Lecture 9: Sampling Distributions

MSU-STT-351-Sum-19B

Statistics & Their Distributions

Let $X = (X_1, \ldots, X_n)$ be a **random sample** from $F(x|\theta)$, where θ is the unknown parameter. That is, each X_i has the *cdf* $F(x|\theta)$ and X_i's are independent.

(i) A statistic T is any value that can be calculated from sample data, that is, $T = T(X_1, \ldots, X_n)$ is a function of $X_1, \ldots, X_n$. For example, $\bar{X}$ and S^2 are sample statistics.

(ii) A statistic $T(X)$, when takes a real value, is also random variable. For an observed $X = x$, $T(x)$ denotes a numerical value.

(iii) The probability distribution of $T(X)$ is called a sampling distribution.

Random Samples

(i) The distribution of a statistic T calculated from a sample with an arbitrary **joint** distribution can be very difficult.

(ii) Often, we assume that our data is a random sample $X_1, \ldots, X_n$ from a distribution $F(x|\theta)$. This means that (a) The X_i's are independent. (b) All the X_i's have the same probability distribution.

Example 1 (Ex 37): A particular brand of dishwasher soap is solid in three sizes; 25oz, 40oz, and 65oz. Twenty percent of all purchasers select a 25oz box, 50% select a 40oz box, and the remaining 30% choose a 65oz box. Let X_1 and X_2 denote the package sizes selected by two **independently** selected purchasers.

(a) Find the sampling distribution of $\overline{X}$, $E(\overline{X})$, and compare it with μ.
(b) Determine the sampling distribution of the sample variance S^2, calculate $E(S^2)$ and compare to σ^2.

Solution: Note both $X_1, X_2 \in \{25, 40, 65\}$ and have the **same** distribution as that of the rv X with

$$P(X = 25) = .2,\ P(X = 40) = .5,\ P(X = 65) = .3$$

which has the mean $\mu = 44.5$ and the variance $\sigma^2 = 212.25$.

Since X_1 and X_2 are independent, their joint distribution can be found and it is given below. Note
$P(X_1 = 25; X_2 = 25) = P(X_1 = 25)P(X_2 = 25) = 0.2 \times 0.2 = 0.04.$

	$p(x_1)$	0.20	0.50	0.30
$p(x_2)$	$X_2 \vert X_1$	25	40	65
0.20	25	0.04	0.10	0.06
0.50	40	0.10	0.25	0.15
0.30	65	0.06	0.15	0.09

(a) Also, $\overline{X} = \frac{X_1+X_2}{2}$. The distribution of $\overline{X}$ is given below:

$\overline{x}$	25	32.5	40	45	52.5	65
$p(\overline{x})$	0.04	0.20	0.25	0.12	0.30	0.09

The mean of the above distribution is

$$E(\overline{X}) = (25)(.04) + (32.5)(.20) + \ldots + (65)(.09) = 44.5 = \mu.$$

(b) Similarly, the distribution of S^2 based on X_1 and X_2 is is

s^2	0	112.5	312.5	800
$p(s^2)$	0.38	0.20	0.30	0.12

The mean of the distribution of S^2 is

$$E(S^2) = 212.25 = \sigma^2.$$

Example 2 (Ex 40): A box contains ten sealed envelopes numbered $1, \ldots, 10$. The first five contain no money, the next three each contains \$5, and there is a \$10 bill in each of the last two. A sample of size 3 is selected **with replacement** and you get the largest amount of the envelopes selected.

If X_1, X_2 and X_3 denote the amounts in the selected envelopes, the statistic of interest is $M =$ the maximum of X_1, X_2 and X_3.

(a) Obtain the probability distribution of this statistic.

(b) Describe how you would carry out a simulation experiment to compare the distributions of M for various sample sizes. How would you guess the distribution would change as n increases?

Solution:
(a) Possible values of M are: $0, 5, 10$. Note $M = 0$ when all 3 envelopes contain 0 money, hence $P(M = 0) = (0.5)^3 = 0.125$. Also, $M = 10$ when there is **at least one** envelope with \$10. Hence,
$P(M = 10) = 1 - P(\text{no envelopes with } \$10) = 1 - (0.8)^3 = 0.488$.
Finally, $P(M = 5) = 1 - [0.125 + 0.488] = 0.387$.

Thus, we obtain the sampling distribution of M as

m	0	5	10
$p(m)$	0.125	0.387	0.488

An alternative solution would be to list all 27 possible combinations using a tree diagram and computing probabilities directly from the tree.

