0.1 Summary

- Statistical inference is the process of making inferences about populations from information provided by samples.

Statistical inference is the process of making inferences about populations from information provided by samples.

0.2 What is Statistical Inference?

- Statistical inference is about inferring from the data about parameters that describe an assumed model for the data.
- Solution: In statistics, a model for the mechanism that has produced data is as- sumed. The model is characterized by some parameters that are unknown. Having data, statistics tries to infer from them some information about these parameters.

0.2.1 Underlying rational of hypothesis testing

- If, under a given observed assumption, the probability of setting the sample is exceptionally small, we conclude that the underlying assumption is not correct.
- When testing a clainm, we make an assumption (i.e. the null hypothesis) that contains, wholly or partially, a supposition of equality.
- We then compare the assumption and the sample results and we form one of the following conclusions.
- If the sample can easil occur when the assumption (null hypothesis) is true, we attribute that relatively small discrepanncy between the assumption and the sample results to chance.
- If the sample is an uncommon occurence when that assumption is true, we explain the relatively large discrepancy between the assumption and the sampke, by concluding that the assumption is not true.

Statistical Inference : Estimation

- When a parameter is being estimated, the estimate can be either a single number or it can be a range of numbers.
- When the estimate is a single number, such as a sample mean, the estimate is called a ***point estimate***.
- When the estimate is a range of values, the estimate is called an ***interval estimate***.
- ***Confidence intervals*** are used for interval estimation.
- As we will soon see, point estimates are not usually as informative as confidence intervals.

Statistical Inference : Confidence Intervals

- Confidence intervals allow us to use sample data to estimate a parameter value, such as a population mean.

- A confidence interval is a range of values for which we can be confident (at a specific level) that parameter value (such as the population mean) lies within.
- A confidence level will have a specified level of confidence, commonly 95%.
- The 95% confidence interval is a range of values which contains the parameter value of interest with a probability of 0.95.
- We can expected that a 95% confidence interval will not contain the parameter value of interest with a probability of 0.05.
- It is natural to interpret a 95% confidence interval on the mean as an interval with a 0.95 probability of containing the population mean.
- However, the proper interpretation is not that simple.
- Consider the case in which 1,000 studies estimating the value of μ in a certain population all resulted in estimates between 30 and 40.
- Suppose one more study was conducted and the 95% confidence interval on μ was computed to be $40 \leq \mu \leq 50$ (based on that one study).
- The probability that μ is between 40 and 50 is very low, the confidence interval not withstanding.

0.2.2 Types of inference procedures

- The two main types of inference procedures are confidence intervals and hypothesis tests .
- There are two ways of conducting a hypothesis test.
- One method is to compute the test statistic, and compare to the critical values.
- The second method is to compute the probability value (p-value), and compare it to the significance level.
- Nearly all computer programs use the p-value approach. In this course we will focus more on the p-value approach.

0.2.3 The probability value

- The probability value (sometimes called the p value) is the probability of obtaining a statistic as different from or more different from the parameter specified in the null hypothesis as the statistic obtained in the experiment.
- The precise meaning of the p-value

0.3 Assumptions for testing claims about population means

- The sample is a simple random sample
- The value of the population variance σ is known.
- Either one or both of these conditions is satisfied
  - The Population is normally distributed.
  - The sample size n is greater than 30.

To test the null hypothesis that the true mean difference is zero, the procedure is as follows: 1. Calculate the difference ($di = y_i x_i$) between the two observations on each pair, making sure you distinguish between positive and negative differences. 2. Calculate the mean difference, $\bar{d}$.

3. Calculate the standard deviation of the differences, s_d, and use this to calculate the standard error of the mean difference, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$

4. Calculate the t-statistic, which is given by $T = \frac{\bar{d}}{SE(\bar{d})}$.

Under the null hypothesis, this statistic follows a t-distribution with $n1$ degrees of freedom.

0.4 Point Estimates

Statisticians use sample statistics to estimate population parameters.

