Probability Lecture I (August, 2006)

# 1 Probability

Flip a coin once. Assuming the coin we use is a fair coin, the probability of getting a head (H) and a tail (T) on a given toss should be equal (we say H and T are equally likely outcomes). In short,

$$P(\{H\}) = P(\{T\}) \tag{1}$$

where $P$ is the probability function assigning equal probability to heads and tails.

Let $\Omega = \{H, T\}$ .

**Definition 1** *The set, $\Omega$, of all possible outcomes of a particular experiment is called the sample space for the experiment.*

**Definition 2** *A probability function is a function $P$ from subsets of $\Omega$ to the real numbers that satisfies the following axioms:*

1. $P(A) \geq 0$ for any event $A$;
2. If $A_1, A_2, ...$ are pairwise disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$;
3. $P(\Omega) = 1$ where $\Omega$ is the sample space.

It is an immediate result from axiom 2, 3 and equation (1) that $P(\{H\}) = P(\{T\}) = \frac{1}{2}$.

When $\Omega$ is a countable set, i.e., it can be written as $\Omega = \{\omega_1, \omega_{2,...}\}$, and every subset of $\Omega$ is assigned a probability, then, by axiom 2, for any event $A$,

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\}). \tag{2}$$

The coin-toss example above gives a special case in which $\Omega$ is finite, say of size $N$, and all elements in $\Omega$ are equally likely. Then, $P(\{\omega_i\}) = \frac{1}{N}, i = 1, 2, ..., N$,for any element $\omega_i \in \Omega$, and

$$P(A) = \frac{\text{Number of elements in } A}{N}, \text{ where } N \text{ is the total number of elements in } \Omega.$$

The following properties of probability functions are consequences of the axioms above:

1. $P(A^c) = 1 - P(A)$.
2. $P(\emptyset) = 0$.
3. $0 \leq P(A) \leq 1$.
4. If $A \subset B$, then $P(A) \leq P(B)$.
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
6. $P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$.

**Remark 3** *Property 5 gives a useful inequality for the probability of an intersection. Since by axiom 1 we have $P(A \cup B) \leq 1$, it follows from property 5 that $P(A \cap B) \geq P(A) + P(B) - 1$ which is a special case of the Bonferroni's Inequality and is useful for getting a bound for the intersection probability when probabilities for the individual events are known and large . The general version of Bonferroni's Inequality is*

$$\left(\cap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} P(A_i^c). \tag{3}$$

One can check that property 5 is consistent with (3).

**Example 4** *Now, consider another experiment: tossing a fair coin 3 times. The sample space corresponding to this experiment is*

$\Omega = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$

What is the probability of observing $HHH$ ?

# 2 Conditional Probability

Suppose we observe $H$ for the first toss, what is the probability of getting $H$ again on the second toss? We calculate regular probabilities (e.g., $P(\{T\}) = \frac{1}{2}$) with respect to the sample space. But in the case we have some information about the outcomes (e.g., in the experiment in example 4, we might have observed that the outcome of the first toss is $H$), we would like to update the sample space to the given event and calculate the *conditional probabilities* with respect to the given event.

**Definition 5** *If $A$ and $B$ are events in $\Omega$ and $P(B) > 0$ , then the conditional probability of $A$ given $B$, written $P(A|B)$ is defined to be*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4}$$

Note that $P(B|B) = 1$ , and $P(A|B) = 0$ if $A$ and $B$ are disjoint.

A little algebra manipulation on (4) gives a useful formula for calculating intersection probability

$$P(A \cap B) = P(A|B)P(B) \tag{5}$$

## 2.1 Bayes Rule

Suppose $A_1, A_2, ...A_n$ are pairwise disjoint events of positive probability and their union is $\Omega$. Since $B = \cup_{j=1}^{n}(A_i \cap B)$,

$$\begin{aligned}
P(B) &= P(\cup_{j=1}^{n}\{A_i \cap B\}) \\
&= \sum_{j=1}^{n} P(A_i \cap B) \text{ by axiom 2 in definition 2} \\
&= \sum_{j=1}^{n} P(A_j)P(B|A_j) \text{ by (5).} \qquad (6)
\end{aligned}$$

If $P(B)$ is positive, then

$$\begin{aligned} P(A_j|B) &= \frac{P(A_j \cap B)}{P(B)} \\ &= \frac{P(B|A_j)P(A_j)}{\sum_{j=1}^n P(A_j)P(B|A_j)} \text{ by (5) and (6).} \end{aligned}$$

