COMP4121 Lecture Notes - Week 0

A (very brief) probability and statistics refresher

LiC: Aleks Ignjatovic

ignjat@cse.unsw.edu.au

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

CAVEAT:

This is a VERY BRIEF REFRESHER ***ONLY*** *of basic Probability and Statistics, which we will be expanding as we progress through the course by introducing more notions as we need them. It is definitely* ***NOT*** *meant to be a replacement for a proper course in Probability and Statistics; if you have not taken such a course you should seriously consider taking one, or, at the very least, you should read a good book on Statistics, because statistical methods are* ***essential*** *for design and testing of modern algorithms.*[1]

1 Basic Definitions

• The **sample space**, denoted Ω, is the set of all possible outcomes of a process (usually called *an experiment*) whose outcomes are non-deterministic, i.e, which might be different if the experiment were repeated.

Examples: Consider the following two experiments:

1. we toss a coin 3 times in a row;
2. we measure the amplitude of a signal in the presence of noise (which is always present!)

In the first example the sample space is the set

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

Note that the outcomes of each individual tossing of the coin are NOT in the sample space but only the triplets belong to it. In the second example the sample space is the set of real numbers $\mathbb{R}$ (or a subset of $\mathbb{R}$ which corresponds to the range of your measuring instrument).

• **An event** is any subset of the sample space. Events which we are interested in are usually definable by a some property which each individual sample might have or might not have. For example, in Example 1 above, one such property could be *getting at least two heads* and the samples which satisfy this property are HHH, HHT, HTH, THH; thus, this event is the set $\{HHH, HHT, HTH, THH\}$. In the second example an event might be "*getting a measurement in the range* $0 - 1$"; thus the corresponding set of samples is the interval $[0, 1] \subset \mathbb{R}$. An event "has happened" (has been realised) in an experiment if the sample which is the outcome of the experiment belongs to this event.

[1] One of the best books on statistics for computer scientists, which we closely follow here, is Larry Wasserman's "*All of Statistics*" which dispenses with all academic "niceties" and "pedagogical examples" and cuts straight into the core of the subject matter, covering essentially all of the topics which are important for the present day computer science, in particular for algorithms in machine learning, information retrieval and data-mining, etc.

• **Probability of an event** is given by a probability distribution function $\mathcal{P}$ which assigns numbers to events[‡] such that the assigned values are all in the range $[0, 1]$ and, in particular, the probability assigned to the entire set Ω of all outcomes is 1. More over, if $A = \bigcup_{i=1}^{\infty} A_i$ and if all A_i are pairwise disjoint sets, then $\mathcal{P}(A) = \sum_{i=1}^{\infty} \mathcal{P}(A_i)$ must hold.

• **Events $\{A_i\}_{1 \le i \le n}$ are independent** if

$$\mathcal{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathcal{P}(A_i).$$

• **Conditional probability of A given B** is defined as

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

It is the probability that event A has happened *assuming* that event B has happened; see the figure 1. Thus, B becomes the universe of all possible events (because we assume that B has happened); the probability that A has also happened is then simply equal to whatever fraction $A \cap B$ is of the entire B (with the probability of a set as the measurement of the "size" of a set).

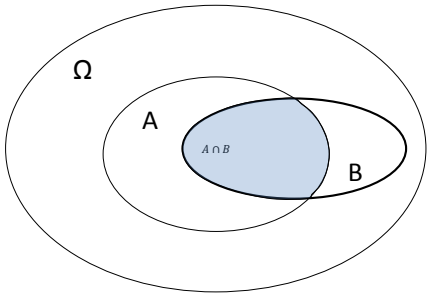

Figure 1.1: $\mathcal{P}(A|B) = \mathcal{P}(A \cap B)/\mathcal{P}(B)$

Note that the above implies that if A is independent of B then

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

i.e, the assumption that B has happened has no impact on the probability that A has also happened, which justifies the term.

