6. An Introduction to Statistics and Resampling

Ali Ataullah and Hang Le

21/04/2020

Contents

1 Introduction

Probability theory offers a representation of uncertainty and provides several models that may approximate real-world scenarios. For example, probability models are used to approximate returns on stocks, volcanic eruptions, spread of infectious diseases and so forth. However, to be good approximations, these models must *learn* from the real-world data. Mathematical statistics provides tools to learn about random processes that resulted in data that we observe. This session provides a brief introduction to essential concepts from mathematical statistics. Our recommended readings are Wasserman (2013) and Chihara and Hesterberg (2018).

2 Intended Learning Outcomes

By the end of this session, students should be able to

1. understand the difference between an estimator and an estimate,
2. construct sampling distribution for a statistic,
3. carry our simple bootstrap procedure to construct bootstrap distribution for a statistic, and
4. construct bootstrap confidence intervals.

```
# Load packages needed for this session


library(tidyverse)
library(gridExtra) # To display multiple plots
library(moments) # To compute skewness
```

3 Estimators and Estimates

In finance (and in many other disciplines), we use a variety of probability distributions to approximate variations in variables under study. However, to make these approximations work, we need to **estimate** parameters of probability distributions using real data. For example, suppose we need to select an optimal portfolio of stocks issued by Dorian plc and Mixolydian plc. As discussed in the previous session, the portfolio selection problem within the *mean-variance* framework requires expected values, standard deviations and correlations between assets.

Suppose we are willing to assume that the log return on a stock issued by Dorian and Mixolydian, denoted by r_D and r_M are normally distributed. That is, $r_D \sim N(\mu_D, \sigma_D)$ and $r_M \sim N(\mu_M, \sigma_M)$. Furthermore, the correlation between the two is ρ_{DM}. Then, to use the mean-variance framework, we need to estimates these parameters from observed data (e.g. historical returns).

Statistics provides various algorithm that use data as input and provide estimates of parameters as outputs. These algorithms or procedures are called **estimators** and the outputs that they produce are called **estimates**. There are usually several estimators available for parameters that we may wish to estimate. Therefore, it is important to have good criteria to compare these estimators and choose the most appropriate one. The mathematical statistics literature highlights **statistical properties** that a good estimator should have.

We can spend many years (a lifetime) studying estimators and their properties. A detailed study of methods to estimate parameters is beyond the scope of these sessions. However, we do provide a simple example to describe the **Maximum Likelihood Estimation**, one of the most widely approach to derive estimators for parameters. For a succinct review, see Wasserman (2013).

3.1 Maximum Likelihood Estimation

Suppose we have a coin for which the probability of heads is p and the probability of tails is $(1-p)$. We do not know the what p is (it may not be a fair coin!). Let Y_i denote the outcome of a toss of this coin. This is an example of a discrete random variable as the outcome is unknown. Let us assume that Y_i has a **Bernoulli distribution** with parameter p. The probability mass function of a Bernoulli random variable is

$$f(y) = p^y.(1-p)^{1-y} \tag{1}$$

For a coin toss, the outcome of a flip is not related to the outcome of the previous flip. Moreover, each flip has the same probability distribution. We say that the coins flips are **independent** and are **identically distributed**. That is, each $Y_i \sim iid\ Benoulli(p)$.

Suppose we flip the coin N times and observe x heads and $N-x$ tails. Consequently, the probability of observing the x heads and $N-x$ tails is :

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

We can use the above to define the **Likelihood Function** as

$$L(p) = p^x(1-p)^{N-x} \tag{2}$$

$L(p)$ is a function of p (its value varies with p). The maximum likelihood estimator of p, denoted by $\hat{p}$, is the value that maximises the above likelihood function. In this simple example, we can use basic calculus to determine $\hat{p}$. Specifically, take the first derivative of the likelihood function and set it equal to 0. In fact, it is usually easier to maximise the log of the likelihood function, called the **log likelihood**. For our example, the log likelihood is

$$l(p) = x\ log\ p + (N-x)\ log(1-p) \tag{3}$$

The first derivative of this log likelihood is:

$$\frac{dl(p)}{dp} = \frac{x}{p} - \frac{N-x}{1-p}$$

Setting the first derivative of the above expression, with respect to p, equal to 0 obtains:

$$\frac{x}{\hat{p}} - \frac{N-x}{1-\hat{p}} = 0$$

This leads to our Maximum Likelihood estimator for p

$$\hat{p} = \frac{x}{N} \tag{4}$$

We have not used the maximum likelihood approach to derive an estimator $\hat{p}$ for p. The estimator is a formula in which we can input the data to obtain an estimate of p. Once we have the estimator, our task is to estimate p using real data. To collect data, we flip the coin 5 times and observe the following sequence:

Head, Tail, Head, Head, Tail. So, $N = 5$ and $x = 3$. Using the above *estimator*, our *estimate* is simply

$$\hat{p} = \frac{x}{N} = \frac{3}{5}$$

.

Note that our estimator in the above example is derived using simple calculus. But in most real-world problems, mathematics required to derive and assess estimators is quite demanding. We will not derive any other estimators in our sessions.

3.2 Properties of Our Estimation Procedures

Our estimate $\hat{p} = \frac{x}{N}$ depends on the data that we have. In our example above, we have $N = 5$ and $x = 3$. This is one **sample** out of infinitely many sample of different sizes that we can generate by flipping the coin again and again. Each sample will generate a different estimate of p. This means that our estimator $\hat{p}$ is a random variable because its outcome varies from sample to sample.

More generally, suppose we wish to estimate an unknown parameter θ of a model. We derive an estimator $\hat{\theta}$, which is a formula that will use real data as input and provide an estimate of the true but unknown θ. Notice the difference in notation for the true parameter and its estimator. The estimate will vary from sample to sample. We say that our estimator $\hat{\theta}$ is a *random variable* and the its variation from sample to sample is capture by its probability distribution called the **sampling distribution**. Note that it is the estimator $\hat{\theta}$ that is random and not the estimate obtained by plugging in a particular data in the estimator.

