Probability

Chance is a part of our everyday lives. Everyday we make judgements based on probability:

- There is a 90% chance Real Madrid will win tomorrow.
- There is a $1/6$ chance that a dice toss will be a 3.

Probability Theory was developed from the study of games of chance by Fermat and Pascal and is the mathematical study of randomness. This theory deals with the possible outcomes of an event and was put onto a firm mathematical basis by Kolmogorov.

The Kolmogorov axioms

Kolmogorov

For a *random experiment* with *sample space* $\mathbf{\Omega}$, then a probability measure P is a function such that

1. for any event $A \in \mathbf{\Omega}$, $P(A) \geq 0$.

2. $P(\mathbf{\Omega}) = 1$.

3. $P\left(\cup_{j \in J} A_j\right) = \sum_{j \in J} P(A_j)$ if $\{A_j : j \in J\}$ is a countable set of incompatible events.

Set theory

The sample space and events in probability obey the same rules as sets and subsets in set theory. Of particular importance are the distributive laws

$$\begin{aligned} A \cup (B \cap C) &= (A \cup B) \cap (B \cup C) \\ A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \end{aligned}$$

and De Morgan's laws:

$$\begin{aligned} \overline{A \cup B} &= \bar{A} \cap \bar{C} \\ \overline{A \cap B} &= \bar{A} \cup \bar{C} \end{aligned}$$

Laws of probability

The basic laws of probability can be derived directly from set theory and the Kolmogorov axioms. For example, for any two events A and B, we have the addition law,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Laws of probability

The basic laws of probability can be derived directly from set theory and the Kolmogorov axioms. For example, for any two events A and B, we have the addition law,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof

$$\begin{aligned} A &= A \cap \Omega \\ &= A \cap (B \cup \bar{B}) \\ &= (A \cap B) \cup (A \cap \bar{B}) \quad \text{by the second distributive law, so} \\ P(A) &= P(A \cap B) + P(A \cap \bar{B}) \quad \text{and similarly for } B. \end{aligned}$$

Also note that

$$\begin{aligned}
A \cup B &= (A \cup B) \cap (B \cup \bar{B}) \\
&= (A \cap \bar{B}) \cup B \quad \text{by the first distributive law} \\
&= (A \cap \bar{B}) \cup \big(B \cap (A \cup \bar{A})\big) \\
&= (A \cap \bar{B}) \cup (B \cap \bar{A}) \cup (A \cap B) \quad \text{so} \\
P(A \cup B) &= P(A \cap \bar{B}) + P(B \cap \bar{A}) + P(A \cap B) \\
&= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\
&= P(A) + P(B) - P(A \cap B).
\end{aligned}$$

Partitions

The previous example is easily extended when we have a sequence of events, $A_1, A_2, \ldots, A_n$, that form a *partition*, that is

$$\bigcup_{i=1}^{n} A_i = \Omega, \quad A_i \cap A_j = \phi \text{ for all } i \neq j.$$

In this case,

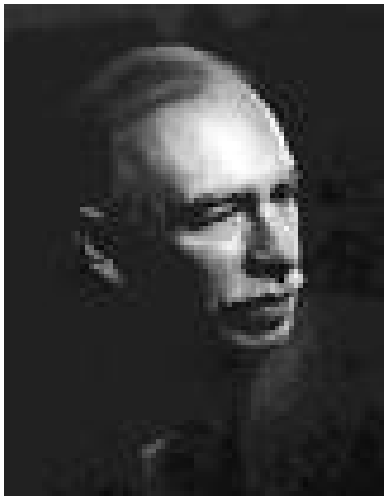
$$\begin{aligned} P\left(\cup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} P(A_i) - \sum_{j>i=1}^{n} P(A_i \cap A_j) + \sum_{k>j>i=1}^{n} P(A_i \cap A_j \cap A_k) + \ldots \\ &\quad +(-1)^n P(A_1 \cap A_2 \cap \ldots \cap A_n). \end{aligned}$$

Interpretations of probability

The Kolmogorov axioms provide a mathematical basis for probability but don't provide for a real life interpretation. Various ways of interpreting probability in real life situations have been proposed.

- Frequentist probability.
- The classical interpretation.
- Subjective probability.
- Other approaches; logical probability and propensities.

Weird approaches



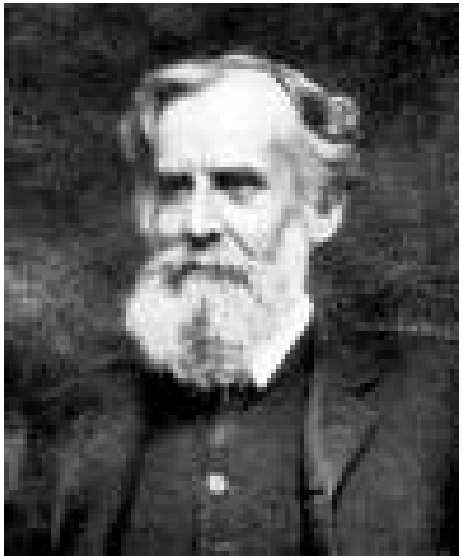
Keynes

- *Logical probability* was developed by Keynes (1921) and Carnap (1950) as an extension of the classical concept of probability. The (conditional) probability of a proposition H given evidence E is interpreted as the (unique) degree to which E logically entails H.

Popper

- Under the theory of *propensities* developed by Popper (1957), probability is an innate disposition or propensity for things to happen. Long run propensities seem to coincide with the frequentist definition of probability although it is not clear what individual propensities are, or whether they obey the probability calculus.

Frequentist probability



Venn

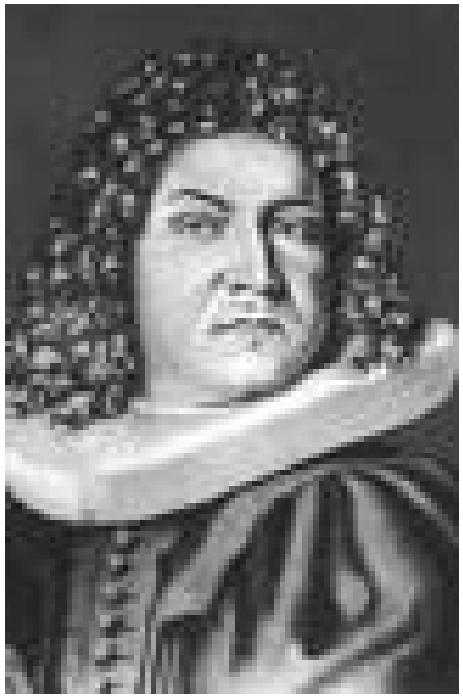
Von Mises

The idea comes from Venn (1876) and von Mises (1919).

