A Probability Refresher.

Moulinath Banerjee

University of Michigan

October 2, 2002

1 Probability

In talking about probabilities, the fundamental object is Ω, the sample space. Points (elements) in Ω are denoted (generically) by ω.

We assign probabilities to subsets of Ω.

Assume for the moment that Ω is finite or countably infinite. Thus Ω could be the space of all possible outcomes when a coin is tossed three times in a row or say, the set of positive integers.

A probability P is then a function from the power set (the class of all possible subsets) of Ω, which we will denote by $\mathcal{A}$, to the interval $[0,1]$ satisfying the following properties:

- (i) $P(\Omega) = 1$.
- (ii) $P(\phi) = 0$.
- (ii) If $\{A_n\}$ is a sequence of mutually disjoint subsets of Ω, then,

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

In general however Ω can be uncountably infinite (for example when Ω is $[0,1]$) in which case (for certain technical reasons that we do not need to go into) $\mathcal{A}$ is not taken to be the entire power set, but is chosen to be a smaller class. Thus we do not assign probabilities to all subsets of Ω, but only to

those that belong to $\mathcal{A}$. The class $\mathcal{A}$ is assumed to contain Ω and to be closed under complementation and countable union. P is then, a function from $\mathcal{A}$ to $[0,1]$ with the same properties as above. The properties (i), (ii) and (iii) are called the **basic axioms** of probability.

From these properties, it is easy to deduce the following:

- (a) For finitely many disjoint sets $A_1, A_2, \ldots, A_n$,
$$P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)\,.$$
- (b) $P(A) = 1 - P(A^c)\,.$
- (c) If $A \subset B$, then $P(A) \leq P(B)$.
- (d) For any two sets A and B (not necessarily disjoint),
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)\,.$$

Exercise: Using (i), (ii), (iii), deduce (a), (b), (c) and (d).

There is an interesting generalization of (d) to the case of more than 2 sets.

Proposition: For (not necessarily disjoint) sets $A_1, A_2, \ldots A_n$,
$$\begin{aligned} P(\cup A_i) &= \sum P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \\ &\quad - \ldots\ldots + (-1)^{n+1} P(\cap_{i=1}^n A_i)\,. \end{aligned}$$

A general proof of this will be skipped. It can be done using induction. However, a proof of the above equality, when Ω is a finite set and each element in Ω is equally likely follows from a slick combinatorial argument.

So, let $\Omega = \{1, 2, \ldots, N\}$ and let P be a probability such that for any i, $P(i) = 1/N$. Then, clearly for any subset A, $P(A) = \#(A)/N$. Proving the above proposition then boils down to establishing that,
$$\begin{aligned} \#(\cup A_i) &= \sum \#(A_i) - \sum_{i<j} \#(A_i \cap A_j) + \sum_{i<j<k} \#(A_i \cap A_j \cap A_k) \\ &\quad - \ldots\ldots + (-1)^{n+1} \#(\cap_{i=1}^n A_i)\,. \end{aligned}$$

So consider some element s belonging to $\cup A_i$. We need to show that s is counted exactly once on the right side of the above expression. Suppose that s belongs to k of the n sets. Then, on the right side of the above expression s is counted

$$k - \binom{k}{2} + \binom{k}{3} - \ldots + (-1)^{k+1} \binom{k}{k}$$

times. Call this number m. Then,

$$\begin{aligned} m &= \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} \\ &= 1 - 1 + \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} \\ &= 1 - \left(1 - \sum_{j=1}^{k} (-1)^{j} \binom{k}{j} \right) \\ &= 1 - \left(\sum_{j=0}^{k} \binom{k}{j} (-1)^{j} (1)^{k-j} \right) \\ &= 1 - (1-1)^{k} \\ &= 1 \,. \end{aligned}$$

This finishes the proof.

In statistical/practical contexts it helps to think of Ω as the set of all possible outcomes of a chance (random) experiment. Probabilities assigned to subsets of Ω are governed by the nature of the chance mechanism (like tossing a coin, throwing a die, shuffling a pack of cards etc.)

Conditional probability: Consider a fixed event B with $P(B) > 0$. The conditional probability of $A \mid B$ (read, A given B) is,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \,.$$

It is not difficult to show that this defines a valid probability. In other words if we define a function P_B from $\mathcal{A}$ to $[0, 1]$ as follows:

$$P_B(A) = P(A \mid B) \,,$$

then P_B satisfies the basic axioms of probability.

The key idea behind defining conditional probabilities is to update the probabilities of events given the knowledge that B happened. Note that for any event $C \subset B^c$, $P(C \mid B) = 0$, which is in accordance with common sense. On the other hand, for any $C \subset B$,

$$P(C \mid B) \geq P(C)\,;$$

in other words, the conditional probability of a sub-event of B, given that B happened is generally higher than the unconditional probability.

Joint probabilities can be expressed in terms of conditional and marginal probabilities in the following manner:

$$P(A \cap B) = P(A \mid B)\, P(B)\,,$$

provided $P(B) > 0$. Also,

$$P(A \cap B) = P(B \mid A)\, P(A)\,,$$

provided $P(A) > 0$. More generally, for sets $A_1, A_2, \ldots A_n$, we have,

$$P(\cap_{i=1}^{n} A_i) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2)\, \ldots\, P(A_n \mid \cap_{i=1}^{n-1} A_i)\,.$$

One of the most important consequences of the notion of conditional probability is **Bayes Rule**. Simple as it looks, its ramifications are profound. Bayes Rule gives us a way of obtaining the conditional probability of one event given another, when we know the marginal probabilities and the conditional probability of the second event given the first. We state Bayes Rule in the form of the following theorem.

Theorem: Suppose that $\{B_j\}$ is a (finite/infinite) partition of Ω and A is any event with $P(A) > 0$. Also suppose that $P(B_i) > 0$ for each i (without any loss of generality). Then,

$$P(B_j \mid A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A \mid B_j)\, P(B_j)}{\sum P(A \mid B_i)\, P(B_i)}\,.$$

Illustrating Bayes Rule: There are three cabinets, A, B and C, each of which has two drawers. Each drawer has one coin. A has two gold coins, B has two silver coins and C has one gold and one silver coin. A cabinet is

chosen at random; one drawer is opened and a silver coin found. What is the chance that the other drawer also has a silver coin ?

To fit this in the framework of the above theorem, take the B_j's to be the events that A is chosen, B is chosen and C is chosen. For brevity we shall denote the events by A, B and C. Note that these events are indeed a partition of the sure event that a cabinet is chosen, and furthermore

$$P(A) = P(B) = P(C) = 1/3\,.$$

Let $S1$ denote the event that the opened drawer of the chosen cabinet has a silver coin. Clearly, we are required to find $P(B \mid S1)$. Now, using Bayes Rule, we have

$$\begin{aligned} P(B \mid S1) &= \frac{P(S1 \mid B)\,P(B)}{P(S1 \mid A)\,P(A) + P(S1 \mid B)\,P(B) + P(S1 \mid C)\,P(C)} \\ &= \frac{1 \times 1/3}{0 \times 1/3 + 1 \times 1/3 + 1/2 \times 1/3} \\ &= \frac{1}{1 + 1/2} \\ &= 2/3\,. \end{aligned}$$

Independence of events: Events A and B are said to be independent if $P(A \cap B) = P(A) \times P(B)$. In general, events $A_1, A_2, \ldots, A_n$ are said to be mutually independent if for any subcollection $\{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}$ of $\{A_1, A_2, \ldots, A_n\}$, it is the case that,

$$P(A_{i_1} \cap A_{i_2} \cap \ldots A_{i_k}) = P(A_{i_1}).P(A_{i_2}).\cdots P(A_{i_k})\,.$$

Exercise: If $A_1, A_2, \ldots, A_n$ are independent, then so are $B_1, B_2, \ldots B_n$ where each B_i is either A_i or A_i^c.

Consider events A and B with $P(A) > 0$, $P(B) > 0$. Then the independence of A and B is equivalent to saying that,

$$P(A \mid B) = P(A)\,,$$

and equivalently,

$$P(B \mid A) = P(B)\,.$$

In other words knowledge about B does not affect the probability of the event A and vice-versa, which is intuitively compatible with the notion of (stochastic) independence.

2 Random Variables

A **random variable** X is a function from Ω to the real numbers. Thus, for each ω, $X(\omega)$ is a real number. In talking about the value of a random variable at a particular sample point, the argument ω is usually suppressed.

Most of probability and statistics deals with the study of random variables. Random variables can broadly be classified into two types. These are (i) Discrete and (ii) Continuous. Discrete random variables are those that assume finitely many or countably many values with positive probability. On the other hand, for a continuous random variable X the probability that X assumes a particular value x is 0 for any x. However, the probability that X lives in some interval $[a, b]$ can be positive; in fact, the probability that X lives in $\mathbb{R}$ is 1. This apparently smells paradoxical; if the chance of any particular value in $[a, b]$ being realized is 0, how can the probability of an interval $[a, b]$ even be positive? This is where the subtlety of the theory lies. An interval contains uncountably many points and the addition rule for probabilities only holds for finitely or countably many events; thus the probability of an interval cannot be computed by adding up the probability of each point in that interval, namely 0. Yet chances can be meaningfully assigned to subsets of $\mathbb{R}$ and be meaningfully interpreted. If we couldn't do this, there wouldn't have been much use for the theory.

Example 1: This is an example of a discrete random variable. Consider the coin-tossing experiment, where a coin is flipped once. The sample space $\Omega = \{0, 1\}$. Let X be the random variable that assumes the value 1 if heads comes up, and 0 if tails comes up. Thus $X(H) = 1$ and $X(T) = 0$. If the chance of the coin landing heads up is p, then clearly,

$$P(X = 1) = p \text{ and } P(X = 0) = q = 1 - p\,.$$

The random variable X is called a Bernoulli random variable.

Example 2: This is also a discrete random variable that is composed by adding independent Bernoulli random variables. Suppose you flip a coin n times and record the outcomes of each toss. The space of all possibilities is the sample space Ω and is made up of sequences of length n where each slot is either H (heads) or T (tails). What is the size of Ω ?

