Ma3 — Probability and Statistics
Kim C Border

Yubo Su

Contents

1 05/01/15 — Introduction 2

2 07/01/15 — Properties of Probability, Conditional Probability, Independence 3
2.1 Conditional probability, Inclusion-Exclusion . . . 3
2.2 Counting and probabilities . . . 4
2.3 Matching balls and bins . . . 4

3 12/01/15 — Binomial Distribution, Independence 5
3.1 Multinomial distribution, "Negative binomial distribution" . . . 5
3.2 Independence, Bayes' Rule . . . 5

4 14/01/15 — More Bayes' Law, expectation of random variables 6
4.1 Random Variables . . . 6

5 16/01/15 — Higher expectations, densities 7
5.1 Continuous distributions . . . 7
5.2 Expectation Properties . . . 7
5.3 Jensen's Inequality . . . 7
5.4 Variance and higher moments . . . 7

6 21/01/15 — Normal Distribution 8
6.1 Proving the Principle of Inclusion/Exclusion . . . 8
6.2 Inclusion/Exclusion Principle . . . 8
6.3 Variances, standardized random variables . . . 8

Chapter 1

05/01/15 — Introduction

Time to go home and toss 128 coins! Apparently 51% of the time coins land on the same side they started on, similar to the empirical probability of being born male.

Types of probabilities: Empirical, observed frequencies in large samples; subjective, that probability exists only in the mind.

Border claims the digits of π are random; to justify this claim, we count the number of occurrences of 31415 in the first billion (and one) digits of π. We count 10010 in fact, very close to the theoretically expected value of 10000. Moreover, the following 100 possible two digit number combinations occur with equal frequency. Finally, if we examine the occurrences of the digits and apply the χ-squared test, we find that a random 1 billion sequence gets similar deviations about 25% of the time. Conclusion, π is random. Faskinating!

Monte Carlo methods are all the rage nowadays given the existence of pseudorandom numbers.

Finally administrative details! Website is `http://www.math.caltech.edu/~2014-15/2term/ma003/`, Border's email is `kcb@caltech.edu` while the lead TA, Marius Lemm, is `mlemm@caltech.edu`.

The set of outcomes of a random experiment is the sample space, S or Ω usually. An event is a subset of the sample space, and the set of all events is denoted ε or Σ. The set of all events is an algebra of sets, and we usually assume it is a σ-*algebra*, which means

- $\emptyset \in \Sigma, S \in \Sigma$
- If $E \in \Sigma$ then $E^c \in \Sigma$
- If E and F belong in Σ then $E \cap F$ and $E \cup F$ belong to Σ.

A probability measure is a set function $P : \Sigma \to [0, 1]$ that satisfies $P(\emptyset) = 0, P(S) = 1$.

Chapter 2

07/01/15 — Properties of Probability, Conditional Probability, Independence

Side note, the odds against an event E is the ratio $\dfrac{P(E^c)}{P(E)}$ while the probability of an event E is $P(E)$. Let's get a few more definitions in.

- Countable additivity — Probability measure satisfying

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \tag{2.1}$$

- Probability space — Triple (S, Σ, P) where S is the outcome space, Σ is the set of events, and P is a countably additive probability measure.
- Random variable — Real valued function on S satisfying that for every interval $I \in X$, I^c is also an event. Note that if Σ comprises all subsets of S then this requirement is automatically satisfied.

Some elementary identities

- $P(A^c) = 1 - P(A)$
- If $B \subset A$ then $P(A \backslash B) = P(A) - P(B)$.
- If $\{A_i\}$ are *pairwise disjoint* then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \tag{2.2}$$

  If the events are not pairwise dispoint the above equality becomes $\leq$.
- Let a sequence $\{E_i\}$ be $E_n \uparrow$ if $E_i \subset E_{i+1}$ and vice versa for $E_n \downarrow$. Then P is countably additive if and only if $E_n \downarrow$ implies

$$P\left(\bigcap_{n} E_n\right) = \lim_{n\to\infty} P(E_n) \tag{2.3}$$

  and union instead of intersection for $E_n \uparrow$.