This is the *Bayes rule.*

**Exercise 6** *There are two boxes. Box I has 2 red balls and 3 blue balls, and box II has 4 red balls and 1 blue ball. Suppose I randomly select a box and take a ball at random from that box. If the ball I drew was red, what would be the probability that the ball was from box B?*

# 3 Independent Events

**Definition 7** *Two events are statistically independent if the occurrence of one event has no effect on the probability of another; i.e.,*

$$P(A|B) = P(A) \tag{7}$$

By (5), this means

$$P(A \cap B) = P(A)P(B) \tag{8}$$

which we will use as the definition of *statistical independence.*

The advantage of (8) is that it treats the events symmetrically and will be easier to generalize to more than two events. Events $A_1, A_2, ...A_n$ are mutually independent, if for any subcollection $A_{i_1}, A_{i_2}, ...A_{i_k}$, we have

$$P(\cap_{j=1}^k A_{i_j}) = \Pi_{j=1}^k P(A_{i_j}) \tag{9}$$

Note: $P(\cap_{j=1}^n A_j) = \Pi_{j=1}^n P(A_j)$ is not a sufficient condition for $A_1, A_2, ...A_n$ to be mutually independent. See Appendix for explanations and counter examples.

Since it is apparent that what we get on one toss should not have effect on the probability of the outcomes on any other tosses, the outcome on each toss should be independent of each other. Thus,

$$\begin{aligned} P(\{HHH\}) &= P(H \text{ on 1st toss })P(H \text{ on 2nd toss and } H \text{ on 3rd toss}|H \text{ on 1st toss}) \\ &= P(H \text{ on 1st toss })P(H \text{ on 2nd toss and } H \text{ on 3rd toss}) \\ &= P(H \text{ on 1st toss })P(H \text{ on 2nd toss})P(H \text{ on 3rd toss}|H \text{ on 2nd toss}) \\ &= P(H \text{ on 1st toss })P(H \text{ on 2nd toss})P(H \text{ on 3rd toss}) \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{8} \end{aligned}$$

Similarly, we can see that each of the outcomes in the sample space $\Omega$ has this probability. $\square$

## 3.1 Binomial and Multinomial Distribution

**Example 8** *Consider betting on a series of $n$ independent games, each with winning (success) probability $p$, (and losing probability $1-p$). Let $X$ = the number of successes in $n$ trials. $X$ could be, for example, the number of heads in $n$ tosses of a coin, or the number of times "6" appears when rolling a die $n$ times. We say that $X$ has Binomial $(n,p)$ distribution. Find $P(X=k)$ for $k=0,1,2,...,n$.*

Since the trials are independent, the probability of getting an outcome with successes for the first $k$ trials and failures for the last $n-k$ trials is

$$p^k(1-p)^{n-k},\ k=0,1,2,...,n.$$

But there are $\left(\begin{array}{c} n \\ k \end{array}\right) = \frac{n!}{k!(n-k)!}$ ways that an outcome with $k$ success and $n-k$ failures can happen; (think about how many ways we can arrange an array of $k$ "S" and $(n-k)$ "F"; see discussion in class.)

Thus,

$$P(X=k) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k(1-p)^{n-k},\ k=0,1,2,...,n.$$

**Exercise 9** *A more general case of the Binomial distribution is the multinomial distribution. Suppose instead of only the success and failure outcomes, an experiment could result in one of the q possible outcomes, $\omega_1,...,\omega_q$, each with probability $p_1,...,p_q$, respectively. Let $A_{k_1,...,k_q}$ be the event that exactly $k_i$ $\omega_i$ are observed in $n$ trials, $i=1,...,q$. Derive a formula for $P(A_{k_1,...,k_q})$.*

## 3.2 Hypergeometric Distribution

The binomial distribution could be modeled by putting tickets labeled with "S" and "F", each with proportion $p$ and $1-p$, into a box and drawing $n$ tickets *with replacement* from the box. Since the tickets are taken with replacement, the draws are independent of each other, What about if the tickets were drawn *without replacement*? If $n$ tickets are drawn without replacement from a box of $N \geq n$ tickets labeled either "S", or "F", with the proportion of the tickets with "S" to be $p$ and that with "F" to be $1-p$, the draws are dependent; Let $G$ = the number of "S" observed in $n$ draws, then

$$P(G=k) = \left(\begin{array}{c} n \\ k \end{array}\right) \frac{(Np)_k(N(1-p))_{n-k}}{N_n} = \frac{\left(\begin{array}{c} Np \\ k \end{array}\right)\left(\begin{array}{c} N(1-p) \\ n-k \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)},$$

where $N_n = \left(\begin{array}{c} N \\ n \end{array}\right) n!$ and $\max(0, N(1-p)) \leq k \leq \min(n, Np)$.

