5. INEQUALITIES, LIMIT THEOREMS AND GEOMETRIC PROBABILITY

5.1 Inequalities

Suppose that $X \geqslant 0$ is a random variable taking non-negative values and that $c > 0$ is a constant. Then

$$\mathbb{P}(X \geqslant c) \leqslant \frac{\mathbb{E}(X)}{c},$$

is **Markov's inequality**. It follows because

$$\mathbb{P}(X \geqslant c) = \mathbb{E}\left(I_{(X \geqslant c)}\right) \leqslant \mathbb{E}\left(\frac{X}{c} I_{(X \geqslant c)}\right) \leqslant \mathbb{E}\left(\frac{X}{c}\right) = \frac{\mathbb{E}(X)}{c}.$$

From this result we may deduce **Chebyshev's inequality**: for any random variable X and constant $c > 0$,

$$\mathbb{P}(|X| \geqslant c) \leqslant \frac{\mathbb{E}\left(X^2\right)}{c^2}.$$

This follows by observing that

$$\mathbb{P}(|X| \geqslant c) = \mathbb{P}\left(X^2 \geqslant c^2\right),$$

and applying Markov's inequality for the random variable $Y = X^2$ and constant c^2. We should note the following points about Chebyshev's inequality:

1. The inequality is 'distribution free'; it holds for all random variables irrespective of the distribution of the random variable.
2. If $\mathbb{E}(X^2) \geqslant c^2$, then the inequality provides no bound on the probability.
3. If $\mathbb{E}\left(X^2\right) < c^2$, the inequality is the best possible in the sense that given c there is a random variable X for which the inequality holds with equality. To see this suppose that $d < c^2$ and let X be the random variable

$$X = \begin{cases} c & \text{with probability } \dfrac{d}{2c^2}, \\ -c & \text{with probability } \dfrac{d}{2c^2}, \\ 0 & \text{with probability } 1 - \dfrac{d}{c^2}. \end{cases}$$

Then $\mathbb{E}\left(X^2\right) = d$ and $\mathbb{P}(|X| \geqslant c) = d/c^2 = \mathbb{E}\left(X^2\right)/c^2$.

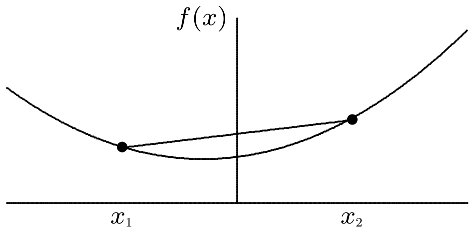
Suppose that $\phi : [0, \infty) \to [0, \infty)$is a non-decreasing function, with $\phi(x) > 0$ for $x > 0$, then we obtain the **generalized Chebyshev's inequality**: for any random variable X and $c > 0$,

$$\mathbb{P}(|X| \geqslant c) \leqslant \frac{\mathbb{E}\left(\phi(|X|)\right)}{\phi(c)}.$$

This follows in the same way by observing that $\mathbb{P}(|X| \geqslant c) \leqslant \mathbb{P}(\phi(|X|) \geqslant \phi(c))$, and using Markov's inequality with $Y = \phi(|X|)$. As an example, take $\phi(x) = x^4$ and we obtain $\mathbb{P}(|X| \geqslant c) \leqslant \mathbb{E}\left(X^4\right)/c^4$.

The next inequality involving random variables that we will consider is based on the concept of convexity. A function $f : \mathbb{R} \to \mathbb{R}$ is **convex** if for all $x_1, x_2 \in \mathbb{R}$ and $\lambda_1 \geqslant 0$, $\lambda_2 \geqslant 0$ with $\lambda_1 + \lambda_2 = 1$, we have

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leqslant \lambda_1 f(x_1) + \lambda_2 f(x_2).$$



Thus a function is convex if the chord joining any two points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ on the graph of the function lies above the function between the points. It is easy to see that if $f(x)$ is a convex function then for $x_1 < x_2 < x_3$, the slope of the chord joining the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ is less than or equal to the slope of the chord joining the points points $(x_2, f(x_2))$ and $(x_3, f(x_3))$; that is

