CSMMA16

Mathematics and Statistics

Probability and Statistics

29th September 2016

Introduction

Probability theory:

- Predict likelihood of future events under some known model of possible outcomes
- Model may be theoretical or result of previous observations

Statistical Inference:

- Analyse past events to develop an understanding of outcomes
- Evaluates and/or constructs models of possible outcomes

Statistical methods apply mathematics to measured (collected) data assuming stochastic data generating mechanism

- What data to collect and when?
- How and where to collect the data?
- How much data to collect?

Samples and Populations

A population is the entire collection of units (people, animals, fields) under investigation.

In many situations it is impossible (or undesirable) to measure the entire population

- opinions of voters before a general election
- width of leaves on a tree
- genomes of all humans

A sample is a subset of the population from which we collect information on variables of interest

A variable is any characteristic which may vary from one unit in the population to another

- An unbiased sample is representative of the population from which it is drawn
- A random sample is where each population unit has the same chance of selection

Types of Data

Variables which we measure can fall into a variety of categories

- they can be quantitative
  - discrete or continuous
- they can be qualitative
  - ordinal (ordered) or nominal (no order)

Which categories do outcomes from the following fit into?

- tossing a coin
- severity of a flu
- height of men living in England
- width of leaves on an oak tree
- votes at a general election

Can you give an example of data which could be classified non-random and continuous?

Traffic example

Below is a sample set of the speed of cars passing a certain point within a 30 mph zone:

$$31, 29, 28, 24, 29, 35, 37, 27, 29, 30, 32, 30, 29, 20, 45, 27, 35, 34, 33, 29$$

Questions we may ask about this data set:

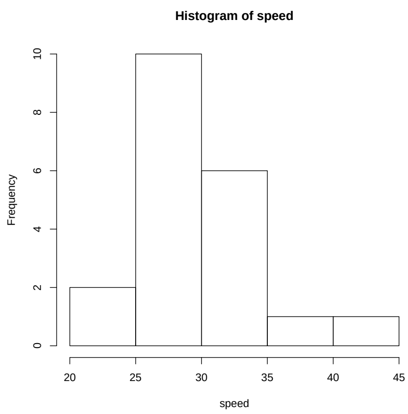
- Is it representative of speed on the road (at the point)?
- How can we use this data
  - to understand speed at the point?
  - to predict the number of cars exceeding the speed limit by 5mph?
  - to predict the speed of the next car?
  - ...

Graphical summaries

- Tables
  - If thought through properly can be a concise way of publishing a lot of detailed information
- Graphs
  - Useful for more visual impact
- Either way we are interested in how a variable varies from observation to observation
  - this is called the variables distribution
  - The distribution tells us what values the variable takes and how often each value occurs
- Qualitative Variables: Barplots, Contingency Tables
- Quantitative Variables: Histograms, Box plots

Traffic histogram

Below is a histogram of the car speed data.
What can we say about this?



In R:

```
speed<-c(31,29,28,...,34,33,29)
hist(speed)
```

Numerical summaries

Numerical summaries descriptive statistics helps us understand key features of the data

- They can also help us identify any errors - very useful
- They can help us interpret/check the results of more complicated analysis

The descriptive statistics we choose to summarise each variable depend on the variable type

- Qualitative (categorical) data - counts, percentages, mode
- Quantitative data - mean, median, variance, standard deviation, range, IQR

Measures of location

- Mean - the arithmetic average of all values for the variable

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

- Median - the middle value of the ordered observations for the variable
  - if there is an even number of values, then the median is half-way between the two middle values
- Mode the most commonly occurring value
  - a variable may have more than one mode

Note that the mean uses every observed value but can be influenced by extreme "outlying" values. It is most useful for symmetric data

The median does not use the actual observed values - just the order. It is most useful for skewed data

Measures of spread

- Range - the difference between the smallest and largest values
- Inter-quartile range (IQR) $= Q_3 - Q_1$ where
  - Q_1 is the first quartile - the value for which 25% of observations lie below and
  - Q_3 is the third quartile - the value for which 75% of observations lie below.
- Variance