Some commonly used point estimates are

- Sample means $\bar{x}$ are used to estimate population means μ
- Sample proportions $\hat{p}$ are used to estimate population proportions π.

In statistics, estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample. There are two types of estimations:

- Point Estimation
- Interval Estimation

An estimate of a population parameter may be expressed in two ways:

Point estimate. A point estimate of a population parameter is a single value of a statistic.

Point estimation refers to the process of estimating a parameter from a probability distribution, based on observed data from the distribution.

You have seen that the sample mean $\bar{x}$ is an unbiased estimate of the population mean μ. Another way to say this is that $\bar{x}$ is the best point estimate of the true value of μ.

Some error is associated with this estimate, howeverthe true population mean may be larger or smaller than the sample mean. Instead of a point estimate, you might want to identify a range of possible values p might take, controlling the probability that μ is not lower than the lowest value in this range and not higher than the highest value. Such a range is called a confidence interval.

0.5 Interval Estimates

- Confidence Intervals
- Prediction Intervals and Tolerance Intervals
- The Student $t-$ distribution
- Concluding Remarks

0.6 Sampling Distribution of the Sample Means

Instead of working with individual scores, statisticians often work with means. What happens is that several samples are taken, the mean is computed for each sample, and then the means are used as the data, rather than individual scores being used. The sample is a sampling distribution of the sample means.

When all of the possible sample means are computed, then the following properties are true:

- The mean of the sample means will be the mean of the population
- The variance of the sample means will be the variance of the population divided by the sample size.
- The standard deviation of the sample means (known as the standard error of the mean) will be smaller than the population mean and will be equal to the standard deviation of the population divided by the square root of the sample size.
- If the population has a normal distribution, then the sample means will have a normal distribution.
- If the population is not normally distributed, but the sample size is sufficiently large, then the sample means will have an approximately normal distribution. Some books define sufficiently large as at least 30 and others as at least 31.

0.6.1 Sampling Distribution

- A probability distribution of a statistic obtained through a large number of samples drawn from a specific population.
- The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

0.7 Confidence Intervals

ν is the degrees of freedom. For large samples (samples of a size greater than thirty) $\nu = \infty$. For small sample (samples of a size thirty or less) $\nu = n - 1$

0.8 Hypothesis Testing and p-values

- In hypothesis tests, the difference between the observed value and the parameter value specified by H_0 is computed and the probability of obtaining a difference this large or large is calculated.
- The probability of obtaining data as extreme, or more extreme, than the expected value under the null hypothesis is called the ***p-value***.
- There is often confusion about the precise meaning of the p-value probability computed in a significance test. It is not the probability of the null hypothesis itself.
- Thus, if the probability value is 0.0175, this does not mean that the probability that the null hypothesis is either true or false is 0.0175.
- It means that the probability of obtaining data as different or more different from the null hypothesis as those obtained in the experiment is 0.0175.

Significance (Die Throw Example)

- Suppose that the outcome of the die throw experiment was a sum of 401. In previous lectures, a simulation study found that only in approximately 1.75% of cases would a fair die yield this result.
- However, in the case of a crooked die (i.e. one that favours high numbers) this result would not be unusual.
- A reasonable interpretation of this experiment is that the die is crooked, but importantly the experiment doesn't prove it one way or the other.
- We will discuss the costs of making a wrong decision later (Type I and Type II errors).

0.8.1 Hypothesis Testing : Two Populations

Two samples drawn from two populations are independent samples if the selection of the sample from population 1 does not affect the selection of the sample from population 2. The following notation will be used for the sample and population measurements:

- p_1 and p_2 = means of populations 1 and 2,
- σ_1 and σ_2 = standard deviations of populations 1 and 2,
- n_l and n_2 = sizes of the samples drawn from populations 1 and 2 ($n_1 > 30$, $n_2 > 30$),
- x_1 and x_2, = means of the samples selected from populations 1 and 2,
- s_1 and s_2 = standard deviations of the samples selected from populations 1 and 2.

