QM1 - Exam Notes - Summary Quantitative Methods 1

Quantitative Methods 1 (University of Melbourne)

Quantitative Methods 1

Exam Notes

Data Visualisation

Introduction – Data

Interval Data
Measurements or observations where the difference between two values is meaningful.

Categorical Data
Measurements or observations that fall into a set of mutually exclusive categories.

Ordinal Data
Measurements or observations that possess a natural ordering or ranking but the difference between values is not informative.

Introduction – Statistics & Numbers

A statistic is a known rule for combining data. Another way of saying this is that a statistic is a function of the data.

- A set of data needs to have a functional form

1. $f(X_1, X_2, \ldots, X_N)$ is not a statistic because $f(.)$ is unspecified.
2. $g(X_1, X_2, \ldots, X_N) = \omega_1 X_1 + \omega_2 X_2 + \cdots + \omega_N X_N$ is not a statistic unless we know $\omega_1, \omega_2, \ldots, \omega_N$.
3. $h(X_1, X_2, \ldots, X_N) = \overline{X} = N^{-1}(X_1 + X_2 + \cdots + X_N)$ is a statistic because we know N (the number of observations).
4. Note that $h(.)$ is a special case of $g(.)$, where the weights are known:
$$\omega_1 = \omega_2 = \cdots = \omega_N = \frac{1}{N}$$
5. The original set of numbers $\{X_1, X_2, \ldots, X_N\}$ is also a statistic.

- X-bar means average
    - Is also an example of data reduction
    - A full data set contains more information than x-bar
    - We have to trade up between a loss of data and something that we can handle more easily

Sigma/Summation Notation

- X_i is called the argument of the sum.
- The symbol i is the index of summation. Common symbols used to represent the index are: i, j, k, l, m, n, s, t. (The index doesn't have to be represented by 'i'.)
- L and U are the lower and upper limits of the sum, respectively. They are typically integers.
- $\sum$ is the summation operator.
- $X_1, X_2, \ldots, X_N$ are symbols representing the numbers to be added. The role of the index is to distinguish between the different numbers to be summed.

Basic Rules of Summation

- The summation operator is **linear**
  - $\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$
  - $\sum_{i=1}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$
- Constant terms do not change with the index of the sum
  - $\sum_{i=1}^{n} c = cn$
- If we are summing along multiple indexes, the order of summation does not matter
  - $\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j = \sum_{j=1}^{m}\sum_{i=1}^{n} x_i y_j$

- If lower limit is higher than upper limit then the answer must be 0

Histograms

- Class intervals = bins
- Sturge's formula – states that given n observations, the number of class intervals K should be:

$$K = 1 + 3.3\log_{10}(n)$$

- Similarly, class width can be determined using the following formula:

$$\text{Width} = \frac{\text{Largest Value} - \text{Smallest Value}}{K}$$

- Histograms can be scaled in terms of: absolute or relative frequencies
  - Number of observations of representative percentage
- Histogram is a graphic depiction of the frequency distribution of data
- Shape
  - Location – in what range does data lie
  - Spread – is my data spread evenly across range or concentrated
  - Skew – are there more large or small observations (pos or neg)
  - Peakedness (Kurtosis) – Are there proportionately more extreme valued observations? Or vice versa.

Descriptive Statistics

Measures of Central Tendency

- Poplar measure of central tendency = sample mean

Sample Mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

- Lower case letters = sample objects (stats and data)
- Upper case letter = population objects (random variables)
- Sample mean becomes unreliable if there is an outlier

Median

- In situations where outliers are present we can resort to the use of the median
- Median = value that partitions the ordered set into two equal parts
    - Observation that lies in the middle

Deviations

Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Sample variance is the sum of squares, to rescale back into the same units as our data we need to take the square root:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{\frac{1}{2}}$$

- For data sets that exhibit a bell shaped histogram, we can use the empirical rule:
    - 68% fall within 1 SD of mean
    - 95% fall within 2 SD of mean
    - 99.7% fall within 3 SD of mean
- **Chebyshev's Theorem** which states that the proportion of observation that lie within k SD of the means is at least

$$1 - \frac{1}{k^2} \qquad \text{for} \quad k > 1$$

The IQR

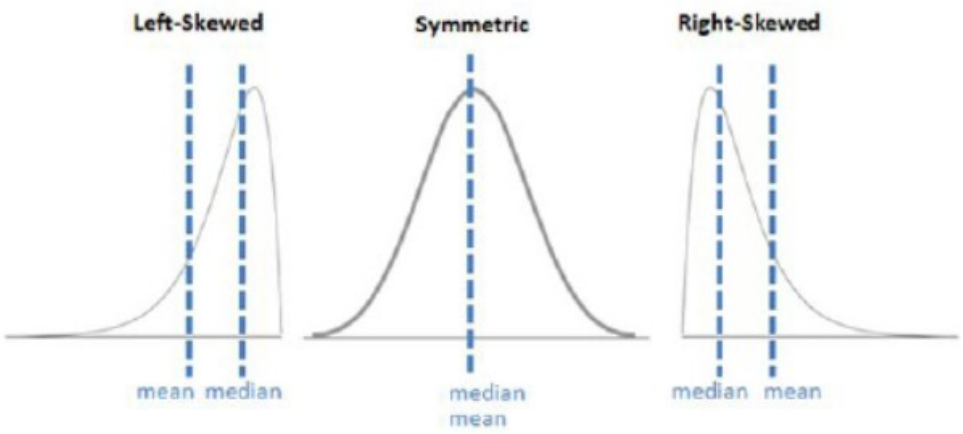
Sample variance is sentence to the presence of outliers, so the IQR can be used to show a measure of spread that uses relative position of values.

Box and Whisker Plots

- Whiskers extend to the minimum and maximum values that are not outliers
- Outliers are more or less than 1.5*(IQR)

Measures of Skew

- Positive skew means that the observations in the data set are concentrated below the mean and the right tail is longer.