[‡] *You can ignore the footnotes marked with §, like the present one; they are there just to keep the mathematicians quiet, because they state something irrelevant to engineers, such as the one which now follows*: It is generally NOT possible to assign a probability to every single event (i.e., to every subset of the sample space), but such "pathological" sets can be safely ignored by engineers simply because they do not correspond to any events that make sense in practice; their "existence" relies on fancy infinitary set theory (Axiom of Choice) which, for an engineer, is as relevant as theology.

Let B be an arbitrary event, and assume that the entire universe Ω of all outcomes has been partitioned into disjoint subsets $\{A_i\}_{1\leq i\leq n}$, i.e., assume that A_i satisfy: for all $i, j \leq n$,

$$A_i \cap A_j = \emptyset \quad \text{and} \quad \bigcup_{1\leq i\leq n} A_i = \Omega;$$

see figure 1. Then,

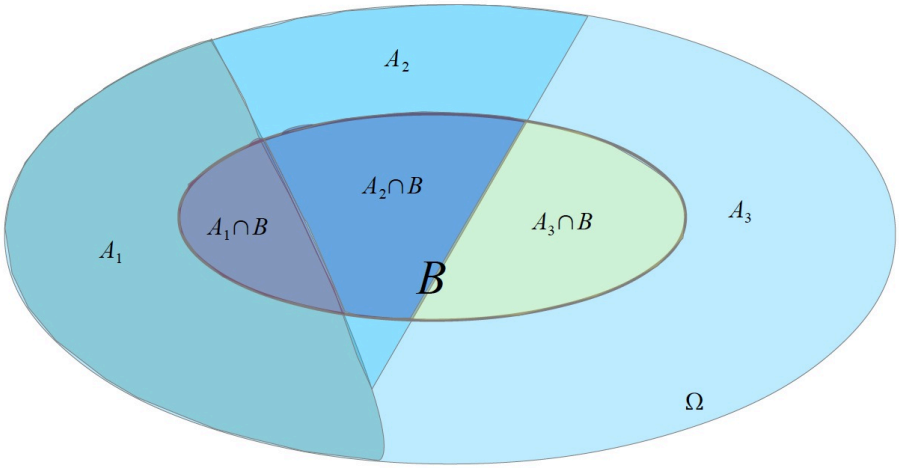

Figure 1.2:

$$\mathcal{P}(B) = \sum_{i=1}^{n} \mathcal{P}(A_i \cap B) = \sum_{i=1}^{n} \mathcal{P}(B \,|A_i)\mathcal{P}(A_i)$$

Thus, we get Bayes' Theorem:

$$\mathcal{P}(A_i \,|\, B) = \frac{\mathcal{P}(A_i \cap B)}{\mathcal{P}(B)} = \frac{\mathcal{P}(B \,|\, A_i)\mathcal{P}(A_i)}{\mathcal{P}(B)} = \frac{\mathcal{P}(B \,|\, A_i)\mathcal{P}(A_i)}{\sum_{i=1}^{n} \mathcal{P}(B \,|A_i)\mathcal{P}(A_i)}. \qquad (1.1)$$

Note that $\mathcal{P}(A_i)$ is called the *prior probability* of A (before B has been observed to have happened) and $\mathcal{P}(A_i \,|\, B)$ is the *posterior probability* (after event B was observed). Bayes' Theorem is important for Computer Science because of the *Bayesian Belief Revision*, in which our belief changes according to what events have been observed to have happened.

Example: A computer manufacturing company has two suppliers of hard disk drives. Company C1 supplies 40% and C2 supplies 60% of all disks. Estimates from the past experience suggest that .4% of $C1's$ drives and .3% of $C2's$ drives are defective.

1. What is the probability that a randomly selected disk is from company C1?

2. Such randomly selected disk has subsequently been tested and found to be defective. What is now the probability that it is from C1?

