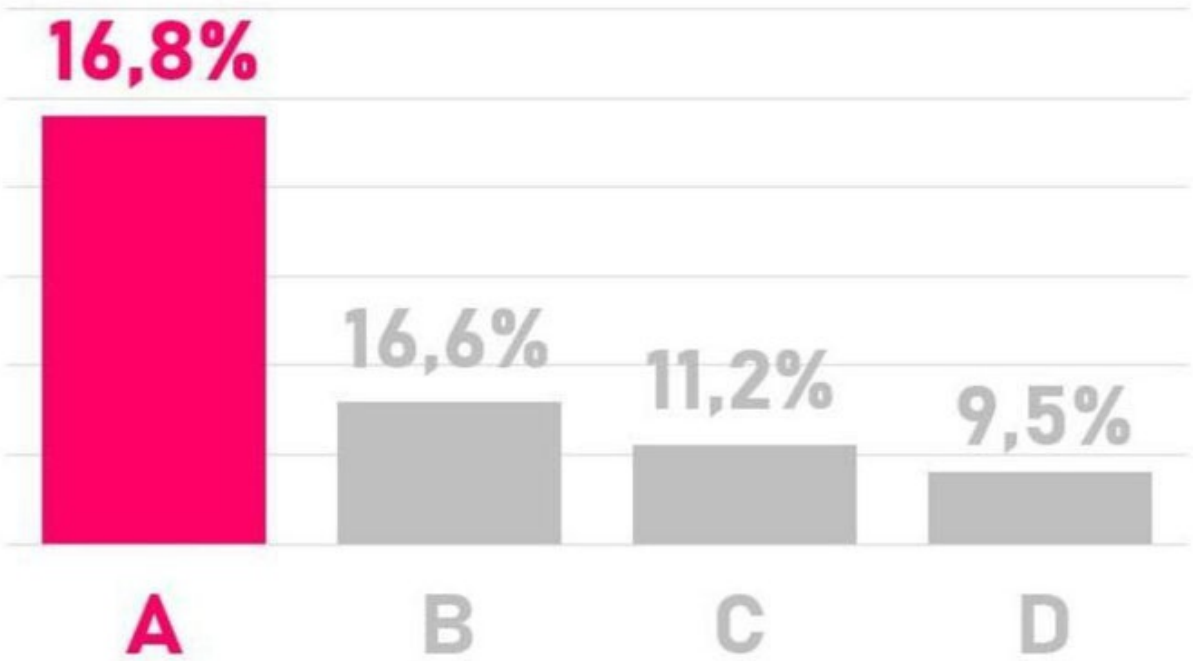
Detect Misinformation, Understand the World Deeper, and Make Better Decisions.

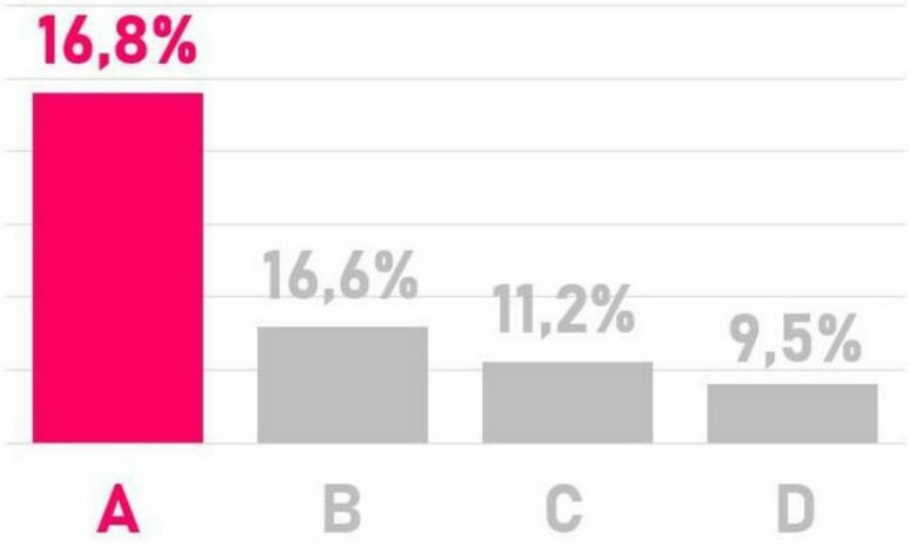
The Art of Statistical Thinking



Albert Rutherford & Jae H.Kim, PhD

Detect Misinformation, Understand the World Deeper, and Make Better Decisions.

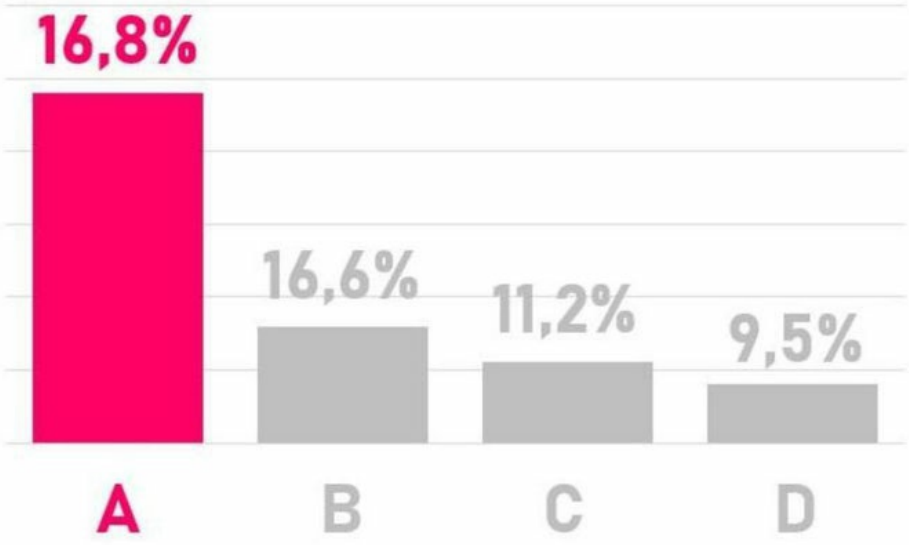
The Art of Statistical Thinking



Albert Rutherford & Jae H.Kim, PhD

Detect Misinformation, Understand the World Deeper, and Make Better Decisions.

The Art of Statistical Thinking



Albert Rutherford & Jae H.Kim, PhD

The Art of Statistical Thinking

Detect Misinformation, Understand the World Deeper,

and Make Better Decisions.

By Albert Rutherford and Jae H. Kim, PhD

I have a gift for you...

_____

Thank you for choosing my book, Practice Game Theory! I would like to show my appreciation for the trust you gave me by giving The Art of Asking Powerful Questions – in the World of Systems to you!

In this booklet you will learn:

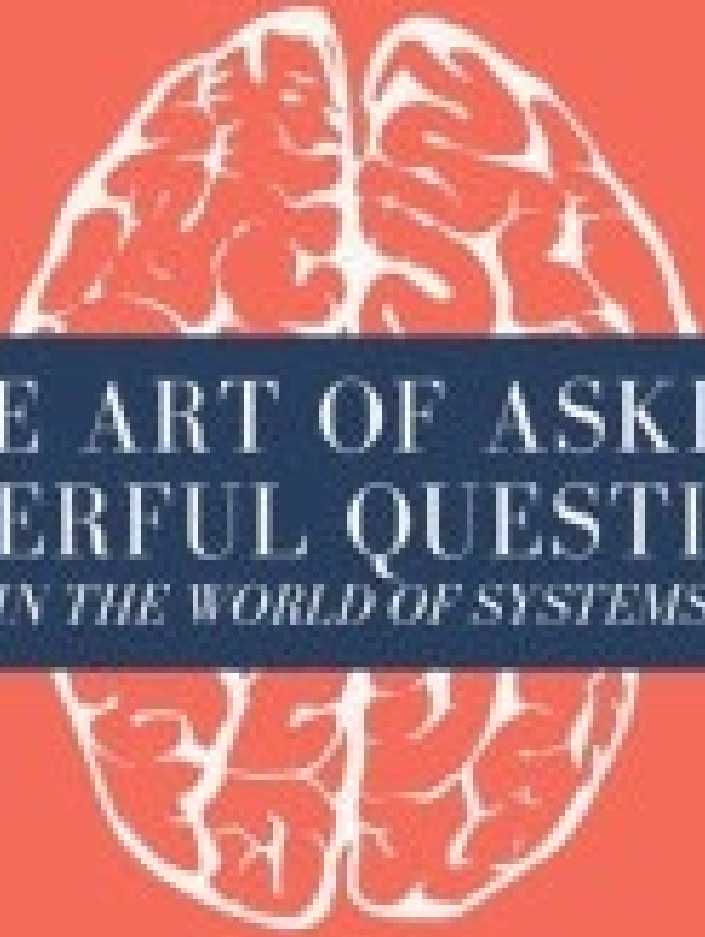
-what bounded rationality is,

-how to distinguish event- and behavior-level analysis,

-how to find optimal leverage points,

-and how to ask powerful questions using a systems thinking perspective.

Written by Albert Rutherford



THE ART OF ASKING POWERFUL QUESTIONS

IN THE WORLD OF SYSTEMS

FREE GIFT

WWW.ALBERTRUTHERFORD.COM

Click here for your FREE GIFT: The Art of Asking Powerful Questions in the World of Systems

Table of Contents

Chapter 3: Statistical Thinking

Chapter 4: How is Statistics Applied in Real Life?

Introduction

We make decisions every day - some can change our lives and those of our loved ones. But it is not only the individuals who make decisions. Companies, courts of law, governments and international organizations also make decisions, often on a large scale, that can affect our jobs, the justice system, and everyday life in a positive or negative way. Such decisions usually are made under incomplete information and uncertainty. The decision-makers often make correct decisions that will benefit our society, but they make incorrect decisions too. The cost of the latter can sometimes be devastating, starting from personal tragedies to changing the course of human history. But let's not run so far ahead.

Suppose you are making an investment decision for your retirement. Investment funds report their average returns for the past 5 years; you read a media report about the recent growth of the real estate market, and you hear about overnight millionaires who have made big from investing in cryptocurrency. You also hear about those who lost their life savings because of wrong investments or scams. And there is always a catch in the fine print: "Past performance is not necessarily indicative of future performance." This means you are facing uncertainty in your investment decisions, and you should learn how to make a well-informed decision under this circumstance.

If you make a decision after you sampled a range of different funds, compared them with those of real estate markets, and studied the future prospect of the world economy, learned from the investment gurus such as Warren Buffet and listened to your friends and advisors, then it is most likely that you have made an informed decision that will bring handsome payoff eventually. This is, in a way, "statistical thinking"; you sample the population and learn from it to make an informed decision. The more diverse and informative your sample's elements are, the more likely it is that you have made the right decision.

This book will show you how to understand statistics as a layman and make informed decisions with the help of statistical thinking. The problem is that statistics can easily be manipulated and misinterpreted. If statistical findings were always presented and utilized in an honest and correct way, the results wouldn't always be as rosy. We often see distorted and misguided numbers and outcomes, even though that was not the intention of those who report statistics. This book is intended to help readers gain better understanding and decision-making skills – the kind that professional statisticians possess. In the first chapter, we will review the definitions and basic concepts of statistics. As a book on statistics, it is inevitable to introduce mathematical details. However, these details will only be presented when necessary, without providing the full theoretical background.

Chapter 1: Definition and Basic Concepts

1. Sample versus population.

An investor wishes to know the five-year average return from investing in the U.S. stock market. There are nearly 2,400 stocks (as of August 2022) listed on the NYSE (New York Stock Exchange), and they must select a manageable number of stocks to form a portfolio of stocks. However, they don't need to calculate the average return of all 2400 stocks. There are stocks not worth investing in – too low return or too risky. Our investor will need to select a set of stocks that suits their investment style.

In this example, the collection of all stocks in the NYSE is called the population in statistical jargon, and a subset of all stocks is called a sample. Collecting the information from all the members of the population is too costly and time-consuming and even unnecessary. We can obtain a good indicator of average return by looking at a sample. The way we select the sample is critically important, and it depends largely on the purpose of the study or the aim of the statistical task at hand.

Suppose the investor's aim is to achieve a steady return with relatively low risk by investing in big and stable companies. Then a good sample is the Dow Jones index, which comprises the stocks of 30 prominent companies, such as Boeing, Coca-Cola, Microsoft, and Proctor & Gamble. If the investor's goal is to achieve a higher return with higher growth, albeit taking a higher risk, the NASDAQ-100 index is a good sample that mainly includes the top technology and IT stocks, such as Amazon, Apple, eBay, and Google. By looking at the average returns of these indices, the investor can get a clear indication and impression of the performance of these stocks. Seasoned investors can select their own sample based on their aim and risk-return preference.

The important point is that the sample should be a good representation of the target population. If the investor wants safe and steady investment returns, but their sample represents high-risk stocks, they may not effectively achieve

the aim of their investment. Hence, the target population should be determined in consideration of the aim of the statistical study.

A sample that is a good representation of the population can be obtained by pure random sampling. The members of the population are selected randomly with an equal chance. For example, in political polls, all eligible voters should be treated equally. In this situation, the most effective way of selecting an unbiased and representative sample is random sampling, where the members of the eligible voters are selected with equal chance, with no pre-selection or exclusions. In a later chapter, we will discuss an example of one of the most disastrous polling outcomes in the history, which occurred due to a violation of this random sampling principle.

2. Descriptive statistics.

Descriptive statistics is a branch of statistics where the sample features are presented with a range of summary statistics and visualization methods. The summary statistics include the mean and median, which describe the centre of the sample values, and the variance and standard deviation are the measures of the variability of the sample values. Visualization methods include plots, charts, and graphs, which are used to make a visual impression about the distribution of the sample values.

1.1. Mean and median.

The mean refers to the average of a set of values. It is computed by adding the numbers and dividing the total by the number of observations. The mean is the average of the sample values of size n, with each individual point given the weight of 1/n. The formula for the mean can be written as,

(1)

where (X1, X2,..., Xn) represent the data points and n is called the sample size. That is, the sample mean is the sum of all sample points divided by the sample size. Alternatively, it can be interpreted as a weighted sum of all data points with an equal weight of 1/n.

The median is the middle number in a sequence of numbers. To find the median, organize each number in order by size; the number in the middle is the median.[i] In statistical terms, the median is defined as the middle value of (X1, X2, ..., Xn) when sorted in ascending or descending order. Consider a simple example of (X1, ..., Xn) = (1, 2, 3, 4, 5) and n = 5. The sum of all X's is 15 (1+2+3+4+5=15), and the sample mean is 3 (15/5=3). The middle value of (1, 2, 3, 4, 5) is 3. In this case, the sample's mean and median are the same.

In general, the mean and median values are different, and the median is widely used where there are possible extreme values in the sample points. Consider the sample points with an extreme observation (X1, ..., Xn) = (1, 2, 3, 4, 20), then the sample mean is 6 (1+2+3+4+20 = 30; 30/5=6), and the median is still 3 as the middle value of the distribution (1, 2, 3, 4, 20). If this extreme value is unusual and does not represent the target population, then the sample mean of 6 can be a misleading value because it was distorted by the presence of 20. In this case, the median should be preferred to the mean.

A practical example of using the median over the mean is the case for house prices. For example, the researcher is interested in the average house price in a middle-class suburb. In such a suburb, there is still a chance that a big mansion or two in a large block of land may be included in the sale. However, these houses do not represent the general characteristics of the suburb, and it is reasonable to use the median in this case to find the average value free from the effect of these extreme values[1].

The mean vs. median is closely related with the "skewedness" of the distribution. If the distribution of the numbers you have is (more or less) symmetric around the mean as in (X1, ..., Xn) = (1, 2, 3, 4, 5), the mean and median will be identical or practically the same. However, when the distribution of the numbers is asymmetric or skewed, then the mean and median can be different. For example, if the distribution is asymmetric, as in

(X1, ..., Xn) = (1, 2, 3, 4, 20), then the two values can be different.

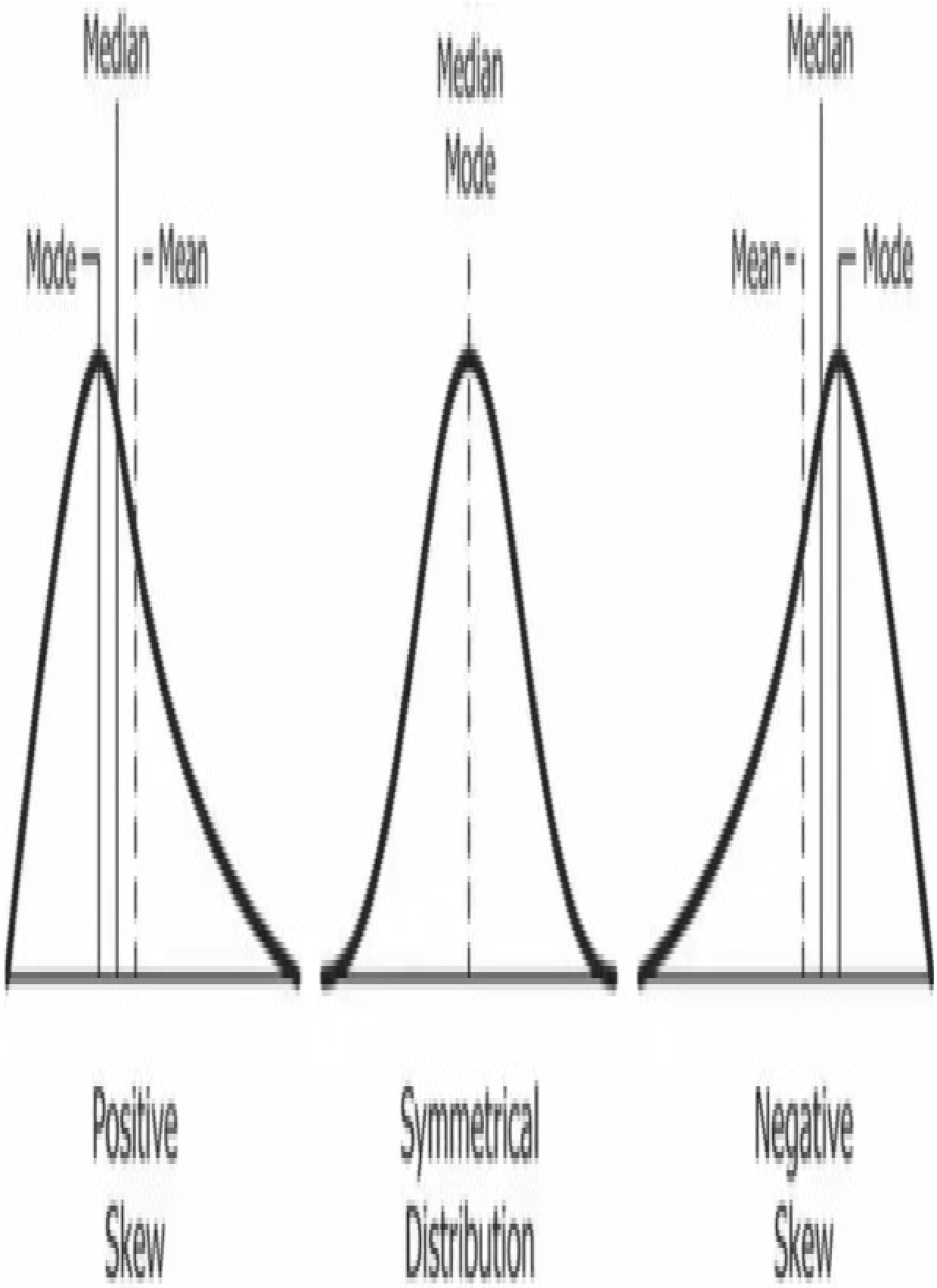

Median
Mode
Mean
Positive
Skew
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew

Photo source: Study.com[ii]

Graphical illustrations of the different shapes of the distribution and the positions of the mean and median are given above. Suppose the above is the distribution of the performance of all salespeople in a company. A symmetric distribution means the higher performers and lower performers are in the same or similar proportion; in which case the mean and median are almost identical. A positive skewed distribution means the presence of a small number of extremely capable performers. In this case, the mean of the sales is inflated by their performance. If the sales manager wants an average value that represents the performance of the "average salesperson", then the use of median is appropriate. If she wants to know the average sales, including the performance of all salespeople in the company, then the use of the mean is appropriate. A similar interpretation can also be made from a negatively skewed distribution illustrated above.

1.2. Variance and standard deviation.

When analyzing or presenting a set of numbers, it is important to know the centre of the distribution. But understanding their dispersion and variability is also important. Consider two salespeople with the same or a similar number of mean sales in the past year. In evaluating who was a more consistent performer, the manager will compare the dispersions in their sales throughout the year.

Measures of variability, variance, and standard deviation present how widespread the sample points are around the mean. The distance of the sample point from the mean is calculated as , and they are squared to make them all positive. The average of all the squared distances from the mean is called the variance, which can be written as,

(2)

How this formula works will be explained in the table below. But it is, in a way, the average of the squared distance of the data points from the mean, i.e., . The standard deviation (s) is defined as the square root of the variance, namely,

(3)

Since the variance is the distance of the sample points from the mean in squares, the standard deviation converts the value into the same unit as the original value of the sample points by taking the square root.

X		
1	-2 (=1-3)	$-2^2 = 4$
2	-1(=2-3)	$-2^2 = 1$
3	0 (=3-3)	$0^2 = 0$
4	1 (=4-3)	$1^2 = 1$
5	2 (=1-3)	$2^2 = 4$
Sum		10

=3

Using the example we used above as an illustration, X= (1, 2, 3, 4, 5) and The variance is the sum of the numbers in the last column on the chart above divided by 4, which is 10/4 = 2.5. The standard deviation is . The interpretation is that the sample points are, on average, 1.58 units away from the mean value of 3.

Why the division (or weight) is by (n-1), not by n, is beyond the scope of this book, but it is to make the calculation more accurate when the sample size is small. When the sample size is large, the division by n or by (n-1) makes no practical difference. There are other variability measures around the median (i.e., interquartile range), and they will be introduced in this book later.

3. Sample statistics and population parameters.

The sample mean ( ) and standard deviation (s) are the statistics calculated from a sample. The sample is a subset of the population, which also has the mean and standard deviation (the median and variance as well). When we use statistics, what we eventually want to know is the population values (also called the population parameters), such as the mean and standard deviation. The population mean and standard deviation are often written with Greek letters as μ and σ, values that are never known.

Suppose you want to know the mean household income of California. If you visit all the households in California to find their mean income, as in a census, you are looking for the value of μ. However, such an exercise is often neither feasible nor necessary. A good representative sample can tell us a lot about μ, as we shall see later. We can gather a random sample of 1,000 households to find their income, and this will give the value of the sample mean ( ). If the sample was a good representation of the population, it is likely the sample mean is a good indicator for the population mean.

The population and variance (and standard deviation) can be written formally as,

(4) ————
(5) ————
(6)

where N is the population size and represent the population values. The formulae above are similar to their sample counterparts in (1) to (3), hence their interpretations are similar, but they are the values of the population.

In our example, N is the number of the total households in California, and are their incomes. If 1,000 households are selected randomly and their mean income is found to be $75,000, then with n =1,000. It is hoped that this value of the sample mean is in close neighbourhood of the true value of the population mean.

Let us take another example. Consider a fictitious country with 1 million (N) eligible voters who are voting for their President. A candidate should have the support rate of more than 0.5 to get elected. The true value of the support rate (μ) is unknown, and what matter is this value on the election date. A poll is conducted from a sample of 1000 (n) eligible voters, 10 days before the election date. This value is the sample mean ( ). Suppose this sample value ( ) is 50.1 per cent. This value is called an estimate of the population parameter (μ). If the sample is a good representation of the population, this estimate of sample mean is an indicator for the value of μ, 10 days before the election date.

4. Descriptive statistics for relative position.

Suppose your IQ score is 115. A natural question is how smart are you (according to the IQ score only) relative to the other people in the sample or population. Suppose your annual income is $50,000. You want to know how rich or how poor you are relative to the others in the sample or population. You ran a marathon, and you completed the race with a record of 3 hours. You want to know your rank in the race and where your rank stands relative to all the participants of the race.

These questions are asking for a relative position, another important question in statistics. The popular measures of relative positions are percentiles (sometimes called quantiles) and quartiles.

Percentiles (quantiles)

With percentiles, we divide the distribution of the numbers into 100 positions. For example, the 90th percentile represents the value in the sample that has 10% of the sample points higher and 90% of the values lower than it. That is, if your IQ score of 115 is said to be the 90th percentile, this means you are at the top 10% of the distribution of all IQ scores.

Suppose your income of $50,000 is the 40th percentile of the distribution, then it means your income is at the bottom 40% of the distribution. That is, if there were 1000 people in the sample, your income stands at the 400th position when all incomes are sorted in ascending order.

Similarly, among the 100 runners who participated in the marathon event, suppose your record of 3 hours is at the 75th percentile. This means your record is at the top 25%, and there are 24 runners who finish the race with a better record than yours, and 74 of them were behind you.

Quartiles

Quartiles are similar to percentile, but instead of dividing the distribution of the numbers into 100 positions, they are based on the division into 4, as the following table shows:

Quartile	Lower	Higher		
1st	25%	75%	25th percentile	
2nd	50%	50%	50th percentile	Median
3rd	75%	25%	75th percentile	
4th	100%	0%	100th percentile	Maximum

The first quartile is the value whose position is at the bottom 25%, and it is the same as the 25th percentile. The second quartile is the 50th percentile, which is also the median. If we go back to your marathon record, your record of 3 hours is the third quartile of the distribution.

Interquartile range

An interquartile range is defined as the difference between the third and 1st quartile of the distribution. It is a measure of variability or dispersion of a distribution alternative to the standard deviation. As the difference between the 3rd and 1st quartiles, the length of the interval contains the (middle) 50% of the data points around the median.

Similarly to the median, the interquartile range is not sensitive to a few extreme values in the distribution, while standard deviation can be inflated by extreme values. More examples will follow for the interquartile range.

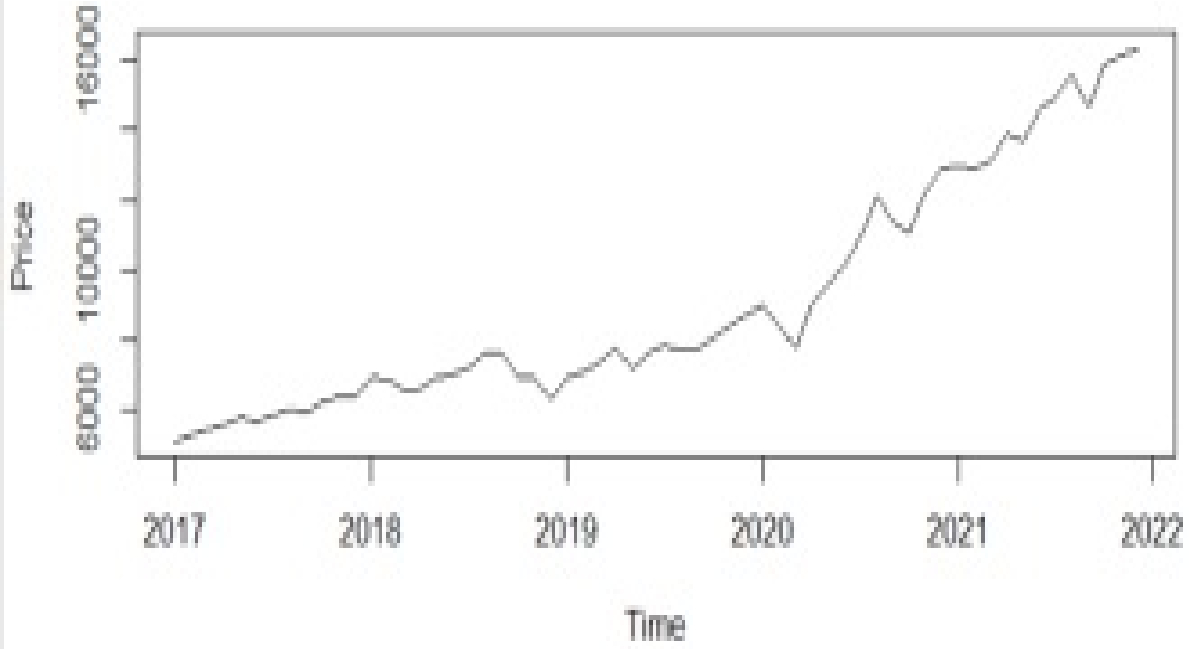
As an example, consider two suburbs whose median house prices are similar at 1 million dollars. The researcher finds the first suburb has the 1st quartile at the $750,000 and the 3rd quartile at $1.25 million, with the interquartile range of $500,000 ($1.25 million - $750,000). The second suburb has the 1st quartile at the $500,000 and the 3rd quartile at $1.5 million, with the interquartile range of 1 million dollars ($1.5 million – $500,000). The interval that contains the middle 50% of the house prices are much longer in the second suburb, which indicates the variability of house prices is substantially larger in the second suburb.

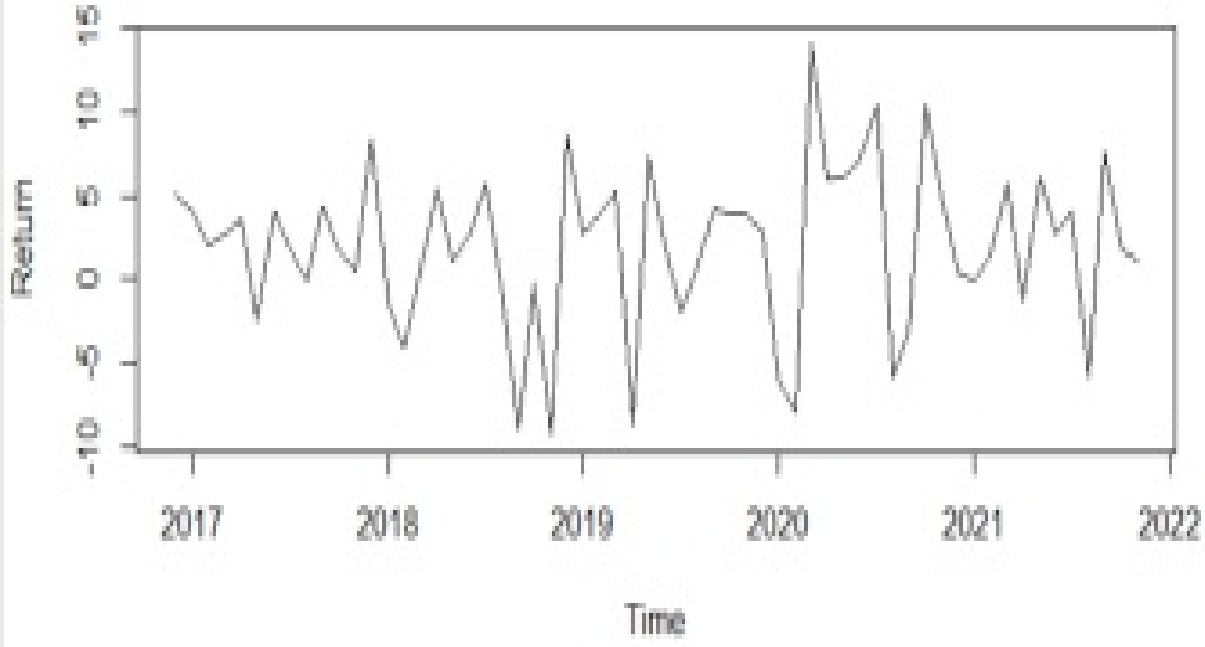
5. Data Visualization

Visualization is a powerful way of understanding the key features of a sample and making impressions. It often makes a better and stronger impression about the data characteristics than a table full of numbers.

Consider an investor who wishes to invest in U.S. stocks. They gather the sample for NASDAQ-100 index and want to know how the index and its return have performed in the last 5 years to December 2021. Figure 1 presents the line charts (time plots) of and return (growth rate) in percentage, monthly from 2017 to 2021. The index has been growing with an upward trend for the last 5 years, and the trend gets steeper from early 2020. The monthly return fluctuates around 0, with most values between -10% and 10%. These plots provide a clear impression of how the index has performed in the last five years.

Figure 1: Time plots of NASDAQ-100 index and return


Price
16000
10000
8000
2017
2018
2019
2020
2021
2022
Time

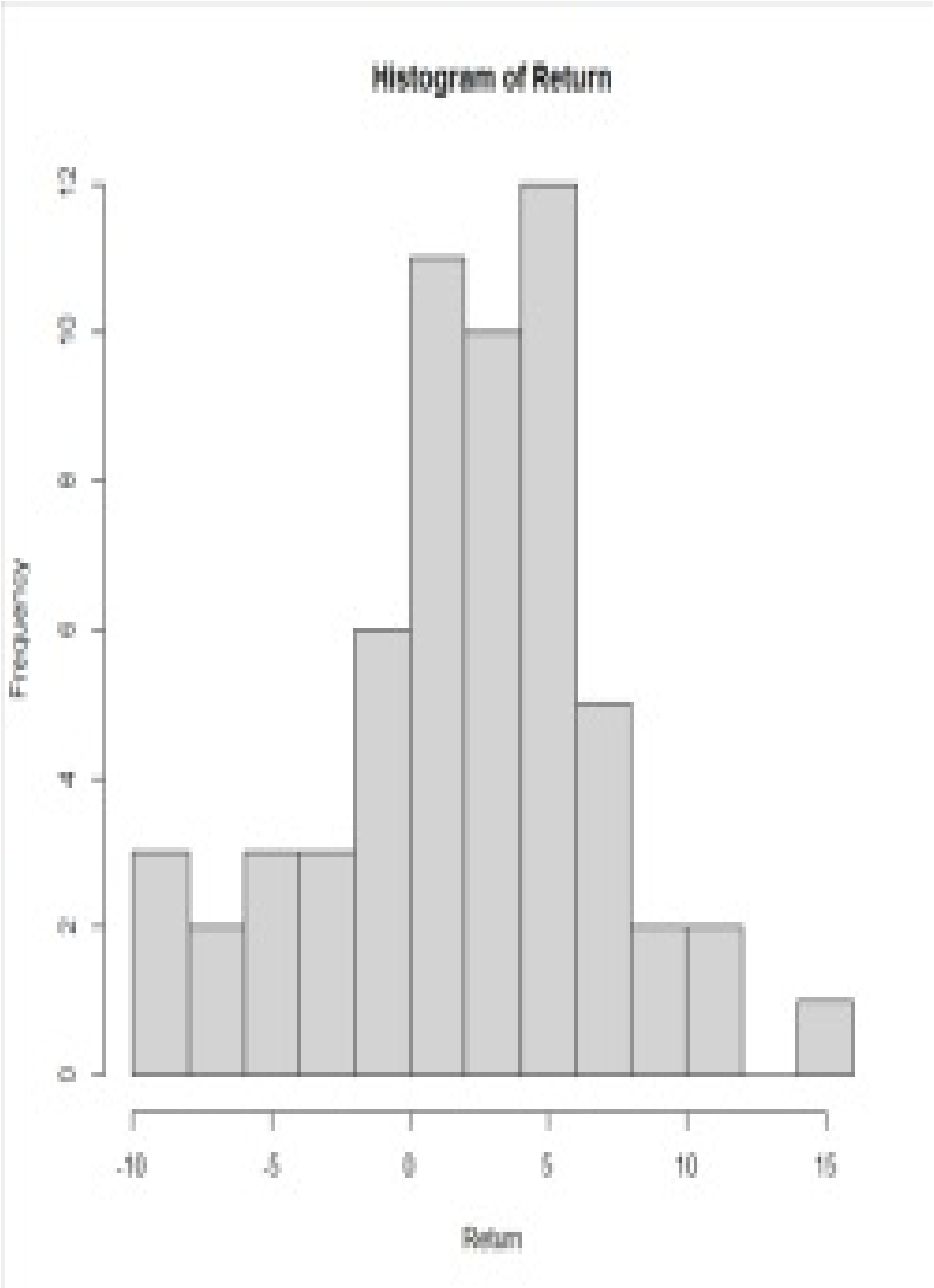

Return
15
10
5
0
-5
-10
2017
2018
2019
2020
2021
2022
Time

Data source: Yahoo Finance.

A histogram is another popular method of data visualization that presents the frequencies of data points over the intervals of sample points. It is a useful method of presenting the distributional shape of the sample points. Figure 2 presents the histogram of the monthly returns, which shows the monthly returns are centred between 0% and 5%, and most of the values are in the range of -10% and 10%.

The sample mean value of the monthly return is 2.02%, and their median is 2.68%, so the index has been increasing at an average growth rate of just higher than 2%. The standard deviation is 4.92%, which indicates the average deviation of the monthly returns from the mean has been around 5%. By combining the plots and summary statistics, the investor can learn about the performance of the index in detail.

Figure 2: Histogram of Returns from NASAQQ-100 index.


Histogram of Return
Frequency
0
2
4
6
8
10
12
-10
-5
0
5
10
15
Return

Data source: Yahoo Finance.

6. Comparing alternative distributions.

Now suppose the investor wishes to compare the performance of the NASDAQ-100 with the Apple stock (APPL) for the same period. The following table compares the basic statistics discussed so far.

Monthly returns for two alternative investments

	NASDAQ-100	APPL
Mean	2.01	3.02
Median	2.68	5.00
Standard Deviation	4.92	8.34
1st Quartile	-0.18	-1.66
3rd Quartile	5.13	9.25
10th percentile	-5.89	-7.35
90th percentile	7.37	12.27

Data source: Yahoo finance.

The figures in this table reveal many details of the two investment alternatives:

The average return from NASDAQ-100 is substantially lower than APPL. The mean and median of the former is 2.01% and 2.68% per month, but those of APPL 3.02% and 5.00%.

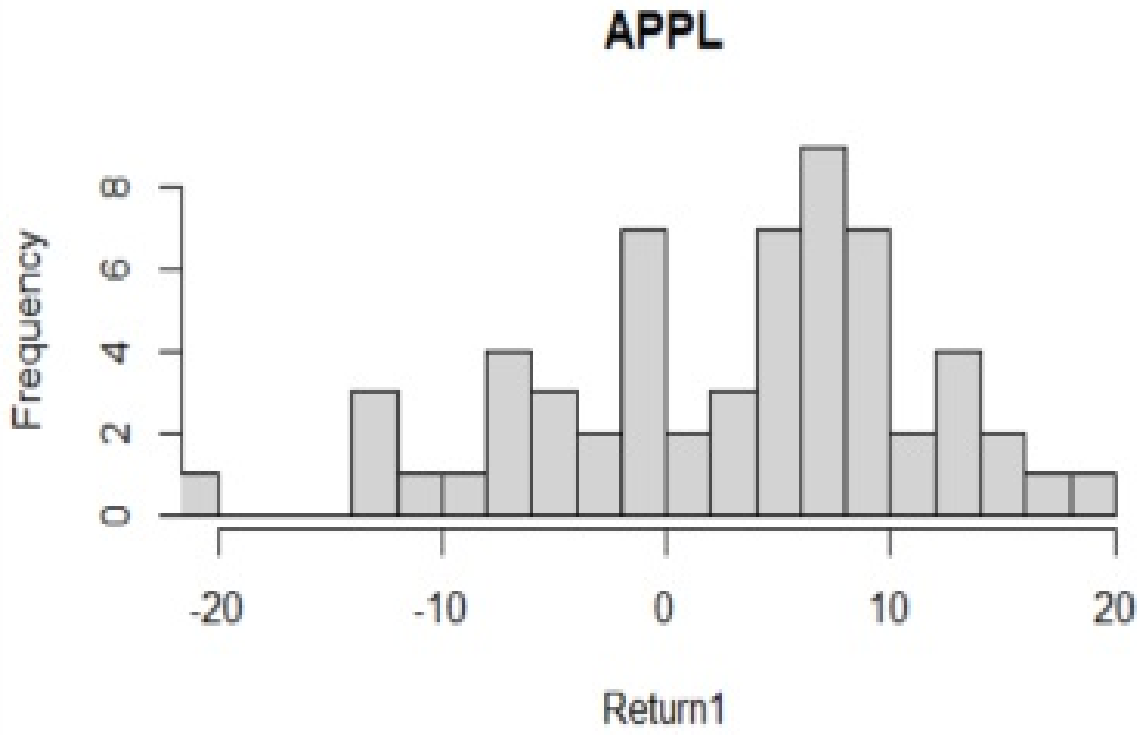
For both cases, the median is larger than the mean, especially the APPL. This means the distribution is skewed to the left, with the presence of extremely low returns. This means, when they go down, they can go down deep! (Especially APPL!)

