Overview: Statistical Inference

In most data analysis problems, the data set we have at hand consists of cases which are drawn from a much larger set of potential cases. In this case, the data we have are known as a sample (or subset) of a population of potential cases. For example, a data set on house selling price in a city could consists of houses sold on a certain platform or period of time. It is drawn from the population of all houses sold in that city. The population could be defined based on the subject matter, problem statement or objectives of analysis. There are various ways of obtaining a sample, known as sampling methods as explained in the video by Simple Learning Pro (2015) below.

Types of Sampling Methods (4.1)

https://www.youtube.com/watch?v=pTuj57uXWlk

Source: (Simple Learning Pro, 2015)

Inferential statistics is the branch of statistics in which we use information from sample data at hand to make informed decisions or derive conclusions about the population of interest. Keep in mind that other samples might have been drawn from the same population, leading to variation thus there is a possibility of arriving at different conclusions although the samples are drawn from the same population. A sample could be given (most often in secondary data sources) or consciously selected (such as from data collection through survey). When selecting a sample, representativeness is very important to avoid sample bias. One common type is self-selection bias in voluntary response sample, a scenario often encountered in product reviews because people are usually more inclined to write reviews when they have either very good or very poor experiences with the product. Therefore, these voluntary response or self-selection samples may be not be able to reflect the true state of affairs. One of the approaches to ensure representativeness is through random sampling.

In statistical inference, both data quality and data quantity should be taken into consideration. In data science, data quality refers to completeness, consistency of format, cleanliness and accuracy. Although some may think that more data is always better, this is not the case. Even though ample data is available, we may sometimes

choose to work with a smaller set of data particularly at the testing stage. If the sample data is of good quality, it is often more efficient since data exploration and pre-processing would take significantly less time with a smaller set of data. On the other hand, massive amounts of data would be useful for some algorithms such as learning search engine queries.

Knowledge on the normal probability distribution is important in this week's topic, as some of the concepts that you will be encountering in this week are closely related to the normal distribution.

2.1 Point Estimation

Introduction

One of the most common statistical inference procedures is point estimation. Point estimation works by identifying a single statistical measure calculated from the sample as an estimate of the parameter of interest in the population. In Week 1, we have discussed measures of central tendency and dispersion such as the mean and standard deviation, respectively. In point estimation, the statistical measure of interest for the population such as population mean (μ) is estimated by its sample counterpart known as sample mean ($\bar{x}$).

Sampling Distribution

The sample mean is a point estimate and an example of a sample statistic, which refers to a statistical measure calculated for a sample data. There are many possible samples which can be drawn from a single population thus leading to variation in the sample statistics. This leads to sampling distribution of a statistic, as explained by the following video by jbstatistics (2012).

Sampling Distributions: Introduction to the Concept

https://www.youtube.com/watch?v=Zbw-YvELsaM

Source: (jbstatistics, 2012)

A sampling distribution of a statistic refers to the frequency distribution of the sample statistic over many samples drawn from the same population. The standard

error describes the variability of a sample statistic over many samples. At this point it is important to note the difference between data distribution (for individual data points) and sampling distribution (for sample statistic).

Central Limit Theorem

The Central Limit Theorem states that the sampling distribution tends to take on a normal shape as sample size increases. Specifically:

If random samples are taken from a population with mean μ and standard deviation σ, then the distribution of the sample means will be approximately normally distributed with mean μ and estimated standard deviation $\sqrt{\frac{\sigma}{n}}$ where *n* is the sample size.

The video below by New York Times gives a light-hearted and informative explanation on Central Limit Theorem and takes around 3 minutes to complete watching.

Bunnies, Dragons and the 'Normal' World: Central Limit Theorem | The New York Times

https://www.youtube.com/watch?v=jvoxEYmQHNM

Source: (The New York Times, 2013)

The Central Limit Theorem plays a key role in subsequent topics in this week on confidence interval estimations and hypothesis tests.