To compute a numerical measure of skewness, we use the following formula:

$$g = \frac{n^2}{(n-1)(n-2)}\frac{m_3}{s^3}$$

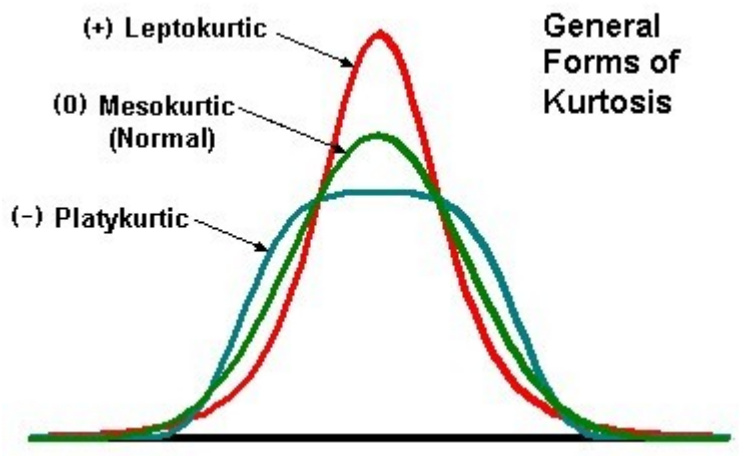
Where

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- The cubing effect amplifies the skew in the formula
- A negative result indicates a negative (left) skew
- A positive result indicates a positive (right) skew

Measures of Kurtosis



We can compute a numerical measure of kurtosis via the following formula

$$k = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

Scatterplots

Simple way to get an immediate impression of the relationship between two variables

The Sample Covariance

- Describes the degree to which two variables are related

Given n observations on two variables x and y:

$$cov(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

- If x and y are positively related, then small observations of x are associated with small observations of y and large with large
- If negatively related, then small observations of x will be associated with large observations of y and vice versa

Coefficient of Correlation

To remove scale from covariance we divide by the standard deviations of the two variables. Becomes the coefficient of correlation.

$$r = \frac{s_{xy}}{s_x s_y}$$

Simple Linear Regression

Linear Relationships

If two variables are perfectly linearly related, then we can represent their relationship using a linear function. Where we have y = dependent variables and x = explanatory variable

- We can determine a line of best fit:

$$y_i = b_0 + b_1 x_i + e_i$$

Line of Best Fit

- B_0 = intercept
- B_1 = slope
- S_{xy} = sample covariance of x and y

So, given a set of n observations on variables x and y, we obtain a line of best fit by:

1. First, computing $\overline{x}$ and $\overline{y}$
2. Then, computing s_{xy} and s_x^2
3. Finally, computing $b_1 = \frac{s_{xy}}{s_x^2}$ and $b_0 = \overline{y} - b_1\overline{x}$

Interpreting the Slope Coefficient - B_1

- There exists a relationship between the correlation coefficient and slope coefficient

  r = $b_1 \frac{s_x}{s_y}$

The Degree of Linear Relationship

- Y hat = fitted value (value that lies on the linear equation)

Total sum of squares (another representation of variation):

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \\ SST &= SSE + SSR \end{aligned}$$

- SSR = sum of squares for regression (amount of variation in y that is explained by x)
- SSE = sum of squares of errors (amount of variation in y that is unexplained by y)

Coefficient of determination:

$$R^2 = \frac{\text{Explained variation in } y}{\text{Total variation in } y} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Probability

Sets and Events

Population as experiments:

- Population can be considered as terms drawn from a random experiment → each experiment is a trial → the outcomes of trials are mutually exclusive

Sample Space/Set Operations

We will typically denote the sample space by the symbol $\mathscr{S}$.

Equality If $\mathscr{A} = \mathscr{B}$ then the two sets are comprised of exactly the same elements.

Union The union of two sets $\mathscr{A}$ and $\mathscr{B}$, written $\mathscr{A} \cup \mathscr{B}$, is the set of distinct elements contains in either or both sets.

Example: $\{a, b, c\} \cup \{c, d, e\} = \{a, b, c, d, e\}$.
Observe that c is not included twice.

Intersection The intersection of two sets $\mathscr{A}$ and $\mathscr{B}$, written $\mathscr{A} \cap \mathscr{B}$, is the set of distinct elements contained both sets.

Example: $\{a, b, c\} \cap \{c, d, e\} = \{c\}$.
So, if c is the outcome of a trial then both events $\mathscr{A}$ and $\mathscr{B}$ are said to have occurred.

Difference The difference between two sets $\mathscr{A}$ and $\mathscr{B}$, written $\mathscr{A} \setminus \mathscr{B}$, is the collection of elements in $\mathscr{A}$ that are not also in $\mathscr{B}$.

Example: $\{a, b, c\} \setminus \{c, d, e\} = \{a, b\}$.

Complement The complement of a set $\mathscr{A}$, written $\overline{\mathscr{A}}$, is those elements of the sample space that are not in $\mathscr{A}$. That is, $\overline{\mathscr{A}} = \mathscr{S} \setminus \mathscr{A}$.

Example: If $\mathscr{S} = \{a, b, c, d, e\}$ and $\mathscr{A} = \{a, b, c\}$ then $\overline{\mathscr{A}} = \{d, e\}$.

Notes:

1. If the event $\overline{\mathscr{A}}$ has occurred then $\mathscr{A}$ has not occurred.
2. $\mathscr{A} \setminus \mathscr{B} = \mathscr{A} \cap \overline{\mathscr{B}}$ and $\overline{\mathscr{S}} = \mathscr{S} \setminus \mathscr{S} = \emptyset$.

Subset If $\mathscr{A}$ is a subset of $\mathscr{B}$, written $\mathscr{A} \subset \mathscr{B}$, then every element of $\mathscr{A}$ is contained in $\mathscr{B}$.

Example: If $\mathscr{S} = \{a, b, c, d, e\}$ and $\mathscr{A} = \{a, b, c\}$ then $\mathscr{A} \subset \mathscr{S}$.

