All of Statistics - Key Concepts

Albert Richard Caputo

November 2018

1 Probability

1.1 Sample Space

A sample space Ω is the set of possible outcomes of an experiment. Points ω in Ω are called sample outcomes, realizations, or elements. Subsets of Ω are called Events.

1.2 Probability Distribution

Definition 1.1. *We will assign a real number $\mathbb{P}(A)$ to every event A, called the probability of A. We also call $\mathbb{P}$ a probability distribution or probability measure if it satisfies the following axioms:*

1. ***Axiom 1**: $\mathbb{P}(A) \geq 0$ for every A.*
2. ***Axiom 2**: $\mathbb{P}(\Omega) = 1$.*
3. ***Axiom 3**: If $A_1, A_2, ...$ are disjoint, then*

$$\mathbb{P}\Big(\cup_{i=1}^{\infty} A_i \Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Lemma 1.2. *For any events A and B,*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Remark 1.3. If Ω is finite and if each outcome is equally likely, then

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

which is called a uniform probability distribution. Also, it is always true that

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^C)$$

where A^C is the complement of A.

For convenience, define $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ to read n choose k as the number of distinct ways of choosing k objects from n.

1.3 Independent Events

Definition 1.4. *Two events A and B are independent if (and only if)*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

and write $A \perp\!\!\!\perp B$. A set of events $\{A_i : i \in I\}$ is independent if

$$\mathbb{P}\left(\cap_{i \in J} A_i \right) = \prod_{i \in J} \mathbb{P}(A_i).$$

Some important things to note about independence:

1. Independence is sometimes assumed and sometimes derived.
2. Disjoint events with positive probability are not independent.

1.4 Conditional Probability

Definition 1.5. *If $\mathbb{P}(B) > 0$, we define the conditional probability of A given B as*

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

In simplest terms, $\mathbb{P}(A|B)$ is the fraction of times A occurs among those in which B occurs.

Some important things to note about conditional probabilitty:

1. $\mathbb{P}(\cdot|B)$ satisfies the axioms of probability for fixed B. In general. $\mathbb{P}(A|\cdot)$ does not satisfy the axioms of probability for fixed A.

2. In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.

3. A and B are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$.

4. $$\begin{aligned}\mathbb{P}(A \cap B) = \mathbb{P}(AB) &= \mathbb{P}(A|B)\mathbb{P}(B) \\ &= \mathbb{P}(B|A)\mathbb{P}(A)\end{aligned}$$

1.5 Bayes' Theorem

Theorem 1.6 (The Law of Total Probability)**.** *Let $A_1, ..., A_k$ be a partition of Ω. Then, for every event B,*

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Proof. Define $C_j = BA_j$ and note that $C_1, ..., C_k$ are disjoin and $B = \cup_{j=1}^{k} C_j$. Hence,

$$\mathbb{P}(B) = \sum_j \mathbb{P}(C_j) = \sum_j \mathbb{P}(BA_j) = \sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)$$

since $\mathbb{P}(BA_j) = \mathbb{P}(B|A_j)\mathbb{P}(A_j)$ by the definition of conditional probability. □

Theorem 1.7 (Bayes' Theorem)**.** *Let $A_1, ..., A_k$ be a partition of Ω such that $\mathbb{P}(A_i) > 0$ for each i. If $\mathbb{P}(B) > 0$ them, for each $i = 1, ..., k$,*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

Remark 1.8. We call $\mathbb{P}(A_i)$ the prior probability of A and $\mathbb{P}(A_i|B)$ the posterior probability of A.

Proof. We apply the definition of conditional probability twice, followed by the law of total probability:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_iB)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

□

2 Random Variables

Definition 2.1. *A random variable is a mapping*

$$X : \Omega \to \mathbb{R}$$

that assigns a real number $X(\omega)$ to each outcome ω.

2.1 Distribution Functions and Proabability Functions

Given a random variable X, we define the cumulative distribution function as follows.

