

”All knowledge degenerates into probability.”

David Hume

1 Discrete Probability

Recall the **Binomial Theorem**

The number of distinct permutations of length k that can be formed from a set of n distinct elements, (NOT allowing repetitions), is

$${}_nP_k = \frac{n!}{(n-k)!}$$

The number of ways to permute a set of n (distinct) objects is ${}_nP_n = n!$

The number of distinct combinations of size k that can be formed from a set of n distinct elements, (NOT allowing repetitions), is

$${}_nC_k = \frac{n!}{k!(n-k)!}$$

When order DOES matter, we want a permutation. When order DOES NOT matter, we want a combination.

Example A boy has 2 red, 4 yellow, and 3 green marbles. In how many ways can the boy arrange the marbles in a line if all marbles of the same color are indistinguishable?

Solution First, we observe that there are $(2 + 4 + 3 =) 9$ slots that the marbles can be arranged in. The number of ways the red marbles can be arranged in each of the different slots is $\binom{9}{2} = \frac{9!}{2!7!}$. Then, we have $\binom{7}{4} = \frac{7!}{4!3!}$

ways to arrange the yellow marbles. Then because all the red marbles are identical, we only have one possible arrangement of red marbles to consider. Our final answer is $\frac{9!}{2!7!} \cdot \frac{7!}{4!3!} \cdot 1$.

1.0.1 The Binomial Distribution

Consider a series of independent trials that yield either success or failure. Let $p = P$ (success occurs at any given trial) and assume that p remains consistent from trial to trial. Then

$$P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 1, 2, 3, \dots, n$$

1.1 Bayes' Theorem

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

From this we have

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$\begin{aligned} P(A \cap B \cap C) &= P(D \cap C) \\ &= P(C|D)P(D) \\ &= P(C|A \cap B)P(A \cap B) \\ &= P(C|A \cap B)P(B|A)P(A) \end{aligned}$$

1.2 Independence

Two events A and B are said to be independent if any of the following logically equivalent statements are satisfied:

- $P(A|B) = P(A) \cdot P(B)$

- $P(A) = P(A|B) = P(A|B^C)$
- $P(B) = P(B|A) = P(B|A^C)$ ¹

To generalize the notion of independence to three variables, we first lay out the conditions:

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C) \quad \text{and} \quad \begin{aligned} P(A \cap B) &= P(A) \cdot P(B) \\ P(B \cap C) &= P(B) \cdot P(C) \\ P(C \cap A) &= P(C) \cdot P(A) \end{aligned}$$

However, neither of these conditions imply the other. The single condition that does imply the independence of the events A_1, A_2, \dots, A_n is if for every set of indices i_1, i_2, \dots, i_n between 1 and n inclusive, we have

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \dots \cdot P(A_{i_k})$$

1.3 Combinatorics

1.4 Probability Density and Cumulative Distribution Functions

A probability function P on a set of real numbers S is called continuous if there exists a function $f(t)$ such that for any closed interval $[a, b] \subset S$, $P([a, b]) = \int_a^b f(t)dt$ where $f(t)$ has the following two properties:

- $f(t) \geq 0$ for all t ,
- Satisfy $\int_{-\infty}^{\infty} f(x)dx = 1$

¹Proof: Let $P(A), P(B) \neq 0$ and $P(A|B) = P(A)$. Then

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

Let Y be a function from a sample space S to the real numbers. Y is called a continuous random variable if there exists a real-valued function $f_Y(y)$ such that for any real numbers a and b with $a < b$

$$P(a \leq Y \leq b) = \int_a^b f_Y(y)dy$$

The function $f_Y(y)$ is the probability density function for Y . As in the discrete case, the cumulative distribution function is defined by

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^y f_Y(t)dt = P(\{s \in S | Y(s) \leq y\})$$

We have several theorems about the CDF:

- $P(Y > s) = 1 - F_Y(s)$
- $P(r < Y \leq s) = F_Y(s) - F_Y(r)$
- $\lim_{y \rightarrow \infty} F_Y(y) = 1$
- $\lim_{y \rightarrow -\infty} F_Y(y) = 0$

1.5 The Median

If X is a discrete random variable, the median m is the point for which $P(X < m) = P(X > m)$. In the event that $P(X \leq m) = 0.5$ and $P(X \geq m') = 0.5$, the median is defined to be the arithmetic average, $(m + m')/2$.

If Y is a continuous random variable, its median is the solution to the integral equation $\int_{-\infty}^m f(x)dx = \int_m^{\infty} f(x)dx = 0.5$.