Notes:

1. If $\mathscr{A} = \mathscr{B}$ then $\mathscr{A} \subset \mathscr{B}$ and $\mathscr{B} \subset \mathscr{A}$.
2. People sometimes use the notation $\mathscr{A} \subseteq \mathscr{B}$ to allow the outcome $\mathscr{A} = \mathscr{B}$, reserving $\mathscr{A} \subset \mathscr{B}$ for 'strict' or 'proper' subsets where $\mathscr{A}$ is a smaller set than is $\mathscr{B}$. We won't bother.
3. We can also define a superset along the following lines: If $\mathscr{A}$ is a superset of $\mathscr{B}$, written $\mathscr{A} \supset \mathscr{B}$, then every element of $\mathscr{B}$ is contained in $\mathscr{A}$.

   Example: In the above example, $\mathscr{S} \supset \mathscr{A}$.

Sets obey the following rules:

1. Commutative Laws
   $\mathscr{A} \cup \mathscr{B} = \mathscr{B} \cup \mathscr{A}$ and $\mathscr{A} \cap \mathscr{B} = \mathscr{B} \cap \mathscr{A}$.
2. Associative Laws
   $\mathscr{A} \cup (\mathscr{B} \cup \mathscr{C}) = (\mathscr{A} \cup \mathscr{B}) \cup \mathscr{C}$ and
   $\mathscr{A} \cap (\mathscr{B} \cap \mathscr{C}) = (\mathscr{A} \cap \mathscr{B}) \cap \mathscr{C}$.
3. Distributive Laws
   $\mathscr{A} \cup (\mathscr{B} \cap \mathscr{C}) = (\mathscr{A} \cup \mathscr{B}) \cap (\mathscr{A} \cup \mathscr{C})$, and
   $\mathscr{A} \cap (\mathscr{B} \cup \mathscr{C}) = (\mathscr{A} \cap \mathscr{B}) \cup (\mathscr{A} \cap \mathscr{C})$.
4. $(\overline{\mathscr{A}}) = \mathscr{A}$ The complement of the complement of $\mathscr{A}$ is $\mathscr{A}$.
5. Sundry rules for unions:
   $\mathscr{A} \cap \mathscr{S} = \mathscr{A}, \mathscr{A} \cup \mathscr{S} = \mathscr{S}, \mathscr{A} \cap \emptyset = \emptyset$, and $\mathscr{A} \cup \emptyset = \mathscr{A}$.
6. Sundry rules for intersections:
   $\mathscr{A} \cap \overline{\mathscr{A}} = \emptyset, \mathscr{A} \cup \overline{\mathscr{A}} = \mathscr{S}, \mathscr{A} \cap \mathscr{A} = \mathscr{A}$, and $\mathscr{A} \cup \mathscr{A} = \mathscr{A}$.
7. De Morgan's Laws:
   $\overline{\mathscr{A} \cup \mathscr{B}} = \overline{\mathscr{A}} \cap \overline{\mathscr{B}}$ and $\overline{\mathscr{A} \cap \mathscr{B}} = \overline{\mathscr{A}} \cup \overline{\mathscr{B}}$.
8. $\mathscr{A} \setminus \mathscr{B} = \mathscr{A} \cap \overline{\mathscr{B}}$

Axioms of Probability

$$\mathrm{P}(\mathscr{A}) = \lim_{n \to \infty} \frac{n_{\mathscr{A}}}{n}.$$

Note: If n finite then $P(\mathscr{A})$ approximated by $n_{\mathscr{A}}/n$.

The Addition Rule: For any two events $\mathscr{A}$ and $\mathscr{B}$

$$\mathrm{P}(\mathscr{A} \cup \mathscr{B}) = \mathrm{P}(\mathscr{A}) + \mathrm{P}(\mathscr{B}) - \mathrm{P}(\mathscr{A} \cap \mathscr{B}).$$

A random experiment is comprised of:

- A sample space which we represent as a set of the sample space of all the possible outcomes of the experiment
- A set of numbers called probabilities that are assigned to the outcomes our events

Marginal Probability

The marginal probability of any single event, $\mathscr{A}$ say, is the probability of that event $\mathscr{A}$ occurring, where

$$\mathrm{P}(\mathscr{A}) = \sum_{O_i \in \mathscr{A}} \mathrm{P}(O_i),$$

with the $O_i \in \mathscr{A}$ being distinct outcomes of the experiment that belong to $\mathscr{A}$

Joint Probability

The joint probability of a set of events is the probability that all events in the set occur simultaneously.

Example: Take our fair six-sided die experiment and again, let $\mathscr{A} = \{2,4,6\}$. Now let's define another event $\mathscr{B} = \{4,5,6\}$. Now from our knowledge of set operations, we know that $\mathscr{A} \cap \mathscr{B} = \{4,6\}$ therefore

$$P(\mathscr{A} \cap \mathscr{B}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Conditional Probability

$$\mathrm{P}(\mathscr{A}|\mathscr{B}) = \frac{\mathrm{P}(\mathscr{A} \cap \mathscr{B})}{\mathrm{P}(\mathscr{B})}, \quad \text{provided that } \mathrm{P}(\mathscr{B}) \neq 0.$$

Independent

Two events are said to be independent if

$$\mathrm{P}(\mathscr{A} \cap \mathscr{B}) = \mathrm{P}(\mathscr{A}) \times \mathrm{P}(\mathscr{B}).$$

This implies that $\mathrm{P}(\mathscr{A}|\mathscr{B}) = \mathrm{P}(\mathscr{A})$

- We can check independence by:
  - Seeing is joint probability is equal to the product of their marginal
  - Conditional probability is equal to the marginal

Mutually Exclusive Events

- Mutually exclusive events are dependent
  - If one has occurred than the other hasn't and such knowledge is evidence of dependence

Random Variables

- Can have outcomes that are numerical or categorical
- Discrete random variables can take a countable number of distinct values
  - E.g. – number of applicants to a university

Probability mass function shows all the values that the random variable can possible take and their probabilities, must equal 1

Cumulative distribution function gives the probability of being less than or equal to some value c

Notation: $F_X(c) = P(X \leq c)$, where c is any real number.