Mathematical statistics provides details of how variations in an estimator $\hat{\theta}$ can be studied. We know from our session on probability that the variation in a random variable can be measured using standard deviation of its probability distribution. The variation in an estimator $\hat{\theta}$ is usually measured by the standard deviation of its sampling distribution. This standard deviation is called the **standard error** of the estimator and it is denoted by $se(\hat{\theta})$.

In most useful cases, deriving a formula to compute $se(\hat{\theta})$ is mathematically demanding. Standard errors of many common estimators are readily available in statistical software like `R`. For example, standard errors of parameters of a variety regression models are easily obtained. We will see this when we use linear regression to estimate the expected return for a stock.

One way is to obtain standard error of estimators is to use the **resampling methods**, which are computationally demanding but require less mathematics compared to the traditional analytical approaches. In the rest of this session, we will study basics of one particular method, the **bootstrap method**, to study variations in our estimator based on samples of data. For details, see Chihara and Hesterberg (2018). To understand basics of bootstrap approach, we develop a simple hypothetical example related to the literature on the relationship between stock returns and directors' trades.

4 Population and Samples

This section uses an example related to the role of directors' trades to explain sampling distribution of estimators and bootstrap procedure.

4.1 Directors' Trades and Trading Strategy

The financial theory lays emphasis on the significance of information asymmetries in determining stock returns (see Easley and O'hara 2004). Broadly speaking, the literature suggests that in the presence of *informed traders* in the market, investors demand a higher rate of return. One strand of literature seeks to examine the significance of trades undertaken by directors of companies who may have better information about the prospects of their companies that outside investors (see Ataullah, Goergen, and Le 2014).

Suppose you are a financial analyst. You are presently examining whether directors in the finance sector differ from those in the retail sector in terms of the amount they investor in the shares of their own companies. This information will help you determine your trading strategy. You are interested in the value of share purchases net of shares sales of directors in the two sector. The variable that you are interested in is called **net purchase**, which is computed as follows.

Suppose for company A, the value of directors' purchases of shares of their own company is £100,000, while the value of sales is £50,000. The net purchase for this company is then £50,000. You seek to address the following question: On a particular day, does the average value of net purchases in the finance sector differ from that in the retail sector?

Table 1: Directors' Net Purchases (in £)

Sector	Purchase
Retail	3014
Finance	122162
Retail	61880
Finance	16338
Finance	3860
Finance	17065
Retail	15104
Finance	34532
Finance	6089
Finance	34680

4.2 Population vs Sample

It may seem that answering the above question is quite straight forward. All you have to do is to compare the averages for two groups. However, in almost all countries, directors' trades are not instantly reported. In other words, there is a lag between the time when directors trade and the time when other potential investors become aware of these trades. Thus, as an investor trying to trade on the basis of directors' purchases on any specific day, you will not observe all trades by directors. Consequently, you will have to base your strategy on a subset of trades that are observed on a particular day. In statistical terms, we say that we not observe the whole *population* and, therefore, you will estimate the net purchases using a *sample* of data.

4.3 A Hypothetical Population

Let us assume that we do observe the whole population of directors' trades. The data for our population is available in `"directorsPurchases.csv"`.

```r
directorsTrades <- read_csv("directorsPurchases.csv")
```

This hypothetical data contains two columns. Column 1 shows the sector that a company operates in, while column 2 contains the value of the net purchases of directors of that company. The first ten observations from the data are given in Table 1.

The whole population of net purchases is given in our dataframe `directorsTrades`. Our aim is to compare the average net purchases of directors in the two sectors. Let μ_F denote the average net purchase of directors in the finance sector, and μ_R denote the average net purchase of directors in the the retail sector. These are the true parameters of the distribution of directors' net purchases in the two sector.

4.4 Exercise

1. In the exercise below, use the `group_by()` and `summarise()` functions to obtain average values in our population given in `directorsTrades`. Also, count the number of observations in each sector.

```
# Your code

# Hint: You will need three functions to compute
# relevant mean differences in the code below

# directorsTrades %>%
#  ???(Sector) %>% ???(???(Purchase))
```

2. Histograms for net purchases for the retail and finance sectors are given in Figure 1 below. Notice how we use the `filter()` function to extract data for each sector. Comment on the shape of the two histograms.

```
hist_retail <- ggplot(data = filter(directorsTrades, Sector == "Retail")) +
  geom_histogram(aes(Purchase), color = "white", fill = "thistle3")

hist_fin <- ggplot(data = filter(directorsTrades,Sector == "Finance")) +
  geom_histogram(aes(Purchase), color = "white", fill = "steelblue")



grid.arrange(hist_retail,hist_fin)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
# Remove histograms to clear R environment

rm(hist_retail,hist_fin)
```