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$$

- Standard deviation is square root of the variance

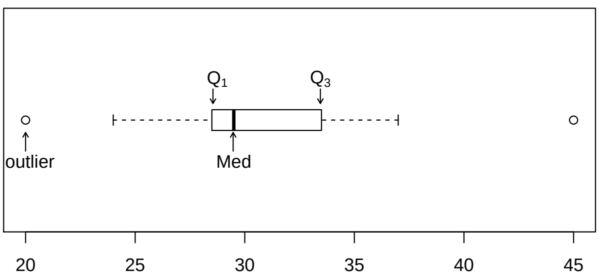
Note that the range is easy to calculate and interpret but it does not use all the data values and is badly affected by extreme values. Also, it tends to increase as the sample size increases.

The variance (and standard deviation) uses all the data values and has nice mathematical properties (unbiased) but is also sensitive to extreme values.

Box plots

Box plots are a good way of visualising data

- it can help you identify the centre, spread, departure from symmetry and any outliers



- Outliers are values outside the interval $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)$
- the whiskers extend from each end of the box to most extreme data points within above interval

Probability

Probability is a measure of chance of an outcome within a given setting (trial)

Let Ω represent the set of all possible outcomes in a trial and $P(A)$ the probability of $A \subset \Omega$.

- $P(A) \geq 0$ for any A
- $P(\Omega) = 1$
- If $A_1, A_2, \ldots$ are disjoint subsets of Ω, the probability of at least one is
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

The above implies that

- $P(\phi) = 0$, where ϕ is the empty set
- $0 \leq P(A) \leq 1$ for any A
- $P(A^c) = 1 - P(A)$
- If $A \subset B$ then $P(A) \leq P(B)$
- If A and B are disjoint then $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$

Properties

$P(A) = 1$ means that event A is certain to happen

$P(A) = 0$ means that event A is certain NOT to happen

It follows from the previous slide that if A and B are mutually exclusive then $P(A \text{ and } B) = P(AB) = 0$

And for any A and B, $P(A \cup B) = P(A) + P(B) - P(AB)$

But how do we measure the chance of an outcome?

This depends on our interpretation of probability.

Interpretations of probability

(a) Gambler's (frequency) interpretation:

$$P(A) = \frac{\text{number of times an event } A \text{ occurs}}{\text{total number of outcomes}}$$

- Based on counts of actual outcomes and assumes each outcome is equally likely
- possible outcomes that do not occur are not counted

(b) Frequentist interpretation:

$$P(A) = \lim_{n\to\infty} \frac{\text{number of occurrences of A}}{\text{number of tries (n)}}$$

- this probability can never be observed and is only estimated
- usually the repetitions are imaginary

Interpretations (cont'd)

(c) Bayesian interpretation:
If someone is willing to accept the odds of $1 - P(A) : P(A)$ on the occurrence of an event A, then

$P(A)$ is the probability that A will happen

- Depends on the person's belief about the chance of A occurring
- As this belief may be different for different people this interpretation is subjective

Conditional probability

Events A and B are independent if the outcome of A has no effect on the outcome of B and vice-versa. Otherwise A and B are dependent events.

The probability of B occurring given that A has occurred (so $P(A) > 0$) is called a conditional probability defined as

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

The concept of conditioning is important and useful in areas such as prediction theory and artificial intelligence

Some conditional probability results

(a) If $A_1, \ldots, A_n$ are disjoint, with union Ω, then

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

(b)

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

(c) *Bayes Theorem:*

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

- This forms the basis for Bayesian Inference

Independence

Definition

Two events A and B are said to be **independent** if
$P(B|A) = P(B)$

Some properties:

- The concept of independence is **well defined**.
  e.g. if A and B are independent (meaning $P(B|A) = P(B)$) then B and A independent (meaning $P(A|B) = P(A)$) and *vice versa*
- Since $P(B|A) = P(A \cap B)/P(A)$ it follows that A and B **independent** if and only if

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

  - Can be extended to any number of independent events.

