Set 20: Statistics and Their Distributions

Stat 260: July 5, 2024

A ***statistic*** is a function of the data.

- For example, if we observed the velocities (in km/hr) of three cars on a busy highway, we might observe the following data:

$$x_1 = 95 \qquad x_2 = 106 \qquad x_3 = 98$$

  The sample mean $\bar{x} = 99.67$ and sample standard deviation $s = 5.69$ of these three measurements are both statistics.

- If we repeat the experiment with three different cars, we might end up with different data:

$$x_1 = 112 \qquad x_2 = 90 \qquad x_3 = 106$$

  now with $\bar{x} = 102.67$ and sample standard deviation $s = 11.37$.

- The three measurements can take different values. Thus, we can represent them as ***random variables*** X_1, X_2, X_3. Likewise, we can treat the mean $\overline{X}$ of the random variables and their standard deviation S as random variables too.
  - ▷ A ***statistic*** can be any function of random variable(s). Some common statistics include:
    - ◇ The sample mean:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots x_n}{n}$$

    - ◇ The sample sum (total):

$$T = X_1 + X_2 + X_3 + \cdots X_n$$

  - ▷ For specific observed values of statistics, we use lower-case letters:

    $\bar{x} \rightarrow$ specific mean from a specific dataset

    $\overline{X} \rightarrow$ random variable for mean
  - ▷ The probability distribution of a statistic is called ***sampling distribution***.

The random variables $X_1, X_2, \ldots, X_n$ form a **random sample** if all of $X_1, X_2, \ldots, X_n$

- have the same mean and the same variance.
- have the same distribution.
- are independent of one another.

We say $X_1, X_2, \ldots, X_n$ are **independent and identically distributed (iid)**.

If $X_1, X_2, \ldots, X_n$ are iid, with $E[X_i] = \mu$ and $V[X_i] = \sigma^2$ for each $1 \le i \le n$, then

$$E[\overline{X}] = E[x_i] \qquad\qquad V[\overline{X}] =$$

$$V(\bar{x}) = V\left(\frac{x_1 + x_2 + \cdots x_n}{n}\right) = V\left(\frac{x_1}{n} + \frac{x_2}{n} + \cdots \frac{x_n}{n}\right)$$

$$= V\left(\frac{x_1}{n}\right) + V\left(\frac{x_2}{n}\right) + \cdots + V\left(\frac{x_n}{n}\right) \leftarrow \text{because } x_1, x_2 \ldots x_n \text{ are independent}$$

$$= \frac{1}{n^2}V(x_1) + \frac{1}{n^2}V(x_2) + \cdots \frac{1}{n^2}V(x_n)$$

$$= \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n}) = \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n} = \frac{V(x_i)}{n}$$

Fact: If $X_1, X_2, \ldots, X_n$ are all normally distributed random variables, then any linear combination of $X_1, X_2, \ldots, X_n$ is also a normal random variable.

That is, if $X_1, X_2, \ldots, X_n$ are all normal random variables then, $(c_1X_1 + c_2X_2 + \cdots + c_nX_n)$ is a normal random variable.

Example 1: Suppose that X, Y, and W are independent normally distributed random variables with,

$$\mu_X = 8, \qquad \sigma_X = 2, \qquad \mu_Y = 3, \qquad \sigma_y = 1, \qquad \mu_W = -6, \qquad \sigma_W = 3.$$

Determine $P(-2X + 5Y - 3W > 20)$.

$(-2X+5Y-3W) \sim \text{Normal}(\mu = E(-2X+5Y-3W), \sigma^2 = V(-2X+5Y-3W))$

$\mu = E(-2X+5Y-3W) = E(-2x) + E(5y) + E(-3w)$

$= -2E(x) + 5E(y) - 3E(w) = -2(8) + 5(3) - 3(-6) = 17$

$\sigma^2 = V(-2x+5y-3w) = V(-2x) + V(5y) + V(-3w)$ ← since x, y, w are independent

$= (-2)^2 V(x) + (5)^2 V(y) + (-3)^2 V(w)$

$= (-2)^2(2^2) + (5)^2 \cdot (1^2) + (-3)^2 \cdot 3^2 = 122$

$P((-2x+5y-3w) > 20)$

$= P\left(\frac{(-2x+5y-3w) - \mu}{\sigma} > \frac{20-17}{\sqrt{122}}\right)$ $\qquad Z = \frac{x-\mu}{\sigma}$

$= P(Z > 0.27) = 1 - P(Z < 0.27)$

$= 1 - 0.6064 = 0.3936$ → from table

Example 2: Suppose that the amount of time a customer spends at a bank's ATM is normally distributed with a mean of $\mu = 98$ seconds and a standard deviation of $\sigma = 15$ seconds. If a random sample of 48 customers are observed at the ATM, what is the probability that their mean time at the machine is between 92 and 97 second?

$$\left.\begin{array}{l} \text{Let } X_1 = \text{time customer 1 spends at the ATM} \\ \text{Let } X_2 = \text{time customer 2 spends at the ATM} \\ \quad\vdots \\ \text{Let } X_{48} = \text{time customer 48 spends at the ATM} \end{array}\right\} \text{id}$$

$X_i \sim \text{Normal}(\mu = 98, \sigma^2 = 15^2)$

Let $\bar{X}$ = mean time at ATM of 48 customers

$\mu_{\bar{X}} = E[\bar{X}] = E[X_i] = 98$, $\sigma^2_{\bar{X}} = V(\bar{X}) = \frac{V(X_i)}{n} = \frac{15^2}{48}$

$\bar{X} = \frac{1}{48}(X_1 + X_2 + \cdots + X_{48}) \leftarrow$ linear combination of $X_1, X_2 \ldots X_{48}$

$\bar{X} \sim \text{Normal}\left(\mu_{\bar{X}} = 98, \sigma^2 = \frac{15^2}{48}\right)$

$P(92 \le \bar{X} \le 97) = P(\bar{X} \le 97) - P(\bar{X} \le 92)$

$= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \le \frac{97-98}{15/\sqrt{48}}\right) - P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \le \frac{92-98}{15/\sqrt{48}}\right)$

$= P(Z \le -0.46) - P(Z \le -2.77)$

$= 0.3228 - 0.0028 = 0.3200$

Extra Example: In a quality control laboratory for wood pulp fibre, a fibre sample must be tested in three independent machines, each of which has a normally-distributed runtime with a mean of 3.4 minutes and standard deviation of 1.1 minutes. Suppose that it costs \$5 per minute to run the first machine, \$2 per minute to the run the second, and \$1 per minute to run the third.

Determine the probability the cost of testing one fibre sample through the three machines is more than \$30.

Answer: 0.3228

Readings: Swartz 5.5 [EPS 4.3]
Practice problems: EPS 4.13, 4.14, 4.19