2 Expected Value

Definition (Mean / Expected Value) Let X be a discrete random variable with probability function $p_X(k)$. The **mean** or **expected value** of X

is

$$E(X) = E[X] = \mu = \mu_X = \sum_{\text{all } k} k \cdot p_X(k)$$

Similarly, if Y is a continuous random variable with PDF $f_Y(y)$,

$$E(Y) = \int_{-\infty}^{\infty} y \cdot p_Y(y)$$

Note: we assume that both the sum and the integral converge absolutely:

$$\sum_{\text{all } k} |k| \cdot p_X(k) < \infty \quad \int_{-\infty}^{\infty} |y| \cdot p_Y(y) < \infty$$

If not, then we say that the associated random variable has no finite expected value. We also require absolute convergence because a convergent sum that is not absolutely convergent depends on the order in which the terms are added; order obviously shouldn't be considered when defining an average.

2.1 Expected Value of a Function of a Random Variable

Suppose X is a discrete random variable with PDF $p_X(k)$. Let $g(X)$ be a function of X . Then the expected value of the random variable $g(X)$ is given by

$$E[g(X)] = \sum_{\text{all } k} g(k) \cdot p_X(k)$$

provided that $\sum_{\text{all } k} |g(k)| \cdot p_X(k) < \infty$.

If Y is a continuous random variable with PDF $f_Y(y)$ and $g(Y)$ is a continuous function, then the expected value of the random variable $g(Y)$ is

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y) \cdot p_Y(y)$$

provided that $\int_{-\infty}^{\infty} |y| \cdot p_Y(y) < \infty$.

For any random variable W , $E(aW + b) = aE(W) + b$ where a and b are constants.

3 The Variance

Definition (Variance) The **variance** of a random variable is the expected value of its squared deviations from μ . If X is discrete with PDF $p_X(k)$,

$$\text{var}(X) = \sigma^2 = E[(X - \mu)^2] = \sum_{\text{all } k} (k - \mu)^2 p_X(k)$$

If Y is a continuous random variable with PDF $f_Y(y)$,

$$\text{var}(Y) = \sigma^2 = E[(Y - \mu)^2] = \int_{-\infty}^{\infty} (y - \mu)^2 \cdot f_Y(y) dy$$

If $E(X^2)$ or $E(Y^2)$ is not finite, the variance is undefined.

Let W be any random variable, discrete or continuous, having mean μ for which $E(W^2)$ is finite. Then $\text{var}(W) = \sigma^2 = E(W^2) - \mu^2$.

Let W be any random variable, discrete or continuous, having mean μ for which $E(W^2)$ is finite. Then $\text{var}(aW + b) = a^2 \text{var}(W)$.

3.1 The Standard Deviation

It should be noted that the units for the variance are the square of the units for the random variable. Unfortunately, this causes problems when relating the variance to the sample values. For that reason, we also use the standard deviation σ , which is the square root of the variance. If X is a discrete random variable,

$$\sigma = \sqrt{\sum_{\text{all } k} (k - \mu)^2 p_X(k)}$$

If Y is a continuous random variable,

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (y - \mu)^2 \cdot f_Y(y) dy}$$

4 Probability Density and Cumulative Distribution Functions

5 Joint Distributions

Proposition (Marginal Densities) If $p_{X,Y}(a,b)$ is the joint probability density function for the discrete random variables X and Y , then for any a and b we may compute the single-variable probability density functions for X and Y as $p_X(a) = \sum_y p_{X,Y}(a,y)$ and $p_Y(b) = \sum_x p_{X,Y}(x,b)$

- In general, a probability density function obtained by restricting a given probability distribution to a subset is called a **marginal probability distribution**. The above proposition gives the procedure for computing the marginal probability distribution on the subsets $X = a$ and $Y = b$ (i.e. when one of the random variables is fixed).
- This formula can be extended to an arbitrary number of variables.

6 Independence

Intuitively, we say two random variables X and Y are independent when knowing the value of one gives no additional information about the value of the other. In other words, $P(X = a \mid Y = b) = P(X = a)$ for every value of a and b , which in turn is equivalent to saying $P(X = a, Y = b) = P(X = a) \cdot P(Y = b)$.

In the language of probability density functions, this is the same as saying $P_{X,Y}(a,b) = P_X(a) \cdot P_Y(b)$; in other words, the probability that $X = a$ and $Y = b$ is the product of those two separate events.