Permutations and Combinations

- multiplication rule:
  - If you are drawing one element from each of k sets of distinct elements, with sizes of the sets $n_1, n_2, \ldots, n_k$ respectively, then the number of different possible outcomes is

$$n_1 n_2 \cdots n_k$$

- permutations rule:
  - if you are drawing k elements from a set of n and arranging the k elements in a distinct order, the number of different possible results is

$$^{n}P_k = \frac{n!}{(n-k)!}$$

Permutations and Combinations (cont'd)

- partitions rule:
  - if you are partitioning the elements of a set of n elements into k groups of $n_1, n_2, \dots, n_k$ elements, the number of different results is

$$\frac{n!}{n_1! n_2! \cdots n_k!}$$

- combinations rule:
  - if you are drawing k elements from a set of n elements without regard to the order of the k elements, the number of different possible results is

$$^nC_k = \frac{n!}{k!(n-k)!}$$

Random variable

A random variable X is a function that assigns values to the outcomes of a trial:

$$X : \Omega \to \mathbb{R}$$

Example: Suppose that two dice each with outcomes $\{1, \dots, 6\}$ are thrown. Then

$$\Omega = \{(1,1), (1,2), \dots, (1,6), (2,1), \dots, (6,6)\} \text{ and } P(\omega) = \frac{1}{36} \ .$$

Now let the random variable X = sum of the points. Then the values of X are

$$\Omega^* = \{2, 3, \dots, 12\}$$

and the associated probability of observing $X = x$ is

$$P(X = x) = \begin{cases} \frac{x-1}{36} & \text{for } x = 2,3,4,5,6,7 \\ \\ \frac{12-(x-1)}{36} & \text{for } x = 8,9,10,11,12 \end{cases}$$

Probability mass function

- If a random variable is discrete, its possible values and their associated probabilities can be modelled by a probability mass function (p.m.f)
  - this is a mathematical expression which covers all possibilities

For example, the set of outcomes of a test of light-bulbs after 800 hours use are $\Omega = \{\text{working}, \ \text{failed}\}$.

Suppose we let $X = 1$ if "working" and $X = 0$ if "failed" and suppose $P(\text{working}) = p$.

The pmf of X is

$$f(x) = P(X = x) = \begin{cases} p^x(1-p)^{1-x} & \text{for } x = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

Probability density function

A random variable X is continuous if it takes values between a and b means that it can take any value between a and b.

For a continuous random variable X, we call any non-negative, integrable function $f(x)$ whose integral over any interval gives the probability X lies within this interval a probability density function (pdf) of X.

In particular, we have that $\int_{-\infty}^{+\infty} f(x)dx = 1$

Cumulative distribution function

- For discrete X the **cumulative distribution function** (cdf) is

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u),$$

- For continuous X, the cdf is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

  - It follows that $P(a \leq X \leq b) = F(b) - F(a)$

Expectation of a random variable

Let g be a function of a random variable X.

- If X is discrete, the expectation of $g(X)$ is defined as

$$E[g(X)] = \sum_{x} g(x)f(x) = \sum_{x} g(x)P(X = x)$$

  - In particular $E(X) = \sum_{x} xP(X = x)$

- If X is continuous the expectation of $g(X)$ is

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

  - In particular $E(X) = \int_{-\infty}^{\infty} xf(x)dx$

Variance and Covariance

Let $E(X) = \mu_X$ and $E(Y) = \mu_Y$

The covariance of the random variables X and Y is given by

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

If X and Y are independent, $\text{Cov}(X, Y) = 0$.

The variance of X is

$$\text{Var}(X) = E[(X - \mu_X)^2]$$

The correlation of X and Y is

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Bernoulli probability model

A study with only two outcomes, say "success" and "failure", is a Bernoulli trial

Let $p = P(\text{success})$

The associated random variable X takes one of two values (0 and 1) and is a Bernoulli random variable with parameter p.

We met the distribution of X earlier (light bulb example).

Other examples of Bernoulli trials:

- tossing a coin
- a patient's response to treatment (recover or not)

$$E(X) = \sum_{x=\{0,1\}} xP(X = x) = 0 \times (1 - p) + 1 \times p = p$$

$$\text{Var}(X) = E[(X - p)^2] = \sum_{x=\{0,1\}} (x - p)^2 P(X = x)$$

$$= (-p)^2 \times (1 - p) + (1 - p)^2 \times p = p(1 - p)$$

Binomial probability model

This is used when there are a set of Bernoulli trials which are independent of each other.