Given a repeatable experiment, the probability of an event is defined to be the limit of the proportion of times that the event will occur when the number of repetitions of the experiment tends to infinity.

This is a restricted definition of probability. It is impossible to assign probabilities in non repeatable experiments.

Classical probability



Bernoulli

This derives from the ideas of Jakob Bernoulli (1713) contained in the *principle of insufficient reason* (or *principle of indifference*) developed by Laplace (1812) which can be used to provide a way of assigning epistemic or subjective probabilities.

The principle of insufficient reason

If we are ignorant of the ways an event can occur (and therefore have no reason to believe that one way will occur preferentially compared to another), the event will occur equally likely in any way.

Thus the probability of an event, S, is the coefficient between the number of favourable cases and the total number of possible cases, that is

$$P(S) = \frac{|S|}{|\Omega|}.$$

Calculating classical probabilities

The calculation of classical probabilities involves being able to count the number of possible and the number of favourable results in the sample space. In order to do this, we often use variations, permutations and combinations.

Variations

Suppose we wish to draw n cards from a pack of size N without replacement, then the number of possible results is

$$V_N^n = N \times (N-1) \times (N-n+1) = \frac{N!}{(N-n)!}.$$

Note that one variation is different from another if the order in which the cards are drawn is different.

We can also consider the case of drawing cards with replacement. In this case, the number of possible results is $VR_N^n = N^n$.

Example: The birthday problem

What is the probability that among n students in a classroom, at least two will have the same birthday?

Example: The birthday problem

What is the probability that among n students in a classroom, at least two will have the same birthday?

To simplify the problem, assume there are 365 days in a year and that the probability of being born is the same for every day.

Let S_n be the event that at least 2 people have the same birthday.

$$\begin{aligned} P(S_n) &= 1 - P(\bar{S}_n) \\ &= 1 - \frac{\#\text{ elementary events where nobody has the same birthday}}{\#\text{ elementary events}} \\ &= 1 - \frac{\#\text{ elementary events where nobody has the same birthday}}{365^n} \end{aligned}$$

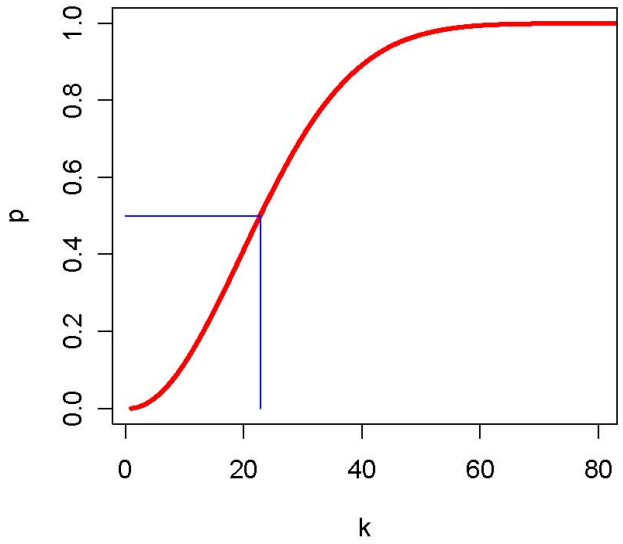
because the denominator is a variation with repetition.

$$P(\bar{S}_n) = \frac{\frac{365!}{(365-n)!}}{365^n}$$

because the numerator is a variation without repetition.

Therefore, $P(S_n) = 1 - \frac{365!}{(365-n)!365^n}$.

The diagram shows a graph of $P(S_n)$ against n.



The probability is just over 0.5 for $n = 23$.

Permutations

If we deal all the cards in a pack of size N, then their are $P_N = N!$ possible deals.

If we assume that the pack contains R_1 cards of type one, R_2 of suit 2, ... R_k of type k, then there are

$$PR_N^{R_1,\ldots,R_k} = \frac{N!}{R_1! \times \cdots \times R_k!}$$

different deals.

Combinations

If we flip a coin N times, how many ways are there that we can get n heads and $N - n$ tails?

$$C_N^n = \left(\begin{array}{c} N \\ n \end{array} \right) = \frac{N!}{n!(N-n)!}.$$

Example: The probability of winning the Primitiva

In the Primitiva, each player chooses six numbers between one and forty nine. If these numbers all match the six winning numbers, then the player wins the first prize. What is the probability of winning?

Example: The probability of winning the Primitiva

In the Primitiva, each player chooses six numbers between one and forty nine. If these numbers all match the six winning numbers, then the player wins the first prize. What is the probability of winning?

The game consists of choosing 6 numbers from 49 possible numbers and there are $\binom{49}{6}$ ways of doing this. Only one of these combinations of six numbers is the winner, so the probability of winning is

$$\frac{1}{\binom{49}{6}} = \frac{1}{13983816}$$

or almost 1 in 14 million.

A more interesting problem is to calculate the probability of winning the second prize. To do this, the player has to match exactly 5 of the winning numbers and the bonus ball drawn at random from the 43 losing numbers.

A more interesting problem is to calculate the probability of winning the second prize. To do this, the player has to match exactly 5 of the winning numbers and the bonus ball drawn at random from the 43 losing numbers.

The player must match 5 of the six winning numbers and there are $C_6^5 = 6$ ways of doing this. Also, they must match exactly the bonus ball and there are $C_1^1 = 1$ ways of doing this. Thus, the probability of winning the second prize is

$$\frac{6 \times 1}{13983816}$$

which is just under one in two millions.