3. Compute the average values of net purchases for the two sectors.

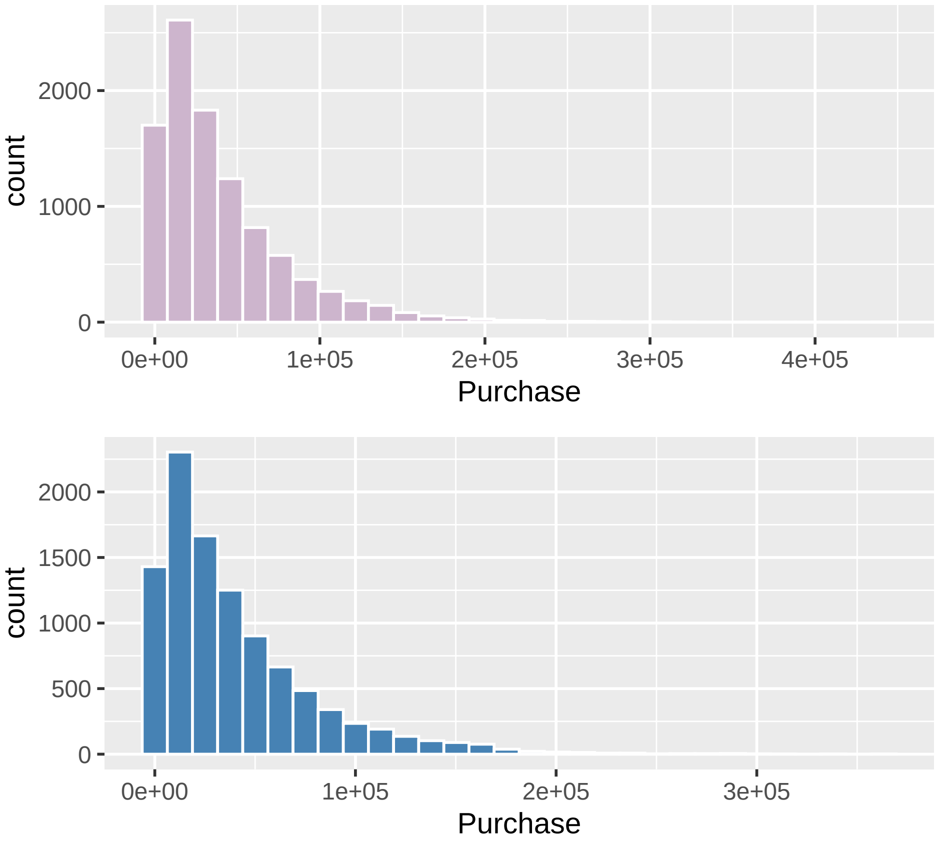

Figure 1: Histograms for Directors' Net Purchases in Two Sectors

```
# Your code
```

Your computations above should inform you that in the population, the average value of net purchases in the finance sector is £40151.25, while the average value of net purchases in the retail sector is £40125.94. We now assume that these are the true population averages. So, the true parameters are $\mu_F = 40151.25$ and $\mu_R = 40125.94$. This computation shows that in the population, the true difference in means is $\mu_F - \mu_R = 40151.25 - 40125.94 = 25.31$ .Thus, in the population, the average difference between net purchases is negligible. We store this population difference in our environment

```
population_diff <-  (40151.25 - 40125.94)
```

4.5 Sample and Estimator

In the computations above, we assume that we have access to the whole population of trades. Our population contained $20,000$ trades (10,000 for Finance and 10,000 for Retail). In the real-world, it is virtually impossible to observe the whole population. In fact, it is sometimes hard to clearly specify the population that we interested in.

In empirical analysis, we always have to rely on a subset of observations drawn (hopefully randomly) from the population. This subset is called a **sample**. We can consider the sample averages $\bar{X}_F$ and $\bar{X}_R$ as *estimators* of the true population parameters μ_F and μ_R, respectively. Let us draw a few random samples from our population of $10,000$ firms in the finance sector and $10,000$ firms in the retail sector.

```
# Create dataframe for finance and retail


Retail_population <- directorsTrades %>%
  filter(Sector == "Retail")
Finance_population <- directorsTrades %>%
  filter(Sector == "Finance")
```

We use `sample_n()` from `dplyr` to draw a random sample (with replacement)of size 500 from the retail sector and 400 from the finance sector. Imagine that these are the only trades that are reported on a particular day.

```
# Retail sector
```

```r
set.seed(1)

sample_Retail <- sample_n(Retail_population, size = 500, replace = TRUE)
```

```r
# Finance sector

set.seed(1)

sample_Finance <- sample_n(Finance_population, size = 400, replace = TRUE)
```

We draw the histograms of the two samples. These histograms are crude estimates of the distributions of purchases in the population.

```r
hist_retail1 <- ggplot(sample_Retail) +
  geom_histogram(aes(Purchase), color = "white", fill = "steelblue")

hist_finance1 <- ggplot(sample_Finance) +
  geom_histogram(aes(Purchase), color = "white", fill = "thistle3")

grid.arrange(hist_retail1,hist_finance1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```r
# Remove histograms
rm(hist_finance1,hist_retail1)
```

We now compute the average net purchase for the two samples and take their difference.

```r
sample_Retail %>%
  summarise(Count = n(),
            mean_retail = mean(Purchase))
```

Count	mean_retail
500	39937.96

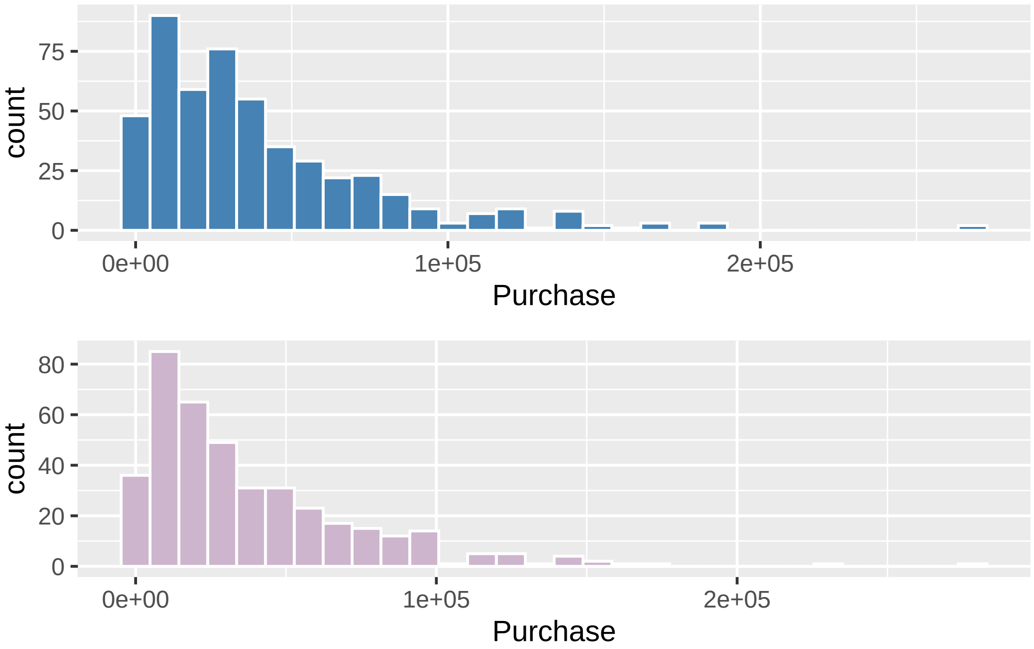

Figure 2: Histgrams for Net Purchases

```
sample_Finance %>%
  summarise(Count = n(),
            mean_retail = mean(Purchase))
```

