

# A Rough Guide to Probability

Elisha Peterson

January 19, 2005

## Contents

<b>1</b>	<b>Getting Oriented</b>	<b>2</b>
<b>2</b>	<b>Combinatorics</b>	<b>2</b>
<b>3</b>	<b>Basic Probability</b>	<b>2</b>
3.1	Conditional Probability . . . . .	2
3.2	The Odds Ratio . . . . .	3
<b>4</b>	<b>Random Variables</b>	<b>3</b>
4.1	Discrete Random Variables . . . . .	3
4.2	Continuous Random Variables . . . . .	4
4.3	Jointly Distributed Random Variables . . . . .	5
4.4	Properties of Expectation . . . . .	5
<b>5</b>	<b>Going Further</b>	<b>5</b>
5.1	Limit Theorems . . . . .	5
<b>6</b>	<b>The Road Ahead</b>	<b>6</b>

# 1 Getting Oriented

## 2 Combinatorics

Before one can ask “how likely?” one must ask “how many?”. The answer to that question is given by combinatorics. The basic question is: how many ways are there to choose  $k$  objects out of a set of  $n$ ? Well, the answer depends on whether order matters, whether they are distinct objects, and so on:

- There are  $n!$  ways to order a set of  $n$  objects;
- **Combination:** there are  ${}_nC_k = \binom{n}{k} = \frac{n!}{(n-k)!k!}$  ways to select  $k$  objects out of  $n$  when order does not matter;
- **Permutation:** there are  ${}_nP_k = \frac{n!}{(n-k)!}$  ways to make this selection when order matters;
- **Multichoose:** there are  $\binom{n}{n_1 \dots n_r} = \frac{n!}{n_1! \dots n_r!}$  ways to partition  $n$  objects into sets of size  $n_1, n_2, \dots, n_r$ .

These are related to powers of polynomials by the **Binomial Theorem**

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$$

and the **Multinomial Theorem**

$$(x_1 + \dots + x_r)^n = \sum_{n_1 + \dots + n_r = n} \binom{n}{n_1 \dots n_r} x_1^{n_1} \dots x_r^{n_r}.$$

## 3 Basic Probability

Probability theory uses the notion of an experiment, and an outcome for that experiment, also called an event. The **sample space** of events is the set of possible outcomes. We may, of course, talk about unions, intersections (denoted  $EF$  rather than  $E \cap F$ ), and complements of events (and sets of events). The probability of certain outcomes is given by the *probability measure*:

**Probability Measure:**

---

a function  $P : S \rightarrow [0, 1]$ , where  $S$  is the sample space, such that: (1)  $P(S) = 1$ ; and (2)  $P(\cup E_i) = \sum_i P(E_i)$  if the events  $E_i$  are disjoint.

---

For example, if all outcomes in a given sample space are equally likely, then the probability of event  $A$  occurring is  $P(A) = \frac{|A|}{|S|}$ .

One can conclude from these axioms that  $P(E^c) = 1 - P(E)$  and that if  $E \subset F$  then  $P(E) \leq P(F)$ . We can also show that for an increasing (decreasing) sequence of events  $E_n$ , we have  $\lim_{n \rightarrow \infty} P(E_n) = P(\lim_{n \rightarrow \infty} E_n)$ . The **Principle of Inclusion and Exclusion**, or **PIE**, states that  $P(E \cup F) = P(E) + P(F) - P(EF)$ , and can easily be extended to any number of sets.

### 3.1 Conditional Probability

If an experiment is repeated, the first outcome may affect the probability of the second. In an extreme example, if someone draws two balls out of a bag contains a black ball and a white ball, then the color of the second ball is completely determined by the color of the first. This is an example of *conditional probability*

### Conditional Probability:

the probability of an event  $E$  given that an event  $F$  has already occurred, given by

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{P(E)P(F|E)}{P(F)}.$$

Note that conditional probability  $P_F(E)$  is itself a probability measure.

The above formulae give rise to several computational rules: for two events, we have  $P(EF) = P(E)P(F|E)$  and  $P(E) = P(EF) + P(EF^c) = P(E|F)P(F) + P(E|F^c)P(F^c)$ . Both of these may be generalized for several events.

If the first outcome does not effect the second, then the two events are **independent** and  $P(EF) = P(E)P(F)$ , in which case  $P(E|F) = P(E)$ .

## 3.2 The Odds Ratio

The **odds ratio** of an event  $F$  is given by

$$\frac{P(F)}{P(F^c)} = \frac{P(F)}{1 - P(F)},$$

which uniquely determines the probability. If an event  $E$  has already occurred, the odds ratio becomes  $\frac{P(H|E)}{P(H^c|E)} = \frac{P(H)}{P(H^c)} \frac{P(E|H)}{P(E|H^c)}$ , so is just multiplied by some factor.