Definition 2.2. *The cumulative distribution function, or CDF, is the function $F_X : \mathbb{R} \to [0,1]$ defined by*

$$F_X(x) = \mathbb{P}(X \leq x)$$

Theorem 2.3. *Let X have CDF F and let Y have CDF G. If $F(x) = G(x)$ for all x, then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all A.*

Theorem 2.4. *A function F mapping the real line to $[0,1]$ is a CDF for probability $\mathbb{P}$ if and only if F satisfies the following conditions:*

1. *F is non-decreasing: $x_1 < x_2$ implies $F(x_1) \leq F(x_2)$.*
2. *F is normalized:*

$$\lim_{x \to -\infty} F(X) = 0$$
$$\lim_{x \to \infty} F(X) = 1$$

3. *F is right-continuous: $F(x) = F(x^+)$ for all x, where*

$$F(x^+) = \lim_{y \to x, y > x} F(y);$$

Proof. Suppose that F is a CDF. Let us show that (3) holds. Let x be a real number and let $y_1, y_2, ...$ be a sequence of numbers such that $y_1 > y_2 > ...$ and $\lim_i y_i = x$. Let $A_i = (-\infty, y_i]$ and let $A = (-\infty, x]$. Note that $A = \cap_{i=1}^{\infty} A_i$ and also note that $A_1 \supset A_2 \supset ....$ Because the events are monotone, $\lim_i \mathbb{P}(A_i) = \mathbb{P}(\cap_i A_i)$. Thus,

$$F(x) = \mathbb{P}(A) = \mathbb{P}\Big(\cap_i A_i\Big) = \lim_i \mathbb{P}(A_i) = \lim_i F(y_i) = F(x^+)$$

Showing (1) and (2) is similar. Proving the other direction requires some deep tools in analysis. □

Definition 2.5. *X is discrete if it takes countably many values $\{x_1, x_2, ...\}$. We define the probability function or probability mass function for X by $f_X(x) = \mathbb{P}(X = x)$.*

Thus, $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ and $\sum_i f_X(x_i) = 1$. Sometimes we write f instead of f_X. The CDF of X is related to f_X by

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i).$$

Definition 2.6. *A random variable X is continuous if there exists a function f_X such that $f_X(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and for every $a \leq b$,*

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx.$$

The function f_X is called a probability density function (PDF). We have that

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

and $f_X(x) = F'_X(x)$ at all points x at which $F_X(x)$ is differentiable.

Lemma 2.7. *Let F be the CDF for a random variable X. Then:*

1. *$\mathbb{P}(X = x) = F(x) - F(x^-)$ where $F(x^-) = \lim_{y \uparrow x} F(y)$*
2. *$\mathbb{P}(x < X \leq y) = F(y) - F(x)$*
3. *$\mathbb{P}(X > x) = 1 - F(x)$*
4. *If X is continuous then*

$$F(b) - F(a) = \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b)$$
$$= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b)$$

Let us now define the inverse CDF (or quantile function).

Definition 2.8. *Let X be a random variable with CDF F. The inverse CDF or quantile function is defined by*

$$F^{-1}(q) = inf\{x : F(x) > q\}$$

for $q \in [0, 1]$. If F is strictly increasing and continuous then $F^{-1}(q)$ is the unique real number x such that $F(x) = q$.

Remark 2.9. Two random variables are equal in distribution, written $X =^d Y$, if $F_X(x) = F_Y(x)$ for all x. This does not mean X and Y are equal. Rather, it means that all probability statements about X and Y will be the same.

2.2 Some Important Discrete Distributions

THE POINT MASS DISTRIBUTION. X has a point mass distribution at a, written $X \sim \delta_a$, if $\mathbb{P}(X = a) = 1$ in which case

$$F(x) = \begin{cases} 0 & x < a \\ 1 & x \geq a. \end{cases}$$

The probability mass function is $f(x) = 1$ for $x = a$ and 0 otherwise.

THE DISCRETE UNIFORM DISTRIBUTION. Let $k > 1$ be a given integer. Suppose that X has a probability mass function given by

$$f(x) = \begin{cases} 1/k & \text{for } x = 1, ..., k \\ 0 & \text{otherwise.} \end{cases}$$

We say that X has a uniform distribution on $\{1, ..., k\}$.

THE BERNOULLI DISTRIBUTION. Let X represent a binary coin flip. Then $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some $p \in [0, 1]$. We say that X has a Bernoulli distribution written $X \sim \text{Bernoulli}(p)$. The probability function is

$$f(x) = p^x(1 - p)^{1-x}$$

for $x \in \{0, 1\}$.