Count	mean_retail
400	38332.71

The above computation suggests that in our sample, the average net purchases for the retail and finance sector are £39, 937.96 and 38332.71, respectively. Thus, in the sample, the difference is

$$\bar{X}_F - \bar{X}_R = 38332.71 - 39937.96 = -1605.25$$

This is much larger than the true difference in the average net purchases in the population (which is only 25).

4.6 Exercise

Draw another sample of size 500 from `Retail_population`, and of size 400 from `Finance_population`. Draw histograms of net purchases and compute the difference in average net purchases between the two sectors. This time, use `set.seed(2)`. Comments on your results.

```
# set.seed(2)

# Your code
```

The above exercise should convince you that the results obtained from the second sample are very different from those obtained from the first sample. Specifically, the difference in average net purchase in the second sample is:

$$\bar{X}_F - \bar{X}_R = 41095.99 - 40874.94 = 221.05$$

Given that our results vary from sample to sample, how much confidence do we have in these sample results? This is the question that we wish to address in this session. We will rely on the standard error of our sampling distribution for guidance.

5 Using the `replicate()` Function

This subsection shows how to use the `replicate()` function, which we will use shortly to perform a task repeatedly. Let us generate a random sample of size 5 from the Standard Normal distribution.

```
set.seed(123)

rnorm(5, mean = 0, sd = 1)
```

```
## [1] -0.56047565 -0.23017749  1.55870831  0.07050839  0.12928774
```

Let us now use the `replicate()` function to generate 5 random samples from the standard normal distribution. The output of the `replicate()` function is a matrix. We can convert it into a dataframe using `as.data.frame()` function.

```
set.seed(123)


df1 <- replicate(n = 5, expr = rnorm(5, mean = 0, sd = 1))


head(df1)
```

```
##             [,1]      [,2]      [,3]      [,4]       [,5]

## [1,] -0.56047565  1.7150650  1.2240818  1.7869131 -1.0678237

## [2,] -0.23017749  0.4609162  0.3598138  0.4978505 -0.2179749
## [3,]  1.55870831 -1.2650612  0.4007715 -1.9666172 -1.0260044
## [4,]  0.07050839 -0.6868529  0.1106827  0.7013559 -0.7288912
## [5,]  0.12928774 -0.4456620 -0.5558411 -0.4727914 -0.6250393
```

```
# Convert the matrix df1 into a dataframe

df1 <- as.data.frame(df1)
```

Let us compute mean of each sample and save it as a vector `s_means1`.

```
df1 <- data.frame(x = colMeans(df1))

df1
```

	x
V1	0.1935703
V2	-0.0443190
V3	0.3079017
V4	0.1093422
V5	-0.7331467

We now plot a histogram of the sample means in Figure 3.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

5.1 Exercises

1. Generate a random sample of size 10 from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 2$. Use `set.seed(1234)`.

```
# set.seed(1234)
# Your code
```

2. Create a dataframe `df2` that contains 200 random samples from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 2$. Use `set.seed(1234)`.

```
# set.seed(1234)
# Your code
```

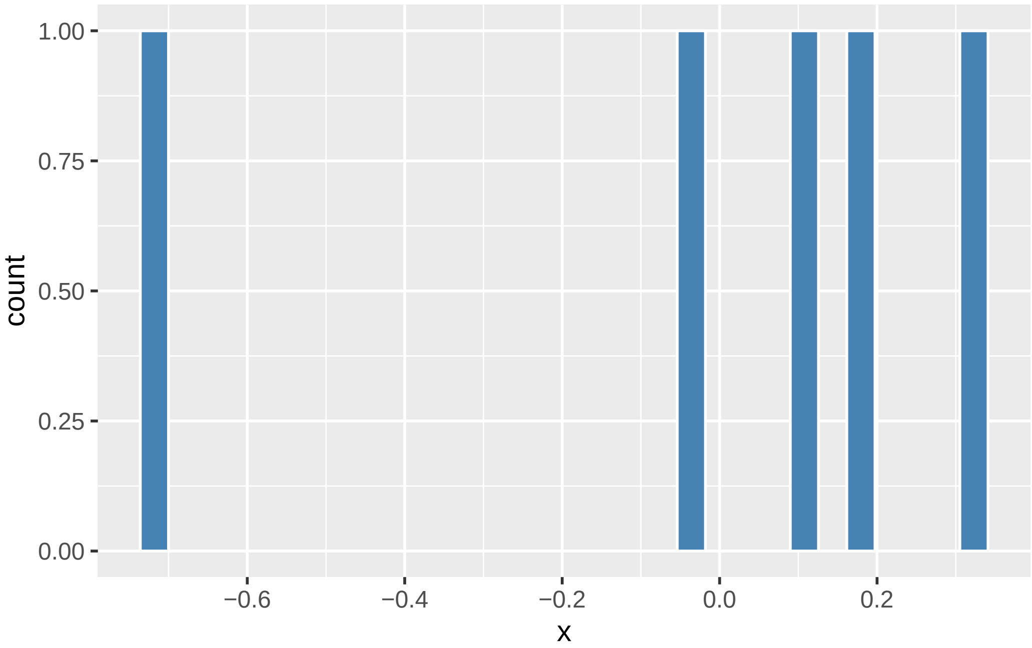

Figure 3: A Histogram of Sample Means

3. Compute the mean and median of each sample in your dataframe `df2`. Save these into a dataframe called `df2`.

```
# Your code
```

4. Plot histograms of sample means and sample medians. What do you observe by comparing the histogram of sample means that you plot with the one drawn earlier.

```
# Your code
```

6 Sampling Distribution

In order to use the sample data to assess if the average value of net purchases is different in the two sectors, we can compute the difference in sample averages:

$$\bar{X}_F - \bar{X}_R$$

This is a **sample statistic** that can shed light on the difference in true averages.