Cumulative distribution functions satisfy the following properties:

1. $F_X(-\infty) = 0$
2. $F_X(\infty) = 1$, and
3. $0 \leq F_X(c) \leq 1$ for all $-\infty < c < \infty$.

Expected Values

The expectation (mean) of a random variable is the weighted sum of all the possible outcomes in which the weights are the associated probabilities.

- Gives us a sense of location

$$\mu = E[X] = \sum_{i=1}^{k} x_i P(X = x_i)$$

Variance

- Shows us how spread out or dispersed our observations are

$$\sigma^2 = V[X] = E[(X-\mu)^2] \sum_{i=1}^{k} (x_i - \mu)^2 P(X = x_i)$$

Or

$$\begin{aligned} \sigma^2 &= E[(X-\mu)^2] \\ &= E[X^2 - 2X\mu + \mu] \end{aligned}$$

Discrete Bivariate Distributions

If we add row and column sums to the previous table then we obtain marginal probability distributions (which is where marginal probabilities they get their name from).

$X \setminus Y$	y_1	y_2	y_3	$P(X)$
x_1	$P(x_1,y_1)$	$P(x_1,y_2)$	$P(x_1,y_3)$	$P(x_1)$
x_2	$P(x_2,y_1)$	$P(x_2,y_2)$	$P(x_2,y_3)$	$P(x_2)$
$P(Y)$	$P(y_1)$	$P(y_2)$	$P(y_3)$	1

Specifically,

$$P(x_1) = P(x_1,y_1) + P(x_1,y_2) + P(x_1,y_3)$$
$$P(x_2) = P(x_2,y_1) + P(x_2,y_2) + P(x_2,y_3)$$

and

$$P(y_1) = P(x_1,y_1) + P(x_2,y_1)$$
$$P(y_2) = P(x_1,y_2) + P(x_2,y_2)$$
$$P(y_3) = P(x_1,y_3) + P(x_2,y_3).$$

Each outcome is a joint event (intersection)

Conditional Probabilities

$$P(Y = y_1|X = x_1) = \frac{P(x_1,y_1)}{P(x_1)}, \quad P(Y = y_2|X = x_1) = \frac{P(x_1,y_2)}{P(x_1)},$$
$$P(Y = y_3|X = x_1) = \frac{P(x_1,y_3)}{P(x_1)}$$

Independence

$$P(x_i,y_j) = P(x_i) \times P(y_j)$$

Measures of Association

- If two random variables are independent, then their correlation is zero

$$\sigma_{XY} = COV[X,Y] = \sum_{i=1}^{k}\sum_{j=1}^{l}(x_i - \mu_x)(y_j - \mu_y)P(X = x_i \cap Y = y - j)$$

Bernoulli & Binomial Distributions

The Bernoulli Distribution

A random distribution that has two characteristics:

1. Two possible outcomes, Success or Failure
2. The probability of success is p and the probability of failure is q = 1-p

A Bernoulli random variable takes the value 1 in the event of a success and 0 in the event of a failure. It's probability mass function looks like this

$$B = \begin{cases} 1 & \text{with probability} \quad p \\ 0 & \text{with probability} \quad 1-p=q \end{cases}$$

With mean $E[B] = p$ and variance $\sigma_B^2 = p(1-p)$

The Binomial Distribution

- A sequence of Bernoulli trials will occur with probability:

  $p^x(1-p)^{n-x}$

- The number of sequences can be calculated through the combinations formula:

$$C_x^n = \frac{n!}{x!(n-x)!}$$

Therefore,

$$P(X=x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

- E(X) = np
- V(X) = np(1-p) S
- Skewness of a binomial =
  If p constant and make n large we get more symmetric and less kurtosis

$$\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}$$

- Kurtosis of a binomial =

$$\gamma_2 = \frac{1-6p(1-p)}{np(1-p)}$$

Continuous Random Variables

- A random variable is said to be discrete if the set of all its possible values is countable
- A set is countable if:
  - It is finite; OR
  - It is infinite but shares the same size as natural numbers
- For our purposes, it is sufficient to characterise a random variable as continuous if it can take any possible value along some continuum. (e.g. [−1,1], (∞,0], (−∞,∞))

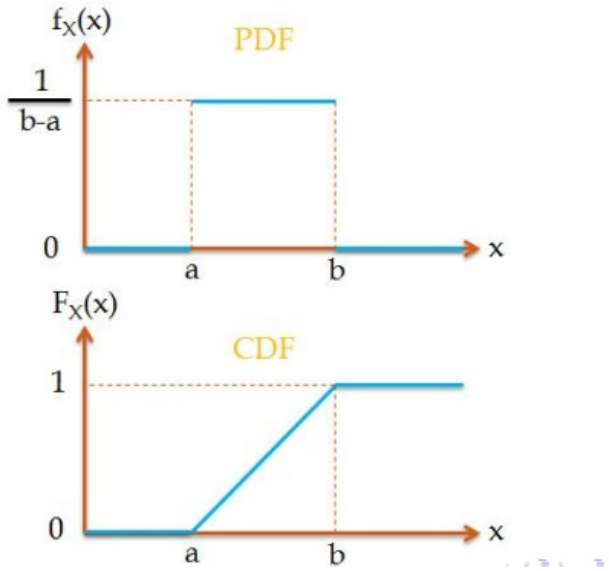
Probability for Continuous Random Variables

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

-

Also, if F_X is the cumulative distribution function of X, then

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



Mean and Variance

In the case of a continuous random variable X, the expected value is an integral

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

Similarly, the variance of a continuous random variable X is defined by

$$\sigma_X^2 = V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)dx$$

Uniform Distribution

Uniform distribution is where:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- Visually it appears in the shape of a rectangle
- The mean, E(X) = (b+a)/2
- Variance = E(X^2) – E(X)^2
    - E(X^2) = (B^3-A^3)/3(b-a)
- Hence, variance = (b-a)^2/12

Normal Distribution

A continuous random variable X is said to be **Normal** or possess a **Normal Distribution**, if it has the following density function