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leqslant \frac{f(x_3) - f(x_2)}{x_3 - x_2}; \tag{5.1}$$

this follows from the definition of convexity, since

$$x_2 = \left(\frac{x_3 - x_2}{x_3 - x_1}\right) x_1 + \left(\frac{x_2 - x_1}{x_3 - x_1}\right) x_3,$$

so that

$$f(x_2) \leqslant \left(\frac{x_3 - x_2}{x_3 - x_1}\right) f(x_1) + \left(\frac{x_2 - x_1}{x_3 - x_1}\right) f(x_3), \tag{5.2}$$

and rearranging (5.2) gives (5.1). Furthermore, from (5.1), it is immediate that for points $x_1 < x_2 < x_3 < x_4$, we have

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leqslant \frac{f(x_4) - f(x_3)}{x_4 - x_3}, \tag{5.3}$$

since we may apply (5.1) twice to obtain

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leqslant \frac{f(x_3) - f(x_2)}{x_3 - x_2} \leqslant \frac{f(x_4) - f(x_3)}{x_4 - x_3}\,. \tag{5.4}$$

It is the case that a function f is convex if and only if (5.1) holds for all choices of $x_1 < x_2 < x_3$.

Moreover, when f is differentiable then f being convex is equivalent to the derivative $f'(x)$ being non-decreasing (or $f''(x) \geqslant 0$); this may be seen by letting $x_2 \to x_1$ and $x_4 \to x_3$ in (5.3). With a similar argument, we may see that, when f is convex, then

$$f(y) - f(x) \geqslant (y - x) f'(x), \quad \text{for all} \quad x, y. \tag{5.5}$$

Now if f is a convex function, for each $n \geqslant 2$, and for any points $x_1, \ldots, x_n \in \mathbb{R}$ and any $\lambda_i \geqslant 0$, $1 \leqslant i \leqslant n$, with $\lambda_1 + \cdots + \lambda_n = 1$, we have

$$f\left(\lambda_1 x_1 + \cdots + \lambda_n x_n\right) \leqslant \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n). \tag{5.6}$$

The proof of the inequality (5.6) is by induction on n. The case $n = 2$ is just the definition of convexity. So assume that (5.6) holds for any n points $x_1, \ldots, x_n \in \mathbb{R}$ and any $\lambda_i \geqslant 0$, $1 \leqslant i \leqslant n$, with $\lambda_1 + \cdots + \lambda_n = 1$, and suppose that we are given $x_1, \ldots, x_{n+1} \in \mathbb{R}$ and $\lambda_i \geqslant 0$, $1 \leqslant i \leqslant n+1$, with $\lambda_1 + \cdots + \lambda_{n+1} = 1$. We may assume that $\lambda_i > 0$ each i, otherwise the result follows by the inductive step immediately. Then, by first using the case $n = 2$ and then the inductive hypothesis, we have

$$\begin{aligned}
f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left((1 - \lambda_{n+1}) \sum_{i=1}^{n} \left(\frac{\lambda_i}{1 - \lambda_{n+1}}\right) x_i + \lambda_{n+1} x_{n+1}\right) \\
&\leqslant (1 - \lambda_{n+1}) f\left(\sum_{i=1}^{n} \left(\frac{\lambda_i}{1 - \lambda_{n+1}}\right) x_i\right) + \lambda_{n+1} f(x_{n+1}), \\
&\leqslant (1 - \lambda_{n+1}) \left(\sum_{i=1}^{n} \left(\frac{\lambda_i}{1 - \lambda_{n+1}}\right) f(x_i)\right) + \lambda_{n+1} f(x_{n+1}) = \sum_{i=1}^{n+1} \lambda_i f(x_i),
\end{aligned}$$

completing the induction.

Jensen's Inequality For a random variable X and a convex function f,

$$f(\mathbb{E}\,X) \leqslant \mathbb{E}\,f(X).$$

The proof of Jensen's Inequality in the case when X takes on just finitely many values $x_1, \ldots, x_n$, with probabilities $p_i = \mathbb{P}(X = x_i)$, $1 \leqslant i \leqslant n$, with $p_i \geqslant 0$ and $\sum_{i=1}^n p_i = 1$, is just a restatement of (5.6), since

$$f(\mathbb{E}\,X) = f\left(\sum_{i=1}^n p_i x_i\right) \leqslant \sum_{i=1}^n p_i f(x_i) = \mathbb{E}\,f(X).$$

In general, for any random variable, we may use (5.5), to see that

$$f(X) - f(\mathbb{E}\,X) \geqslant (X - \mathbb{E}\,X) f'(\mathbb{E}\,X),$$

and taking the expectation of both sides we see that

$$\mathbb{E}\,f(X) - f(\mathbb{E}\,X) \geqslant \mathbb{E}\,\left(X - \mathbb{E}\,X\right) f'(\mathbb{E}\,X) = 0,$$

which gives the result.