The distribution of $G$ is called *hypergeometric* $(D,N,n)$ distribution, where $D=Np$.

# 4 Random Variables and Vectors

Some comments on notations:

1. If $(a_1,b_1),...,(a_k,b_k)$ are $k$ open intervals, we shall call the set $(a_1,b_1) \times ... \times (a_k,b_k) = \{(x_1,...x_k) : a_i < x_i < b_i, 1 \leq i \leq k\}$ an open $k$ rectangle.
2. The Borel field in $R^k$, which we denote by $\mathcal{B}^k$, is defined to be the smallest sigma field having all open $k$ rectangles as members. Any subset of $R^k$ we might conceivably be interested in turns out to be a member of $\mathcal{B}^k$. We will write $R$ for $R^1$ and $\mathcal{B}$ for $\mathcal{B}^1$.

Let $X$ = number of " $H$ " for the experiment as described in example 4. Then, $X$ is a function mapping each outcome in $\Omega$ into a number on the real line as depicted in figure 1.

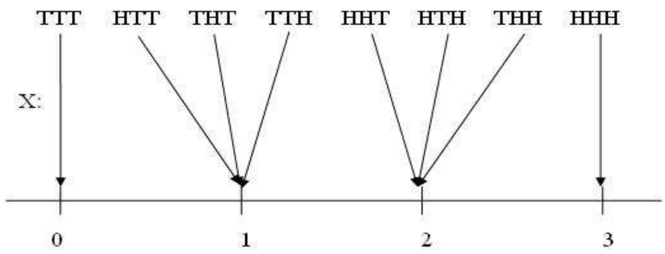


Figure 1: Random Variable X mapping from S to the real line

**Definition 10** *A random variable $X$ is a mapping from a sample space $\Omega$ into the real numbers, and a random vector $\mathbf{X} = (X_1, ..., X_k)^T$ is k-tuple of random variables.*

Let $\chi = \{0, 1, 2, 3\}$ =range of $X$. The probability function $P_X$ associates with $X$ is

$$P_X(X = x_i) = P(\{\omega_i \in \Omega : X(\omega_i) = x_i\}) \; \forall x_i \in \chi. \tag{10}$$

Note that the function $P_X$ is an induced probability function defined in terms of the original function $P$. Because of the equivalence in (10), we will simply write $P_X(X = x_i)$ as $P(X = x_i)$. Also, note that if we know the values returned by the probability function for all $x$, we know the probability distribution of $X$.

Random variables can be categorized into two types: discrete and continuous. While a discrete variable takes values in a countable set, a continuous variable takes values in an uncountable set. Our $X$ defined above for the number of " $H$ " for the experiment in example 4 is a discrete variable since it takes values in the countable set $\chi = \{0, 1, 2, 3\}$.

## 4.1 Distribution Functions

**Definition 11** *Another function that associates with a random variable is called cumulative distribution function (cdf) of $X$, denoted by $F_X$,*

$$F_X(x) = P(X \leq x) \forall x. \tag{11}$$

For example 4, the cdf of $X$ is

| $x$ | $F_X(x)$ |
|---|---|
| $(-\infty, 0)$ | 0 |
| $[0, 1)$ | $\frac{1}{8}$ |
| $[1, 2)$ | $\frac{1}{2}$ |
| $[2, 3)$ | $\frac{7}{8}$ |
| $[3, \infty)$ | 1 |

Table 1: Table for probability distribution function of X

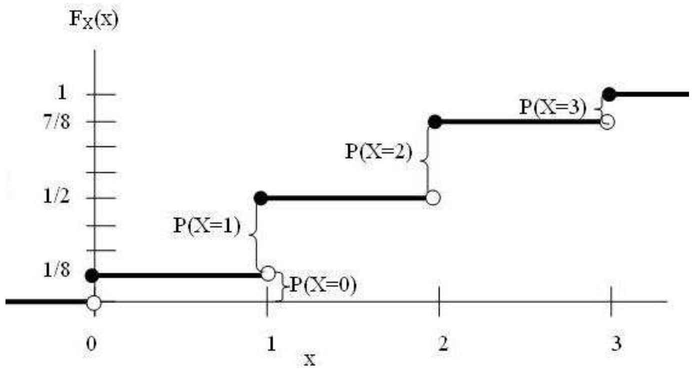


Figure 2: Graph of cdf of X

From figure (2), we see that

$$F_X(x) = P(X \leq x) = \sum_{\{k \in \mathcal{X}: k \leq x\}} P(X = k), \tag{12}$$

and it is not hard to see that this formula holds in general for any discrete variable. We also write $F_x$ as $F(x)$ sometimes for simplicity.