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

The Poisson Distribution

- The Binomial distribution is useful for when we want to represent an underlying phenomenon that generates a set of independent observations that arise from one of two outcomes
- Not all phenomena can be represented this way. Some are better described as number of occurrence arising in a given internal of time/space
    - Number of calls in an hour
    - Potholes in a given stretch of road

Let λ be the average number of occurrences (or events) in a given interval in time and space, then the probability distribution of a Poisson random variable X is given by

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

Deriving the Poisson Distribution

- Poisson random variable is described as the number of occurrences arising in a given interval

So what we want to do is rewrite the distribution function in terms of λ and take the limit $n \to \infty$. Using the fact that $p = \frac{\lambda}{n}$, we write

$$\lim_{n\to\infty} P(X = x) = \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

As a first step, let's factor out all the terms that do not depend on n

$$\left(\frac{\lambda^x}{x!}\right) \lim_{n\to\infty} \frac{n!}{(n-x)!} \frac{1}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

Now let's rewrite this in the following way

$$\left(\frac{\lambda^x}{x!}\right) \lim_{n\to\infty} \frac{n!}{(n-x)!} \frac{1}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-x}$$

1. $\lim_{n\to\infty} \frac{n!}{(n-x)!}\left(\frac{1}{n^x}\right) = 1$
2. $\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$
3. $\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{-x} = 1$

Therefore

$$\lim_{n\to\infty} P(X=x) = \left(\frac{\lambda^x}{x!}\right) \lim_{n\to\infty} \frac{n!}{(n-x)!}\frac{1}{n^x}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-x}$$

Simply collapses to our desired result!

$$\lim_{n\to\infty} P(X=x) = \left(\frac{\lambda^x}{x!}\right) e^{-\lambda} = \left(\frac{e^{-\lambda}\lambda^x}{x!}\right)$$

Moments of the Poisson Distribution

$$\begin{aligned} E[X] &= \lambda \\ V[X] &= \lambda \\ \gamma_1 &= \frac{1}{\sqrt{\lambda}} \\ \gamma_2 &= \frac{1}{\lambda} \end{aligned}$$

Population & Sample Moments

	Sample	Population
Mean	$\frac{1}{n}\sum_{i=1}^n x_i = \overline{x}$	$E[X] = \mu$
Variance	$\frac{1}{n-1}\sum_{i=1}^n (x_i-\overline{x})^2 = s^2$	$E[(X-\mu)^2] = \sigma^2$
Skewness	$\frac{n^2}{(n-1)(n-2)}\frac{m_3}{s^3} = g$	$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \gamma_1$
Kurtosis (Ex)	$\frac{m_4}{m_2^2} - 3 = k$	$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3 = \gamma_2$

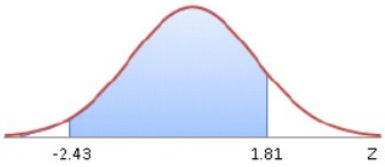
The population moments describe the shape of probability distribution functions while the sample moments describe the shape of histograms!

Approximating a Binomial

Shape of a Normal

Mean	μ
Variance	σ^2
Skewness	0
Kurtosis	3
Excess Kurtosis	0

Suppose we wanted to compute the following area:



This is simply

$$\begin{aligned} P(-2.43 \leq Z \leq 1.81) &= P(-\infty \leq Z \leq 1.81) \\ &\quad -P(-\infty \leq Z \leq -2.43) \\ &= 0.9649 - 0.0075 \\ &= 0.9574 \end{aligned}$$

Using the Standard Normal

Using what we know about Z, how can we use this to compute probabilities with the mean and variance

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

Approximating a Binomial Using a Normal

Compute mean and variance

Mean =np

Variance = np(1-p)

Then, use the standard normal formula for calculating interval (see above)

Approximating a Poison Using a Normal

Find mean and variance for the Poisson and input to standard normal interval formula.

Inference

1. Let X be a random variable with a mean (u) and variance (o2) that represents a data generating process. U and o2 are unknown quantities (not random)
2. Now suppose we have observed a sample of n observations generated by X

The Sample Mean

One very important property of the sample mean is that it will converge to the population mean as n gets closer to infinity

Law of Large Numbers:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow \mu \qquad \text{as} \quad n \rightarrow \infty$$

- The sample mean is known as a point estimate of the population mean (u)
- If we take numerous sample means from a population they well tend toward the true population mean
    - Larger the sample size, the closer to true population mean

$\overline{X}$ is known as an **estimator** while $\overline{x}$ is known as an **estimate**.

- The distribution of an estimator is known as a sampling distribution

Distribution of X-BAR

CENTRAL LIMIT THEOREM:

Let $\{X_1, X_2, \ldots, X_n\}$ be a sequence of independently and identically distributed random variables with $E[X_i] = \mu$ and $V[X_i] = \sigma^2$ for $i = 1, \ldots, n$. Then, as $n \rightarrow \infty$:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \overrightarrow{\text{distribution}} \qquad N(\mu, \frac{\sigma^2}{n})$$

- As n becomes infinitely large, the sampling distribution X-BAR will converge to a normal with mean (u) and variance (o2/n)
- Large is generally sample sizes above 30

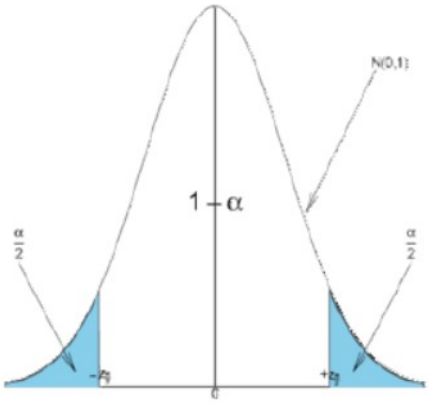
Therefore:

- If X_i is distributed as a *Binomial* with parameters m and p, then $\overline{X}$ will be approximately distributed as $N\left(mp, \frac{mp(1-p)}{n}\right)$
- If X_i is distributed as a *Poisson* with parameter λ then $\overline{X}$ will be approximately distributed as $N\left(\lambda, \frac{\lambda}{n}\right)$
- If X_i is distributed as a *Uniform* with parameters $[a, b]$, then $\overline{X}$ will be approximately distributed as $N\left(\frac{a+b}{2}, \frac{(b-a)^2}{12n}\right)$

Confidence Intervals

- We don't know how close or far away the sample mean is from the population mean
- We can construct an interval around the point estimator that has some likelihood of containing the unknown mean. This is known as an **interval estimator**.