The random variable X records the number of heads in a sequence. Thus if ω denotes a sequence, then $X(\omega) =$ number of heads in ω. Thus X can

assume any value between 0 and n. It is easy to see that,

$$P(\{\omega\}) = p^{X(\omega)}\, q^{n-X(\omega)}\,,$$

and consequently,

$$P(X = m) = \binom{n}{m} p^m\, q^{n-m}\,.$$

Note that X is the sum of n independent Bernoulli random variables.

Example 3: This is a discrete random variable that takes infinitely many values unlike the previous two. Suppose, we keep on flipping a coin, till we get a heads and then stop. Let the probability of the coin landing heads up in a flip be p. The sample space is then given by

$$\Omega = \{H, TH, TTH, TTTH, \cdots\}\,.$$

Let X denote the number of flips needed to get the first head. Clearly X can be any positive integer. Then X is clearly a random variable and it is easily seen that,

$$P(X = m) = q^{m-1}\, p\,.$$

The geometric distribution has an interesting property called the **memoryless** property. This can be stated as follows: Let X be a geometric random variable. Given any two positive integers m and n,

$$P(X > m + n \mid X > m) = P(X > n)\,.$$

The reason why the above is referred to as the memoryless property is fairly clear. Given the information that you haven't seen a head in the first m tosses, the chance that you won't see a head in n further tosses, is the same as the probability that you don't see a head in the first n tosses. So in a sense you are starting from square one, from the $m + 1$'st toss onwards.

A distribution that arises very naturally from the geometric is the negative binomial distribution. Often useful in analyzing discrete data, it can be described very simply, in terms of the following random experiment: Suppose I keep tossing a coin till I get the r'th heads (success). Let X be the number of tails that I get before the r'th heads. Clearly X is a random variable. What is the chance that $X = x$? To compute this, note that for X to be equal to x, there must have been a total of $x + r$ tosses including the r'th heads and in the first $x + r - 1$ tosses there must have been $r - 1$ heads. Now, the total number of (distinct) sequences of H's and T's that are $x + r$

long and end in an H and have $r-1$ heads in the first $x+r-1$ slots is just $\binom{x+r-1}{r-1} \equiv \binom{x+r-1}{x}$. The chance than any such sequence is actually realized is just $p^r\, q^x$. Hence,

$$P(X = x) = \binom{x+r-1}{x} p^r\, q^x\,.$$

Note that X can assume any value greater than or equal to 0 with positive probability.

Exercise: If $W = X + r$ is the total number of tosses needed to get r heads, write down the p.m.f. of W and show that $X_1 + X_2 + \ldots + X_r$ where the X_i's are i.i.d. geometric (as defined above) is distributed like W.

Exercise: Prove that the geometric distribution satisfies the memoryless property.

The interesting fact is that among all discrete random variables supported on $\{1, 2, 3, \ldots\}$ the geometric distribution is the only one that satisfies the memoryless property. In other word, the memoryless property completely characterizes the geometric distribution.

Other examples of discrete random variables include Poisson, Negative Binomial, etc.

The **probability mass function** of a discrete random variable x taking values in, say, $\{0, 1, 2, 3, \ldots\}$ is the function p from the set $\{0, 1, 2, 3, \ldots\}$ to the set $[0, 1]$ defined by,

$$p(i) = P(X = i)\,.$$

Example 4: This is an example of a continuous random variable. Let the sample space $\Omega = (0, 1)$ and let P be a probability defined (uniquely) on Ω in the following way. For any $0 < x < 1$,

$$P((0, x]) = x\,.$$

This is a good probability model for drawing a point at random from the interval $(0, 1)$. What it says is that the chance of a randomly chosen point belonging to an interval is precisely given by the length of the interval. Let

U be a random variable defined as follows:

$$U(\omega) = \omega\,.$$

Then, it is easy to see that,

$$P(U \leq u) = u\,,$$

for any $0 < u < 1$. The random variable U is said to have the uniform distribution on $(0,1)$. In some sense, U is the most basic or fundamental random variable, because, as we shall see presently, all random variables can be generated from U.

The probabilistic behavior of a random variable is captured completely by its **distribution function**. If X is a random variable, it's distribution function F_X, henceforth denoted by F, is defined by,

$$F(x) = P(X \leq x)\,.$$

F has the following properties.

- (i) $0 \leq F(x) \leq 1\,.$
- (ii) If $x < y$, then $F(x) \leq F(y)$. Also,

$$\lim_{y \to x+} F(y) = F(x)\,.$$

  In other words F is monotone increasing and right continuous.

- (iii) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1\,.$

F need not be left-continuous as can be shown by counter-example. Formally, a continuous random variable is defined as one whose distribution function is continuous.

For any $0 < p < 1$, the p'th quantile of F is any number x_p such that

$$F(x_p-) \equiv \lim_{y \to x_p-} F(y) = P(X < x_p) \leq p \leq P(X \leq x_p) = F(x_p)\,.$$

Clearly, the p'th quantile need not be unique. However, if F is a strictly increasing continuous function, in which case it is one-one, it's inverse function F^{-1} is well defined on $(0,1)$. For any $0 < p < 1$, $F^{-1}(p)$ is the unique number x, such that $F(x) = p$, and $F^{-1}(p)$ is the unique p'th quantile of F.

When $p = 0.5$, we refer to the quantile as the median.

Fortunately, there is a neat way to define the inverse function F^{-1} even when F is not strictly increasing and continuous. For any $0 < t < 1$, we set,

$$F^{-1}(t) = \text{smallest element of } \{x : F(x) \geq t\}\,.$$

The fact that the above set does indeed have a smallest element can be shown. It is now easy to see that $F^{-1}(t)$ is indeed a t'th quantile of F, though not necessarily the unique one. Note that,

$$F^{-1}(t) \leq x \Leftrightarrow F(x) \geq t\,.$$

We can now state a very crucial theorem.

Theorem: Let X be a random variable with distribution function F. Let U be a Uniform random variable on $(0,1)$. Then $Y = F^{-1}(U)$ is also a random variable and its distribution function is F.

Proof: We have,

$$P(F^{-1}(U) \leq x) = P(F(x) \geq U) = P(U \leq F(x)) = F(x)\,.$$

Thus, by knowing F, and hence in principle F^{-1}, one can generate a random variable with distribution function F, provided one can generate from a $U(0,1)$ distribution. Another related theorem follows.

Theorem: If X is a continuous random variable, then $F(X)$ has the uniform distribution.

The proof of this theorem, in the case that F is strictly increasing and continuous, is given in Rice. The general case is omitted.

For discrete random variables, we can obtain the distribution function F from the mass function p in the following way: Suppose that X is a discrete random variable taking on the values $x_1, x_2, x_3, \ldots$ with probabilities $p_1, p_2, p_3, \ldots, .$ If p denotes the probability mass function, then $p(x_i) = p_i$. The distribution function F of X is then given by,

$$F(x) = \sum_{i:x_i \leq x} p(x_i)\,.$$

Many continuous distribution functions (and in fact, the vast majority of those which we deal with in statistics) possess a **probability density function**, which is basically the analogue of the probability mass function in the discrete case. A continuous distribution function F is said to possess a density f if there exists a non-negative real-valued function f, such that for all $a < b$,

$$F(b) - F(a) = P(a < X \leq b) = \int_a^b f(x)\,dx\,.$$

If f is continuous, then it is easy to see that $f(x)$ is the derivative of F at x. Also, note that the integral of f over the entire line must be 1; i,e.

$$\int_{-\infty}^{\infty} f(x)\,dx = 1\,.$$

The uniform distribution on $(0,1)$ is easily seen to have density function f, given by $f(x) = 1$ whenever $0 < x < 1$ and 0 elsewhere. Note that if f is a density function for a distribution function F, and g is obtained by changing the values of f at finitely many points, then g is also a density function.

Many distributions are specified using density functions rather than the distribution functions themselves. In particular, one of the most crucial distributions in statistics, the normal distribution, which arises all over the place, is specified through it's density. The distribution function of the normal cannot be written in closed form. Recall that a normal random variable with mean μ and standard deviation σ has density,

$$\phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\,\pi}\,\sigma} \exp\left(-\frac{1}{2\,\sigma^2}(x-\mu)^2\right) .$$

The exponential distribution, which is used to model waiting times, has density function given by,

$$f(x) = \lambda \exp(-\lambda\, x)\,,$$

where λ is some positive constant. Its distribution function is available in closed form as,

$$F(x) = 1 - \exp(-\lambda\, x)\,.$$

Among all continuous distributions on $(0,\infty)$, the exponential distribution is the only one that has the memoryless property.

Exercise: Show that the exponential distribution has the memoryless property; i.e. if X is an exponential random variable and $s, t > 0$, then,

$$P(X > s + t \mid X > s) = P(X > t)\,.$$

The converse is more difficult to prove, but a fun exercise. We'll skip it for now.

Exercise: Given a uniform random variable U, how would you generate from an exponential distribution with parameter λ ?

Transforming Random Variables: In statistics, we are often interested in deriving the distribution of a given function of a random variable. Thus, if X is a random variable, we might be interested in finding the distribution

of $g(X)$, where g is a fixed pre-specified function. We consider, first, the case when X is discrete.

Theorem: Suppose X is a discrete random variable assuming values $\{x_1, x_2, x_3, \ldots\}$ with probabilities $\{p_1, p_2, p_3, \ldots\}$. Let $Y = g(X)$, where g is a given function and let $\{y_1, y_2, y_3, \ldots\}$ be the values assumed by Y. Then the p.m.f. of Y is given by:

$$P(Y = y_i) \; = \sum_{j:g(x_j)=y_i} p_j \, .$$

The proof is immediate.

For continuous random variables with a density f, the **Jacobian Theorem** gives us the distribution of a transformed variable under some regularity conditions on the transformation g. We state the theorem below (without proof).