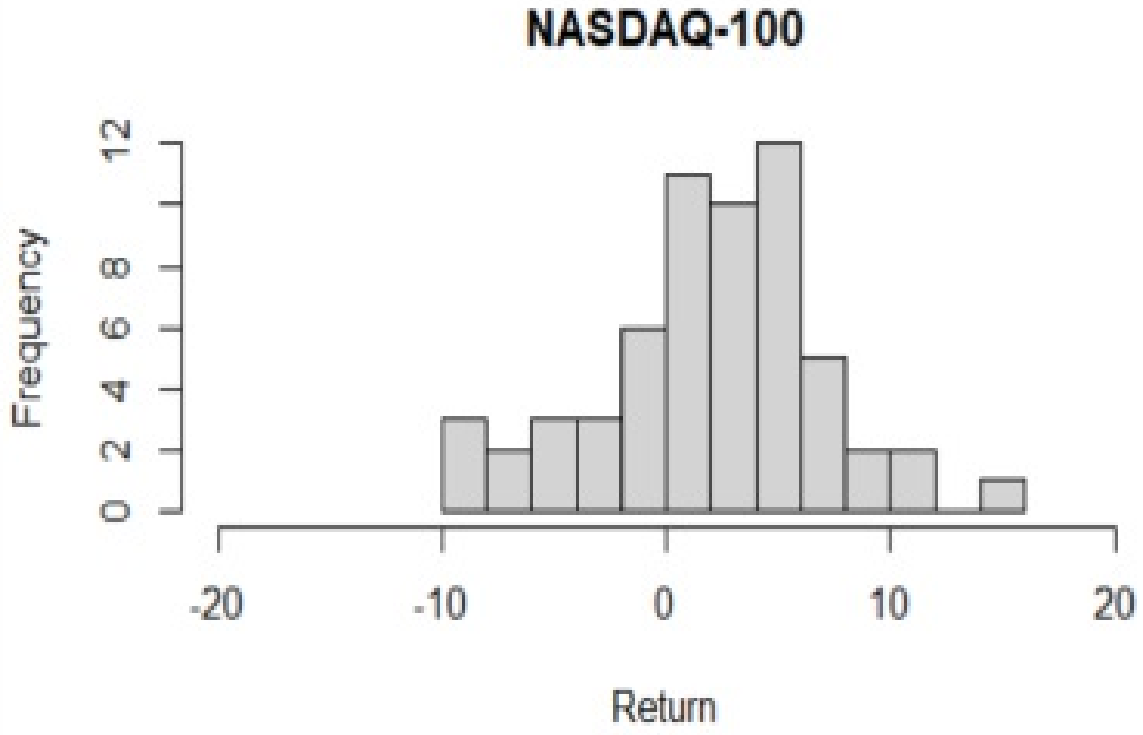
The variability is a lot higher for the returns from APPL. The standard deviation of APPL (8.34) is nearly twice larger than that of NASDA-100 (4.92). This means APPL has a lot larger variation around the mean.

The interquartile range for APPL is 10.91 (9.25 + 1.66) and that of NASDAQ-100 is 5.31 (5.13+0.18). The length of interval that contains the middle 50% of the returns around the median is again nearly twice larger for the APPL.

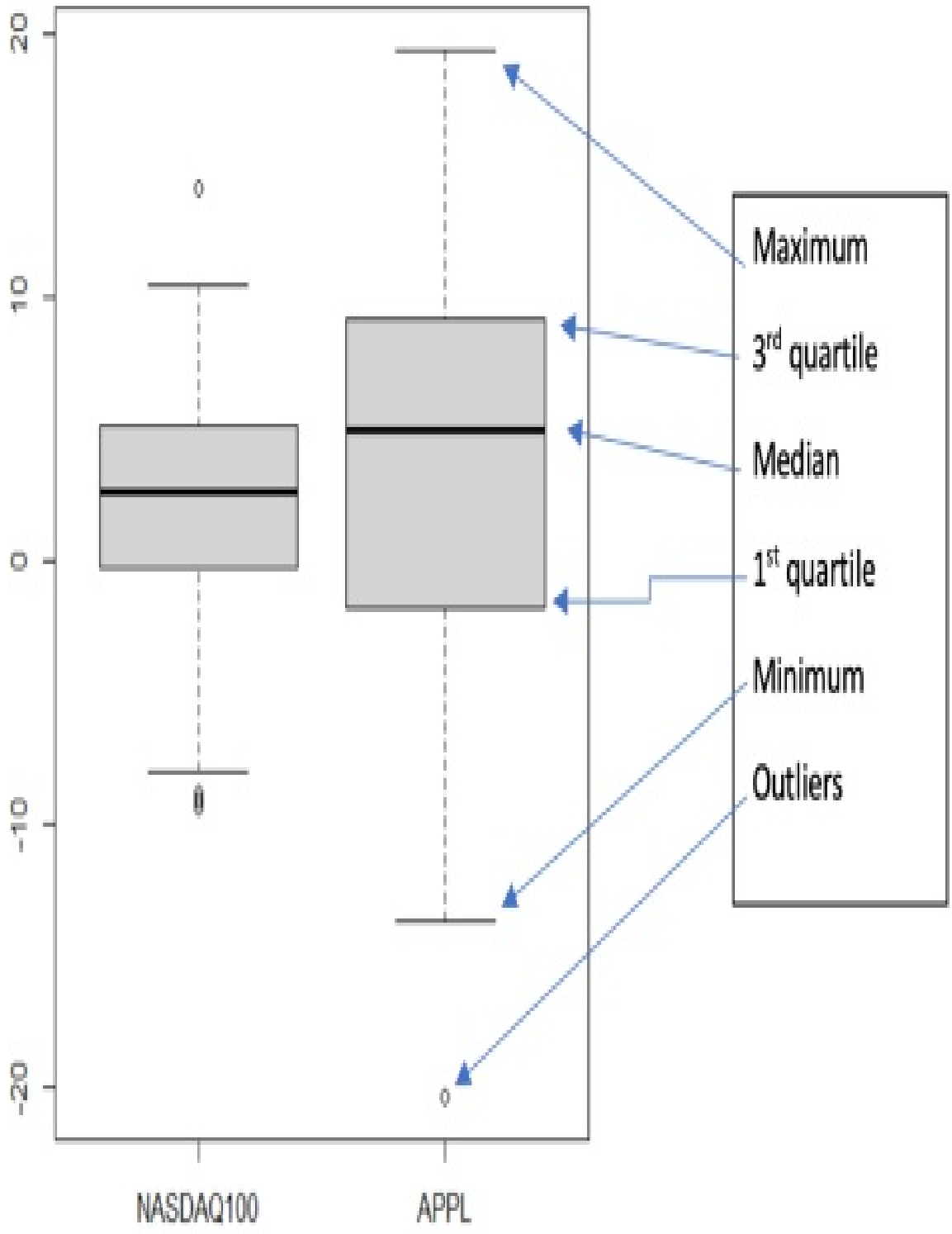
The worst possible outcome with 10% chance for APPL has been -7.35%, and that for NASDAQ-100 has been -5.89%. The best possible outcome with 10% chance for APPL has been 12.27% a month, and that for NASDAQ-100 has been 7.37%.

The comparison of these descriptive statistics reveals that monthly returns are a lot higher for APPL investment, but it shows substantially higher variability or risk. This is a well-known principle in finance: a higher return is compensation for taking a higher risk.


APPL
Frequency
0
2
4
6
8
-20
-10
0
10
20
Return1


NASDAQ-100
Frequency
0
2
4
6
8
12
-20
-10
0
10
20
Return

The above plots present the histograms for the two investments. A larger variability of the APPL with a heavier skew to the left of the distribution than NASDAQ-100 is clear. While the summary statistics tell the difference with the numbers, these histograms can make a visual comparison.


20
10
0
-10
-20
NASDAQ100
APPL
Maximum
3rd quartile
Median
1st quartile
Minimum
Outliers

To make a further visual comparison, another method of visualisation called the “Box-Whisker” plot is introduced. It plots the mean, the median, the 1st quartile, the 3rd quartile, maximum and minimum, along with outliers. The box in the middle is based on the 3rd quartile and 1st quartile, and the height of the box represents the interquartile range. Outliers are determined by a certain criterion (i.e., the outliers are defined as those lying three standard deviations away from the mean).

Again, the APPL investment gives a substantially higher median return per month, but its monthly variability is much higher than NASDAQ-100. Which investment to choose depends on how risk-averse or risk-tolerant the investor is. If you are a Braveheart and enjoy a roller coaster ride, investing in APPL is not a bad choice; otherwise, stick to the NASDAQ-100 for a safer option.

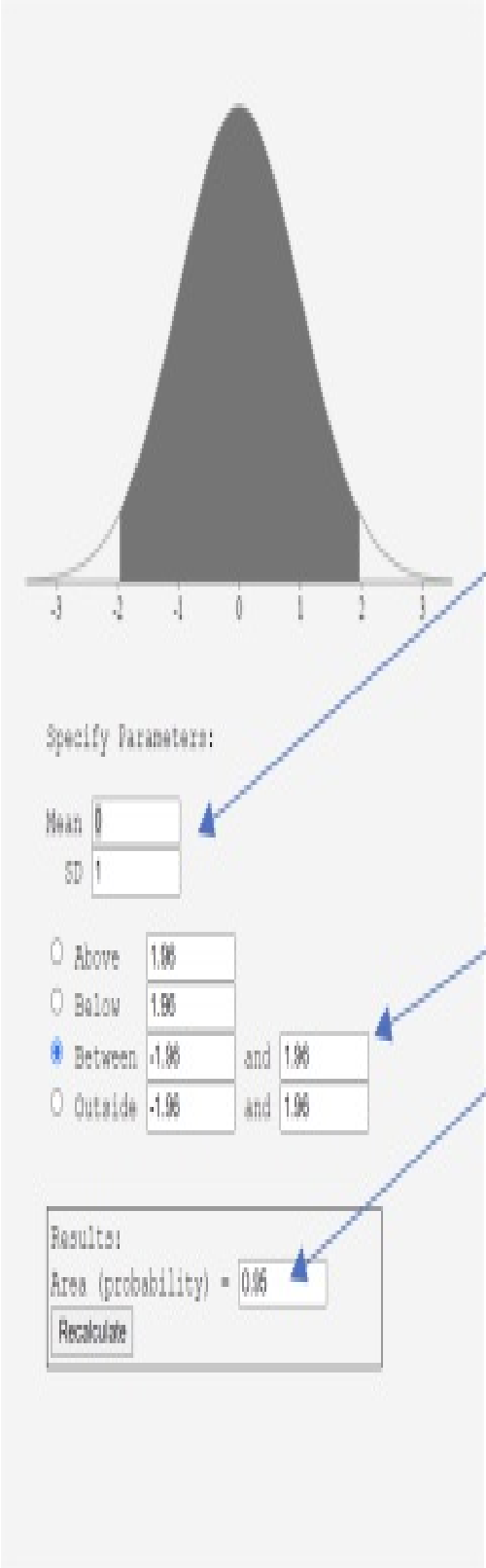
7. Normal distribution.

Figure 2 presents a distribution of the sample points using a histogram. In statistics, distribution is an important feature for both the sample and the population. While we can observe a distribution of the sample as in Figure 2, that of the population is often unknown and not observable. Understanding the features of a distribution is one of the fundamental questions of statistics. For example, what is the chance that investing in the NASDAQ-100 index will provide a return greater than 2%? What proportion of the households in California has a lower annual income than $50,000? We can only guess using the distribution of the sample we observe. Again, if the sample is a fair representation of the population, the distribution of the sample can well reflect the distribution of the population.

On the other hand, there are several known distributions in statistics where the probability can be calculated using the given values of the parameters, such as the mean and standard deviation. Among them, the most fundamental and popular is the normal distribution. It is also a key distribution in the inferential statistics to be discussed in the next chapter.

Normal distribution is a bell-shaped distribution, symmetric around its mean (or median), and the probability at any point of the distribution is known. A normal distribution with a mean μ and standard deviation of σ is written as $N(\mu,\sigma)$. In the special case of the mean being zero and the standard deviation 1, it is called standard normal distribution, and it is denoted as $N(0,1)$. Figure 3 is a screenshot from an online calculator.[2]

Figure 3: Standard normal distribution.



This is the standard normal distribution with 0 mean and 1 standard deviation (SD).

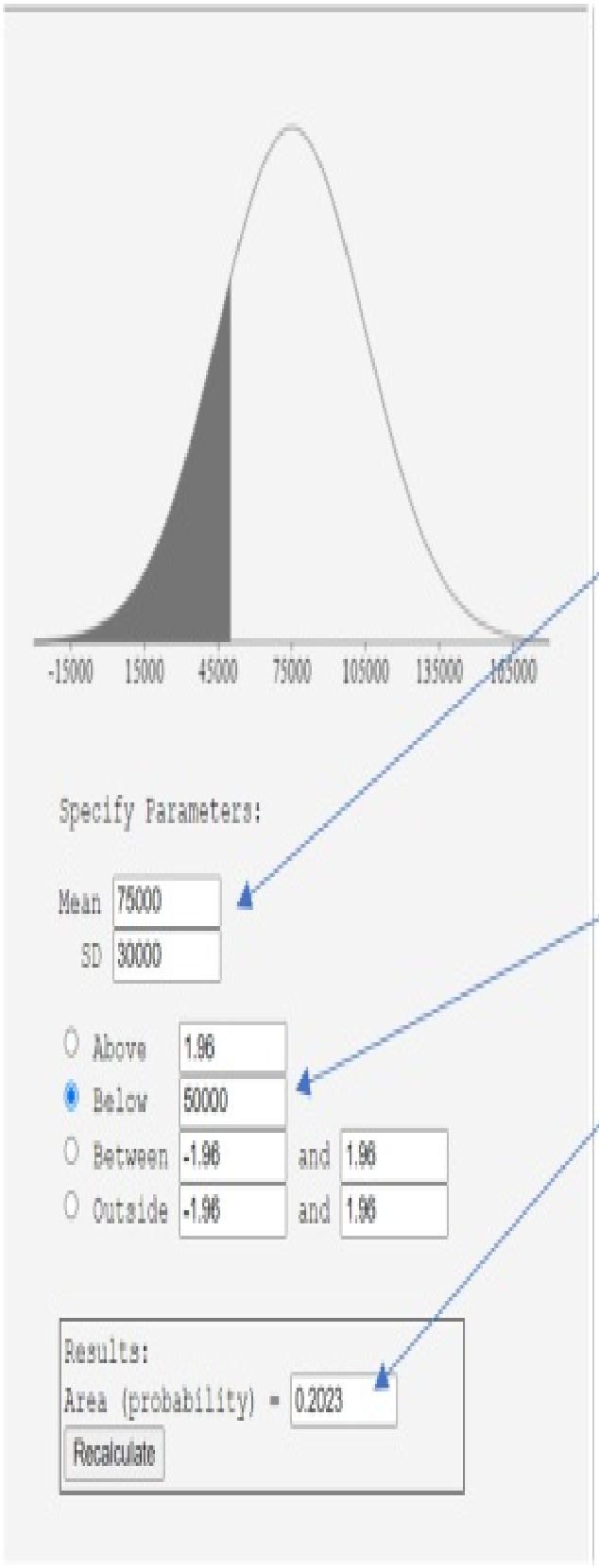
The probability of the distribution taking a value between -1.96 and 1.96 is 0.95

Given the values of the mean and standard deviation, any probability between an interval can be calculated. Figure 3 shows a normal distribution with zero mean and a standard deviation of 1 (called the standard normal distribution).

Suppose your return (in percentage) from an investment follows the standard normal distribution. The probability that your return is between -1.96% and 1.96% is calculated to be 0.95 (dark area on the bell illustration). This also means the probability of the tail areas is 5% (white area on the bell illustration). Your investment return can be lower than -1.96% with the probability of 0.025 and can take a value greater than 1.96% with the probability 0.025.

Let’s assume the household income in California follows a normal distribution of $75,000 with the standard deviation of $30,000 (see Figure 4). Then, the household income distribution of California is represented by the bell curve in Figure 4. The probability that a household income is less than $50,000 or the proportion of the households with income less than $50,000 is represented by the dark area in the distribution, which is 0.20 approximately. In other words, if you pick a household at random, you have a 0.20 chance to bump into one with an income less than $50,000. This also means the chance of a randomly selected household having an income higher than $50,000 is around 0.80 (= 1- 0.2023)

Figure 4: Application of a normal distribution.



Income follows the normal distribution with mean of 75000 and standard deviation of 30000.

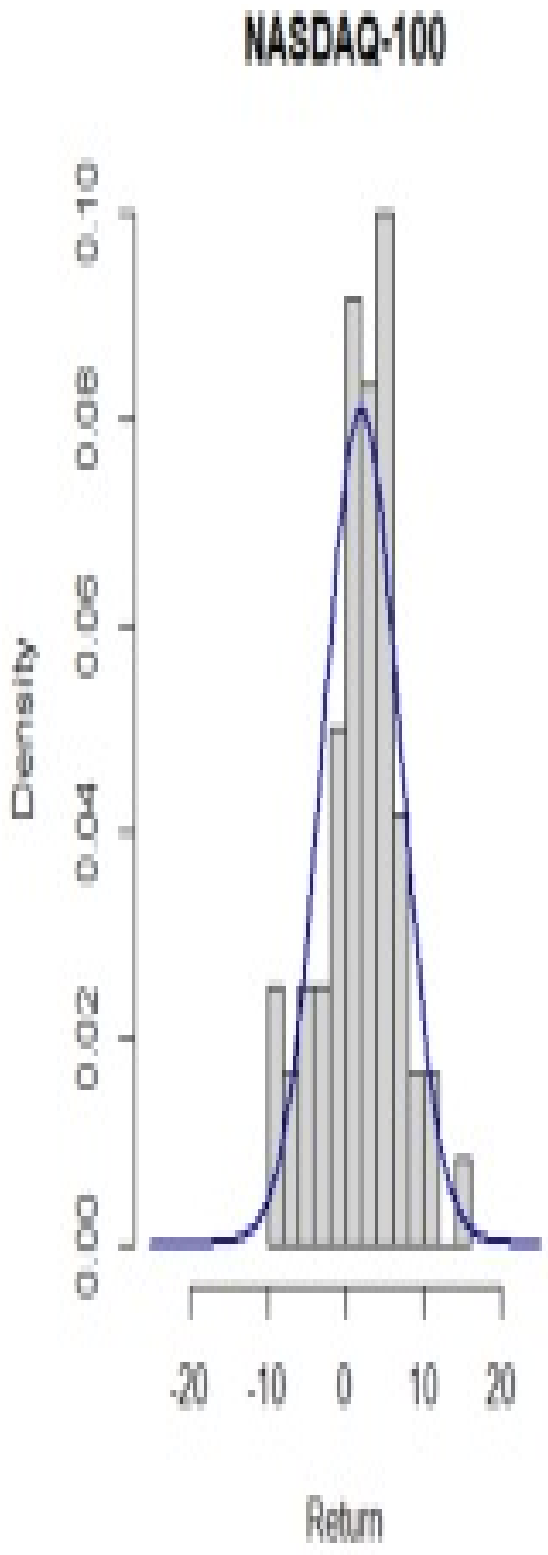
The probability that a randomly selected household has income less than 50000 is 0.2023.

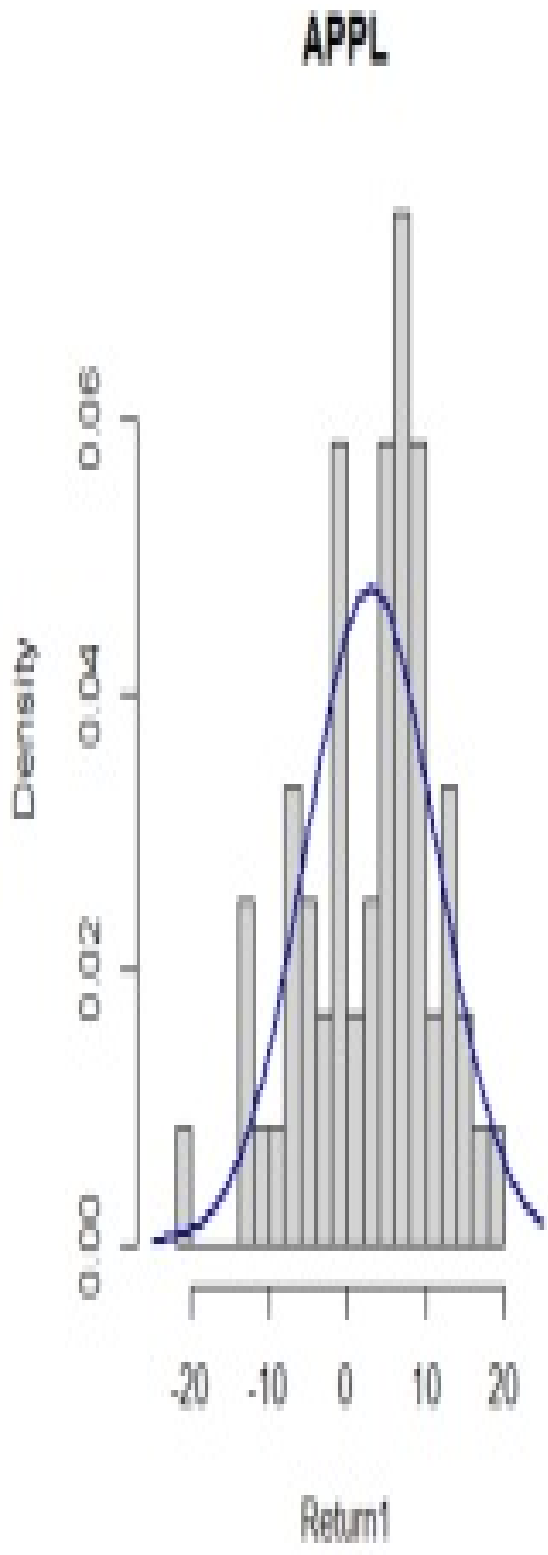
8. Checking the normality of a distribution.

Normal distribution is the most fundamental and popular distribution in statistics, and it is widely used as a “benchmark” distribution or as an “approximation” to the true distribution when it is unknown. Being a benchmark or approximation means it may be sometimes useful, but sometimes not, depending on the context and situation.

Figure 5 is the histogram we have seen in Figure 2, the returns from NASDAQ-100 investment, overlayed with the normal distribution with the same mean and standard deviation values of the returns. While the histogram shows a similar shape to the normal distribution, with near symmetry and bell curve, the fine details are not impressively consistent with the normal distribution. While an approximation by a normal distribution to a stock return distribution is sometimes used, it is generally accepted that a stock return distribution shows a clear departure from a normal distribution.

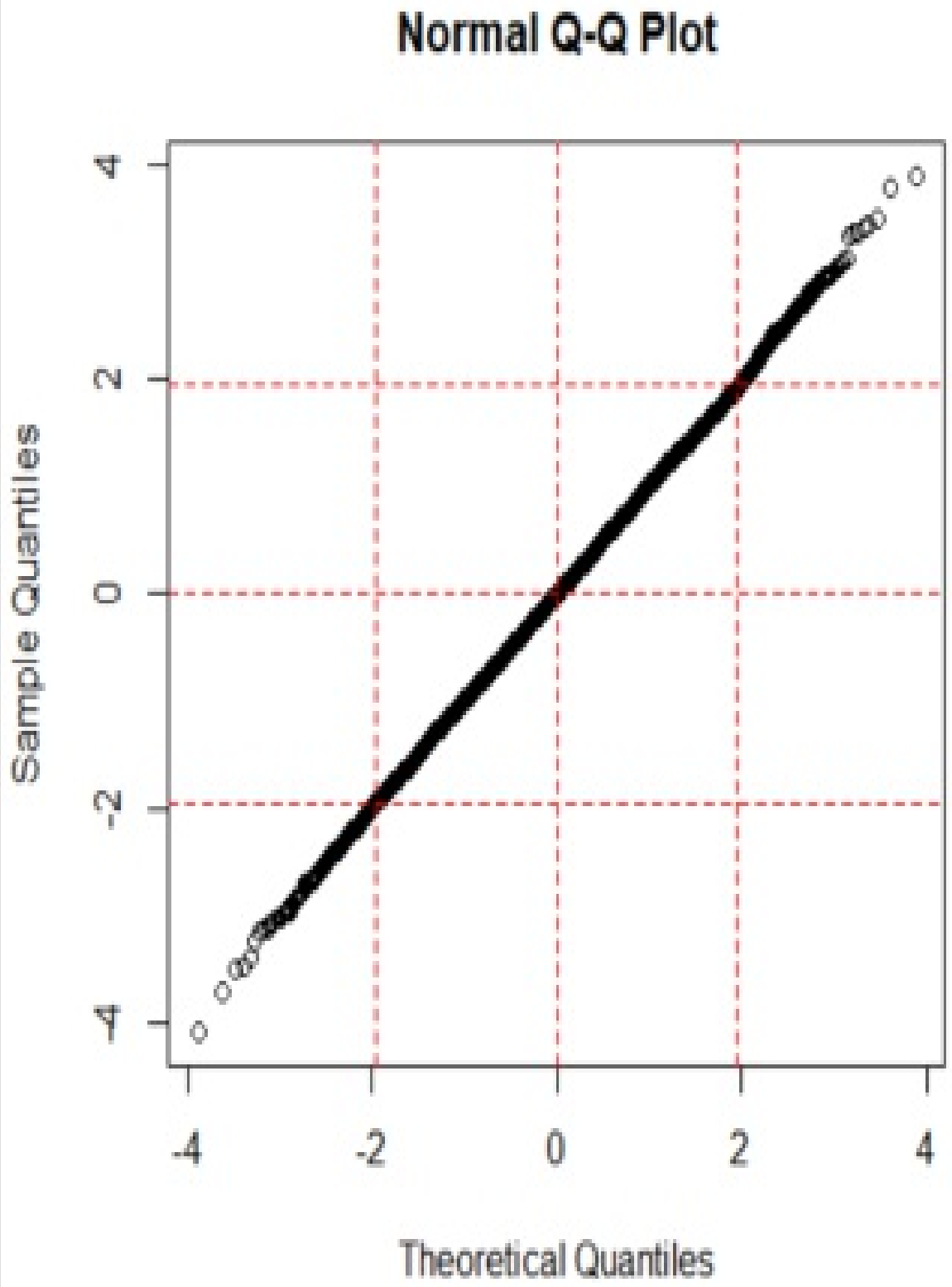
Figure 5: Histogram of the NASDAQ-100 and APPL returns and a normal curve


NASDAQ-100
Density
0.00
0.02
0.04
0.06
0.08
0.10
-20
-10
0
10
20
Return


APPL
Density
0.00
0.02
0.04
0.06
-20
-10
0
10
20
Return1

The Q-Q (quantile-quantile) plot provides a clearer way of checking the normality of a sample distribution using a graphical method. It connects the sample quantiles (or percentiles) with the (theoretical) quantiles from the normal distribution. If the sample follows the standard normal distribution, then its percentiles should match the percentiles from the normal distribution with the same mean and standard deviation. The 95th percentile from the sample distribution (which is normal) should match the 1.96, and the 50th percentile from the sample distribution should be 0, which is the 50th percentile from the normal distribution.

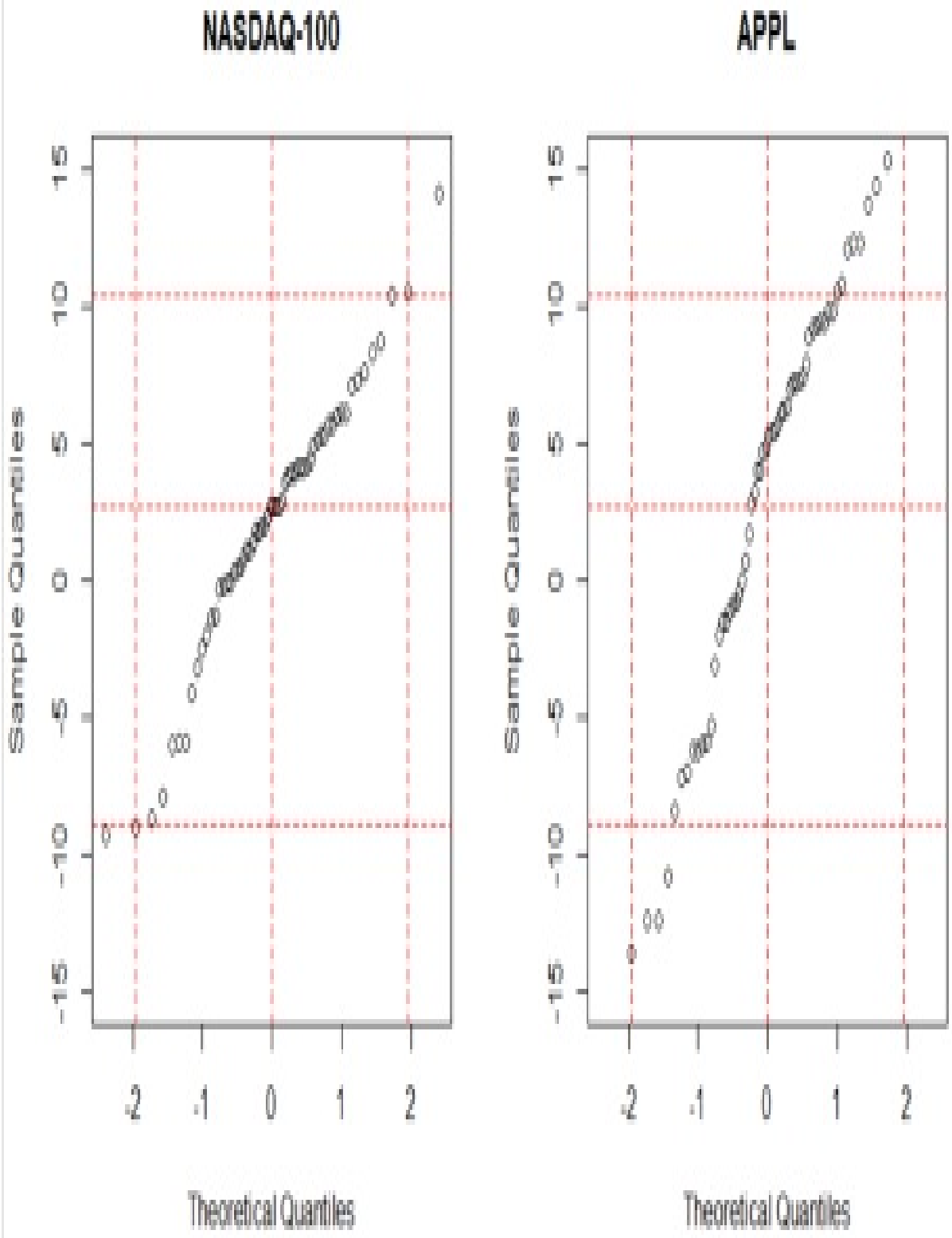
An example of the Q-Q plot is given here:


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
4
2
0
-2
-4
-4
-2
0
2
4

The grid lines are at (-1.96, 0, 1.96) for both axes, which are the 2.5th, 50th, and 97.5th percentiles from the standard normal distribution.

The y-axis (vertical) represents the sample quantile, and the x-axis (horizontal) represents the theoretical quantiles from the normal distribution. The grid lines are at (-1.96, 0, 1.96) for both axes, which match exactly. Hence, any sample that shows a Q-Q plot like the one above can be well approximated by a normal distribution.

Figure 6. Q-Q plots for NASDAQ-100 and APPL returns


NASDAQ-100
APPL
Sample Quantiles
Theoretical Quantiles
15
10
5
0
-5
-10
-15
-2
-1
0
1
2

The grid lines are at (-1.96, 0, 1.96) for both axes.

Figure 6 presents the Q-Q plots for the NASDAQ-100 and APPL returns. The return from the NASDAQ-100 return shows a reasonable match with the normal quantiles, while the quantiles of the APPL return show substantial departures from the normal quantiles. This indicates that, while the NASDAQ-100 returns may be approximated by a normal distribution with reasonable accuracy, a normal distribution will be a poor approximation to the APPL return distribution.

9. Concluding remarks

As an opening chapter, the basic concepts and descriptive measures of statistics were discussed with the following keywords:

Sample and population

Mean and Median

Standard deviation and Inter-quartile range

Percentile or quartiles

Histogram, Time plots, Q-Q plot, Box-Whisker plot

Normal distribution

If you understand the listed concepts and methods, and you can apply them to real-world situations, you already have made big steps into the world of statistical thinking! You can produce these statistics using popular tools such as Excel.

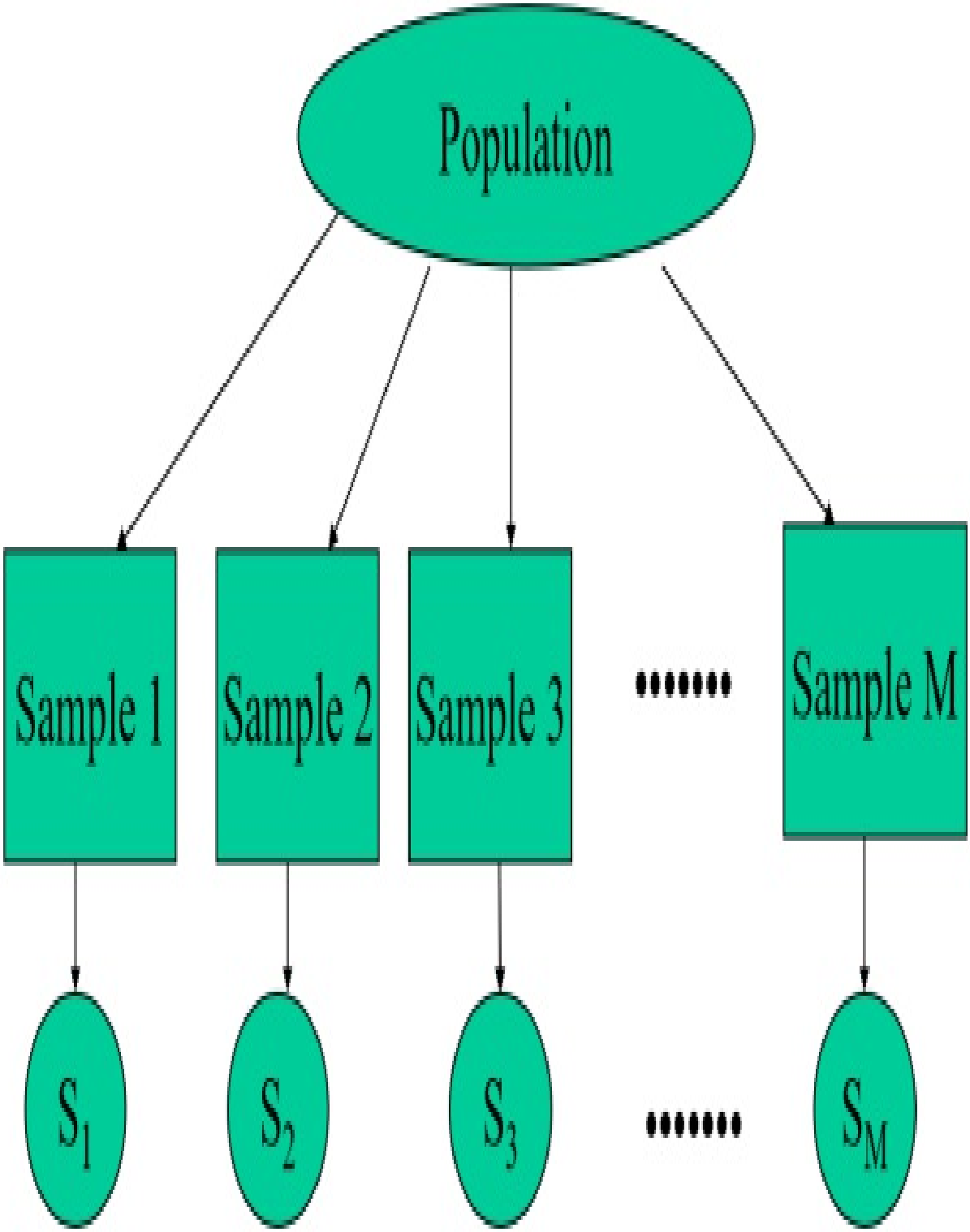
Chapter 2: Inferential statistics

This chapter is dedicated to inferential statistics. While descriptive statistics are used for illustration mostly, inferential statistics is mainly used as an aid to decision-making. It is also at the centre of the problems related to the misuse or abuse of the statistics we will see later. To understand the problems associated with modern statistics, it is essential to understand the elements and concepts of inferential statistics.

The concepts and methods covered include random sampling and the sampling distribution of a statistic, hypothesis testing, confidence interval, and p-value. Further in-depth discussion and problems with inferential statistics are discussed in the next chapter.

1. Random (Repeated) Sampling and Sampling Distribution

_____


Population
Sample 1
Sample 2
Sample 3
Sample M
S_1
S_2
S_3
S_M

Modern statistical methods we use are based on the idea of repeated random sampling from the population. A sample is a small subset of the entire population, and the researcher can conceptually take a large number of samples of the same size repeatedly (with replacement). Practically, however, repeated sampling can be costly and may not even be possible, and the researchers often have a single realization of the sample.

For the household example, the researcher can take the first sample with 1,000 households (say S1), and the second sample of the same size (S2), and so one. The process continues to SM where M is the number of repeated samples from the population. The process is illustrated in the above diagram.

From each realization of (S1, ..., SM), the researchers can calculate the sample statistics, such as the mean and standard deviation. Then, they would have M sets of these statistics, such as the sample mean, i.e., This collection of sample mean is called the sampling distribution of the sample mean. One can also have the sampling distribution of the standard deviation, if required, i.e., .

If the population follows a normal distribution, the sampling distribution of the sample mean also follows a normal distribution. This sampling distribution supports the inferential statistics widely used in statistical research, as we shall see in the next subsections.