0.9 Summary of Inference Procedures

Point Estimates:

$$\hat{p}_1 = \frac{x_1}{n_1}$$
$$\hat{p}_2 = \frac{x_2}{n_2}$$

`Hypotheses:`

$$H_0 : \quad \pi_1 \leq \pi_2$$
$$H_1 : \quad \pi_1 > \pi_2$$

- The population proportion for group 1 does not exceed the corresponding value for group 2.
- The population proportion for group 1 does exceed (is greater than) the corresponding value for group 2.

$$H_0 : \quad \pi_1 - \pi_2 \leq 0$$
$$H_1 : \quad \pi_1 - \pi_2 > 0$$

Standard Error First we computed the aggregate sample proportion $\bar{p}$.

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

The Standard Error is

$$S.E.(\pi_1 - \pi_2) = \sqrt{\bar{p} \times (100 - \bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

(*Given in formula sheet*) **Standard Error**
The Test Statistic is therefore

$$TS = \frac{(\hat{p}_1 - \hat{p}_2) - (\pi_1 - \pi_2)}{S.E.(\pi_1 - \pi_2)}$$

Critical Vale

- $\alpha = 0.05$
- One-tailed Procedure (refer back to H_1) k=1
- Large sample ($x_1 + x_2 > 30$)

Descision is $|TS| > CV$?
Comclusion: We can reject the null hypothesis, We can reasonably conclude that....

0.10 Inference

0.10.1 Standard Errors

$$H_o : p = p_0$$
$$H_a : p \neq p_0$$

$$S.E.(\hat{p}) = \sqrt{\frac{(p_o(1 - p_o)}{n}} \tag{0.10.1}$$

Important:
Formulae at back of Exam Paper

- Murdoch Barnes Table 3 (The Z distribution)
- Murdoch Barnes Table 7 (The Student t distribution)

Important Considerations

- Significance α / confidence 1- α
- Number of tails
  - one tailed procedure or two tailed procedure
  - Confidence intervals are always two tailed
- Sample size
  - degrees of freedom depends on sample size

0.10.2 Two tailed test

$H_0 := H_1 :\neq$
α is divided equally between the two tail of the critical region.
$\neq$ i.e. *"not equal to"* can also mean *"less than or greater than"*.

0.11 Assumptions for testing claims about population means, with unknown variance.

- The sample is a simple random sample
- The value of the population variance σ is not known.
- Either one or both of these conditions is satisfied

– The Population is normally distributed.

– The sample size n is greater than 30.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

degrees of freedom (df or ν) = n-1
What is the standard error S.E.($\hat{p}$)

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

When computing the standard error for a hypothesis test, we use this formula.

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{\hat{p_0} \times (1 - \hat{p_o})}{n}}$$

Where p_0 is the population proportion, as proposed by the null hypothesis.

0.12 Inferences around two proportions

0.12.1 Assumptions

- We have proportions from two independent simple random samples
- For both samples the conditions $np \geq 5$ $n(1-p) \geq 5$ are met.

For population 1, let

- p_1 population proportion
- n_1 sample size
- x_1 number of successes in sample 1
- $hatp_1$ is the sample proportion, an estimate for p_1.

0.13 Confidence Intervals

0.13.1 Margin of Error

$$E = \frac{\text{upper conf. limit} - \text{ lower conf. limit}}{2}$$

0.13.2 Point Estimates

- The sample proportion $\hat{p}$ is the best point estimate of the population proportion p.
- The sampk

0.13.3 The pooled estimate

The pooled estimate of p_1 and p_2 is denoted by $\bar{p}$.

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

i.e. The overall proportion of successes in the aggregate sample, regardless of population.