6.1 Sampling Variation of a Statistic

We have two randomly drawn samples above. In the first sample, $\bar{X}_F - \bar{X}_R < 0$, while in the second sample, $\bar{X}_F - \bar{X}_R > 0$. This means that results of our sample statistics vary from sample to sample. So, if we base our trading strategy by following directors' net purchases, then these estimates may recommend opposite strategies.

Therefore, in order to determine our confidence in the results obtained from a sample, we must measure how much our estimates vary from sample to sample. In other words, we need to estimate the variation in the **sampling distribution** of our statistic to determine our confidence in results obtained from a sample. This is an important idea for all quantitative subjects (including finance) and so we discuss this a bit further in the next section.

7 Sampling Distributions of Differences in Sample Means

Let us now draw multiple samples of different sizes from our populations of net purchases in the finance and the retail sector. Our aim is to understand how sampling distribution of differences in sample means changes with sample size.

7.1 Repeated Samples from the Population

Let us draw several samples of different sizes from our population. Specifically, we draw $10,000$ samples of size 500 in the retail sector. Similarly, we draw $10,000$ samples of size 400 from our population of $10,000$ directors' purchases in the finance sector. Note that this exercise is done only to explain the idea of sampling distribution of a statistic. In the real world data analysis situations, we normally draw only one sample from the population. We will come back to this point in the next section.

```
# 10,000 samples of size 500 each for the retail sector


set.seed(123)


sample500_retail <- bind_rows(
  replicate(n = 10000, expr = sample_n(Retail_population, size = 500,replace = TRUE),
            simplify=FALSE),
  .id="Sample_Number")
```

```
# 10,000 samples of size 400 each for the finance sector

set.seed(123)

sample400_finance <- bind_rows(
  replicate(n = 10000, expr = sample_n(Finance_population, size = 400,replace = TRUE),
            simplify=FALSE),
  .id="Sample_Number")
```

7.2 Computing Sample Averages

We now use `group_by()` and `summarise()` functions that we learned earlier to compute sample averages of each sample and each sector. We need to include two arguments in the `group_by()` function as follows.

```
sample500_retail <- sample500_retail %>%
  group_by(Sample_Number) %>%
  summarise(sampleMean_R = mean(Purchase)) %>%
  ungroup()

sample400_finance <- sample400_finance %>%
  group_by(Sample_Number) %>%
  summarise(sampleMean_F = mean(Purchase)) %>%
  ungroup()
```

We now take this opportunity to learn the `join()` function in `R`. This is an extremely useful function that helps us join different dataframes.

```
sample1 <- left_join(sample500_retail, sample400_finance, by = "Sample_Number")
```

Finally, we compute the differences in samples means for 10,000 samples.

```
sample1 <- sample1 %>%
  mutate(sample_diff = sampleMean_F - sampleMean_R)
```

7.3 Visualising the Sampling Distribution

We now have a sampling distribution on difference in sample means in `sample1` with 10,000 random samples drawn from the original population. Let us study this sampling distribution.

First, we visualise it using a histogram in Figure 4.

```
ggplot(sample1, aes(sample_diff)) +
  geom_histogram(color= "white", fill = "steelblue", bins = 50) +
   geom_vline(xintercept = population_diff, color = "red")
```