2.1 Conditional probability, Inclusion-Exclusion

If $P(A) > 0$, the conditional probability of B given A, written $P(B|A)$ is

$$P(B|A) = \frac{P(BA)}{P(A)} \tag{2.4}$$

Two events are considered *independent* if $P(AB) = P(A) \cdot P(B)$. Note that if A, B are independent, so are pairwise combinations of them and A^c, B^c.

We then arrive at the Principle of Inclusion-Exclusion, which states that in general

$$P\left(A \cup B\right) = P(A) + P(B) - P(AB) \tag{2.5}$$

and of course the generalisation.

2.2 Counting and probabilities

First, distinguish *lists* from *sets* because the first are ordered. For choosing subsets, we use binomial notation $\binom{n}{k}$. The very useful identity

$$\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k} \tag{2.6}$$

is simply a restatement of Pascal's Triangle.

2.3 Matching balls and bins

A cool problem that we've seen frequently is that of numbered balls and bins; what is the probability that at least one ball matches its bin? Index the balls and bins i, and let A_i be the probability that the i-th ball matches the i-th bin. We want $P\left(\bigcup_i A_i\right)$ which is a union of non-disjoint events! So we have to apply the principle of inclusion-exclusion,

$$P\left(\bigcup_i A_i\right) = \sum_i p(A_i) - \sum_{i<j} P(A_i A_j) + \dots \tag{2.7}$$

For the k-th such term, there are $n-k$ balls unrestricted, which gives $(n-k)!$ arrangements for these remaining balls. Since there are $n!$ total arrangements, the k-th term simplifies to $\sum \frac{(n-k)!}{n!}$. Since there are $\binom{n}{k}$ ways to choose the k balls that match, the summation must just be a product by this and so the k-th term is $\binom{n}{k}\frac{(n-k)!}{n!} = \frac{1}{k!}$. Finally, this gives our full answer

$$P\left(\bigcup_i A_i\right) = \sum_i p(A_i) - \sum_{i<j} P(A_i A_j) + \dots = \sum_k (-1)^{k+1}\frac{1}{k!} \tag{2.8}$$

$$= 1 - \frac{1}{e} \tag{2.9}$$

Chapter 3

12/01/15 — Binomial Distribution, Independence

A *Bernoulli trial* is a random experiment with only two outcomes; the mean is typically denoted p. The probability of k successes in n *independent* Bernoulli trials is given by

$$P = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.1}$$

We can introduce an example of this in flipping a coin $2n$ times; what's the probability we get heads n times? Clearly $\binom{2n}{n}\frac{1}{2^{2n}}$. To examine how this goes for large n we need Stirling's formula

$$n! \approx e^{-n} n^n \sqrt{2\pi n} \tag{3.2}$$

for $n \to \infty$. We can then apply it trivially to our problem above as $\binom{2n}{n} \approx \frac{2^{2n}}{\sqrt{\pi n}}$. Then we find that $P \propto n^{-1/2}$, which means this actually vanishes as we take the $n \to \infty$ limit. This shouldn't surprise us because it becomes harder and harder to have *exactly* one half be heads/tails.

3.1 Multinomial distribution, "Negative binomial distribution"

Straightforward generalization of binomial distribution,

$$P(k_i \text{ outcomes of type } i, i = 1, \dots m) = \frac{n!}{k_1! k_2! \dots k_m!} p_1^{k_1} p_2^{k_2} \dots p_m^{k_m} \tag{3.3}$$

Note also that if we want the rth success to occur on trial t, we need $t-r$ failures, $r-1$ successes (!) and a success on trial t. This gives a slightly different expression

$$P = \binom{t-1}{r-1} p^r (1-p)^{t-r} \tag{3.4}$$

3.2 Independence, Bayes' Rule

We denote two events to be independent if $P(B|A) = P(B)$, or $P(AB) = P(B)P(A)$. Then the total sample space of the two random experiments is simply the Cartesian product plus enough sets to make a σ-algebra.