Change of variable theorem: Let X be a continuous random variable with density function f and g be a real-valued function defined on some open interval I, such that $P(X \in I) = 1$. Assume further that g is continuously differentiable and that $g^{'}(x) \neq 0$ for $x \in I$(these assumptions actually entail that g is a strictly monotone transformation on I). Let $Y = g(X)$. Then the density function of Y can be computed as,

$$f_Y(y) = f(g^{-1}(y)) \mid \frac{d}{dy} g^{-1}(y) \mid = f(g^{-1}(y)) \mid \frac{1}{g^{'}(g^{-1}(y)} \mid, \;\; y \in g(I) \, ,$$

and 0 otherwise.

Proof: Suppose $I = (a, b)$ and g is increasing. Then $a < X < b$ with probability 1 and the density f can be taken to be identically 0, outside of I. Also $g(a) < Y \equiv g(X) < g(b)$ with probability 1 and the density f_Y of Y can be taken to be identically 0 outside of $g(I)$. Let F_Y denote the distribution function of Y. Let $g(a) < y < g(b)$. Then,

$$F_Y(y) = P(Y \leq y) = P(g(a) \leq Y \equiv g(X) \leq y) = P(a \leq g^{-1}(Y) \leq g^{-1}(y)) \, .$$

Since, $X \equiv g^{-1}(Y)$,

$$F_Y(y) = P(a \leq X \leq g^{-1}(y)) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \, .$$

On differentiating the above, with respect to y, we obtain,

$$f_Y(y) = f(g^{-1}(y))\,\frac{d}{d\,y}\,g^{-1}(y) = f(g^{-1}(y))\,\left|\frac{1}{g^{'}(g^{-1}(y)}\right|\,,$$

where the last equality follows from the chain rule and the fact that g and g^{-1} have positive non-vanishing derivatives. $\Box$

The equality in the above display clearly implies that for any subset S of I:

$$\int_S f(x)\,dx = \int_{g(S)} f(g^{-1}(y))\;\mid \frac{1}{g^{'}(g^{-1}(y)}\mid dy\,. \tag{1}$$

In fact, the above formula is valid very generally; f does not need to be a density function. Any non-negative function, or for that matter, any function, with a well-defined (finite) integral will do. Let's see how we can apply this formula to compute an integral, that one might come across in calculus. Suppose we wish to compute,

$$\mathcal{S} = \int_0^1 u\,sin\,u^2\,du\,.$$

We write

$$\int_0^1 u\,sin\,u^2\,du = \int_0^1 \frac{1}{2}\,2u\,sin(u^2)\,du\,,$$

and then set $w = u^2$, noting that as u runs from 0 t0 1, so does w in the same direction. But $w = u^2$ means that $dw = 2u\,du$ and thus

$$\int_0^1 \frac{1}{2}\,2u\,sin(u^2)\,du = \int_0^1 \frac{1}{2}\,sin\,w\,dw\,,$$

by direct substitution and we obtain,

$$\mathcal{S} = \frac{1}{2}\,(1 - cos\,1)\,.$$

Basically, what we have done here is use (1) informally. To see this, let

$$f(x) = \frac{1}{2}\,2x\,sin\,x^2 \quad I = (0,1) \;\text{ and }\; g(x) = x^2\,.$$

Clearly, g is continuously differentiable on I and $g^{'}$ is non-vanishing (since $g^{'}(x) = 2x$)). Now, the inverse transformation g^{-1} from $g(I) = (0,1)$ to I is

$$g^{-1}(y) = \sqrt{y}\,.$$

Also,

$$f(g^{-1}(y)) = \frac{1}{2}\, 2\sqrt{w}\, sin\, w\,.$$

Thus,

$$\begin{aligned}
\mathcal{S} &= \int_I f(x)\, dx \\
&= \int_{g(I)} f(g^{-1}(y)) \mid \frac{1}{g'(g^{-1}(y)} \mid dy \\
&= \int_0^1 \frac{1}{2}\, 2\sqrt{y}\, sin\, y\, \frac{1}{2\sqrt{y}} \\
&= \frac{1}{2} \int_0^1 sin\, y\, dy \\
&= \frac{1}{2}\, (1 - cos\, 1)\ .
\end{aligned}$$

There is an useful extension of the Jacobian theorem to the case where the transformation is not one-one. This is stated below.

Extension of the change of variable theorem: Let X be a continuous random variable with density function f and g be a real-valued function (but not necessarily one-one) defined on some open set I, such that $P(X \in I) = 1$. Let $V = g(I)$. Suppose that there exist open-intervals $I_1, I_2, \ldots I_k$ which are mutually disjoint, with $I = I_1 \cup I_2 \cup \ldots \cup I_k$, such that the restriction of g to I_i, say g_i is a one-one continuously differentiable function from I_i onto $g(I)$ with non-vanishing derivative g_i'. Let $Y = g(X)$. Then the density function of Y can be computed as,

$$f_Y(y) = \sum_{i=1}^{k} f(g_i^{-1}(y)) \mid \frac{d}{dy}\, g_i^{-1}(y) \mid, \quad y \in V = g(I)$$

and 0 otherwise.

We can use the above theorem to obtain the density of $Y = X^2$ where X is a random variable with density f on $(-\infty, \infty)$. We let $I = (-\infty, 0) \cup (0, \infty)$ and note that $P(X \in I) = 1$. Note that $g(x) = x^2$ is not one-one but g restricted to $I_1 \equiv (-\infty, 0)$ is and so is g restricted to $I_2 \equiv (0, \infty)$. Let g_i be the restriction of g to I_i. Note that $V = g(I) = (0, \infty)$. Now,

$$g_1^{-1}(y) = -\sqrt{y}$$

and

$$g_2^{-1}(y) = \sqrt{y}\,.$$

The density of Y is thus given by,

$$f_Y(y) = f(-\sqrt{y})\,\frac{1}{2\sqrt{y}} + f(\sqrt{y})\,\frac{1}{2\sqrt{y}},\quad y > 0\,,$$

and 0 otherwise. Alternatively, the density can be computed from first principles as in Rice.

3 Multidimensional Random Variables

By a multidimensional random variable we mean a (2 or more dimensional) vector, such that each component of the vector is a real-valued random variable. Each component of the random vector must be defined on the same probability space. In other words, the components of the vector represent numerical features of the same underlying random experiment. Consider, for example, two consecutive (independent) throws of a die. Let X denote the sum of the numbers on the two throws and Y denote the absolute magnitude of the difference of the numbers on the two throws. Thus $X = X_1 + X_2$ and $Y = \mid X_1 - X_2 \mid$ where X_1 and X_2 are the outcomes of the two throws. Then both (X_1, X_2) and (X, Y) are 2 dimensional random variables.

We can talk about the joint distribution of (X_1, X_2) or (X, Y). The random vectors we are concerned with here are discrete random vectors which assume finitely many values with positive probability and the joint distribution is described completely by the joint probability mass function. Consider, firstly, (X_1, X_2). This assumes values in the set $\mathcal{S}_1 = \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\}$ and the p.m.f. is,

$$p_{X_1,X_2}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)\,P(X_2 = x_2) = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}\,,$$

for $(x_1, x_2) \in \mathcal{S}_1$ and is 0 otherwise. Here, we have used the independence of the random variables X_1 and X_2. However, the random variables X and Y are not independent. This is readily seen on noting that,

$$P(X = 12, Y = 1) = 0 \neq P(X = 12)\,.\,P(Y = 1) > 0\,.$$

The random vector (X, Y) takes values in the set $\{2, 3, \ldots, 12\} \times \{0, 1, 2, 3, 4, 5\}$ though not all pairs of values have positive probability. The p.m.f. of (X, Y)

is obtained readily as,

$$p_{X,Y}(x,y) = P(X = x, Y = y) = \frac{n_{x,y}}{36} \;\; (x,y) \in \mathcal{S}_2 \,,$$

where $n_{x,y}$ is the total number of pairs $(i,j) \in \mathcal{S}_1$ such that $i + j = x$ and $\mid i - j \mid = y$. To find the marginal probability mass functions of X and Y, we proceed thus. For $x \in \{2, 3, \ldots, 12\}$,

$$p_X(x) = P(X = x) = \sum_{y \in \{0,1,\ldots,5\}} P(X = x, Y = y) = \sum_{y \in \{0,1,\ldots,5\}} p_{X,Y}(x,y) \,.$$

Similarly, for $y \in \{0, 1, \ldots, 5\}$,

$$p_Y(y) = P(Y = y) = \sum_{x \in \{2,3,\ldots,12\}} P(X = x, Y = y) = \sum_{x \in \{2,3,\ldots,12\}} p_{X,Y}(x,y) \,.$$

The joint distribution function of (X,Y) is obtained as,

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = \sum_{(u,v): u \leq x, v \leq y} p_{X,Y}(u,v) \,.$$

Let (X,Y) be a two-dimensional continuous random vector. Thus $P(X = x, Y = y) = 0$ for all (x,y). Also assume that (X,Y) has a density function $f(x,y)$. What this means is the following: There exists a non-negative function $f(x,y)$, such that for any nice ("measurable") subset of $\mathbb{R}^2$, the probability that (X,Y) assumes values in $\mathbb{R}^2$ can be represented as

$$P((X,Y) \in A) = \int_A f(x,y)\, dx\, dy \,.$$

Thus, the volume enclosed by the surface $\{x, y, f(x,y)\}$ in x–y–z space over the area A gives the chance that (X,Y) takes values in A. The distribution function of (X,Y) can be computed as

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = \int_{(-\infty,x] \times (-\infty,y]} f(u,v)\, du\, dv \,.$$

Let's discuss a concrete example. Consider a helicopter landing randomly inside a circular helipad. Take the center of the helipad to be the origin (0,0) and assume that we scale the unit of distance in such a way, that the circular helipad is the unit circle, given by the equation $x^2 + y^2 \leq 1$. Then the point (X,Y) at which the helicopter lands is a two-dimensional continuous

random variable that is distributed uniformly inside the unit circle and has density,

$$f_{X,Y}(x,y) = \frac{1}{\pi}\, 1\{x^2 + y^2 \le 1\}\,,$$

where $1\{x^2+y^2 \le 1\}$ is the indicator function of the unit circle, and assumes the value 1 if (x, y) lies in the unit circle (which is the same as saying that $x^2 + y^2 \le 1$) and 0 otherwise. The chance that (X, Y) lies in a subregion $\mathcal{R}$ of the unit circle is simply Area$(\mathcal{R})/\pi$. To find the marginal densities of X and Y, which are denoted by f_X and f_Y respectively, we proceed thus:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy \;\text{ for } (-1 \le x \le 1)\,.$$