Let us denote by A_1 (A_2, respectively) the event that the randomly chosen disk is from company C1 (C2, respectively) and by B the event that the disk is defective; thus,

$$\mathcal{P}(A_1) = 40/100;\ \mathcal{P}(A_2) = 60/100;\ \mathcal{P}(B \mid A_1) = 4/1000;\ \mathcal{P}(B \mid A_2) = 3/1000.$$

Clearly, the answer to question 1 is $\mathcal{P}(A_1) = 40/100$ which is the *prior probability* that the disk is from company C1. After the new piece of information has arrived, knowing that the selected disk is defective, we can revise the degree of our belief that the disk is from company C1, by using Bayes' Theorem: since

$$\begin{aligned} \mathcal{P}(B) &= \mathcal{P}(B \mid A_1)\mathcal{P}(A_1) + \mathcal{P}(B \mid A_2)\mathcal{P}(A_2) \\ &= 4/1000 \times 40/100 + 3/1000 \times 60/100 = 0.0034 \end{aligned}$$

we get

$$\mathcal{P}(A_1 \mid B) = \frac{\mathcal{P}(B \mid A_1)\mathcal{P}(A_1)}{\mathcal{P}(B)} = \frac{4/1000 \times 40/100}{0.0034} \approx 0.47$$

Thus, after the new piece of information has arrived, our belief that the disk comes from C1 has increased from 0.4 to 0.47 which is to be expected, because disks from company C1 are more frequently defective than the disks from company C2.

2 Random Variables

• **A random variable** X is a mapping from the set of all samples into the set of real numbers. Thus, X assigns a value $X(\omega)$ to every sample ω.[‡]

• **A discrete random variable** X is a random variable which can have at most countably many values $\{x_0, \ldots, x_n, \ldots\}$; a random variable which can have continuum many values is called a **continuous random variable**.

• **The cumulative distribution function (CDF)** of a random variable X is the mapping from $\mathbb{R}$ into interval $[0, 1]$ given by $F_X(x) = \mathcal{P}(\{\omega : X(\omega) \leq x\})$ or, in a more usual notation, $F_X(x) = \mathcal{P}(X \leq x)$.

• **The probability mass function** $f_X(x)$ of a discrete random variable X specifies the probability that X attains each of its possible values x_i, i.e., $f_X(x_i) = \mathcal{P}(X = x_i) = p_i$; clearly, $\sum_{1 \leq i \leq n} p_i = 1$.

• **The probability density function (PDF)** of a continuous random variable is a function $f_X(x) \geq 0$ satisfying:

$$\int_{-\infty}^{a} f_X(x)dx = \mathcal{P}(X \leq a). \tag{2.1}$$

[‡]Just as with probability distributions, not all such mappings are legitimate random variables, but only those which have the property that the inverse images $X^{-1}[a, b]$ of intervals $[a, b]$ are in the domain of the corresponding probability distribution on the (subsets of the) sample space. Since the inverse image of an interval can be "weird" only for "very weird" random variables X, all mappings from the sample space into $\mathbb{R}$ which make physical sense are in fact legitimate random variables. Thus, the engineers can safely let the mathematicians worry about the rest.

This implies that for all a, b such that $b > a$,

$$\mathcal{P}(a \leq X \leq b) = \mathcal{P}(X \leq b) - \mathcal{P}(X \leq a) = \int_{-\infty}^{b} f_X(x)dx - \int_{-\infty}^{a} f_X(x)dx$$
$$= \int_{a}^{b} f_X(x)dx \tag{2.2}$$

Note that, by definition, a random variable is a fully deterministic mapping from the set of all possible outcomes of an experiment into real numbers; we can think of a random variable as a (deterministic) measurement performed on the outcome of an experiment. The values of a random variable are non-deterministic only because its inputs are such.

However, we usually do not care about particular samples which are the outcomes of an experiment, but only about the values of the "measurements" performed on these random samples, i.e., we only care about the values of the random variables representing the values of these measurements. Thus, we usually ignore the samples altogether and consider a random variable itself as a "probabilistic object" specified by the composition of X^{-1} and the probability distribution $\mathcal{P}$: for every $x \in \mathbb{R}$ let

$$A_x = \{\omega : X(\omega) \leq x\} = X^{-1}\big((-\infty, x]\big);$$

then x is assigned the value $\mathcal{P}(A_x) = \mathcal{P}(X \leq x)$.

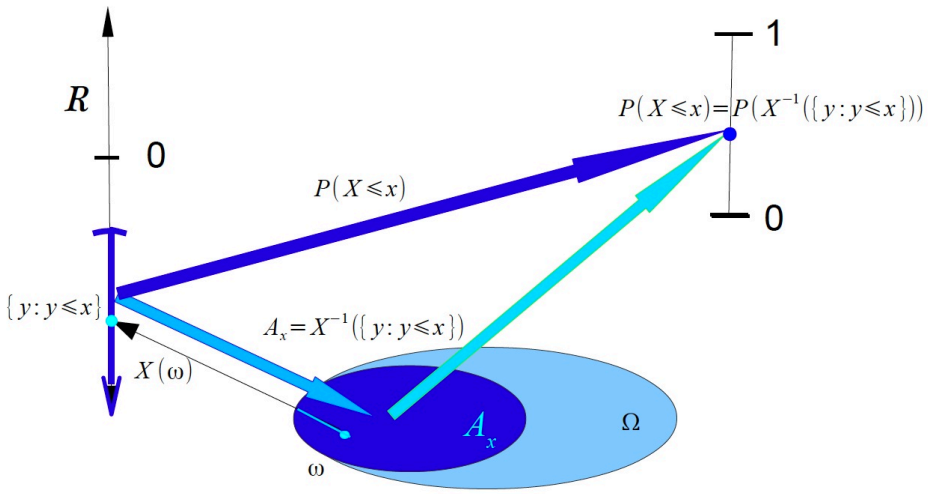

Figure 2.1: A random variable seen as a mapping from $\mathbb{R}$ into $[0, 1]$

Using the first equality of (2.2) we get the probability of the event that the value of a random variable X belongs to an interval $[a, b]$; this can be extended to obtain the probability that the value of X belongs to any "reasonable" subset $A \subset \mathbb{R}$. Thus, every random variable induces a probability distribution on $\mathbb{R}$ and we can now ignore the set of samples Ω and deal only with random variables and their corresponding probability distributions $\mathcal{P}(X \in A)$ and their

corresponding probability density functions $f_X(x)$; see figure 2.[‡‡]

Warning: Random variables CANNOT be identified with their probability distributions: for example, if a distribution of X is symmetric around the origin, in the sense that $\mathcal{P}(X \leq a) = \mathcal{P}(X \geq -a)$ for every $a \in \mathbb{R}$, then

$$\mathcal{P}(X \leq a) = \mathcal{P}(X \geq -a) = \mathcal{P}(-X \leq a).$$

Thus, X and $-X$ have the same probability distribution, but are clearly not equal as mappings from samples into reals.

• Two random variables X, Y are **equally distributed** if they have the same probability distribution function, in which case we write $X \sim Y$.

If $F(x)$, ($f(x)$, respectively) is a probability distribution (probability density function, respectively) and X has probability distribution $F(x)$ (probability density function $f(x)$, respectively) we abuse the notation and also write $X \sim F$ ($X \sim f$, respectively).

3 Expectation and Variance of a Random Variable

The expected value of a random variable is simply its mean. Thus, if X can take countably many values $v_0, v_1, \ldots$ with the corresponding probabilities $p_0, p_1, \ldots$, then

$$E(X) = \sum_{i=1}^{\infty} v_i \cdot p_i$$

providing that the sum converges. Thus, not every random variable has a finite mean value.

If X is a continuous random variable with probability density function $f(x)$, then

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

again assuming that such integral converges.