Example 5.7 *Arithmetic-Geometric Mean Inequality* For positive real numbers $x_1, \ldots x_n$,

$$\left(\prod_{i=1}^n x_i\right)^{1/n} \leqslant \frac{1}{n}\sum_{i=1}^n x_i.$$

This follows by Jensen's inequality applied to the convex function $f(x) = -\log x$, and the random variable X which takes the value x_i with probability $\frac{1}{n}$, so that

$$-\frac{1}{n}\sum_{i=1}^n \log x_i = \mathbb{E}\,\left(-\log(X)\right) \geqslant -\log(\mathbb{E}\,X) = -\log\left(\frac{1}{n}\sum_{i=1}^n x_i\right)$$

from which we see that

$$\log\left(\left(\prod_{i=1}^n x_i\right)^{1/n}\right) \leqslant \log\left(\frac{1}{n}\sum_{i=1}^n x_i\right),$$

which gives the result, since log is an increasing function. $\square$

5.2 Weak Law of Large Numbers

Theorem 5.8 Weak Law of Large Numbers *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with $\mathbb{E}\, X_1 = \mu$ and $\mathbb{V}\mathrm{ar}\,(X_1) < \infty$. For any constant $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) \to 0, \quad \textit{as} \quad n \to \infty.$$

Proof. By Chebyshev's inequality we have

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) &\leqslant \frac{1}{\epsilon^2}\,\mathbb{E}\,\left(\frac{X_1 + \cdots + X_n}{n} - \mu\right)^2 \\
&= \frac{1}{n^2\epsilon^2}\,\mathbb{E}\,(X_1 + \cdots + X_n - n\mu)^2 \\
&= \frac{1}{n^2\epsilon^2}\,\mathbb{V}\mathrm{ar}\,(X_1 + \cdots + X_n) = \frac{n}{n^2\epsilon^2}\mathbb{V}\mathrm{ar}\,(X_1) \\
&= \frac{1}{n\epsilon^2}\,\mathbb{V}\mathrm{ar}\,(X_1) \to 0,
\end{aligned}$$

as required. □

Notes 1. The statement in Theorem 5.8 is normally referred to by saying that the random variable $(X_1 + \cdots + X_n)/n$ 'converges in probability' to μ, written

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{P} \mu, \quad \text{as} \quad n \to \infty.$$

2. This result should be distinguished from the Strong Law of Large Numbers which states that

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n}{n} \to \mu, \quad \text{as} \quad n \to \infty\right) = 1.$$

As the name implies, the Strong Law of Large Numbers implies the Weak Law. The mode of convergence in the Strong Law is referred to as 'convergence with probability one' or 'almost sure convergence'.

3. Notice that the requirement that the random variables in the Weak Law be independent is not one that we may dispense with. For example, suppose that $\Omega = \{\omega_1, \omega_2\}$ has just two points and let $X_n(\omega_1) = 1$ and $X_n(\omega_2) = 0$ for each n, so that the random

variables are identically distributed, but not of course independent. Let $p = \mathbb{P}(\{\omega_1\}) = 1 - \mathbb{P}(\{\omega_2\})$, where $0 < p < 1$, then $\mathbb{E}\, X_1 = p$, and we have

$$\frac{X_1(\omega_1) + \cdots + X_n(\omega_1)}{n} = 1 \quad \text{and} \quad \frac{X_1(\omega_2) + \cdots + X_n(\omega_2)}{n} = 0, \quad \text{for all } n,$$

so that the conclusion of Theorem 5.8 cannot hold.

4. By giving a more refined argument it is possible to dispense with the requirement in the statement of the Theorem that $\mathbb{V}\mathrm{ar}\,(X_1) < \infty$. The conclusion still holds provided $\mathbb{E}\,|X_1| < \infty$.