Since $f(x, y)$ is non-zero only when $x^2 + y^2 \le 1$, the values of y that contribute to the integral above are $-\sqrt{1-x^2} \le y \le \sqrt{1-x^2}$. Thus,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\, dy \\ &= \int_{-\infty}^{\infty} \frac{1}{\pi}\, 1\{x^2 + y^2 \le 1\}\, dy \\ &= \int_{-\infty}^{\infty} \frac{1}{\pi}\, 1\{-\sqrt{1-x^2} \le y \le \sqrt{1-x^2}\} \\ &= \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\, dy \\ &= \frac{2}{\pi}\sqrt{1-x^2}\,. \end{aligned}$$

So we can write down the marginal density of X as,

$$f_X(x) = \frac{2}{\pi}\sqrt{1-x^2}\; 1\{-1 \le x \le 1\}\,.$$

By symmetry, we can immediately write down the marginal density of Y as,

$$f_Y(y) = \frac{2}{\pi}\sqrt{1-y^2}\; 1\{-1 \le y \le 1\}\,.$$

You can deduce quite easily that X and $-X$ have the same distribution and also that Y and $-Y$ have the same distribution by using Problem (1) from Homework 2. Thus X and Y are both symmetrically distributedaround 0 and therefore have 0 means. In fact, using Problem (2) from Homework 2, you can also deducethat (X, Y) has the same distribution as $(e_1\, X, e_2\, Y)$

where e_1 is 1 or -1 and e_2 is also 1 or -1. It follows that X and Y are uncorrelated (once again, from Problem (2)). The fact that X and Y have the same marginal distributions is tied to the fact that you can swap the values of x and y in $f(x,y)$ without changing the value of the density. If the helipad was elliptic instead of being circular, this would no longer be the case, since $f(x,y)$ would no longer necessarily equal $f(y,x)$. For example, consider the uniform distribution on the ellipse with major axis having length 2 and minor axis having length 1. The density function is given by,

$$g(x,y) = \frac{1}{2\,\pi}\,1\left\{\frac{x^2}{4} + y^2 \leq 1\right\}.$$

If (X,Y) has the uniform distribution on this ellipse, then you can still show that changing the sign of one or more of the components will not change the distribution; thus $(-X,Y),(X,-Y),(-X,-Y)$ all have the same distribution as (X,Y), and consequently X and Y are uncorrelated as in the previous case. However X and Y no longer have the same marginal distribution; you can figure this out by simply looking at the shape of the ellipse and noting that it is more stretched out along the horizontal axis than the vertical (which is NOT the case with the circle). See Figure 1 where the elliptical helipad is sketched. If you think a bit, you'll also see that the marginal distribution of X must concentrate on $(-2,2)$ whereas the marginal distribution of $(-Y,Y)$ concentrates on $(-1,1)$. I would advise working out the marginal distributions as an exercise. Choosing $(x,y) = (1.5,0)$ we see that $g(x,y) = \frac{1}{2\pi} \neq 0 = g(y,x)$.

Let's return to the circular helipad example. We want to compute the conditional distribution of X given $Y = y$; so, contingent on the information that the Y co-ordinate of the point where the helicopter landed is y, what is the chance that the X co-ordinate is less than some specificied value, say x? In other words, what is $P(X \leq x \mid Y = y) \equiv F_{X|Y=y}(x)$? Here $F_{X|Y=y}$ is the *conditional distribution function* of X given that $Y = y$. We find this by evaluating the conditional density of X given $Y = y$, which is

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

we only define the conditional density of X given $Y = y$ for y such that the marginal density of Y at y is positive. In this case, for example, Y never lies outside of $(-1,1)$, so it does not make sense to condition on Y being equal

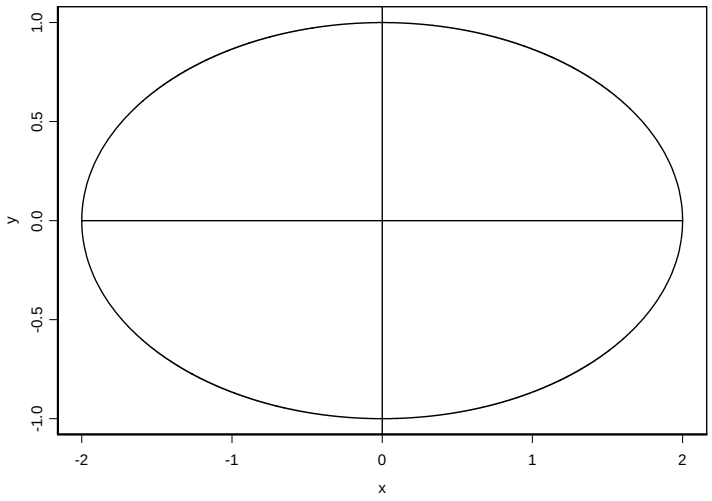

Figure 1: The uniform distribution on the ellipse $x^2/4 + y^2 \leq 1$.

to 1.5. For $-1 < y < 1$, we have,

$$\begin{aligned} f_{X|Y=y}(x) &= \frac{f(x,y)}{f_Y(y)} \\ &= \frac{1}{\pi} \frac{1}{(2/\pi)\sqrt{1-y^2}} 1\{x^2 \leq 1-y^2\} \\ &= \frac{1}{2\sqrt{1-y^2}} 1\left\{-\sqrt{1-y^2} \leq x \leq \sqrt{1-y^2}\right\}. \end{aligned}$$

Thus, X given $Y = y$ has the uniform distribution on $(-\sqrt{1-y^2}, \sqrt{1-y^2})$.

We extend this example a little. Suppose now that two helicopters land randomly inside the circular helipad. The chance that they land at the same point is 0, so we'll ignore that event. Let (X_1, Y_1) and (X_2, Y_2) denote their landing points. If we assume that the helicopters land independently of each other (may not sound too realistic, but then this is more of a thought experiment!! A helicopter seldom occupies only a single point..but a point can be a good approximation for a helicopter if the dimensions of the helipad are much much larger than the helicopter itself.) then (X_1, Y_1) and (X_2, Y_2) are i.i.d. and each is distributed uniformly on the unit circle. Let D denote

the distance between the copters. Then

$$D = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}\,.$$

We want to find the average squared distance between the copters. Thus, we seek to compute,

$$E(D^2) = E\left[(X_1 - X_2)^2\right] + E\left[(Y_1 - Y_2)^2\right].$$

Noting that $E(X_1 - X_2) = E(Y_1 - Y_2) = 0$, and that for a random variable Z, $\mathrm{Var}(Z) = E(Z^2) - (EZ)^2$, we get,

$$E(D^2) = \mathrm{Var}(X_1 - X_2) + \mathrm{Var}(Y_1 - Y_2)\,.$$

Note that (X_1, X_2) and (Y_1, Y_2) are two pairs of independent observations from the marginal distribution of X or Y, and therefore the joint distribution of (X_1, X_2) is the same as the joint distribution of (Y_1, Y_2). Thus,

$$\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(Y_1 - Y_2)$$

and consequently,

$$\begin{aligned}
E(D^2) &= 2\,\mathrm{Var}(X_1 - X_2) \\
&= 2\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) - 2\,\mathrm{Cov}(X_1, X_2)\right. \\
&= 2\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2)\right) \ (\text{since } X_1 \text{ independent of } X_2) \\
&= 4\,\mathrm{Var}(X_1)\,,
\end{aligned}$$

on using the fact that X_1 and X_2 are identically distributed. It remains to compute the variance of X_1 and this is slightly involved (though not really messy). We have,

$$\begin{aligned}
\mathrm{Var}(X_1) &= E(X_1^2) \\
&= \int_{-1}^{1} 2\,x^2\,\frac{1 - x^2}{\pi}\,dx \\
&= \frac{1}{\pi}\int_0^{\pi} 2\,cos^2(\theta)\,sin^2(\theta)\,d\,\theta \quad (\text{setting } x = sin(\theta)) \\
&= \frac{1}{2\,\pi}\int_0^{\pi} sin^2(2\,\theta)\,d\,\theta \quad \text{since } 2\,sin\,\theta\,cos\,\theta = sin\,2\,\theta \\
&= \frac{1}{8\,\pi}\int_0^{\pi} (2\,sin^2(2\,\theta))\;2\,d\,\theta \\
&= \frac{1}{8\,\pi}\int_0^{2\,\pi} (2\,sin^2(\phi))\;d\,\phi \quad (\text{setting } 2\,\theta = \phi)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{8\,\pi}\int_0^{2\,\pi}(1-cos\,2\,\phi))\;d\,\phi \;\;(\text{since}\;\; 1-2\,sin^2(\phi)=cos\,2\,\phi)\\
&= \frac{1}{4}-\frac{1}{8\,\pi}\int_0^{2\,\pi}cos\,2\,\phi\;\;d\,\phi\\
&= \frac{1}{4}-\frac{1}{16\,\pi}\int_0^{4\,\pi}cos\,\psi\;\,d\,\psi\\
&= \frac{1}{4}\,;
\end{aligned}
$$

this follows on noting that the integral on the penultimate line of the above display is 0 by the periodicity of the cosine function.

We will now discuss the change of variable theorem, which enables us to find the density of the random vector (U,V) which is a "nice" transformation of (X,Y). Nice will be made precise in what follows.

Change of variable theorem: Let (X,Y) be jointly distributed continuous random variables with density function $f_X(x,y)$. Let S be an open subset of $\mathbb{R}^2$, such that $P((X,Y)\in S)=1$ (so the density f can be assumed to be concentrated on S). Let g be a transformation from S to $\mathbb{R}^2$. Thus we can write,

$$(Y_1,Y_2)\equiv g(X_1,X_2)=(g_1(X_1,X_2),g_2(X_1,X_2))\,,$$

where g_1 and g_2 are both real-valued. Now assume that,

(1) g has continuous first partial derivatives on S.