Consider a set $X_1, X_2, \ldots, X_n$ of n independent Bernoulli trials each with probability of success p and let the random variable Y be the total number of successes. Then

$$Y = X_1 + X_2 + \cdots + X_n$$

is said to follow a binomial distribution with parameters n and p, written $Y \sim B(n, p)$

The pmf of Y is

$$f(y) = P(Y = y) = \begin{cases} {}^nC_y p^y (1-p)^{n-y} & \text{for } y = 0, 1, \ldots, n \\ \\ 0 & \text{otherwise.} \end{cases}$$

$$E(Y) = np \quad \text{and} \quad \mathrm{Var}(Y) = np(1-p)$$

Poisson distribution

Let X denote the number of occasions an outcome is observed during a fixed period of time (or region of space) and suppose that

1. the numbers of occasions in any two disjoint intervals are independent
2. the probability of the outcome in a very short interval is proportional to the length of the interval
   - the probability of the outcome in any interval is a constant λ
3. the probability of two or more occurrences of the outcome in a very short interval is negligible.

then X is said to follow a Poisson distribution with parameter λ, written $X \sim P(\lambda)$.

The pmf of X is

$$f(x) = P(X = x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{for } x = 0, 1, 2, \ldots \\ \\ 0 & \text{otherwise.} \end{cases}$$

$$E(X) = \lambda \quad \text{and} \quad \mathrm{Var}(X) = \lambda$$

Note: $E(X) = \mathrm{Var}(X)$

Poisson approximation to binomial

Suppose X is binomial with parameters n and p where n is large and p is close to 0. It can be shown that

$$P(X = x) \approx \frac{e^{-np}(np)^x}{x!}$$

That is $X \sim P(\lambda = np)$ approximately.

There are many situations when an event is rare - there are a large number of samples n and the probability p of occurrence of the event is low

- failure of electronic components
- defects in manufacturing
- arrival of call for a particular number at a telephone exchange
- number of accidents in a factory

These situations are can be dealt with by using the Poisson distribution

Normal distribution

The pdf for a variable X with a normal distribution, parameters μ and σ^2, is

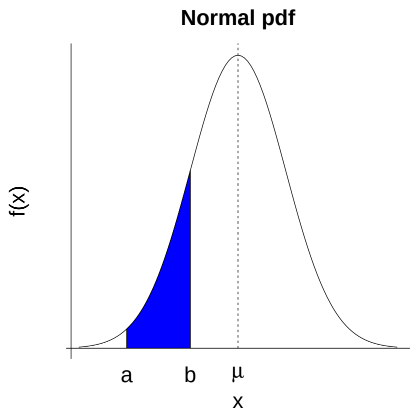
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

We write $X \sim N(\mu, \sigma^2)$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \mu$$

$$\mathrm{Var}(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \sigma^2$$

Normal pdf



$$P(a < X < b) = \int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

Standard normal distribution

Suppose $X \sim N(\mu, \sigma^2)$ and let

$$Z = \frac{X - \mu}{\sigma} .$$

Then Z has a standard normal distribution with pdf

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

We write $Z \sim N(0, 1)$

Integrals associated with normal distribution are difficult to solve in closed form, so usual to compute probabilities using the standard normal distribution:

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

Notation: We use z_α to denote the α percentile point of the standard normal distribution:

$$P(Z < z_\alpha) = \alpha. \text{ By symmetry } z_{1-\alpha} = -z_\alpha$$

Gamma distribution

A random variable X has a gamma distribution if the pdf is of the form

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0. \end{cases}$$

$$E(X) = \frac{\alpha}{\beta} \text{ and } \mathrm{Var}(X) = \frac{\alpha}{\beta^2}$$

Exponential distribution

If $\alpha = 1$ then X is said to have an exponential distribution with parameter β

Developed as a model for time between events in processes where events occur continuously and independently at constant rate

χ^2 distribution

Another particular type of gamma distribution is the chi-square (χ^2) distribution.