- Our goal is to define an interval estimator that has some level of confidence which we shall express as a percentage (1-α) %
- if we were to observe k samples and compute a (1- α) % confidence interval for each sample, then the proportion of samples that will produce an interval estimate that contains the unknown mean m will be (1-α) % as 'k' tends toward infinity

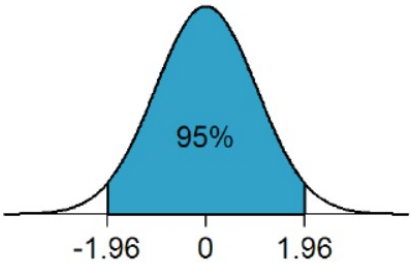


Confidence Interval Calculation:

$$\overline{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- X-bar corresponds to the z value of the tail probability of α/2

The values of z that corresponds to a tail probability of $\frac{\alpha}{2} = 0.025$ is ± 1.96



- The higher the level of confidence, the wider the interval
- The greater the variance of the sampling distribution, the wider the interval

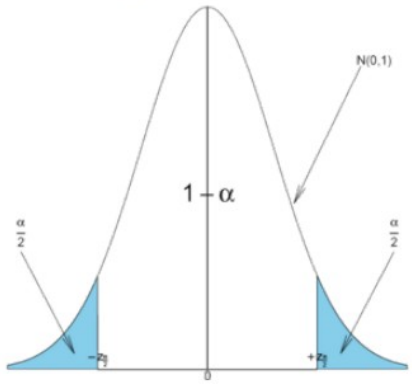
Performing Inference

If we define an interval as $\overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ there is a 95% chance of such an occurrence.

Another way of saying this is that I would be 95% confident that this interval will contain the unknown mean u.

General Rule for performing inference with confidence intervals:

In general, to construct an interval that has a $(1-\alpha\%)$ confidence level, we need to first find the value of the standard normal that corresponds to a tail probability of $\frac{\alpha}{2}$. This we shall denote by $z_{\frac{\alpha}{2}}$



Then a $(1-\alpha\%)$ confidence interval is given by $\overline{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Inference with an Unknown Variance

We need to find the distribution of the following object:

$$\frac{\overline{X}-\mu}{S/\sqrt{n}}$$

Instead of working with σ^2 we work with the sample variance.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

To determine the remaining parts of this object, we can define the following:

$$\begin{aligned} \overline{X} &= \frac{1}{n}(X_1+X_2+\cdots+X_n) \\ S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \\ T &= \frac{\overline{X}-\mu}{S/\sqrt{n}} \end{aligned}$$

<u>Defining Object T (student t)</u>

- V = n-1
- Construction of Student t relies on the fact that the random variable X is normal

$$t_j = \frac{\overline{x}_j - \mu}{s_j/\sqrt{n}}$$

This means that a $(1-\alpha\%)$ confidence interval can be constructed as

$$\overline{X} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$$

Where $t_{\frac{\alpha}{2},n-1}$ is the value of a Student t random variable with $n-1$ degrees of freedom that corresponds to a tail probability of $\frac{\alpha}{2}$

The values of the student t that correspond to a tail probability of $\frac{\alpha}{2}$ can be read off a t-table:

Degrees of freedom	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
⋮	⋮	⋮	⋮	⋮	⋮

- For large sample, n>30 we can use the z value to construct intervals instead of t

$$\overline{x} \pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$$

Estimating a Population Proportion '$\hat{p}$'

For a sample size of n, let Y be the number of observations that are of interest to the analysis (e.g. out of n individuals, Y is the number of liberal voters)

The point **estimator** of the population proportion p is the sample proportion (*this is a random variable!*)

$$\hat{p} = \frac{Y}{n}$$

Now recall that if Y is binomial, then $E[Y] = np$ and $V[Y] = np(1-p)$. Therefore

$$\begin{aligned} E[\hat{p}] &= E\left[\frac{Y}{n}\right] = p \\ V[\hat{p}] &= V\left[\frac{Y}{n}\right] = \frac{1}{n^2}V[Y] = \frac{p(1-p)}{n} \end{aligned}$$

Distribution of '$\hat{p}$'

Therefore, for a finite n, we can say that a sample proportion $\hat{p}$ is approximately distributed as a Normal with mean p and variance $\frac{p(1-p)}{n}$

Sampling Distribution of '$\hat{p}$'

Now remember, p is a population parameter. It is unknown! Therefore the sampling distribution that we work with in practice is

$$\hat{p} \sim N\left(p, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

Or equivalently,

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$$

Confidence Interval for '$\hat{p}$'

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Hypothesis Testing

Specifying a Hypothesis

As a first step, we are going to concentrate our attention on hypothesis about the unknown population mean μ

We are required to think about the population mean in one of the following ways:

Is the population mean μ greater/less/different than some value k?

- These questions are known as research hypotheses

Two Tailed

"Is the population mean different than some value k?"