THE BINOMIAL DISTRIBUTION. Suppose we have a coin which falls heads up with probability p for some $0 \leq p \leq 1$. Flip the coin n times and let X be the number of heads. Assume the tosses are independent. Let $f(x) = \mathbb{P}(X = x)$ be the mass function. It can be shown that

$$f(x) = \begin{cases} \binom{n}{x} p^x(1 - p)^{n-x} & \text{for } x = 0, ..., n \\ 0 & \text{otherwise.} \end{cases}$$

A random variable with this mass function is called a Binomial random variable and we write $X \sim \text{Binomial}(n, p)$. If $X_1 \sim \text{Binomial}(n_1, p)$ and $X_2 \sim \text{Binomial}(n_2, p)$ then $X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$.

THE GEOMETRIC DISTRIBUTION. X has a geometric distribution with parameter $p \in (0, 1)$, written $X \sim \text{Geom}(p)$, if

$$\mathbb{P}(X = k) = p(1 - p)^{k-1}, \quad k \geq 1.$$

We have that

$$\sum_{k=1}^{\infty} \mathbb{P}(X = k) = p \sum_{k=1}^{\infty} (1 - p)^k = \frac{p}{1 - (1 - p)} = 1.$$

Think of X as the number of flips needed until the first head when flipping a coin.

THE POISSON DISTRIBUTION. X has a Poisson distribution with parameter λ, written $X \sim \text{Poisson}(\lambda)$ if

$$f(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x \geq 0.$$

Note that

$$\sum_{x=0}^{\infty} f(x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

The Poisson is often used as a model for counts of rare events. If $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

2.3 Some Important Continuous Distributions

THE UNIFORM DISTRIBUTION. X has a Uniform(a, b) distribution, written $X \sim$ Uniform(a, b), if

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

where $a < b$. The distribution function is

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b. \end{cases}$$

NORMAL (GAUSSIAN). X has a Normal (Gaussian) distribution with parameters μ and σ, denoted by $X \sim N(\mu, \sigma^2)$, if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\big(-\frac{(x-\mu)^2}{2\sigma}\big), \quad x \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. The parameter μ is the center, or mean, of the distribution and σ is the spread, or standard deviation, of the distribution. We say that X has a standard Normal distribution if $\mu = 0$ and $\sigma = 1$. Traditionally, a standard normal random variable is denoted by Z. The PDF and CDF of a standard normal are denoted by $\phi(z)$ and $\Phi(z)$, respectively. Some useful facts:

1. If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.
2. If $Z \sim N(0, 1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.
3. If $X_i = N(\mu_i, \sigma_i^2), i = 1, ..., n$ are independent, then

$$\sum_{i=1}^{n} X_i \sim N\big(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\big).$$

It follows from (1) that if $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} \mathbb{P}(a < X < b) &= \mathbb{P}(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) \\ &= \Phi(\frac{a-\mu}{\sigma}) - \Phi(\frac{b-\mu}{\sigma}). \end{aligned}$$

Thus we can compute any probabilities we want so long as we can compute the CDF $\Phi(z)$ of the standard Normal.

THE EXPONENTIAL DISTRIBUTION. X has an Exponential Distribution with parameter β, denoted by $X \sim$ Exp(β), if

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \quad x > 0.$$

The exponential distribution is used to model the lifetimes of electronic components and the waiting times between rare events.

THE GAMMA DISTRIBUTION. For $\alpha > 0$, the Gamma function is defined by $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$. X has a Gamma distribution with parameters α and β, denoted by $X \sim$ Gamma(α, β), if

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad x > 0$$

where $\alpha, \beta > 0$. The exponential distribution is just a Gamma$(1, \beta)$ distribution. If $X_i \sim$ Gamma(α_i, β) are independent, then $\sum_{i=1}^n X_i \sim$ Gamma(α_i, β).

t AND CAUCHY DISTRIBUTION. X has a t distribution with v degrees of freedom, written $X \sim t_v$ if

$$f(x) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \frac{1}{(1 + \frac{x^2}{v})^{\frac{v+1}{2}}}.$$

The t distribution is similar to a Normal but it has thicker tails. In fact, the Normal corresponds to a t with $v = \infty$. The Cauchy distribution is a special case of the t distribution corresponding to $v = 1$. The density is

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

THE χ^2 DISTRIBUTION. X has a χ^2 distribution with p degrees of freedom, written $X \sim \chi^2_p$, if

$$f(x) = \frac{1}{\Gamma(\frac{p}{2})2^{\frac{p}{2}}} x^{\frac{p}{2}-1} e^{-\frac{x}{2}}, \quad x > 0.$$

If $Z_1, ..., Z_p$ are independent standard Normal random variables then $\sum_{i=1}^p Z_i^2 \sim \chi^2_p$.