Figure 7 presents the sampling distribution (in the form of histograms) of the sample mean, i.e., , when the sample size (n) increases from 10 to 1000 with M = 5,000. The population is assumed to follow a normal distribution with zero mean and standard deviation of 1. That is, for example, when n = 10, the sampling of 10 households is repeated 5000 times to collect their incomes. When n = 1000, the sampling of 1000 households is repeated 5000 times. Each histogram plots the distribution of , for the respective sample size from 10 to 1000, when M = 5,000.

The sampling distributions of the sample mean shrink around the population mean of 0, as the sample size increases. That is, as the sample size increases, the accuracy of the sample means as an estimate of the population mean gets higher. This is the fundamental property that modern statistical methods rely heavily on.

_____

Figure 7. Sampling distribution of sample mean ( ) when

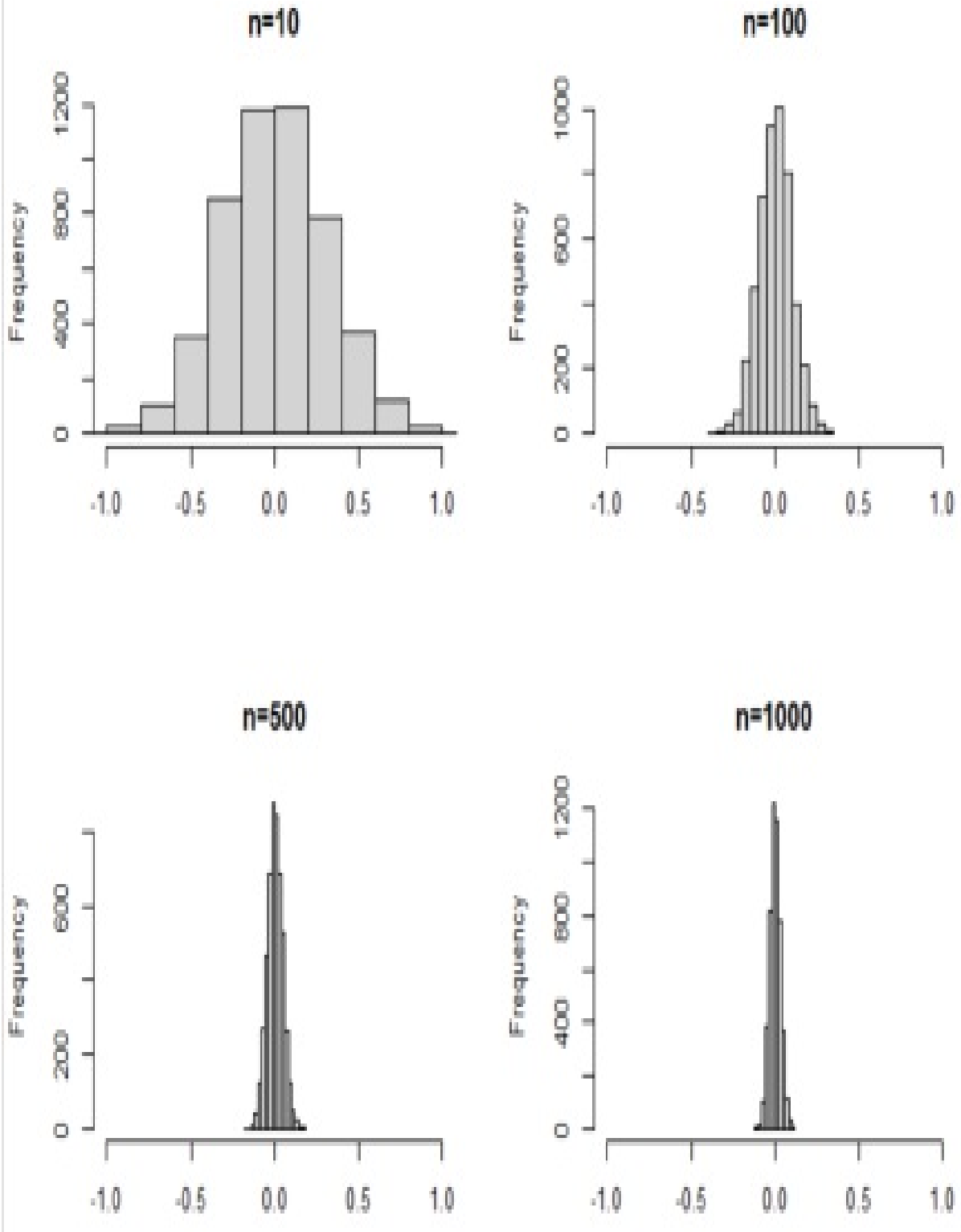
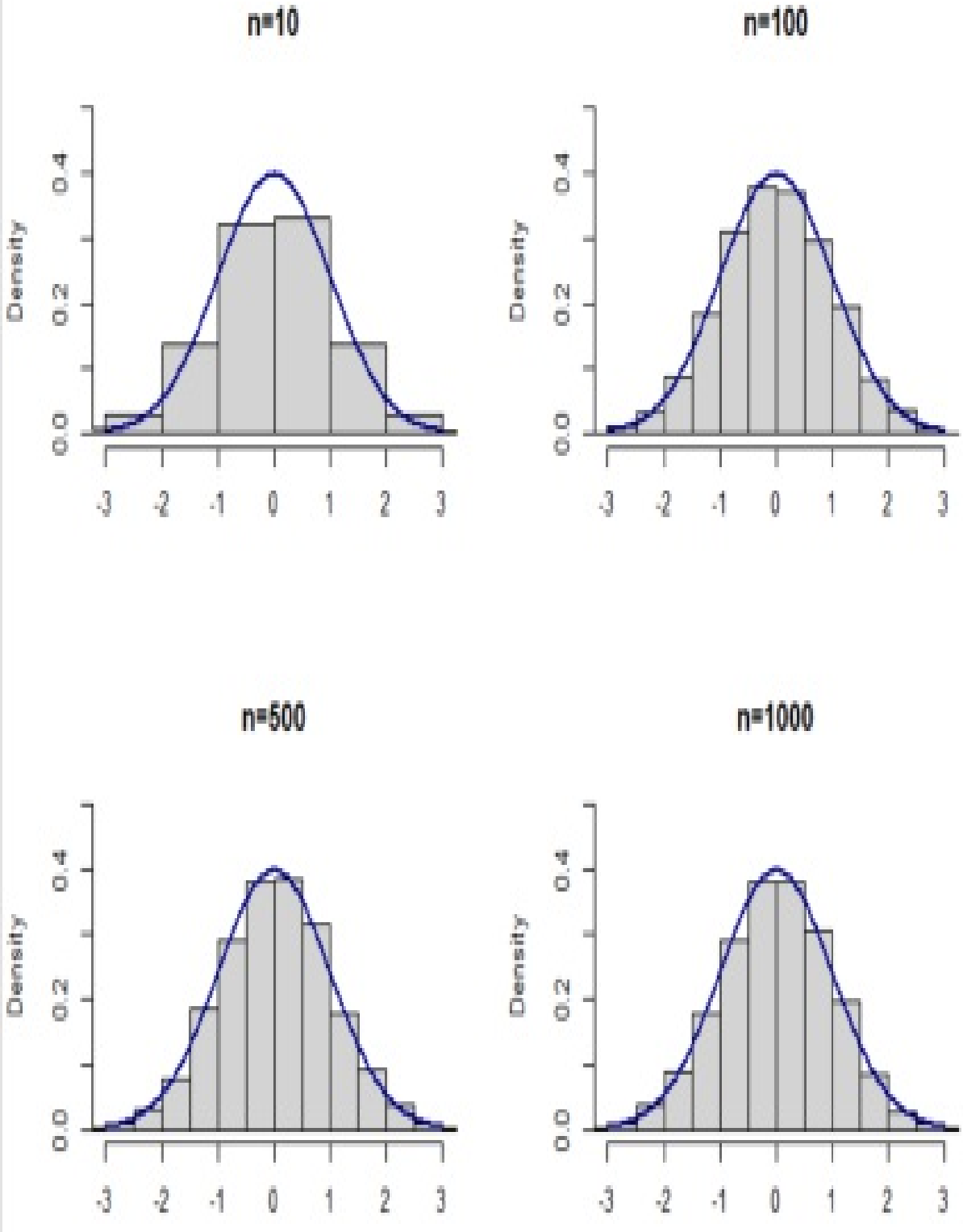

n=10
Frequency
0
400
800
1200
-1.0
-0.5
0.0
0.5
1.0
n=100
Frequency
0
200
600
1000
-1.0
-0.5
0.0
0.5
1.0
n=500
Frequency
0
200
600
-1.0
-0.5
0.0
0.5
1.0
n=1000
Frequency
0
400
800
1200
-1.0
-0.5
0.0
0.5
1.0

Figure 8: Sampling distribution of the Z statistic (


n=10
n=100
n=500
n=1000
Density
0.0
0.2
0.4
-3
-2
-1
0
1
2
3

_____

In fact, the sampling distribution is often presented and utilized by using a scaled or transformed version of . This scaling is called standardization, which can be achieved by this transformation:

This transformation, from to Z, can be achieved by taking the mean away from then dividing it by what is called the standard error Note that in our example in Figure 7, without loss of generality. This transformation is called standardization because the resulting distribution becomes the standard normal distribution N(0,1).

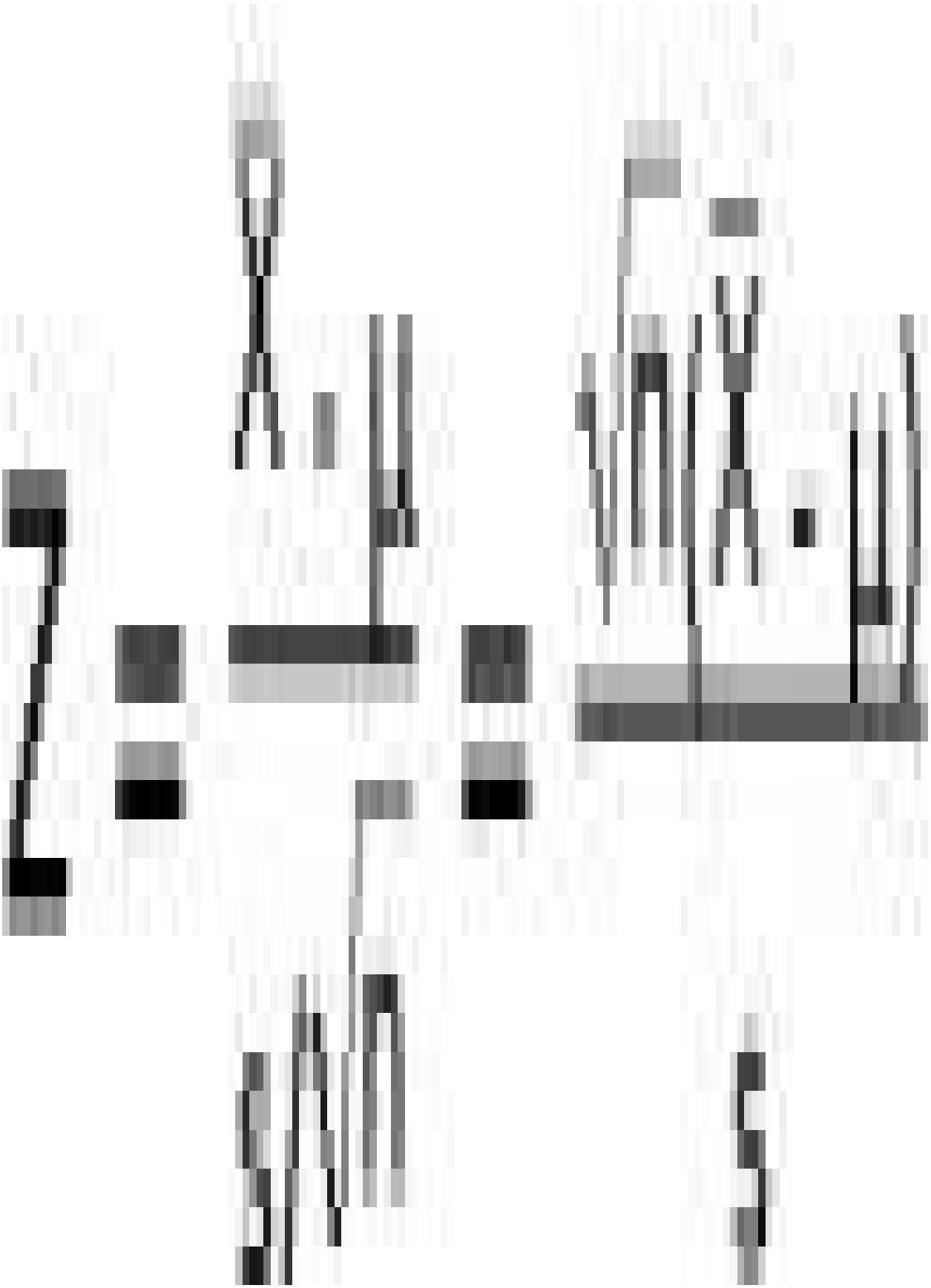
Figure 8 presents the sampling distributions in the form of Z. That is, (Z1, ..., ZM) values obtained from M set of samples (S1, ..., SM) are plotted (M=5000) when the sample size n increases from 10 to 1000. The distribution of shrinks into a single number of μ as the sample size increases as in Figure 7, but the distribution of (Z1, ..., ZM) does not shrink, but it follows the standard normal distribution N(0,1) for all sample sizes as in Figure 8. This is quite a nice property because we have full knowledge about the N(0,1) distribution as we have seen in the previous chapter. Not only its mean standard deviation values, but we also know the value of its percentiles and probabilities at all percentiles for any given value of μ and at any sample size. As a result, we can make some inferences about the value of μ, and this is what inferential statistics is about.

2. Understanding the test statistic.

In the previous section, it is demonstrated that Z statistics, a scaled version of the sample mean, follows the standard normal distribution. If we repeat random sampling from the population M times, (Z1, ..., ZM) follows the standard normal distribution, provided that the population follows a normal distribution. This property holds even if we do not know the value of μ and at all sample sizes we might have.

However, Figures 7 and 8 represent what is happening in a hypothetical world where repeated sampling M times is possible. Situations like this can exist only in statistics books. We have a single realization of S, a single realization of the sample mean, and a single realization of Z. Instead of thousands of realizations as we saw in Figures 7 and 8, we should work with only a single realization of (S, , Z). What we know is this, if we had repeated the sampling many times (such as M = 5000), all the Z values we'd have would follow a nice distribution, which is the standard normal. Based on this result, we should make an educated guess about the value of , the population mean, even if we have a single value of a single realization of (S, , Z) in reality. This is what we do in inferential statistics.

How do we do this? We should make use of the information in the Z statistic. Let us look closely at the Z statistics:



Signal: the second term of the numerator may be called a signal. It will tell us how much our sample information (differs from the population value μ.

Noise (s): this is the standard deviation of the sample, measuring the variability of the sample.

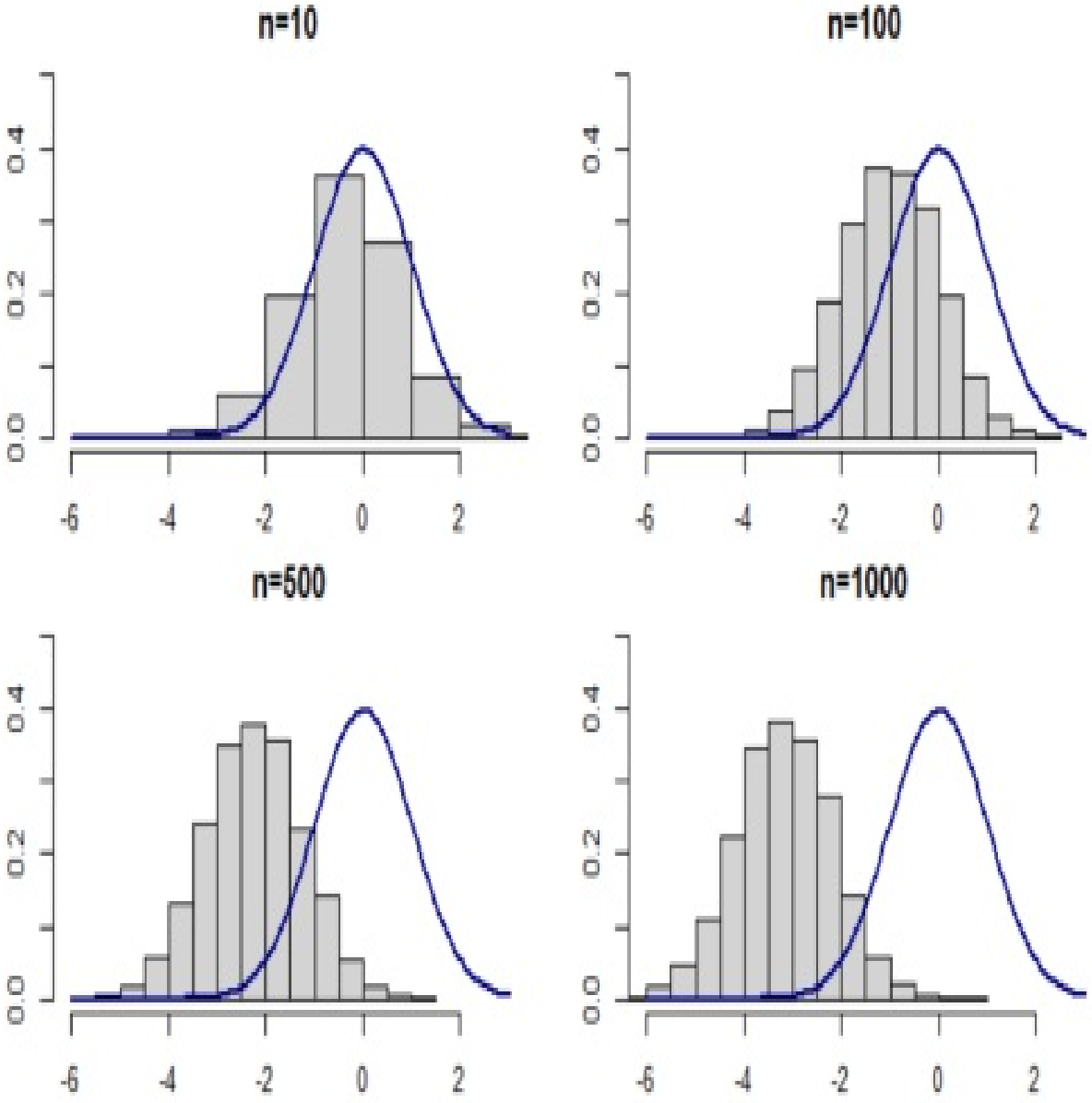
Scaling ( ): This is the first term of the numerator. This scaling is necessary because, without this adjustment, the distribution will shrink towards the value of μ, as the sample size increases, as we have seen in Figure 7.

Hence, the Z-statistic can be interpreted as a (scaled) signal-to-noise ratio: how much is the signal from the sample away from the value of μ per unit of noise. If this value is close to 0, it means the sample value is a good indicator for the population value of μ, when the sampling noise is considered. If this value is far away from 0, the sample value is a poor indicator for the population value of μ, when the sampling noise is considered.

Figure 8 is the case where the sampling distribution of Z is calculated with the known value of $\mu = 0$. Suppose our guess on μ was correct, then most of the Z values are concentrated around 0, and any value far away from 0 represents very unusual cases. In this case, we are likely to have Z values as the most typical and central values of the N(0,1) distribution and it is likely to say that the sample value is consistent with the value of $\mu = 0$.

Now suppose we incorrectly guess that $\mu = 0.1$. As in Figure 9, the sampling distribution deviates further away from the standard normal distribution as the sample size increases. Given the noise level, we have a stronger and stronger signal that the sample value differs from 0.1 as the sample size increases. In this case, most values of Z will differ from the most typical and central values of N(0,1) distribution, especially when the sample size is 500 or 1000. Hence, there is a high chance that we decide the population mean differs from 0.1.

Figure 9: Sampling distribution of the Z statistic (


n=10
n=100
n=500
n=1000
0.0
0.2
0.4
-6
-4
-2
0
2

3. Effect size vs. sample size.

As discussed in the previous section, the test statistic has three key elements as a signal-to-noise ratio scaled by the (square root of) sample size. Here, the most important element is the signal, and in statistics, it is also called the effect size, which is defined as a quantitative measure of the magnitude of the effect.

For example, a pharmaceutical company is measuring the effectiveness of a new drug. From the participants of the trial (sample), the researchers measure the proportion of those who showed the effect. If this proportion is large enough or the value of is convincingly larger than 0 (μ), then the researchers receive a clear signal about the clinical effectiveness of the new drug. Hence, in any statistical analysis, it is imperative to evaluate the effect size , and the key question is whether the signal is large enough practically.

Consider an economic policy designed to reduce the unemployment rate. Economists are interested in measuring its effect on the unemployment rate. Suppose they are not convinced that the effect size is not large enough economically: the policy may reduce the unemployment rate to a degree, but it may not be large enough to justify the cost of the policy. In this case, it is likely that the policy is abandoned.

The effect size or signal should be the key element of a statistical decision. A decision to implement a policy or to approve a new drug should be made based on the effect size or the signal.

As we have seen above, the test statistic does contain the signal in its formulation, but it is also affected by sample size and noise. Note it is sensitive to the value of sample size. A range of problems arises because the modern inferential statistical method is designed so the test statistic is used to make a decision, not the signal or effect size. Hence, when the sample size is large or massive, any small effect size or signal from data can be judged to be important (or statistically significant). This also means that any effect or

signal, albeit so small or practically negligible, can be made statistically important if the researcher has a large enough sample size. Any small effect can be inflated by sample size to produce a large value of test statistic. This is one of the critical problems of modern statistics, especially in the era of big data.

Many problems we are going to see in a later chapter, such as

the correlation vs. causation,

conflict between practical importance and statistical importance, and

misinterpretation of statistics

come down to this problem. Hence, a good statistical thinker should be able to distinguish the effect size from the test statistic and make a statistical decision based on effect size in combination with the test statistic. Sounds simple, but many top-level professional statisticians often fail to do so.

4. Inferential statistics.

Utilizing a range of summary statistics and visualization tools, descriptive statistics provides the researchers with information or impressions about the key characteristics of the sample. While the information gathered from descriptive statistics is highly suggestive of the features of the population, it is not a direct statement about the population features. Inferential statistics allows you to test a hypothesis about the population parameters or assess whether your sample is compatible with the features of the population.

For example, is the claim that the monthly average return from the NASDAQ-100 investment higher than 2% in the population supported by the sample? Is the claim that the mean household income in California is equal (or close) to $75,000 in the population supported by the sample? These questions can be more decisively answered using inferential statistics in a more systematic and objective way.

There are two alternative but equivalent ways of inferential statistics: hypothesis testing and confidence interval.

Hypothesis testing.

The main elements of hypothesis testing are:

Null hypothesis

Alternative hypothesis

Test statistics and its distribution under H0

Level of significance and critical values.

With hypothesis testing, two hypotheses are proposed: the null hypostasis (H0) and an alternative hypothesis (H1). The null hypothesis represents “a belief we will maintain until we are convinced by the sample evidence that it is not true,” while the alternative is the hypothesis we accept if the null hypothesis is not supported by the sample. These hypotheses are formulated for the population parameters such as the mean and standard deviation.

The test statistic carries the information about the sample evidence, and it measures the distance of the sample statistics from the value of the population parameter being tested, with appropriate scaling. The statistic is assessed against its sampling distribution under the null hypothesis. If it is too far away from the value of the population parameter under H0 (null hypothesis), then H0 is rejected.

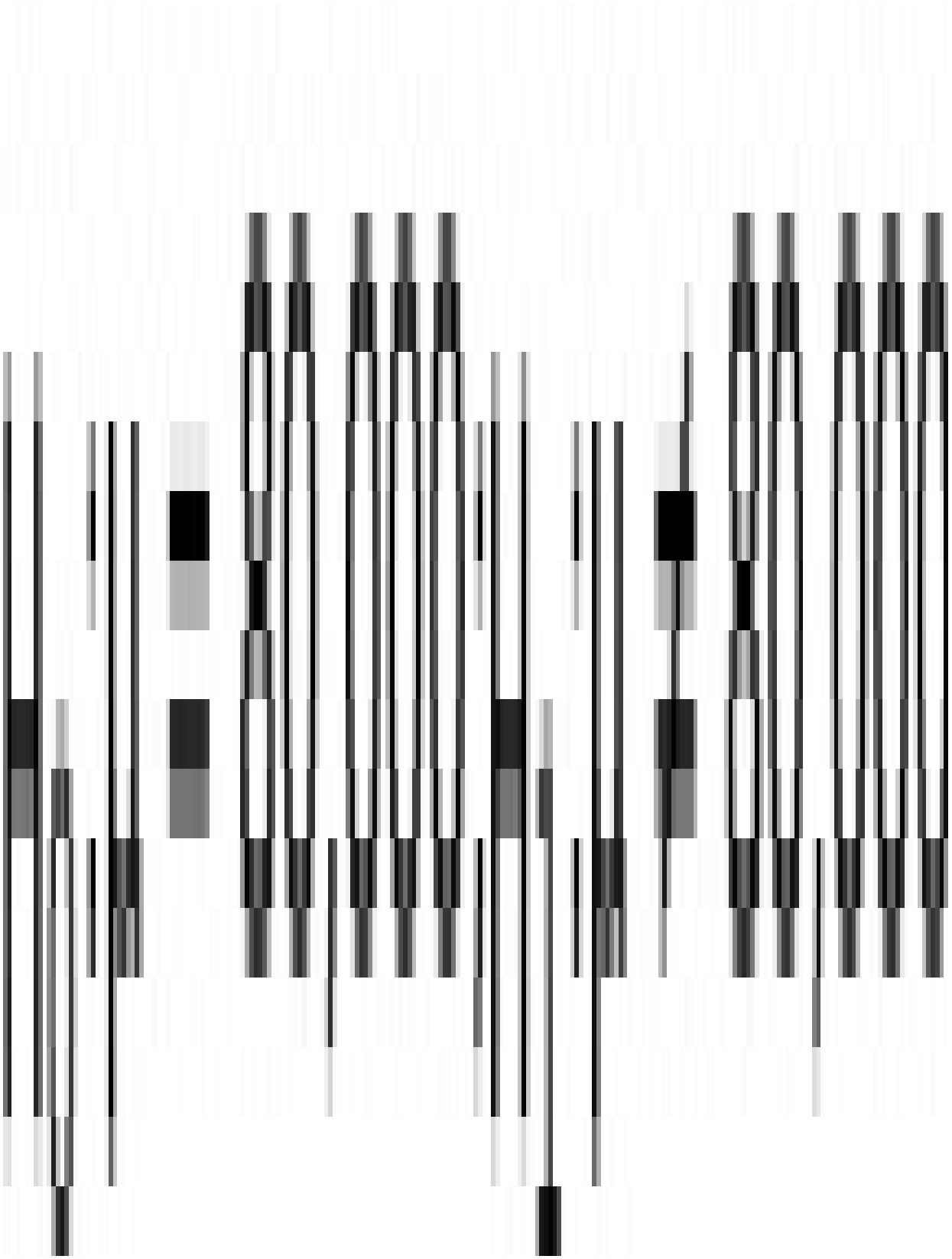
How far is too far for the test statistic from the value of the population parameter under H0 to reject H0? Sampling distribution of the test statistic is used to obtain the critical values. The critical values are determined by the level of significance, which represents the probability of rejecting the true null hypothesis (Type I error). This probability (denoted α) is conventionally set at 5% (sometimes 1% or 10%).

If the test statistics is beyond these critical values, then the null hypothesis is rejected; if not, the null hypothesis is maintained at the significance level of α. We know this may be challenging to follow, so let’s look at a practical example.

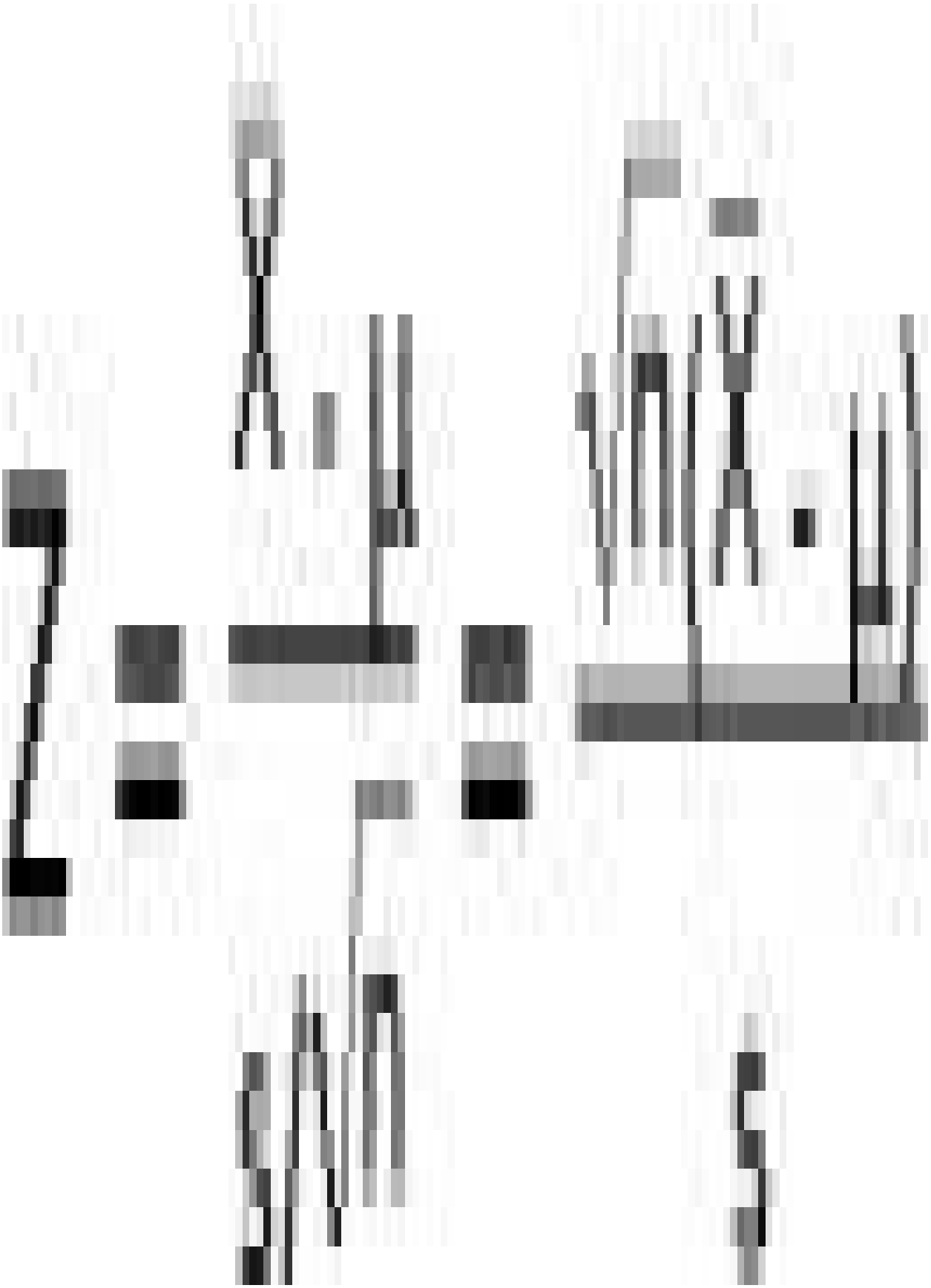
Example 1

Consider the case of household income distribution in California. Suppose the population follows a normal distribution with an unknown mean and a standard deviation of μ and σ. Let’s assume economists believe the mean income is around $80,000, and they want to test whether the hypothesis is

supported by the sample. A random sample of 1,000 households is taken to find that the sample mean is $75,000 with a standard deviation of $60,000. Then the following null and alternative hypotheses can be tested:



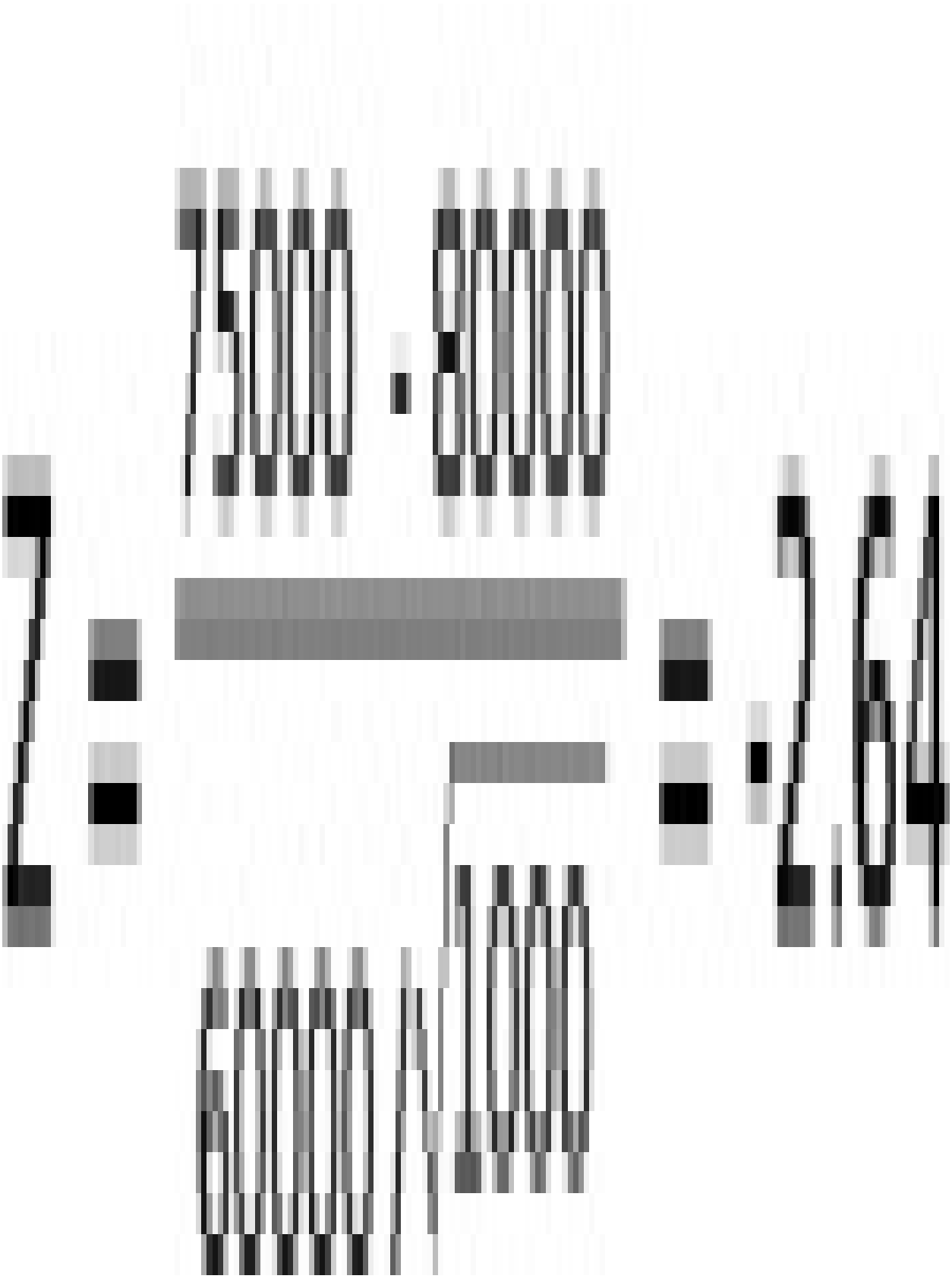
The test statistic is given by,

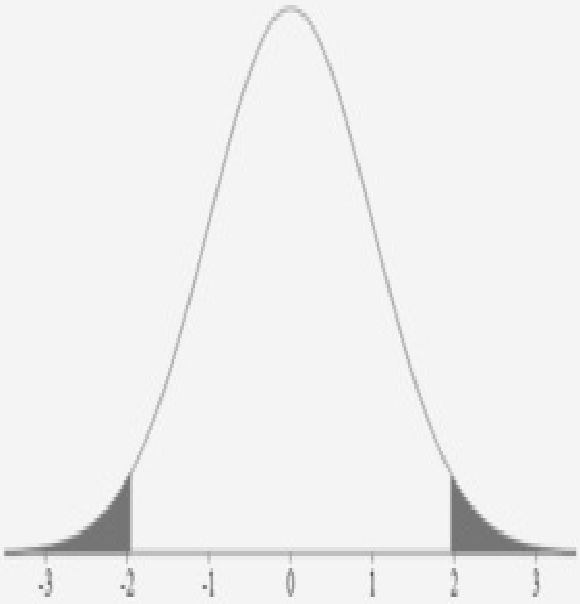


The Z statistic above measures the signal, the distance of the sample mean ( ) from the value of the population mean ( ), divided by the noise (s) and scaled by the sample size ( ), while is also called the standard error. While we have seen in Figure 7 that the sampling distribution of shrinks to the population mean value, this scaling makes the Z statistic follow the standard normal distribution [N(0,1)] in repeated sampling. That is, the sampling distribution of Z is standard normal distribution, which is illustrated in Figure 8.

The next step is to calculate the value of Z from the sample and check how large the sample value is compared to the critical values from N(0,1).

The sample value is calculated as,





Specify Parameters:

Mean 0
SD 1

- Above 1.96
- Below 1.96
- Between -1.96 and 1.96
- Outside -1.96 and 1.96

Results:
Area (probability) = 0.05
Recalculate

This test is called a two-tailed test because the critical region is on both tails of the distribution, reflecting the structure of H1.

The critical values with a level of significance 5% is -1.96 and 1.96. Hence, the null hypothesis is rejected at the 5% level of significance because it is less than -1.96. The sample does not support the claim or hypothesis that the population mean income is equal to $80,000.

_____

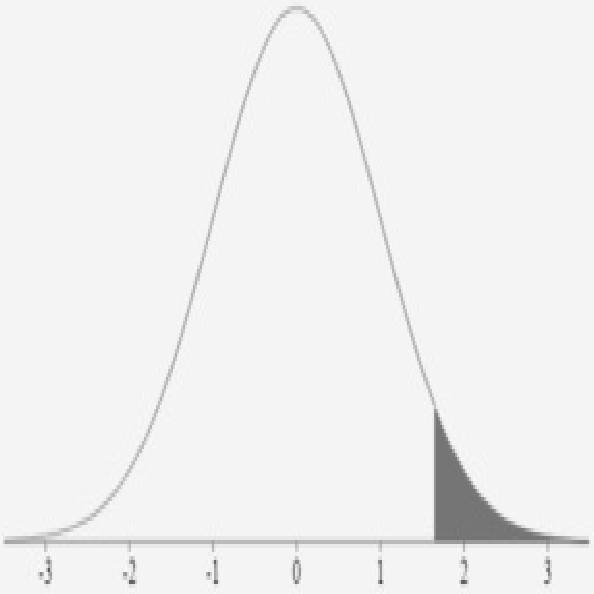
Example 2

Suppose the investor of the NASDAQ-100 portfolio wishes to test for the following hypothesis:

.

Test the null hypothesis where the investment return is equal to 2% against the alternative that it is greater than 2%. Recall that the mean value of the monthly return is 2.02% and the sample standard deviation is 4.92% from a sample of 60 monthly returns (n=60). Assuming a random sample from the population that follows a normal distribution,

=0.03.