(2) g is a 1–1 function.

(3) Let $A(x_1,x_2)$ be the 2×2 matrix whose first row is

$$\left(\frac{\partial\,g_1}{\partial\,x_1}(x_1,x_2),\frac{\partial\,g_2}{\partial\,x_1}(x_1,x_2)\right)\equiv\left(\frac{\partial\,y_1}{\partial\,x_1}(x_1,x_2),\frac{\partial\,y_2}{\partial\,x_1}(x_1,x_2)\right)\,,$$

and whose second row is

$$\left(\frac{\partial\,g_1}{\partial\,x_2}(x_1,x_2),\frac{\partial\,g_2}{\partial\,x_2}(x_1,x_2)\right)\equiv\left(\frac{\partial\,y_1}{\partial\,x_2}(x_1,x_2),\frac{\partial\,y_2}{\partial\,x_2}(x_1,x_2)\right)\,.$$

Let

$$
\begin{aligned}
J_g(x_1,x_2) &= \text{abs}\;(\det A(x_1,x_2))\\
&= \left|\frac{\partial\,y_1}{\partial\,x_1}(x_1,x_2)\,\frac{\partial\,y_2}{\partial\,x_2}(x_1,x_2)-\frac{\partial\,y_2}{\partial\,x_1}(x_1,x_2)\,\frac{\partial\,y_1}{\partial\,x_2}(x_1,x_2)\right|\,,
\end{aligned}
$$

be the Jacobian of g. Then, $J_g(x_1, x_2)$ does not vanish for any $(x_1, x_2) \in S$.

Let h denote the inverse transformation of g. Thus h is defined on $g(S)$ and $h(y_1, y_2) \equiv (h_1(y_1, y_2), h_2(y_1, y_2))$ for (y_1, y_2) in $g(S)$ is the unique (x_1, x_2) in S such that $(g_1(x_1, x_2), g_2(x_1, x_2)) = (y_1, y_2)$. Then h itself has continuous first partial derivatives on $g(S)$ and is clearly 1–1. Also, if $B(y_1, y_2)$ denotes the matrix of first partial derivatives of h, then the Jacobian of h,

$$J_h(y_1, y_2) = \left| \frac{\partial x_1}{\partial y_1}(y_1, y_2) \frac{\partial x_2}{\partial y_2}(y_1, y_2) - \frac{\partial x_2}{\partial y_1}(y_1, y_2) \frac{\partial x_1}{\partial y_2}(y_1, y_2) \right| ,$$

where $x_1 = h_1(y_1, y_2)$ and $x_2 = h_2(y_1, y_2)$, does not vanish on $g(S)$ and in fact

$$J_h(y_1, y_2) = J_g(h_1(y_1, y_2), h_2(y_1, y_2))^{-1} .$$

Also, the density of the random vector (Y_1, Y_2) is given by,

$$f_Y(y_1, y_2) = f(h_1(y_1, y_2), h_2(y_1, y_2))\, J_h(y_1, y_2) , \quad (y_1, y_2) \in g(S)$$

$$f_Y(y_1, y_2) = 0 \text{ otherwise} .$$

Thus, for any nice subset I of S, we have,

$$\begin{aligned} \int_I f_X(x_1, x_2)\, dx_1\, dx_2 &= P((X_1, X_2) \in I) \\ &= P((Y_1, Y_2) \in g(I)) \\ &= \int_{g(I)} f(h_1(y_1, y_2), h_2(y_1, y_2))\, J_h(y_1, y_2) . \end{aligned}$$

We now do an application of the change of variable theorem, that will clearly illustrate what is going on. The theorem looks big and messy at first shot but really has a nice pattern, once you keep staring at it. Those of you who remember your advanced calculus well, will probably spot resemblances to the change of variable theorem in calculus (for two variables). In fact, this is precisely what the above theorem, which we will subsequently refer to as the Jacobian theorem, is, but in a different garb. The theorem extends readily to the case of more than 2 variables but we shall not discuss that extension.

Suppose that (X_1, X_2) are i.i.d. Exponential(λ) random variables. Thus,

$$f_X(x_1, x_2) = \lambda e^{-\lambda x_1}\, \lambda e^{-\lambda x_2} = \lambda^2 e^{-\lambda (x_1 + x_2)} , \ (x_1, x_2) \in S ,$$

where S is the open set $\{x_1 > 0, x_2 > 0\}$. Consider the following transformation, g, of (X_1, X_2).

$$(Y_1, Y_2) = g(X_1, X_2) = (g_1(X_1, X_2), g_2(X_1, X_2)) = (X_1 + X_2, X_1/(X_1 + X_2)) .$$

Then, $g(S)$, the open set in which the random vector (Y_1, Y_2) assumes values is,

$$g(S) = \{(y_1, y_2) : 0 < y_1, 0 < y_2 < 1\} .$$

Computing the partial derivatives of g we have,

$$\frac{\partial g_1}{\partial x_1} = 1 \quad , \quad \frac{\partial g_2}{\partial x_1} = \frac{x_2}{(x_1 + x_2)^2} ,$$

and

$$\frac{\partial g_1}{\partial x_2} = 1 \quad , \quad \frac{\partial g_2}{\partial x_2} = -\frac{x_1}{(x_1 + x_2)^2} .$$

Clearly, the partial derivatives are continuous functions of (x_1, x_2); also, g is clearly a 1–1 function on S and furthermore,

$$J_g(x_1, x_2) = \left| -\frac{x_1 + x_2}{(x_1 + x_2)^2} \right| = \frac{1}{x_1 + x_2} > 0 ,$$

for every (x_1, x_2) in S. Thus, all conditions of the Jacobian theorem are satisfied.

To obtain the density function of (Y_1, Y_2) we need to find the inverse transformation. This amounts to expressing (X_1, X_2) in terms of (Y_1, Y_2). Note that, $Y_2 (X_1 + X_2) = X_1$; but $Y_1 = X_1 + X_2$. Thus $Y_2 Y_1 = X_1$. Consequently, $X_2 = Y_1 - X_1 = Y_1 - Y_1 Y_2 = Y_1(1 - Y_2)$. Thus, we obtain the function h from $g(S)$ to S as,

$$h_1(y_1, y_2) = y_1 y_2 \quad , \quad h_2(y_1, y_2) = y_1 - y_1 y_2 .$$

The density of (Y_1, Y_2) at the point (y_1, y_2) in $g(S)$ is then computed as,

$$\begin{aligned} f_Y(y_1, y_2) &= f_X(h_1(y_1, y_2), h_2(y_1, y_2))\, J_h(y_1, y_2) \\ &= \lambda^2\, e^{-\lambda (h_1(y_1, y_2) + h_2(y_1, y_2))}\, J_g(h_1(y_1, y_2), h_2(y_1, y_2))^{-1} \\ &= \lambda^2\, e^{-\lambda (y_1 y_2 + y_1 - y_1 y_2)}\, y_1 , \end{aligned}$$

on noting that

$$J_g(x_1, x_2)^{-1} = x_1 + x_2 .$$

Thus we can rewrite the density of (Y_1, Y_2) as

$$f_Y(y_1, y_2) = (\lambda^2 \, e^{-\lambda \, y_1} \, y_1) \, 1\{y_1 > 0\} \, 1\{0 < y_2 < 1\} \, .$$

The above shows immediately that Y_1 and Y_2 are independent and that Y_1 follows $\Gamma(2, \lambda)$ while Y_2 follows $U(0, 1)$. Here I am tacitly using propositions on factorization of joint densities as a product of marginal densities as a necessary and sufficient condition for independence of random variables, a fact you must have learnt in Stat/Math 425. See, Example 7c on page 233 of Sheldon and Ross's book for a related (and more general) example. Verify that we would have gotten the same answer (as we must!!) had we computed $J_h(y_1, y_2)$ directly and plugged that in to the expression for the joint density of (Y_1, Y_2).

Here is another application of the Change of Variable Theorem and one that gives a way of generating observations from a Normal distribution. Let (X, Y) be i.i.d. $N(0, 1)$ random variables. Let R be the radius vector corresponding to the point (X, Y) and let Θ be the angle that R subtends with the positive direction of the x–axis. Thus (R, Θ) represents the vector (X, Y) in polar co-ordinates and we have the following equations:

$$X = R cos \, \theta \;\; \text{and} \;\; Y = R sin \, \theta \, .$$

(Recall the picture that I drew in class). We want to find the joint density of (R, θ). Note that (R, Θ) lives, with probability 1, in the open set $(0 \, \infty) \times (0, 2 \, \Pi)$. When we express X and Y in terms of R and Θ we are looking at the inverse transformation h; the transformation g that maps (X, Y) to (R, Θ) is a "nice" transformation in the sense that it satisfies the assumptions (1), (2) and (3) of the Change of Variable Theorem.