2.1.1 The Bootstrap

The Bootstrap

Application of the Central Limit Theorem requires distributional assumptions on the data set. Traditionally, the theorem has been used to estimate the sampling distribution of a statistic. In recent years, the advent of computing power has spurred the popularity of nonparametric approaches such as the bootstrap (Efron and Tibshirani, 1986). The idea behind the bootstrap is to think of our sample as if it

were the population. Bootstrapping involves drawing many new samples from our original sample – a process known as resampling. When resampling, we allow sample with replacement (duplicating any of the observations in the resampled data). Watch the following 2-minute video (Salford Systems, 2017) for a quick illustration on how a bootstrap sample is selected from the original data set.

Bootstrap Sampling

https://www.youtube.com/watch?v=tTZybQTE0dw

Source: (Salford Systems, 2017)

A nonparametric approach to estimate the sampling distribution of a statistic, or of model parameters, is to draw additional samples with replacement from the sample itself and recalculate the statistic or model for each resample. This procedure is known as the bootstrap and the sample which is taken with replacement is known as a bootstrap sample. The number of iterations (or resample) is set somewhat arbitrarily, although more iterations would lead to more accurate estimates of the standard error or confidence interval.

Intro to Bootstrapping in R

In Week 4, we will be exploring linear regression in statistical modelling. One of the key concepts involved in linear regression analysis is hypothesis testing about the regression parameters β.

Watch the 14-minute 47-second video below to unpack the use of bootstrapping in relation to parameter estimation. For the interested reader, you may revisit this video after completing the topics in Week 4.

- MarinStatsLectures-R Programming & Statistics. (2018). Bootstrap hypothesis testing in R with example | R video tutorial 4.4 | MarinStatsLecutres [Video]. YouTube. https://www.youtube.com/watch?v=Zet-qmEEfCU

2.2 Confidence Interval Estimation

Confidence Intervals

In 2.1, we discussed the role of point estimation in statistical inference. Instead of presenting the estimate as a single point, an alternative approach is using a confidence interval estimate. A confidence interval gives us an idea of how variable a sample result might be and enables us to communicate the potential error in a point estimate or to recommend if a larger sample. A coverage level is attached to confidence intervals, expressed as a high percentage (usually 90%, 95% or 98%). For example, a 95% confidence interval around a sample estimate should, on average, contain similar sample estimates 95% of the time when a similar sampling procedure is followed.

As an example, confidence interval estimates for the population mean are computed through the use of two formula which are derived from the Central Limit Theorem:

- z-interval estimate: $\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\sigma}{n}}$
- t-interval estimate: $\bar{x} \pm t_{\alpha/2}\sqrt{\frac{\sigma}{n}}$

where $\bar{x}$ is the sample mean, σ is the population standard deviation and s is the sample standard deviation. The values of $z_{\alpha/2}$ and $t_{\alpha/2}$ are read from the statistical tables.

The choice between using a z-interval or t-interval estimate depends on the knowledge of population standard deviation. In real world scenarios, it is almost by default that the t-interval estimate is used since the population standard deviation would be unknown. The z-values and t-values are based on the normal and t distribution, respectively and can be read from statistical tables for a given confidence level. The value of α is known as the significance level and is related to the confidence level as $(100-\alpha)$. For example, a 95% z-confidence interval has $\alpha=0.05$ thus $z_{\alpha/2} = 0.025 = 1.96$.

The z-interval and t-interval estimates require underlying distributional assumption or requirements on the sample size. With increasing accessibility of computing facilities, confidence interval estimates can also be computed using the bootstrap method discussed in 2.1.1 which do not require any underlying distribution assumptions on the data.

We will revisit confidence interval estimates in upcoming topics on statistical modelling, such as estimation of the parameters in a linear regression model.

2.3 Hypothesis Testing

Introduction

Hypothesis tests, also known as significance tests, play an important role in decision making by providing statistical information on whether an observed effect is due to random chance. As such, a hypothesis test is the statistical answer to the need of a proof that an observed difference (or effect) is more extreme than what chance might reasonably produce. A hypothesis test consists of the following key concepts:

- Null hypothesis: the “baseline” statement
- Alternative hypothesis: counter-statement to the null hypothesis
- Significance or p-value

Example 1:

Null hypothesis: Average time Malaysians spend on social media is 3 hours a day.
Alternative hypothesis: Average time Malaysians spend on social media is not equal to 3 hours a day.