Specify Parameters:

Mean 0
SD 1

Above 1.645
Below 1.96
Between -1.96 and 1.96
Outside -1.96 and 1.96

Results:
Area (probability) = 0.05
Recalculate

This test is called a one-tailed test because we consider only the one side of the distribution due to the structure of H1.

Since the alternative hypothesis is concerned only with the "greater" than 2, we only need to consider the upper side of the distribution. The null hypothesis is rejected if the test statistics is sufficiently larger than 0. The critical value, in this case, is 1.645 at the 5% level of significance.

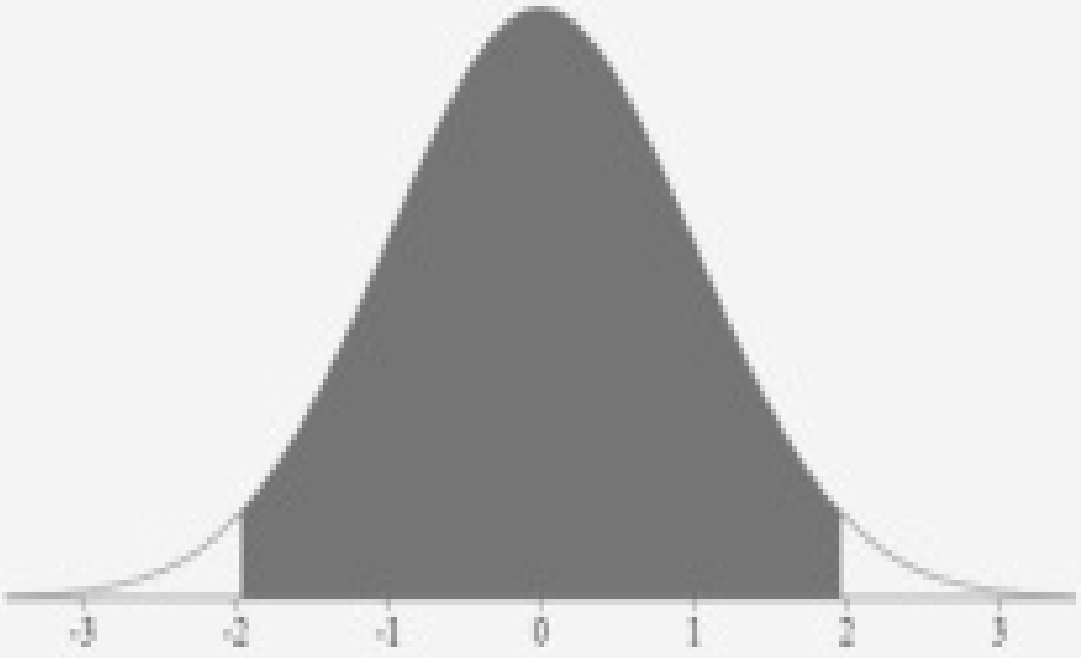
_____

Since the value of Z statistics is less than 1.645, the sample does support the null hypothesis that it is equal to or around 2%, and there is no evidence to suggest it is greater.

_____

Confidence interval.

_____



Specify Parameters:

Mean: 0

SD: 1

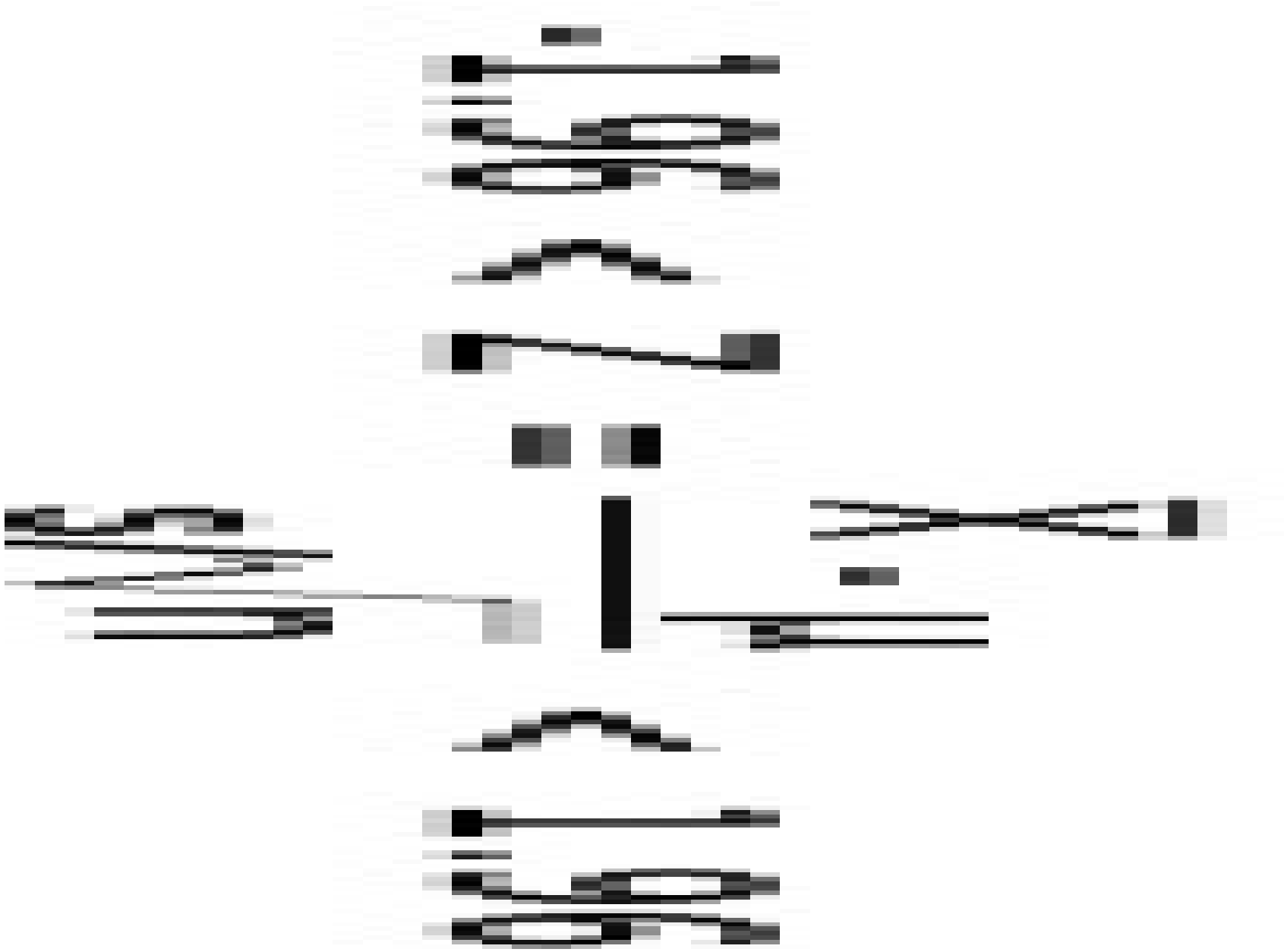
○	Above	1.96		
○	Below	1.96		
◉	Between	-1.96	and	1.96
○	Outside	-1.96	and	1.96

Results:

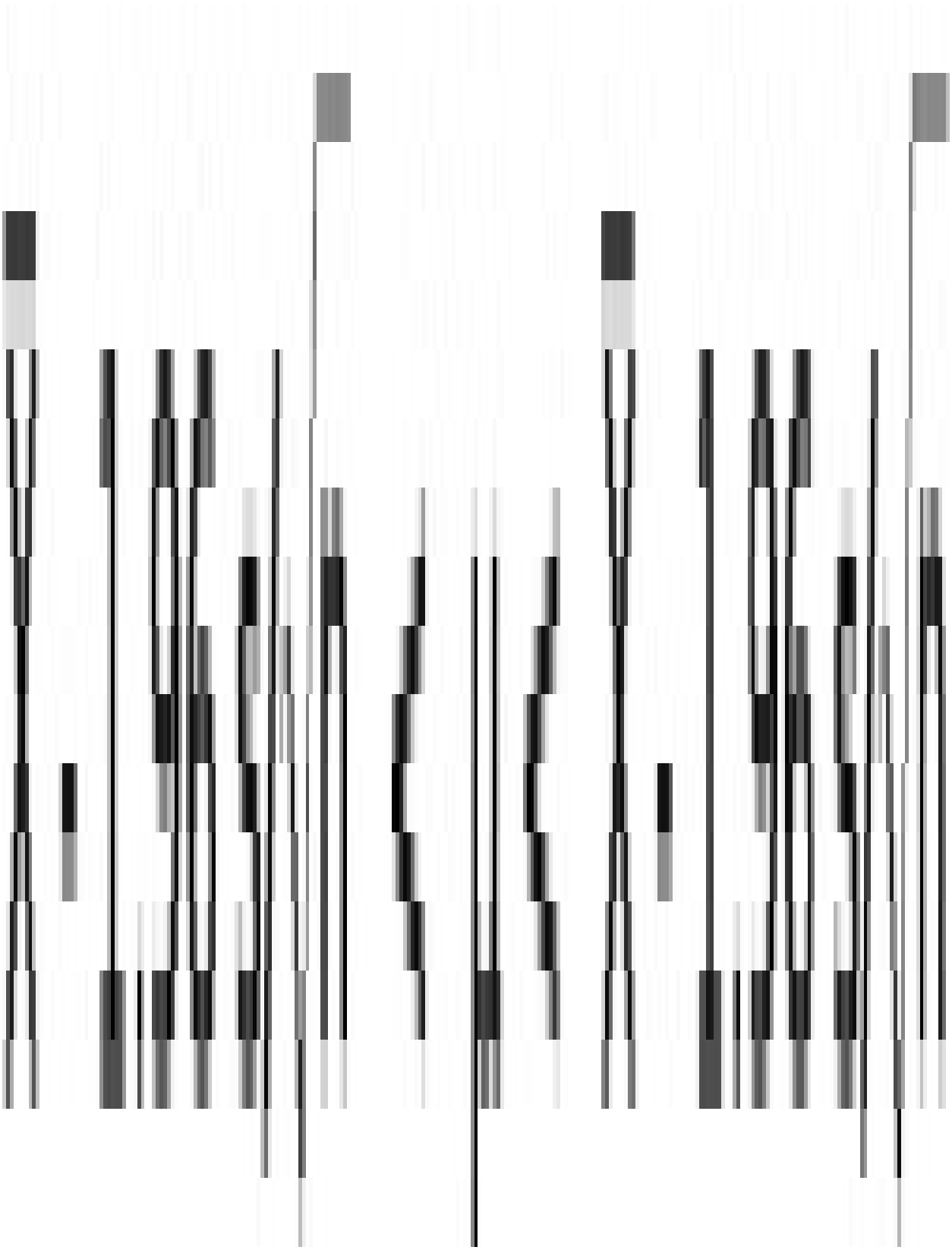
Area (probability) = 0.95

Recalculate

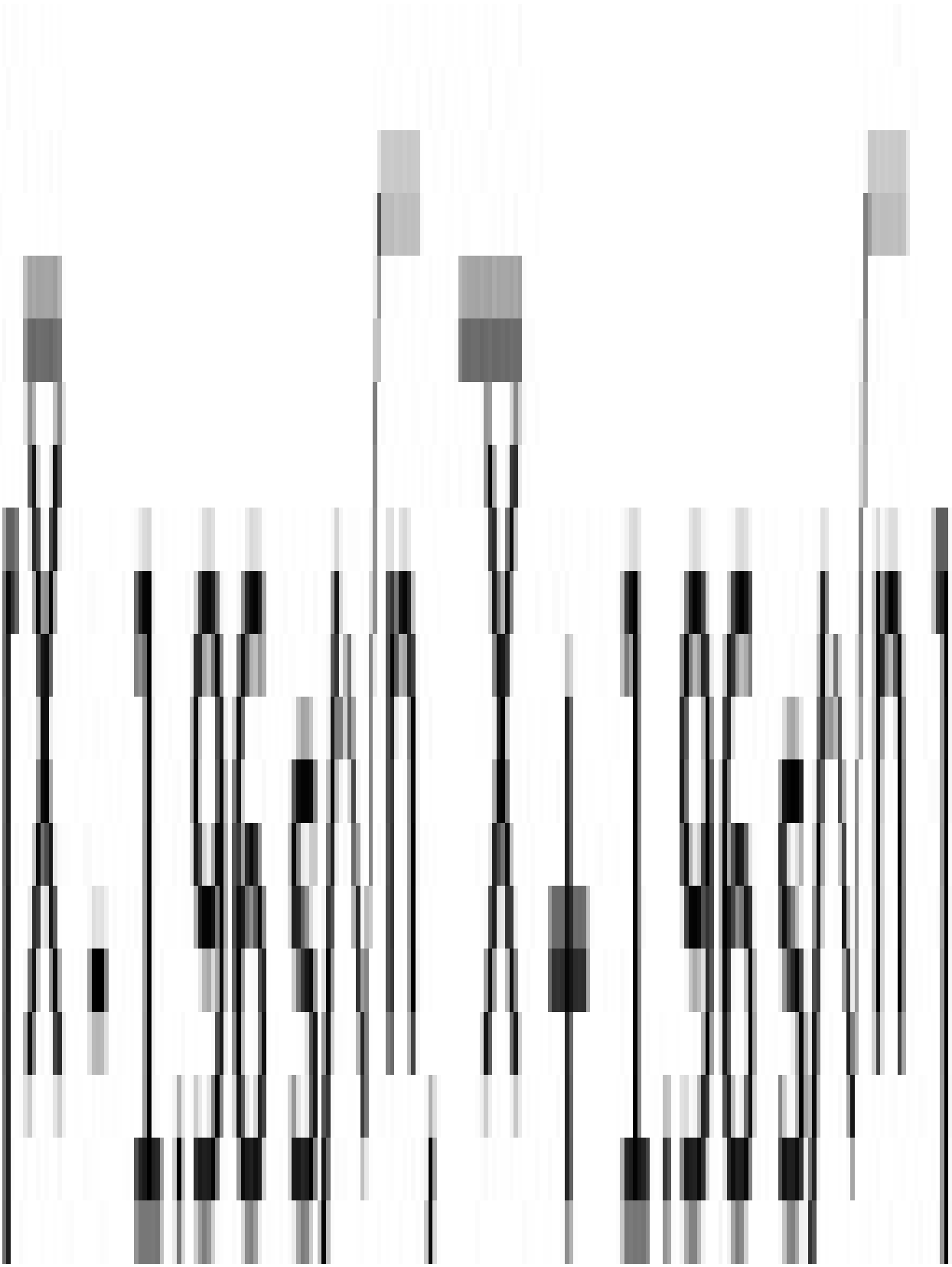
The above probability from the standard normal distribution tells us the probability of the Z value between -1.96 and 1.96 is 0.95. That is, the chance of Z being



is 0.95. The above expression can be re-arranged for μ, which can be written as

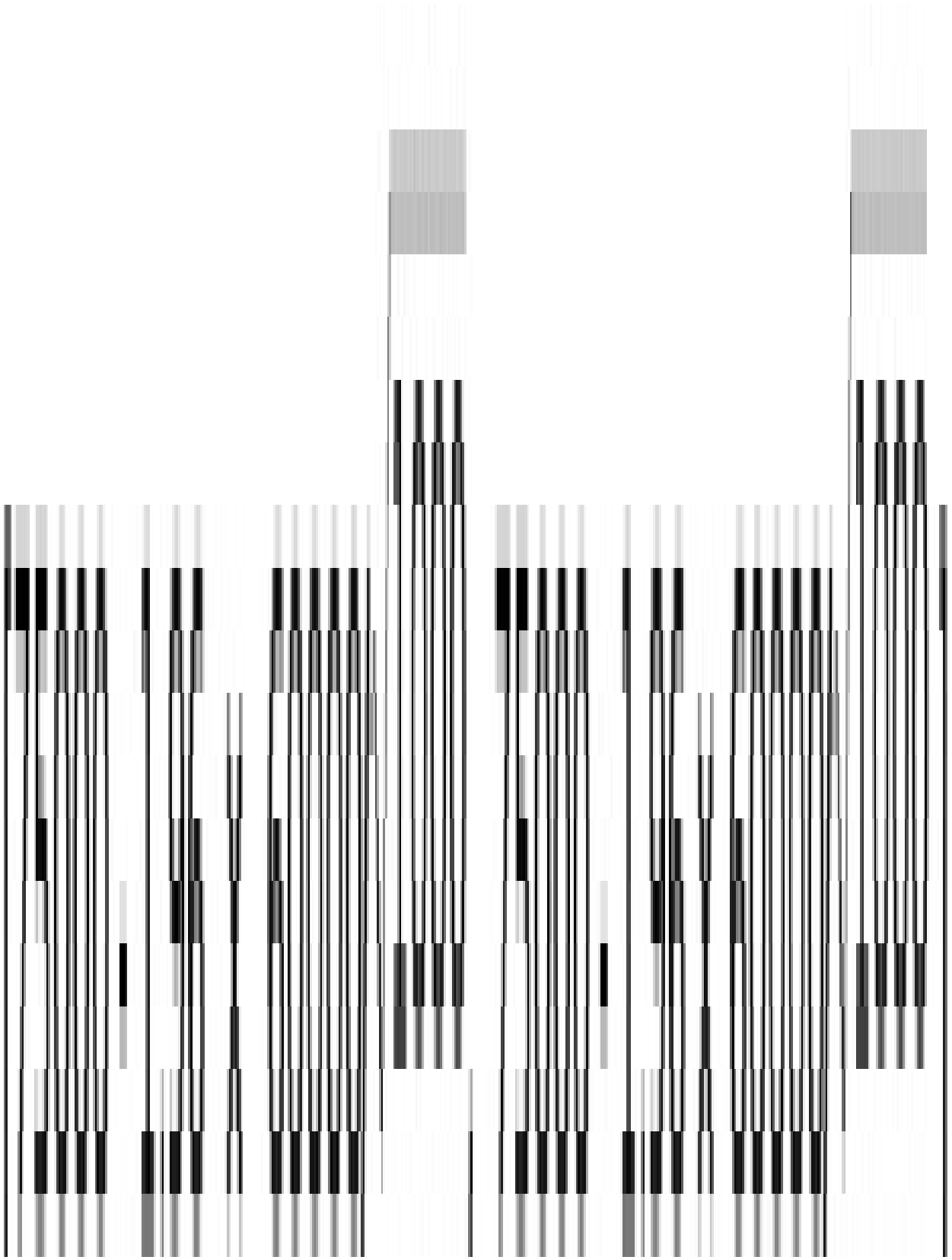


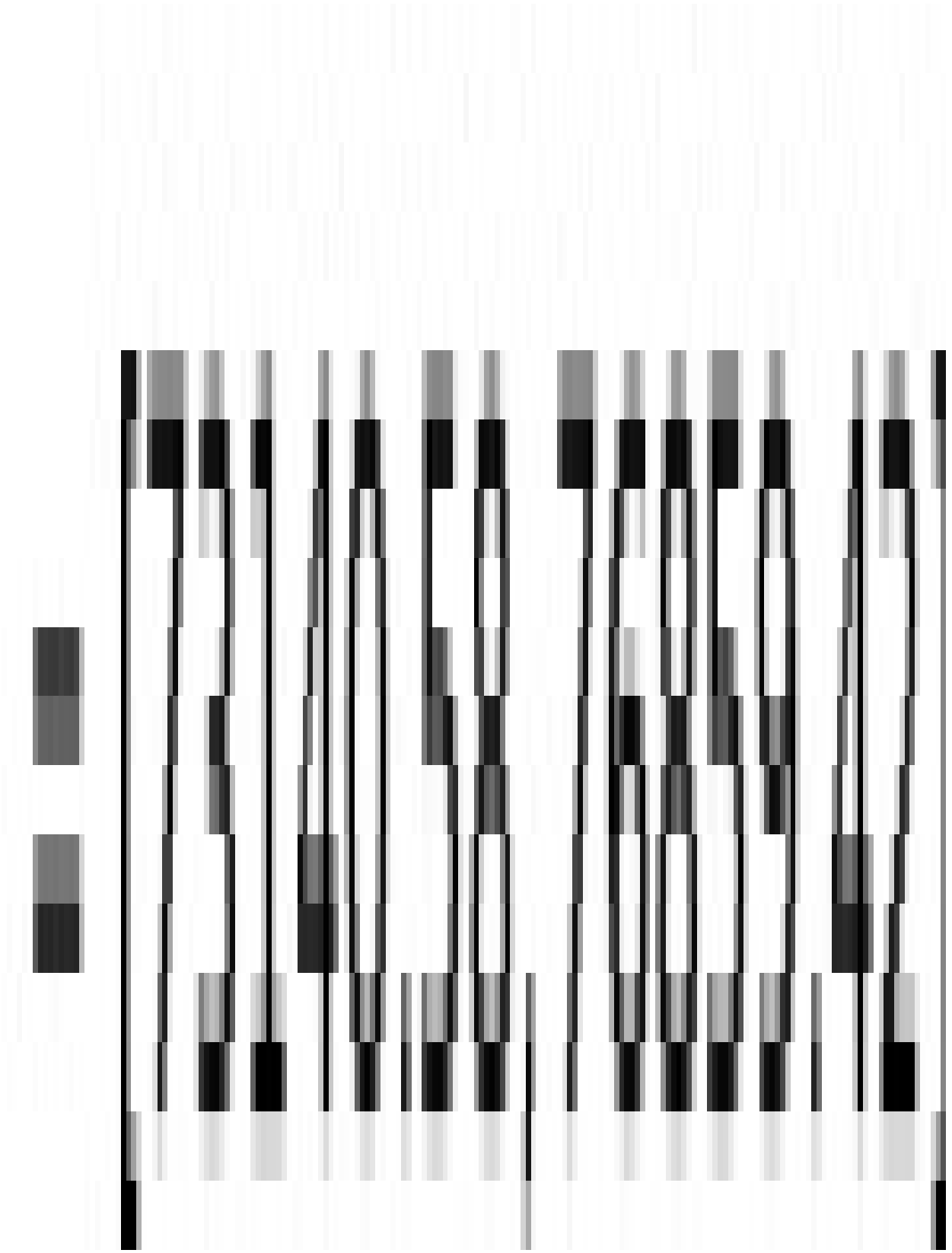
This means the chance of the population value of μ covered by the interval of



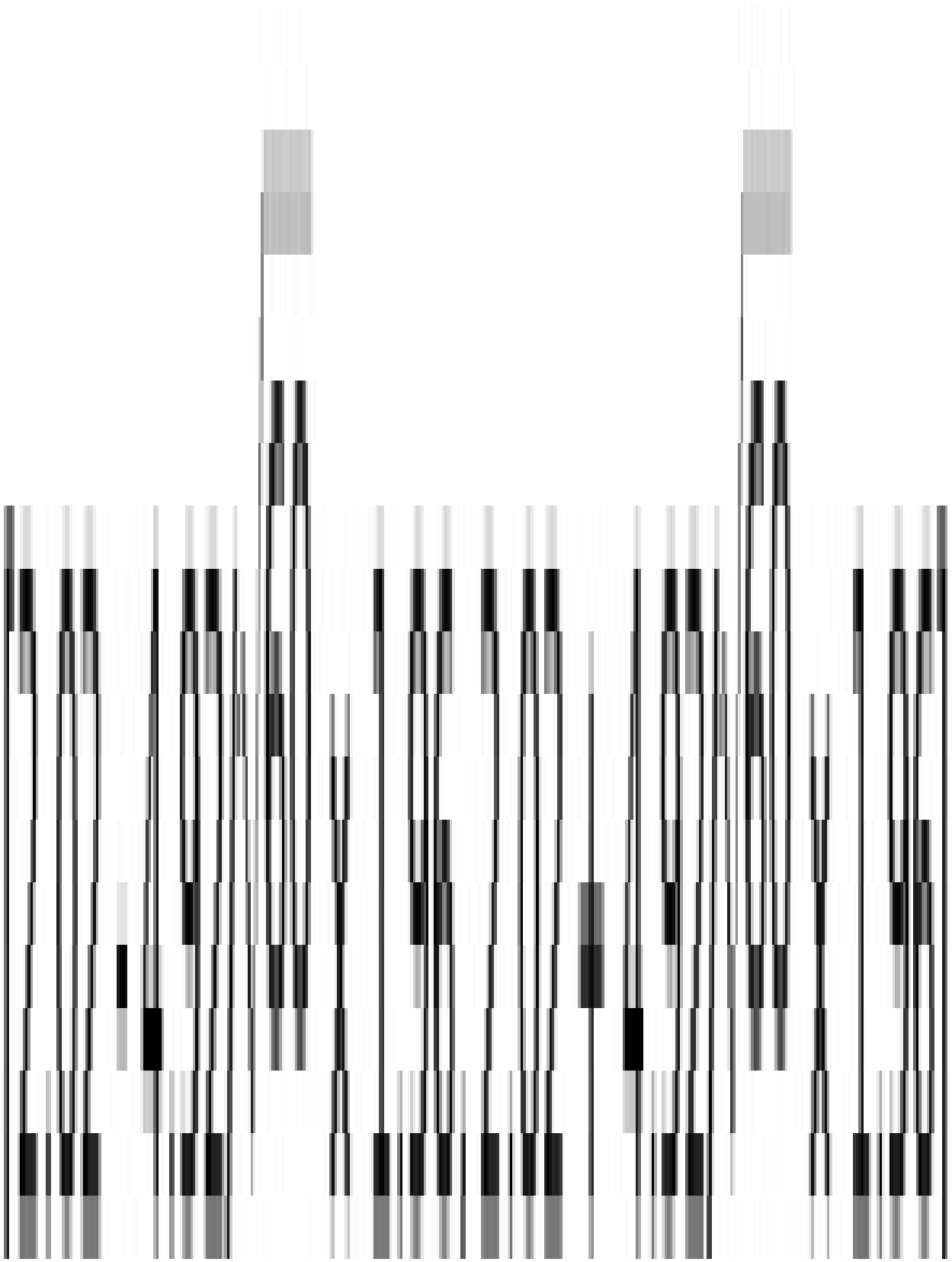
is 0.95. This interval is defined as an interval that covers the true value of the population parameter with 95% confidence in repeated sampling.

In our example for household income, the 95% confidence interval for the population mean is





That is, we are 95% confident in repeated sampling that the population mean household income in California is covered by this interval. For the investment example, the 95% confidence interval for the mean return is,



= [0.76,3.28],

which covers the true mean value with 95% confidence in repeated sampling.

The advantage of the confidence interval over hypothesis testing is that it gives an impression about the possible location of the population parameter with a level of confidence. A wide interval is not informative about the location of the population value, representing a high degree of estimation variability, while a tight interval is informative about the location of the population value and indicative of a high degree of estimation accuracy. A tighter interval is informative, and the more likely our estimations are correct and vice versa.

For the case of the Californian mean income (Example 1), the 95% confidence interval does not cover the value of $80,000. We are 95% confident the population mean value of $80,000 is not supported by the sample. This feature is consistent with the rejection of at the 5% level of significance. Similarly, for the investment example (Example 2), the 95% confidence interval does cover the value of $\mu = 2\%$, and we are 95% confident this value of the population mean return is supported by the sample, which is consistent with the acceptance of at the 5% level of significance. Hence, there is a close connection between confidence interval and hypotheses testing.

p-value.

The p-value is another indicator for statistical evidence as to whether the null hypothesis is rejected, given a level of significance. Although it is equivalent to the decision based on the critical values we have seen above, it is widely used due to its simplicity and universality. The p-value can be computed using any statistical package (such as Excel) for nearly any statistical test, and the decision can be made simply by comparing it with the level of significance.

To state the decision rule first, the null hypothesis is rejected at a level of significance (say, α) if

p-value $< \alpha$.

Otherwise, the null hypothesis cannot be rejected at the α level of significance.

It is conventional to choose the level of significance α at 0.05, or 0.10, or 0.01. But there will be more discussions about this later in this book. The attraction to the p-value is that the researcher doesn't need to look for the critical values for the test as in the standard normal distribution. Instead, the decision to reject or not reject the null hypothesis can be made almost immediately by comparing the p-value with the level of significance α.

The p-value is defined as the probability of observing a test statistic as extreme or more extreme than its observed value under the null hypothesis and can be interpreted as a measure of how incompatible the sample is with the null hypothesis.[iii]

Specify Parameters:

Mean	0
SD	1

○	Above	1.96		
○	Below	1.96		
○	Between	-1.96	and	1.96
◉	Outside	-2.64	and	2.64

Results:

Area (probability) = 0.0083

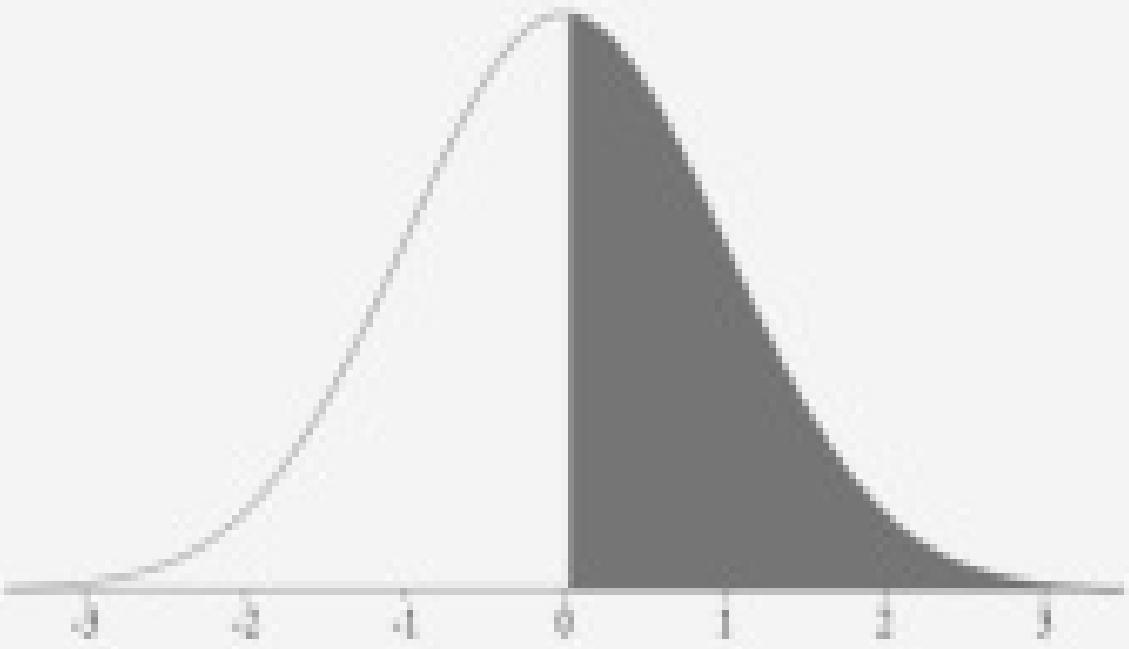
Recalculate

In Example 1 for the household income in California with

,

the value of Z-statistic was -2.64. The probability of observing a test statistic as extreme or more extreme than its observed value under the null hypothesis is 0.0083, as the above figure shows. The sample is highly unlikely to be compatible with the null hypothesis with the probability of 0.0083. Since this value is less than 0.05, we reject the null hypothesis at the 5% level of significance, consistent with what we have seen in hypothesis testing using the critical values.

_____



Specify Parameters:

Mean: 0

SD: 1

- (•) Above: 0.03
- ( ) Below: 1.96
- ( ) Between: -1.96 and 1.96
- ( ) Outside: -2.84 and 2.84

Results:

Area (probability) = 0.488

Recalculate

In Example 2, the investor with a NASDAQ-100 portfolio wished to test for the following hypothesis:

.

The Z-statistic was 0.03 and the p-value 0.488., so the sample is likely to be compatible with the null hypothesis with the probability of 0.488. Since this value is greater than 0.05, we cannot reject the null hypothesis at the 5% level of significance, consistent with what we have seen in hypothesis testing using the critical values.

_____

Here, we re-state the decision rule based on the p-value as this:

Reject H0 at the α-level of significance if p-value < α.

This decision rule is called the p-value criterion, and the decision based on it is equivalent to the method based on critical values. The p-value is widely used because of its universality. Any statistical packages or programs, such as Excel, provide the p-value automatically for any statistical test. Hence, the researcher needs to compare their p-value with the chosen level of significance, without having to worry about the critical values and the distribution they are obtained from. Due to its simplicity, the p-value is widely reported in many statistical reports or articles but sometimes misused or abused. For example, a low p-value can tell us the sample is not compatible with the null hypothesis, but it does not necessarily mean the effect or signal is practically large. However, it is often the case that a low p-value is misinterpreted as the indication of strong effect. The readers should be reminded that a low p-value can be associated with practically negligible effect or signal.

5. Concluding Remarks.

This chapter has presented the basic elements of a modern statistical method of inferential statistics. The concepts of random sampling and sampling distributions are the fundamental concepts that support the widely used (or abused) statistical methods, such as hypothesis testing, p-value, and confidence intervals. Many decisions in business and government are made based on these methods, and many scholarly articles are published based on the outcomes of these inferential statistics. If you understand these concepts and methods, you may use popular tools, such as Excel, to calculate these inferential statistics for your data.

However, as briefly mentioned in this chapter, these methods have deficiencies and can cause a range of problems. The test statistic and p-value contain the signal from sample or effect size in their formulation, but they are often masked by a large sample size. There is a danger that statistical decisions can be made solely based on test statistic and p-value, ignoring the effect size or signal from sample. Any effect, albeit so negligible practically, can be found to be statistically important, especially when the sample size is large. Another problem is arbitrariness of the conventional level of significance ($\alpha = 0.05$, 0.01, or 0.10), which can lead to further problems. These issues will be discussed in the next chapter.

Chapter 3: Statistical Thinking

In the last two chapters, we have reviewed the basic concepts of statistics and the methods of descriptive statistics and inferential statistics. In this chapter, we evaluate what decision-makers should consider and how they should control different elements of hypothesis testing for sound statistical thinking under uncertainty. We also discuss the problems associated with the contemporary methods of hypothesis testing used in practical applications and professional research. The problems include,

the researchers do not fully consider the statistical uncertainty related with Type I and II errors,

they often ignore the effect size or the signal from the sample, and

they set the decision threshold arbitrarily with no scientific or practical justifications.

This is because the teaching of contemporary statistics does not take the issues above seriously for some unknown reason. The issues are not covered in detail in our textbooks or lectures of modern statistics. As we shall see later in this book, this has caused many problems in modern statistical research, such as replication crisis, data-snooping bias, and publication bias.

1. Understanding uncertainty.

One key to sound statistical thinking is to understand the degree of uncertainty involved in hypothesis testing. Like any other decision-making, statistical outcomes are uncertain, and the decisions are subject to possible errors. In many cases, the researcher must make a decision under uncertainty, but they often make a decision without fully assessing the degree of uncertainty and the consequences of errors. Since these errors are unavoidable in probability, statistical decisions should be made taking these errors and their consequences into consideration. For sound statistical thinking, the researchers should understand the degree of uncertainty, the chance of errors, and their consequences.

Type I and Type II errors.

There are two types of errors in hypothesis testing. Type I error is the rejection of the null hypothesis that is true, and Type II error is the failure to reject the alternative hypothesis that is false, as summarized in the table below:

	H0 is true	H1 is true
Accept Ho	Correct Decision	Type II error
Accept H1	Type I error	Correct Decision

The classical example is the verdict in the court of law where the defendant is assumed to be innocent until proven guilty. That is,

H0: the defendant is innocent; H1: the defendant is guilty.

The Type I error here is delivering a “guilty” verdict to an innocent person, and the Type II error is a “not guilty” verdict to a defendant who is guilty.

Let us take another real-world example, testing for pregnancy. A doctor assumes that a patient is not pregnant until tested otherwise. That is,

H0: the patient is not pregnant; H1: the patient is pregnant.

The Type I error here is judging the patient to be pregnant when she is not, and the Type II error is judging the patient not to be pregnant when she is pregnant.

Although we do our best to prevent such errors, we know these errors are unavoidable. A court of law employs several lawyers and a jury and takes a long time for prudent discussions and deliberation, but we see that an error of judgment happens sometimes. Pregnancy tests have now become highly accurate, but still, there is a chance of a false positive (Type I error) and a false negative (Type II error).

In these examples, the chance of making a Type I or II error may be low. However, such errors can also occur in any statistical decisions, possibly with a lot higher chance than that of a court making an ill judgment or the failure of a pregnancy test. But the main issue is that statistical researchers rarely consider the probabilities of these errors seriously, not to mention their consequences.

Consequences of Type I and Type II errors.

We all know that errors have consequences, and they can be costly. When making a judgment with a possibility of errors, the decision should be made considering such losses.

Let’s go back to the example with a court making a judgment. Suppose that,

if found guilty, the defendant will receive the death penalty. If not, the defendant is freed. A Type I error means an innocent is sent to death row, while a Type II error means justice is not served. While both errors have consequences, a Type I error is not only a personal tragedy and a huge loss for society, but also a serious failure of justice.

Now, let's look at the case of the pregnancy test. A Type I error means a false positive pregnancy, which may not have severe consequences. But a Type II error can lead to serious complications and omissions because the failure of detecting a true pregnancy can compromise the health and welfare of both mother and baby.

It should be noted that both Type 1 and Type 2 errors can lead to severe consequences. It is not like a Type 1 error is always worse than a Type II or vice versa.

Threshold of decisions.

A rational decision-maker would think: how do we minimize the losses from Type I and Type II errors? One way of achieving this is to control the threshold of decisions to reject the null hypothesis. If this threshold is too low, then we will have a higher chance of making a Type I error, but that will see a low chance of making a Type II error. Conversely, if the threshold is too high, the chance of a Type I error is low, but we must face a high chance of making a Type II error.

Consider the case of a court of law where the guilty verdict is the death penalty. The court will require "beyond reasonable doubt" as a burden of proof to pass the guilty verdict to prevent a Type I error. But this way, the probability of a not guilty verdict for a defendant who could be guilty (Type II error) is higher. The court will not choose guilty if they are not convinced about the evidence presented to pass the threshold of "beyond reasonable doubt." The court does not want to make a Type I error, even at the cost of risking more Type II errors, and takes a great burden of proof.

Burden of Proof	Description	Trials
Preponderance of evidence	Greater the 50% chance	Civil,

Clear and convincing evidence	Highly and substantially probable	Civil,
Beyond reasonable doubt	No plausible reason to believe otherwise	Crimi

The above table shows various burdens of proof or decision thresholds for different types of trials. The court is taking a more lenient burden of proof for the trials where Type I errors are less consequential. That is, "beyond reasonable doubt" is applied to the trials where the consequences of a Type I error are dire, while a preponderance of evidence is required for the trials where a Type I error is not as significant and final. By adjusting these burdens of proof, the court is balancing the chances of Type I and Type II errors, considering their outcomes.

Here there are two important points to note:

There is a trade-off between the chance of a Type I error and the chance of a Type II error;

The decision threshold should be adjusted to reflect the consequences of Type I and II errors and their chances.

The trade-off means that, if you increase the chance of a Type I error, the chance of a Type II error will decrease, and vice versa. And the two chances cannot become zero or be made small, simultaneously. Hence, there should be a balance between the two, in consideration of their possible consequences. Our legal system is rational, and it is doing this nicely.