Subjective probability

Ramsey

A different approach uses the concept of ones own probability as a subjective measure of ones own uncertainty about the occurrence of an event. Thus, we may all have different probabilities for the same event because we all have different experience and knowledge. This approach is more general than the other methods as we can now define probabilities for unrepeatable experiments. Subjective probability is studied in detail in Bayesian Statistics.

Conditional probability and independence

The probability of an event B conditional on an event A is defined as

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

This can be interpreted as the probability of B given that A occurs.

Two events A and B are called *independent* if $P(A \cap B) = P(A)P(B)$ or equivalently if $P(B|A) = P(B)$ or $P(A|B) = P(A)$.

The multiplication law

A restatement of the conditional probability formula is the *multiplication law*

$$P(A \cap B) = P(B|A)P(A).$$

Example 12

What is the probability of getting two cups in two draws from a Spanish pack of cards?

Write C_i for the event that draw i is a cup for $i = 1, 2$. Enumerating all the draws with two cups is not entirely trivial. However, the conditional probabilities are easy to calculate:

$$P(C_1 \cap C_2) = P(C_2|C_1)P(C_1) = \frac{9}{39} \times \frac{10}{40} = \frac{3}{52}.$$

The multiplication law can be extended to more than two events. For example,

$$P(A \cap B \cap C) = P(C|A, B)P(B|A)P(A).$$

The birthday problem revisited

We can also solve the birthday problem using conditional probability. Let b_i be the birthday of student i, for $i = 1, \ldots, n$. Then it is easiest to calculate the probability that all birthdays are distinct

$$\begin{aligned} P(b_1 \neq b_2 \neq \ldots \neq b_n) &= P(b_n \notin \{b_1, \ldots, b_{n-1}\} | b_1 \neq b_2 \neq \ldots b_{n-1}) \times \\ & P(b_{n-1} \notin \{b_1, \ldots, b_{n-2}\} | b_1 \neq b_2 \neq \ldots b_{n-2}) \times \cdots \\ & \times P(b_3 \notin \{b_1, b_2\} | b_1 \neq b_2) P(b_1 \neq b_2) \end{aligned}$$

Now clearly,

$$P(b_1 \neq b_2) = \frac{364}{365}, \quad P(b_3 \notin \{b_1, b_2\} | b_1 \neq b_2) = \frac{363}{365}$$

and similarly

$$P(b_i \notin \{b_1, \ldots, b_{i-1}\} | b_1 \neq b_2 \neq \ldots b_{i-1}) = \frac{366 - i}{365}$$

for $i = 3, \ldots, n$.

Thus, the probability that at least two students have the same birthday is, for $n < 365$,

$$1 - \frac{364}{365} \times \cdots \times \frac{366 - n}{365} = \frac{365!}{365^n (365 - n)!}.$$

The law of total probability

The simplest version of this rule is the following.

Theorem 3

For any two events A and B, then

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}).$$

We can also extend the law to the case where $A_1, \ldots, A_n$ form a partition. In this case, we have

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

Bayes theorem

Theorem 4

For any two events A and B, then

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

Supposing that $A_1, \ldots, A_n$ form a partition, using the law of total probability, we can write Bayes theorem as

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} \quad \text{for } j = 1, \ldots, n.$$

The Monty Hall problem

Example 13

The following statement of the problem was given in a column by Marilyn vos Savant in a column in Parade magazine in 1990.

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Simulating the game

Have a look at the following web page.

http://www.stat.sc.edu/~west/javahtml/LetsMakeaDeal.html

Simulating the game

Have a look at the following web page.

http://www.stat.sc.edu/~west/javahtml/LetsMakeaDeal.html

Using Bayes theorem

http://en.wikipedia.org/wiki/Monty_Hall_problem

Random variables

A random variable generalizes the idea of probabilities for events. Formally, a random variable, X simply assigns a numerical value, x_i to each event, A_i, in the sample space, Ω. For mathematicians, we can write X in terms of a mapping, $X : \Omega \to \mathbb{R}$.

Random variables may be classified according to the values they take as

- discrete
- continuous
- mixed

Discrete variables

Discrete variables are those which take a discrete set range of values, say $\{x_1, x_2, \ldots\}$. For such variables, we can define the *cumulative distribution function*,

$$F_X(x) = P(X \leq x) = \sum_{i, x_i \leq x} P(X = x_i)$$

where $P(X = x)$ is the *probability function* or *mass function*.

For a discrete variable, the *mode* is defined to be the point, $\hat{x}$, with maximum probability, i.e. such that

$$P(X = x) < P(X = \hat{x}) \text{for all } x \neq \hat{x}.$$

Moments

For any discrete variable, X, we can define the mean of X to be

$$\mu_X = E[X] = \sum_i x_i P(X = x_i).$$

Recalling the frequency definition of probability, we can interpret the mean as the limiting value of the sample mean from this distribution. Thus, this is a measure of location.

In general we can define the expectation of any function, $g(X)$ as

$$E[g(X)] = \sum_i g(x_i) P(X = x_i).$$

In particular, the variance is defined as

$$\sigma^2 = V[X] = E\left[(X - \mu_X)^2\right]$$

and the standard deviation is simply $\sigma = \sqrt{\sigma^2}$. This is a measure of spread.

Probability inequalities

For random variables with given mean and variance, it is often possible to bound certain quantities such as the probability that the variable lies within a certain distance of the mean.

An elementary result is *Markov's inequality*.

Theorem 5

Suppose that X is a non-negative random variable with mean $E[X] < \infty$. Then for any $x > 0$,

$$P(X \geq x) \leq \frac{E[X]}{x}.$$

Proof

$$
\begin{aligned}
E[X] &= \int_0^\infty u f_X(u)\,du \\
&= \int_0^x u f_X(u)\,du + \int_x^\infty u f_X(u)\,du \\
&\geq \int_x^\infty u f_X(u)\,du \quad \text{because the first integral is non-negative} \\
&\geq \int_x^\infty x f_X(u)\,du \quad \text{because } u \geq x \text{ in this range} \\
&= xP(X \geq x)
\end{aligned}
$$

which proves the result.

∎

Markov's inequality is used to prove *Chebyshev's inequality*.

Chebyshev's inequality

It is interesting to analyze the probability of being close or far away from the mean of a distribution. *Chebyshev's inequality* provides loose bounds which are valid for any distribution with finite mean and variance.