We first write down the joint density of $X, Y)$.

$$f_{X,Y}(x, y) = f_X(x) \, f_Y(y) = \frac{1}{\sqrt{2 \, \pi}} \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2 \, \pi}} \exp\left(-\frac{y^2}{2}\right) = \frac{1}{2 \, \pi} \exp\left(-\frac{x^2 + y^2}{2}\right) \, .$$

Now,

$$(x, y) = (h_1(r, \theta), h_2(r, \theta)) \equiv (r \, cos \, \theta, r \, sin \, \theta) \, .$$

We next compute the Jacobian of h at the point (r, θ). This is,

$$J_h(r, \theta) \;\; = \;\; \left| \frac{\partial \, x}{\partial \, r} \frac{\partial \, y}{\partial \, \theta} - \frac{\partial \, y}{\partial \, r} \frac{\partial \, x}{\partial \, \theta} \right|$$

$$
\begin{aligned}
&= |cos\,\theta\ r\,cos\,\theta - \sin\,\theta\ (-r\,sin\,\theta)| \\
&= \left|r\,cos^2(\theta) + r\,sin^2(\theta)\right| \\
&= r\,.
\end{aligned}
$$

Thus the joint density of (R, θ) is,

$$
\begin{aligned}
f_{R,\Theta}(r,\theta) &= \frac{1}{2\,\pi}\exp\left(-\frac{h_1(r,\theta)^2 + +h_2(r,\theta)^2}{2}\right) J_h(r,\theta)\,1\{r>0\}\,1\{0<\theta<2\,\pi\} \\
&= \frac{1}{2\,\pi}\exp\left(-\frac{r^2\,cos^2(\theta) + r^2\,sin^2(\theta)}{2}\right) r\,1\{r>0\}\,1\{0<\theta<2\,\pi\} \\
&= \frac{1}{2\,\pi}\,1\{0<\theta<2\,\pi\}\ r\exp\left(-r^2/2\right)\ 1\{r>0\}\,.
\end{aligned}
$$

This immediately shows that R and Θ are independent, and that Θ has the uniform distribution on $(0, 2\,\Pi)$ with marginal density,

$$
f_\Theta(\theta) = \frac{1}{2\,\pi}\,1\{0<\theta<2\,\pi\}\,.
$$

The density of R is,

$$
f_R(r) = r\exp\left(-r^2/2\right)\ 1\{r>0\}\,.
$$

Thus, if we generate R and Θ independently, with marginal distributions given as above, then $X = R\,cos\,\theta$ and $Y = R\,cos\,\theta$ are i.i.d. $N(0,1)$ random variables. To generate R and Θ we proceed as follows: Recall that if F is the distribution function of a random variable X, then $F^{-1}(U)$ has the same distribution as X, where U is a random variable distributed uniformly on $(0,1)$. Now, it is easy to show (by using the change of variable theorem in 1 dimension discussed in the previous section) that R^2 follows exponential$(1/2)$ (this is left as an exercise). If F denotes the distribution function of exp$(1/2)$, we have,

$$
F(w) = 1 - \exp\left(-w/2\right),
$$

so that

$$
F^{-1}(p) = -2\,\log(1-p)\,.
$$

Thus if U_1 and U_2 be i.i.d U(0,1) random variables, then $-2\,\log(1-U_1)$ follows exp$(1/2)$ and $2\,\pi\,U_2$ has a uniform distribution on $(0, 2\,\pi)$. Consequently, we can take,

$$
R = \sqrt{-2\,\log(1-U_1)}\ \text{ and }\ \Theta = 2\,\pi\,U_2\,.
$$

You can use this example to solve Problem 6(i) of Homework 1 quite easily.

4 Some inequalities, Law of Large Numbers, Central Limit Theorem

We first introduce some very useful probability inequalities.

Markov's inequality: Let X be a non-negative random variable and let g be a increasing non-negative function defined on $[0, \infty)$. Suppose that $E(g(X))$ is finite. Then, for any $\epsilon > 0$,

$$P(X \geq \epsilon) \leq \frac{E(g(X))}{g(\epsilon)} .$$

Proof: The proof is fairly straightforward. We will prove the inequality assuming that X is a continuous random variable with density function$f$; an analogous proof holds in the discrete case. The theorem of course holds more generally, but a completely rigorous proof is outside the scope of this course. Note that ,

$$x \geq \epsilon \Rightarrow g(x) \geq g(\epsilon) \quad (\star)$$

since g is increasing. Now,

$$\begin{aligned} E(g(X)) &= \int_0^\infty g(x)\, f(x)\, dx \\ &= \int_{[0,\epsilon)} g(x)\, f(x)\, dx + \int_{[\epsilon,\infty)} g(x)\, f(x)\, dx \\ &\geq \int_{[\epsilon,\infty)} g(x)\, f(x)\, dx \\ &\geq \int_{[\epsilon,\infty)} g(\epsilon)\, f(x)\, dx \quad \text{by } \star \\ &= g(\epsilon)\, P(X \geq \epsilon) . \end{aligned}$$

This is equivalent to the assertion of Markov's inequality.

Note that Markov's inequality gives us an upper bound on the "tail-probailities" of X and is more useful for smaller values of ϵ (and consequently smaller values of $g(\epsilon)$). Since probabilites are always bounded above by 1, the inequality will not be useful for ϵ's for which $g(\epsilon)$ is larger than $E(g(X))$. By choosing g carefully, Markov's inequality can be employed to deduce other important results, like Chebyshev's inequality, which we discuss below.

Chebyshev's inequality: Let Y be a random variable such that $E(Y^2)$ is finite. Let $\mu \equiv E(Y)$ and let $\sigma^2 = \mathrm{Var}(Y)$. Then, for any $\epsilon > 0$,

$$P\left(|Y - \mu| > \epsilon\right) < \frac{\sigma^2}{\epsilon^2}\,.$$

The proof of this is almost immediate from Markov's inequality. Let,

$$X = |Y - \mu| \quad \text{and} \quad g(x) = x^2\,,$$

and note that by the definition of variance,

$$E(g(X)) = \sigma^2\,.$$

The assumption that $E(Y^2)$ is finite enables us to talk meaningfully about the variance. There are random variables which do not have finite variance; for example, if X is Cauchy (with p.d.f. $(\pi\,(1+x^2))^{-1}$) then X has neither a mean, nor a variance. Chebyshev's inequality is important in that it gives an upper bound on the chance that a random variable Y deviates to a certain extent away from the mean. Among its numerous applications is the (weak) law of large numbers that we will discuss currently. Often, in statistics, we standardize a variable, by centering around its mean and dividing by σ, the standard deviation of the variable. The resulting variable,

$$Z \equiv \frac{Y - \mu}{\sigma}\,,$$

is free of the underlying unit of measurement and has mean 0 and variance 1. By standardizing two variables (and thereby getting rid of the original units in which they were measured), we can meaningfully define mathematical measures of association between these variables, like the correlation. The correlation between two random variables U and V is simply the average value of the product of the standardized versions of U and V; thus,

$$\rho_{U,V} = E\left(\frac{U - E\,U}{S.D.(U)}\,\frac{V - E\,V}{S.D.(V)}\right)\,.$$

Chebyshev's inequality leads to an upper bound on the chance that the standardized version of any random variable assumes a large value. If Y_{st} denotes the standardized version of Y, then from Chebyshev's inequality,

$$P\left(|Y_{st}| > \epsilon\right) = P\left(\left|\frac{Y - \mu}{\sigma}\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}\,.$$

Thus the chance that Y deviates from its mean by more than k standard deviations is less than $1/k^2$ for any random variable Y. For $k = 1$ this is non-informative, since $k^2 = 1$. For $k = 2$ this is 0.25 – in other words, the chance is less than a quarter that Y deviates by more than 2 standard deviations from its mean. If we knew that Y was normal, so that Y_{st} was $N(0, 1)$, then the chance that $\mid Y_{st} \mid > 1$ is .32 (appx), and the chance that $\mid Y_{st} \mid > 2$ is .05 (appx). Thus, the bounds provided by Chebyshev's inequality are not very sharp in this case. Thus, it does not make sense to use Chebyshev's inequality if the underlying variable is approximately normal (as is the case, in general, with sums and averages). Nevertheless, the inequality is useful, since the distribution is allowed to be arbitrary and can therefore be applied in a variety of very general situations to deduce important results.

Law of Large Numbers: To formulate the law of large numbers, we first introduce the concepts of convergence in probability and almost sure convergence. Consider a probability space $(\Omega, \mathcal{A}, P)$, where Ω, the sample space is the set of all outcomes of a random experiment, $\mathcal{A}$ is a class of subsets of Ω on which P is a probability. For $A \in \mathcal{A}$, the quantity $P(A)$ is the chance that the event A happens. A generic point in Ω is denoted by ω. Let $T_1, T_2, T_3, \ldots$ be an infinite sequence of random variables defined on Ω. Thus, each T_i is a function from $\Omega \to \mathbb{R}$ and $T_i(\omega)$ is the value of T_i at the point ω.

The sequence $\{T_n\}$ is said to *converge in probability* to the random variable T (which is also defined on Ω) if, for every $\epsilon > 0$,

$$\lim_{n \to \infty} P\left(|T_n - T| > \epsilon\right) = 0\,.$$

In particular T can be some constant c; in that case, T_n converges in probability to c if,

$$\lim_{n \to \infty} P\left(|T_n - c| > \epsilon\right) = 0\,.$$

The sequence T_n is said to *converge almost surely* to T (or c) is there is a set $\Omega_0 \subset \Omega$, with $P(\Omega_0) = 1$, such that for each $\omega \in \Omega_0$, the sequence $\{T_n(\omega)\}$ converges to $T(\omega)$ (or c).

Almost sure convergence implies convergence in probability but we will not bother about this here.

Example 1: Suppose that $\{T_n\}$ is a sequence of random variables, such that $P(T_n = 1/n) = 1/n^2$ and $P(T_n = 1) = 1 - 1/n^2$. Then, clearly T_n

converges in probability to the constant 1.

Example 2: Consider an infinite sequence of flips of a fair coin and let X_n denote what you get (1 if Heads and 0 if Tails) on the n'th flip. Then $X_1, X_2, \ldots$ are i.i.d. Bernoulli(1/2). Define a sequence T_n in the following way:

$$T_n = \frac{X_1 + X_2 + \ldots + X_n}{n}\,.$$

Thus $T_n \equiv \overline{X}_n$ is the proportion of Heads in the first n tosses of the coin. It can be shown that X_n converges in probability to 1/2 (we will prove this in a more general context). We write

$$\overline{X}_n \rightarrow_p \frac{1}{2}\,,$$

to denote this phenomenon. In fact, a stronger result holds. The sequence $\{\overline{X}_n\}$ actually converges almost surely to 1/2.

We now formally state the laws of large numbers.

WLLN – Weak Law of Large Numbers: Let $X_1, X_2, \ldots$ be an infinte sequence of i.i.d. random variables, such that $E(\mid X_1 \mid) < \infty$. Let $\mu \equiv E\,X_1$. Then,

$$E\,\left(\mid \overline{X}_n - \mu \mid\right) \rightarrow 0\,.$$

It follows from an easy application of Markov's inequality that $\overline{X}_n$ converges in probability to μ. We will not prove the WLLN as stated above. Instead, we will deduce a weaker conclusion, that is also referred to as WLLN. We state this formally.