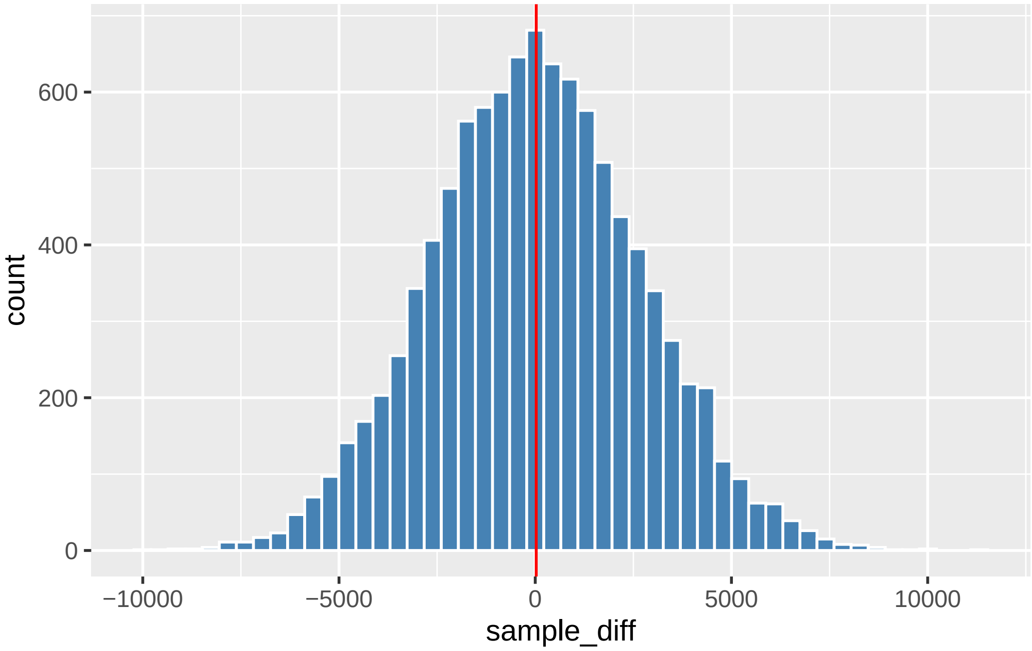

Figure 4: Sampling Distibution based on the Population

7.4 QQ-Plots

We can draw a **Quantile-Quantile plot (QQ-plot)** for the sampling distribution of our statistic. A QQ-plot is a scatterplot that compares the quantiles of our sampling distributions with the quantiles of normal distribution. If the sampling distribution is approximately normal, then the points in the QQ-plot should lie roughly on a straight line. In `R`, we can draw a QQ-plot using `stat_qq()` and `stat_qq_line()` function in the `ggplot2` package as follows. The QQ-plot is shown in Figure **??**(fig:qqplotsam).

```
ggplot(sample1, aes(sample = sample_diff)) +
  stat_qq(color = "steelblue") +
  stat_qq_line(color = "red") +
  labs(x = "Theoretical Normal Quantiles", y = "Sample")
```

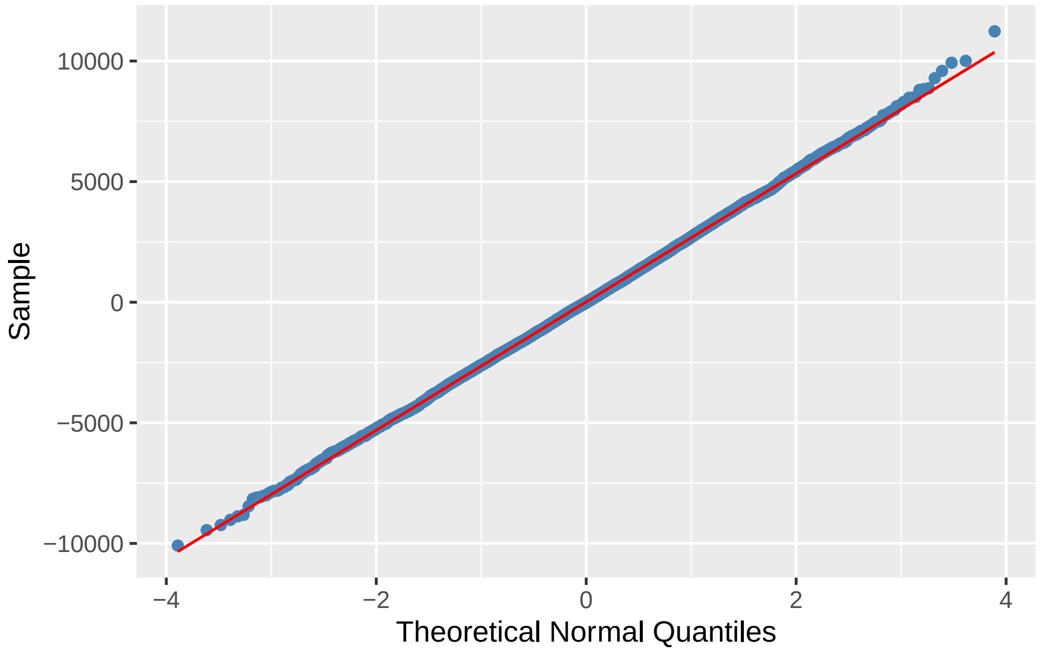

Figure 5: QQ-Plot for Sampling Distribution

7.5 Unbiased Estimator

We now compute the average of differences in means obtained from the above sampling distribution. Recall that the difference in population is £25.31.

```
mean(sample1$sample_diff)
```

```
## [1] 25.38355
```

Our computations show that on average, our sample differences in means, $\mathbb{E}(\hat{\theta})$, is very close to the true population difference, θ \$. That is,

$$\mathbb{E}(\hat{\theta}) = \theta \tag{5}$$

We say that the ratio of sample difference in means, is an **unbiased estimator** of the true population difference. If the above does not hold, then our estimator is biased and the bias is

$$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta \tag{6}$$

Thus, it appears that if we are to draw repeated samples, then on average, we will be very close to the true population value.

7.6 Examining the Dispersion of the Sampling Distribution

In our session on probability, we learned that standard deviation can be used as a measure of spread or dispersion around the mean of a random variable. Our computations show that the standard deviation (2665.90) is considerable compare to the average difference in net purchases. This suggests that there is considerable variation from sample to sample. It would be very interesting to compare this standard error with the bootstrap sampling distribution that we will obtain later. So, we store the sample standard deviation in our `R` environment.

```
sample_sd <- sd(sample1$sample_diff)
```

```
rm(sample1, sample400_finance,sample500_retail)
```

7.7 Exercise

Repeat the above computation by drawing a sample of size 500 from the retail sector and 400 from the finance sector. For your exercise, use `set.seed(2345)`. Draw histogram and QQ-plot. Also, comment on the bias and dispersion of the sampling distribution.

```r
# Your code
```

8 Bootstrap Distribution

In our examples above, we had access to the whole population. We could draw several random samples from our population and visualise the sampling distribution of our statistic $\bar{X}_F - \bar{X}_R$. The sampling distribution gives us an indication of how much our estimates vary from sample to sample. The standard deviation of the sampling distribution gives us a measure of the dispersion of the distribution. Recall that the standard deviation of the sampling distribution is called the standard error, which tell us the dispersion of our random variable around the mean value.

However, in the real world, we never observe the whole population. Thus, we cannot draw repeated samples as we did earlier. In empirical work, we normally have only one sample of data. How do we then obtain the sampling distribution of our statistics and measure its spread (standard error)? In this section, we will learn how to use one and only one sample to study how much our estimate is likely to vary from sample to sample.

8.1 The Bootstrap Procedure

he key idea behind the the bootstrap approach is that "original sample approximate the population from which it is drawn. So, resamples from this sample approximate what we would get if we took many samples from the population" (Chihara and Hesterberg 2018, p. 104). Bootstrap sampling distributions is constructed as follows (Chihara and Hesterberg 2018, p. 106):

1. Given a sample of size N, draw a sample (called resample) of size N from the original sample.
2. Use the sample to compute the statistic of interest.
3. Repeat the above a large number of times (say 10,000).
4. The sampling distribution obtained from above steps is called the bootstrap sampling distribution of the statistic.