For any positive integer n, the gamma distribution with $\alpha = \frac{n}{2}$ and $\beta = \frac{1}{2}$ is called the χ^2 distribution with n degrees of freedom (df).

It follows from the previous slide that if $X \sim \chi^2_n$,

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0. \end{cases}$$

$$E(X) = n \text{ and } \mathrm{Var}(X) = 2n$$

Suppose that Z is standard normal, then the distribution of

$V = Z^2$ is chi-square on 1 df

$$V \sim \chi^2_1$$

The χ^2 distribution is closely related to the normal distribution

Student's t distribution

Another distribution closely related to the normal distribution is the t distribution.

Like the χ^2 distribution, the t distribution is widely applied in statistical inference.

Consider two independent random variables V and Z such that $V \sim \chi^2_n$ and $Z \sim N(0,1)$ and suppose

$$X = \frac{Z}{\sqrt{\frac{W}{n}}}$$

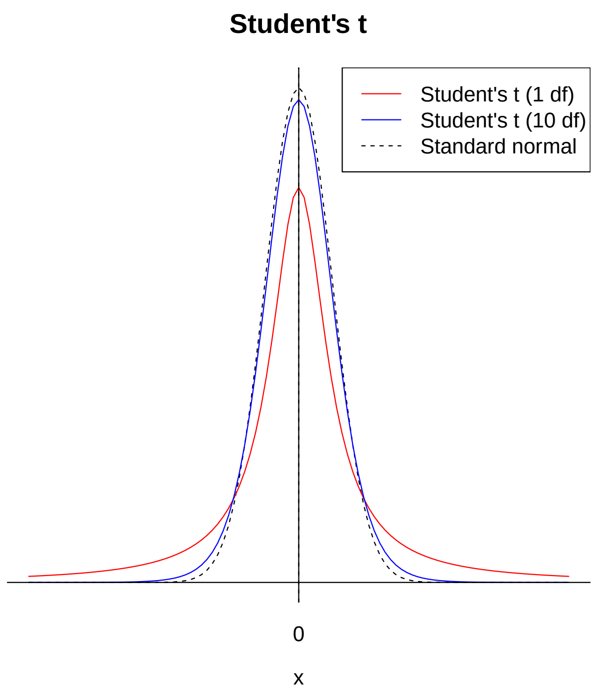
Then X follows a t distribution with n df: $X \sim t_n$.

The pdf is

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{(n\pi)^{\frac{1}{2}}\Gamma(\frac{n}{2})}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad \text{for } -\infty < x < \infty$$

$$E(X) = 0 \text{ for } n > 1 \quad \text{and} \quad \mathrm{Var}(X) = \frac{n}{n-2} \text{ for } n > 2$$

Student's t pdf



Expectation and variance of sum of random variables

Suppose that $X_1, X_2, \ldots, X_n$ are n random variables with respective means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$.

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

That is expectation of the sum of a set of random variables is the sum of the expectations.

$$\begin{aligned}\mathrm{Var}(X_1 + X_2 + \cdots + X_n) =& \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) + \\ & 2\mathrm{Cov}(X_1, X_2) + 2\mathrm{Cov}(X_1, X_3) + \cdots + 2\mathrm{Cov}(X_{n-1}, X_n)\end{aligned}$$

In particular, if $X_1, X_2, \ldots, X_n$ are independent then $\mathrm{Cov}(X_i, X_j) = 0, \quad i \neq j$ and

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Sum of iid random variables

Suppose that $X_1, X_2, \ldots, X_n$ are n independent identically distributed (iid) random variables with mean μ and variance σ^2 and let

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

.

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n\mu \text{ and hence } E(\bar{X}) = \mu \ .$$

Further,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\sigma^2$$

and hence

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$$

Note that these results hold if $X_1, X_2, \ldots, X_n$ is a randomly drawn sample of size n from some population.

Independent normal variables

Suppose we have two independent normal variables $X \sim N\left(\mu_X, \sigma_X^2\right)$ and $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$ that we wish to compare.