Suppose hypothesis testing as described in many statistics textbooks and in the previous chapter is implemented in our legal system. This means the threshold of 0.05 is almost universally applied to our decision. The threshold may be chosen to be 0.01 or 0.10, and the choice is arbitrary with no scientific or legal justification. Under the p-value threshold of 0.05, the legal system allows for a Type I error in 1 in 20 cases. If the death penalty is given to an innocent defendant with such a high frequency, it is a failure of the legal system. If the same threshold is applied to a matter for a small claims court, then it is allowing too many Type II errors. And if this threshold can be changed arbitrarily to another fixed level such as 0.01 or 0.10 with no

justifications, our legal system will lose its credibility and integrity.

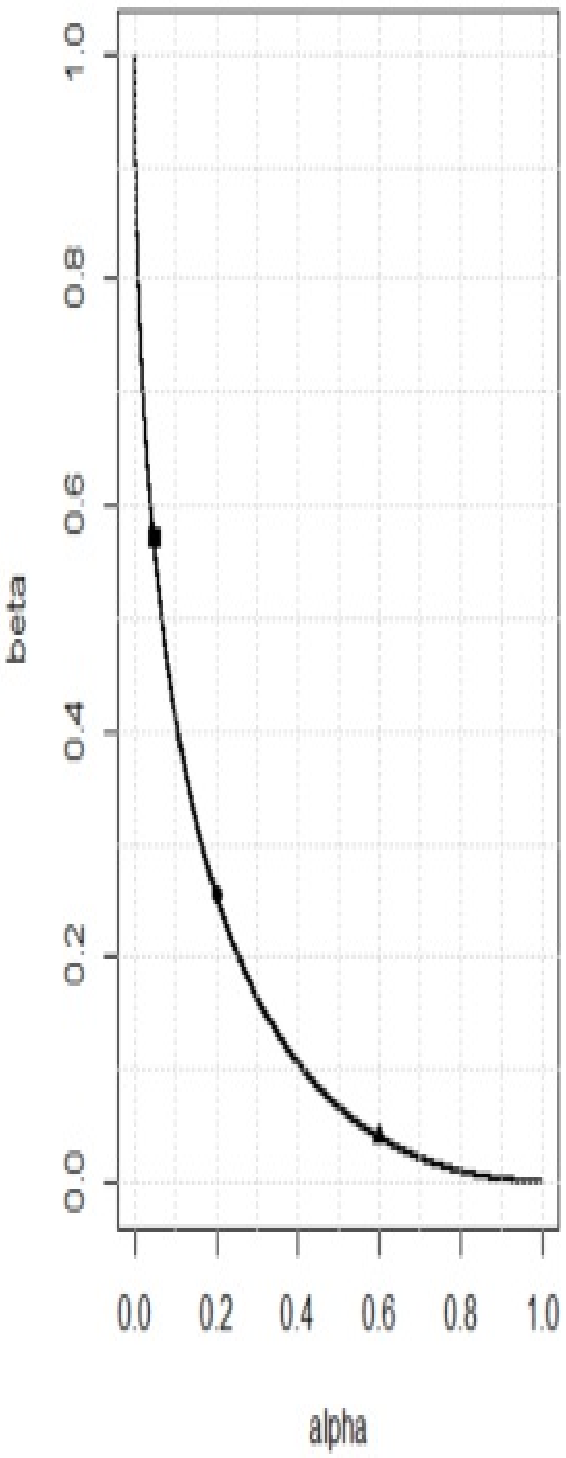
Statistical decisions based on inferential statistics are no different. The researcher is facing the uncertainty of possible errors. If they can control the probability of these errors and understand their consequences, they can make the best decisions. One way of achieving this goal is to adjust the threshold of decision, just as our legal system does.

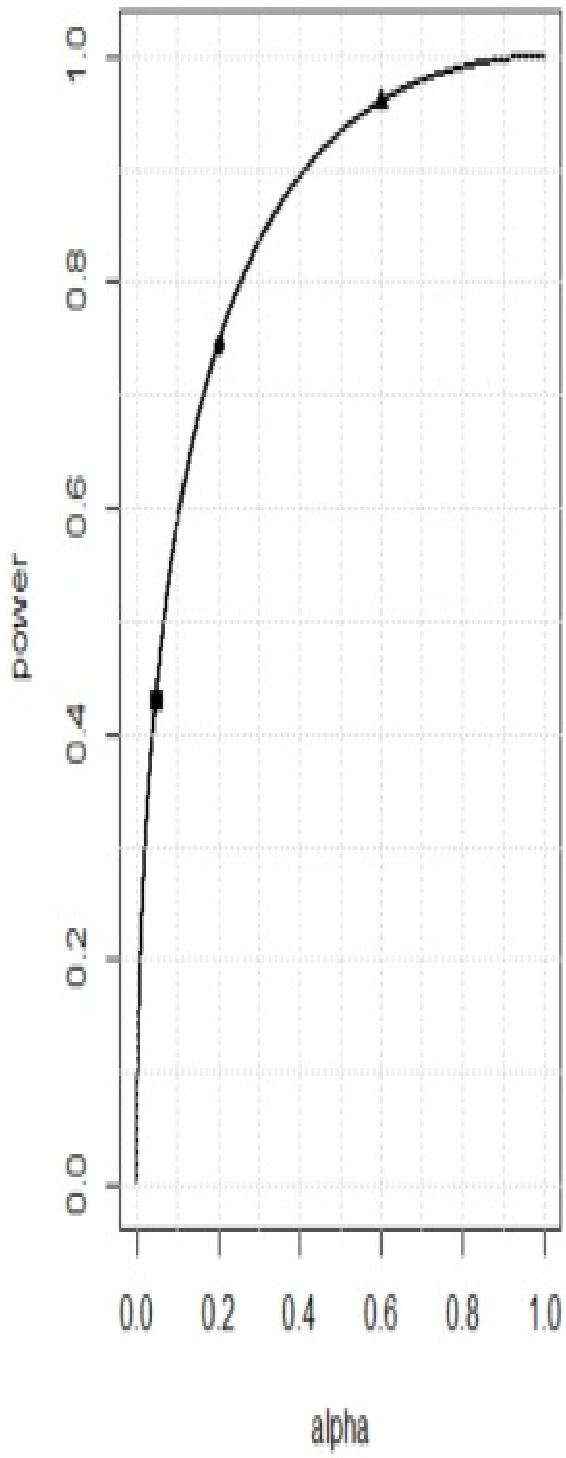
Under the conventional hypothesis testing widely adopted in our statistical research, statistical decisions are made using the p-value with a decision threshold of 0.05, sometimes 0.01 or 0.10, with little justification. In what follows, we will see how silly this can be and how we can do better.

2. Research design.

In statistics, it is a convention to use α to denote the probability of a Type I error, which is the probability of rejecting the true null hypothesis (false positive). The probability of a Type II error is denoted as β, which represents the probability of accepting the null hypothesis which is false (false negative). The (statistical) power is defined as 1 – β, which is the probability to reject the null hypothesis that is false (a correct decision).

It is natural that, as a decision-maker, you want the chance of making errors small as possible, maintaining a reasonable chance of making a correct decision. As discussed above, there is a trade-off between α and β. A higher (lower) value of α is associated with a lower (higher) values β, and this will lead to a higher (lower) value of power.


beta
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
alpha


power
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
alpha

The trade-off and the power function are illustrated in the figure above. The square dot is associated with $(\alpha,\beta) = (0.05, 0.57)$ with the power of 0.43; the circle dot is $(\alpha,\beta) = (0.20, 0.26)$ with the power of 0.74; and the triangular dot is $(\alpha,\beta) = (0.60, 0.05)$ with the power of 0.95.

A rational decision-maker would naturally wish to minimize the error probabilities $(\alpha+\beta)$ and maintain a high level of statistical power $(1-\beta)$ simultaneously. They will also need to consider the losses or consequences of making incorrect decisions. If an innocent person will die because of a Type I error, then the decision-maker will need to control the value of α at a very low level. If a disaster with a huge financial implication will occur because of a Type II error, then the decision-maker has to control the value of β and keep it to a minimum level. This process is called research or experimental design. Sound statistical research should be conducted with a prudent and careful research design before a sample is collected.

The conventional hypothesis testing, however, almost always chooses the value of alpha at 0.05 or sometimes at 0.01 or 0.10. That means, in the above setting, the researcher is allowing for a relatively high value of β and a low value of statistical power. Are there any good reasons for this, not to mention scientific justification? The answer is absolutely no. Conventional hypothesis testing is not based on such reasoning or scientific justification. This is how we teach statistics at our universities, and how they conduct statistical research at the top professional levels. This is as silly as a court of law that applies the same burden of proof for all legal cases every day and all the time, ignoring the characteristics of each individual case.

3. Alpha, beta, and power.

In inferential statistics, the most critical choice for the decision-maker is the level of significance. In hypothesis testing, it will determine the critical values or critical region based on which decision is made. This choice should be made before the researcher collects the data, like a judge chooses the burden of proof before they enter the courtroom to hear the details.

To understand how this value should be chosen, we will need to know the variables that affect the value of β or statistical power (1-β), given a value of α. They are

the sample size, and

a plausible value under H1.

These are the two key parameters of inferential statistics to be chosen carefully. Under inferential statistics outlined in the previous chapter, however, the value under H1 need not be known, and the sample size can be as many as available. As we shall see later, a larger sample size will deflate the p-value, and it will become more and more likely to reject the null hypothesis as the sample size gets larger.

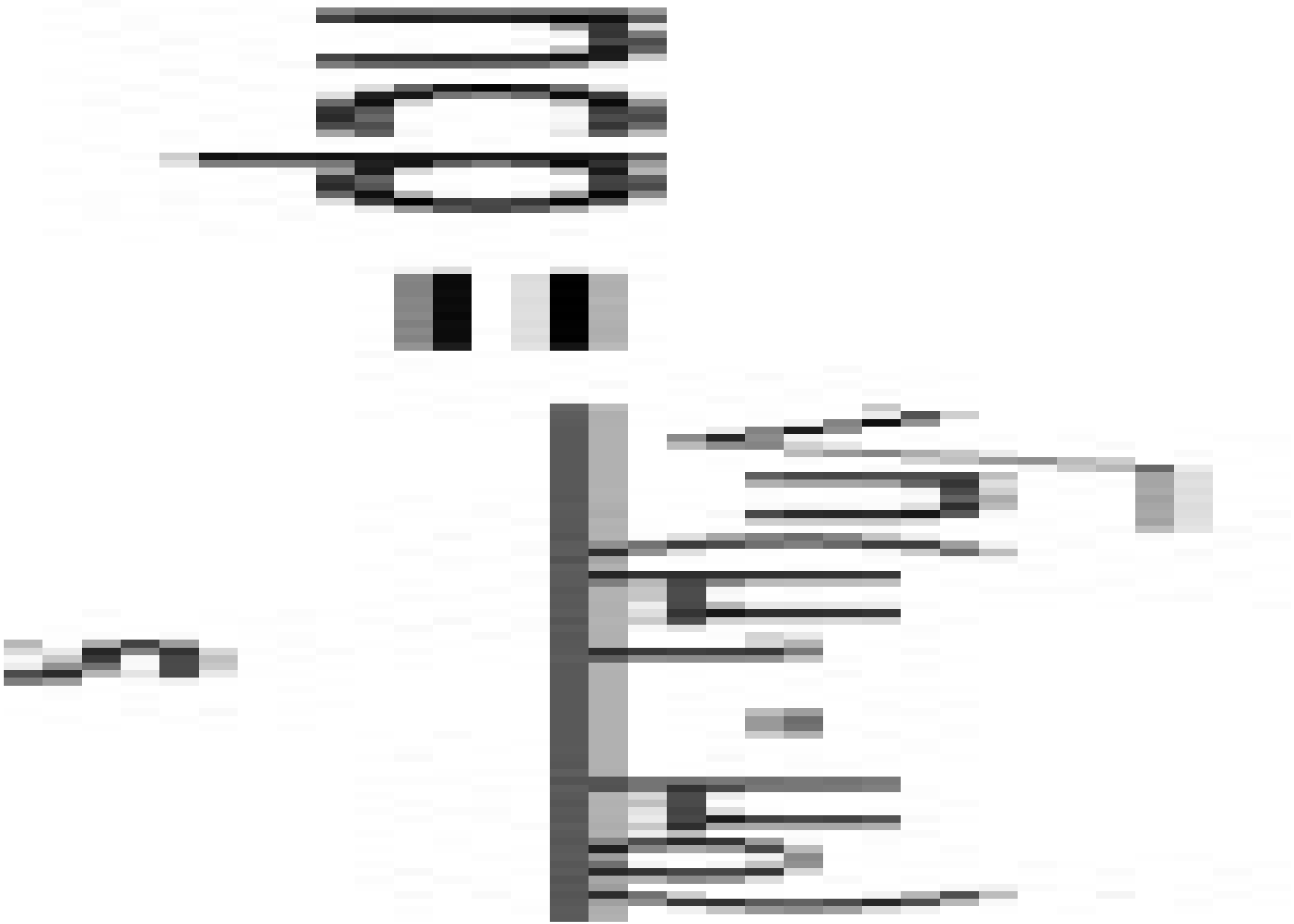
Consider the case of an investor on NASDAQ-100. The null and alternative hypotheses tested in Chapter 2 are:

.

Note that the value under H1 is not specified here. For sound statistical research with statistical thinking, the value under H1 should be specified. This is the value under which the investor can take an action or the value that will influence their behaviour. To know this value, they will need to study hard. That could be the value that will determine future payoff and the decision to invest. Let's assume that, after careful investigation, the value is found to be 3%. Then, the null and alternative hypotheses should be set up with a specified value under the latter as

.

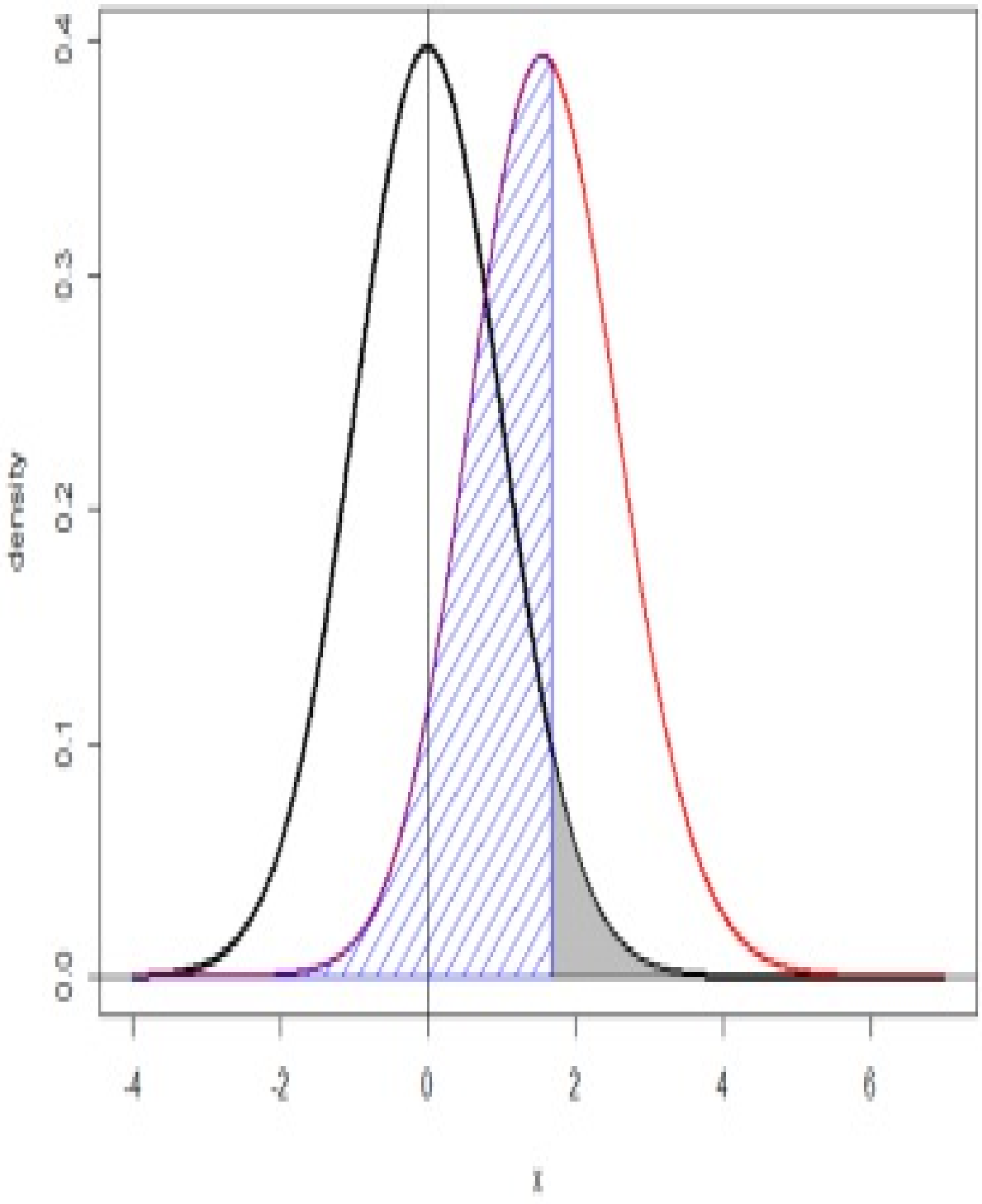
Then the distribution of the sample mean under H0 is N(0,1) and that under H1 is N(ncp,1) where



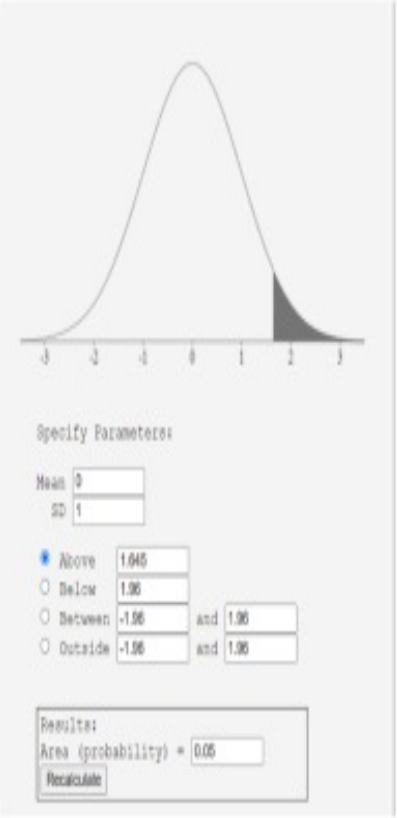
Note μ0 is the value under H0 and μ1 is the value under H1, and ncp refers to the non-centrality parameter.

The distribution under H0 is a normal distribution with mean zero and standard deviation equal to 1 (black distribution on the left), and the distribution under H1 is the normal distribution with the mean of and standard deviation equal to 1 (distribution on the right), which are given in Figure 10 below.

Figure 10: Distributions under H0 and H1(; n = 60)

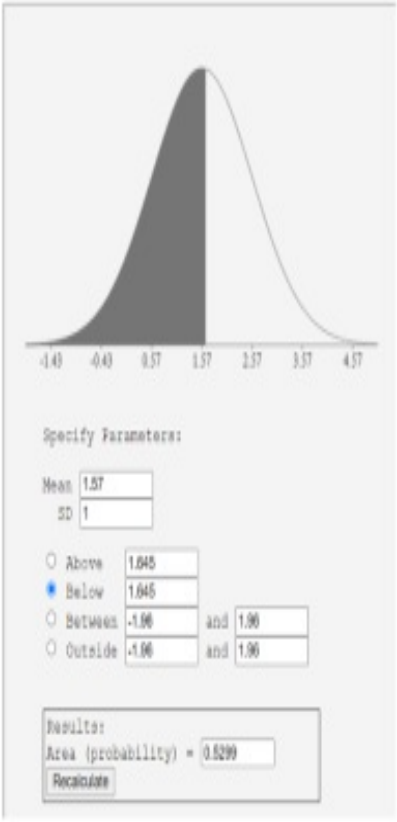

density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
6
x

The curve on the left is N(0,1), and the curve on the right is N(ncp,1) with ncp=1.57.



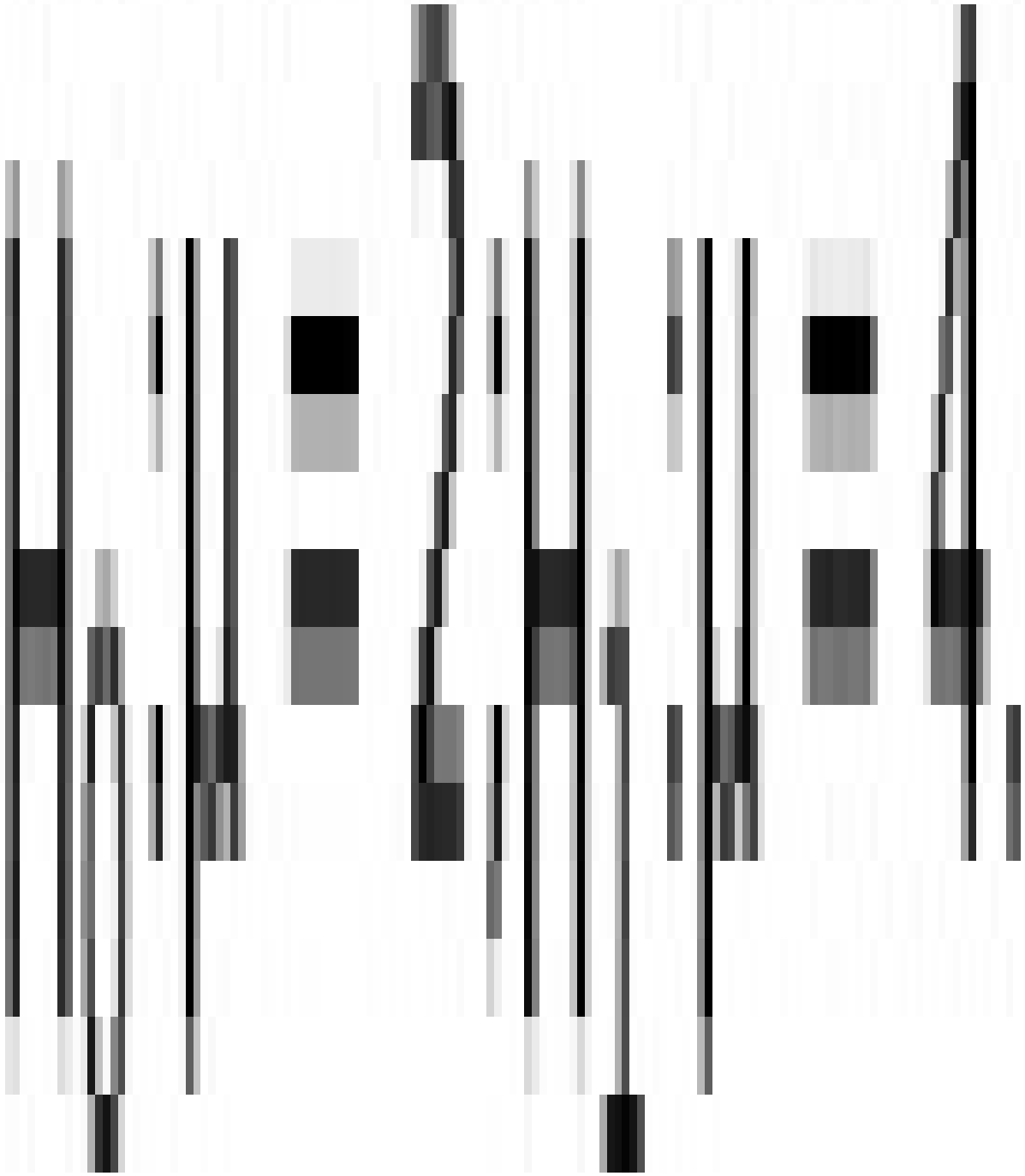
The probabilities on the left curve above can be explained as below:

If the researcher chooses $\alpha = 0.05$ with the critical value of 1.645 from the normal distribution, then the probability of 5% is indicated by the grey, filled, area under the black curve.



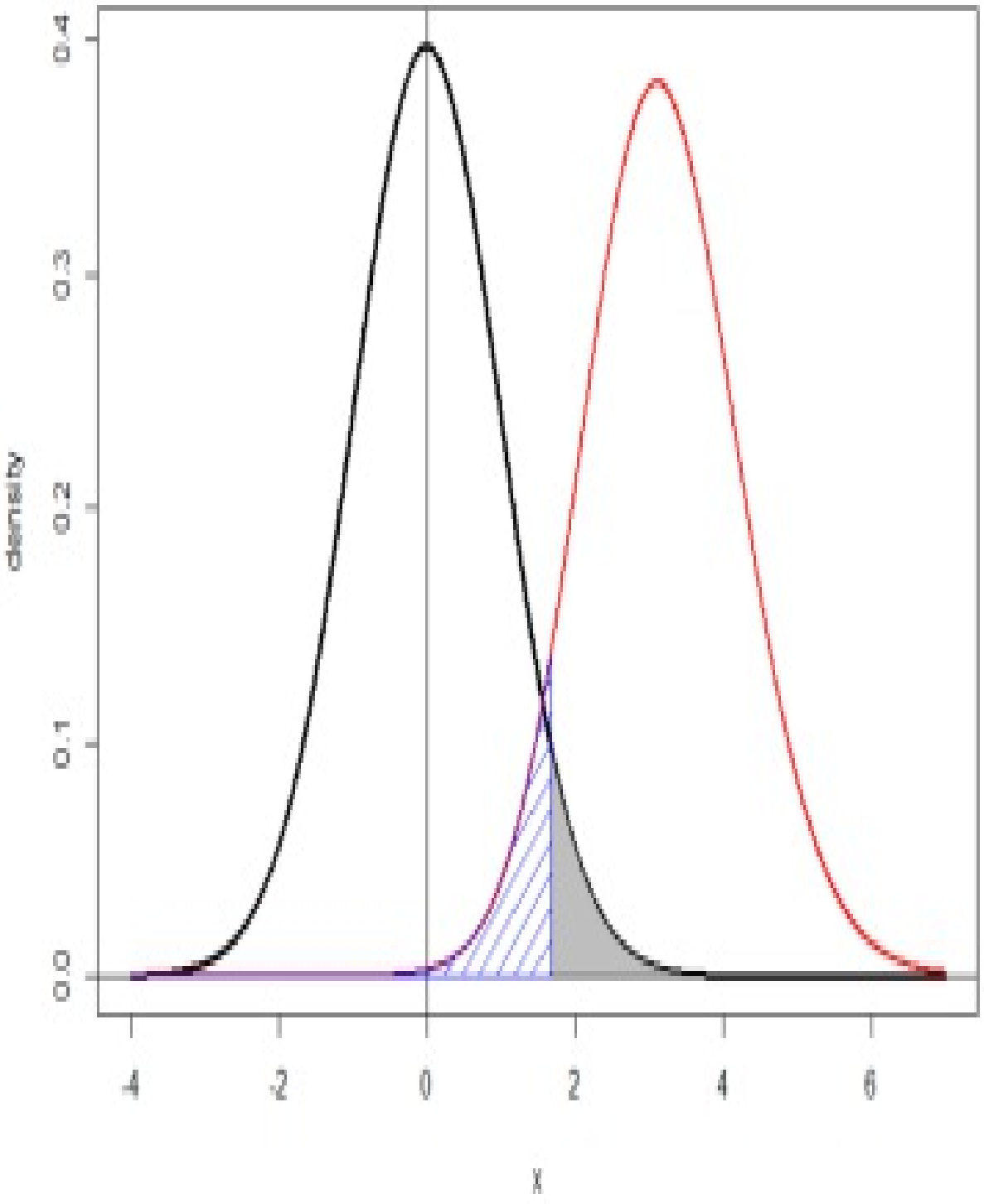
The curve on the left panel is the normal distribution with *ncp* = 1.57. The value of β is the striped area under the curve, which is 0.53, and the power is 0.47.

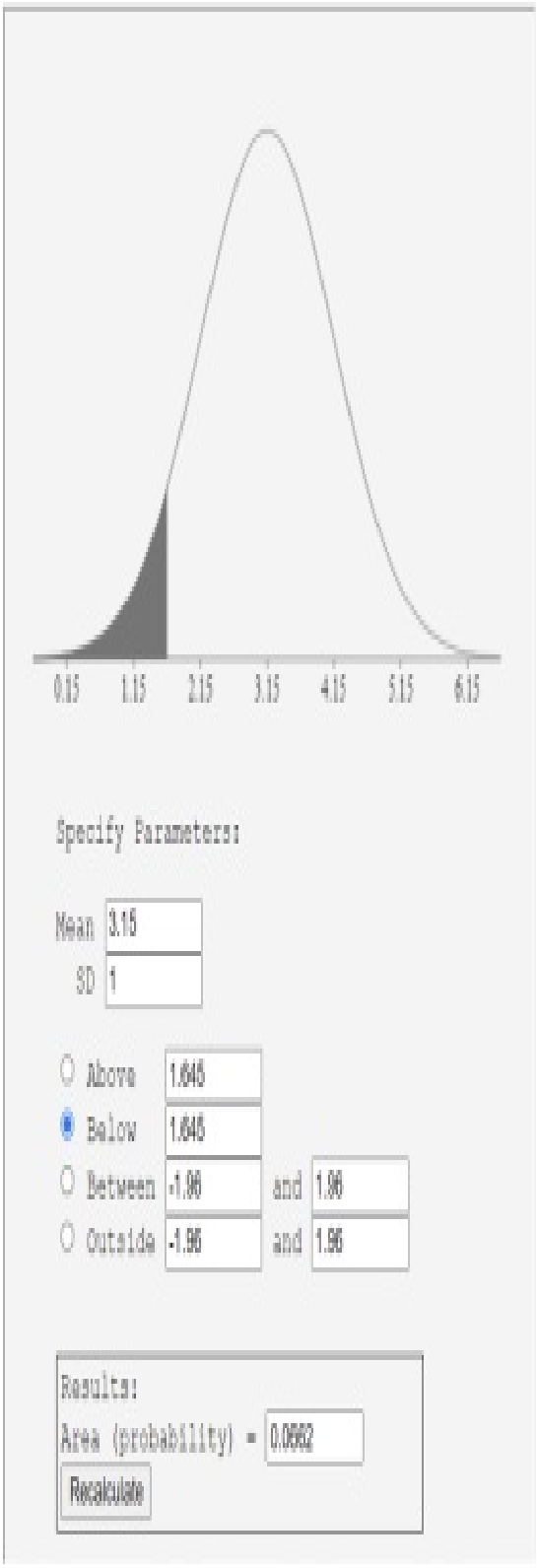
Now suppose the value under H1 is set at 4%. That is,



Then the distribution under H1 is the curve ( ) on the right-hand. Again, the dark grey area under the curve on the left represents the level of significance of 5%. The Type II error probability is around 0.07, indicated by the striped area under the curve on the right, and the power of the test is 93%.

Figure 11: Distributions under H0 and H1 (; n = 60)


density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
6
x



With *ncp*=3.15, the area under the curve on the right is the probability of Type II error, which is around 0.07.

Figure 12: Distributions under H0 and H1(; n = 180)

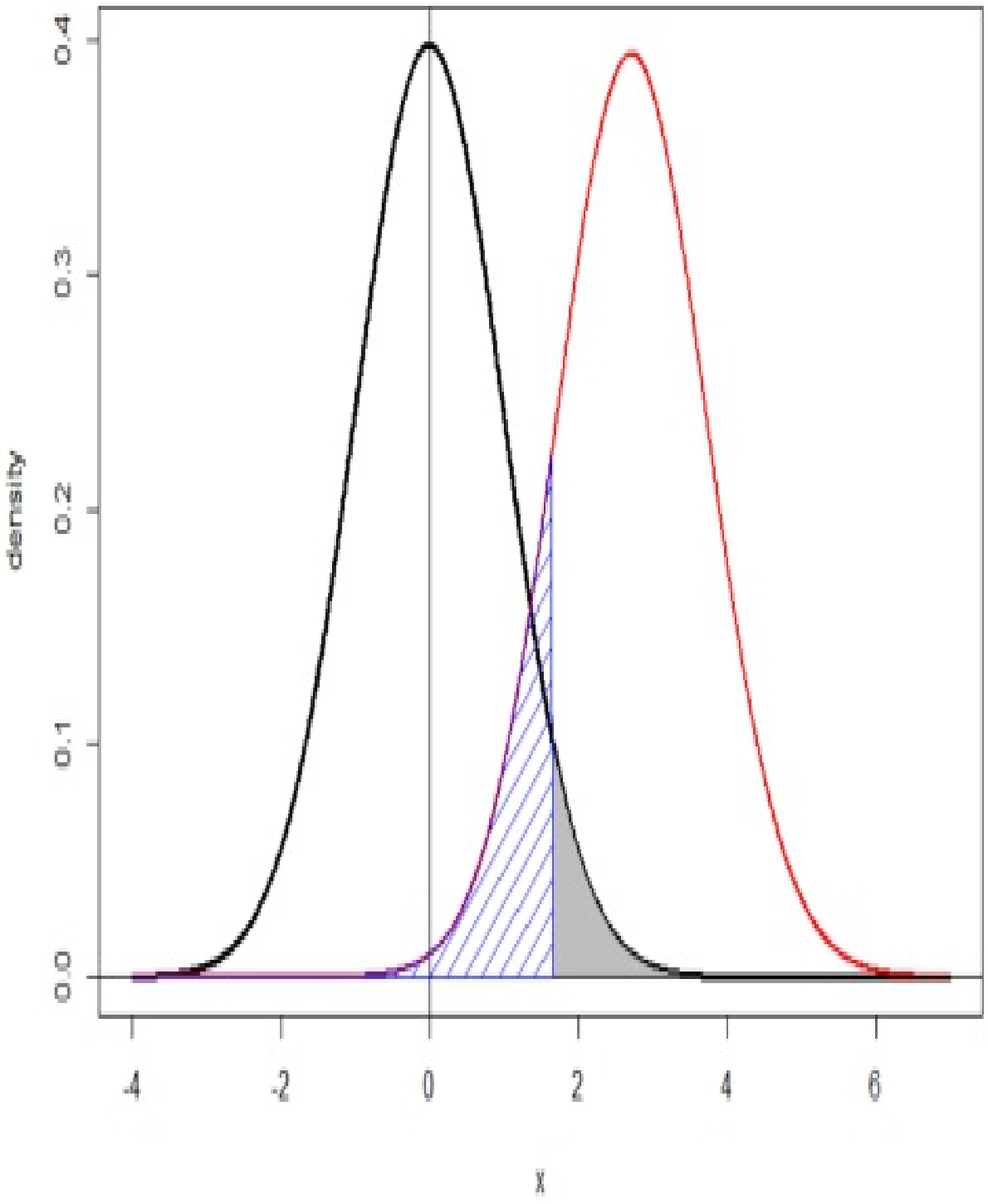

density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
6
x

Figure 12 is the case of with the sample size increased to 180 from 60. The distribution on the right is the normal distribution with the mean of With a larger sample size, the Type II error probability is 0.14 (striped area under the curve on the right) with the power of 0.86. Compared to the case where n = 60, the probability of a Type II error is much smaller, and the power is a lot higher.

Hence, the two important factors that control the power and Type II error probability at a given value of Type I probability are

the value under H1, and

the sample size.

In other words, given the value of α, a sound research design requires a careful choice of sample size and a meaningful value under H1. The sample size should be chosen by looking at the balance between the Type I and II error rates, also in consideration of the cost of sampling. The value under H1 should be the value that matters practically. In our investment example, the chosen value of 3% under H1 is the value that triggers the investment decision. If such research design is applied to the drug trial, the value under H1 should be the effective rate of drug at which the researchers would recommend the approval of the drug for public use.

4. Implications to research with Big Data.

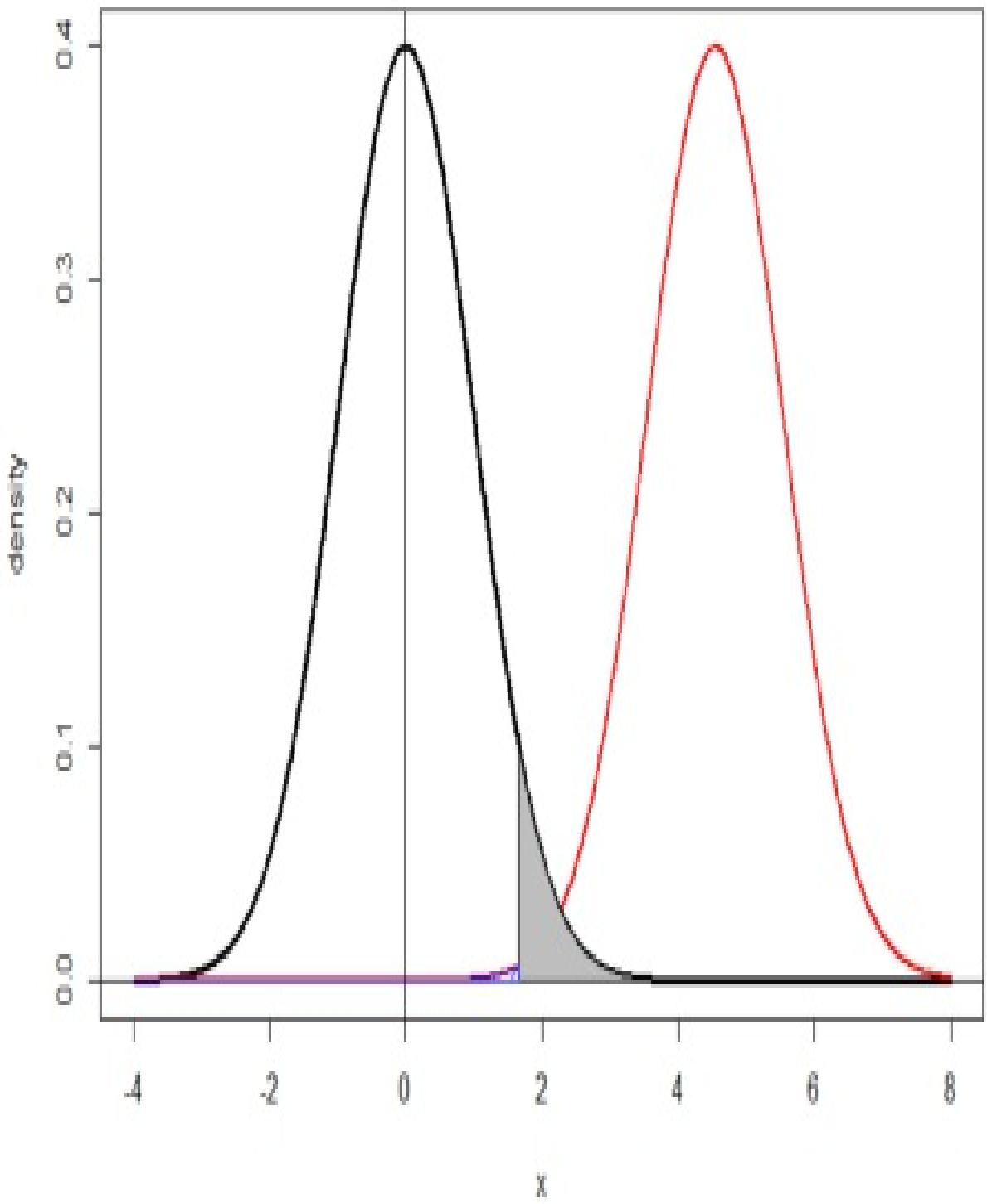
We are living in the big data era, when data is abundant and cheap. With fast internet and an efficient system of data storage, it is nearly effortless to obtain tens of thousands of data points, if not millions. If researchers have easy access to such big data sets, it is natural they want to use as many data points as available in their research and decision-making. Then what are the implications for statistical inference?