There are excellent `R` packages available for bootstrap (e.g. `boot` and `infer`). We use functions that you have learned in these sessions to construct bootstrap sampling distribution.

8.2 Draw A Sample from the Population

Let us implement the above steps using one sample of size 500 from the retail sector, and one sample of size 400 from the finance sector from our population.

```r
# Retail sector

set.seed(12345)

sample_Retail <- sample_n(Retail_population, size = 500, replace = TRUE)
```

```r
# Finance sector

set.seed(12345)

sample_Finance <- sample_n(Finance_population, size = 400, replace = TRUE)
```

The difference in means in the two samples is now computed. This is our estimate of the true population difference in means.

```r
mean_diff <- mean(sample_Finance$Purchase) - mean(sample_Retail$Purchase)
```

Thus, our estimate of $\mu_F - \mu_R$ is

$$\bar{X}_F - \bar{X}_R = -509.791$$

We know, however, that this is not a good estimate because in the population, the true difference is only 25. So, how much confidence do we have? Our confidence will be quite low if we knew that our estimate vary substantially from sample to sample. That is, the standard error is high. But we do not have the population to draw samples repeatedly.

The bootstrap procedure begins by treating the original sample as *plug-in* for the population. We then draw repeated samples (with replacement) from our sample. We compute our statistic from each sample and construct the **bootstrap distribution**. This process is valid if it captures the characteristics of our true sampling distribution (which we happen to know in our example). Let us start bootstrapping!

8.3 Draw Resamples

We now draw 10,000 samples (called resamples) from our original sample of 500 purchases for retail and 400 purchases for finance.

```
# Resamples : Retail
set.seed(12357)

Resample_Retail <- bind_rows(
  replicate(n = 10000, expr = sample_n(sample_Retail, size = 500,replace = TRUE),
            simplify=FALSE),
  .id="Sample_Number")
```

```
# Resamples : Finance
set.seed(12357)

Resample_Finance <- bind_rows(
  replicate(n = 10000, expr = sample_n(sample_Finance, size = 400,replace = TRUE),
            simplify=FALSE),
  .id="Sample_Number")
```

We now have 10,000 resamples drawn from the original samples. We now compute means for each sample, and then compute differences in means.

```
Resample_Retail <- Resample_Retail %>%
  group_by(Sample_Number) %>%
  summarise(Retail_Mean = mean(Purchase)) %>%
  ungroup()



Resample_Finance <- Resample_Finance %>%
  group_by(Sample_Number) %>%
  summarise(Finance_Mean = mean(Purchase)) %>%
  ungroup()
```

We have 10,000 sample means for the retail sector and 10,000 sample means for the finance sector. We now join the dataframes for finance and retail and compute the difference in means.

```
# Join the two dataframe

Resample_All <- left_join(Resample_Retail, Resample_Finance, by = "Sample_Number")



# Compute the difference in means

Resample_All <- Resample_All %>%
  mutate(mean_difference = Finance_Mean - Retail_Mean) %>%
  select(mean_difference)
```

We now have our **bootstrap** distribution of the difference in means for the two sectors. Let us examine this distribution and compare its characteristics with the sampling distribution that we obtained earlier.

8.4 Visualising Bootstrap Distribution

We now plot the histogram for the bootstrap sampling distribution. This is shown in Figure 6. The red line in the figure represent the difference in sample means. This histogram is very similar to the histogram that we obtained by repeatedly sampling from the original population. The histogram in Figure 6 looks very similar to the sampling distribution in Figure 4. However, it is not centred near the true mean difference of 25.

```
ggplot(Resample_All) +
  geom_histogram(aes(mean_difference),
                 color = "white",
                 fill = "steelblue",
                 bins = 30) +
  labs(x = "Mean Differences") +
  geom_vline(xintercept = mean_diff, color = "red")
```