The cdf has the following properties:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$ .
2. $F(x)$ is a nondecreasing function of $x$.
3. $F(x)$ is right-continuous. That is, for every number $x_0$, $\lim_{x \downarrow x_0} F(x) = F(x_0)$ .

These three properties can easily be verified by writing $F$ in terms of the probability function.

#### 4.1.1 Geometric Distribution

**Example 12** *Consider an experiment that consists of tossing a coin until a head appears. Let p = probability of a head on any given toss and define a random variable $X$ =number of tosses required to get a head. Then, for any $x = 1, 2, ...$,*

$$P(X = x) = (1 - p)^{x-1} p, \tag{13}$$

since we must get $x-1$ tails followed by a head for the event to occur and all trials are independent. From (13), we calculate, for any positive integer $x$,

$$\begin{aligned} P(X \le x) &= \sum_{i=1}^{x} P(X = i) \\ &= \sum_{i=1}^{x} (1-p)^{i-1} p, x = 1, 2, ... \end{aligned}$$

Recall that the partial sum of the geometric series is

$$\sum_{k=1}^{n} t^{k-1} = \frac{1-t^n}{1-t}, t \neq 1.$$

Applying this formula to our probability, we find that the cdf of the random variable $X$ is

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= \frac{1-(1-p)^x}{1-(1-p)} p \\ &= 1-(1-p)^x, x = 1, 2, .... \end{aligned}$$

From this expression of the cdf for $X$, we can check that $F_X$ satisfies the properties for cdf above. $F_X(x)$ is the cdf of a distribution called the *geometric distribution.* $\square$

## 4.2 Probability Density(Mass) functions

**Definition 13** *The probability mass function (pmf) $f_X(x)$ of a discrete random variable $X$ is the function given by*

$$f_X(x) = P(X = x)\ \forall x. \tag{14}$$

For our random variable $X$ defined for example 4, the pmf is

| $x$ | $P(X = x)$ |
|---|---|
| 0 | $\frac{1}{8}$ |
| 1 | $\frac{3}{8}$ |
| 2 | $\frac{3}{8}$ |
| 3 | $\frac{1}{8}$ |

Table 2: Table for probability mass function of X

An analogy to the probability mass function of a discrete variable is the probability density function (pdf) of a continuous random variable.

**Definition 14** *The probability density function (pdf) $f_X(x)$ of a continuous random variable $X$ is the function that satisfies*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt\ \forall x. \tag{15}$$

We can think of the cdf as "adding up" the "point probability" $f_X(x)$ to obtain interval probabilities.

**Remark 15** *Recall that for a discrete variable $X$, $F_X(x) = P(X \leq x) = \sum_{\{k \in \mathcal{X}: k \leq x\}} P(X = k)$. For a continuous variable $X$, as the usual, sums becomes integrals. If $X$ has density $f_X(x)$, then $F_X(x)$ is the area from the left up to $x$ under the curve of the density $f_X(x)$.*

Similarly, discrete difference become derivatives,

$$dF_X(x) = F_X(x + dx) - F_X(x) = P(X \in dx) = f(x)dx$$

Thus,

$$f_X(x) = \frac{dF_X(x)}{dx} = F_X^{'}(x).$$

Therefore, if the cdf $F_X(x)$ is everywhere continuous, and differentiable at all except at a finite number of points, then the corresponding distribution has density $f_X(x) = F_X^{'}(x)$.

In summery, we have seen the cdf and pdf (pmf) for continuous (discrete) variables.

| | Discrete variable | Continuous variable |
|---|---|---|
| pmf/pdf | $f_X(x) = P(X = x)\ \forall x.$ | $f_X(x)$ that satisfies $F_X(x) = \int_{-\infty}^{x} f_X(t)dt\ \forall x$ |
| cdf | $F_X(x) = P(X \leq x) = \sum_{\{k \in S: k \leq x\}} P(X = k)\ \forall x$ | $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt\ \forall x$ |

**Definition 16** *Random variables $X$ and $Y$ are identically distributed if they have the same range and for every set $A$ in the range*

$$P(X \in A) = P(Y \in A).$$

The next example shows that two random variables that are identically distributed are not necessarily equal.

