ldots + X_n$ is between 6 and 8 min?
- Clearly, $T = n\bar{X}$. We know that T is normal with the mean $n\mu = 7.5$ and the variance $n\sigma^2 = .6125$.

- Then,

$$P(6 \le T \le 8) = P\left(\frac{6-7.5}{.783} \le Z \le \frac{8-7.5}{.783}\right)$$
$$= \Phi(0.64) - \Phi(-1.92) = .7115$$

- To find the probability that the average time to reach the maze exit is at most 2.0 min we need to remember that $\bar{X}$ is normal with the mean $\mu_{\bar{X}} = \mu = 1.5$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = .35/\sqrt{5} = .1565$.
- Then,

$$P(\bar{X} \le 2.0) = P\left(Z \le \frac{2.0-1.5}{.1565}\right)$$
$$= P(Z \le 3.19) = \Phi(3.19) = .9993$$

Central Limit Theorem (CLT)

- Let $X_1, \ldots, X_n$ be a random sample from some distribution with mean value μ and variance σ^2. Then, if n is sufficiently large, $\bar{X}$ is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$.
- Note that, unlike the case where the distribution of X itself is normal, this is only *approximately* true. The quality of approximation improves with large n.

Example

- The amount of a particular impurity in a batch of a certain chemical product is a random variable with mean $\mu = 4$ g and standard deviation $\sigma = 1.5$ g.
- What is the approximate probability $P(3.5 < \bar{X} < 3.8)$ if $n = 50$?
- We assume that, approximately, $\bar{X}$ is normal with mean $\mu_{\bar{X}} = 4$ and standard deviation $\sigma_{\bar{X}} = \frac{1.5}{\sqrt{50}} = .2121$
- Then,

$$P(3.5 < \bar{X} < 3.8) \approx P\left(\frac{3.5 - 4.0}{.2121} < \bar{X} < \frac{3.8 - 4.0}{.2121}\right)$$
$$= \Phi(-.94) - \Phi(-2.36) = .1645$$

Remark

- The quality of approximation depends greatly on how close the original distribution of X is to the normal.
- The usual rule of thumb is to use the CLT when $n \geq 30$.