WLLN (weaker version): Let $X_1, X_2, \ldots$ be an infinite sequence of i.i.d. random variables, such that $E(\mid X_1 \mid) < \infty$. Let $\mu = E(X_1)$ and let σ^2, the variance of X_1, be finite. Then $\overline{X}_n$ converges in probability to μ.

Proof: We will use Chebyshev's inequality. Choose and fix any $\epsilon > 0$. Note that,

$$P\,\left(\left|\overline{X}_n - \mu\right| > \epsilon\right) \leq \frac{E\left((\overline{X}_n - \mu)^2\right)}{\epsilon^2}\,.$$

But note that $E\,(\overline{X}_n)$ is μ, so that,

$$E\left((\overline{X}_n - \mu)^2\right) = \mathrm{Var}(\overline{X}_n)\,.$$

But,

$$\mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X_1 + X_2 + \ldots + X_n)}{n^2} = \frac{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n)}{n^2} = n\,\frac{\mathrm{Var}(X_1)}{n^2}\,.$$

Thus,

$$E\left((\overline{X}_n - \mu)^2\right) = \frac{\sigma^2}{n}\,,$$

and we get,

$$P\left(\left|\overline{X}_n - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{n\,\epsilon^2} \to 0\,,$$

which finishes the proof.

Strong Law of Large Numbers (SLLN): Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. random variables such that $E(\mid X_1 \mid) < \infty$ and let $\mu \equiv E(X_1)$. Then,

$$E\,\overline{X}_n \to_{a.s} \mu\,.$$

(In words, $\overline{X}_n$ converges almost surely to μ.) The strong law of large numbers, as its name suggests, is a strong results – it says that, outside of a set of probability 0 (and therefore a negligible one), the sample average, in the long run, is going to hit the population average. To see the strong law "in action" we perform a computer simulation experiment:

We "tossed a fair coin 100 times" independently (on a computer) and recorded the sequence of outcomes $(X_1, X_2, \ldots, X_{100})$. Thus each X_i takes the value 1 or 0 with probability 0.5. How does one simulate a coin tossing experiment on a computer? Statistical software packages or standard C or Fortran compilers have a uniform random number generator, which can be used to generate observations from the uniform distribution. So, to simulate 100 independent coin flips, you generate 100 observations using the random number generator; call these $U_1, U_2, \ldots, U_{100}$. Then, the U_i's are i.i.d. uniform random variables and define X_i to be 1 if $U_i < 0.5$ (this happens with probability 1/2) and 0 otherwise. Then X_i has a Bernoulli(1/2) distribution. And of course, the X_i's are independent.

Having obtained the X_i's, we computed the proportion of heads at each stage i from 1 to 100. Thus, for each i, we obtained

$$\overline{X}_i = \frac{X_1 + X_2 + \ldots + X_i}{i}\,,$$

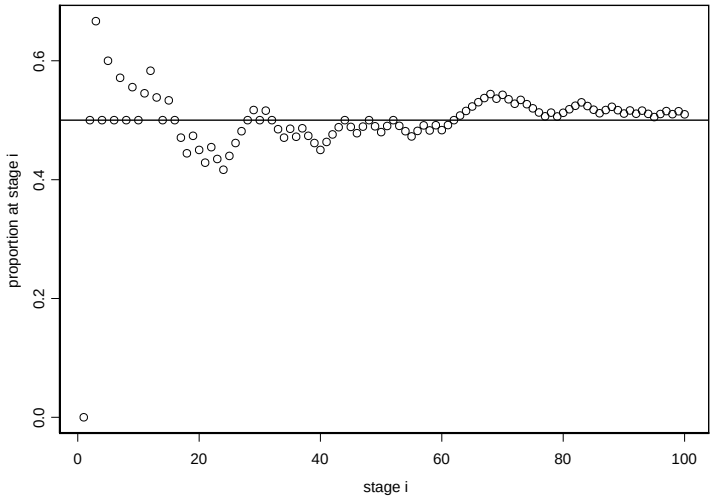

Figure 2: The behavior of the sample proportion of heads in 100 tosses of a fair coin.

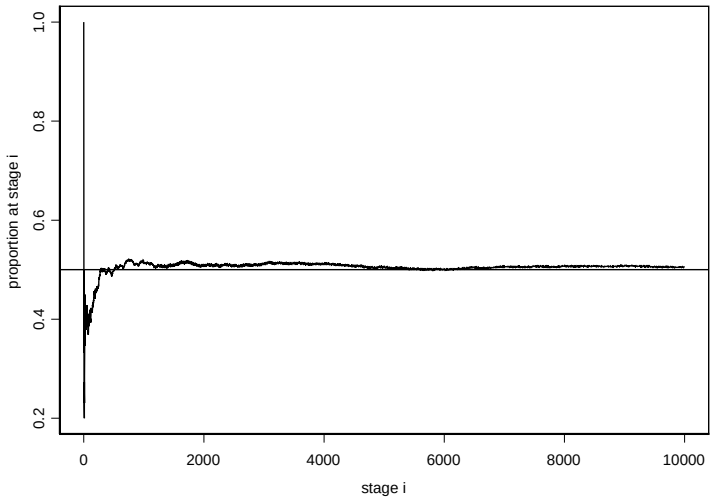

Figure 3: The behavior of the sample proportion of heads in 10000 tosses of a fair coin.

and then plotted $\{i, \overline{X}_i\}_{i=1}^{100}$. The plot is shown in 2. The horizontal line sketched in the figure is the line $y = 0.5$. It is clearly seen that the sample proportion of tosses approaches the population proportion (0.5) with an increasing number of tosses, in accordance with the strong law of large numbers. Of course, the higher the number of tosses, the closer is the sample proportion going to be to the population proportion. Figure 3 shows the behavior of the sample proportion of heads with increasing sample size for 10000 tosses. Once again, it is clear that the sample proportion in the long run is very close to 0.5. With 100 tosses, the largest absolute deviation between the sample proportion and the population proportion (0.5) for the last 30 stages (the last 30%) is around .04; with 10000 tosses the largest absolute deviation between the sample proportion and the population proportion for the last 3000 stages (the last 30%) is around .008. Increasing n leads to higher precision.

We will discuss some applications of the strong law shortly. But before that, we discuss a motivation for the Central Limit Theorem (CLT). In the above example, we have seen that $\hat{p}_n$, the sample proportion at stage n, converges to $p = 0.5$, the chance of heads on a single toss (as the SLLN tells us). However, the SLLN does not tell us anything about the behavior of the (random) fluctuations $\hat{p}_n - p$, with increasing n (sample size). However, an idea of this behavior is often needed to gauge the accuracy of $\hat{p}_n$ as an estimate of p. This is where the Central Limit Theorem (CLT) comes in. Very broadly, what the CLT tells us is that, under appropriate conditions, sums or averages of a large number of random variables is approximately normally distributed. We have referred to *the* CLT above, but this is not strictly correct. There are many CLT's formulated under various different situations, but the unifying theme in all these theorems is the appearance of a normal distribution as a limit. We will deal in this course, with the most ubiquitous (and the most famous) CLT of all – the Lindeberg Levy Central Limit Theorem.

We quickly define the concept of *convergence in distribution* or *weak convergence* in what follows. Let $X_1, X_2, \ldots$ be a sequence of random variables (note that we do not require these to be independent or identically distributed; indeed, these need not be even defined on the same probability space). Let F_n be the distribution function of X_n. Let X be a random variable with distribution function F. Then, X_n is said to converge in dis-

tribution to X if for all x, such that x is a continuity point of F, we have,

$$\lim_{n\to\infty} F_n(x) = F(x)\,.$$

Note that, though a distribution function F is right continuous, it is not necessarily continuous at all points. Distributional convergence requires the above limit to hold only at those points x where F is continuous. If X assumes a certain value, say x_0 with positive probability, then F is not continuous at x_0 and the above convergence is not required to hold. You can easily convince yourself that the points at which the distribution function F fails to be continuous are precisely those values that X assumes with positive probability. For example, if X is Bernoulli(1/2), X assumes the values 0 and 1, each with probability 1/2 and the distribution function $F(x)$ is 0 for $x < 0$, $F(x)$ is 0.5 for $0 \le x < 1$ and $F(x)$ is 1 for $1 \le x < \infty$ and the points of (jump) discontinuity of F are just 0 and 1 where F jumps by 0.5. We next formally state the Lindeberg-Levy CLT.

Central Limit Theorem (de Moivre 1706, Laplace 1812, Lindeberg 1922): Let $X_1, X_2, \ldots$ be i.i.d. random variables with ANY common distribution, with finite mean μ and variance $\sigma^2 > 0$. Let $S_n = X_1 + X_2 + \ldots + X_n$. Let $S_n^\star$ denote the standardized version of S_n. Thus,

$$S_n^\star = \frac{S_n - E(S_n)}{\sqrt{\mathrm{Var}(S_n)}} = \frac{S_n - n\,\mu}{\sigma\,\sqrt{n}}\,.$$

Then $S_n^\star$ converges in distribution to $N(0,1)$. In other words, for every x, we have,

$$\lim_{n\to\infty} P\ (S_n^\star \le x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\,\pi}}\, e^{-t^2/2}\, dt\,.$$

Thus the CLT tells us that in the long run, the standardized sum $S_n^\star$ behaves approximately like a N(0,1) random variable. It follows, that the actual sum S_n behaves like a $N(n\,\mu, \sigma\,\sqrt{n})$ random variable, since,

$$P(S_n \le y) = P(S_n^\star \le (y - n\,\mu)/\sigma\,\sqrt{n}) \to \Phi((y - n\,\mu)/\sigma\,\sqrt{n}) = \mathrm{Prob}(N(n\,\mu, \sigma\,\sqrt{n}) \le y)\,.$$

Here are some other comments/observations about the CLT.

- A. Note that the sample average $\overline{X}_n$ is also approximately normal, with parameters $\mu, \sigma/\sqrt{n}$, by the CLT. Since,

$$S_n^\star = \frac{S_n - n\,\mu}{\sigma\,\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \overline{X}_n^\star\,,$$

where $\overline{X}_n^\star$ is the standardized version of $\overline{X}_n$. Indeed, the CLT could be stated equivalently, in terms of the sample average, rather than the sum.