Theorem 6

For any random variable, X, with finite mean, μ, and variance, σ^2, then for any $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Therefore, for any random variable, X, we have, for example that $P(\mu - 2\sigma < X < \mu + 2\sigma) \geq \frac{3}{4}$.

Proof

$$\begin{aligned} P(|X-\mu| \geq k\sigma) &= P\left((X-\mu)^2 \geq k^2\sigma^2\right) \\ &\leq \frac{E\left[(X-\mu)^2\right]}{k^2\sigma^2} \quad \text{by Markov's inequality} \\ &= \frac{1}{k^2} \end{aligned}$$

Chebyshev's inequality shows us, for example, that $P(\mu - \sqrt{2}\sigma \leq X \leq \mu + \sqrt{2}\sigma) \geq 0.5$ for any variable X.

Important discrete distributions

The binomial distribution

Let X be the number of heads in n independent tosses of a coin such that $P(\text{head}) = p$. Then X has a binomial distribution with parameters n and p and we write $X \sim \mathcal{BI}(n, p)$. The mass function is

$$P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.$$

The mean and variance of X are np and $np(1-p)$ respectively.

An inequality for the binomial distribution

Chebyshev's inequality is not very tight. For the binomial distribution, a much stronger result is available.

Theorem 7
Let $X \sim \mathcal{BI}(n, p)$. Then

$$P(|X - np| > n\epsilon) \leq 2e^{-2n\epsilon^2}.$$

Proof See Wasserman (2003), Chapter 4. ∎

The geometric distribution

Suppose that Y is defined to be the number of tails observed before the first head occurs for the same coin. Then Y has a geometric distribution with parameter p, i.e. $Y \sim \mathcal{GE}(p)$ and

$$P(Y = y) = p(1-p)^y \quad \text{for } y = 0, 1, 2, \ldots$$

The mean any variance of X are $\frac{1-p}{p}$ and $\frac{1-p}{p^2}$ respectively.

The negative binomial distribution

A generalization of the geometric distribution is the negative binomial distribution. If we define Z to be the number of tails observed before the r'th head is observed, then $Z \sim \mathcal{NB}(r, p)$ and

$$P(Z = z) = \begin{pmatrix} r+z-1 \\ z \end{pmatrix} p^r (1-p)^z \quad \text{for } z = 0, 1, 2, \ldots$$

The mean and variance of X are $r\frac{1-p}{p}$ and $r\frac{1-p}{p^2}$ respectively.

The negative binomial distribution reduces to the geometric model for the case $r = 1$.

The hypergeometric distribution

Suppose that a pack of N cards contains R red cards and that we deal n cards without replacement. Let X be the number of red cards dealt. Then X has a hypergeometric distribution with parameters N, R, n, i.e. $X \sim \mathcal{HG}(N, R, n)$ and

$$P(X = x) = \frac{\begin{pmatrix} R \\ x \end{pmatrix}\begin{pmatrix} N-R \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \quad \text{for } x = 0, 1, \ldots, n.$$

Example 14

In the Primitiva lottery, a contestant chooses 6 numbers from 1 to 49 and 6 numbers are drawn without replacement. The contestant wins the grand prize if all numbers match. The probability of winning is thus

$$P(X = x) = \frac{\begin{pmatrix} 6 \\ 6 \end{pmatrix}\begin{pmatrix} 43 \\ 0 \end{pmatrix}}{\begin{pmatrix} 49 \\ 6 \end{pmatrix}} = \frac{6!43!}{49!} = \frac{1}{13983816}.$$

What if N and R are large?

For large N and R, then the factorials in the hypergeometric probability expression are often hard to evaluate.

Example 15

Suppose that $N = 2000$ and $R = 500$ and $n = 20$ and that we wish to find $P(X = 5)$. Then the calculation of $2000!$ for example is very difficult.

What if N and R are large?

For large N and R, then the factorials in the hypergeometric probability expression are often hard to evaluate.

Example 15

Suppose that $N = 2000$ and $R = 500$ and $n = 20$ and that we wish to find $P(X = 5)$. Then the calculation of $2000!$ for example is very difficult.

Theorem 8

Let $X \sim \mathcal{HG}(N, R, n)$ and suppose that $R, N \to \infty$ and $R/N \to p$. Then

$$P(X = x) \to \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n.$$

Proof

$$
\begin{aligned}
P(X=x) &= \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}} = \frac{\binom{n}{x}\binom{N-n}{R-x}}{\binom{N}{R}} \\
&= \binom{n}{x}\frac{R!(N-R)!(N-n)!}{(R-x)!(N-R-n+x)!N!} \\
&\to \binom{n}{x}\frac{R^x(N-R)^{n-x}}{N^n} \to \binom{n}{x}p^x(1-p)^{n-x}
\end{aligned}
$$

In the example,$p = 500/2000 = 0.25$ and using a binomial approximation, $P(X=5) \approx \binom{20}{5} 0.25^5 0.75^{15} = 0.2023$. The exact answer, from Matlab is 0.2024.

The Poisson distribution

Assume that rare events occur on average at a rate λ per hour. Then we can often assume that the number of rare events X that occur in a time period of length t has a Poisson distribution with parameter (mean and variance) λt, i.e. $X \sim \mathcal{P}(\lambda t)$. Then

$$P(X = x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \quad \text{for } x = 0, 1, 2, \ldots$$

The Poisson distribution

Assume that rare events occur on average at a rate λ per hour. Then we can often assume that the number of rare events X that occur in a time period of length t has a Poisson distribution with parameter (mean and variance) λt, i.e. $X \sim \mathcal{P}(\lambda t)$. Then

$$P(X = x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \quad \text{for } x = 0, 1, 2, \ldots$$

Formally, the conditions for a Poisson distribution are

- The numbers of events occurring in non-overlapping intervals are independent for all intervals.
- The probability that a single event occurs in a sufficiently small interval of length h is $\lambda h + o(h)$.
- The probability of more than one event in such an interval is $o(h)$.