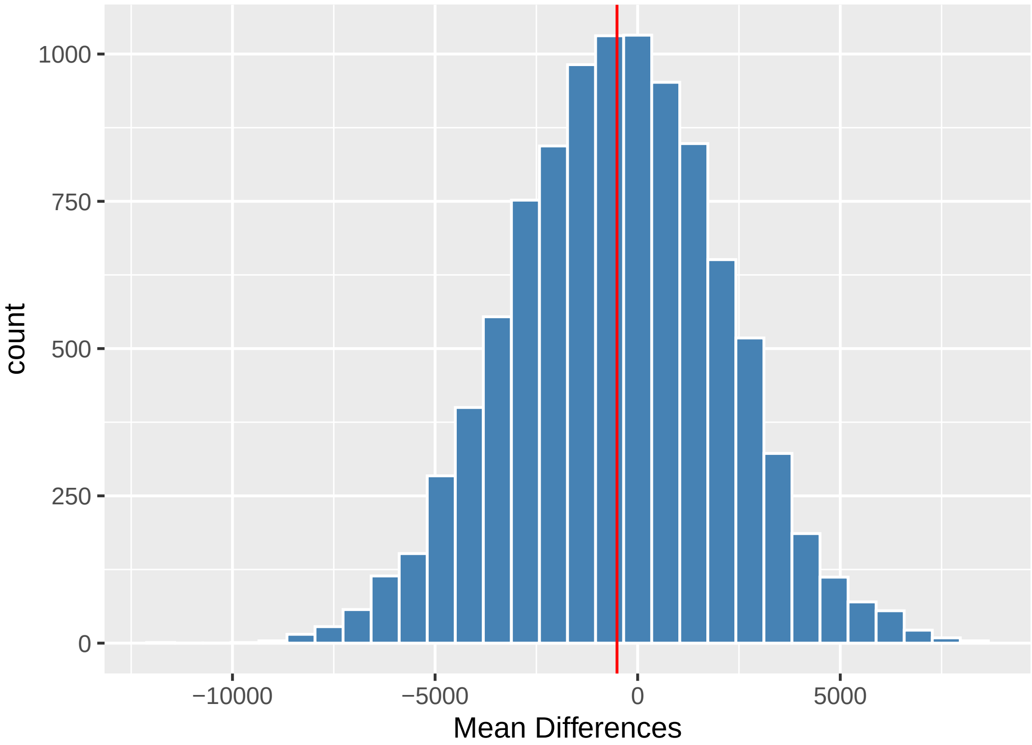

Figure 6: Histogram of Bootstrap Distribution

8.5 Usefulness of the Bootstrap Distribution

We have now seen the sampling distribution of our statistic (the difference in sample means) obtained from repeated samples from the true population and the bootstrap distribution. Let us compare the two in terms of the centre, spread, bias and skewness (Chihara and Hesterberg 2018, p. 114).

8.5.1 Centre

The average of the sampling distribution obtained from the population is approximately equal to the true population parameter. However, as the computation below shows, the average of the bootstrap distribution is not close to the true parameter. So, the centre of the bootstrap distribution does not approximate the centre of the true sampling distribution.

```
mean(Resample_All$mean_difference)
```

```
## [1] -519.8637
```

8.5.2 Bias

Equation defines the bias of a statistic. For our sampling distribution based on the true population, bias is close to 0. The computation below show that the bias for bootstrap distribution is 27. Thus, our bootstrap distribution captures the bias of the sampling distribution quite well.

```
mean(Resample_All$mean_difference) - mean_diff
```

```
## [1] -10.07267
```

8.5.3 Spread

We now have the bootstrap sampling distribution. Let us compute the standard error of this distribution. The computation of standard error below shows that our bootstrap standard error is around 2657.993. Recall that the standard error of sampling distribution when we drew repeated samples from the population was around 2665.91. Our bootstrap standard error is remarkably close to the true standard error!

```
boot_sd <- sd(Resample_All$mean_difference)
```

8.5.4 Skewness

The computation below show that the skewness of our bootstrap distribution is reasonably similar to that of the sampling distribution. This is clear from the histogram as well (looks like a normal distribution).

```
skewness(Resample_All$mean_difference)
```

```
## [1] -0.01992418
```

8.6 Key Message

The key message that we can draw from the above analysis is that for our statistic (difference in means), the bootstrap distribution provides a very good approximation to the spread (measured by standard error), skewness and bias of the sampling distribution obtained by drawing several samples from the true population. Thus, we can use the bootstrap distribution to assess how much confidence we have in our result based on one sample.

9 Bootstrap Confidence Interval

In this section, we seek to address perhaps the most important question in statistical inference: “how much confidence do we have in our estimate obtained from a sample given that estimates vary from sample to sample”? Recall that in our one sample, the mean difference is

$$\bar{X}_F - \bar{X}_R = -509.791$$

In statistical inference, instead of relying on one estimate of the quantity of interest, we construct a **confidence interval** or **interval estimate** to represent the variation that we will get in repeated samples. A 95% confidence interval is a random interval (set of possible values) such that there is 95% probability that this interval will capture the true unknown parameter (Ruppert and Matteson 2015, p. 690). If the interval is too wide, then we will have very little confidence in our sample results to provide reliable information. This is because other samples will provide very different results (i.e. too much sampling variation).

9.1 Confidence Interval Based on Percentiles

The simplest way to construct a 95% confidence interval is to determine the 2.5th and 97.5th percentiles (or 0.025 and 0.975 quantiles). This is done as follows for our bootstrap distribution.

```r
quantile(Resample_All$mean_difference, c(0.025, 0.975))
```

```
##      2.5%     97.5% 
## -5710.470  4609.135
```

As you can see, the interval is too wide. It means that other samples are likely to provide very different results compared to the result that we obtained from our sample.

9.2 Confidence Interval Based on Bootstrap Standard Error

Another way to construct 95% confidence interval is

$$\hat{\theta} \pm 2.se(\hat{\theta})$$

For us,

$$\hat{\theta} = \bar{X}_F - \bar{X}_R = -509.791$$

and

$$se(\hat{\theta} = se(\bar{X}_F - \bar{X}_R) = 2624.897$$

The 95% confidence interval is then [-5825.778, 4806.196]. This interval is also very large, which provides very little confidence in our sample estimate.

```
-509.791 - 2*boot_sd
```

```
## [1] -5759.585
```

```
-509.791 + 2*boot_sd
```

```
## [1] 4740.003
```

Based on this 95% confidence interval, we can conclude that the results obtained from our sample does not provide evidence that we can confidently conclude that the two sectors are different in terms of net purchases by directors.

9.3 Other Ways to Construct Confidence Intervals

There are other ways to construct confidence interval based on t distribution. They perform better properties. We do not discuss these here and encourage students to consult (Chihara and Hesterberg 2018, Chapter 7).

10 Next Step

This session provides a brief introduction to a key challenge in statistical inference, namely, how confident can we be that result from a sample provides evidence for a question about the real world. We provide brief introduction to sampling distribution and bootstrap distribution. Bootstrap distribution is then used to construct confidence intervals. In the next session, we learn how to use `R` for a simple study.

References

Ataullah, Ali, Marc Goergen, and Hang Le (2014). “Insider Trading and Financing Constraints”. In: *Financial Review* 49.4, pp. 685–712.

Chihara, Laura and Tim Hesterberg (2018). *Mathematical Statistics with Resampling and R*. Wiley Online Library.

Easley, David and Maureen O’hara (2004). “Information and the Cost of Capital”. In: *The journal of finance* 59.4, pp. 1553–1583.

Ruppert, David and David S Matteson (2015). *Statistics and Data Analysis for Financial Engineering with R examples*. Springer.

Wasserman, Larry (2013). *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media.