It can be shown that the sum

$$X + Y \sim N\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\right)$$

and the difference

$$X - Y \sim N\left(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2\right)$$

If a random sample $X_1, X_2, \dots, X_n$ of size n is *to be* taken from a normal population with mean μ and variance σ^2, then the sample mean $\bar{X}$ and the sample variance $S^2 = \sum(X_i - \bar{X})^2/(n-1)$ are independent and

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Central Limit Theorem

Simple version:

If a random sample $X_1, X_2, \dots, X_n$ of size n is to be taken from *any* population with mean μ and variance σ^2, the sample mean $\bar{X}$ is *approximately* normally distributed with mean μ and variance σ^2/n.

In other words

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1) \text{ approximately}$$

A consequence of the c.l.t is that the total of the samples $T_n = X_1 + X_2 + \cdots + X_n$ is also normally distributed:

$$T_n \sim N\left(n\mu, n\sigma^2\right)$$

Normal approximation to binomial

If X is binomial $B(n, p)$ then the distribution of X can be approximated by the normal distribution

$$X \sim N(np, npq) \text{ approximately}$$

where q = 1 - p.

The approximation is appropriate when n is large and p not close to zero or one.

A rule of thumb is np and nq greater than 5.

Recall that if n is large and p is close to zero so that np is not large, the binomial distribution is approximated by the Poisson

Normal approximation to Poisson

If X is Poisson $P(\lambda)$ then the distribution of X can be approximated by the normal distribution

$$X \sim N(\lambda, \lambda) \text{ approximately}$$

The approximation is appropriate when n is large and λ not too small.

A rule of thumb is λ greater than 10.

Estimation

Statistical inference uses data (sample) to draw conclusions about unknown features of a population.

Any summary calculated from a sample is called a statistic

Suppose that random sample $X_1, X_2, \ldots, X_n$ of size n is taken from a population with **unknown** mean μ and **unknown** variance σ^2.

Estimates of the parameters μ and σ^2 are

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ and } \hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

These estimates are more likely to be accurate if the sample is reasonably representative of the population.

A measure of the accuracy of $\bar{X}$ as an estimate of μ is provided by its standard error

$$s.e.(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} \text{ estimated by } \frac{s}{n}$$

Notice this is simply $\sqrt{\mathrm{Var}(\bar{X})}$

Confidence Interval for the mean

A confidence interval for a parameter is an interval estimate for value of the parameter.

It provides an interval for which we have a certain level of confidence contains the **true** value of the parameter

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a population with **unknown** mean μ and **known** variance σ^2.

A $(1-\alpha)100\%$ confidence interval for the mean μ is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

When variance σ^2 is **unknown** a $(1-\alpha)100\%$ confidence interval is

$$\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

Note: The normal approximation is appropriate when $n > 30$

Hypothesis testing

Concerned with using samples from a population to establish whether or not there is sufficient evidence to reject a hypothesis about some feature of the population

Null Hypothesis H_0: Usually a precise statement about a/several population parameter/s.

For example

- $\mu = 4mg$ - the population mean equals 4mg
- $\mu_A = \mu_B$ - no difference between the means of groups A and B.

Alternative Hypothesis H_1: Usually a statement of what we will accept if there is insufficient evidence to support H_0.

For example

- $\mu \neq 4mg$ or $\mu < 4mg$ or $\mu > 4mg$- the population mean is not/less than/more than 4mg
- $\mu_A \neq \mu_B$ - there is a difference between the means of group A and B.

Hypothesis test procedure

We test to see how much evidence there is to support the null hypothesis H_0.

1. Summarise evidence in the sample by calculating a value t for the test statistic T assuming H_0 is true
2. Establish *the likelihood* of values of T if H_0 is true
    - Statement of probability
    - Requires having a distribution of the test statistic T
3. Compute the probability ($p - value$) that a value at least as extreme as $T = t$ would be observed under H_0
4. Make conclusions regarding your hypotheses

Weighing up the evidence

- $0.05 < p - value < 0.10$
  - Weak/some evidence against H_0; not significant testing at the 5% level
- $0.01 < p - value < 0.05$
  - Statistically significant evidence against H_0
- $0.001 < p - value < 0.01$
  - Strong statistically significant evidence against H_0
- $p - value < 0.001$
  - Very strong statistically significant evidence against H_0

One Sample t-test

Example:
Suppose that a random sample $x_1, x_2, \ldots, x_n$ of size n is observed from a population and we wish to test whether the population mean is some value μ_0.