**Example 17** *Consider the experiment of tossing a fair coin three times as in example 4. Define the random variables $X$ and $Y$ by*

$$X = \text{number of heads observed}$$
$$Y = \text{number of tails observed}$$

We can verify that $X$ and $Y$ have the same distribution; it is sufficient to check that for each $k = 0, 1, 2, 3$, we have $P(X = k) = P(Y = k)$. However, $Y = 3 - X$ and for no sample points do we have $X(k) = Y(k)$. □

In the example above, $Y$ is a function of $X$ since $Y = g(X)$. In general, let $g$ be any function from $R^k$ to $R^m$, $k, m \geq 1$, such that $g^{-1}(B) = \{y \in R^k : g(y) \in B\} \in \mathcal{B}^k$ for every $B \in \mathcal{B}^m$. Then, the random transformation $Y = g(\mathbf{X})$ is defined by

$$Y = g(\mathbf{X})(\omega) = g(\mathbf{X}(\omega)).$$

Some examples of such a transformation are $g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}$, where $g_1(\mathbf{X}) = \frac{1}{k}\sum_{i=1}^{k} X_i = \overline{X}$, and $g_2(\mathbf{X}) = \frac{1}{k}\sum_{i=1}^{k}(X_i - \overline{X})^2$, and $g(\mathbf{X}) = \begin{pmatrix} \max\{X_i\} \\ \min\{X_i\} \end{pmatrix}$

The probability distribution of $g(\mathbf{X})$ is completely determined by that of $\mathbf{X}$ through

$$P[g(\mathbf{X}) \in B] = P[\mathbf{X} \in g^{-1}(B)].$$

If $\mathbf{X}$ is discrete with mass function $p_X$, then $g(\mathbf{X})$ is discrete and has mass function

$$p_{g(\mathbf{X})}(\mathbf{t}) = \sum_{\{\mathbf{X}: g(\mathbf{X})=\mathbf{t}\}} p_{\mathbf{X}}(\mathbf{X}).$$

**Remark 18** *If two random vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent, so are $g(\mathbf{X})$ and $h(\mathbf{Y})$.*

**Theorem 19** *Suppose that $X \in R$ is continuous with density $p_X$ and $g$ is real-valued and one-to-one on an open set $S$ such that $P[X \in S] = 1$. Furthermore, assume that the derivative $g'$ of $g$ exists and does not vanish on $S$. Then $g(X)$ is continuous with density given by*

$$p_{g(X)}(t) = \frac{p_X(g^{-1}(t))}{|g'(g^{-1}(t))|}$$

*for $t \in g(S)$, and $0$ otherwise. This is called the change of variable formula.*

**Example 20** *If $g(X) = \sigma X + \mu$, where $\sigma \neq 0$, and $X$ is continuous, then*

$$p_{g(X)}(t) = \frac{1}{\sigma} p_X\left(\frac{t-\mu}{\sigma}\right).$$

□

**Exercise 21** *Suppose $X$ has density $f_X(x)$. Find a formula for the density of $Y = X^2$.*

# 5 Expected Values

**Definition 22** *The expected value or mean of a random variable $g(X)$, denoted by $Eg(X)$ is*

$$Eg(X) = \begin{cases} \sum_{x \in X} g(x) f_X(x) = \sum_{x \in X} g(x) P(X = x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{if } X \text{ is continous} \end{cases} \tag{16}$$

provided the integral or sum exists. If $E|g(X)| = \infty$, we say that $Eg(X)$ does not exist.

**Example 23** *Roll a fair die once. Let $X$ =the number shown on the die. Then, by the definition of $E(X)$*

$$E(X) = 1P(X=1) + 2P(X=2) + ... + 6P(X=6) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + ... + 6 \times \frac{1}{6} = 3.5$$

$E(X)$ gives the expected average of the rolls when rolling the fair die a large number of time. □

**Example 24** *Roll a fair die repeatedly. You and I are betting on the number shown on each roll. If the number is 4 or less, you win \$1; otherwise, you pay me \$2.5. Is this a fair game? By "fair", we mean that if we play this game repeatedly for a long time, we can expect that our winnings turn out to be even.*

Let $Y$=your winning. Then

$$\begin{aligned} E(Y) &= 1P(Y=1) + (-2.5)P(Y=-2.5) \\ &= 1P(\text{die shows 1, 2, 3, or 4}) - 2.5P(\text{die shows 5, or 6}) \\ &= 1 \times \frac{4}{6} - 2.5 \times \frac{2}{6} \\ &= -\frac{1}{6} \end{aligned}$$

□

This says that if we play this game many times, on average, you will loss about $\$\frac{1}{6}$ to me per bet. Thus, this is not a fair game. On a fair game, the winnings of the players should turn out even in the long run; i.e., $E(Y) = 0$.

