§5.3: Statistics and Their Distributions

Def: A statistic is a random variable that is a function of other (joint) random variables

$$Z = h(X, Y)$$

Example: Suppose (X, Y) is a joint random variable

The following functions are all "statistics"

- $h(X,Y) = X$
- $h(X,Y) = Y$
- $h(X,Y) = X+Y$
- $h(X,Y) = |X-Y|$
- $h(X,Y) = \min(X,Y)$
- $h(X,Y) = \max(X,Y)$

Previously we computed things like $E[X+Y]$ & $E[X]$

But now we are interested in the full pmf.

→ **Note:** Marginal p.m.f. $p(x,y) \rightsquigarrow p_X(x)$ is the p.m.f. of statistic $h(X,Y) = X$

(p.m.f. of statistic is like a generalization of marginal p.m.f.)

Similar to marginals, p.m.f. of statistic is given by adding probabilities of all points where the statistic has a given value

- p.m.f. of Discrete statistic $Z = h(X,Y)$

$$p_h(z) = \sum_{\substack{x,y \\ h(x,y)=z}} p(x,y)$$

$p_h(z) = P(h(X,Y) = z)$ "Probability that statistic equals z"

- p.d.f. of Continuous statistic $Z = h(X,Y)$

$$f_h(z) = \int_{C_z} f(x,y)\, ds$$

Line integral of $f(x,y)$ along level curve of h $C_z = \{h(x,y) = z\}$

Equal by "Fund. Thm of Line Int."

$$= \frac{d}{dz} \iint_{R_z} f(x,y)\, dA$$

$R_z = \{h(x,y) \le z\}$ the region inside C_z

$\iint_{R_z} f\, dA = P(h \le z)$

$p(x,y)$		X = 0	2	4
Y	1	1/10	1/10	2/10
	3	2/10	3/10	1/10

Example: Joint pmf is:

What is p.m.f. of $X+Y$?

$x+y$	0	2	4
1	0+1	2+1	4+1
3	0+3	2+3	4+3
	$h=3$	$h=5$	$h=7$

Values of statistic $h = X+Y$

p	0	2	4
1	1/10	1/10	2/10
3	2/10	3/10	1/10

Probability values

(Example continues...)

p.m.f. of statistic $h = X+Y$ is total probability for each value of h

p.m.f.:

$h = X+Y$	1	3	5	7
$p(x+y)$	$1/10$	$2/10 + 1/10$	$3/10 + 3/10$	$1/10$

Example: Compute p.m.f. of $|X-Y|$ for same joint (X,Y).

$\|x-y\|$	0	2	4
1	$\|0-1\|$	$\|2-1\|$	$\|4-1\|$
3	$\|0-3\|$	$\|2-3\|$	$\|4-3\|$

$h=3$ $h=1$ $h=1$

$p(x,y)$	0	2	4
1	$1/10$	$1/10$	$3/10$
3	$3/10$	$3/10$	$1/10$

Values of statistic $h = |X-Y|$

Probability values

p.m.f.

$h = \|X-Y\|$	1	3
$p(\|x-y\|)$	$1/10 + 1/10 + 3/10 + 1/10$	$2/10 + 2/10$

Example: Compute p.m.f. of $\min(X,Y)$ for same (X,Y)

(**Note:** $\min(x,y)$ = smaller of x & y

Ex: $\min(1,2) = 1$

$\min(2,0) = 0$)

$\min(x,y)$	0	2	4
1	$\min(0,1)$	$\min(2,1)$	$\min(4,1)$
3	$\min(0,3)$	$\min(2,3)$	$\min(4,3)$

$\min=0$ $\min=2$ $\min=3$ $\min=1$

$p(x,y)$	0	2	4
1	$1/10$	$1/10$	$3/10$
3	$2/10$	$3/10$	$1/10$

Values of statistic $h = \min(X,Y)$

Probability values

p.m.f.

$h = \min(X,Y)$	0	1	2	3
Prob	$1/10 + 2/10$	$1/10 + 2/10$	$3/10$	$1/10$

Note: In practice it is not necessary to write the $h(x,y)$ table ("Values of statistic")

→ For most problems, you can just group together probability values in your head and then directly write the answer...

Our favorite statistics are ones arising from "sampling" a random variable X:

Given a random variable X we can sample (measure) X multiple times, creating many independent copies

$$\begin{cases} X_1 = \text{result of first measurement} \\ X_2 = \text{result of second measurement} \\ \vdots \\ X_n = \text{result of } n^{\text{th}} \text{ measurement} \end{cases}$$

Example: Let X = height of person (know $X \sim$ Normal)

Measure the height of multiple people to get:

$$\begin{cases} X_1 = \text{height of } 1^{\text{st}} \text{ person} \\ X_2 = \text{height of } 2^{\text{nd}} \text{ person} \\ \vdots \\ X_n = \text{height of } n^{\text{th}} \text{ person} \end{cases}$$

All of the X_k have the same distribution as X

This is called being "Identically Distributed"

Usually the measurements are also Independent of each other.

"IID" = "Independent & Identically Distrib."

Important Statistics for IID $(X_1, X_2, \ldots, X_n)$:

- "Sample Mean"
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$
- "Sample Variance"
$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$
- "Sum"
$$T_0 = X_1 + X_2 + \cdots + X_n$$
- "Sum of Squares"
$$T_1 = X_1^2 + X_2^2 + \cdots + X_n^2$$
- "Maximum"
$$X^+ = \max(X_1, X_2, \ldots, X_n)$$
- "Minimum"
$$X^- = \min(X_1, X_2, \ldots, X_n)$$

We can work with continuous distributions, too:

Example: If $X_k \sim \text{Exponential}(\lambda)$ are IID

Then $T_0 = \sum_1^n X_k \sim \text{Gamma}(n, 1/\lambda)$

$\bar{X} = \frac{1}{n} T_0 \sim \text{Gamma}(n, \frac{1}{n\lambda})$

(See "Chapter 4 Summary" Notes)