Let us go back to our simple example of

.

Assume the true value of μ is 2.10, which represents a negligible deviation from the value of μ under H0. Suppose the researcher has 50,000 data points or more.

Figure 13: Distributions under H0 and H1(; n = 10000)


density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
6
8
x

As illustrated above, the distribution under H0 (left curve) is the standard normal, while that of H1 (right curve) is the normal distribution with . At the 5% level of significance, the value of $\beta = 0.0019$, and the power is 0.9981.

If the true value of μ is 2.10, then the test statistic is always generated from the distribution under H1. The implication is that, at the 5% level of significance, the null hypothesis is almost always rejected, with the power being almost equal to 1. The problem here is that the value of μ of 2.10 may not differ from 2. The investment return of 2.1% is larger than 2%, but it is likely that they are practically the same. However, the researcher can show that the mean return is statistically larger than 2%.

The implication is that, with a large enough sample size, any negligible deviation from the value under H0 can be shown to be statistically significant, as the above example shows. The non-centrality parameter (ncp) is an increasing function of the sample size, and it will push the distribution under H1 away from that under H0, as the sample size increases. This is the reason almost any variables are correlated with statistical significance in a big data set. A prudent researcher will measure the effect size or signal from the data and decide that such a small deviation from the value under H0 is not practically or substantively important. However, many researchers are also misguided by apparent statistical significance and make erroneous decisions.

If we compare the Type I error and Type II error probabilities, the Type I error rate of 5% (or 0.05) is over 26 times higher than the Type II error rate of 0.0019. This clear imbalance means a Type I error is likely in this situation: i.e., reject H0 when it is (practically) true.

One possible and sensible solution is to reduce the level of significance to a much lower level, such as 0.1% (or 0.001), which will balance the Type I and II error probabilities. When the sample size is massive, it is sensible to increase the decision threshold to reject H0 by lowering the level of significance. However, this is a clear departure from the convention of using the 5% (sometimes 1% or 10%) level of significance that many statistical researchers are reluctant to adopt, and such a recommendation is usually met with strong resistance.

5. Choosing the level of significance.

In the previous section, we reviewed the method of evaluating the probability of a Type II error and statistical power, given the probability of a Type I error. The purpose of the exercise is to understand the degree of uncertainty associated with decision-making. By doing this exercise, the researcher knows the probability of making errors and making a correct decision.

The big question is how we can set the level of significance α. It should be chosen in consideration of β (and power) and sample size and the consequences of making incorrect decisions.

Let us go back to the example of the legal trial where the guilty verdict is the death penalty. The court adopts "beyond reasonable doubt" as the burden of proof because it does not want an innocent person on death row because of the court committing a Type I error. A level of significance that will set the decision threshold as high as "beyond reasonable doubt" in this case is a low value like 0.001. We may want to make a Type I error only with a chance of 1 in 1000. We now know this will make the probability of a Type II error unreasonably high (as high as 0.99) with a low power, but this may be justifiable because an innocent human life is so costly.

Let us take another example:

H0: Climate is not changing

H1: Climate is changing

We maintain our belief there is no climate change until proven otherwise. A Type I error judges that the climate is changing while it actually is not; a Type II error judges that the climate is not changing when it is. We all know a Type II error could have serious consequences, a lot more than a Type I error. A Type II error will encourage no or inadequate actions for climate change, but a Type I error may implement the actions that can improve life

on our planet, even if the climate is not changing.

In this case, a wise action is to set the value of α at a high value so we have a low value of β and high power. We want to control the chance of a Type II error at a low level so it can happen only with a small chance. For example, we can set the value of $\alpha = 0.95$ so the value of β is as low as 0.1. That means we will reject the null hypothesis with a low burden of proof. Any faint sign of climate change should be taken seriously, and this will implement a range of actions to save the environment, which will benefit the current and future generations, whether or not climate is actually changing.

From these two examples, the message is clear. We need a cost-benefit analysis in choosing the level of significance; cost-benefit analysis under uncertainty of Type I and II errors. Any reasonable decision-maker with clear statistical thinking who understands the uncertainty involved would do this: this is how we teach our students in our high school and universities in all disciplines but statistics. In statistics, unfortunately, they teach the level of significance to be set at a conventional level of significance.

As discussed in the previous section, it is also ludicrous to stick to a conventional level of significance in the big data era. Hence, we need a new and revised standard of statistical evidence when a large or massive sample size is routinely used. On this basis, recently, there are researchers calling for much lower levels of significance, such as 0.5% or 0.1%, instead of 5% or 1%[3].

6. A brief history of modern statistics.

So far in this chapter, we have discussed how a rational decision-maker should consider the chance of Type I and Type II errors and how she should set the level of significance as a decision threshold in consideration of the consequences of these errors. This was the teaching of the pioneers of modern statistics. However, this is not how the researchers in modern times make their statistical decisions. To understand this, a brief review of the history of modern mainstream statistics will help.

Student

William Sealy Gosset was a pioneer of modern statistics. After graduating with honours from Oxford University in mathematics and chemistry, he was hired as a brewer by Guinness Brewery in Dublin in 1899. His job was to control the quality of the final product by monitoring the process of manufacturing. As the Head of Experiment Brewer of Guinness, he had to develop a mathematical method of judging whether an effect was by treatment or by accident. For example, his problem was to determine which types of barley provided the best quality beer when appropriately controlled with several factors. The types of barley were limited, and his problem was to make a decision from a small-sized sample. His sabbatical to the University of London has led to a paper published under the pen name of Student, from which many fundamental concepts of statistical inference come to light. An important point to note is that the starting point of the modern statistical method was to develop a way to tackle a problem with a small sample size.

Fisher

Sir Ronald Fisher further established the method of hypothesis testing by introducing the null hypothesis and using the p-value. In his influential book published in 1925, he proposed the p-value as a measure of evidence against the null hypothesis and suggested 0.05 (1 in 20 chance) as a benchmark threshold of the decision to reject the null hypothesis. This decision rule was

recommended only when the researcher knows little about the problem at hand. Fisher believed the p-value "as an objective aid to assess the plausibility of a hypothesis and ultimately the conclusion of differences or associations to be drawn remained to the scientist who had all the available facts at hand."[iv][4] This 0.05 cut-off has become an almost universal threshold to reject or not reject the null hypothesis in modern statistics. However, in 1956, Fisher further stated,

"No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."

While Fisher did recommend 0.05 as a benchmark threshold for a small sample analysis, it seems he never intended it to be used as a universal threshold. A part of the deep problem we have is this 0.05 threshold recommended for a small sample analysis is still being routinely and mindlessly used in the era of big data.

Neyman and Pearson

The next generation of pioneers were Jersey Neyman and Egon Pearson, who introduced a decision-theoretic approach to hypothesis testing in their paper published in 1933. They introduced the alternative hypothesis in addition to the null hypothesis and concepts such as the level of significance (Type I error rate), statistical power, and Type II error rate. The sample size and the level of significance are critical elements in their method, and they should be chosen before the researcher observes the data. These choices also determine the critical values or critical regions where the researcher accepts the null hypothesis or the alternative hypothesis. Note that "accepting a hypothesis does not mean that you believe in it, but only that you act as if it were true."[v][5] According to their teaching, the sample size and the level of significance should be chosen in explicit consideration of the losses or consequences of incorrect decisions.

As we have seen in this chapter, the concepts such as Type I error rate, Type II error rate, and statistical power are important for the paradigm of Neyman and Pearson and largely because it has a substantive alternative hypothesis that specifies the population value we are testing for. This is the point that was different from Fisher's method and the method being adopted in modern statistics.

In addition, these pioneers recommended inferential statistics as an aid to make the final decision. The inferential outcome should not drive the whole decision process. They recommend being "modest and thoughtful" in making statistical decisions, evaluating the outcome of statistical inference carefully considering all information.

Null ritual

The statistical methods adopted by modern statistical researchers are rather different from the teaching of the above pioneers. It is called "the null ritual" by Gigerenzer (2004) in his article, "Mindless Statistics". Following his description, the null ritual is conducted in the following way:

Set up a statistical null hypothesis of "no mean difference" or "zero correlation." Don't specify the predictions of your research hypothesis or of any alternative substantive hypotheses.

Use 5% as a convention for rejecting the null hypothesis. If the difference is significant, accept your research hypothesis. Report the result as $p < 0.05$, $p < 0.01$, or $p < 0.001$ (whichever comes next to the obtained p-value).

Always perform this procedure.

With the null ritual, the researchers' primary concern is the p-value and whether it is less than 0.05 or sometimes 0.01 and 0.10. If it is less than 0.05, they are satisfied with the research outcome, and they can stop there. They only need to justify their research outcome under the banner of "statistical

significance". The substantive importance (the size of the effect), plausibility, and implication of the research outcome are a secondary issue. They are only interested in selling the sign of the effect (whether it is positive or negative), with several asterisks attached to the values they report. In economics, this practice is labelled "sign econometrics" and "asterisk econometrics".[vi]

If the p-value is not less than 0.05, they believe they might have done something wrong, and they keep trying (data collection, different models, or methods, etc.) until it becomes less than 0.05. This process is called p-hacking or data snooping as we shall see in a later chapter. Hence, the readers of academic journals and statistical reports only see a distorted picture of research outcomes (called publication bias).

Law of the instrument: "if all you have is a hammer, everything looks like a nail."

The crux of the deep problem we have is that the above "null ritual" is being used everywhere for (nearly) everything, from the first textbook of statistics to top academic journal articles. And the world has taken this null ritual as a single-hammer approach to statistical decision-making: it is not an aid to decision-making anymore; it has become a silver bullet. This problem is also known as Maslow's Hammer, which is a cognitive bias that involves an over-reliance on a familiar tool.[vii]

The null ritual is not what Fisher, Neyman, and Pearson had in mind as an instruction. It is rather a hybrid of their proposed methods, removing the researcher's thoughts and judgments. Nevertheless, most textbooks and statistical lectures teach statistical inference following this ritual as a single hammer approach. They rarely teach any alternatives, nor do they inform the students and researchers of the drawbacks of this approach. Most academic papers also report their statistical results following the above ritual adopting a single-hammer approach. As a result, "the world is awash in bullshit,"[viii] acquired from mindless statistics, and this was the main reason the American Statistical Association acted with their statements we will discuss in Chapter 5.

No one knows exactly where the null ritual has come from and how it slipped into the mainstream statistical research in the post Neyman-Pearson era.

Nevertheless, it silently sneaked into our textbooks and lecture notes and has become a collective habit of nearly all statistical researchers. Breaking this habit will be a real challenge for years to come for the statisticians of the future.

7. Concluding remarks

Statistical thinking and decision-making require careful evaluation of the degree of uncertainty associated. This chapter presented the methods of evaluating the probabilities of Type I and Type II errors. These errors are consequential and can incur serious losses. Sound decision-making should evaluate these probabilities and balance the possible losses from the errors. The key elements of hypothesis testing, such as the decision threshold or the level of significance, should be determined in explicit consideration of these probabilities and losses. The effect size or the signal from sample should also be carefully investigated. Recently, Kim (2021) has proposed a decision-theoretic approach to hypothesis testing, in which the level of significance is chosen so the expected loss from incorrect decisions is minimized, following Leamer (1978)[ix] and the spirit of the Neyman-Pearson decision theory. Interested readers are pointed to Kim (2021)[x] for further details.

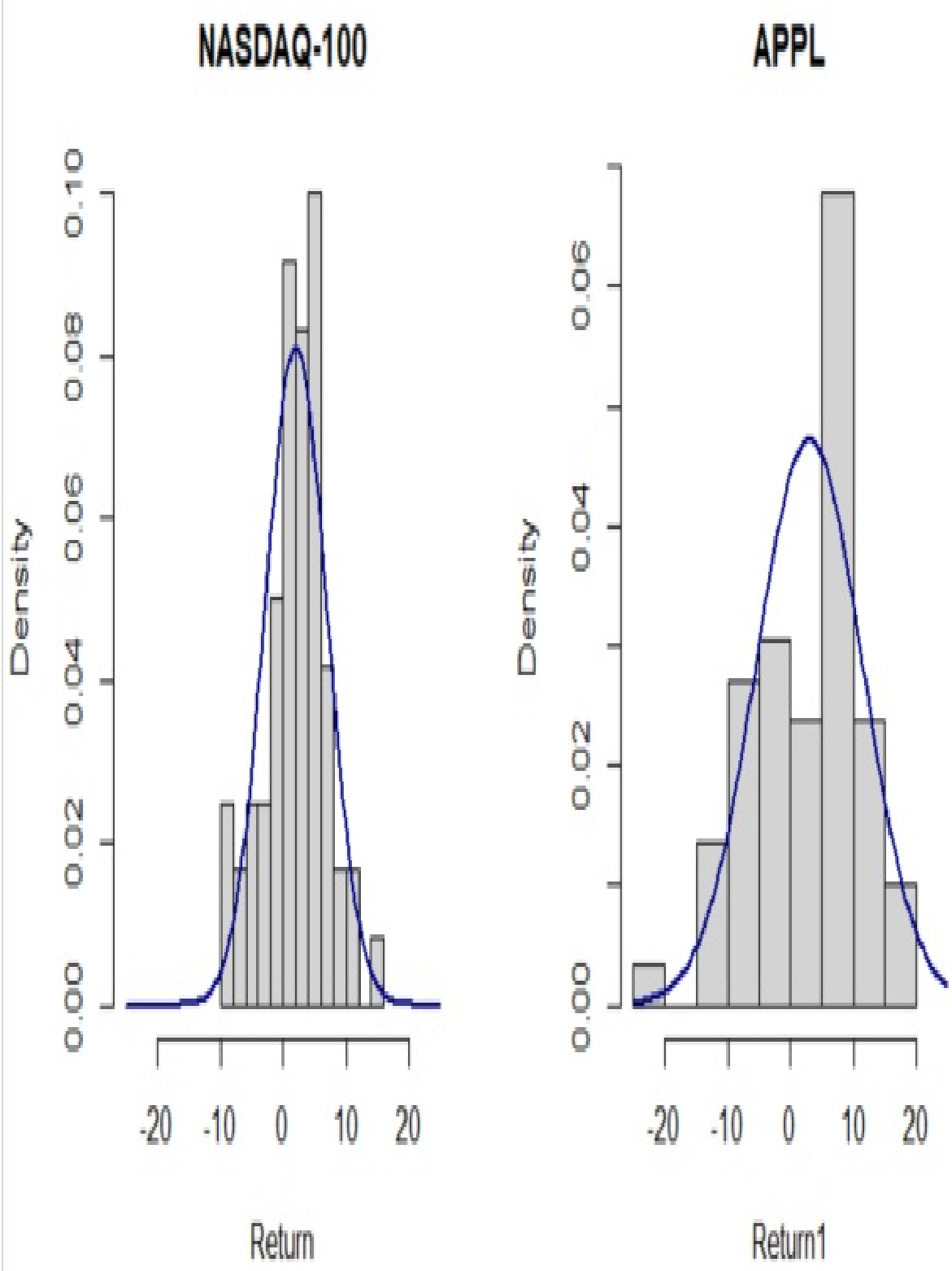
While the pioneers of statistical thinking and methods, such as Student, Fisher, Neyman, and Pearson, have proposed sound methods of statistical inference, modern statisticians have been taught to use a hybrid of their proposals, called the null ritual. They were also taught to use this ritual mindlessly, training future generations to follow this null ritual with a single-hammer approach. A serious change required to restore credibility is our statistical methods. This is particularly so in the era of big data, when traditional statistical methods show critical limitations and deficiencies under a massive sample size.

Chapter 4: How is Statistics Applied in Real Life?

In this chapter, we will see how statistical methods discussed in Chapters 1 and 2 are applied to real-world problems using examples from different fields of social and natural sciences. Where appropriate, we will consider two types of researchers: one is the null ritualist who follows the null ritual as described in Chapter 3, and the other is a decision-maker following the Neyman-Pearson paradigm.

Investment Decision

Consider the investor who has an interest in investing in the NASDAQ-100, as discussed in Chapter 1. They found this investment for the last 5 years - to December 2021 – had an average return of 2.02% per month (median = 2.67%) with a standard deviation of 4.92%. Now they consider an alternative by investing in Apple stocks (APPL), which had an average return of 3% per month (median = 5.26%) with a standard deviation of 8.44%.


NASDAQ-100
Density
0.00
0.02
0.04
0.06
0.08
0.10
-20
-10
0
10
20
Return
APPL
Density
0.00
0.02
0.04
0.06
-20
-10
0
10
20
Return1

While the APPL stock has a higher mean (or median) return, it has a larger variability. The standard deviation of the NASDAQ-100 is 4.92% in comparison with that of APPL, which8.44%. This means the APPL returns vary a lot more around the mean than NASDAQ-100, as the above histograms show. The investor can get a higher average profit from APPL, but it can be a riskier investment, with monthly returns sometimes lower than -20%.

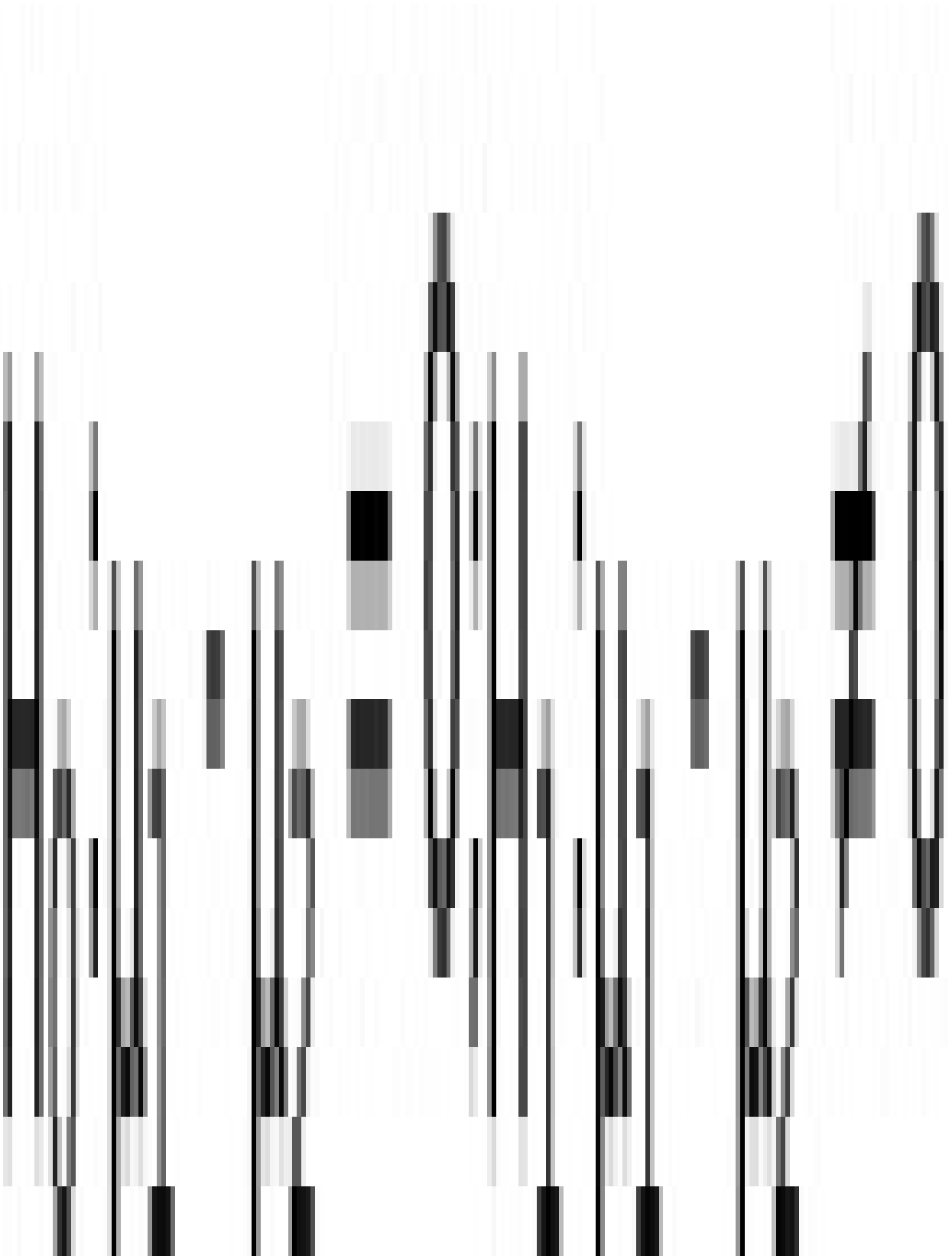
		s	S
NASDAQ-100	2.03	4.92	2.03/4.92 = 0.41
APPL	3.00	8.44	3.00/8.44 = 0.36

The investor can take a descriptive approach by using the ratio of the mean to the standard deviation. That is,

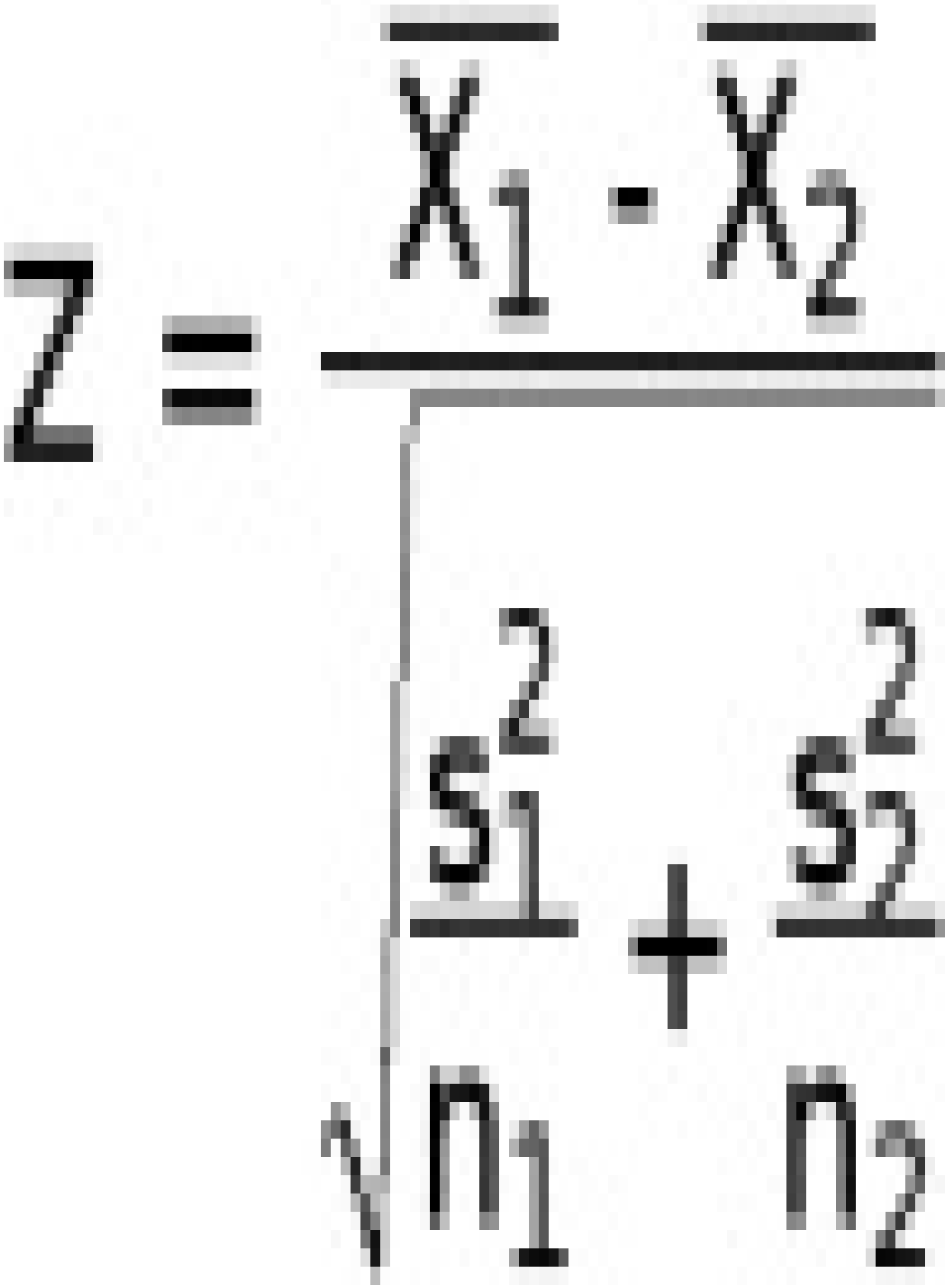
,

where S (called the Sharpe ratio) can be interpreted as a standardized measure of return, i.e., the average return per unit of, in this case, risk. As the above table shows, the value of S for NASDAQ-100 is 0.41, which means the average return of 0.41% per unit of risk, and the APPL is 0.36% per unit of risk. The difference is not substantial, so the choice depends on how the investor perceives the risk of each investment and its prospects.

In inferential statistics, let us first follow the way of the null ritualist who wants to test whether the population mean from the two alternative investments are the same. The null and alternative hypotheses are written as:



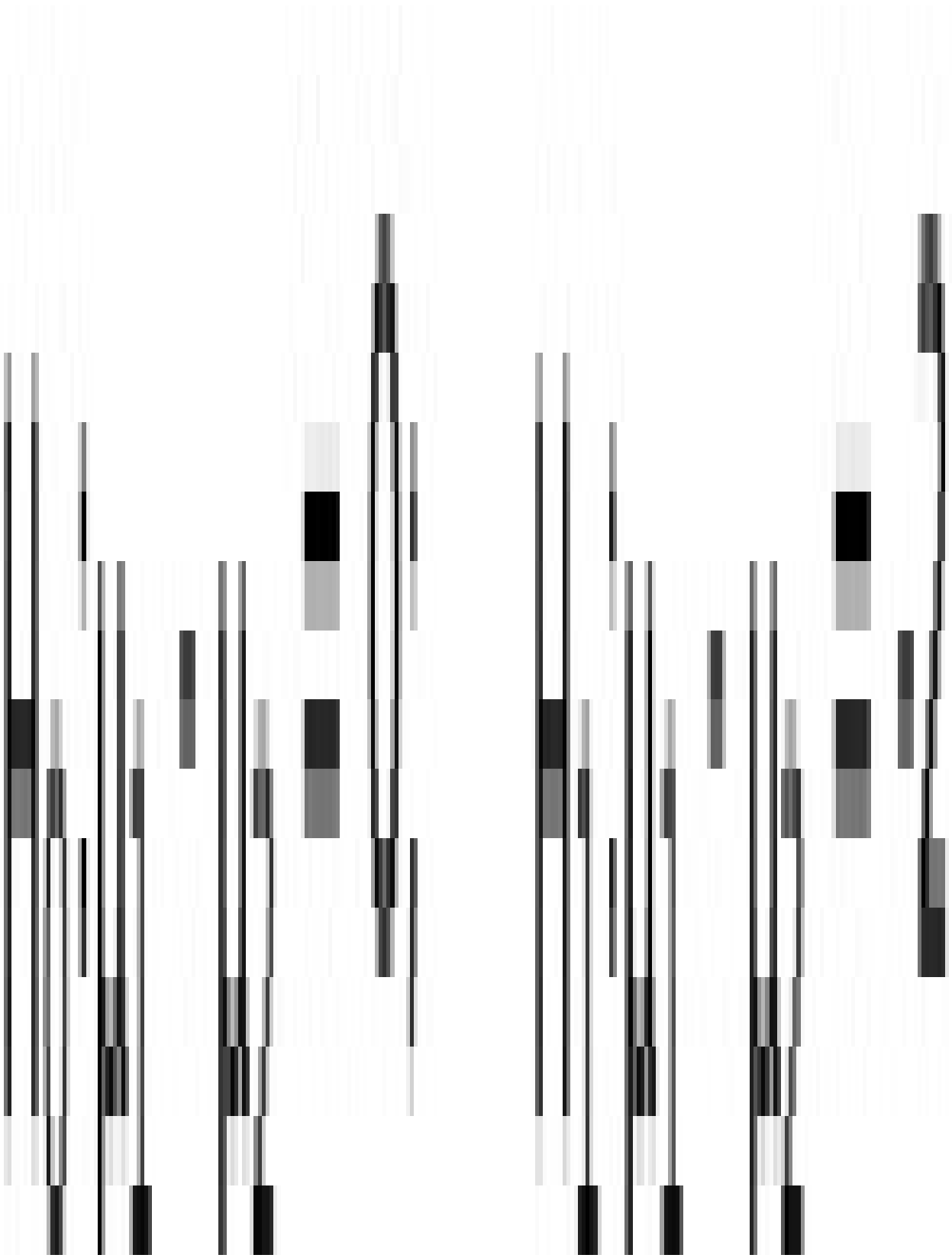
The test statistic is written as:



where and are the sample means from different investments, and and are their sample variances, and n1 and n2 are the sample sizes. This is an extended version of the Z-statistics we presented in Chapter 2, but again it can be interpreted as a signal-to-noise ratio: the numerator is how much the sample value differs from 0 (the value under H0), and the denominator is the noise with scaling by the sample size.

Inserting the values into the above formula gives the Z-statistic of 0.7731 with the p-value of 0.4408. Hence, we cannot reject the null hypothesis at the 5% level of significance. The mean difference of around 1% is due to sampling variability. We found no evidence that it represents the difference in population. Hence, the researcher can be indifferent between these two options.

Now consider an Neyman-Pearson decision-theoretic approach. The null and alterative hypotheses are set up as:



The investor will choose the APPL stock if their return from APPL investment is at least over 2% higher than that from NASDAQ-100. This specific value under H1 should be chosen based on the careful study of the investment positions and risk involved. The level of significance (probability of Type I error) and probability of Type II error should be determined considering the possible losses from each investment. This approach requires a lot more investigation and careful design of the study than the null ritual. However, as we have discussed in the previous chapter, the world has forgotten this approach, and this type of hypothesis testing is hardly used in academic and professional statistics.

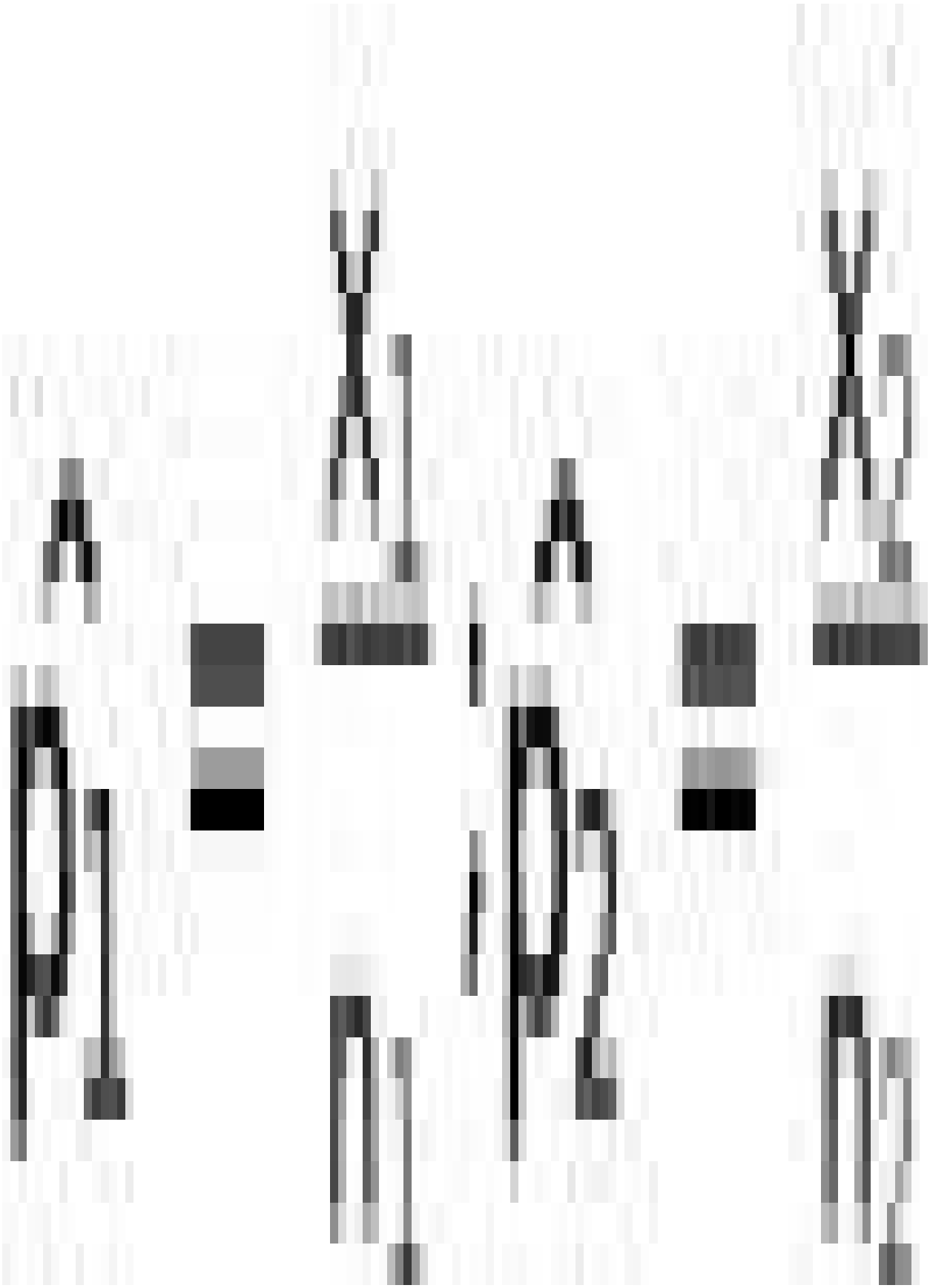
Opinion Polls

An opinion poll is one of the most widely available applications of statistics. It is usually “designed to represent the opinions of a population by conducting a series of questions and then extrapolating generalities in ratio or within confidence intervals.”[xi] Political polls often hit news headlines during election times, attracting enormous attention from politicians, news agencies, decision-makers in business, policymakers, and voters. Multiple surveys are being held every day to understand the expectations, behaviors, and preferences of a group of people or organizations.

With political polls, the pollsters usually take the sample from the population of eligible voters of the size varying between 1000 and 3000. This range of sample size is chosen to make the task manageable with the time and cost available and ensures the sampling error is within a reasonable limit.

As an example, let us take the poll results for the 2024 presidential election between Presidents Biden and Trump.[xii] On polls ending 23 August conducted by Emerson college with a sample of 1000 voters, 43% of the sampled voters support Biden and 42% Trump. A key statistical question is whether the difference of 1% has any significance statistically. Does this difference represent a real difference in population or occur because of sampling variability?

Let p1 be the population proportion of Biden support and p2 be of Trump. The sample proportions are estimated as:

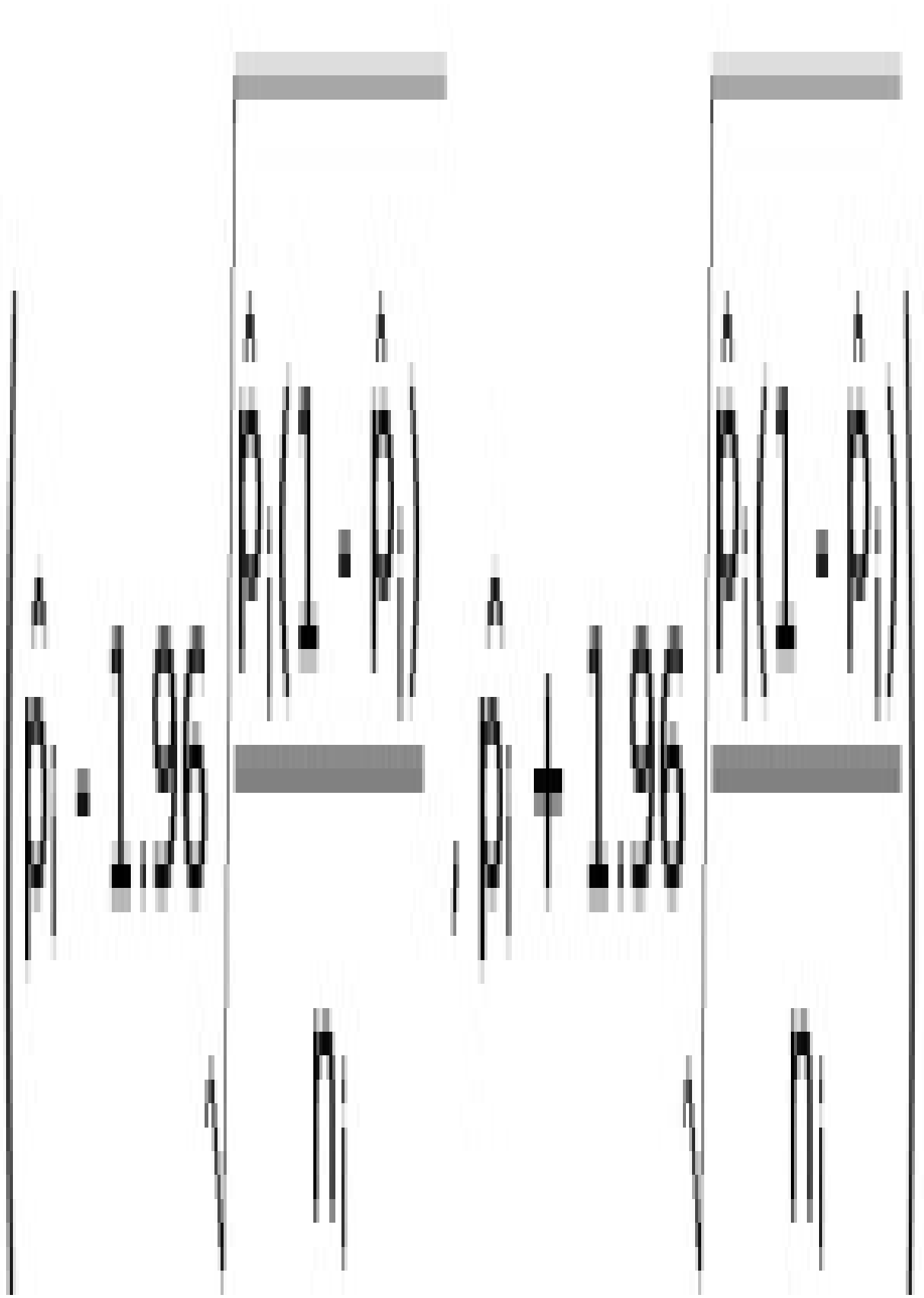


where X1 and X2 are respectively the number of Biden and Trump supporters among the total number of respondents (n1=n2=1000).

The sampling distribution of the proportion follows a normal distribution (for i = 1, 2):

.

This means the sample proportion follows a normal distribution with the mean the same as the population value and the standard error of . Based on the 95% confidence interval for the population, the value is calculated as,



This interval is (0.40, 0.46) for Biden support and (0.39, 0.45) for Trump support. Here, the two confidence intervals overlap, and the difference of 1% is within sampling variability. There is no evidence to uphold that the support rates are different in population.