2.4 Bivariate Distributions

Given a pair of discrete random variables X and Y, define the join mass function by $f(x, y) = \mathbb{P}(X = x \text{ and } Y = y) = \mathbb{P}(X = x, Y = y)$.

In the continuous case, we call a function $f(x, y)$ a PDF for the random variables (X, Y) if

1. $f(x, y) \geq 0$ for all (x, y),
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ and,
3. for any set $A \subset \mathbb{R} \times \mathbb{R}$, $\mathbb{P}((X, Y) \in A) = \int \int_A f(x, y) dx dy$.

2.5 Marginal Distributions

Definition 2.10. *If (X, Y) have joint distribution with mass function $f_{X,Y}$, then the marginal mass function for X is defined by*

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f(x, y)$$

and the marginal mass function for Y is defined by

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y) = \sum_x f(x, y).$$

For continuous random variables, the marginal densities are

$$f_X(x) = \int f(x, y) dy$$

and

$$f_Y(y) = \int f(x, y) dx.$$

The corresponding marginal distribution functions are denoted by F_X and F_Y.

2.6 Independent Random Variables

Definition 2.11. *Two random variables X and Y are independent if, for every A and B,*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

and we write $X \perp\!\!\!\perp Y$. Otherwise we say X and Y are dependent.

Theorem 2.12. *Let X and Y have joint PDF $f_{X,Y}$. Then $X \perp\!\!\!\perp Y$ if and only if $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ for all values of x and y.*

2.7 Conditional Distributions

Definition 2.13. *The conditional probability mass function is*

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

if $f_Y(y) > 0$.

For continuous random variables, the conditional probability density function is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

assuming that $f_Y(y) > 0$. *Then,*

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx.$$

2.8 Multivariate Distributions and IID Samples

Let $X = (X_1, ..., X_n)$ where $X_1, ...X_n$ are random variables. We call X a random vector. Let $f(x_1, ..., x_n)$ denote the PDF. It is possible to define their marginals, conditionals, etc. much the same way as in the bivariate case. We say that $X_1, ..., X_n$ are independent if, for every $A_1, ..., A_n$,

$$\mathbb{P}(X_1 \in A_1, ..., X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

It suffices to check that

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

Definition 2.14. *If* $X_1, ..., X_n$ *are independent and each has the same marginal distribution with CDF* F*, we say that* $X_1, ..., X_n$ *are IID (independent and identically distributed) and we write*

$$X_1, ..., X_n \sim F.$$

If F *has density* f *we also write* $X_1, ..., X_n \sim f$. *We also call* $X_1, ..., X_n$ *a random sample of size* n *from* F.

See All of Statistics p.39-p.40 for some important multivariate distributions.

2.9 Transformations of Random Variables

Suppose that X is a random variable with PDF f_X and CDF F_X. Let $Y = r(X)$ be a function of X.

Three steps for transformations:

1. For each y, find the set $A_y = \{x : r(x) \leq y\}$.
2. Find the CDF
$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) \\ &= \mathbb{P}(\{x; r(x) \leq y\}) \\ &= \int_{A_y} f_X(x)dx. \end{aligned}$$
3. The PDF $f_Y(y) = F'_Y(y)$.

When r is strictly monotone increasing or strictly monotone decreasing then r has an inverse $s = r^{-1}$ and in this case one can show that

$$f_Y(y) = f_X(s(y))\left|\frac{ds(y)}{dy}\right|$$

2.10 Transformations of Several Random Variables

Let $Z = r(X, Y)$ be the function of interest. Three steps for transformations:

1. For each z, find the set $A_y = \{(x, y) : r(x, y) \le z\}$.
2. Find the CDF

$$\begin{aligned} F_Z(y) &= \mathbb{P}(Z \le z) = \mathbb{P}(r(X, Y) \le z) \\ &= \mathbb{P}(\{(x, y); r(x, y) \le z\}) \\ &= \int\int_{A_y} f_{X,Y}(x, y) dx dy. \end{aligned}$$

3. The PDF $f_Z(z) = F'_Z(z)$.