## 4 Random Variables

### 4.1 Discrete Random Variables

A **random variable**  $X$  may encode the outcomes of an experiment if they are numerical. For discrete outcomes, the **probability mass function** is  $p(x) = P(X = x)$ , and the **cumulative distribution function** is  $f(x) = P(X \leq x) = \sum_{a \leq x} p(a)$ . Note that  $\lim_{x \rightarrow -\infty} f(x) = 0$ ,  $\lim_{x \rightarrow \infty} f(x) = 1$ , and that  $f$  is right continuous. Given a random variable, we have:

- **expected value:** the linear function  $E[X] = \sum_x xp(x)$ ;
- expected value of a function  $g(X)$ :  $E[g(X)] = \sum_x g(x)p(x)$ ;
- **variance:**  $\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$ ;
- **standard deviation:**  $\text{StDev}(X) = \sqrt{\text{Var}(X)}$ .

### Examples

Many experiments have just two outcomes: success and failure. This simple experiment gives rise to several random variables:

- **Bernoulli random variable**  $X$ : a single experiment is performed, with probability  $p$  of success;
- **Binomial random variable**  $X^n$ :  $n$  independent Bernoulli trials are performed, with probability of  $i$  successes given by  $P(X^n = i) = \binom{n}{i} p^i (n - p)^{n-i}$ . This has expected value  $E[X^n] = np$  and variance  $\text{Var}(X^n) = np(1 - p)$ ;
- **Poisson random variable**  $X^\infty$ : an approximation of  $X^n$  for  $n$  large, given by  $P(X^\infty = n) = \frac{\lambda^n}{n!} e^{-\lambda}$ . The parameter  $\lambda$  is both the expected value and the variance;
- **Geometric random variable**  $\hat{X}$ : measures the time until the first success in independent Bernoulli trials, given by  $P(\hat{X} = n) = (1 - p)^{n-1} p$ . Here,  $E[\hat{X}] = \frac{1}{p}$  and  $\text{Var}(\hat{X}) = \frac{1-p}{p^2}$ ;

- **Negative binomial random variable**  $\hat{X}^r$ : measures the number of trials before  $r$  successes, given by  $P(\hat{X}^r = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$ . We have  $E[\hat{X}^r] = \frac{r}{p}$  and  $\text{Var}(\hat{X}^r) = \frac{r(1-p)}{p^2}$ .

When taking  $k$  objects from a sample of  $m$  white and  $n-m$  black objects, the probability of getting  $i$  white objects is given by the **hypergeometric random variable**  $X_g$ , with  $P(X_g = i) = \frac{\binom{m}{i} \binom{n-m}{k-i}}{\binom{n}{k}}$ . The expected value is  $E[X_g] = \frac{km}{n}$ .

## 4.2 Continuous Random Variables

Random variables may also be continuous (take the speeds of cars along a highway for example), in which case the *probability mass function*  $f(x)$  is defined on a continuous set. Rather than exact values, we usually speak of the random variable lying in some range of values  $B$ :  $P(X \in B) = \int_B f(x) dx$ . When  $f$  is continuous, the probability of a precise value being taken is always zero, but if  $f$  has a discontinuity at  $x$ , then the probability of  $X$  taking the value  $x$  is just the size of the jump. We have:

- *cumulative distribution function*:  $F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$  (so  $f(x) = F'(x)$ );
- *expected value of  $X$* :  $E[X] = \int_{-\infty}^{+\infty} x f(x) dx$ ;
- *expected value of a function  $g(X)$* :  $E[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx$ ;
- *variance*: as before,  $\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$ .

A function  $Y = g(X)$  of a continuous random variable  $X$  has its own probability mass function:  $f_Y(y) = f_X(g^{-1}(y)) \left( \frac{d}{dy} g^{-1}(y) \right)$ .

### Examples

- **Uniform Distribution**: all outcomes in an interval  $[a, b]$  are equally probable, giving a mass function  $f(x) = \frac{1}{b-a}$ ;
- **Normal Distribution**: given by  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$ , with parameters  $\mu$  (the mean) and  $\sigma$  (the standard deviation). It is an amazing fact that this distribution approximates the behavior of almost all experiments when the number of trials is very high;
- **Exponential Distribution**: given by  $f(x) = \lambda e^{-\lambda x}$ , with parameter  $\lambda$ . Its mean and variance are  $1/\lambda$  and  $1/\lambda^2$ .
- **Gamma Distribution**: given by  $f(x) = \lambda e^{-\lambda x} (\lambda x)^{t-1} / \Gamma(t)$  for  $x \geq 0$ , where  $\Gamma(t) = \int_0^{+\infty} e^{-x} x^{t-1} dx$  is the standard *gamma function*. It has expected value  $t/\lambda$  and variance  $t/\lambda^2$ ;
- **Beta Distribution**: given by  $f(x) = x^{a-1} (1-x)^{b-1} / B(a, b)$  for  $0 \leq x \leq 1$ , where  $B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$ . Its mean and variance are  $\frac{a}{a+b}$  and  $\frac{ab}{(a+b)^2(a+b+1)}$ .