[‡‡]Again, not every continuous random variable has a corresponding probability density function, but only those which induce a "well-behaved" probability distribution $\mathcal{P}(X \leq x)$, the so called *absolutely continuous probability distributions*. Roughly speaking, "well behaved" probability distributions are those which assign probability zero to all sets whose "geometric size" is zero (such as, for example, sets of isolated points). Since all physically meaningful random variables have this property, for engineers all continuous random variables have a probability density function, which the mathematicians call the *Radon - Nikodym derivative* of the probability distribution $\mathcal{P}(X < x)$ - a term justified by taking the derivatives with respect to a of both sides of equation (2.1).

Sometimes expectation is written as $E(X) = \int x dF_X(x)$ where $F(x)$ is the cumulative distribution function, $F(x) = \mathcal{P}(X \leq x)$; this is not just a notation but it can be given a precise meaning using more sophisticated mathematical machinery.[‡]

Properties of expectation

- Let $Y = g(X)$; then $E(Y) = \int g(x) dF_X(x)$.
- Let $X_1, \ldots, X_n$ be random variables and $c_1, \ldots, c_n$ real numbers; then

$$E(c_1 X_1 + \ldots + c_n X_n) = c_1 E(X_1) + \ldots + c_n E(X_n).$$

- Assume $X_1, \ldots, X_n$ are independent; then

$$E(X_1 X_2 \cdots X_n) = E(X_1) E(X_2) \ldots E(X_n)$$

The k^{th} moment of a distribution X is defined as $E(X^k)$, assuming that $E(|X|^k)$ exists, i.e. assuming that

$$E(|X|^k) = \int |x|^k dF_X(x) < \infty$$

Clearly, if $E(|X|^k) < \infty$, then also $E(X^k) < E(|X|^k) < \infty$. The stronger assumption, namely $E(|X|^k) < \infty$ rather than just $E(X^k) < \infty$ ensures that the moments are "well behaved". For example one can easily verify that if $E(|X|^k) < \infty$, then also $E(X^m) \leq E(|X|^m) < \infty$ for all $m \leq k$, i.e., the existence of a higher order moment guarantees the existence of all lower order moments.

The k^{th} central moment of a distribution X is defined as the k^{th} moment of $Y = X - E(X)$.

The moment generating function of a random variable X is the function $m(t) = E(e^{t\,X})$. The reason for such a name is that the Taylor Expansion of $m(t)$, obtained by expanding the exponential function into a series and using the linearity of expectation,

$$m(t) = E(e^{tX}) = E\left(\sum_{k=1}^{\infty} \frac{t^k X^k}{k!}\right) = \sum_{k=1}^{\infty} \frac{t^k E(X^k)}{k!}$$

and $E(X^k)$, $(1 \leq k)$, are the moments of X.

Variance and Standard Deviation of a Random Variable

The variance of a random variable is defined as $V(X) = E(X - E(X))^2$, assuming that both expectations involved are finite; the standard deviation of a

[‡]To make such notation fully meaningful one needs an extension of the notion of a function, called *a generalized function* (also called *a distribution*). Generalized functions allow us to take derivatives of step-like functions such as $F_X(x)$ when X is a discrete random variable.

random variable X is given by $\sigma = \sqrt{V(X)}$.

Properties of variance

- $$\begin{aligned} V(X) &= E(X - E(X))^2 && (3.1)\\ &= E(X^2 - 2XE(X) + (E(X))^2) && (3.2)\\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 && (3.3)\\ &= E(X^2) - (E(X))^2 && (3.4) \end{aligned}$$

- Let $X_1, \ldots, X_n$ be **independent** random variables and $c_1, \ldots, c_n$ real numbers; then

$$V(c_1X_1 + \ldots + c_nX_n) = c_1^2V(X_1) + \ldots + c_n^2V(X_n).$$

4 A few most common random variables

We first list a few common discrete random variables.

4.1 The Discrete Uniform Distribution

The Discrete Uniform Distribution on the set of samples of size n assigns probability $1/n$ to any of the n possible outcomes.

