Set 20: Section 5.5, Statistics and their Distributions

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

Let $X_1, X_2, \ldots, X_n$ be n random variables. Examples:

- $\bar{X} = \sum_{i=1}^{n} X_i/n$ **is a statistic**
- $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$ **is a statistic**
- $T = X_1 + X_2 + \cdots + X_n$ (the sample sum)
- $X_{max} = max(X_1, X_2, \ldots, X_n)$ (the sample maximum)
- $X_{min} = min(X_1, X_2, \ldots, X_n)$ (the sample minimum)
- $\tilde{X} = median(X_1, X_2, \ldots, X_n)$ (the sample median)

For the observed value of these statistics, we use lower-case letters (e.g. $\overline{x}, t, x_{max}, x_{min}, \tilde{x}$)

Since X's are rv's, statistics are rv's.
The probability distribution of a statistic is called a *sampling distribution*.

Typically, a statistic is a function of a *random sample*, i.e. a sequence of independent, identically distributed (*iid*) random variables, e.g. $X_1, X_2, \ldots, X_n$.

Example: Obtain the distribution of the statistic $Q = X + Y$ where the joint pmf of X and Y is given in the following table.

	X=1	**X=2**	**X=3**
Y=1	0.1	0.1	0.2
Y=2	0.2	0.3	0.1

Example: Suppose two people flip a coin three times. Let X_1, X_2 denote the number of tails flipped by person 1 and 2. Find the sampling distribution of the sample mean.

Since statistics are random variables, they have an expected value and variance.

Example: Find $E(\overline{X})$

Example: Two machines operate independently. The first and second machines make an average of 6 and 3 defective items per hour. Find the sampling distribution of T, the sample total of defective items made by the machines.

We have that $X_1 \sim Poisson(6), X_2 \sim Poisson(3)$. Then $f(x_1, x_2)$ has joint pmf:

$$f(x_1, x_2) = \begin{cases} \frac{6^{x_1} \ e^{-6}}{x_1!} \ \frac{3^{x_2} \ e^{-3}}{x_2!} & x_1 \geq 0, x_2 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

We wish to find $g(t) = P(T = t) = P(X_1 + X_2 = t)$.

If $t < 0$, then $g(t) = 0$.

If $t \geq 0$ then $g(t)$ will equal to the sum of all $f(x_1, x_2)$ where $x_1 + x_2 = t$.

For example, $g(3) = P(T = 3)$ is,

$$f(0,3) + f(1,2) + f(2,1) + f(3,0)$$

$$= \frac{6^0 3^3 e^{-9}}{0!3!} + \frac{6^1 3^2 e^{-9}}{1!2!} + \frac{6^2 3^1 e^{-9}}{2!1!} + \frac{6^3 3^0 e^{-9}}{3!0!} \approx 0.1499$$

In general:

$$g(t) = \sum_{i=0}^{t} f(i, t-i) = \sum_{i=0}^{t} \frac{6^i 3^{t-i} e^{-9}}{i!(t-i)!}$$

Proposition: Linear combinations of normal rv's are normal.

Corollary: Suppose that $X_1, \ldots, X_n$ is a sample from the $\text{Normal}(\mu, \sigma^2)$ distribution. Then

$$\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$$

Example: Determine the distribution of the rv $Y = 2X_1 - X_2 + 3X_3 + 3$ where X_1, X_2 and X_3 are <u>independent</u>, $X_1 \sim \text{Normal}(4, 3)$, $X_2 \sim \text{Normal}(5, 7)$ and $X_3 \sim \text{Normal}(6, 4)$.

Example: Determine the distribution of the rv $Y = X_1 - X_2$ where $\text{Cov}(X_1, X_2) = 6$, $X_1 \sim \text{Normal}(5, 10)$ and $X_2 \sim \text{Normal}(3, 8)$.

Example: When Z_1 and Z_2 are independent standard normal, then $Y = Z_1 + Z_2 \sim \text{Normal}(0, 2)$.

Problem: Suppose that the waiting time for a bus in the morning is uniformly distributed on [0,8] whereas the waiting time for a bus in the evening is uniformly distributed on [0,10]. Assume that the waiting times are independent.

(a) If you take a bus each morning and evening for a week, what is the total expected waiting time?

(b) What is the variance of total waiting time?

(c) What are the expected value and variance of how much longer you wait in the evening than in the morning on a given day?