We need to imagine two states:

- Where μ = k (Null hypothesis = H_0: μ = k)
- Where $\mu \neq$ k (Alternative hypothesis = H_A: $\mu \neq$ k)

How do we decide between these two states? We need:

1. Some data

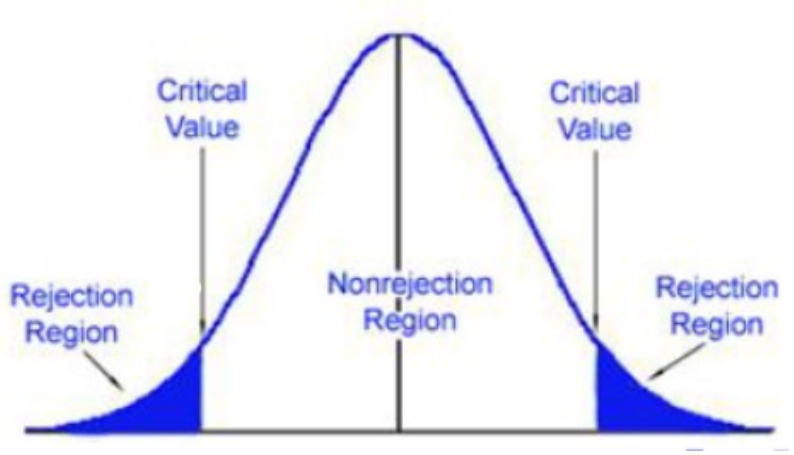
2. A decision rule that tells u how to use the data to decide between null and alternative (specifies the conditions under which we would reject the null in favour of alternative)

For the data we can use sample mean (x-bar).

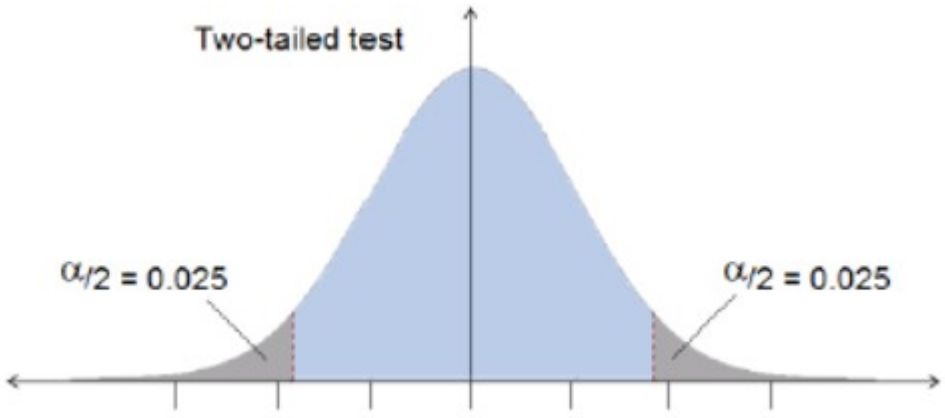
If we assume that the null $H_0 : \mu = k$ were true, then $\overline{X} \sim N\left(k, \frac{\sigma^2}{n}\right)$. For now, let's assume that the variance of the parent distribution σ^2 is known.

- We can reject the null hypothesis μ = k if the observed sample mean (x-bar) is sufficiently far away from the hypothesised value k

Two-Tailed Test



- The rejection region will depend on our significance level $\propto$ %
- If we choose a significance level of $\propto$ %, we are saying that we will only reject the null hypothesis if we observe a sample that has $\propto$ % or less chance of occurring
    - These values can be found by finding the standard normal value corresponding with $\frac{\propto}{2}$ %



Example

Is the population mean μ greater/less/different than some value k?

Let X have an unknown mean μ and known variance σ^2 = 9. Sample of n=16 and a sample mean of x-bar= 10

We would like to test the following set of hypotheses at the $\alpha\% = 10\%$ level of significance

$$H_0 : \mu = 8$$
$$H_A : \mu \neq 8$$

Therefore, if our sample mean has a 10% or less chance of occurring under the we will reject, H_0

By calculating the Z-score using the below formula we can determine that x-bar=10 is clearly above the hypothesised mean of 8 and is therefore rejected in favour of the alternative.

➔ This being the case, our alternative become μ>8and we only look for the right positive tail

Testing with an Unknown σ^2

We can relax the assumption that the variance is known and use Student t (with n-1 freedom)

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

If we compute a T (test statistic) that is greater than our 'student t' upper limit, then we can reject the null. '

P-Values

Instead of comparing the test statistic to the critical values (testing hypotheses) we can compute the tail probability associated with the test statistic. We will call this the p-value.

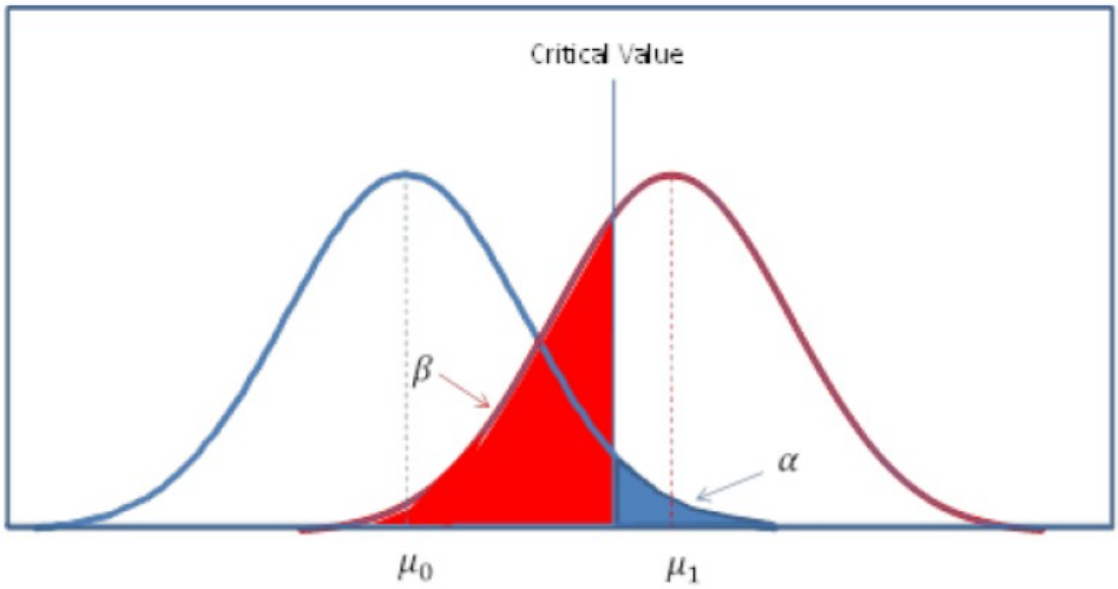
- If the p-value of a test statistic is smaller than the significance level, we reject the null
- Therefore, we would reject the null If:

$$\text{p-value} < \alpha \quad \text{or} \quad |\text{test statistic}| > |\text{critical value}|$$