4.2 The Bernoulli Distribution

This is just the probability of getting a head when tossing a (biased) coin. Thus, the set of all outcomes is $\{0, 1\}$ and $\mathcal{P}(X = 1) = p$; $\mathcal{P}(X = 0) = 1 - p$, for some $0 \leq p \leq 1$.

Note: $p = 0$ does NOT mean that the event $X = 1$ is physically impossible; it only means that if you repeated the coin tossing experiment n times, then the number of heads you get divided by the total number of trials will converge to 0 as n approaches infinity.

4.3 The Binomial Distribution

Let the experiment be "tossing a (biased) coin n times", and let the random variable X be "the number of heads obtained". Then the values of X are $\{0, 1, \ldots, n\}$, and $\mathcal{P}(X = k) = \binom{n}{k}p^k(1 - p)^{n-k}$. We denote such a random variable by Binomial(n, p).

Assume that we toss a coin n times and count the number of heads X_1 and then we toss it then m times and again count the number of heads X_2. What is the probability distribution of the sum of these two numbers, i.e., what is the probability distribution of $X_1 + X_2$? Clearly, since the above experiment can be also seen as a single instance of the experiment consisting of tossing the

coin $n+m$ times and counting the number of heads, we have that $X_1 + X_2 = \text{Binomial}(n+m, p)$, i.e.,

$$\text{Binomial}(n,p) + \text{Binomial}(m,p) \sim \text{Binomial}(n+m,p).$$

The expected value of Binomial distribution can be found by considering it as the sum of n Bernoulli distributed independent random variables, $X = Y_1 + \ldots + Y_n$; then $E(X) = nE(Y_1) = np$; the variance is likewise $V(X) = nV(Y_1) = nE(X-p)^2 = n(E(x^2) - p^2) = n(p - p^2) = np(1-p)$.

4.4 The Poisson Trials

The Poisson Trials generalize the Binomial Distribution by allowing different probabilities at each trial. Thus, in this case the random variable is defined as $X = \sum_{i=1}^{n}$ where X_i are independent Bernoulli Trials, each with probability of success of p_i. Clearly if $p_i = p$ for all i we get just the Binomial Distribution. Clearly, $E(X) = \sum_{i=1}^{n} p_i$.

4.5 The Geometric Distribution

The Geometric Distribution is just the distribution of the random variable corresponding to the number of tossing of a (biased) coin you need to perform to get a head. Clearly, its values are all positive integers, and $\mathcal{P}(X = k) = p(1-p)^{k-1}$.

Note that

$$\sum_{k=1}^{\infty} p(1-p)^{k-1} = p\sum_{k=1}^{\infty}(1-p)^{k-1} = \frac{p}{1-(1-p)} = 1$$

4.6 The Poisson Distribution

A discrete random variable X has the Poisson Distribution with parameter λ if

$$\mathcal{P}(X = n) = e^{-\lambda}\frac{\lambda^n}{n!}$$

for all integers $n \geq 0$. Note that the definition is correct because

$$\sum_{n\geq 0} e^{-\lambda}\frac{\lambda^n}{n!} = e^{-\lambda}\sum_{n\geq 0}\frac{\lambda^n}{n!} = e^{-\lambda}e^{\lambda} = 1$$

We now list a few common continuous random variables.

4.7 The Uniform Distribution

The uniform distribution on an interval $[a,b]$ is given by a constant density on $[a,b]$, i.e., if X has the following density function $f(x)$:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

4.8 The Normal (Gaussian) Distribution

X has a normal distribution with the mean μ and standard deviation σ, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$ if

$$f(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

If X is a Binomial distribution corresponding to Bernoulli trials with probability of success equal to p, then, if n is reasonably large and p is not too close to 1 or 0, then X can be reasonably well approximated by a normal distribution with the same mean and standard deviation, $X \sim \mathcal{N}(np, np(1-p))$