Similarly, we can extend this to more than two random variables. The discrete random variables X_1, X_2, \dots, X_n are independent when, for any subsets Y_1, \dots, Y_k of X_i , the joint distribution $p(Y_1, \dots, Y_k)(a_1, \dots, a_k) = p_{Y_1}(a_1) \cdot \dots \cdot p_{Y_k}(a_k)$. This condition directly implies independence.

Proposition (Variance and Independence) If X and Y are independent discrete random variables whose expected values exist, then $E(XY) = E(X) \cdot E(Y)$ and $var(X + Y) = Var(X) + Var(Y)$.

Corollary (Binomial Variance) Let X be the binomially-distributed random variable representing the total number of successes obtained by performing n independent Bernoulli trials each of which has success probability p . Then $E(X) = np$, $var(X) = np(1 - p)$, and $\sigma(X) = \sqrt{np(1 - p)}$.

7 Covariance and Correlation

Definition (Covariance) If X and Y are random variables whose expected values exist and are μ_X and μ_Y respectively, then the covariance of X and Y is defined as $cov(X, Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)] = E(XY) - E(X)E(Y)$

Notes:

- Roughly speaking, the covariance measures how well a change in the value of X relative to its average value correlates with a change in the value of Y relative to its average value.
- If the covariance is large and positive, then when X increases, Y will also tend to increase, and when X decreases, Y will also tend to decrease.
- If X and Y are independent, their covariance is zero. However, the converse does NOT hold.

Properties of the Covariance: If X, Y, Z , are discrete random variables whose expected values exist, then for any a and b we have

- $Cov(X, X) = Var(X)$, $Cov(Y, X) = Cov(X, Y)$
- $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$
- $Cov(aX + b, Y) = a \cdot cov(X, Y)$

- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

Definition (Pearson Correlation) If X and Y are two discrete random variables whose variances exist and are nonzero, then the Pearson correlation between X and Y is defined as $corr(X, Y) = \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}$

8 Continuous Random Variables

Definition (Continuous Probability Density Function) A continuous probability density function $p(x)$ is a piecewise-continuous, nonnegative, real-valued function such that $\int_{-\infty}^{\infty} p(x)dx = 1$.

Note The integral $\int_{-\infty}^{\infty} p(x)dx$ is improper. However, since $p(x)$ is nonnegative, the value of the integral is always well-defined (though it may be ∞).

Definition (Continuous Random Variable) We say that X is a continuous random variable if there exists a continuous probability density function $p(x)$ such that for any interval I on \mathbb{R} , we have $P(X \in I) = \int_I p(x)dx$.

Definition (The Exponential Distribution) The exponential distribution with parameter $\lambda > 0$ is the continuous random variable with probability density function $p(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and is 0 for negative x .

8.0.1 Expected Value, Variance, Standard Deviation

Definition (Expected Value) $E(X) = \int_{-\infty}^{\infty} xp(x) dx$ (assuming that the original integral converges).

Definition (The Variance) When μ exists and is finite, $Var(X) = E[(X - \mu)^2] = E(X^2) - E(X)^2$.

Definition (Standard Deviation) When μ exists and is finite, the standard deviation is $\sigma(X) = \sqrt{var(X)}$.

Proposition (Properties of the Variance) If X is a continuous random

variable and a and b are any real numbers, then $\text{var}(aX + b) = a^2 \cdot \text{var}(X)$ and $\sigma(aX + b) = |a|\sigma(X)$.

Theorem (Chebyshev's Inequality) If X is a random variable with expected value μ and standard deviation σ , $P(|X - \mu| \geq k\sigma) \leq 1/k^2$ for any positive real k .

- This says that the probability that X takes a value at least k standard deviations away from its mean is at most $1/k^2$.
- Chebyshev's inequality applies to both discrete or continuous random variables.
- **Example** For $k = 2$, Chebyshev's inequality says the value of X can be at most 2 standard deviations from the mean at most $1/4$ of the time.

Joint Distributions

Definition : If X_1, X_2, \dots, X_n are continuous random variables, then the function $p_{X_1, X_2, \dots, X_n}(a_1, a_2, \dots, a_n)$ defined on ordered n -tuples of real numbers, such that $\iint_R da_n da_{n-1} \cdots da_1 = P[(X_1, X_2, \dots, X_n) \in R]$ for every region R in n -dimensional space is called the **joint probability density function** of X_1, X_2, \dots, X_n .

Proposition (Marginal Densities) If $p_{X,Y}(a, b)$ is the joint probability density function for the continuous random variables X and Y , then for any a and b we may compute the single-variable probability density functions for X and Y