Type I and Type II Errors

Type I Error: Reject the null when the null is true
Type II Error: Fail to reject the null when the null is false

- Significance level ∝ can be thought of as the probability of Type I error
- Probability of a Type II error is given by the area under the distribution centred around the alternative hypothesis to the left of the critical value



Statistical Power

Probability of correctly rejecting the null hypothesis is (1-β) → this is known as a Power Test

Types of errors

Truth

Decision based on the *p* value		**No diff** H_0 to be not rejected	**Diff** H_0 to be rejected (H_1)
	H_0 not rejected **No diff**	Right decision 1-α	**β Type II error**
	H_0 rejected (H_1) **Diff**	**α Type I error**	Right decision 1-β

- **H_0 is "true" but rejected: Type I or α error**
- **H_0 is "false" but not rejected: Type II or β error**

Testing a Population Proportion 'p'

Sampling distribution for $\hat{p}$ is:

$$\hat{p} \sim N\left(p, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

Hypotheses Testing for a proportion 'p'

$$H_0 : p = c$$
$$H_A : p \neq c$$

With test statistic:

$$Z = \frac{\hat{p} - c}{\sqrt{\frac{c(1-c)}{n}}} \sim N(0,1)$$

Estimators, Estimates and Sampling Distributions

A good estimator, ψ of an unknown population parameter Ψ has two basic properties:

1. Unbiasedness: $E[\psi] = \Psi$
2. Consistency: $plim(\psi) = \Psi$ as $n \to \infty$

- An estimator is unbiased if its sampling distribution is centred around the true population parameter
- An estimator is consistent if the sampling distribution collapse to the true population parameter as n → ∞

The CLT shows us that the estimator X-bar is an unbiased and consistent estimator of μ.

Alternative Estimators of μ

1. $\tilde{X} = \frac{1}{n}\sum_{i=1}^{n} X_i + 5$
2. $\hat{X} = X_1$
3. $\check{X} = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{2}{n}$

1. The estimator is biased and inconsistent as the sampling distribution is not centred around μ and does not collapse to μas n gets larger
2. The estimator is unbiased and inconsistent. Since $E[X_1] = \mu$, the sampling distribution is correctly centre but its variance dos not collapse as n gets larger since it is not a function of n
3. The estimator is biased and consistent. The sampling distribution for a finite n is not centred around μ , there is a bias of 2/n however this disappears as n → meaning that it is consistent

Inference with Point and Interval Estimators

Sometimes it is useful to infer the value of a population parameter using a range of values. We do this by building an interval around the point estimator. We call this an interval estimator.

Performing Inference on σ^2

An unbiased and consistent estimator of σ^2 is the sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

One unresolved issue we have is why we divide by 'n-1' as opposed to 'n'.

So we have shown explicitly that $\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is a biased estimator of the population variance σ^2 since

$$E[\tilde{S}^2] = \frac{n-1}{n}\sigma^2$$

Therefore an obvious way to correct for this bias is to multiply $\tilde{S}^2$ by $\frac{n}{n-1}$

$$E\left[\frac{n}{n-1}\tilde{S}^2\right] = \frac{n}{n-1}E[\tilde{S}^2] = \sigma^2$$

When we write things out, this correction simply returns to us the definition of the sample variance!

$$\frac{n}{n-1}\tilde{S}^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = S^2$$

Summary from Previous Lecture

- This is a biased estimator of the population variances which means that it will always underestimate the population variance

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- However, the bias will disappear as N → infinity

Once the bias has been corrected, our unbiased formula for the sample variance is:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

The Sampling Distribution of S^2

Parameter of Interest	Estimator	Sampling Distribution
μ (known σ^2)	$\overline{X}$	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
μ (unknown σ^2)	$\overline{X}$	$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim T(v)$
p	$\hat{p}$	$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$
$\mu_1 - \mu_2$ (known σ_1^2, σ_2^2)	$\overline{X}_1 - \overline{X}_2$	$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1)$
σ^2	S^2	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(k)$

- Where $v = k = n-1$ degrees of freedom

We use the Chi-Squared distribution when we want to find the population variance but only have access to the sample variance.

N.B – A Chi-Squared distribution is not symmetrical

Computing an Interval Estimator for S^2

A (1-∝)% confidence interval for σ^2 is given by:

$$\left[\frac{(n-1)S^2}{\chi^2_{k,\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{k,1-\frac{\alpha}{2}}}\right]$$

Performing a Hypothesis Test on σ^2

As with any hypothesis test we begin by specifying a null and an alternative

Null:

$$H_0 : \sigma^2 = \sigma_0^2$$

Alternatives:

$$\begin{aligned} H_A &: \sigma^2 \neq \sigma_0^2 \quad \text{(two tailed test)} \\ &: \sigma^2 < \sigma_0^2 \quad \text{(one tailed test)} \\ &: \sigma^2 > \sigma_0^2 \quad \text{(one tailed test)} \end{aligned}$$

Recapping Empirical Measures of Association

1. The sample covariance:

$$cov(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

2. The sample correlation:

$$r = \frac{s_{xy}}{s_x s_y}$$

3. The simple linear regression coefficient:

$$b_1 = \frac{s_{xy}}{s_x^2}$$

4. The coefficient of determination:

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

Covariance

All our objects/parameters so far:

Parameter	Estimator	Estimate
μ_x	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
σ_x^2	$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$
σ_{xy}	$S_{xy} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$	$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1}$
$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$	$R = \frac{S_{xy}}{S_x S_y}$	$r = \frac{s_{xy}}{s_x x_y}$

The covariance and correlation coefficient describe the shape of the joint density function.

Covariance and Correlation

We can visualise a joint density function as a surface in a three dimensional space. In the case of a bivariate normal density: