Higher Theory of Statistics
Math2901 UNSW

Jeremy Le *

2023T2

Contents

*With some inspiration from Hussain Nawaz's Notes

1 Probability

1.1 Experiment, Sample Space, Event

Experiment An experiment is any process leading to recorded observations.

Outcome An outcome is a possible result of an experiment.

Sample Space The set Ω of all possible outcomes is the sample space of an experiment. Ω is discrete if it contains a countable (finite or countably infinite) number of outcomes.

Events An event is a set of outcomes, i.e. a subset of Ω. An event occurs if the result of the experiment is one of the outcomes in that event.

Mutual Exclusion Events are mutually exclusive (disjoint) if they have no outcomes in common.

Set Operations If you have trouble remembering the above rules, then one can essentially replace $\cup$ by multiplication and $\cap$ by addition.
(The associative law) If A, B, C are sets then

$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$

(Distributive Law) If A, B, C are sets then

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

1.2 Sigma-algebra

The $\sigma-$algebra must be defined for rigorously working with probability. The $\sigma-$algebra can be thought of as the family of all possible events in a sample space. Analogously, this may be conceptualised as the power set of the sample space.

Probability A probability is a set function, which is usually denoted by $\mathbb{P}$, that maps events from the $\sigma-$algebra to $[0, 1]$ and satisfies certain properties.

Probability Space The triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is the probability/sample space where

- Ω is the sample space,
- $\mathcal{A}$ is the σ-algebra,
- $\mathbb{P}$ is the probability function.

Properties of Probability Given the probability/sample space $(\Omega, \mathcal{A}, \mathbb{P})$, the probability function $\mathbb{P}$ must satisfy

- For every set $A \in \mathcal{A}$, $\mathbb{P}(A) \geq 0$
- $\mathbb{P}(\Omega) = 1$
- (Countably additive) Suppose the family of sets $(A_i)_{i\in\mathbb{N}}$ are mutually exclusive, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Probability Lemmas

- Given a family of disjoint sets $(A_i)_{i=1,\ldots,k}$

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \mathbb{P}(A_i)$$

- $\mathbb{P}(\phi) = 0$
- For any $A \in \mathcal{A}, \mathbb{P}(A) \leq 1$ and $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- Suppose $B, A \in \mathcal{A}$ and $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Continuity from Below Given an increasing sequence of events $A_1 \subset A_2 \subset \ldots$ then,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

Continuity from Above Given a decreasing sequence of events $A_1 \supset A_2 \supset \ldots$ then,

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

1.3 Conditional Probability and Independence

Conditional Probability The conditional probability that an event A occurs given that an event B has occurred is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cup B)}{\mathbb{P}(B)}, \quad \mathbb{P}(B) > 0$$

Independence Events A and B are independent if $\mathbb{P}(A \cup B) = \mathbb{P}(A)\mathbb{P}(B)$.

Lemma - Given two events A and B then $\mathbb{P}(A|B) = \mathbb{P}(A)$ if and only if $\mathbb{P}(B|A) = \mathbb{P}(B)$.

Independence of Sequences

- A countable sequence of event $(A_i)_{i=\mathbb{N}}$ is pairwise independent if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $i \neq j$.
- A countable sequence of events $(A_i)_{i=\mathbb{N}}$ are independent if for any sub-collection $A_{i_1}, \ldots, A_{i_n}$ we have
$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cdots \cap A_{i_n}) = \prod_{j=1}^{n} \mathbb{P}(A_{i_j})$$

Independence implies pairwise independence, but pairwise independence does not imply independence.

Multiplicative Law Given events A and B then

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

and similarly, if you have events A, B, C then

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_3|A_2 \cap A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_1)$$

Additive Law Let A and B be events then

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

Law of Total Probability Suppose $(A_i)_{i=1,\ldots,k}$ are mutually exclusive and exhaustive of Ω, that is $\bigcup_{i=1}^{k} A_i = \Omega$, then for any event B, we have

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

Bayes Formula Given sets B, A and a family of disjoint and exhaustive sets $(A_i)_{i=1,\ldots,k}$ then

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

1.4 Descriptive Statistics

Categorical Data can be sorted into a finite set of (unordered) categories. e.g. Gender

Quantitative Responses are measured on some sort of scale. e.g. Weight.

Numerical Summaries of the Quantitative Data Given observations $x = (x_1, \dots, x_n)$. The sample mean (estimated mean) or average is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample variance (estimated variance)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

2 Random Variables

2.1 Random Variables

Random Variables A random variable (r.v) X is a function from Ω to $\mathbb{R}$ such that $\forall \mathbf{x} \in \mathbb{R}$, the set $A_{\mathbf{x}} = \{\omega \in \Omega, X(\omega) \leq \mathbf{x}\}$ belongs to the σ-algebra $\mathcal{A}$.

Cumulative Distribution Function The cumulative distribution function of a r.v X is defined by

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

Cumulative Distribution Theorems Suppose F_X is a cumulative distribution function of X, then

- it is bounded between zero and one, and

$$\lim_{x \downarrow -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \uparrow \infty} F_X(x) = 1$$

- it is non-decreasing, that is if $x \leq y$ then $F_X(x) \leq F_X(y)$

- for any $x < y$,

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

- it is right continuous, that is

$$\lim_{n \uparrow \infty} F_X(x + \frac{1}{n}) = F_X(x)$$

- it has finite left limit and

$$\mathbb{P}(X < x) = \lim_{n \to \infty} F_X(x - \frac{1}{n})$$

  which we denote by $F_X(x-)$.

Discrete Random Variables A r.v X is said to be discrete if the image of X consists of countable many values x, for which $\mathbb{P}(X = x) > 0$.

Discrete Probability Function The probability function of a discrete r.v X is the function $\boldsymbol{\nabla} F_X(x) = \mathbb{P}(X = x)$ and satisfies

$$\sum_{\text{all possible } x} \mathbb{P}(X = x) = 1$$

Continuous Random Variables A r.v X is said to be continuous if the image of X takes a continuum of values.

Continuous Probability Density Function The probability density function of a continuous r.v is a real-valued function f_X on $\mathbb{R}$ with the property that

$$\mathbb{P}(X \in A) = \int_A f_X(y)\,dy$$

for any 'Borel' subset of $\mathbb{R}$.

For a function $f : \mathbb{R} \to \mathbb{R}$ to be a valid density function, the function f must satisfty the following properties.

1. for all $x \in \mathbb{R}, f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

Useful Properties (for continuous random variable) For any continuous random variable X with the density f_X,

1. by taking $A = (-\infty, x], \mathbb{P}(X \in (-\infty, x]) = \mathbb{P}(X \leq x)$ and
$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy$$
2. For any $a < b \in \mathbb{R}$, one can compute $\mathbb{P}(a < X \leq b)$ by
$$F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx$$
3. From the fundamental theorem of calculus and 1, we have
$$F_X'(x) = \frac{d}{dx} \int_{-\infty}^{x} f_X(y)\,dy = f_X(x).$$

2.2 Expectation and Variance

Expectation The expectation of a r.v X is denoted by $\mathbb{E}(X)$ and it is computed by

1. Let X be a discrete r.v. then

$$\mathbb{E}(X) := \sum_{\text{all possible } x} x\mathbb{P}(X = x) = \sum_{\text{all possible} x} x \boldsymbol{\nabla} F_X(x)$$

2. Let X be a continuous r.v. with density function $f_X(x)$ then

$$\mathbb{E}(x) := \int_{-\infty}^{\infty} x f_X(x)\, dx$$

Expectation of Transformed Random Variables Suppose $g : \mathbb{R} \to \mathbb{R}$, then the expectation of the transformed r.v $g(X)$ is

$$\mathbb{E}(g(X)) = \begin{cases} \int_{\mathbb{R}} g(x) f_X(x)\, dx & \text{continuous} \\ \sum_x g(x)\mathbb{P}(X = x) & \text{discrete} \end{cases}$$

usually one is interested in computing $\mathbb{E}(X^r)$ for $r \in \mathbb{N}$, which is called the r-th moment of X.

Linearity of Expectation The expectation $\mathbb{E}$ is linear, i.e., for any constants $a, b \in \mathbb{R}$,

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

Variance Let X be a r.v and we set $\mu = \mathbb{E}(X)$. The variance is X is denoted by $\operatorname{Var}(X)$ and

$$\operatorname{Var}(X) := \mathbb{E}((X - \mu)^2)$$

and the standard deviation of X is the square root of the variance.