The exponential distribution is closely related to the *hazard* or *failure rate function*  $\lambda(t) = \frac{f(t)}{1-F(t)}$ . Here,  $\lambda(t) dt$  represents the probability that an item which is  $t$  years old will fail within an additional time  $dt$ . For the exponential distribution,  $\lambda(t)$  is constant, so failure does not become more likely over time. Actually, the exponential distribution is the only one having a constant failure rate.

### 4.3 Jointly Distributed Random Variables

This section generalizes both random variables and conditional probability. When the outcome of two random values (discrete or continuous) are not independent, we have a **joint probability mass function** given by  $f(x, y) = P(X = x, Y = y)$  and a corresponding cumulative distribution function  $F(x, y)$ . Note that the cumulative functions for  $x$  and  $y$  can be found from  $F(x, y)$  by  $F_X(x) = \lim_{y \rightarrow \infty} F(x, y)$  and  $F_Y(y) = \lim_{x \rightarrow \infty} F(x, y)$ .

The random variables  $X$  and  $Y$  are independent iff their mass function factors into separate functions of  $x$  and  $y$ :  $f(x, y) = g(x)h(y)$ . Otherwise, we have a **conditional probability mass function** given by  $f_{X|Y}(x) = \frac{f(x, y)}{f_Y(y)}$ .

The distribution of the sum of two random variables is given by their **convolution**  $F_{X+Y}(a) = \int_{-\infty}^{+\infty} F_X(a-y)F_Y(y)dy$ . This allows us to show that the distribution of several independent normal random variables with parameters  $(\mu_i, \sigma_i^2)$  is again a normal random variable with parameters  $(\sum_i \mu_i, \sum_i \sigma_i^2)$ . Parameters also add for the sum of independent Poisson variables.

### 4.4 Properties of Expectation

The expected value of a function  $g(X, Y)$  is  $E[g(X, Y)] = \sum_{x, y} g(x, y)p(x, y)$  in the discrete case, and with an integral in the continuous case. Clearly, the sum of expected values is the expected value of the sums.

The **covariance** of  $X$  and  $Y$  is defined as  $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$ , which can be shown to equal  $E[XY] - E[X]E[Y]$ . Note that  $\text{Cov}(\sum_i X_i, \sum_j Y_j) = \sum_{i, j} \text{Cov}(X_i, Y_j)$ . Covariance is related to variance by the identity  $\text{Var}(\sum_i X_i) = \sum_i \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j)$ . We define the **correlation** between  $X$  and  $Y$  to be  $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$ .

Assuming that  $Y = y$ , we can define the **conditional expected value** of  $X$ . This is given by  $E[X|Y = y] = \sum_x xP(X = x|Y = y)$  in the discrete case and  $\int_{-\infty}^{+\infty} xf_{X|Y}(x)dx$  in the continuous case. Note that  $E[X] = E[E[X|Y]]$ ; as a consequence we have  $E[X] = \sum_y E[X|Y = y]P(Y = y)$  in the discrete case, and a corresponding formula in the continuous case.

We can also define the **conditional variance**:  $\text{Var}(X|Y = y) = E[(X - E[X|Y = y])^2|Y = y]$ , giving the formula  $\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$ .

Given a random variable  $X$ , the **moment generating function** is given by  $M_X(t) = E[e^{tX}]$ . The moments of  $X$  can then be found by differentiating  $M(t)$  and evaluating the result at  $t = 0$ . This uniquely determines the distribution of the random variable. We also have  $M_{X+Y}(t) = M_X(t)M_Y(t)$ , if  $X$  and  $Y$  are independent.

The **multivariate normal distribution** is defined for linear combinations of a finite set of independent standard normal random variables. They have **sample mean**  $\bar{X} = \sum_i X_i/n$  and **sample variance**  $S^2 = \sum_i \frac{(X_i - \bar{X})^2}{n-1}$ . Both  $\bar{X}$  and  $S^2$  are random variables, for independent identically distributed  $X_i$ , and they are independent. The variable  $\bar{X}$  has a normal distribution, with mean  $\mu$  and variance  $\sigma^2/n$ .

## 5 Going Further

### 5.1 Limit Theorems

The main idea in this section is proving that as the number of trials of an experiment grows very large, we will see a normal distribution.

First, we have two inequalities. The **Markov Inequality** states that  $P(X \geq a) \leq E[X]/a$  for nonnegative random variables. The **Chebyshev Inequality** states that  $P(|X - \mu| \geq k\sigma) \leq 1/k^2$  for all positive  $k$ . These can be used to prove:

**Central Limit Theorem:**

given independent identically distributed random variables  $X_i$  with mean  $\mu$  and variance  $\sigma^2$ , we have  $P\left(\frac{X_1+\dots+X_n-n\mu}{\sigma\sqrt{n}} \leq a\right)$  approaches  $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^a e^{-x^2/2} dx$  as  $n \rightarrow \infty$ . This just means that the sum  $\sum_i X_i$  approaches the normal distribution.

As a consequence, we have the **Strong Law of Large Numbers**, which states that the average success of  $n$  trials of such variables approaches their mean, as  $n \rightarrow \infty$ .

## 6 The Road Ahead