Example 2:

Null hypothesis: In year 2022, 65% of Malaysians are working from home.
Alternative hypothesis: In year 2022, less than 65% of Malaysians are working from home.

Idea behind hypothesis testing

https://www.youtube.com/watch?v=cn4S3QqEBRg

Source: (Khan Academy, 2020)

The null and alternative hypothesis must account for all possibilities for the scenario. The structure of the hypothesis test (either a one-tail or two-tail test) is determined by the null hypothesis hence it is important to state the null hypothesis appropriately.

Statistical significance is a measure used to gauge if an experiment or study based on existing data yields a result more extreme than what might be produced by chance. If the result is beyond the realm of chance variation, it is said to be statistically significant.

Important terminologies used for discussing statistical significance are:

- P-value: the probability of obtaining results as unusual or extreme as the observed results
- Alpha α: the probability threshold of "unusualness" that chance results must surpass for actual outcomes to be deemed statistically significant. Typical values are 5% or 1%.
- Type 1 error: mistakenly concluding an effect is real or mistakenly rejecting the null hypothesis
- Type 2 error: mistakenly concluding an effect is due to chance or mistakenly not rejecting the null hypothesis

The basis function of a hypothesis test is to protect against being fooled by random chance; thus they are typically structured to minimize Type 1 errors.

Typically, a hypothesis testing procedure consists of the following steps:

1. Define the null and alternative hypothesis.
2. Construct a test statistic that summarizes the strength of evidence against the null hypothesis. For a hypothesis test on the population mean, the test statistic will be based on the sample mean.
3. Compute the p-value.
4. Decide whether to reject the null hypothesis based on the p-value

In recent years, much caution has been placed around the use of p-values for decision making due to the extent of misunderstanding surrounding its use (https://www.nature.com/articles/520612a). The American Statistical Association has released a statement for researchers and journal editors stressing six principles on the use of p-values (Wasserstein and Lazar, 2016, https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108). More recent guidance (Wasserstein, Schirm, and Lazar 2019, https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913) suggested the ATOM proposal: "Accept uncertainty, be Thoughtful, Open and Modest". For data scientists, a p-value is a used as a decision tool in an experiment hence it should not be considered controlling, but merely another point of information bearing on a decision. For example, p-values are sometimes used as intermediate inputs in some statistical or machine learning models.

In general, the outcome of a hypothesis test depends on the real difference between the groups being tested and also the sample size used. To distinguish a small difference, more data will be needed to detect it.

Traditionally, the normal distribution or t-distribution plays an important role in drawing conclusion for a hypothesis test. This approach is also known as parametric tests. Non-parametric approaches are preferred alternatives when the distribution of the parameter cannot be assumed to be normally distributed or unknown. With the increasing availability of computer-intensive methods, resampling methods such as permutation test becomes an attractive alternative since it does not require underlying distributional assumptions on the dataset.

Power and Sample size

One may ask, how large a sample size do I need, or how many runs of an experiment is required in a hypothesis test? There is no one-size-fits-all answer.
Power is the probability of detecting a specified effect size with specified sample characteristics (size and variability). Power is also related to β, which is the probability of Type 2 error through the relationship Power = 1 – β. There is special-

purpose statistical software to calculate power. The most common use of power calculations is to estimate how big a sample you will need.

Recall earlier that we also have Type 1 error (or alpha) of the test. Naturally, we would like to minimize both errors at the same time. Unfortunately, both errors somewhat work against each other. This is also the reason why commonly used significance or alpha levels are 0.01, 0.05 or 0.10 which have been found to serve as a good balance between these two errors. Graphical methods have also been proposed to investigate size and power of hypothesis tests (Davidson and MacKinnon, 1998, https://openurl.ebsco.com/c/lzo7om/EPDB%3Abth%3A9%3A9123945/detailv2?sid=ebsco%3Aplink&id=ebsco%3Abth%3A300861&x-cgp-token=lzo7om).