(b) Let X denote the amount contained in a randomly selected envelope. Its population distribution (also called population distribution) is as follows:

x	0	5	10
p(x)	1/2	3/10	1/5

Write a computer program to generate the digits 0-9 from a **uniform** distribution. Assign a value of 0 to the digits 0-4, a value of 5 to digits 5-7, and a value of 10 to digits 8 and 9. Generate samples of increasing sizes, keeping the number of replications constant and compute M from each sample.

As n, the sample size, increases, $P(M = 0)$ goes to zero, $P(M = 10)$ goes to one. Furthermore, $P(M = 5)$ goes to zero, but at a slower rate than $P(M = 0)$.

Deriving a Sampling Distribution

Example 3: An automobile service charges \$40, \$45 and \$50 for a tune-up of four-, six- and eight-cylinder cars. The revenue (say X) distribution of cars is

x	40	45	50
p(x)	0.2	0.3	0.5

which has $\mu = 46.5$ and $\sigma^2 = 15.25$.

Only two jobs are done in a day. Let X_i =revenue from i-th service, $i = 1, 2$. The distribution of X_1, X_2 is given below:

See the following table for details.

x_1	x_2	$p(x_1, x_2)$	$\overline{x}$	s^2
40	40	0.04	40	0
40	45	0.06	42.5	12.5
40	50	0.10	45	50
45	40	0.06	42.5	12.50
45	45	0.09	45	0
45	50	0.15	47.5	12.5
50	40	0.10	45	50
50	45	0.15	47.5	12.5
50	50	0.25	50	0

Sampling Distributions

The sampling distribution of $\overline{x}$ is

$\overline{x}$	40	42.5	45	47.5	50
$p(\overline{x})$	0.04	0.12	0.29	0.30	0.25

Note $P(\overline{X} = 42.5) = p(42.5) = 0.06 + 0.06$, using the table in the previous page.

Similarly, the sampling distribution of S^2 is

s^2	0	12.5	0
$P_{S^2}(s^2)$	0.38	0.42	0.20

Data from Continuous Distributions

Example 4 (5.21)

Let X_1 and X_2 denote a random sample (service times) of size 2 from exponential distribution with parameter λ (or $G(1, 1/\lambda)$). Then

$$X_1 + X_2 \sim G(2, 1/\lambda).$$

Then the sample mean $\overline{X} = \frac{X_1+X_2}{2} \sim G(2, 1/2\lambda)$ with density

$$f(\overline{x}) = \begin{cases} 4\lambda^2 \overline{x} e^{-2\lambda\overline{x}}, & \text{if } \overline{x} > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Properties of Sample mean and Sample sum

(i) Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean value μ and standard deviation σ. Then

$$E[\overline{x}] = \mu_{\overline{x}} = \mu;$$
$$V(\overline{x}) = \sigma^2_{\overline{x}} = \sigma^2/n;$$
$$\sigma_{\overline{x}} = \sigma/\sqrt{n}.$$

(ii) Let $T_n = X_1 + X_2 + \ldots + X_n$ be the sample total. Then

$$E[T_n] = n\mu;$$
$$V(T_n) = n\sigma^2;$$
$$\sigma T_n = \sqrt{n}\sigma.$$

(iii) If the original distribution of the X_i's is normal, then the distribution of $\overline{X}$ and T_n are also normal.

Central Limit Theorem

(i) Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then for n sufficiency large, $\overline{X} \simeq N(\mu, \sigma^2/n)$.

(ii) Another way of phrasing this is that the distribution of

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1), \quad \text{as } \ n \to \infty.$$

(iii) The larger the value of n, the better the approximation.

(iv) For continuous distributions and reasonably symmetric, the convergence to the normal distribution is good, even for small values of n.

Convergence of means from $U[-1, 1]$ to a normal shape

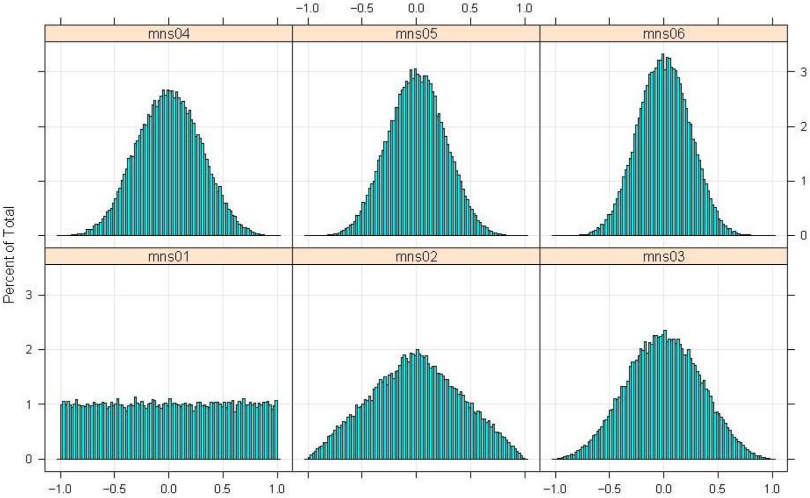
(i) The uniform distribution on the interval $[-1, 1]$ has a mean of 0 and a variance of $1/3$.