We can rearrange some stuff in the definition of conditional probability to arrive at Bayes' Rule

$$P(B|A)P(A) = P(A|B)P(B) \tag{3.5}$$

Jkay, the full form for a more complicated sample space looks naturally like

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum\limits_j P(A|B_j)P(B_j)} \tag{3.6}$$

$$= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c} \tag{3.7}$$

Chapter 4

14/01/15 — More Bayes' Law, expectation of random variables

Bayes' Law is useful for dealing with two-stage experiments. Particularly, in multi-stage experiments conditional probabilities are much easier to work with than absolute probabilities; in such cases both Bayes' Law and tree diagrams can be helpful in solving the problem.

4.1 Random Variables

We define

A *random varibale* on a probability space (S, Σ, P) is a real-valued function on S which has the property that for every interval I the inverse image of I is also an event.

There is a distinction between two variables being equal and having the same distribution; the latter doesn't even require an identical sample space! Consider a coin vs. parity of a die roll.

We define $p_X(x)$ the probability the random variable $X = x$. Moreover, the cumulative distribution function $F_X(x)$ is just $p_X(-\infty, t]$. Interesting note, many respect authors and most of the French define the CDF with an open interval, but we will adhere to Ah-MUHR-ken way and use a closed upper bound.

We define stochastic dominance such that X *stochastically dominates* Y if $P(X \geq t) \geq P(Y \geq t)$ and for at least one T this is a strict inequality. In other words, X is at least as probable as Y and is at least in one case strictly more probable.

The expectation of a random variable is exactly what we expect, $\mathbf{E}(X) = \sum xP(x)$. The expectation is also referred often as the *first moment*. Note that we will eventually claim the Law of Large Numbers, that the expectation in the long run of X is equal to $\mathbf{E}(X)$ provided the expected value is finite.

Chapter 5

16/01/15 — Higher expectations, densities

Two variables are called *independent and identically distributed*, abbreviated iid, if they have a common distribution function and are stochastically independent. Instead of typing out the definition of stochastically independent for the nth time, I will simply point out that pairwise independence doesn't imply independence of the ensemble as a whole: consider X, Y independent and Z their parity.

5.1 Continuous distributions

For continuous distributions, our cumulative distribution function definition doesn't change, but our probability mass function instead becomes a probability density function, such that

$$P(X \in [a, b]) = \int_a^b f(x)\, dx \tag{5.1}$$

Note it doesn't make sense to talk about a probability mass function for a continuous distribution! The expectation also has a similarly simple form $\mathbf{E}(X) = \int x f(x)\, dx$.

5.2 Expectation Properties

- Expectation is linear! $\mathbf{E}(aX + bY) = a\mathbf{E}(X) + b\mathbf{E}(Y)$.
- Expectatiton is *positive*, so if $X \geq 0$ then $\mathbf{E}(X) \geq 0$.

Expectation distributes simply for independent random variables $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$.

5.3 Jensen's Inequality

Let X be a random variable with finite expecation and let $f : \mathbf{R} \to \mathbf{R}$ over the range of X. Then

$$\mathbf{E}(f(X)) \geq f(\mathbf{E}(X)) \tag{5.2}$$

5.4 Variance and higher moments

The variance is defined to be $\mathbf{Var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \mathbf{E}(X^2) - (\mathbf{E}(X))^2$. Note that for independent random variables variances are additive $\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y)$.

In general, the nth central moment of X is $\mathbf{E}\left(X - \mathbf{E}(X)\right)^n$.

