# Lecture 23

**Definition: The *covariance* between the rvs $X$ and $Y$ is given by**

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}(\ (X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\ ) \\ &= \mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y) \end{aligned}$$

**Interpretation:**

- **positive covariance**
  - **large $x$'s occur with large $y$'s**
  - **small $x$'s occur with small $y$'s**
- **negative covariance**
  - **large $x$'s occur with small $y$'s**
  - **small $x$'s occur with large $y$'s**

**Correlation is the scaled and preferred version of covariance.**

**Definition: The *correlation* between the rvs $X$ and $Y$ is given by**

$$\rho = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{V}(X)}\sqrt{\mathrm{V}(Y)}}$$

**Discussion points:**

- $-1 \leq \mathrm{Corr}(X, Y) \leq 1$
- **correlation is location/scale invariant**
- $\rho$ **is the population analogue of** $r$
- $\rho$ **typically relevant to continuous rvs**
- **if** $a > 0$**, then** $\mathrm{Corr}(X, aX + b) = 1$
- **if** $a < 0$**, then** $\mathrm{Corr}(X, aX + b) = -1$

**Example: Obtain the correlation between $X$ and $Y$ where the joint pmf of $X$ and $Y$ is given in the following table.**

| | **X=1** | **X=2** | **X=3** |
|---|---|---|---|
| **Y=1** | 0.1 | 0.2 | 0.3 |
| **Y=2** | 0.0 | 0.2 | 0.2 |

**Proposition: If $X$ and $Y$ are independent, then**

$$\mathrm{Cov}(X, Y) = 0$$

**In addition,** $\mathrm{Corr}(X, Y) = 0$ **provided** $\mathrm{V}(X)$ **and** $\mathrm{V}(Y)$ **are nonzero. The converse is not true.**

**Also, recall that correlation does not imply causation.**

**Proposition:** $\mathrm{V}(X+Y) = \mathrm{V}(X) + \mathrm{V}(Y) + 2\mathrm{Cov}(X,Y)$

**Proposition: More generally,**

$$\mathrm{V}(aX + bY + c) = a^2\mathrm{V}(X) + b^2\mathrm{V}(Y) + 2ab\mathrm{Cov}(X,Y)$$

**Proposition: Even more generally,**

$$\mathrm{V}\left(\sum_{i=1}^{n} a_i X_i + c\right) = \textstyle\sum_{i=1}^{n} a_i^2 \mathrm{V}(X_i) + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)$$

$$\mathrm{E}\left(\sum_{i=1}^{n} a_i X_i + c\right) = c + \textstyle\sum_{i=1}^{n} a_i \mathrm{E}(X_i)$$

**Lets put some of this stuff together to provide a useful result.**

**Corollary: Suppose that the rv's $X_1, \ldots, X_n$ are a *sample*. In other words, the $X$'s are independent and arise from a common distribution with mean $\mu$ and variance $\sigma^2$. Then the sample mean has the following properties:**

- $\mathrm{E}(\bar{X}) = \mu$
- $\mathrm{V}(\bar{X}) = \sigma^2/n$

**Suprisingly, we have reached this point in our Statistics course and we have not yet defined the word *statistic*.**

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

**Some 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**

**Since data are variable, statistics are also variable. Sometimes we are interested in the distributions of statistics.**

**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 |

**The previous example was simple. To generalize, we need to go a little crazy with notation.**

**Suppose that $X_1, \ldots, X_n$ are discrete with joint pmf $p(x_1, \ldots, x_n)$. Then the pmf for the general statistic $Q(X_1, \ldots, X_n)$ is**

$$p_Q(q) = \sum_A p(x_1, \ldots, x_n)$$

**where the sum is a multiple sum and $A$ is the set of $x_1, \ldots, x_n$ such that $Q(x_1, \ldots, x_n) = q$.**

**Suppose that $X_1, \ldots, X_n$ are continuous with joint pdf $f(x_1, \ldots, x_n)$. Then the cdf for the general statistic $Q(X_1, \ldots, X_n)$ is**

$$F_Q(q) = \mathrm{P}(Q \leq q) = \int_A f(x_1, \ldots, x_n)\ dx_1 \ldots dx_n$$

**where the integral is a multiple integral and $A$ is the set of $x_1, \ldots, x_n$ such that $Q(x_1, \ldots, x_n) \leq q$.**

**I have mentioned previously that statistical practice relies heavily on computation. Here is a simulation procedure that can be used to approximate distributions of statistics when the sums and integrals from the previous page are too difficult to obtain analytically.**

- **Repeat the following two steps $M$ times where $M$ is large and let $i$ denote the $i$-th iteration**
  - **generate $x_1, \ldots, x_n$ according to $p(x_1, \ldots, x_n)$ or $f(x_1, \ldots, x_n)$ (depending whether the data are discrete or continuous)**
  - **calculate $Q_i = Q(x_1, \ldots, x_n)$ for the data**
- **approximate the distbn of Q with a histogram based on generated outcomes $Q_1, \ldots, Q_M$**