5. It should be noticed that the Weak Law of Large Numbers is 'distribution free' in that the particular distribution of the summands $\{X_i\}$ only influences the result through the mean, μ, (and, in the form we have stated it, through the fact that the variance is finite) but otherwise the underlying distribution does not enter the conclusion of the Theorem.

6. The Weak Law of Large Numbers underlies the 'frequentist' interpretation of probability. Suppose that we have independent repetitions of an experiment, and we set $X_i = 1$ if a particular outcome occurs on the ith repetition (e..g., 'Heads'), and $X_i = 0$, otherwise (e.g., 'Tails'). Then $\mathbb{E}\, X_i = p$, say, where $p = \mathbb{P}(X_i = 1)$ is the probability of the outcome. Then $(X_1 + \cdots + X_n)/n$ is the average number of occurrences of the outcome in n repetitions and this converges in the above sense to $\mathbb{E}\, X_i = p$; thus the probability p is the long-run proportion of times that the outcome occurs.

5.3 Central Limit Theorem

Theorem 5.9 Central Limit Theorem *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with $\mathbb{E}\, X_1 = \mu$ and $\sigma^2 = \mathbb{V}\mathrm{ar}\,(X_1)$, where $0 < \sigma^2 < \infty$. For any x, $-\infty < x < \infty$,*

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leqslant x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy = \Phi(x),$$

which is the distribution function of the standard $N(0,1)$ distribution.

Sketch of Proof: We will illustrate why the result of the Theorem holds by using moment generating functions in the case when the moment generating function of the $\{X_i\}$, $m(\theta) =$

$\mathbb{E}\left(e^{\theta X_1}\right)$, satisfies $m(\theta) < \infty$ for a non-trivial range of values of θ (that is, an open interval which necessarily contains the point $\theta = 0$). It should be noted that this condition on the moment generating function is not necessary for the result to hold. Let

$$Y_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}, \quad \text{and} \quad m_n(\theta) = \mathbb{E}\left(e^{\theta Y_n}\right),$$

then we will show that as $n \to \infty$, $m_n(\theta) \to e^{\theta^2/2}$, which is the moment generating function of the $N(0,1)$ distribution. This is sufficient to establish the conclusion of the Theorem, although we will not prove that in this course. We now have

$$m_n(\theta) = \mathbb{E}\left(e^{\theta(X_1+\cdots+X_n-n\mu)/(\sigma\sqrt{n})}\right) = e^{-\theta\mu\sqrt{n}/\sigma}\mathbb{E}\left(e^{\theta(X_1+\cdots+X_n)/(\sigma\sqrt{n})}\right)$$

and since the $\{X_i\}$ are independent and identically distributed, this

$$= e^{-\theta\mu\sqrt{n}/\sigma}\left[\mathbb{E}\left(e^{\theta X_1/(\sigma\sqrt{n})}\right)\right]^n = \left[e^{-\theta\mu/(\sigma\sqrt{n})} m\left(\frac{\theta}{\sigma\sqrt{n}}\right)\right]^n.$$

Expand the two terms using Taylor's Theorem to see that $m_n(\theta)$ equals

$$\left[\left(1 - \frac{\theta\mu}{\sigma\sqrt{n}} + \frac{\theta^2\mu^2}{2\sigma^2 n} + O\left(\frac{1}{n^{3/2}}\right)\right)\left(1 + \frac{\theta}{\sigma\sqrt{n}}\, m'(0) + \frac{\theta^2}{2\sigma^2 n}\, m''(0) + O\left(\frac{1}{n^{3/2}}\right)\right)\right]^n;$$

now, using the fact that $m'(0) = \mathbb{E}\, X_1 = \mu$, and $m''(0) = \mathbb{E}\left(X_1^2\right) = \sigma^2 + \mu^2$, this shows that

$$\begin{aligned} m_n(\theta) &= \left[1 - \frac{\theta^2\mu^2}{\sigma^2 n} + \frac{\theta^2(\sigma^2+\mu^2)}{2\sigma^2 n} + \frac{\theta^2\mu^2}{2\sigma^2 n} + O\left(\frac{1}{n^{3/2}}\right)\right]^n \\ &= \left[1 + \frac{\theta^2}{2n} + O\left(\frac{1}{n^{3/2}}\right)\right]^n \to e^{\theta^2/2}, \quad \text{as required.} \end{aligned}$$

□

Notes 1. The mode of convergence described in Theorem 5.9 for the random variables

$$Y_n = (X_1 + \cdots + X_n - n\mu)/(\sigma\sqrt{n})$$

is known as 'convergence in distribution' and the conclusion is written as $Y_n \xrightarrow{\mathcal{D}} Z$, where $Z \sim N(0,1)$.