In hypothesis testing, one can test the following null and alternative hypotheses:

.

Although the polls carry useful information and often provide reasonably correct predictions or representations of the population, it also has been shown to have big failures. For example, in the 2016 Presidential election for Clinton and Trump, national polls predicted a Clinton victory. In 2020, polls predicted a Biden victory by a comfortable margin, but Biden pulled a victory by a narrow margin.

There are polling errors in the outcomes that can be biased in a systematic way. They are called non-sampling errors, which differ from the sampling errors we discussed in Chapter 2. Non-sampling errors include biased sampling, biased survey questions, non-responses, and false information provided by respondents. In the 2016 U.S. election, polls did not represent the opinions of people with lower education and white voters. As they are not sampling errors, they do not disappear or get smaller as the sample size increases. The bias associated with these non-sampling errors can be magnified as the sample size increases, further deteriorating the accuracy of the polling results.

Economics Research

In economics, the relationship between economic variables is estimated and tested. This is to establish the empirical validity of economic theories or to estimate the parameter value of interest implied by the data at hand. In making economic policies or recommendations, the relationship and the estimated parameter values are important. For example, it is a well-established economic theory that, in times of high inflation, the central bank intends to increase the interest rate to put pressure on rising price levels. The question the central bank should answer before they increase the interest rate is how much inflation will decline in response to a 1% increase in interest rate. This “how much” question can be answered by estimating the relationship between interest rate and inflation rate using past data.

Consider a simple case of demand equation. Economic theory tells us that demand is negatively related to price. If consumers face a higher price for a product or service, the demand or sales will decline. They will simply reduce their consumption, or they will look for an alternative product or service. But the question is how much demand will decline if the price is increased by a certain amount. The answer depends on a range of factors, such as the type of product or service and demographic features of the consumers, among others. It can be the case that the demand is highly sensitive to the price, or the demand may not change substantially in response to the change in price. And this “how much” question can be answered by looking at the data.

To estimate the economic relationship, we need a simplified version of reality. The reality of economic relationship can be highly complex and may not even be fully observable, and it may be even impossible to estimate using the observed data. Hence, economists build a model - a simplified version of reality- that is simple enough to be mathematically and statistically tractable. The model enables economists to estimate the simplified relationship using the data, which sheds light on the reality of the relationship. The model can be regarded as an approximation of reality, and it is the hope of the

economists that their model is a good approximation of the true relationship.

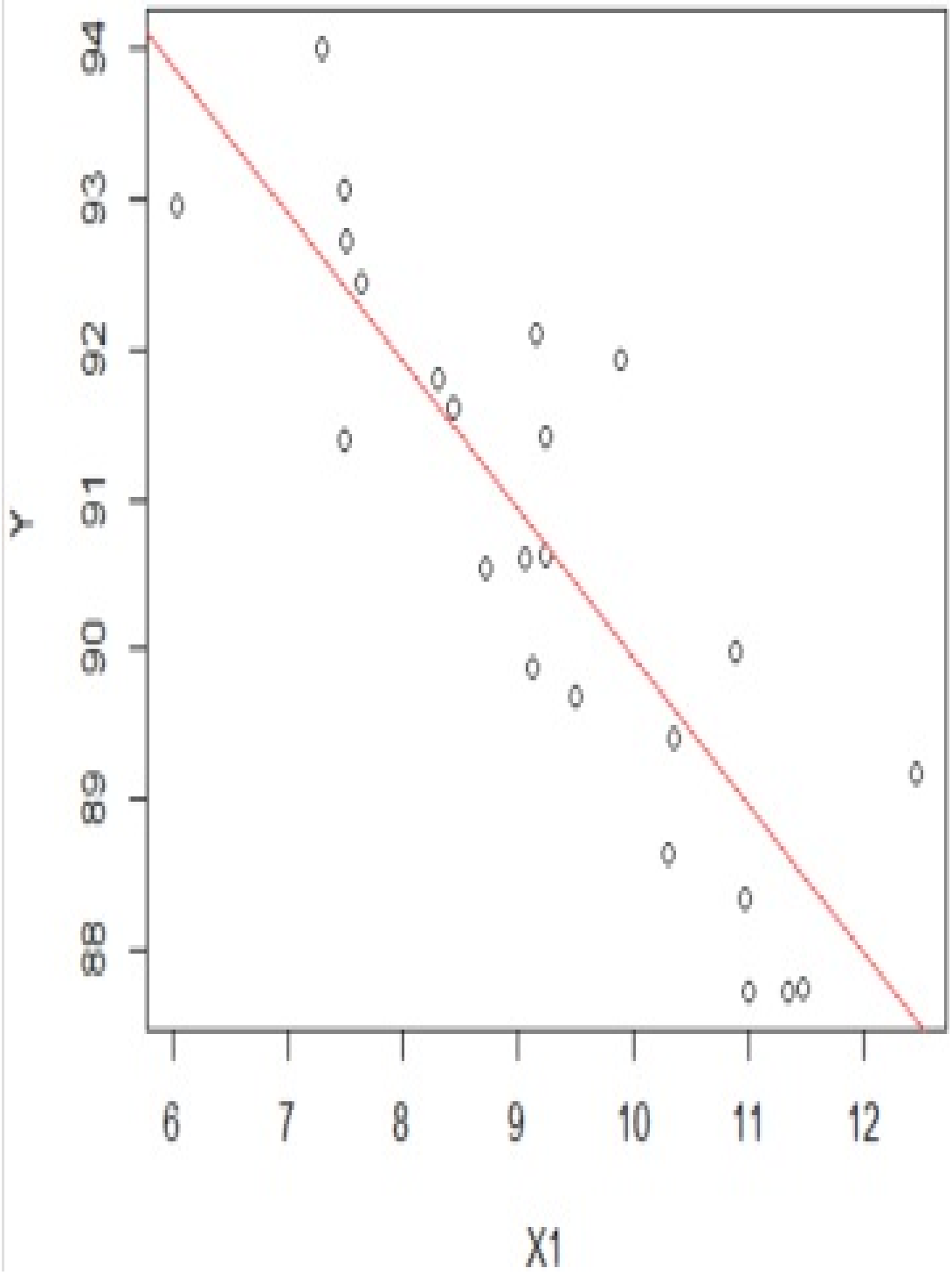
As the famous statistician George Box stated in 1976, “all models are wrong, but some are useful.” The statement means the model is an approximation, which can sometimes be a good representation of reality. In this case, the model can provide useful results about the relationship. However, often the model represents a poor approximation and may provide misleading outcomes of the relationship.

In their training, economists learn how to employ various statistical methods to find a model that is a good approximation of reality. This discipline is called econometrics, which involves statistical and mathematical approaches to modeling economic relationships. A useful model is hard to get, but it provides interesting information from data and can be used by economists to make important predictions and policy decisions.

Suppose an internet content provider is considering an increase in its monthly subscription fees. They are concerned that the number of subscriptions may drop substantially. To estimate this sensitivity, they can consider the following model:

(1)

where Y is the number of monthly subscriptions (in 000’s) and X1 is the price of monthly subscription (in dollars). If the company has the past data for the number of subscriptions and price, then the parameters α and $\beta 1$ can be estimated. Here, Y is called the dependent variable, and X1 is called the explanatory variable. u is an error term that represents the shocks that cannot be explained by the model.


Y
94
93
92
91
90
89
88
6
7
8
9
10
11
12
X1

The past data says that, while the price changed from around $6 to $13 per month, the number of subscriptions dropped from 94000 to 86000. This is a sensitive response, and the diagonal line is the best-fitting line through the scatter of Y and X. The parameter α represents the intercept of this red line, and β1 represents the slope of the line. The line can tell the company the expected change in the number of subscriptions if the price changes from $15 to say $17 per month.

The estimated line is Y = 99.13 – 0.98 X1. This means that, if the price increases by $1, then it is expected that the number of subscriptions drops by 0.98 units or by 980.

This analysis is called regression analysis, which is widely used in many areas of science. It is useful to answer these “how much” questions or to measure the marginal effect of the Y variable regarding X1.

The above model is a simple version of the regression model where only one explanatory variable is used. Economic relations can be complicated, and a dependent variable can be explained by several explanatory variables.

Suppose it is argued that the number of subscriptions also depends on economic conditions. Then the above model may not represent the full economic determinants for Y. Then the model can be written as,

(2)

where X2 is the unemployment rate. It is possible that the deteriorating economic conditions and consumers’ income and confidence have caused the drop in the number of subscriptions. With X2 added to the data set, the following result has been obtained:

Y = 60.12 – 0.11 X1 – 1002.14 X2

This means that, in response to a 1% increase of the unemployment rate, the number of subscriptions has been down by around 1000. The sensitivity to

price is only -0.11, so omitting X2 has over-stated the value of β1. The dependent variable has not been so sensitive to the price, but rather it has been declining because of economic downturn.

Here, the model (1) is a poor approximation to the reality that provided an incorrect and over-stated responses of the dependent variable to a change in the explanatory variable. The model (2) is a better approximation and can be more useful in making correct judgment about the relationship.

For inferential statistics, the researcher can test if H0: $\beta 1 = 0$ using the Z-test or confidence interval. This will reveal statistical significance of the variable X1 for Y. That will tell us whether the estimated value -0.11 is statistically different from 0 or statistically distinguishable from 0, given the sampling variability. A more important question is the economic significance of X1 for Y. That is, whether the sensitivity of the demand or sales to the price is economically important. That can be answered by questioning whether the estimated value of -0.11 matters economically. This is the study of the effect size or signal from data, which we discussed in the previous chapter.

In response to the one dollar increase in the monthly subscription fee, the number of total subscriptions is expected to decrease by around 110, given that the unemployment rate stays at the same level. This response may not be substantial and can deliver the company a higher level of revenue despite losing some customers, if unemployment stays at the same level.

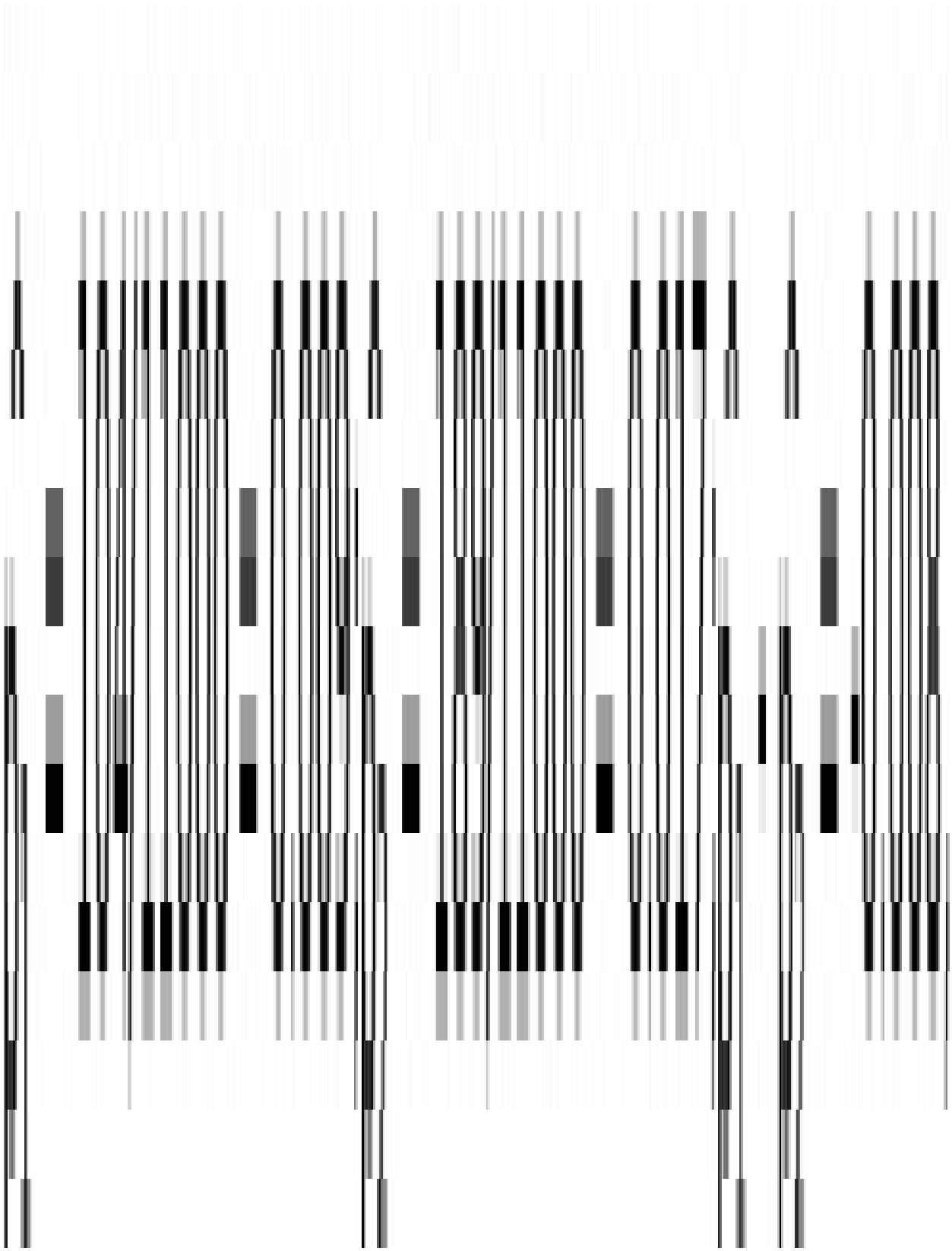
Medical Research

Medical research is another popular application for statistical analysis. When a pharmaceutical company develops a new medicine or vaccine, they need evidence that their new product is effective without a major side effect. This is based on the outcomes of clinical trials that are “research studies performed in people that are aimed at evaluating a medical, surgical, or behavioral intervention.”[xiii] Through clinical trials, researchers learn if a new treatment (a new drug or a medical device) is safe and effective. For this new treatment to get approval from governments or international bodies for public use, they need convincing statistical evidence to show that what they offer is effective and safe.

The structure of a clinical trial is simple. Two samples of participants are selected: one group is given a new drug or treatment, and the other group is given a placebo or no treatment. The effect of the drug or treatment is compared to determine if it makes a significant difference. For example, the COVID-vaccines have been developed, trialed, and approved following a similar process.

As an example, let us take the case of clinical trials for the effectiveness of Aspirin on preventing heart attacks[6]. A three-year study was conducted involving 22,000 males: 11,000 were given regular doses of Aspirin and the other 11,000 a placebo. After the three-year period, the researchers found that 104 in the first group had at least a heart attack, while 189 in the second group.

The statistical evidence can be summarized as follows:

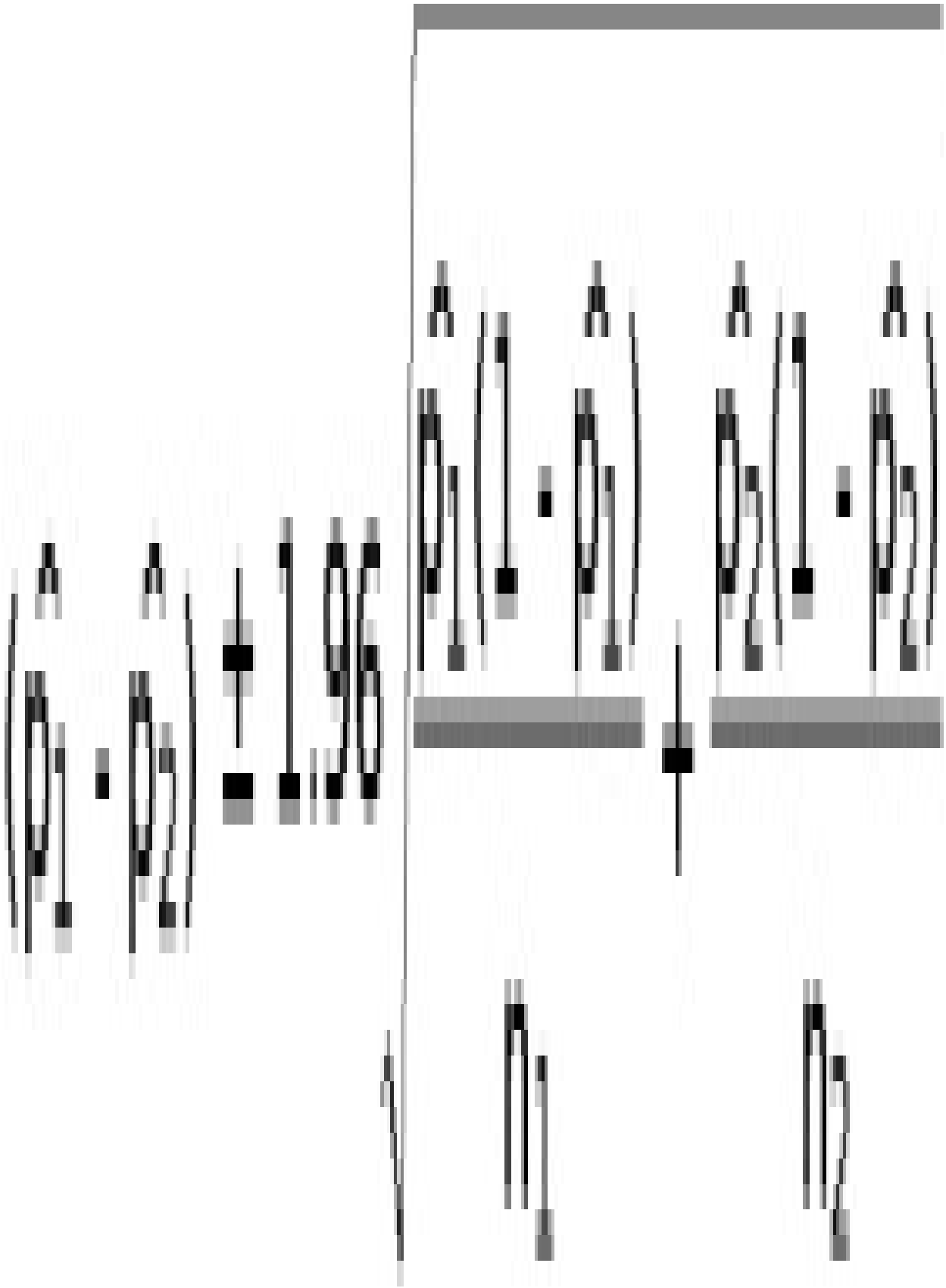


That is, 0.9% of the members in the first group had heart attacks and 1.7% of the members in the second group. If there is sufficient evidence to suggest that 0.9% is statistically different from 1.7%, or the difference -0.008 of (p1-p2) differs from 0, then the effectiveness of Aspirin is established (statistically).

The following hypotheses can be tested:

,

where p1 and p2 represent the population parameters for the proportions. The 95% confidence interval is calculated as:

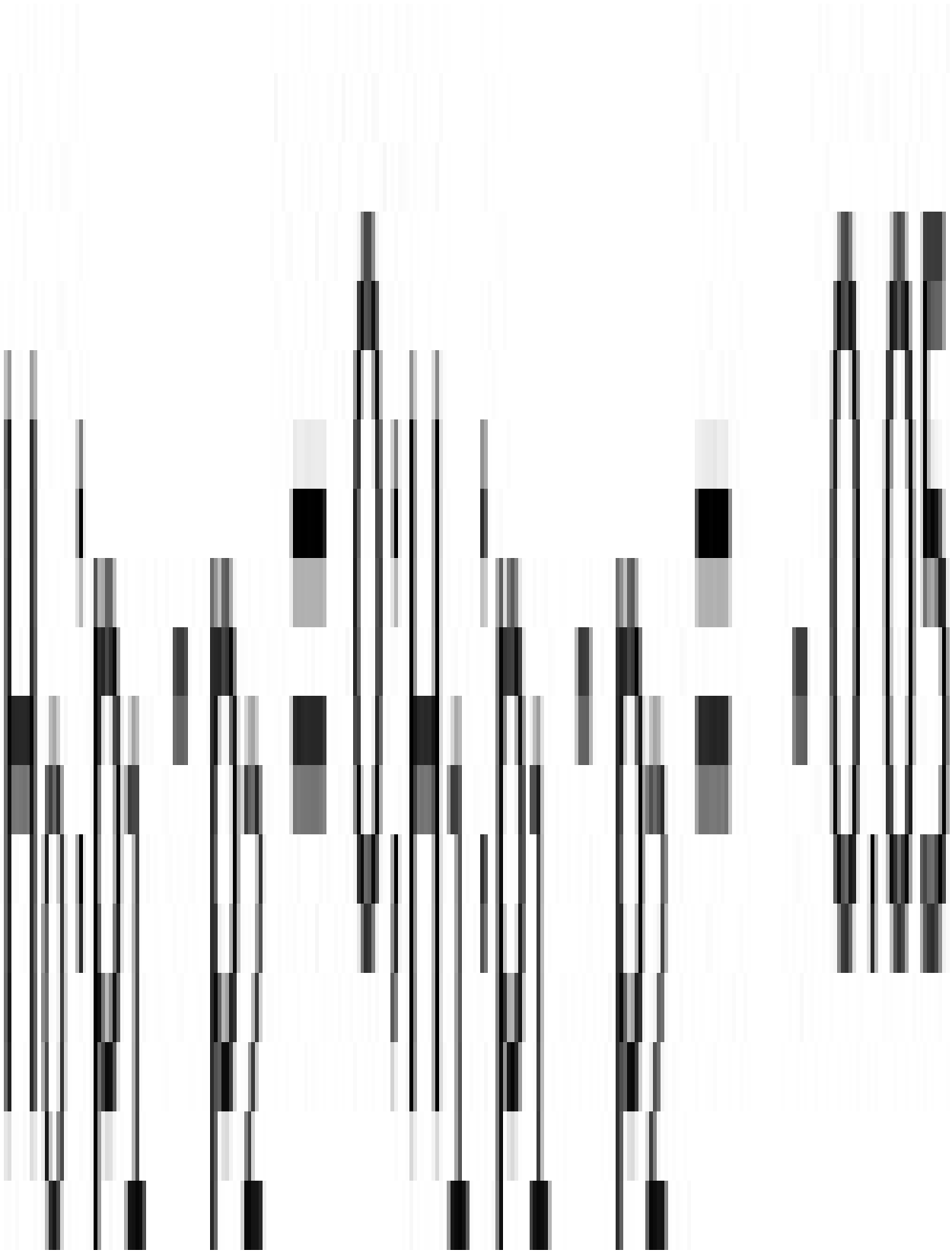


which is (-0.0107, -0.0047). The interpretation is this interval covers the true population value of (p1-p2) with 95% confidence, in repeated sampling. If this kind of trial is repeated 1,000 times, 950 will cover the true value of (p1-p2). Since this one does not cover the value of 0, the statistical evidence suggests the difference in sample proportion differs from 0 statistically, at the 5% level of significance. Hence, the effectiveness of Aspirin has been established statistically. The null hypothesis of is rejected in favor of the alternative hypothesis that . The report may add the headline that Aspirin takers have nearly twice the chance of preventing heart attacks than those not taking it.

The above example is how medical researchers establish statistical significance. Whether the difference of -0.008 is clinically and practically different from 0 is a very important question, but we do not delve into that matter here. However, this issue should be further evaluated critically by qualified health researchers and should be the primary focus of the clinical trial. As stressed, this is the study of the effect size or signal from the sample, which is often disregarded by researchers.

An important point is that statistical evidence should be a secondary issue to practical importance. An ideal situation is where the clinical and practical importance is supported by statistical evidence. But if the former is doubtful or not clearly established, then statistical evidence should be disregarded or treated with caution.

The above method of hypothesis testing is based on the null ritual discussed in Chapter 3. If we follow the Neyman-Pearson decision theoretic approach, hypothesis testing should be done in the following way.



The difference is the alternative hypothesis taking a specific value. This value of -0.05 means the difference should be at least 5% between the two groups to dismiss the null hypothesis that there is no difference and no effectiveness of Aspirin on heart attacks. This value should be chosen by health researchers as the minimum value of difference that is sufficient, practically and clinically, before they obtain the data.

Once this value is chosen, researchers set the value of the level of significance (α), Type II error probability (β), and power, which will determine the sample size. One big problem of the null ritual is that the sample size is chosen with no statistical justification. In the case of the example, no justification is given as to why such a large sample size of 22,000 is chosen. The effect of this will be discussed further in Chapter 5. But, as discussed in Chapter 3, this is a typical case where the test statistic is inflated by the sample size, masking the effect size, which is negligible.

The world has been flooded with such medical research reports. Higher intake of coffee can reduce the chance of cancer, higher intake of processed meat can increase the chance of cancer, and the list goes on. Academic journals publish such clinical studies based largely on statistical significance, with less weight on practical significance. This is largely because the effect size is not fully investigated in statistical research.

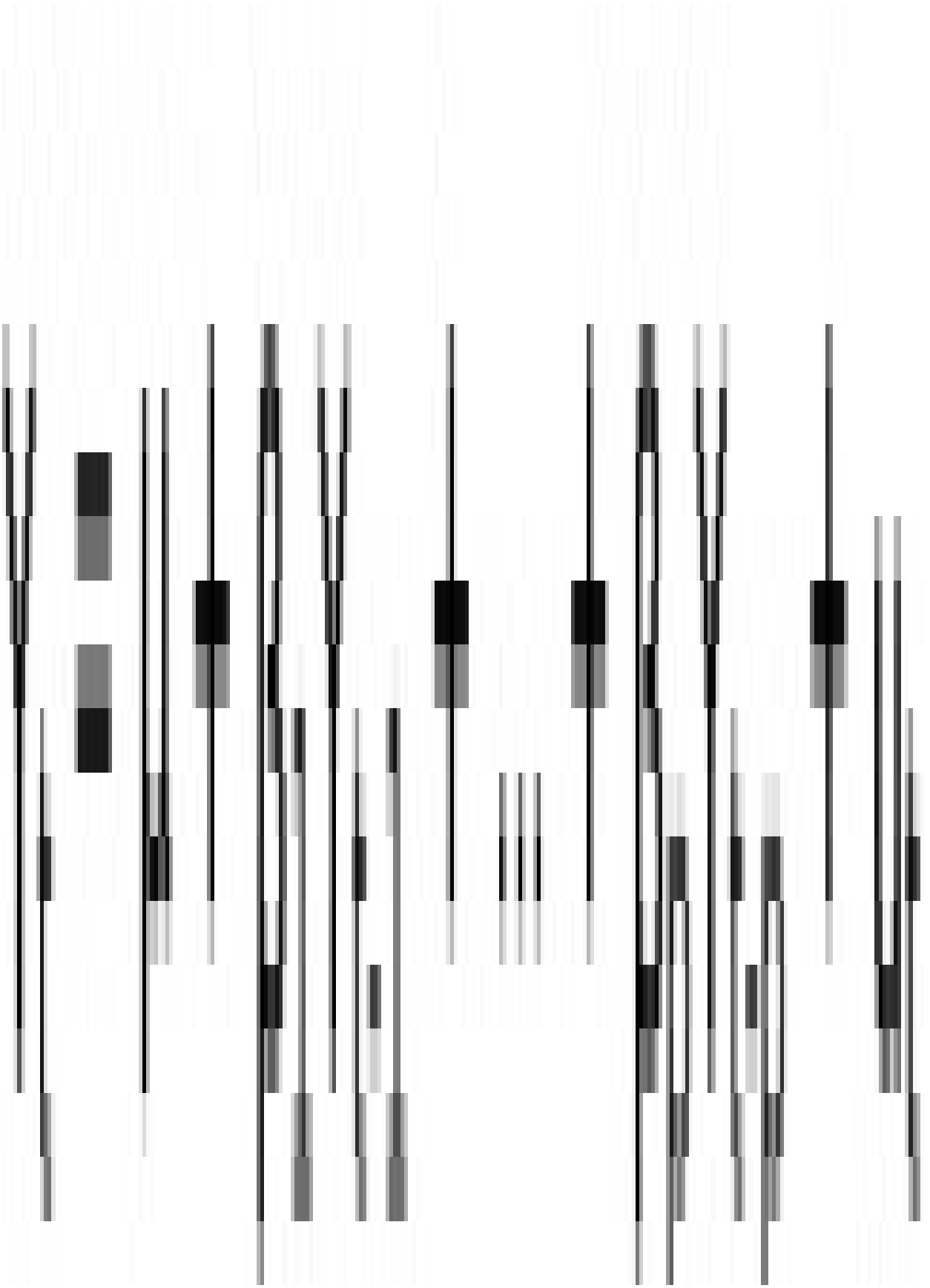
Can this kind of clinical trial be repeated? Statistical significance is a concept and method grounded in the notion of repeated sampling. It is virtually impossible or difficult to expect that a three-year study involving 22,000 men is repeated even twice, not to mention 1,000 times. Can the other independent research studies be conducted in a similar experimental setting? Yes, they can. Can they replicate or reproduce the results? The answer is unlikely. As we shall see in the next chapter, the replication rate in scientific research is less than 50%. That means that a statistical study cannot be replicated by other researchers in most cases. This is called the replication crisis in scientific research, which will be discussed in more detail in the next chapter.

Economic and Business Forecasting

A statistical model or method can be used for the purpose of forecasting. The future values of key economic indicators, such as the real GDP and unemployment rate, are important for decision-makers in business, policymakers in government, and investors.

The most basic and widely used method of statistical forecasting is time series forecasting. Time series is a set of data generated over time often at a regular frequency such as daily, monthly, quarterly and annually. A time series often show a historical pattern. For example, monthly sales of a company can show a steady annual growth of, say, 5%, with peaks in June and December. A steady growth may be related with economic fundamentals, such as long-term economic growth and population growth, and peaks in June and December may coincide with mid-year and end-of-year sales periods. If there is such a strong pattern, then it is likely that it continues. And if we can identify the pattern using a statistical model and past data, the estimated pattern can be projected into the future to generate forecasts.

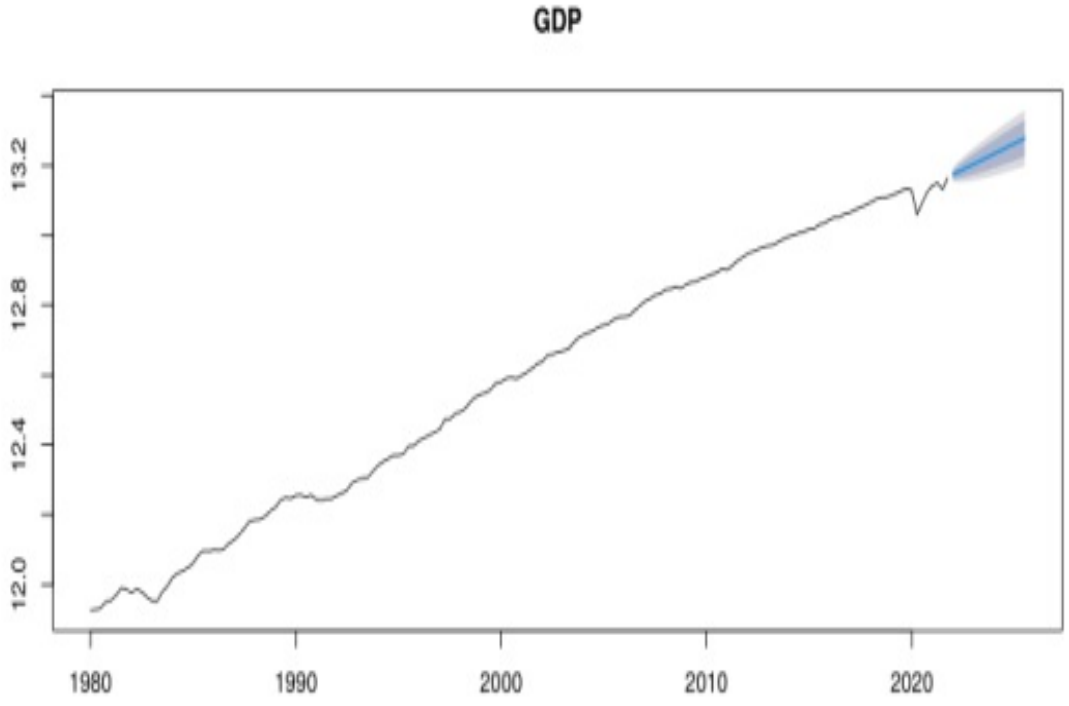
A simple model to illustrate this is an autoregressive model, which can be written as:

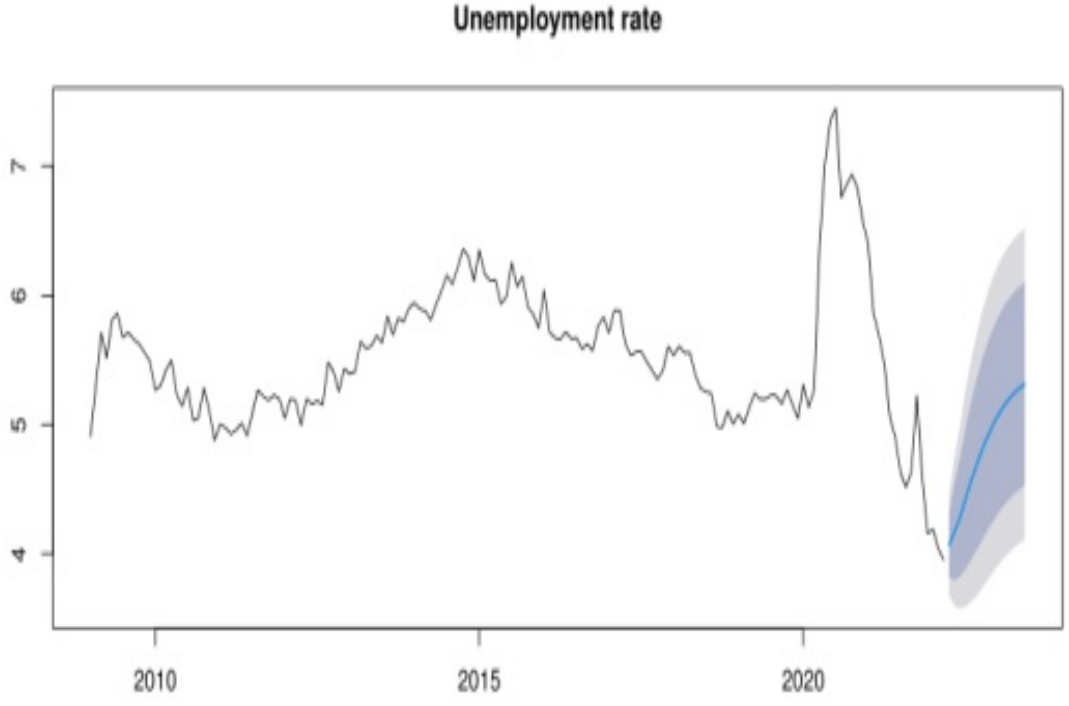


where Yt represents the current value of a time series variable and Yt-k the value of time series k period ago. The above model specifies that the current value of time series depends on the past values of its own with the lags of 1 to p with the long-term mean value of μ and the coefficients of β's. The error term ut captures the unexpected or non-fundamental shocks that cannot be explained by the model.

The value of μ and β's are unknown and should be estimated from the data, along with the value of p. There are several statistical methods to find the value of p, and there are alternative ways to estimate the values of β's. Once the values of these parameters are estimated, the model can be used to generate forecasts.

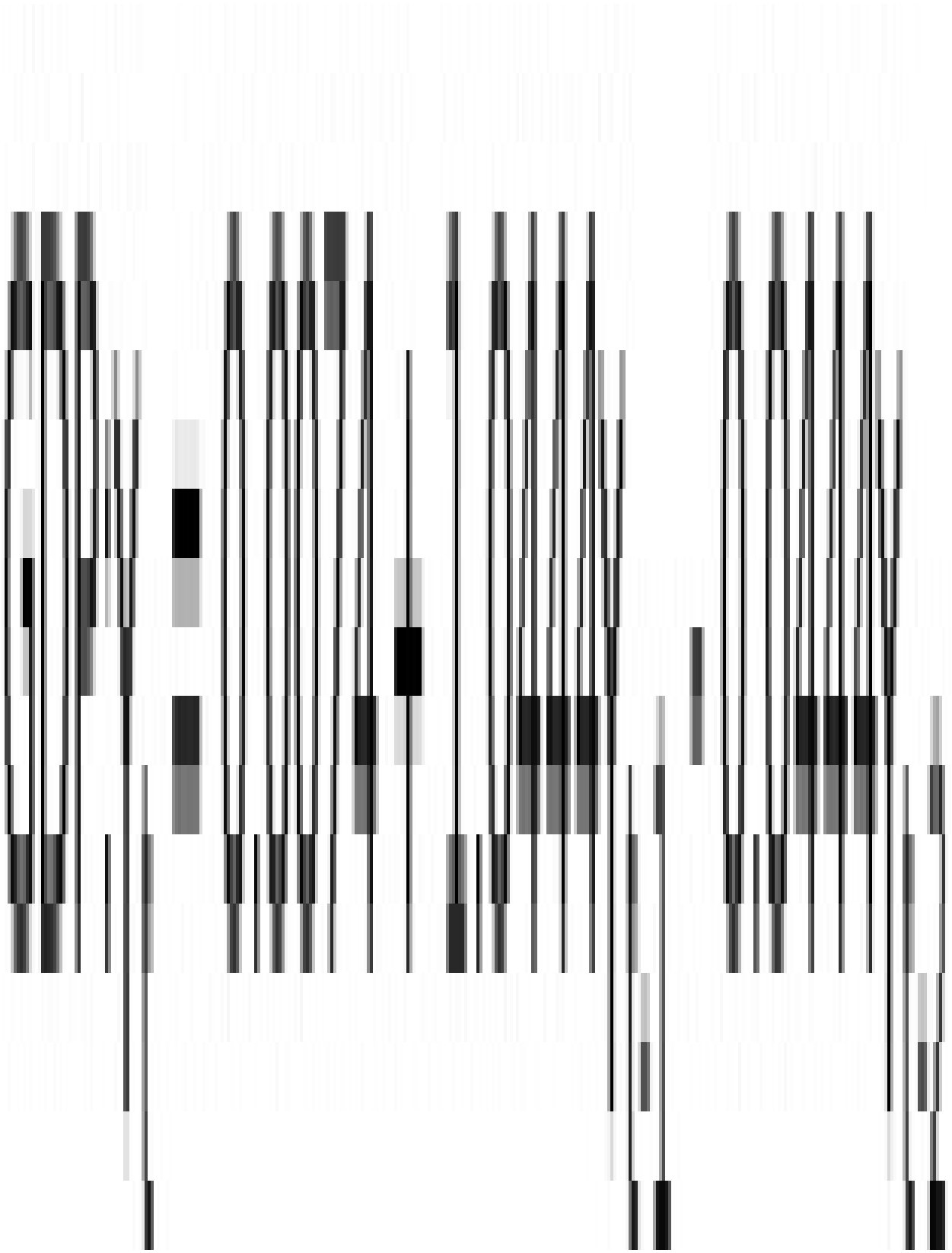
As an example, we use two macroeconomic time series from Australia, obtained from FRED (Federal Reserve Economic Data, https://fred.stlouisfed.org/). The real GDP (seasonally adjusted, domestic currency) for Australia is quarterly from 1980 (quarter 1) to 2021 (quarter 4) with 168 observations, and the unemployment rate (aged 15 over, all persons in Australia, seasonally adjusted) is monthly from January 2009 to March 2022 (267 observations). These time series are plotted in the Figure below.