Our hypotheses are: $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$

and we will have evidence against H_0 if $\bar{x} - \mu$ is large.

Using results from distribution theory, we know the distribution of

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

so this can be used as the test statistic.

Under $H_0 : \mu = \mu_0$, the computed value t of T is

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} \quad \text{and} \quad p - value = P(T > |t|)$$

is found by referring to tabulated values of the student's distribution with $n - 1$ df or using a software package.

Significance level

Alternatively, for testing at $\alpha\%$ significance level, the critical values are percentile points of the t distribution $t_{n-1,\alpha/2}$ and $t_{n-1,1-\alpha/2}$, and we reject H_0 if the value of the test statistic is more extreme than the critical values.

By symmetry of student's distribution $-t_{n-1,\alpha/2} = t_{n-1,1-\alpha/2}$ and so we reject H_0 if $|t| > t_{n-1,1-\alpha/2}$

This approach

- avoids computing $p - value$ but
- we need to decide on significance level before doing the test

Comparing two populations

Suppose we wish to compare the means μ_1 and μ_2 of populations 1 and 2 using two independent random samples $x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}$ and $x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}$.

$$\text{Our null hypothesis is :} \quad H_0 : \mu_1 = \mu_2$$

while the alternative can be the means are not equal ($H_1 : \mu_1 \neq \mu_2$) or one mean is less than the other, e.g. $H_1 : \mu_1 < \mu_2$

Assuming a common variance σ^2 for the two groups, the pooled estimator of this variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_1 - 1)}$$

where s_1^2 and s_2^2 are the respective variance estimates of the two populations.

Two-sample t test

Using distribution theory results for two independent normal variables, the two-sample t test statistic is

$$t = \frac{\sqrt{n}(\bar{x}_1 - \bar{x}_2)}{s\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

which follows the student's distribution with $n_1 + n_2 - 2$ df .

Matched pairs t test

In situations where we have n random pairs of dependent observations (x_{1i}, x_{2i}), $i = 1, \dots, n$, that we wish to compare, we need to account for the dependency so we cannot use the above test.

Instead, we take the difference $d_i = x_{1i} - x_{2i}$ and conduct a **one-sample t test** on this difference.

This is called the matched pairs t test.

Linear models

Suppose in a study we measure a response variable Y_i, $i = 1, \dots, n$ and for each Y_i we also observe the vector $\mathbf{x_i} = \{x_{i,1}, \dots, x_{i,k}\}$ of k explanatory variables, each measured without error.

The explanatory variables may be continuous or discrete covariates or they may be dummy variables indicating levels of a study (experimental) factor.

The general linear model expresses Y_i as a linear combination of $\mathbf{x_i}$:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \epsilon_i, \quad i = 1, \dots, n$$

where β_j, $j = 0, \dots, k$ are unknown parameters and ϵ_i is random noise.

Examples

1. Simple linear regression: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$
   - Straight line relationship between Y and x
   - β_0 is y-intercept, β_1 is slope of the line
2. One way ANOVA:
   $Y_{i,j} = \mu + \tau_j + \epsilon_{i,j}, \quad i = 1, \ldots, n_j; j = 1, \ldots, t$
   - $Y_{i,j}$ is i^{th} response in j^{th} group
   - μ is overall mean and τ_j is j^{th} "treatment" effect - deviation of mean of j^{th} group from overall mean μ

Matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- $\mathbf{Y}$ is $n \times 1$ column vector of responses,
- $\mathbf{X}$ is $n \times (k+1)$ *d*esign matrix,
- $\boldsymbol{\beta}$ is $(k+1) \times 1$ column vector of parameters and
- $\boldsymbol{\epsilon}$ is $n \times 1$ column vector

The (least squares) estimates of $\boldsymbol{\beta}$ are

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

and the *fitted* model is

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$