Question 1

What is the purpose of a hypothesis test?

Ans: Provides statistical information on whether an observed effect is due to random chance

Question 2

Determine the recommended order for conducting a hypothesis testing procedure.

1. Define the null and alternative hypothesis.
2. Construct a test statistic that summarizes the strength of evidence against the null hypothesis.
3. Compute the p-value.
4. Decide whether to reject the null hypothesis based on the p-value.

Question 3

Which one of the following describes "concluding an effect is real or mistakenly rejecting a true null hypothesis"?

Ans: Type I error

2.3.1 Case Study: Estimation and Hypothesis Testing

In some cases, appropriate hypothesis tests need to be determined by the analyst or data scientist based on the objective of the analysis. For example, the t-test, named after Student's t-distribution developed by W. S. Gosset, is commonly for hypothesis testing about the population mean. In some problems, we are interested to investigate the presence of association between two qualitative variables. In this case, a chi-square test of hypothesis for presence of association can be performed.

In the following R code snippet, the t.test() and chisq.test() functions are used for performing one-sample t-test about the population mean and association analysis, respectively. You can try it out with the same variable as practice, or use another variable as required in the Application Activity below.

```r
dat <- read.csv("AmesHouseNormal.csv")
summary(dat$SalePrice)
t.test(dat$SalePrice, mu=100000)  #t-test; H0: mu = 100000
chisq.test(table(dat$LandContour, dat$LandSlope)) #chi-squared test for
```

One-Sample t-test

H_0: $\mu = 100000$

H_1: $\mu \neq 100000$

R's t.test() function computes the test statistic and corresponding p-values. For the one-sample t-test as stated above on the mean of SalePrice, the test statistic is 37.337 with corresponding *p-value < 0.0001*. At a *5%* significance level, we reject H_0 and conclude that statistically the mean of SalePrice is significantly different from 100000.

Chi-Square Test for Association

H_0: LandContour and LandSlope are statistically independent.

H_1: There is an association between $LandContour$ and LandSlope.

R's chisq.test() function performs the Pearson's chi-squared test for association. The $\chi 2$ test statistic is 518.16 with corresponding p-value < 0.0001. At the 5% significance level, we reject H_0 and conclude that there is an association between the two variables LandContour and LandSlope.

Sometimes, R (or other statistical software that you may happen to use) will prompt that the chi-squared approximation may be incorrect. This could be due to several reasons, such as small sample size that leads to sparse contingency tables. In such cases, other alternatives to Pearson's chi-squared test are available such as Fisher's exact test (https://statsandr.com/blog/fisher-s-exact-test-in-r-independence-test-for-a-small-sample/). This test has been used in a study on assessing performance accuracy of different Covid-19 RTK diagnostic tests using different clinical samples (Ahmed et al., 2022).

Application Activity: Estimation and Hypothesis Testing

This case study uses the AmesHouseNormal.csv data set.

Time: 30 minutes

Purpose: This activity provides an opportunity for the student to perform estimation and hypothesis testing using R.

Tasks:

1. Identify a quantitative variable and two qualitative variables of interest. Subsequently, compute the point estimate for the mean of the quantitative variable.
2. Perform a t-test and a chi-square test of hypothesis for your selected qualitative variables from the data set.

hypotest.R

```
2 summary(dat$SalePrice)
3 t.test(dat$SalePrice, mu=10000)
4 chisq.test(dat$SalePrice)
```

/home/hypotest.R 4:26 **Spaces: 4 (Auto)** All changes saved

Console Terminal Run Submit

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39300  130000  160000  175202  205000  755000

        One Sample t-test

data:  dat$SalePrice
t = 82.021, df = 1197, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 10000
95 percent confidence interval:
 171250.6 179153.9
sample estimates:
mean of x
 175202.2


        Chi-squared test for given probabilities

data:  dat$SalePrice
X-squared = 33203970, df = 1197, p-value < 2.2e-16
```

Program exited with code 0