**Properties** There are some properties of the expectation. For random variables $X_1, X_2, ...,$ and $X_n$,

| | | |
|---|---|---|
| Addition Rule | $E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i)$ | however dependent $X_1, X_2, ..., X_n$ are; |
| Multiplication Rule | $E(X_1X_2) = E(X_1)E(X_2)$ | if $X_1$ and $X_2$ are independent; |
| Scaling & Shifting | $E(aX+b) = aEX + b$ | for constants $a$ and $b$. |

## 5.1 Moments

**Definition 25** *The $k^{th}$ moment of $X$ is defined to $E(X^k)$ for $k \in \mathbf{Z}^+$ and $X$ a random variable, provided $E(X^k)$ exists.*

By (16),

$$E(X^k) = \begin{cases} \sum_{x \in X} x^k P(X=x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^k f_X(x)dx & \text{if } X \text{ is continous} \end{cases}$$

**Definition 26** *The $k^{th}$ central moment of $X$ is defined $E[(X-EX)^k]$.*

**Remark 27** *The second central moment of $X$ is called the variance of $X$ and will be introduced next.*

# 6 Variance

Suppose we measure the distance between a random variable $X$ and a constant $b$ by $(X-b)^2$. The closer $b$ is to $X$, the smaller this quantity is. We can now determine the value of $b$ that minimizes $E(X-b)^2$ and this quantity will give us a good predictor of $X$.

We claim that $b = EX$ is the value that minimizes $E(X-b)^2$. This is because, for any value $b$

$$
\begin{aligned}
E[(X-b)^2] &= E[(X-EX+EX-b)^2] \\
&= E[(X-EX)^2] + 2E[(X-EX)(EX-b)] + E[(EX-b)^2] \\
&= E[(X-EX)^2] + 2(EX-b)E(X-EX) + (EX-b)^2 \ \textit{since } (EX-b) \textit{ is a constant} \\
&= E[(X-EX)^2] + (EX-b)^2 \ \textit{since } E(X-EX) = EX - EX \textit{ by the linearity of "}E\textit{"}
\end{aligned}
$$

Since $(EX-b)^2$ is non-negative, $E[(X-b)^2]$ is minimized when $(EX-b)^2 = 0$; i.e., when $EX = b$. □

**Definition 28** *The variance of a random variable $X$ is defined as*

$$\mathrm{Var}(X) = E[(X-EX)^2] = E(X^2) - (EX)^2 \tag{17}$$

which measures the mean squared deviation of $X$ from its expected value $EX$.
Note that the variance of $X$ is finite if and only if the second moment of $X$ is finite.
The standard deviation of $X$, denoted $SD(X)$, is

$$SD(X) = \sqrt{\mathrm{Var}(X)}. \tag{18}$$

**Remark 29** *If $X$ is any random variable with finite mean and variance, the standardized version or $Z$-score of $X$ is the random variable $Z = (X - E(X))/\sqrt{\mathrm{Var}(X)}$. It follows that*

$$E(Z) = 0 \textit{ and } \mathrm{Var}(Z) = 1.$$

If $X_1$ and $X_2$ are random variables, the *product moment* of order $(i,j)$ of $X_1$ and $X_2$ is, by definition, $E(X_1^i X_2^j)$, $i,j \in \mathbf{Z}^+$, and the *central product moment* of order $(i,j)$ of $X_1$ and $X_2$ is $E\left[(X_1 - E(X_1))^i (X_2 - E(X_2))^j\right]$. The central product moment of order $(1,1)$ is called the *covariance* of $X_1$ and $X_2$ and is written as $\mathrm{Cov}(X_1, X_2)$.

$$\mathrm{Cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2). \tag{19}$$

If we put $X_1 = X_2 = X$, we get back the formula for Var as in (17)

The covariance is defined whenever $X_1$ and $X_2$ have finite variances and in that case

$$|\mathrm{Cov}(X_1, X_2)| \leq \sqrt{(\mathrm{Var}\ X_1)\,(\mathrm{Var}\ X_2)}. \tag{20}$$

Correlation of two variables $X_1$ and $X_2$ is defined as

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{(\mathrm{Var}\ X_1)\,(\mathrm{Var}\ X_2)}}. \tag{21}$$

Note that

$$|\mathrm{Corr}(X_1, X_2)| \leq 1 \text{ always.}$$

If $X_1$ and $X_2$ are independent and integrable,

$$\text{Cov}(X_1, X_2) = \text{Corr}(X_1, X_2) = 0 \text{ when } \text{Var}(X_i) > 0, i = 1, 2.$$

However, it is *not* true in general that $X_1$ and $X_2$ that satisfy this (i.e. are *uncorrelated*) need be independent.