0.13.4 Inferences for matched pairs

Assumptions

1 The sample data is comprised of matched pairs.

2 The samples are simple random samples.

0.13.5 Independent samples

- These samples are taken quite independently from two populations. There is no link between observations in one sample and the other.
- The samples may be of different sizes.
- The simplest assumption to make is that both populations have normal distributions, though their means and variances may be different. We may denote the two population distributions by $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$.
- The corresponding samples (of size m and n) give rise to sample means $\bar{x}$ and $\bar{y}$ and sample variances s_x^2 and s_y^2.

0.13.6 Standard Error of a sample mean

The standard error of the sample mean is given as

$$\text{S.E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

0.13.7 Confidence interval of a mean (small sample)

If the data have a normal probability distribution and the sample standard deviation s is used to estimate the population standard deviation σ, the interval estimate is given by:

$$\bar{X} \pm t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}} \tag{0.13.1}$$

where $t_{1-\alpha/2,n-1}$ is the value providing an area of $\alpha/2$ in the upper tail of a Students t distribution with n - 1 degrees of freedom.

0.14 Interval Estimates

- Confidence Intervals
- Prediction Intervals and Tolerance Intervals
- The Student $t-$ distribution
- Concluding Remarks

0.14.1 two sample t-test

Problem: You have obtained the number of years of education from one random sample of 38 police officers from City A and the number of years of education from a second random sample of 30 police officers from City B. The average years of education for the sample from City A is 15 years with a standard deviation of 2 years. The average years of education for the sample from City B is 14 years with a standard deviation of 2.5 years. Is there a statistically significant difference between the education levels of police officers in City A and City B?

- The schedule of formulae that will be at the back of your examination paper will be posted on the SULIS site shortly.
- It is advisable to familiarise yourself with the contents before the examination.
- Please let me know as soon as possible if there are any issues with it.

Type I and Type II errors - Die Example

- Recall our die throw experiment example.
- Suppose we perform the experiment twice with two different dice.
- We dont not know for sure whether or not either of the dice is fair or crooked.
- Suppose we get a sum of 401 from one die, and 360 from the other.

Examples
For the sake of brevity, in the following example the following values are always used or realised.

- Test statistic = 2.7
- Significance level $\alpha = 0.05$. Tests are always two tailed.
- Samples are either large or of size $n = 10$
- In the case of large samples CV = 1.96. For small samples, CV = .
-
- All of the following example will have the same significance level and test statistic.
- Hence the overall outcomes are the same ($\alpha = 0.05$, k=2, TS = 2.7)

- We will use both the p-value method and the critcal value approach.
-

Hypothesis testing for Two Samples

- Paired t test
- Difference of two mean (large samples)
- Difference of two mean (small samples)
- Difference of two propotions (large sample only)

Computing the Standard Error and Test Statistic

- The point estimate is therefore $\bar{x}_1 - \bar{x}_2$
- The standard error is computed from formulae.
- The test statistic is therefore

Difference of proportions

- The point estimate is the difference
- Observed difference = 7%. Under the null hypothesis, the expected difference is 0%
- The standard error formulae

Example 1 A fair die is thrown. The number shown on the die is the random variable X. Tabulate the possible outcomes. Solution X takes the six possible outcomes 1, 2, 3, 4, 5, 6 which each have probability 1/6 (i.e. one sixth).

r 1 2 3 4 5 6 P(X = r)1/6 1/6 1/6 1/6 1/6 1/6 Example 2 Two unbiased spinners, one numbered 1, 3, 5, 7 and the other numbered 1, 2, 3 are spun. The random variable X is the sum of the two results. Find the probability distribution for X.

Solution Listing all the possible outcomes is best done in a table.

0.14.2 Question 3.2

What is the probability of getting a number divisible by 3 in each of 3 throws of a dice?

Solution

Numbers divisible by 3 : 3 and 6 probability of throwing 3 or 6:

Probability of throwing 3 or 6 three times in a row (Each throw of a dice is an independent event.)

P[3T] = P[T]P[T]P[T]=133=127

Question 1

Question 2

Question 3 $\sqrt{(\hat{p} \times (1 - \hat{p}))}$

Pen and Paper calculations are a bit easier when working in terms of percentages.