(ii) We simulate 50000 replications from the original distribution, mns01, from the distribution of the means of samples of sizes $2, 3, 4, 5$, and 6.

(iii) Histograms of the means of the samples will show convergence to a normal shape and decreasing variance.

(iv) If we multiply the means of samples of size n by $\sqrt{n}$ we can put them all on the same scale to see the convergence to a normal shape.

Histograms of raw means of samples from U[-1,1].



Example 5 (Ex 47):

The inside diameter of a randomly selected piston ring is a **normal** random variable with mean value 12 cm and standard deviation .04cm.

(a) Calculate $P(11.99 \leq \overline{X} \leq 12.01)$ when $n = 16$.

(b) How likely is it that the sample mean diameter exceeds 12.01 when $n = 25$?

Solution Given $\mu = 12$cm $\sigma = 0.04$cm.
(a) For $n = 16$, we have

$$\begin{aligned} P(11.99 \leq \overline{X} \leq 12.01) &= P\Big(\frac{11.99 - 12}{0.01} \leq Z \leq \frac{12.01 - 12}{0.01}\Big) \\ &= P(-1 \leq Z \leq 1) \\ &= \Phi(1) - \Phi(-1) \\ &= 0.8413 - 0.1587. \\ &= 0.6826 \end{aligned}$$

(b) For $n = 25$, we have

$$\begin{aligned} P(\overline{X} > 12.01) &= P\Big(Z > \frac{12.01 - 12}{.04/5}\Big) \\ &= P(Z > 1.25) \\ &= 1 - \Phi(1.25) \\ &= 1 - 0.8944 \\ &= .1056. \end{aligned}$$

Example 6 (Ex 54):

Suppose the sediment density (g/cm) of a randomly selected specimen from a certain region is normally distributed with mean 2.65 and standard deviation .85.

(a) If a random sample of 25 specimens is selected, what is the probability that the sample average sediment density is at most 3.00? Between 2.65 and 3.00?

(b) How large a sample size would be required to ensure that the first probability in part (a) is at least .99?

Solution. It is given that

$$\mu_{\overline{X}} = \mu = 2.65,\ \sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{.85}{5} = 0.17$$

Hence,

$$P(\overline{X} \leq 3.00) = P\Big(Z \leq \tfrac{3.00-2.65}{.17}\Big) = P(Z \leq 2.65) = .9803$$
$$P(2.65 \leq \overline{X} \leq 3.00) = P(\overline{X} \leq 3.00) - P(\overline{X} \leq 2.65) = .4803$$

(b) Since,

$$P(\overline{X} \leq 3.00) = P\Big(Z \leq \frac{3.00 - 2.65}{0.85/\sqrt{n}}\Big) = 0.99$$

we have $\dfrac{0.35}{85/\sqrt{n}} = 2.33$, from which $n = 32.02$.

Thus, $n = 33$ will suffice.

Linear Combinations and their means

(i) For n random variables $X_1, \ldots, X_n$ and n constants $a_1, \ldots, a_n$, the random variable

$$Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

is called a linear combination of the X_i's.

(ii) Whether or not the X_i's are independent,

$$E[a_1 X_1 + \ldots + a_n X_n] = a_1 E[X_1] + \ldots + a_n E[X_n].$$

Variances of linear combinations

(i) If $X_1, \ldots, X_n$ are independent with variances $\sigma_1^2, \ldots, \sigma_n^2$, then

$$V(a_1 X_1 + \ldots + a_n X_n) = a_1^2 V(X_1) + \ldots + a_n^2 V(X_n) = a_1^2 \sigma_1^2 + \ldots + a_n^2 \sigma_n^2.$$

(ii) In general,
$V(a_1 X_1 + \ldots a_n X_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j Cov(X_i, X_j)$, where $Cov(X_i, X_j)$ denotes the covariance between X_i and X_j.

The difference between random variables

(i) Note, $Y = X_1 - X_2$ is a special linear combination with $a_1 = 1,\ a_2 = -1$, and

$$E(X_1 - X_2) = E(X_1) - E(X_2);$$

(ii) When X_1 and X_2 are independent,

$$\begin{aligned} Var(X_1 - X_2) &= a_1^2 Var(X_1) + a_2^2 Var(X_2) \\ &= 1^2 Var(X_1) + (-1)^2 Var(X_2) \\ &= Var(X_1) + Var(X_2). \end{aligned}$$

That is, the variance of the difference is the sum of the variances.