**Properties** Some properties of the variance. For random variables $X_1, X_2, ...,$ and $X_n$,

| | | |
|---|---|---|
| Addition Rule | $\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$ | if $X_1, X_2, ...,$ and $X_n$ are independent; |
| Scaling & Shifting | $\text{Var}(aX + b) = a^2\text{Var}(X)$ | for constants $a$ and $b$. |

**Exercise 30** *Let $X_1, X_2, ..., X_n$ be $n$ independent identically distributed (iid) random variables. Define $S_n = \sum_{i=1}^n X_i$, and $\overline{X}_n = S_n/n$ the average. Find $E(S_n), E(\overline{X}_n), \text{Var}(S_n)$, and $\text{Var}(\overline{X}_n)$.*

# 7 Some Named Distributions

## 7.1 More on Binomial Distribution

**Example 31** *(Binomial) Consider betting on a series of $n$ independent games, each with winning (success) probability $p$. Let $I_i$ be the random variable such that* $I_i = \begin{cases} 1; & \textit{if you win (succeed) on the } i^{th} \textit{ bet} \\ 0; & \textit{otherwise} \end{cases}$. *$I_i$ is called the indicator of success on trial $i$. Let $X = \sum_{i=1}^n I_i$; then, $X$ is the number of successes in $n$ trials and $X$ has Binomial $(n, p)$ distribution. Calculate $EX$ and $Var(X)$ by using the $I_i' s$.*

$$EI_i = 1P(I_i = 1) + 0P(I_i = 0) = p$$

$$E(I_i^2) = 1^2P(I_i = 1) + 0^2P(I_i = 0) = p$$

which implies that

$$Var(I_i) = E(I_i^2) - (EI_i)^2 = p - p^2 = p(1 - p) \tag{22}$$

Thus,

$$\begin{aligned}
E(X) &= E(\sum_{i=1}^{n} I_i) \\
&= \sum_{i=1}^{n} E(I_i) \textit{ by the linearity of "E"} \\
&= \sum_{i=1}^{n} E(I_1) \textit{ since } I_1, I_2, ... I_n \textit{ are identically distributed} \\
&= \sum_{i=1}^{n} p \\
&= np
\end{aligned}$$

and

$$\begin{aligned}
Var(X) &= Var(\sum_{i=1}^{n} I_i) \\
&= \sum_{i=1}^{n} Var(I_i) \textit{ since } I_1, I_2, ... I_n \textit{ are independent} \\
&= \sum_{i=1}^{n} Var(I_1) \textit{ since } I_1, I_2, ... I_n \textit{ are indentically distributed} \\
&= np(1-p) \textit{ by (22)}
\end{aligned}$$

□

## 7.2 Poisson Distribution

**Exercise 32** *(Poisson) The Poisson distribution with parameter $\mu$, denoted $Poisson(\mu)$ distribution is the distribution of probabilities $P_\mu(k)$ over $\{0, 1, 2, ...\}$ defined by*

$$P_\mu(k) = e^{-\mu}\mu^k/k! \ (k = 0, 1, 2, ...)$$

show that $EY = \mu$ and $Var(Y) = \mu$ for a random variable $Y$ of $Poisson(\mu)$ distribution.

### 7.3 Exponential Distribution

**Exercise 33** *A random time $T$ has exponential distribution with rate $\lambda$, denoted exponential$(\lambda)$, where $\lambda$ is a positive parameter, if $T$ has probability density*

$$f(t) = \lambda e^{-\lambda t}.$$

Show that $P(T > t) = e^{-\lambda t}$ and $E(T) = SD(T) = \frac{1}{\lambda}$.

**Remark 34** *Memoryless Property of the Exponential Distribution: A positive random variable $T$ has exponential($\lambda$) distribution for some $\lambda > 0$ if and only if $T$ has the memoryless property*

$$P(T > t + s | T > t) = P(T > s); \;\; s \geq 0, t \geq 0$$

This means that given survival to time $t$, the chance of surviving a further time $s$ is the same as the chance of surviving to time $s$ in the first place.

**Binomial$(\mathbf{n}, \mathbf{p})$ & geometric(p) v.s. Poisson$(\lambda t)$ & exponential$(\lambda)$**

A sequence of $n$ independent Bernoulli trials, with probability $p$ of success on each trial, can be characterized in two ways: the number of successes in $n$ trials has binomial$(n, p)$ distribution, and the waiting times between each success and the next are independent with geometric$(p)$ distribution.