$\sqrt{(\hat{p} \times (100 - \hat{p}))}$

Question 4 (Question 4 will be covered again in next weeks tutorial, and does not need to be completed)

Question 7
Point Estimate : The sample mean $\bar{X}$
95% confidence interval quantile 1.96
(Apart from pointing out that the sample is large, the justification for using this number is not required)
Question 8
Point Estimate : $\bar{X}n$
x is the number of successes n is the sample size
$\hat{p} = 68.8\%$
$\sqrt{(}\hat{p} \times (100 - \hat{p}))$

0.15 Revision for Inference Procedures

- Definitions
- Computing Confidence Intervals
- Performing Hypothesis Testing
  - by comparing test statistics to critical values
  - by considering the p-value
  - by using the confidence interval.

0.15.1 Inference : Confidence Intervals

Basic Structure

$$\text{Observed value} \pm [\text{Quantile} \times \text{Standard Error}]$$

0.16 Inference

0.16.1 Inference

- Inferential statistics uses sampple data to make inferences (or generalizations) about a population.

0.16.2 Example

Given the population of men has normal distributed weights with a mean of 172 lbs and a standard deviation of 29 lbs,

0.16.3 Example

Let X be the score from the throw of a die.

It can be shown that, no matter what distribution theunderlying values are from, the statistics will follow a normal distribution.

0.16.4 Standard Error

- As the sampling size increases, the sampling distribution approaches the normal distribution

0.16.5 Central Limit Theorem

- Consider a sample of size n.
- The population variance σ^2 is unknown, but can be estimated by the sample variance s^2

0.17 New Section

Formulae

- The schedule of formulae that will be at the back of your examination paper will be posted on the SULIS site shortly.
- It is advisable to familiarise yourself with the contents before the examination.
- Please let me know as soon as possible if there are any issues with it.

The frequency distribution of the mileage travelled before the first major motor failure for each of 182 trucks is given below

Distance Traveled (Thousands of Miles) Frequency 0- ¡ 20 6 20- ¡ 40 11 40- ¡ 60 15 60- ¡ 80 23 80- ¡ 100 32 100- ¡ 120 43 120- ¡ 140 32 140- ¡ 160 16 160- ¡ 180 2 180- ¡ 200 2

(a) Draw the histogram corresponding to this frequency distribution. 4

(b) Calculate the mean and median for the above data set.6

(c) Complete the above frequency table with the relative and cumulative relative frequencies. Use an Ogive (cumulative relative frequency) to compute the following percentiles (approx): 10 th , 90 th , 95 th . 7

(d) Calculate the interquartile range. 5

(e) Construct a box plot for the above data. 3

(a)In exploratory data analysis it is important to test underlying assumptions about the data. Describe briefly the role of the following techniques in data analysis and indicate in an outline sketch what each would look like for well behaved data. (i)Run sequence plot (ii)Log plot (iii)Histogram (iv)Normal probability plot

(b)Outline the steps required to construct a Box Plot and draw box plots for the following situations: (i)Clean data, symmetrical, not outlines (ii)Skewed data, very large data points as outlines

(c)Construct a Box Plot for the following set of data and comment on its key features:

1, 3, 7, 8, 12, 2, 9, 12, 14, 38

Working with Small Samples

Confidence Intervals (Revision)

- The 95% confidence interval is a range of values which contain the true population parameter (i.e. mean, proportion etc) with a probability of 95%.
- We can expect that a 95% confidence interval will not include the true parameter values 5% of the time.

- A confidence level of 95% is commonly used for computing confidence interval, but we could also have confidence levels of 90%,99% and 99.9%.