Properties of Variance Given a random variable X then for any constant $a, b \in \mathbb{R}$,

1. $\operatorname{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$
2. $\operatorname{Var}(ax) = a^2 \operatorname{Var}(X)$
3. $\operatorname{Var}(X + b) = \operatorname{Var}(X)$
4. $\operatorname{Var}(b) = 0$

2.3 Moment Generating Functions

Moments A moment of the random variable is denoted by

$$\mathbb{E}[X^r], \quad r = 1, 2, \ldots$$

Moments measure mean, variance, skewness, and kurtosis, all ways of looking at the shape of the distribution.

Moment Generating Function The moment generating function (MGF) of a r.v X is denoted by

$$M_X(u) := \mathbb{E}(e^{uX})$$

and we say that the MGF of X exists if $M_X(u)$ is finite in some interval containing zero.

The moment generating function of X exists if there exists $h > 0$ such that the $M_X(x)$ is finite for $x \in [-h, h]$.

Calculating Raw Moments Suppose the moment generating function of a r.v X exists then

$$\mathbb{E}(X^r) = \lim_{u \to 0} M_X^{(r)}(u) = \lim_{u \to 0} \frac{d^r}{du} M_X(u)$$

Equivalence of Moment Generating Functions Let X and Y be two r.vs such that the moment generating function of X and Y exists and $M_Y(u) = M_X(u)$ for all u in some interval containing zero then $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$.

This theorem tells you that if the moment generating function exists then it uniquely characterises the cumulative distribution function of the random variable.

2.3.1 Useful Inequalities

The Markov Inequality (Chebychev's First Inequality) For any non-negative r.v X and $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

Chebychev's Second Inequality Suppose X is any r.v with $\mathbb{E}(X) = \mu, \mathrm{Var}(X) = \sigma^2$ and $k > 0$ then

$$\mathbb{P}(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$$

Convex (Concave) Functions A function h is convex (concave) if for any $\lambda \in [0, 1]$ and x_1 and x_2 in the domain of h, we have

$$h(\lambda x_1 + (1 - \lambda)x_2) \leq (\geq) \lambda h(x_1) + (1 - \lambda)h(x_2)$$

Jensen's Inequality Suppose h is a convex (concave) function and X is a r.v then

$$h(\mathbb{E}(X)) \leq (\geq) \mathbb{E}(h(X))$$

By using Jensen's inequality, one can show

$$\text{Arithmetic Mean} \geq \text{Geometric Mean} \geq \text{Harmonic Mean}.$$

That is given a sequence of number $(a_i)_{i=1,\ldots,n}$, we have

$$\frac{1}{n}\sum_{i=1}^{n} a_i \geq \left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} \geq n\left(\sum_{i=1}^{n} a_i^{-1}\right)^{-1}$$

3 Common Distributions

3.1 Common Discrete Distributions

Bernoulli Trail A Bernoulli trial is an experiment with two possible outcomes. The outcomes are often labelled 'success' and 'failure'. A Bernoulli trial defines a random variable X, given by

$$X = \begin{cases} 1 & \text{if the trail is a success} \\ 0 & \text{if the trail is a failure} \end{cases}$$

- Let $p \in [0, 1]$ be the probability of success
- We write $X \sim \text{Bernoulli}(p)$
- The probability function is given by $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$
- $\mathbb{E}(X) = p$
- $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = p(1 - p)$

Binomial Distribution Consider a sequence of n independent Bernoulli trials each with probability of success p. Let

$$X := \text{ total number of successes}$$

then X is a Binomial r.v with parameter n and p, and we write $X \sim \text{Bin}(n, p)$.

If $(Y_i)_{i=1,\dots,n}$ is a sequence of independent Bernoulli(p) random variable then $X := \sum_{i=1}^{n} Y_i$ is $\text{Bin}(n, p)$. The expectation of a Binomial random variable.

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} \mathbb{E}(Y_i) = np$$

Poisson Distribution A r.v X is said to follow the Poisson distribution with parameter λ, if it's probability function is given

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad k = 0, 1, \dots$$

where $\lambda = \mathbb{E}(X) = \text{Var}(X)$.