Let $n- > \infty$ and $p- > 0$.with $np = \lambda$ a constant Then, the number of arrivals $N(I)$ in a fixed time interval $I$ of length $t$ is Poisson$(\lambda t)$ distributed, and the waiting times between each arrival and the next are independent with exponential$(\lambda)$.

**Exercise 35** *Suppose calls are coming into a telephone exchange at an average rate of 3 per minute, according to a Poisson arrival process. So, for instance, $N(2, 4)$, the number of calls coming in between $t = 2$ and $t = 4$, has Poisson distribution with mean $\lambda(4 - 2) = 3 \times 2 = 6$; and $W_3$, the waiting time between the second and the third calls, has exponential($3$) distribution. Calculate:*

a) The probability that no calls arrive between $t = 0$ and $t = 2$.

b) The probability that the first call after $t = 0$ takes more than 2 minutes to arrive.

c) The probability that no calls arrive between $t = 0$ and $t = 2$ and at most four calls arrive between $t = 2$ and $t = 3$.

d) The probability that the fourth call arrives within 30 seconds of the third.

e) The probability that the first call after $t = 0$ takes less than 20 seconds to arrive, and the waiting time between the first and second calls is more than 3 minutes.

f) The probability that the fifth call takes more than 2 minutes to arrive.

# 8 Appendix

$$P(\cap_{j=1}^{n} A_j) = \Pi_{j=1}^{n} P(A_j)$$

is not a sufficient condition for $A_1, A_2, ...A_n$ to be mutually independent. To see why, consider the following example.

**Example 36** *Toss a coin twice. The table below lists the outcomes in the sample space for the experiment.*

| | | $1^{st}$ | | | Toss | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | 1 | $(1,1)$ | $(1,2)$ | $(1,3)$ | $(1,4)$ | $(1,5)$ | $(1,6)$ |
| | 2 | $(2,1)$ | $(2,2)$ | $(2,3)$ | $(2,4)$ | $(2,5)$ | $(2,6)$ |
| $2^{nd}$ | 3 | $(3,1)$ | $(3,2)$ | $(3,3)$ | $(3,4)$ | $(3,5)$ | $(3,6)$ |
| Toss | 4 | $(4,1)$ | $(4,2)$ | $(4,3)$ | $(4,4)$ | $(4,5)$ | $(4,6)$ |
| | 5 | $(5,1)$ | $(5,2)$ | $(5,3)$ | $(5,4)$ | $(5,5)$ | $(5,6)$ |
| | 6 | $(6,1)$ | $(6,2)$ | $(6,3)$ | $(6,4)$ | $(6,5)$ | $(6,6)$ |

Let $A_1 = \{\text{the sum is even}\} = \{(1,1),(1,3),(1,5),(2,2),(2,4),(2,6),(3,1),(3,3),(3,5),(4,2),(4,4),(4,6),(5,1),(5,3),(5,5),(6,2),(6,4),(6,6),\}$

$A_2 = \{\text{the sum} \geq 10\} = \{(4,6),(5,5),(6,4),(5,6),(6,5),(6,6)\}$

$A_3 = \{\text{the sum is a multiple of } 3\} = \{(1,2),(2,1),(1,5),(2,4),(3,3),(4,2),(5,1),(3,6),(4,5),(5,4),(6,3),(6,6)\}$.

Thus,

$$P(A_1) = \frac{1}{2},\ P(A_2) = \frac{1}{6},\ P(A_3) = \frac{1}{3}.$$

and

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P\{(6,6)\} \\ &= \frac{1}{36} \\ &= \frac{1}{2} \times \frac{1}{6} \times \frac{1}{3} \\ &= P(A_1)P(A_2)P(A_3). \end{aligned}$$

However,

$$\begin{aligned} P(A_2 \cap A_3) &= P\{(6,6)\} \\ &= \frac{1}{36} \\ &\neq P(A_2)P(A_3) \end{aligned}$$

Similarly, one can show that $P(A_1 \cap A_2) \neq P(A_1)P(A_2)$. Thus, the condition $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$ is not strong enough to force the pairwise independence between the events.

The next natural proposal for the sufficient condition for the general definition of the independence might be pairwise independence of the events, $A_1, A_2, ...A_n$. Unfortunately, this condition is still not strong enough. (See Casella and Berger [3])

# References


[1] Peter J. Bickel and Kjell A. Doksum (2001) Mathematical Statistics: Basic Ideas and Selected Topics, Vol. I, 2nd Edition. *Prentice Hall.*

[2] Jim Pitman (1992) Probability, *Springer-Verlag New York, Inc.*

[3] George Casella and Roger L. Berger (1990) Statistical inference, *Brooks/Cole Publishing Company.*