Chapter 6

21/01/15 — Normal Distribution

6.1 Proving the Principle of Inclusion/Exclusion

To do this we must first introduce the multinomial identity

$$1 \pm x_1)(1 \pm x_2) \dots (1 \pm x_n) = \sum_{k=0}^{n} \sum_{i_1 < \dots < i_k} (\pm 1)^k x_{i_1} \dots x_{i_k} \tag{6.1}$$

and indicator functions

$$\mathbf{1}_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \notin A \end{cases} \tag{6.2}$$

which obey $\mathbf{E}\left(\mathbf{1}_A\right) = P(A)$. This also happens to obey $\mathbf{1}_{AB} = \mathbf{1}_A \cdot \mathbf{1}_B$ and $\mathbf{1}_{A^c} = 1 - \mathbf{1}_A$.

6.2 Inclusion/Exclusion Principle

We go about this by

$$P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - P\left(\left(\bigcup_{i=1}^{n} A_i\right)^c\right) \tag{6.3}$$

$$= 1 - P\left(\bigcap_{i=1}^{n} A_i^c\right) \tag{6.4}$$

Because the expectation value operator is just the probability for an indicator function, we can rewrite this as

$$\mathbf{E}\left(\mathbf{1}\left[\left(\bigcup_{i=1}^{n} A_i\right)\right]\right) = 1 - \mathbf{E}\left(\mathbf{1}\left[P\left(\bigcap_{i=1}^{n} A_i^c\right)\right]\right) \tag{6.5}$$

$$= 1 - \mathbf{E}\prod_{i=1}^{n}\left(1 - \mathbf{1}_{A_i}\right) \tag{6.6}$$

Then expanding out using the multinomial identity we can eventually arrive at a proof.

6.3 Variances, standardized random variables

Note that since variance is linear, the variance of a sum is just the sum of the variances, while the variance of an average is the sum of the variances divided by the number of summands *squared*.

We then define the *standardization* of X a random variable to be $X^* = \dfrac{X - \mu}{\sigma}$ so that $\mathbf{E}(X^*) = 0, \mathbf{Var}(X^*) = 1$.

Note that for the binomial distribution, we can now discuss the expectation and variance of it; $\mathbf{E}(X) = np, \mathbf{Var}(X) = np(1-p)$, recalling the binomial distribution to be the distribution of n Bernoulli trials.

6.4 The Normal Distribution

The *normal distribution* or the Gaussian distribution sits atop all statistical distributions because of the Central Limit Theorem. It is parameterized by two parameters μ, σ, the mean and standard deviation respectively. There is no closed form expression for the cdf, but the probability density is given

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.7}$$

Note that this is properly normalized as $\int\limits_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \; dx = \sqrt{2\pi}$. The cdf of the standard normal is denoted by $\Phi(t)$, given below

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{t} e^{-\frac{x^2}{2}} \; dx \tag{6.8}$$

The celebrated and closely related *error function* denoted $\mathrm{erf}(t)$ is actually given by

$$\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}} \int\limits_{0}^{t} e^{-x^2} \; dx \tag{6.9}$$

$$\Phi(t) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{t}{\sqrt{2}}\right) \tag{6.10}$$

It is tradition to denote a standard noraml random variable by $Z \sim N(0,1)$. Note that any two normal distributions differ only by *scale* or *location*, because $N(\mu, \sigma^2) \sim \sigma Z + \mu$. It is worthy to note that the binomial distribution approximates the normal for large n.

6.5 Using the Normal approximation

We can use the normal approximation to determine whether the outcome of a binomial distribution experiment is "close" to the expected value. For example, in the case of our coin toss experiment, we can simply integrate under the normal distribution after standardizing the results to see what the probability of obtaining a result such as ours was.

We can also discuss the DeMoivre-Laplace theorem as to when we are well approximated by the normal distribution, but I won't do that here out of laziness (and because we almost certainly didn't talk about it in class, which I didn't go to). Refer to Lecture07 for the full discussion.