4.9 The Exponential Distribution

X has an exponential distribution with a parameter β, denoted by $X \sim \text{Exp}(\beta)$, if

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \quad \text{for all} \quad x \geq 0.$$

5 Some simple inequalities

$$1 + x \leq e^x \quad \text{for all} \quad x \in \mathbb{R}\,. \tag{5.1}$$

Let $f(x) = e^x - x - 1$, since $f'(x) = e^x - 1$, we have $f'(x) = 0$ if and only if $x = 0$. Since $f''(x) = e^x > 0$, function $f(x)$ is convex and thus it attains a minimum at $x = 0$. Since $f(0) = 0$ we get that $f(x) \geq 0$ for all x. Taking $x = -1/n$ we get the first of the two inequalities below; prove, as an exercise the second inequality.

$$\left(1 - \frac{1}{n}\right)^n \leq \frac{1}{e} \leq \left(1 - \frac{1}{n}\right)^{n-1} \quad \text{for all} \quad n \in \mathbb{N}\,. \tag{5.2}$$

6 Probability Inequalities

Roughly speaking, probability inequalities estimate the probability that a random variable X takes values which deviate from its mean $\mu = E(X)$.

6.1 The Markov Inequality

Let $X > 0$ be a non-negative random variable. Then for all $t > 0$,

$$\mathcal{P}(X \geq t) \leq \frac{E(X)}{t}$$

Thus, taking $t = \lambda E(X)$, we get

$$\mathcal{P}(X \geq \lambda E(X)) \leq \frac{1}{\lambda}.$$

The Markov inequality follows from the observation that

$$\begin{aligned} E(X) &= \int x dF_X x = \int_{x<t} x dF_X x + \int_{x\geq t} x dF_X x \\ &> \int_{x\geq t} x dF_X x > \int_{x\geq t} t dF_X x = t \int_{x\geq t} dF_X x \\ &= t\mathcal{P}(X > t) \end{aligned}$$

Note that, if $g(x)$ is a function which always attains only non-negative values, then for every random variable X we have

$$\mathcal{P}(g(X) \geq t) \leq \frac{E(g(X))}{t}$$

and if $\mathcal{P}(g(X) \geq t) > 0$, then the above proof shows that, in fact,

$$\mathcal{P}(g(X) \geq t) > \frac{E(g(X))}{t}. \tag{6.1}$$

One cannot prove a sharper inequality than the Markov Inequality without assuming something more that non-negativity of X and the existence of its expected value. Assuming that the variance of X also exists, we can get the following *additive* bound (rather than multiplicative, as in the Markov inequality) for how far X is likely to be away from its expected value.

6.2 The Chebyshev Inequality

Let $X > 0$ be a random variable with the expected value $\mu = E(X)$ and standard deviation $\sigma = \sqrt{E((X-\mu)^2)}$. Then for all $\lambda > 0$,

$$\mathcal{P}(|X - \mu| \geq \lambda\, \sigma) \leq \frac{1}{\lambda^2}$$

The Chebyshev inequality follows from the Markov inequality applied to the random variable $(X-\mu)^2$. Note that $(X-\mu)^2 \geq \sigma^2$ iff $|X-\mu| > \sigma$. Thus

$$\mathcal{P}(|X - \mu| \geq \lambda\, \sigma) = \mathcal{P}((X-\mu)^2 \geq \lambda^2\, \sigma^2) \leq \frac{E((X-\mu)^2)}{\lambda^2\, \sigma^2} = \frac{1}{\lambda^2}$$

6.3 The Chernoff bound

The last inequality we will use is an example of a technique called *the Chernoff bound*; it can be seen as an application of the Markov inequality to the moment generating function $m(t) = e^{t\,X}$, where X is a Poisson Trial.