Hypergeometric Distribution A random variable has hypergeometric distribution with parameter N, m, n and we write $X \sim \text{Hyp}(n, m, N)$ if

$$\mathbb{P}(X = x) = \frac{C_x^m C_{n-x}^{n-m}}{C_n^N} \qquad x = 1, \dots, n$$

3.2 Continuous Distribution

Normal Random Variable A random variable X is said to be a normal random variable with parameters μ and σ^2 if its probability density function is

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

and we write $X \sim \mathcal{N}(\mu, \sigma^2)$.

Linear Transform Let X be a r.v with probability density function f_X, let $Y := a + bX$ then for $b > 0$ and $a \in \mathbb{R}$,

$$f_Y(x) = \frac{1}{b} f_X\left(\frac{x-a}{b}\right)$$

Linear Transform of Normally Distributed Random Variable Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ and $a \in \mathbb{R}$ and $b > 0$. The random variable $Y := a + bX$ is also normally distributed with parameter $(a + b\mu, b^2\sigma^2)$, i.e. $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$.

Standardisation Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ then

$$Z := \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Exponential Distribution A random variable X is said to be exponentially distributed with parameter $\lambda > 0$ if its probability density function is given by

$$f_X(x) = \frac{1}{\lambda} e^{-\frac{1}{\lambda}x}, \qquad x > 0$$

and we write $X \sim \exp(\lambda)$. Then $\mathbb{E}(x) = \lambda$ and $\mathrm{Var}(X) = \lambda^2$.

Gamma Distribution A random variable X is said to be Gamma distributed with parameter $\alpha, \beta > 0$ if its probability density function is given by

$$f_X(x; \alpha, \beta) = \frac{e^{\frac{-x}{\beta}} x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}, \qquad x > 0$$

and we write $X \sim \mathrm{Gamma}(\alpha, \beta)$ where $\mathbb{E}(X) = \alpha\beta$ and $\mathrm{Var}(X) = \alpha\beta^2$.

Beta Distribution The Beta function is given by

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt, \qquad x, y > 0$$

and the Beta and Gamma functions satisfies the following relationship

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \qquad x, y > 0$$

A random variable is said to follow a Beta distribution with parameters $\alpha, \beta > 0$ if its density function is given by

$$f_X(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \qquad x \in (0,1)$$

and we write $X \sim \text{Beta}(\alpha, \beta)$.

3.2.1 QQ-plot

Quantile Suppose X is a continuous random variable with CDF given by F_X. The $k\%$-th quantile of X is given by

$$Q_X(k) := F_X^{-1}(k), \qquad k \in [0,1]$$

where F_X^{-1} is the inverse of the CDF F_X.

Quantile Plot Given continuous r.vs X and Y, the theoretical qunatile plot of X against Y is the graph

$$(Q_X(k), Q_Y(k)), \qquad k \in [0,1]$$

Suppose we are given X and $Y = aX + b$ for some $a > 0, b \in \mathbb{R}$ then the quantile plot of X against Y is a straight line.

Given r.v.s X and Y and suppose that the quantile plot of X against Y is a straight line. Then the distribution of X is equal to the distribution of a linear transform of Y.

3.2.2 Indicator Functions

- A indicator function of a set A is defined by

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \in A^c \end{cases}$$

- Indicator function of an interval is given as

$$I_[a,b](x) = I_{\{a \leq x \leq b\}} \qquad \text{or} \qquad I_{(a,b]}(x) = I_{\{a < x \leq b\}}$$

- The indicator unifies expectation $\mathbb{E}$ and probability $\mathbb{P}$ notation since, the probability is the expectation of the indicator function. Therefore, it may be written that

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx = \int_{-\infty}^{\infty} I_A(x) f_X(x) dx = \mathbb{E}(I_A(X)).$$

4 Bivariate Distribution

The joint density function of two continuous random variables X and Y is given by a bivariate function $f_{X,Y}$ with the properties

1. For all $x, y \in \mathbb{R}^2$, $f_{X,Y}(x, y) \geq 0$.
2. The double integral over $\mathbb{R}^2$ is equal to one, that is

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dxdy = 1.$$

3. For any (measurable) set $A, B \in \mathbb{R}$

$$\int_B \int_A f_{X,Y}(x, y)\, dxdy = \mathbb{P}(X \in A, Y \in B).$$