(ii) Remember that "Variances add" in the sense that even when you take the difference of independent random variables, their variances add. But the standard deviations do not add:

$$\sigma_Y = \sqrt{\sigma_1^2 + \sigma_2^2} \neq \sigma_1 + \sigma_2.$$

The Case of Normal Random Variables

If $X_1, \ldots, X_n$ are independent normal $N(\mu_i, \sigma_i)$ variables, then

$$\sum_{i=1}^{n} a_i X_i \sim N\Big(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\Big).$$

Example 7 (Ex 60): Five automobiles of the same type are to be driven on a 300-mile trip. The first two will use an economy brand of gasoline, and the other three will use a name brand. Let X_1, X_2, X_3, X_4 and X_5 be the observed fuel efficiencies (mpg) for the five cars. Suppose these variables are independent and normally distributed with
$\mu_1 = \mu_2 = 20,\ \mu_1 = \mu_2 = \mu_3 = 21$ and $\sigma^2 = 5$ for the economy brand and 3.5 for the name brand. Define on rv Y by

$$Y = \frac{X_1 + X_2}{2} - \frac{X_3 + X_4 + X_5}{3}.$$

So, Y is a measure of the difference in efficiency between economy gas and name-brand gas. Compute $P(Y \geq 0)$ and $P(-1 \leq Y \leq 1)$.

Solution. Note

$$\begin{aligned}\mu_Y &= \frac{1}{2}(\mu_1+\mu_2) - \frac{1}{3}(\mu_3+\mu_4+\mu_5) = -1;\\ \sigma_Y^2 &= \frac{1}{4}\sigma_1^2 + \frac{1}{4}\sigma_2^2 + \frac{1}{9}\sigma_3^2 + \frac{1}{9}\sigma_4^2 + \frac{1}{9}\sigma_5^2 = 3.167;\\ \sigma_Y &= 1.7795.\end{aligned}$$

Thus,

$$\begin{aligned}P(Y \geq 0) &= P\Big(Z \geq \frac{0-(-1)}{1.7795}\Big)\\ &= P(Z \geq 0.56)\\ &= 0.2877.\end{aligned}$$

$$\begin{aligned}P(-1 \leq Y \leq 1) &= P(0 \leq Z \leq \frac{2}{1.7795})\\ &= P(0 \leq Z \leq 1.12)\\ &= 0.3686.\end{aligned}$$

Example 8 (Ex 64): Suppose your waiting time for a bus in the morning is uniformly distributed on $[0, 8]$, whereas waiting time in the evening is uniformly distributed on $[0, 10]$ independent of morning waiting time.
(a) If you take the bus each morning and evening for a week, what is your total expected waiting time?
(b) What is the variance of your total waiting time?
(c) What are the expected value and variance of the difference between morning and evening waiting times on a given day?
(d) What are the expected value and variance of the difference between total morning waiting time and total evening waiting time for a particular week?

Solution

Let $X_1, \ldots, X_5$ denote morning times and $X_6, \ldots, X_{10}$ denote evening times. Then (a)

$$\begin{aligned} E(X_1 + \ldots + X_{10}) &= E(X_1) + \ldots + E(X_{10}) \\ &= 5E(X_1) + 5E(X_6) \\ &= 5(4) + 5(5) = 45 \end{aligned}$$

(b)

$$\begin{aligned} Var(X_1 + \ldots + X_{10}) &= Var(X_1) + \ldots + Var(X_{10}) \\ &= 5Var(X_1) + 5Var(X_6) \\ &= 5\left[\frac{64}{12} + \frac{100}{12}\right] \\ &= \frac{820}{12} = 68.33 \end{aligned}$$

(c) $$E(X_1 - X_6) = E(X_1) - E(X_6) = 4 - 5 = -1.$$

$$Var(X_1 - X_6) = Var(X_1) + Var(X_6) = \frac{64}{12} + \frac{100}{12} = \frac{164}{12} = 13.67.$$

(d) $$E[(X_1 + \ldots + X_5) - (X_6 + \ldots + X_{10})] = 5(4) - 5(5) = -5.$$

$$\begin{aligned} Var\Big[(X_1 + \ldots + X_5) - (X_6 + \ldots + X_{10})\Big] &= Var(X_1 + \ldots + X_5) \\ &\quad + Var(X_6 + \ldots + X_{10}) \\ &= Var(X_1) + \ldots + Var(X_{10}) \\ &= 68.33. \end{aligned}$$

Home Work:

Sect: 5.3 : 39, 42

Sect: 5.4 : 46, 51, 55

Sect: 5.5 : 59, 65, 71, 73