Confidence Level (Revision)

- A confidence level for an interval is denoted to $1 - \alpha$ (in percentages: $100(1-\alpha)\%$) for some value α.
- A confidence level of 95% corresponds to $\alpha = 0.05$.
- $100(1-\alpha)\% = 100(1-0.05)\% = 100(0.95)\% = 95\%$
- For a confidence level of 99%, $\alpha = 0.01$.
- Knowing the correct value for α is important when determining quantiles.

$$(\bar{X} - \bar{Y}) \pm \left[\text{Quantile } \times S.E(\bar{X} - \bar{Y})\right]$$

- If the combined sample size of X and Y is greater than 30, even if the individual sample sizes are less than 30, then we consider it to be a large sample.
- The quantile is calculated according to the procedure we met in the previous class.
- Assume that the mean (μ) and the variance (σ) of the distribution of people taking the drug are 50 and 25 respectively and that the mean (μ) and the variance (σ) of the distribution of people not taking the drug are 40 and 24 respectively.

Difference in Two means

For this calculation, we will assume that the variances in each of the two populations are equal. This assumption is called the assumption of homogeneity of variance.

The first step is to compute the estimate of the standard error of the difference between means ().

$$S.E.(\bar{X} - \bar{Y}) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

- s_x^2 and s_x^2 is the variance of both samples.
- n_x and n_y is the sample size of both samples.

The degrees of freedom is $n_x + n_y - 2$.

CI for Proportion: Example (1)

- $\hat{p} = 0.62$
- Sample Size $n = 250$
- Confidence level $1 - \alpha$ is 95%

CI for Proportion: Example (2)

- First, lets determine the quantile.
- The sample size is large, so we will use the Z distribution.
- (Alternatively we can uses the $t-$ distribution with ∞ degrees of freedom.

Although the sample mean is useful as an unbiased estimator of the population mean, there is no way of expressing the degree of accuracy of a point estimator. In fact, mathematically speaking, the probability that the sample mean is exactly correct as an estimator of the population mean is $P = 0$.

A confidence interval for the mean is an estimate interval constructed with respect to the sample mean by which the likelihood that the interval includes the value of the population mean can be specified.

The *level of confidence* associated with a confidence interval indicates the long-run percentage of such intervals which would include the parameter being estimated.

- Confidence intervals for the mean typically are constructed with the unbiased estimator $\bar{x}$ at the midpoint of the interval.
- The $\pm Z\sigma_x$ or $\pm Zs_x$ frequently is called the ***margin of error*** for the confidence interval.

We indicated that use of the normal distribution in estimating a population mean is warranted for any large sample ($n > 30$), **and** for a small sample ($n \leq 30$) only if the population is normally distributed and σ is known.

- Now we consider the situation in which the sample is small and the population is normally distributed, but σ is not known.
- The distribution is a family of distributions, with a somewhat different distribution associated with the degrees of freedom (df). For a confidence interval for the population mean based on a sample of size n, $df = n - 1$.

Computing the Standard Error

$$S.E.(\hat{p}) \;=\; \sqrt{\frac{\hat{(p)} \times (100 - \hat{p})}{n}}$$

$$\hat{p} = 144/200 \times 100\% = 0.72 \times 100\%. = 72$$

$100\% - \hat{p} = 100\% - 72\% = 28\%$

Computing the Standard Error

$$S.E.(\hat{p}) \;=\; \sqrt{\frac{72 \times 28}{200}}$$

Computing the point estimate

Sample percentage

$$\hat{p} = \frac{x}{n} \times 100$$

- $\hat{p}$ - sample proportion.
- x - number of "successes".
- n - the sample size.

0.18 Inference

0.18.1 Standard Errors

$$H_o : p = p_0$$
$$H_a : p \neq p_0$$

$$S.E.(\hat{p}) = \sqrt{\frac{(p_o(1 - p_o)}{n}} \tag{0.18.1}$$

Important:
Formulae at back of Exam Paper

- Murdoch Barnes Table 3 (The Z distribution)
- Murdoch Barnes Table 7 (The Student t distribution)

Important Considerations

- Significance α / confidence 1- α
- Number of tails
  - one tailed procedure or two tailed procedure
  - Confidence intervals are always two tailed
- Sample size
  - degrees of freedom depends on sample size