Continuous variables

Continuous variables are those which can take values in a continuum. For a continuous variable, X, we can still define the distribution function, $F_X(x) = P(X \leq x)$ but we cannot define a probability function $P(X = x)$. Instead, we have the density function

$$f_X(x) = \frac{dF(x)}{dx}.$$

Thus, the distribution function can be derived from the density as $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$. In a similar way, moments of continuous variables can be defined as integrals,

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

and the mode is defined to be the point of maximum density.

For a continuous variable, another measure of location is the *median*, $\tilde{x}$, defined so that $F_X(\tilde{x}) = 0.5$.

Important continuous variables

The uniform distribution

This is the simplest continuous distribution. A random variable, X, is said to have a uniform distribution with parameters a and b if

$$f_X(x) = \frac{1}{b-a} \quad \text{for } a < x < b.$$

In this case, we write $X \sim \mathcal{U}(a, b)$ and the mean and variance of X are $\frac{a+b}{2}$ and $\frac{(b-a)^2}{12}$ respectively.

The exponential distribution

Remember that the Poisson distribution models the number of rare events occurring at rate λ in a given time period. In this scenario, consider the distribution of the time between any two successive events. This is an exponential random variable, $Y \sim \mathcal{E}(\lambda)$, with density function

$$f_Y(y) = \lambda e^{-\lambda y} \quad \text{for } y > 0.$$

The mean and variance of Y are $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$ respectively.

The gamma distribution

A distribution related to the exponential distribution is the gamma distribution. If instead of considering the time between 2 random events, we consider the time between a higher number of random events, then this variable is gamma distributed, that is $Y \sim \mathcal{G}(\alpha, \lambda)$, with density function

$$f_Y(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y} \quad \text{for } y > 0.$$

The mean and variance of Y are $\frac{\alpha}{\lambda}$ and $\frac{\alpha}{\lambda^2}$ respectively.

The normal distribution

This is probably the most important continuous distribution. A random variable, X, is said to follow a normal distribution with mean and variance parameters μ and σ^2 if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad \text{for } -\infty < x < \infty.$$

In this case, we write $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

- If X is normally distributed, then $a + bX$ is normally distributed. In particular, $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.
- $P(|X-\mu| \geq \sigma) = 0.3174$, $P(|X-\mu| \geq 2\sigma) = 0.0456$, $P(|X-\mu| \geq 3\sigma) = 0.0026$.
- Any sum of normally distributed variables is also normally distributed.

Example 16

Let $X \sim \mathcal{N}(2, 4)$. Find $P(3 < X < 4)$.

$$\begin{aligned} P(3 < X < 4) &= P\left(\frac{3-2}{\sqrt{4}} < \frac{X-2}{\sqrt{4}} < \frac{4-2}{\sqrt{4}}\right) \\ &= P(0.5 < Z < 1) \quad \text{where } Z \sim \mathcal{N}(0, 1) \\ &= P(Z < 1) - P(Z < 0.5) = 0.8413 - 0.6915 \\ &= 0.1499 \end{aligned}$$

The central limit theorem

One of the main reasons for the importance of the normal distribution is that it can be shown to approximate many real life situations due to the central limit theorem.

Theorem 9
Given a random sample of size $X_1, \ldots, X_n$ from some distribution, then under certain conditions, the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ follows a normal distribution.

Proof See later. ∎

For an illustration of the CLT, see

http://cnx.rice.edu/content/m11186/latest/

Mixed variables

Occasionally it is possible to encounter variables which are partially discrete and partially continuous. For example, the time spent waiting for service by a customer arriving in a queue may be zero with positive probability (as the queue may be empty) and otherwise takes some positive value in $(0, \infty)$.

The probability generating function

For a discrete random variable, X, taking values in some subset of the non-negative integers, then the probability generating function, $G_X(s)$ is defined as

$$G_X(s) = E\left[s^X\right] = \sum_{x=0}^{\infty} P(X = x)s^x.$$

This function has a number of useful properties:

- $G(0) = P(X = 0)$ and more generally, $P(X = x) = \frac{1}{x!}\frac{d^x G(s)}{ds^x}\,|_{s=0}$.
- $G(1) = 1$, $E[X] = \frac{dG(1)}{ds}$ and more generally, the k'th factorial moment, $E[X(X-1)\cdots(X-k+1)]$, is

$$E\left[\frac{X!}{(X-k)!}\right] = \left.\frac{d^k G(s)}{ds^k}\right|_{s=1}$$

- The variance of X is

$$V[X] = G''(1) + G'(1) - G'(1)^2.$$

Example 17

Consider a negative binomial variable, $X \sim \mathcal{NB}(r, p)$.

$$\begin{aligned} P(X = x) &= \binom{r+x-1}{x} p^r (1-p)^x \quad \text{for } z = 0, 1, 2, \ldots \\ E[s^X] &= \sum_{x=0}^{\infty} s^x \binom{r+x-1}{x} p^r (1-p)^x \\ &= p^r \sum_{x=0}^{\infty} \binom{r+x-1}{x} \{(1-p)s\}^x = \left(\frac{p}{1-(1-p)s}\right)^r \end{aligned}$$

$$
\begin{aligned}
\frac{dE}{ds} &= \frac{rp^r(1-p)}{(1-(1-p)s)^{r+1}} \\
\left.\frac{dE}{ds}\right|_{s=1} &= r\frac{1-p}{p} = E[X] \\
\frac{d^2E}{ds^2} &= \frac{r(r+1)p^r(1-p)^2}{(1-(1-p)s)^{r+2}} \\
\left.\frac{d^2E}{ds^2}\right|_{s=1} &= r(r+1)\left(\frac{1-p}{p}\right)^2 = E[X(X-1)] \\
V[X] &= r(r+1)\left(\frac{1-p}{p}\right)^2 + r\frac{1-p}{p} - \left(r\frac{1-p}{p}\right)^2 \\
&= r\frac{1-p}{p^2}.
\end{aligned}
$$

The moment generating function

For any variable, X, the moment generating function of X is defined to be

$$M_X(s) = E\left[e^{sX}\right].$$

This generates the moments of X as we have

$$\begin{aligned} M_X(s) &= E\left[\sum_{i=1}^{\infty} \frac{(sX)^i}{i!}\right] \\ \left.\frac{d^i M_X(s)}{ds^i}\right|_{s=0} &= E\left[X^i\right] \end{aligned}$$

Example 18

Suppose that $X \sim \mathcal{G}(\alpha, \beta)$. Then

$$
\begin{aligned}
f_X(x) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } x > 0 \\
M_X(s) &= \int_0^\infty e^{sx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\, dx \\
&= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-s)x}\, dx \\
&= \left(\frac{\beta}{\beta - s}\right)^\alpha \\
\frac{dM}{ds} &= \frac{\alpha\beta^\alpha}{(\beta-s)^{\alpha-1}} \\
\left.\frac{dM}{ds}\right|_{s=0} &= \frac{\alpha}{\beta} = E[X]
\end{aligned}
$$

Example 19

Suppose that $X \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned}
M_X(s) &= \int_{-\infty}^{\infty} e^{sx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[x^2 - 2s\right]\right)\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[x^2 - 2s + s^2 - s^2\right]\right)\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[(x-s)^2 - s^2\right]\right)\, dx \\
&= e^{\frac{s^2}{2}}.
\end{aligned}$$

Transformations of random variables

Often we are interested in transformations of random variables, say $Y = g(X)$. If X is discrete, then it is easy to derive the distribution of Y as

$$P(Y = y) = \sum_{x, g(x)=y} P(X = x).$$

However, when X is continuous, then things are slightly more complicated.

If $g(\cdot)$ is monotonic so that $\frac{dy}{dx} = g'(x) \neq 0$ for all x, then for any y, we can define a unique inverse function, $g^{-1}(y)$ such that

$$\frac{dg^{-1}(y)}{dy} = \frac{dx}{dy} = \frac{1}{\frac{dy}{dx}} = \frac{1}{g'(x)}.$$

Then, we have

$$\int f_X(x)\,dx = \int f_X\left(g^{-1}(y)\right)\frac{dx}{dy}\,dy$$

and so the density of Y is given by

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dx}{dy}\right|$$

If g does not have a unique inverse, then we can divide the support of X up into regions, i, where a unique inverse, g_i^{-1} does exist and then

$$f_Y(y) = \sum_i f_X\left(g_i^{-1}(y)\right)\left|\frac{dx}{dy}\right|_i$$

where the derivative is that of the inverse function over the relevant region.

Derivation of the χ^2 distribution

Example 20

Suppose that $Z \sim \mathcal{N}(0,1)$ and that $Y = Z^2$. Then the function $g(z) = z^2$ has a unique inverse for $z < 0$, that is $g^{-1}(z) = -\sqrt{z}$ and for $z \geq 0$, that is $g^{-1}(z) = \sqrt{z}$ and in each case, $|\frac{dg}{dz}| = |2z|$ so therefore, we have

$$\begin{aligned} f_Y(y) &= 2 \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y}{2}\right) \times \frac{1}{2\sqrt{y}} \quad \text{for } y > 0 \\ &= \frac{1}{\sqrt{2\pi}} y^{\frac{1}{2}-1} \exp\left(-\frac{y}{2}\right) \\ Y &\sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2}\right) \end{aligned}$$

which is a chi-square distribution with one degree of freedom.

Linear transformations

If $Y = aX + b$, then immediately, we have $\frac{dy}{dx} = a$ and that $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$. Also, in this case, we have the well known results

$$\begin{aligned} E[Y] &= a + bE[X] \\ V[Y] &= b^2 V[X] \end{aligned}$$

so that, in particular, if we make the standardizing transformation, $Y = \frac{X-\mu_X}{\sigma_X}$, then $\mu_Y = 0$ and $\sigma_Y = 1$.

Jensen's inequality

This gives a result for the expectations of convex functions of random variables.

A function $g(x)$ is convex if for any x, y and $0 \leq p \leq 1$, we have

$$g(px + (1-p)y) \leq pg(x) + (1-p)g(y).$$

(Otherwise the function is concave.) It is well known that for a twice differentiable function with $g''(x) \geq 0$ for all x, then g is convex. Also, for a convex function, the function always lies above the tangent line at any point $g(x)$.

Theorem 10

If g is convex, then

$$E[g(X)] \geq g(E[X])$$

and if g is concave, then

$$E[g(X)] \leq g(E[X])$$

Proof Let $L(x) = a + bx$ be a tangent to $g(x)$ at the mean $E[X]$. If g is convex, then $L(X) \leq g(X)$ so that

$$E[g(X)] \geq E[L(X)] = a + bE[X] = L(E[X]) = g(E[X]).$$

One trivial application of this inequality is that $E[X^2] \geq E[X]^2$.

Multivariate distributions

It is straightforward to extend the concept of a random variable to the multivariate case. Full details are included in the course on Multivariate Analysis.

For two discrete variables, X and Y, we can define the joint probability function at $(X = x, Y = y)$ to be $P(X = x, Y = y)$ and in the continuous case, we similarly define a joint density function $f_{X,Y}(x, y)$ such that

$$\begin{aligned}
\sum_x \sum_y P(X = x, Y = y) &= 1 \\
\sum_y P(X = x, Y = y) &= P(X = x) \\
\sum_x P(X = x, Y = y) &= P(Y = y)
\end{aligned}$$

and similarly for the continuous case.

Conditional distributions

The conditional distribution of Y given $X = x$ is defined to be

$$f_{Y|x}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Two variables are said to be *independent* if for all x, y, then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ or equivalently if $f_{Y|x}(y|x) = f_Y(y)$ or $f_{X|y}(x|y) = f_X(x)$.

We can also define the conditional expectation of $Y|x$ to be $E[Y|x] = \int y f_{Y|x}(y|x)\,dx$.

Covariance and correlation

It is useful to obtain a measure of the degree of relation between the two variables. Such a measure is the *correlation*.

We can define the expectation of any function, $g(X, Y)$, in a similar way to the univariate case,

$$E[g(X,Y)] = \int \int g(x,y) f_{X,Y}(x,y)\, dx\, dy.$$

In particular, the *covariance* is defined as

$$\sigma_{X,Y} = Cov[X,Y] = E[XY] - E[X]E[Y].$$

Obviously, the units of the covariance are the product of the units of X and Y. A scale free measure is the *correlation*,

$$\rho_{X,Y} = Corr[X,Y] = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

Properties of the correlation are as follows:

- $-1 \leq \rho_{X,Y} \leq 1$
- $\rho_{X,Y} = 0$ if X and Y are independent. (This is not necessarily true in reverse!)
- $\rho_{X_Y} = 1$ if there is an exact, positive relation between X and Y so that $Y = a + bX$ where $b > 0$.
- $\rho_{X_Y} = -1$ if there is an exact, negative relation between X and Y so that $Y = a + bX$ where $b < 0$.

The Cauchy Schwarz inequality

Theorem 11
For two variables, X and Y, then

$$E[XY]^2 \leq E[X^2]E[Y^2].$$

Proof Let $Z = aX - bY$ for real numbers a, b. Then

$$0 \leq E[Z^2] = a^2E[X^2] - 2abE[XY] + b^2E[Y^2]$$

and the right hand side is a quadratic in a with at most one real root. Thus, its discriminant must be non-positive so that if $b \neq 0$,

$$E[XY]^2 - E[X^2]E[Y^2] \leq 0.$$

The discriminant is zero iff the quadratic has a real root which occurs iff $E[(aX - bY)^2] = 0$ for some a and b. ∎

Conditional expectations and variances

Theorem 12

For two variables, X and Y, then

$$\begin{aligned} E[Y] &= E[E[Y|X]] \\ V[Y] &= E[V[Y|X]] + V[E[Y|X]] \end{aligned}$$

Proof

$$\begin{aligned} E[E[Y|X]] &= E\left[\int y f_{Y|X}(y|X)\,dy\right] = \int f_X(x) \int y f_{Y|X}(y|X)\,dy\,dx \\ &= \int y \int f_{Y|X}(y|x) f_X(x)\,dx\,dy \\ &= \int y \int f_{X,Y}(x,y)\,dx\,dy \\ &= \int y f_Y(y)\,dy = E[Y] \quad \blacksquare \end{aligned}$$

Example 21

A random variable X has a beta distribution, $X \sim \mathcal{B}(\alpha, \beta)$, if

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } 0 < x < 1.$$

The mean of X is $E[X] = \frac{\alpha}{\alpha+\beta}$.

Suppose now that we toss a coin with probability $P(\text{heads}) = X$ a total of n times and that we require the distribution of the number of heads, Y.

This is the beta-binomial distribution which is quite complicated:

$$
\begin{aligned}
P(Y=y) &= \int_0^1 P(Y=y|X=x) f_X(x)\,dx \\
&= \int_0^1 \left(\begin{array}{c} n \\ y \end{array} \right) x^y (1-x)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,dx \\
&= \left(\begin{array}{c} n \\ y \end{array} \right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+y-1}(1-x)^{\beta+n-y-1}\,dx \\
&= \left(\begin{array}{c} n \\ y \end{array} \right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y)\Gamma(\beta+n-y)}{\Gamma(\alpha+\beta+n)}
\end{aligned}
$$

for $y = 0, 1, \ldots, n$.

We could try to calculate the mean of Y directly using the above probability function. However, this would be very complicated. There is a much easier way.

We could try to calculate the mean of Y directly using the above probability function. However, this would be very complicated. There is a much easier way.

$$\begin{aligned} E[Y] &= E[E[Y|X]] \\ &= E[nX] \quad \text{because } Y|X \sim \mathcal{BI}(n, X) \\ &= n\frac{\alpha}{\alpha+\beta}. \end{aligned}$$

The probability generating function for a sum of independent variables

Suppose that $X_1, \ldots, X_n$ are *independent* with generating functions $G_i(s)$ for $s = 1, \ldots, n$. Let $Y = \sum_{i=1}^{n} X_i$. Then

$$\begin{aligned} G_Y(s) &= E\left[s^Y\right] \\ &= E\left[s^{\sum_{i=1}^{n} X_i}\right] \\ &= \prod_{i=1}^{n} E\left[s^{X_i}\right] \quad \text{by independence} \\ &= \prod_{i=1}^{n} G_i(s) \end{aligned}$$

Furthermore, if $X_1, \ldots, X_n$ are identically distributed, with common generating function $G_X(s)$, then

$$G_Y(s) = G_X(s)^n.$$

Example 22

Suppose that $X_1, \ldots, X_n$ are Bernoulli trials so that

$$P(X_i = 1) = p \quad \text{and} \quad P(X_i = 0) = 1 - p \quad \text{for } i = 1, \ldots, n$$

Then, the probability generating function for any X_i is $G_X(s) = 1 - p + sp$. Now consider a binomial random variable, $Y = \sum_{i=1}^{n} X_i$. Then

$$G_Y(s) = (1 - p + sp)^n$$

is the binomial probability generating function.

Another useful property of pgfs

If N is a discrete variable taking values on the non-negative integers and with pgf $G_N(s)$ and if $X_1, \ldots, X_N$ is a sequence of independent and identically distributed variables with pgf $G_X(s)$, then if $Y = \sum_{i=1}^{N} X_i$, we have

$$\begin{aligned} G_Y(s) &= E\left[s^{\sum_{i=1}^{N} X_i}\right] \\ &= E\left[E\left[s^{\sum_{i=1}^{N} X_i} \mid N\right]\right] \\ &= E\left[G_X(s)^N\right] \\ &= G_N(G_X(s)) \end{aligned}$$

This result is useful in the study of *branching processes*. See the course in Stochastic Processes.

The moment generating function of a sum of independent variables

Suppose we have a sequence of independent variables, $X_1, X_2, \ldots, X_n$ with mgfs $M_1(s), \ldots, M_n(s)$. Then, if $Y = \sum_{i=1}^{n} X_i$, it is easy to see that

$$M_Y(s) = \prod_{i=1}^{n} M_i(s)$$

and if the variables are identically distributed with common mgf $M_X(s)$, then

$$M_Y(s) = M_X(s)^n.$$

Example 23

Suppose that $X_i \sim \mathcal{E}(\lambda)$ for $i = 1, \ldots, n$ are independent. Then

$$\begin{aligned} M_X(s) &= \int_0^\infty e^{sx} \lambda e^{-\lambda x}\, dx \\ &= \lambda \int_0^\infty e^{-(\lambda - s)x}\, dx \\ &= \frac{\lambda}{\lambda - s}. \end{aligned}$$

Therefore the mgf of $Y = \sum_{i=1}^n X_i$ is given by

$$M_Y(s) = \left(\frac{\lambda}{\lambda - s} \right)^n$$

which we can recognize as the mgf of a gamma distribution, $Y \sim \mathcal{G}(n, \lambda)$.

Example 24

Suppose that $X_i \sim \chi^2_1$ for $i = 1, \ldots, n$. Then

$$\begin{aligned} M_{X_i}(s) &= \int_0^\infty e^{sx_i} \frac{1}{\sqrt{2\pi}} x_i^{\frac{1}{2}-1} \exp\left(-\frac{x_i}{2}\right) dx_i \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty x_i^{\frac{1}{2}-1} \exp\left(-\frac{x_i(1-2s)}{2}\right) dx_i \\ &= \sqrt{\frac{1}{1-2s}} \quad \text{so if } Y = \sum_{i=1}^n X_i, \text{ then} \\ M_Y(s) &= \left(\frac{1}{1-2s}\right)^{n/2} \end{aligned}$$

which is the mgf of a gamma distribution, $\mathcal{G}\left(\frac{n}{2}, \frac{1}{2}\right)$ which is the χ^2_n density.

Proof of the central limit theorem

For any variable, Y, with zero mean and unit variance and *such that all moments exist*, then the moment generating function is

$$M_Y(s) = E[e^{sY}] = 1 + \frac{s^2}{2} + o(s^2).$$

Now assume that $X_1, \ldots, X_n$ are a random sample from a distribution with mean μ and variance σ^2. Then, we can define the standardized variables, $Y_i = \frac{X_i - \mu}{\sigma}$, which have mean 0 and variance 1 for $i = 1, \ldots, n$ and then

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$$

Now, suppose that $M_Y(s)$ is the mgf of Y_i, for $i = 1, \ldots, n$. Then

$$M_{Z_n}(s) = M_Y\left(s/\sqrt{n}\right)^n$$

and therefore,

$$M_{Z_n}(s) = \left(1 + \frac{s^2}{2n} + o(s^2/n)\right)^n \to e^{\frac{s^2}{2}}$$

which is the mgf of a normally distributed random variable.

and therefore,

$$M_{Z_n}(s) = \left(1 + \frac{s^2}{2n} + o(s^2/n)\right)^n \to e^{\frac{s^2}{2}}$$

which is the mgf of a normally distributed random variable.

To make this result valid for variables that do not necessarily possess all their moments, then we can use essentially the same arguments but defining the characteristic function $C_X(s) = E[e^{isX}]$ *instead of the moment generating function.*

Sampling distributions

Often, in order to undertake inference, we wish to find the distribution of the sample mean, $\bar{X}$ or the sample variance, S^2.

Theorem 13

Suppose that we take a sample of size n from a population with mean μ and variance σ^2. Then

$$\begin{aligned} E[\bar{X}] &= \mu \\ V[\bar{X}] &= \frac{\sigma^2}{n} \\ E[S^2] &= \sigma^2 \end{aligned}$$

Proof

$$
\begin{aligned}
E[\bar{X}] &= \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n} \int \dots \int \sum_{i=1}^{n} x_i f_{X_1,\dots,X_n}(x_1,\dots,x_n)\, dx_1,\dots,dx_n \\
&= \frac{1}{n} \int \dots \int \sum_{i=1}^{n} x_i f_{X_1}(x_1)\cdots f_{X_n}(x_n)\, dx_1,\dots,dx_n \quad \text{by independence} \\
&= \frac{1}{n} \sum_{i=1}^{n} \int x_i f_{X_i}(x_i)\, dx_i = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu
\end{aligned}
$$

$$
\begin{aligned}
V[\bar{X}] &= \frac{1}{n} V[\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n} \sum_{i=1}^{n} V[X_i] \quad \text{by independence} \\
&= \frac{n\sigma^2}{n} = \sigma^2 \\
E[S^2] &= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right] \\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (X_i - \mu + \mu - \bar{X})^2\right] \\
&= \frac{1}{n-1} \sum_{i=1}^{n} E\left[(X_i - \mu)^2 + 2(X_i - \mu)(\mu - \bar{X}) + (\mu - \bar{X})^2\right] \\
&= \frac{1}{n-1} \left(n\sigma^2 - 2nE\left[(\bar{X} - \mu)^2\right] + n\frac{\sigma^2}{n}\right) \\
&= \frac{1}{n-1} \left(n\sigma^2 - \sigma^2\right) = \sigma^2 \qquad \blacksquare
\end{aligned}
$$

The previous result shows that $\bar{X}$ and S^2 are $unbiased$ estimators of the population mean and variance respectively.

A further important extension can be made if we assume that the data are normally distributed.

Theorem 14
If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ then we have that $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Proof We can prove this using moment generating functions. First recall that if $Z \sim \mathcal{N}(0, 1)$, then $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$ so that

$$M_X(s) = E[e^{sX}] = E[e^{s\mu + s\sigma Z}] = e^{s\mu} E[e^{s\sigma Z}] = e^{s\mu} M_Z(\sigma s).$$

Therefore, we have $M_X(s) = e^{s\mu} e^{\frac{(s\sigma)^2}{2}} = e^{s\mu + \frac{s^2\sigma^2}{2}}$.

Now, suppose that $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\begin{aligned} M_{\bar{X}}(s) &= E[e^{s\bar{X}}] \\ &= E\left[e^{\frac{s}{n}\sum_{i=1}^{n} X_i}\right] \\ &= M_X\left(\frac{s}{n}\right)^n \\ &= \left(e^{\frac{s}{n}\mu + \frac{s^2\sigma^2}{2n^2}}\right)^n \\ &= e^{s\mu + \frac{s^2\sigma^2/n}{2}} \end{aligned}$$

which is the mgf of a normal distribution, $\mathcal{N}(\mu, \sigma^2/n)$. ∎