- B. Since Φ is a continuous distribution function, the convergence of the distribution function of $S_n^\star$ at the point x to $\Phi(x)$ needs to happen at every point x.

What does the CLT look like in pictures? Once again, we illustrate with our favorite independent sequence of coin flips example. Suppose, we toss a coin n times, where n is fairly large. If $X_1, X_2, \ldots, X_n$ denote the outcomes, then check that,

$$\overline{X}_n^\star = \frac{\hat{p}_n - 1/2}{1/2}.$$

By the CLT, this should have an approximately standard normal distribution. To check this, we draw a large number of observations (say N) from the distribution of $\overline{X}_n^\star$. Each observation is generated by flipping a fair coin n times (independently) and computing $\overline{X}_n^\star$ from the sequence of n flips. These N values from the distribution of $\overline{X}_n^\star$ are then plotted as a histogram, drawn to a probability scale. A histogram is a graphical device that pictorially represents the distribution of values of a variable. We first partition the range of the observed values (in this case, of $\overline{X}_n^\star$) into a number of mutually disjoint (and contiguous) intervals (generally these are all taken to have equal length) . We then count the number of observations that fall in each interval and compute the relative frequency (in this case, dividing by N). We then erect a rectangle on each interval, such that the area of the rectangle above that interval gives the relative frequency of that interval (so the height of each rectangle is the relative frequency in that interval per unit length). Thus, the histogram looks like a series of rectangles of varying heights and the total area under the histogram is 1. As we increase the number of observations from the underlying distribution, and increase the number of intervals to use for the histogram at an appropriate rate, we get a better and better idea of the distribution in the population. In fact, at each stage the histogram can be thought of as a piecewise constant approximation to the density of the underlying random variable. In Figure 4 we show the histogram for $\overline{X}_{200}^\star$ based on 10000 replicates from the distribution and in Figure 5 we show the histogram for $\overline{X}_{1000}^\star$ based on $N = 10000$ replicates. In each of these plots the $N(0,1)$ density is superimposed. It is seen that the shape of the histogram matches up well with the superimposed density, as it should, by the CLT. With $n = 1000$, the fit is seen to be (expectedly) better than with $n = 200$.

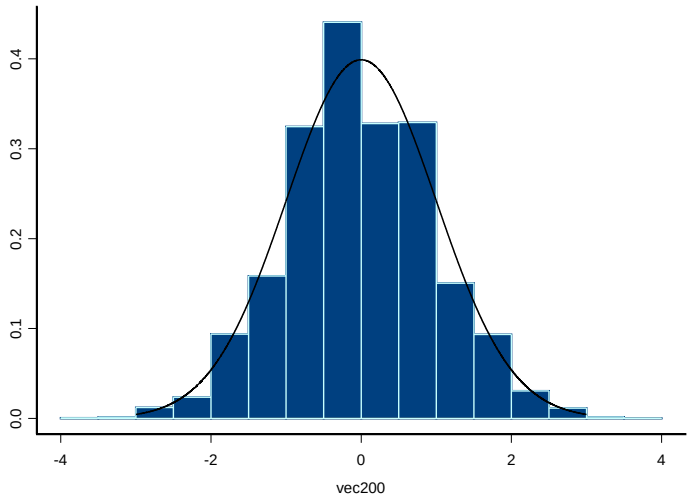

Figure 4: The histogram of the standardized proportion for 200 tosses of a fair coin and the superimposed limit.

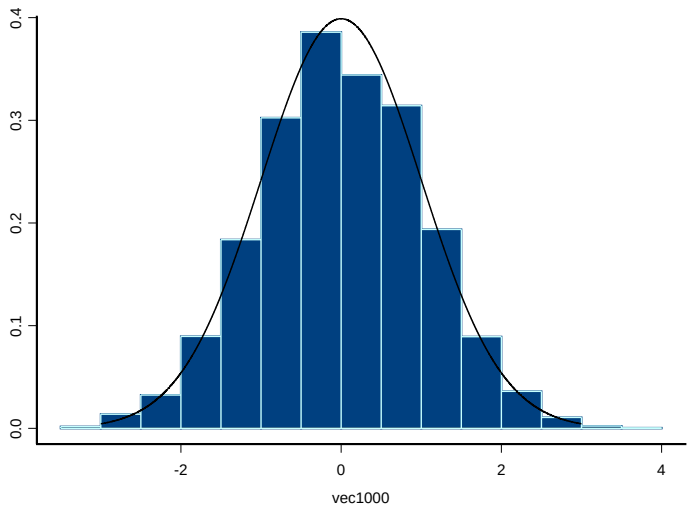

Figure 5: The histogram of the standardized proportion for 1000 tosses of a fair coin and the superimposed limit.

We now discuss a few applications of the SLLN and the CLT. The SLLN is used widely to do, what is called "Monte-Carlo" integration. Suppose we wish to evaluate, the integral,

$$\int_0^1 f(x)\,dx\,;$$

if the function f has a very complicated form, then it might not be possible to obtain an analytic expression for the integral. However, good approximations to the value of the integral can be obtained by using the strong law. Note, that,

$$\int_0^1 f(x)\,dx = E(f(U))\,,$$

where U is a uniform random variable on $(0,1)$. Now, suppose we have an i.i.d. sequence of uniform random variables, $U_1, U_2, \ldots$. Then, $f(U_1), f(U_2), \ldots$ are i.i.d. and each has the same distribution as $f(U)$. By the SLLN, we know that with probability 1,

$$\lim_{n\to\infty} \frac{f(U_1) + f(U_2) + \ldots + f(U_n)}{n} = E(f(U))\,.$$

Thus to compute the integral, we generate $U_1, U_2, \ldots, U_n$ from Uniform (0,1), independently, and for a very large n and use, $(f(U_1) + f(U_2) + \ldots + f(U_n))/n$ to estimate the integral. Note that this is an extremely powerful trick; we did not resort to any complicated numerical integration procedure or a tricky maths exercise. We simply worked out an average of a function of uniforms. Since uniforms can be generated on any decent computer, this gives us an efficient way of computing.

In general, how about computing an integral g over a finite interval of the form (a,b)? In this case, we want to evaluate,

$$\int_a^b g(x)\,dx = (b-a)\int_a^b g(x)\,\frac{1}{b-a}\,dx = (b-a)\,E(g(X))\,,$$

where X has the uniform distribution on (a,b). Thus in this case, all that we need to do is generate $X_1, X_2, \ldots, X_n$ from Uniform(a,b), for a large n and estimate $E(g(X))$ as $n^{-1}\,(g(X_1) + g(X_2) + \ldots + g(X_n))$. How do you generate X from Uniform(a,b)? Check that if U is Uniform(0,1), then $a + (b-a) * U$ is Uniform(a,b).

The CLT is of paramount importance in statistics. We shall encounter

it a lot in this course, in various different contexts. Right now, let's see how it provides good approximations, to some very standard distributions in practice. In Homework 2, you have encountered/will encounter the Poisson process and Gamma distributions. We show here that both Poisson and Gamma distributions can be approximated by normal distributions through the CLT. Let $X_1, X_2, \ldots..$ be i.i.d. Exponential(λ) random variables. The CLT then tells us that for large enough n,

$$\frac{S_n - n\,\lambda^{-1}}{\sqrt{n}\,\lambda^{-1}} \sim_{appx} N(0,1)\,;$$

here $S_n = X_1 + X_2 + \ldots + X_n$ and $E(S_n) = n\,\lambda^{-1}$ and $\mathrm{Var}(S_n) = n\,\lambda^{-2}$. But S_n is distributed as $\Gamma(n,\lambda)$, as you have shown/will show in Homework 2. Thus, it follows, that for large n, S_n is approximately, $N(n\,\lambda^{-1}, n\,\lambda^{-2})$. In other words, for large n, the Gamma distribution can be approximated by the $N(n\,\lambda^{-1}, n\,\lambda^{-2})$ distribution.

A similar approximation works for the Poisson random variable. Let $X_1, X_2, \ldots,$ be i.i.d. Poisson(λ) random variables. Let $S_n = X_1 + X_2 + \ldots + X_n$. Then, from Homework 1, we know that S_n is itself, a Poisson random variable, with mean and variance both equal to $n\,\lambda$. By CLT,

$$\frac{S_n - n\,\lambda}{\sqrt{n\,\lambda}} \sim_{appx} N(0,1)\,,$$

so that S_n is approximately distributed as $N(n\,\lambda, n\,\lambda)$. In other words, the Poisson($n\,\lambda$) distribution, for large n can be approximated by the $N(n\,\lambda, n\,\lambda)$ distribution. Figure 6 provides a histogram of 10000 observations drawn from the Poisson $n = 100$ distribution and the corresponding normal approximation.

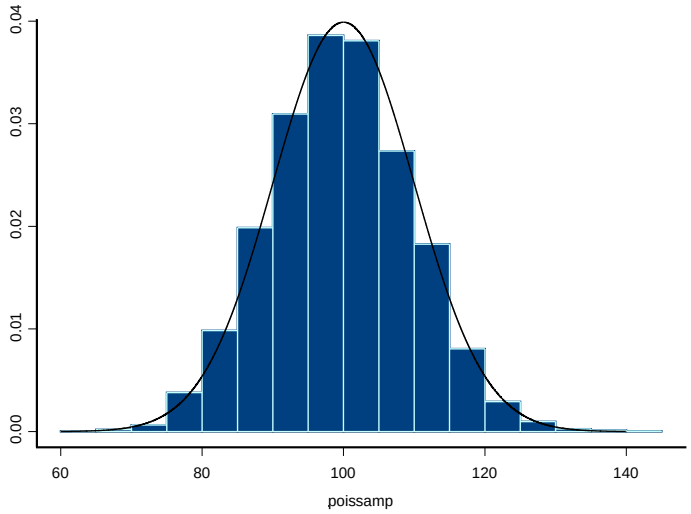

Figure 6: Histogram from the Poisson($n = 100$) distribution and the superimposed Normal(100,100) density.