Theorem: Let $X = \sum_{k=1}^{n} X_k$, where X_k, $1 \leq k \leq n$, are independent Bernoulli trials with the probability of success $\mathcal{P}(X_k = 1) = p_k$, where $0 < p_k < 1$. Thus, $\mu = E(X) = \sum_{k=1}^{n} p_k$. Let $\delta > 0$ be any positive real. Then the following **Chernoff bound** holds:

$$\mathcal{P}(X > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

Proof: Clearly

$$\mathcal{P}(X > (1+\delta)\mu) = \mathcal{P}(e^{t\,X} > e^{t(1+\delta)\mu})$$

and by the Markov inequality to the right side we get

$$\mathcal{P}(e^{t\,X} > e^{t(1+\delta)\mu}) < \frac{E(e^{t\,X})}{e^{t(1+\delta)\mu}}$$

We now use the fact that $X = \sum_{k=1}^{n} X_k$ and obtain

$$\mathcal{P}(X > (1+\delta)\mu) < \frac{E(e^{t\,\sum_{k=1}^{n} X_k})}{e^{t(1+\delta)\mu}} = \frac{E(\prod_{k=1}^{n} e^{t\,X_k})}{e^{t(1+\delta)\mu}}$$

Since X_k are independent, so are random variables e^{tX_k}; thus

$$E\left(\prod_{k=1}^{n} e^{t\,X_k}\right) = \prod_{k=1}^{n} E(e^{t\,X_k})$$

Clearly, random variable $e^{t\,X_k}$ takes the value $e^{t\cdot 1} = e^t$ with probability p_k and the value $e^{t\cdot 0} = e^0 = 1$ with probability $1 - p_k$. Thus,

$$E(e^{t\,X_k}) = p_k\,e^t + 1 - p_k = 1 + p_k(e^t - 1);$$

Consequently,

$$\mathcal{P}(X > (1+\delta)\mu) < \frac{\prod_{k=1}^{n}(1 + p_k(e^t - 1))}{e^{t(1+\delta)\mu}}$$

We now apply the inequality (5.1) with $x = p_k(e^t - 1)$, and obtain

$$\begin{aligned}\mathcal{P}(X > (1+\delta)\mu) &< \frac{\prod_{k=1}^{n} e^{p_k(e^t-1)}}{e^{t(1+\delta)\mu}} = \frac{e^{(e^t-1)\sum_{k=1}^{n} p_k}}{e^{t(1+\delta)\mu}} = \frac{e^{\mu(e^t-1)}}{e^{t(1+\delta)\mu}} \\ &= e^{\mu(e^t-1)-t(1+\delta)\mu}\end{aligned}$$

Note that the above derivations hold for every positive t; we now choose t which yields the tightest bound, which happens if the right hand side achieves a minimum. Thus, we find the stationary points of the function

$$h(t) = e^{\mu(e^t-1)-t(1+\delta)\mu}.$$

Since

$$h'(t) = e^{\mu(e^t-1)-t(1+\delta)\mu}(\mu\,e^t - (1+\delta)\mu)$$

we have

$$h'(t) = 0 \quad \Leftrightarrow \quad \mu\,e^t - (1+\delta)\mu = 0$$

i.e.,

$$t = \ln(1+\delta)$$

It is easy to verify that $h''(t) = e^{\mu(e^t-1-(1+\delta)t)}(\mu^2(1+\delta-e^t)^2 + \mu e^t) > 0$ for all t, which implies that $h(t)$ achieves a minimum at $t = \ln(1+\delta)$.

Thus, we conclude that

$$\begin{aligned}\mathcal{P}(X > (1+\delta)\mu) &< e^{\mu(e^{\ln(1+\delta)}-1)-(1+\delta)\mu\ln(1+\delta)} = e^{\mu((1+\delta)-1)-(1+\delta)\mu\ln(1+\delta)} \\ &= \frac{e^{\mu\,\delta}}{e^{\ln(1+\delta)^{(1+\delta)\mu}}} = \frac{e^{\mu\,\delta}}{(1+\delta)^{(1+\delta)\mu}} \\ &= \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}\end{aligned}$$