2. Note that, like the Weak Law of Large Numbers, the Central Limit Theorem is distribution free, in that the underlying distribution of the $\{X_i\}$ influences the form of the result only through the mean $\mu = \mathbb{E}\, X_1$ and variance $\sigma^2 = \mathbb{V}\text{ar}\,(X_1)$.

3. Note that in the Central Limit Theorem, by subtracting off at each stage the mean of the sum of the random variables $X_1 + \cdots + X_n$, that is $n\mu$, and dividing by its standard deviation, $\sigma\sqrt{n}$, we are ensuring that the random variable Y_n has $\mathbb{E}\, Y_n = 0$ and $\mathbb{V}\mathrm{ar}\,(Y_n) = 1$, for each n.

Example 5.10 *Normal approximation to the binomial distribution* If the random variable $Y \sim \mathrm{Bin}(n, p)$, we may think of the distribution of Y as being the same as that of the sum of n i.i.d. random variables each of which has the Bernoulli distribution; thus the random variable $(Y - np)/\sqrt{np(1-p)}$ has approximately the $N(0,1)$ distribution for large n. Note that here p is being held fixed and $n \to \infty$, unlike the situation we described in the Poisson approximation to the binomial where $n \to \infty$ and $p \to 0$ in such a way that the product $np \to \lambda > 0$. □

Example 5.11 *Normal approximation to the Poisson distribution* When the random variable $Y \sim \mathrm{Poiss}(n)$, where $n \geqslant 1$ is an integer, we may think of Y as having the same distribution as that of the sum of n i.i.d. random variables each with the Poiss(1) distribution. Thus $(Y - n)/\sqrt{n}$ has approximately the $N(0,1)$ distribution for large n. The same conclusion is true for $Y \sim \mathrm{Poiss}(\lambda)$ for non-integer λ; that is, $(Y - \lambda)/\sqrt{\lambda}$ is approximately $N(0,1)$ for λ large. □

Example 5.12 *Opinion polls* Suppose that the proportion of voters in the population who vote Labour is p, where p is unknown. A random sample of n voters is taken and it is found that S voters in the sample vote Labour and we estimate p by S/n. We want to ensure that $|S/n - p| < \epsilon$, for some small given ϵ, with high probability, $\geqslant 0{\cdot}95$, say. How large must n be? Note that $S \sim \mathrm{Bin}(n, p)$, so that $\mathbb{E}\, S = np$ and $\mathbb{V}\mathrm{ar}\,(S) = np(1-p)$. Then by the Central Limit Theorem, we require

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{S}{n} - p\right| < \epsilon\right) &= \mathbb{P}\left(-\epsilon\sqrt{\frac{n}{p(1-p)}} < \frac{S - np}{\sqrt{np(1-p)}} < \epsilon\sqrt{\frac{n}{p(1-p)}}\right) \\
&\approx \Phi\left(\epsilon\sqrt{\frac{n}{p(1-p)}}\right) - \Phi\left(-\epsilon\sqrt{\frac{n}{p(1-p)}}\right) \\
&= 2\Phi\left(\epsilon\sqrt{\frac{n}{p(1-p)}}\right) - 1 \geqslant 0{\cdot}95, \quad \text{since } \Phi(x) = 1 - \Phi(-x),
\end{aligned}$$

so that we need $\epsilon\sqrt{n/p(1-p)} \geqslant 1{\cdot}96$, that is

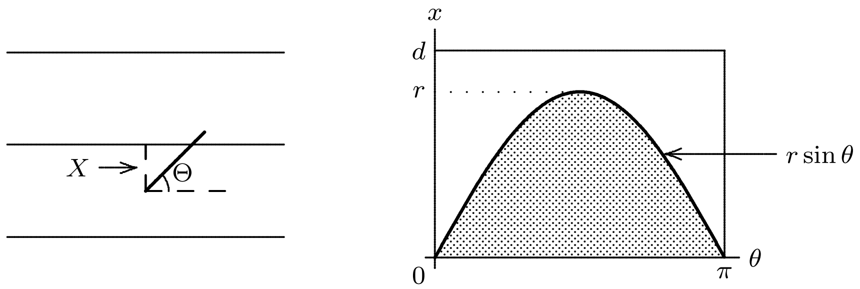
$$n \geqslant (1{\cdot}96)^2 p(1-p)/\epsilon^2.$$

We do not know the value of p, but it is always the case that $p(1-p) \leqslant \frac{1}{4}$, with equality occurring when $p = \frac{1}{2}$, so to ensure that we have the required bound we need $n \geqslant (1{\cdot}96/2\epsilon)^2$. For example, if we take $\epsilon = 0{\cdot}02$, so that the estimate of the percentage of Labour voters is accurate to within 2 percentage points with 95% probability we would need to take a sample with $n \geqslant 2401$. The typical sample size in opinion polls is $n \approx 1000$ which corresponds to an error $\epsilon \approx 0{\cdot}03$. $\square$

5.4 Geometric probability

Buffon's Needle Consider a needle of length r which is thrown at random on to a plane surface on which there are parallel straight lines at distance $d > r$ apart. What is the probability that the needle intersects one of the lines?

Think of the parallel lines running West-East and let X be the distance from the point representing the Southern end of the needle to the nearest line North of that point; if the needle is parallel to the lines, take the right-hand end point. Let Θ be the angle that the needle makes with the West-East lines. Then we will assume that X is uniformly distributed on $[0, d)$ and Θ is uniformly distributed on $[0, \pi]$, and that X and Θ are independent.



The joint probability density of (X, Θ) is

$$f(x, \theta) = \begin{cases} \dfrac{1}{\pi d}, & \text{for } 0 \leqslant x < d \text{ and } 0 \leqslant \theta \leqslant \pi, \\ 0 & \text{otherwise.} \end{cases}$$

Then if A is the shaded area in the $x-\theta$ plane illustrated, the probability that the needle intersects a line is

$$\mathbb{P}\left(X \leqslant r\sin\Theta\right) = \iint\limits_{A} f(x,\theta)dxd\theta = \int_0^{\pi}\int_0^{r\sin\theta}\frac{1}{\pi d}dxd\theta = \frac{r}{\pi d}\int_0^{\pi}\sin\theta d\theta = \frac{2r}{\pi d}\,.$$

This probability was derived in 1777 by the French mathematician and naturalist Georges Louis Leclerc, Comte de Buffon, who suggested a method of approximating the value of π by repeatedly dropping a needle and estimating the probability that a line is intersect (and hence the value of π) by recording the proportion of times that the needle crosses a line.

First note that if $X \sim N(\mu, \sigma^2)$, where σ^2 is small, and $g : \mathbb{R} \to \mathbb{R}$ then

$$g(X) = g(\mu) + (X-\mu)g'(\mu) + \cdots \simeq N\left(g(\mu), \left(g'(\mu)\right)^2\sigma^2\right),$$

where the symbol $\simeq$ may be read as "approximately distributed as". Now, if $S_n = X_1 + \cdots + X_n$ denotes the total number of times that the needle intersects a line in n drops of the needle, where X_i is the indicator of the event that a line is intersected on the ith drop, then $S_n \sim \mathrm{Bin}(n,p)$ where $p = 2r/(\pi d)$. By the Central Limit Theorem we have that $S_n/n \simeq N(p, p(1-p)/n)$. Let $g(x) = 2r/(xd)$, so that $g(p) = \pi$ and $g'(p) = -\pi^2 d/(2r)$. We see that an estimate of π is given by $\hat{\pi} = g(S_n/n)$, where

$$\hat{\pi} \simeq N\left(\pi, \frac{\pi^2}{2rn}\left(\pi d - 2r\right)\right).$$

One small difficulty that arises when one tries to replicate Buffon's procedure for estimating π on a computer is how to do the simulation without using the value of π to take random samples of Θ from the uniform distribution on $(0,\pi]$. One way around this is to generate a sample (X,Y) which has the uniform distribution over the circle centre the origin and of unit radius as follows:

Step 1. generate independent X and Y each with the uniform distribution on $[-1,1]$;

Step 2. if $X^2+Y^2>1$ repeat Step 1, otherwise take (X,Y).

Now set $\Theta = \frac{1}{2}\tan^{-1}(Y/X)$, which will be uniform on $(0,\pi)$. □

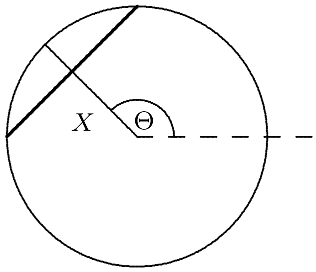
Bertrand's Paradox This results from the following question posed by Bertrand in 1889: What is the probability that a chord chosen at random joining two points of a circle of radius r has length $\leqslant r$?

The difficulty with the question is that there is not a unique interpretation of what it means for a chord to be chosen 'at random'; there are different ways to do this and they lead to different probabilities for the length, C, of the chord being less than r. We will consider two approaches.

Approach 1. Let X be a random variable having the uniform distribution on $(0, r)$ and let Θ be a random variable, independent of X, with the uniform distribution on $(0, 2\pi)$.

The length of the chord is

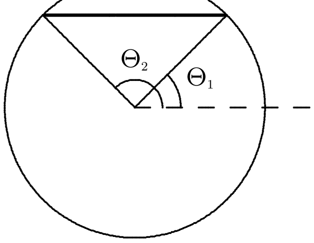
$$C = 2\sqrt{r^2 - X^2}$$

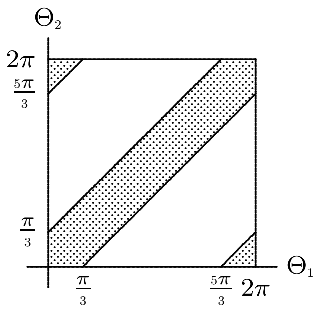


Construct the chord by taking a reference line (the x-axis, say) and drawing the radius at angle Θ with the line. Then take the chord at right angles to this radius at distance X from the centre of the circle. We then have

$$\mathbb{P}(C \leqslant r) = \mathbb{P}\left(4(r^2 - X^2) \leqslant r^2\right) = \mathbb{P}\left(\sqrt{3}r/2 \leqslant X\right) = 1 - \sqrt{3}/2 \approx 0{\cdot}134.$$

Approach 2. Let Θ_1 and Θ_2 be independent random variables each with the uniform distribution on $(0, 2\pi)$. Take the end points of the chord as the points $(r\cos\Theta_i, r\sin\Theta_i)$, $i = 1, 2$, on the circumference of the circle, where the angles are measured from some reference line. The length of the chord is $C = 2r\sin\left(\frac{|\Theta_1 - \Theta_2|}{2}\right)$.





The probability is then

$$\mathbb{P}(C \leqslant r) = \mathbb{P}\left(\sin\left(\frac{|\Theta_1 - \Theta_2|}{2}\right) \leqslant \frac{1}{2}\right) = \mathbb{P}\left(|\Theta_1 - \Theta_2| \leqslant \frac{\pi}{3} \text{ or } |\Theta_1 - \Theta_2| \geqslant \frac{5\pi}{3}\right);$$

this probability is the area of the shaded region in the square divided by $(2\pi)^2$ which gives the probability to be $\frac{1}{3} \approx 0{\cdot}3333$. This probability is the same as when you take one end of the chord as fixed (say on the reference line) and take the other end at a point at angle Θ uniformly distributed on $(0, 2\pi)$.

It should be noted that both probabilities may arise as the outcomes of physical experiments which are choosing the chord 'at random'. For example, the probability in Approach 1 would be found if a circular disc of radius r is thrown randomly onto a table on which parallel lines at distance $2r$ are drawn; the chord would be determined by the unique line intersecting the circle and the distribution of the distance of center of the circle to the nearest line would be uniform on $(0, r)$. By contrast, the probability in Approach 2 is obtained if the disc is pivoted on a point on its circumference, which is on a given straight line and the disc is spun around that point, then the chord would be determined by the intersection of the given line and the circumference of the disc.

9 January 2007