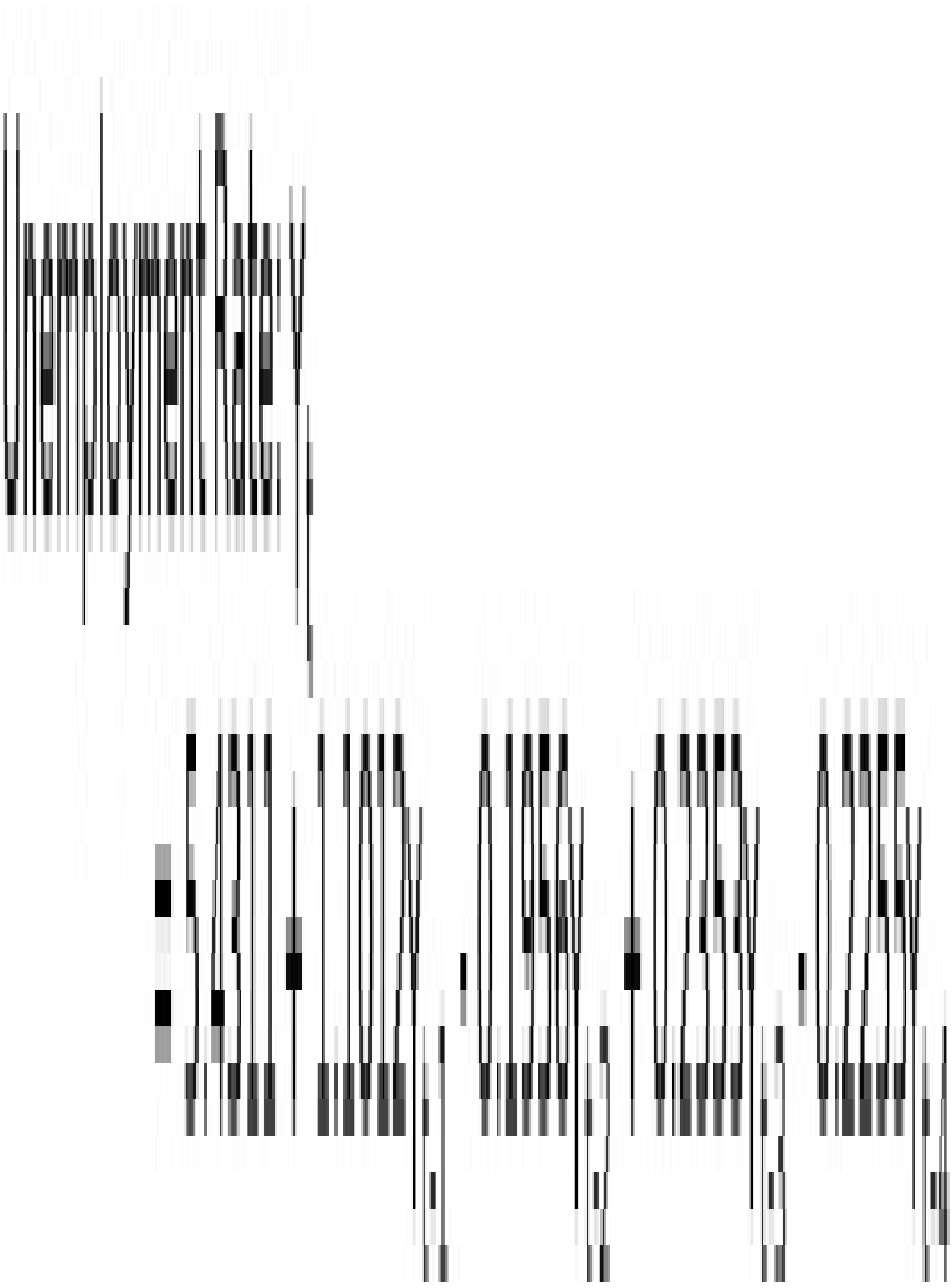

GDP
13.2
12.8
12.4
12.0
1980
1990
2000
2010
2020


Unemployment rate
7
6
5
4
2010
2015
2020

The real GDP (in natural logarithm) is showing a clear upward trend with small cyclical variation around this trend. This means the Australian real income is increasing at a steady rate, but there is a small degree of fluctuations depending on economic booms or recessions. The unemployment rate shows no trend, but it is showing fluctuations around its long-term mean of 5.53%. Because of the COVID pandemic, the unemployment rate is showing high fluctuation around 2021, and the real GDP shows a sudden decline around the same time.

The model estimation results are summarized as follows:





The estimation results show the dynamic patterns of the time series. The real GDP depends on its value in the past quarter and the value two quarters ago with coefficients of 1.04 and -0.04. The unemployment rate is showing a longer degree of dependence with up to 4 quarters.

The above estimation results are projected into the future for forecasting. The line at the end of the second graph shows the mean of the forecast and the gray area shows the distribution of the possible future values with darker areas indicating the area with a higher chance. These models can often generate accurate and reliable forecasts, but again, past performance does not guarantee future performance. As George Box state, “all models are wrong, but some are useful.” The final decision about the future outcome should be made by combining all available information, including the forecasts generated from these time series models.

These cases are simple examples, and the method adopted in the real world of business and economic forecasting can be much more elaborate and complicated. However, the principle is the same. The model identifies the past pattern or history of the time series in a simplified version of the real world, and this identified historic pattern is projected into the future.

The above example is quantitative forecasting based purely on statistical methods, with no input from intuitions or opinions. When there is no or inadequate historical data, then quantitative forecasting methods may not work satisfactorily. When the economy or market faces a new and unexpected situation (such as the outbreak of a pandemic or the onset of a global financial crisis), it is likely the forecasts from quantitative models may be useless. In this case, judgmental forecasting methods should be used. These methods incorporate intuitive judgments, opinions, and subjective probability estimates. A simple example is a survey of expert opinions, which can be useful when quantitative forecasting is likely to fail or be insufficient.

Stock trading and portfolio selection

Is the stock market efficient? This is a big question for many professionals in the fund management and stock trading industry, as well as lay investors. If the market is (perfectly) efficient, all prices fully and instantly reflect all available information, and no investor can make abnormal profits consistently. If the market is inefficient, the price under-reacts or over-reacts to the desired level implied by the available information with a delay, and there is room for abnormal profits for those investors who can time the market.

The trading strategies will be different under the two states of nature. Under market efficiency, there is no point in active trading. Investors will be better off by adopting a buy-and-hold strategy. Only long-term investors will be the winners who will ride out the market volatility, and their profits will be driven by the long-run economic and market fundamentals. Under market inefficiency, active trading can bring abnormal profits, and traders will adopt a range of methods to time the market by selecting the stocks that are undervalued or overvalued. Many fund managers and traders operate under the implicit or explicit assumptions that the market is inefficient. Many quantitative and statistical methods are available to them, and new methods are being developed with the increasing availability of state-of-art statistical methods and investment products.

Are the stock markets efficient or inefficient? Finance academics over many generations from the 1960's have tested hypotheses using a range of statistical methods. A search in Google Scholar with the term "stock market efficiency" reveals over three million academic papers (most are published). However, the evidence is still inconclusive and the controversy ongoing. In the era when stock trading can be made with a click of a mouse, we are still not sure if the pricing of a stock is accurate.

However, one thing is clear. Academics proved that perfect and absolute

efficiency is impossible because, if the price really reflects all information and makes accurate adjustments instantly to the desired level, then there is no motivation to trade. In response to this argument, a modified version of the efficient market hypothesis is proposed, which is called the adaptive markets hypothesis. [xiv]

Some argue the market is adaptive and changes continuously depending on the prevailing market and economic conditions. The state of the market cannot be dichotomous between being efficient or inefficient; rather the market can depart from these two extreme states and show market inefficiency or return predictability occasionally. Hence, the market can be inefficient in some episodes of time depending on the prevailing conditions. The observed inefficiency may disappear as soon as the market participants exploit them. Hence, the consensus is that the market is nearly efficient most times, but episodes of profit opportunities may arise occasionally. Implications to stock trading are that both buy-and-hold and active trading strategies are justified. There are broadly two kinds of fund managers in the asset management industry: those who provide products with a buy-and-hold strategy (such as index fund) and those with active market-timing strategies.

The strategies adopted by active strategies are numerous, ranging from simple ones to what they call state-of-art, including machine learning and artificial intelligence. The latter has gained popularity with the arrival and availability of big data: some methods go down to the transaction levels or minute-by-minute data, estimating models with millions of data points, but it is not certain whether they outperform the simple ones.

Here are a few simple ones. The momentum strategy selects the past winners to form a stock portfolio, with an expectation that the price appreciation will continue. The contrarian strategy selects past losers, with an expectation they will become the next winners. There are chartists who pick up the buy-signal or sell-signal by scrutinizing graphical presentation of the past price movement and develop a trading rule. For example, they smooth out the past price to reveal the trend of the price and determine the current level of under-performance or out-performance using the estimated trend line.

A class of investment strategy that relies heavily on statistical method and thinking is factor investing. The strategy selects the stocks based on the

factors identified as the main drivers of the return based on statistical methods. Their starting point is a fundamental model called the capital asset pricing model, which can be written in its simple form as:

(1)

where Ri is the return for stock i and Rm is the return from a general market index portfolio. The coefficient “alpha” here captures the stock’s outperformance of the market, because it is the level of the stock’s return even if the market returns zero profit. The coefficient “beta” captures the sensitivity of the stock’s return to the risk from the market. If beta = 1, the stock’s riskiness is the same as the market’s since it changes one-to-one with the market movement on average; if beta is > 1, then the stock is riskier than the market because it moves more sensitively and vice versa if beta is < 1.

Suppose there is another risk factor, X1, that can explain the stock’s return and movement, then the asset-pricing model becomes:

(2)

Here the coefficient of the additional factor β1 captures the stock's sensitivity or riskiness to this additional risk factor. Suppose this factor X1 represents the returns from small companies. Then the stock i is said to be driven by the size factor, and its riskiness in relation to this factor is explained by the corresponding beta coefficient. If the stock provides positive alpha, then the stock can provide the profit above the market level, when adjusted with risk associated with the general market conditions and size factor. This information is used when fund managers select stocks to formulate their portfolio.

There are hundreds of factors proposed in the market and from academia. This is called the factor zoo, and finance researchers have identified over 400 such factors published in top finance journals.[xv] There will be more discussions about this later, but such proliferation of factors is closely related to the abuse or misuse of statistics. It is likely that most of these factors are false, and they have been obtained as an outcome of statistical significance without economic significance. Only a handful of factors will make economic sense with persistent performance.

Broadly, there are two group of factors: macroeconomic factors and style factors. The former includes the key macroeconomic indicators, such as economic growth, inflation rate, and interest rate, and the latter includes the size, value, momentum, quality, and volatility. For example, you will see many index funds or ETF (exchange-traded funds) with specific themes, such as momentum, emerging markets, or low volatility. These portfolios are chosen based on factors using the methods similar to the one above.

Risk management

The aftermath of the global financial crisis in 2008 revealed several deficiencies in the practice of risk management and financial regulation of banks and financial institutions. In response to these problems, the Basel Committee of Banking Supervision[xvi] has initiated a range of new measures for a tighter regulatory framework. This is a committee of banking supervisory authorities established by the central bank governors of the "Group of Ten" countries in 1974. As of 2019, it has 28 member countries. They reached what is called the Basel III Accord, which sets new standards for the new regulations that banks and financial institutions need to follow. Most of them are for the banks' capital and liquidity requirements for a range of risks including market risk.

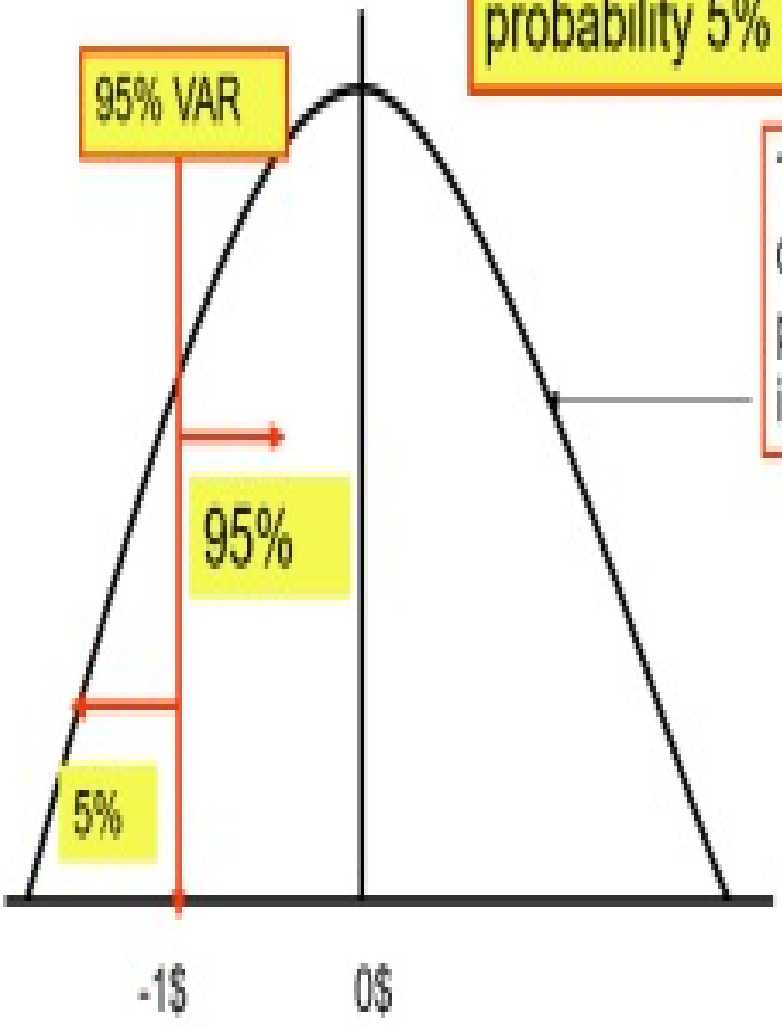
The value-at-risk is a statistical measure of financial risk under normal market conditions. The Basel III Accord requires that banks and trading houses set aside reserve capital based on the value of their value-at-risk, which should be reported in their financial statement.

It is determined in a similar manner to the confidence interval discussed in Chapter 2. The 95% value-at-risk is the minimum expected loss with a 5% confidence level for a given time horizon. It is the minimum loss that would occur once in 20 days.

_____

95% VaR of a portfolio

When 10$ is invested, the worst outcome with probability 5% is losing 1$



The return distribution of a portfolio when 10$ is invested

There is 5% chance that your loss is more than 1$. That is, a loss of more than 1$ is expected once in every 20 days.

The above figure is an illustration of 95% of a portfolio where the return distribution is the black curve. When the investment is $10, the minimum loss is 1$ in 20 days, and the bank should set aside a reserve capital based on this figure of value-at-risk.

Suppose, as a simple example, the portfolio return at time t follows a normal distribution with the mean μt and standard deviation σt. Then, the 95% value-at-risk is defined as

In reality, the return distribution is not normal, and the value of σt changes over time with little-known dynamics. Hence, the challenge of professionals in risk management is to identify the distribution of their portfolio return and find a suitable dynamic model for their standard deviation. A popular one is called the GARCH (generalized autoregressive conditional heterskedasticity) model and widely used in the finance industry.[xvii]

Concluding remarks

This chapter has provided several examples of applying statistical methods using both descriptive and inferential statistics we have discussed in earlier chapters. The aim was to show the readers how these methods are being applied to real-world problems. With the increasing availability of big data, statistical methods are also becoming more and more elaborate and technical. However, the principle is the same. A sample, which is a fair representation of the population, should be examined with a sound research design and statistical thinking. The results should be interpreted with care, combining all available information from a range of descriptive and inferential statistics. The final decision should be made in consideration of the degree of uncertainty and consequences of possible errors. However, the statistical methods are also widely misused, abused, or misinterpreted, which is the topic of the next chapter.

Chapter 5: Misinterpretations of Statistics

_____

Statistics are widely being misinterpreted, misused, and abused. It is rampant in scientific studies, media reports, business decisions, and government policymaking. And it is affecting our lives, justice, and jobs (Ziliak and McCloskey, 2008). It does not get better with big data; it may get worse(Harford, 2014). It is happening mostly unknowingly, but sometimes wilfully, by the academics who have to publish their papers for success in their grant applications or promotion, by the journalists who need eye-catching articles, by the lawyers hoping to win their cases, and by the politicians who need to win the next election.

The problem is so serious that the American Statistical Association, the world's largest community of statisticians, has recently issued two successive statements on this issue (Wasserstein and Lazar, 2016; Wasserstein et al., 2019). The statement raises serious concerns about the distortion and damages that the widespread misinterpretation of statistics generates, also spelling out the principles for "good statistical practice". Yet, many practitioners of statistics, not to mention the lay people, do not understand the nature and extent of the problems and the good practice to follow. Their textbooks and lectures on statistics tell virtually nothing about the problem. This chapter is dedicated to the issue of the misinterpretation of statistics, covering the materials that our textbooks and lectures do not tell us. We hope to get down to the bottom of the problems, which can help us furnish sound statistical thinking and conduct "good statistical practice".

Illusion of Statistical Significance.

As discussed in Chapter 2, statistical decisions are often made based on statistical inference (or inferential statistics), where a null hypothesis (H0) is tested against an alternative hypothesis (H1). A null hypothesis is formulated on the value of a population parameter, such as the correlation coefficient (ρ) that measures the degree of linear association between two variables. A correlation is a unit-free measure that lies between -1 and 1. A value closer to 1 (-1) means a positive (negative) and strong relationship, while a zero value of correlation means no linear association. The decision to reject or not reject H0 is based on the value of Z-statistic, which measures how much the sample correlation coefficient differs from the value of ρ under H0. Readers are encouraged to refer to Chapter 2 for further details.

Consider a simple case of statistical inference where H0: $\rho = 0$, H1: $\rho \neq 0$ where ρ is the population correlation coefficient. In this case, the Z-statistic, again as a scaled signal-to-noise ratio as discussed in Chapter 3, is calculated as

(1)

where r is the sample correlation coefficient. The Z-statistic measures the distance of r from $\rho = 0$ (the value under H0), scaled by a factor that makes it approximately follow the standard normal distribution. If it is too big or too small (if the value of r is far away from 0), we reject H0 in favour of H1; if it is close to 0, we do not reject H0. How big is too big or how small is too small to reject H0? The critical values are determined by the level of significance, which represents the probability of a Type I error (rejecting H0 when it is true). If we choose the conventional level of significance, 0.05, this means we allow a Type I error to occur once in twenty times. The critical values at the 5% level of significance are -1.96 and 1.96, and we reject H0 at the 5% level of significance if Z-statistic is less than -1.96 or greater than

1.96.

When we reject H0: $\rho = 0$, the sample correlation is said to be statistically different from 0 or statistically significant at the 5% level of significance. The p-value measures how incompatible the sample (or Z-statistic) is with the distribution under H0: $\rho = 0$ (Student-t distribution). It is small when the Z-statistic is too large or too small, so the sample is highly incompatible with H0: $\rho = 0$. Equivalently to the decision rule based on the critical values, we reject H0 in favor of H1 at the 5% level of significance if the p-value is less than 0.05.

Table 1. correlation coefficient and Z-test results

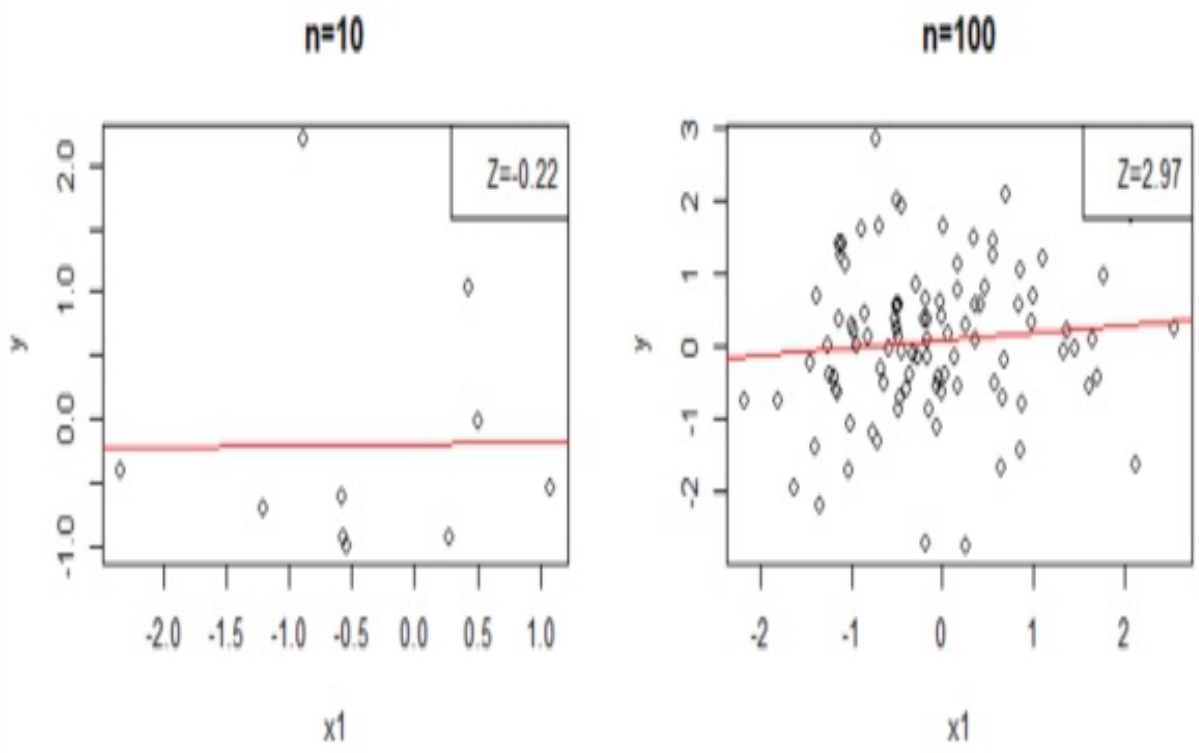
Study	1	2	3	4
n	10	100	500	1000
r	-0.08	0.29***	0.22***	0.17***
Z	-0.22	2.97	4.93	5.54
p-value	0.83	0.00	0.00	0.00

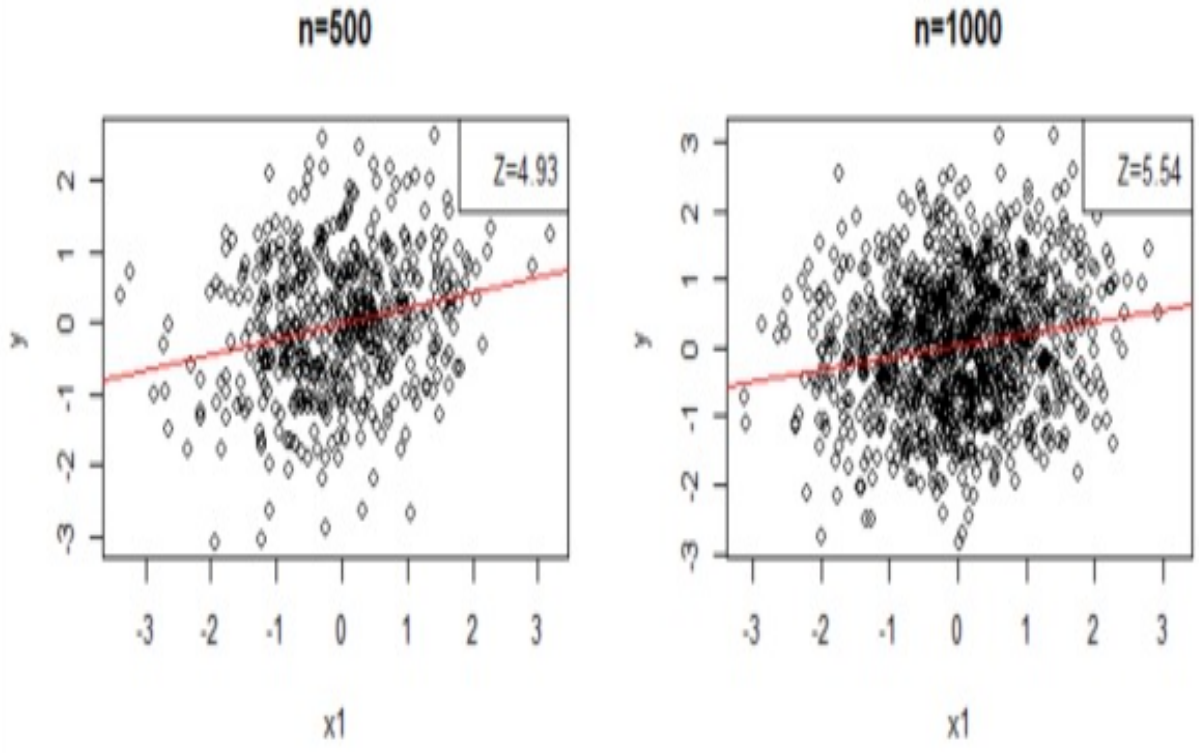
***: statistical significance at the 5% level of significance

_____

Suppose four researchers estimated the correlation coefficient for the variables y and x1 from their independent studies, respectively with sample sizes 10, 100, 500, and 1000. The population correlation coefficient is 0.2, which is unknown to the researchers. Table 1 reports the values of the sample correlation coefficients and their statistical significance. With three asterisks on the sample correlation coefficients, researchers 2 to 4 can be seen to have found strong linear relationships. This impression or illusion can further be reinforced by a large Z-statistic and infinitesimal p-values. Figure 1 presents scatter plots for the variables x1 and y for each study. For all cases, the degree of linear association is negligible visually, all consistent with a relatively low population correlation coefficient of 0.2. The key question is whether the association is strong to justify three asterisks. Those who only have seen the scatter plots may well not agree.

Figure 14. Scatter plot and Z-statistics ($\rho = 0.2$)


n=10
Z=-0.22
y
x1
n=100
Z=2.97
y
x1


n=500
Z=4.93
y
x1
n=1000
Z=5.54
y
x1

This phenomenon occurs since the test is designed to reject H0: $\rho = 0$ when the value of r is sufficiently different from 0. Such ability to reject the value different from 0, albeit small, increases sharply with sample size. It is apparent from the inspection of the formula given in (1), where the t-statistic is an increasing function of sample size. The question is whether we should take seriously the values of sample correlation in the neighborhood of 0.2. With large values of Z-statistic (in absolute value) and small p-values, we are led (or misled) to a belief that such a small value of correlation is practically important and that the two variables are "strongly" correlated.

From a survey of professional economists, Soyer and Hogarth (2012) report that their respondents have shown better performance in predicting the linear relationship among the variables when presented with graphs than when presented with inferential statistics. This result demonstrates that top-level professionals can also fall into the same illusion of statistical significance. A sound statistical analysis should also include a range of descriptive statistics and a study of effect size, besides inferential statistics.

Another point of concern from this illusion is that the researchers are strongly motivated to increase their sample size, only to increase the value of test statistics (in absolute value) or to reduce the p-value and gain statistical significance. With a range of new technologies that make storage and transmission of big data so quick and cheap, using massive sample size (as large as 10,000 or larger) is common (see, for example, Kim and Ji, 2014). The challenging question is whether a meaningful effect is present, and the investigation should go beyond statistical significance to evaluate the practical importance of effect size.

Illusion of a Massive Sample.

Let us revisit the example in Chapter 4. Selvanathan et al. (2017) provide an example of a three-year medical study that examines the effectiveness of

regularly taking aspirin in reducing the incidence of heart attacks. The study involves 22,000 men, with one group of 11,000 taking aspirin and the other half taking a placebo. From the first group, 104 (0.9%) men developed heart attacks and 189 (1.7%) men from the second group. The difference in the proportion is 0.00773, which is statistically significant or different from 0, with a Z-statistic of 4.99 and a p-value of 0. On this basis, the study finds "overwhelming evidence to infer that aspirin reduces the incidence of heart attacks among men," with a conclusion that "if 1 million men start taking aspirin, around 7730 will avoid heart attacks."

The use of a large sample size has given researchers a wrong impression about the effect in this case, misled by the value of Z-statistic heavily inflated by the sample size of 22,000. If 100 men were taking the aspirin, we can save less than 1 man from a heart attack. It is hard to assert this is overwhelming evidence of the health benefit of taking aspirin. There could be other more effective and safer alternatives to reduce heart disease than taking aspirin. Here, statistical significance only refers to the fact that the difference in the proportion of 0.0073 differs from 0 statistically, but it does not necessarily mean the effect is also practically important.

Big data hubris: misinterpretation of the central limit theorem?

One of the first big topics we learn from statistics is the central limit theorem. In layman's terms, the theorem says a larger sample provides a more accurate statistical result. For example, a correlation coefficient can be more accurately measured when the sample size is 1000 than when it is 10. This applies not only to the sample correlation coefficient (r) but also to statistical inference regarding the decision as to reject H0 or not. A widely misunderstood point about the central limit theorem is that we should always seek a larger sample and that a larger sample is always better than a smaller one.

The key point is that the central limit theorem is all about approximating the unknown distribution when the sample size is small, using the distribution known when the sample size is infinite. Its message is not that we should always chase a larger sample. This is related to the hype going on with big data, also known as big data hubris. Big data hubris is “often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.” This misunderstanding is so deep and widespread even in academia that many journal editors often reject a paper simply based on a small sample size.

The central limit theorem is valid under a range of strict assumptions. The most fundamental is a random and independent sampling from a population with the same mean and variance. It is unlikely the data collection process of big data satisfies all these theoretical assumptions. The “found” data collected from various disparate sources, such as the internet, social media, or public reports, may be incompatible with such strict theoretical conditions. More importantly, as the sample size increases, the assumption of randomness may be compromised. This assumption requires that a sample is a small subset of the population, which conflicts with the big data.

Sampling bias.

Conventional statistical methods are built on the assumption that the data is collected from a well-defined population, and sampling is designed in such a way that a sample is a fair representation of the population. If this assumption is violated, the desirable properties associated with the modern state-of-art statistical methods promise will fail to hold. If data is not collected with a careful sampling design, it may well contain systematic bias, and this bias will not die out as the sample size increases. A large sample size can magnify the bias in a dramatic fashion: see Kaplan et al. (2014).

The classical example is the polling debacle that unfolded in the wake of the 1936 U.S. Presidential Election[7].[xviii] [xix] [xx] The poll by the magazine called Literary Digest predicted a landslide victory of the Republican candidate Landon, while a Gallup poll predicted the opposite result of landslide victory of the Democratic incumbent Roosevelt. The Literary Digest is based on 2.27 million responses, big data even by today's standard, while Gallup counted only about 3000. So why did Literary Digest fail? Their sample was simply biased. The magazine sent out over 10 million postal votes drawn from automobile registration lists and telephone books, which were the items of the privileged in 1936. Added to this was the non-response bias, with less than 25% of the total votes returned. A massive sample size did not help, since it further magnified the bias. In contrast, Gallup collected a purely "random" sample based on quota sampling, which was likely a representative subset of the U.S. voters. This example is a classic example of how random sampling is important and how the sampling bias can be magnified with a massive sample.

Cherry-picking.

Many of the published statistical results may have gone through the process of cherry-picking. Often, the results can be unexpected but surprising, grabbing the attention of the journalists, politicians, and even the editors of academic journals.

Figure 15 presents a well-known example of a jellybean study. Scientists hypothesize that intake of jellybeans causes acne. They keep trying the test with different sets of data until they have found a statistically significant outcome. They stop there to write a paper to be submitted to a journal for publication. This astonishing result is what journalists are waiting for. While this seems like a joke, it is not much different from the reality of scientific research. For example, the public is often inundated with conflicting research outcomes in the news about health benefits of something. Taking coffee, for instance, there are research findings that state drinking coffee regularly can increase the chance of certain cancers or reduce the chance of others, or it can even bring certain health benefits.[xxi]

If one keeps repeating the test for significance with different sample realizations, a statistically significant outcome will turn out sooner or later. In the jellybean example, the test is conducted at the 5% level of significance, so the probability of a Type I error is set at 0.05. There is a 1 in 20 chance that the true null hypothesis of no effect is rejected. If the researchers are repeating the same experiment with different data sets twenty times, they are most likely to find at least one false positive result.

Figure 15. The Jellybean Example


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLANNING PHARMACY!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (p > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT PHARMACY!
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (p > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (p > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (p > 0.05).
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (p > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (p > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (p > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (p < 0.05).
WOW!
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (p > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (p > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (p > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Picture source[xxii]

This process is more formally known under the names of data-snooping, data mining, or p-hacking in academia. Researchers fall into this process often unknowingly, but sometimes wilfully. The problem is so serious that a Presidential Address of the American Finance Association is dedicated to the problem of p-hacking: see Harvey (2017). The consequence is the accumulation of false stylized facts (Wasserstein and Lazar, 2016), an embarrassing number of false positives (Harvey, 2017), and a replication crisis (Peng, 2015). The replication crisis refers to a deep problem in many fields of science where many published results cannot be reproduced or replicated by subsequent independent studies. A comprehensive study conducted in psychology concludes that only about 40% of the published studies are replicable (Open Science Collaboration, 2015).

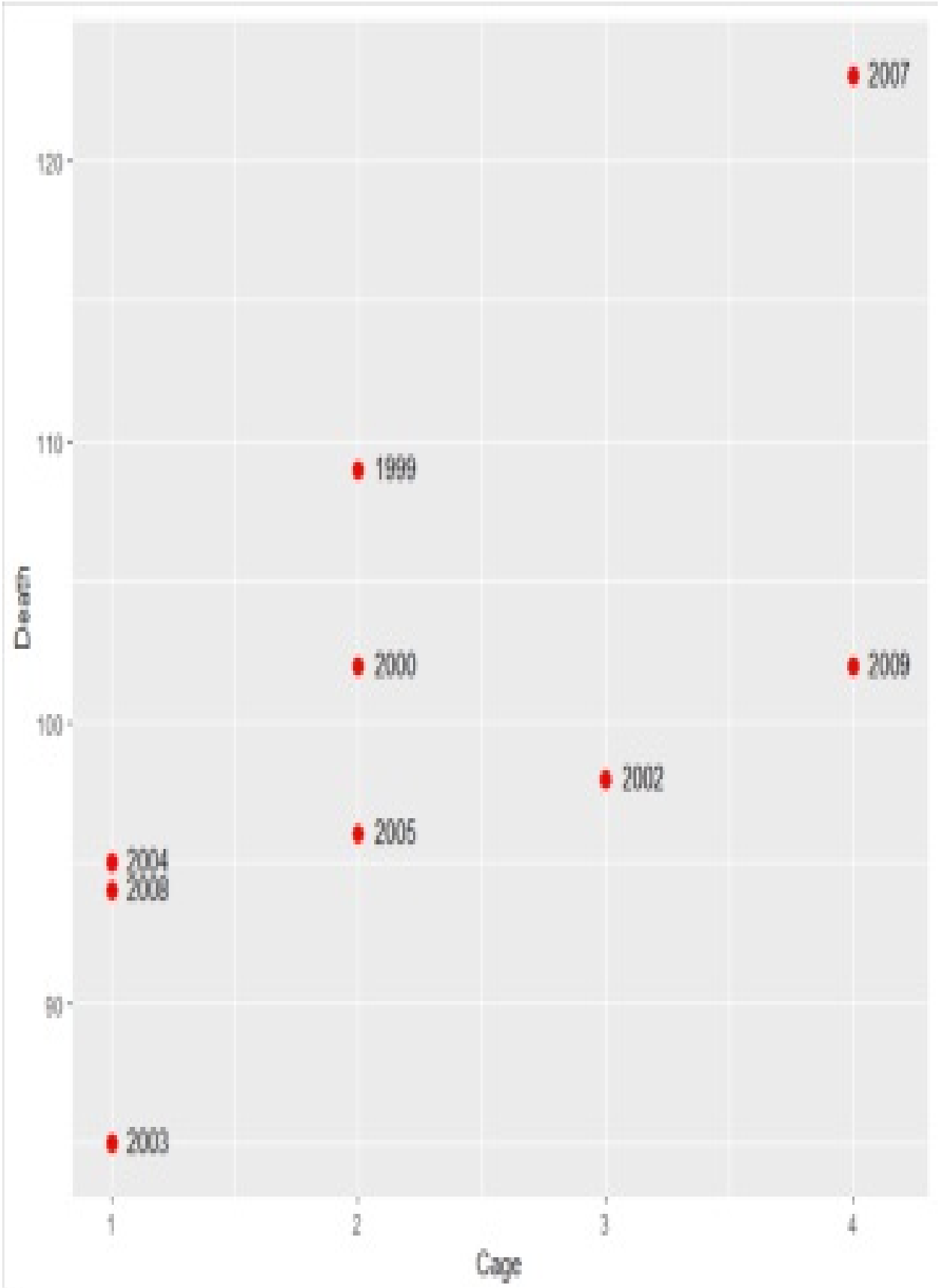
A solution to this crisis is to conduct a sound statistical inference that requires “full reporting and transparency”: see Wasserstein and Lazar (2016). Researchers should reveal the full details of their research, including the number of data sets and hypotheses tested before they come to their conclusion.

Correlation, not causation.

Correlation does not mean causation. However, correlation is often misinterpreted as causation. A good example is the Google Flu Trend, which was launched in 2008 as a forerunner of big data analysis. It was built on a statistical model predicting the number of flu outbreaks based on an aggregate of Google's search queries. For example, if a higher number of search queries with the words such as "fever" and "pharmacy" occurs for a city or region, the model predicts a higher number of flu outbreaks. It started with promising prediction outcomes, but by 2013, it was producing grossly over-estimated predictions for flu outbreaks and finally shut down in 2015. In explaining the epic failure of Google Flu Trends, Harford (2014) argues that Google engineers were merely chasing correlation and patterns, not causation, adding that "a theory-free analysis of mere correlations is inevitably fragile."

A statistically significant correlation can occur as a result of pure coincidence, as the following example shows: the number of drowning deaths in the U.S. and the number of movies in which Nicolas Cage appeared. As reported in Figure 16, a clear positive linear relationship between the two is evident. The correlation is statistically significant with the sample correlation coefficient of 0.67 and Z-statistic of 2.67. While this is a clear coincidence, any decision to implicate Nicolas Cage in drowning deaths should be a Type I error. The null hypothesis of $H0: \rho = 0$ should not be rejected based on practical unimportance. A similar example is the number of drowning deaths in the U.S. and the number of ice creams sold. As in Figure 16, there is a clear positive relationship, but this is caused by a third factor, which is the hot weather of the summer. Failure to recognize such a third factor can lead to a conclusion that a strong association exists between the two events that are independent.

Figure 16. Number of Drowning Deaths in the U.S. and the number of movies Cage appeared


2007
120
110
1999
2000
2009
100
2002
2005
2004
2008
90
2003
1
2
3
4
Cage

The above examples may represent obvious cases of correlation but not causation, but there are more subtle and controversial cases where causation is claimed based on correlation. In empirical finance, there are claims that investors' moods systematically affect the stock market. For example, there are articles that claim stock returns are negatively affected by seasonality (Bouman and Jacobsen, 2002), winter blues (Kamstra et al., 2003), weather (Hirshleifer and Shumway, 2003), and sports sentiment (Edmand et al., 2007), to name a few. Harvey et al. (2016) identified hundreds of similar factors published in the finance literature over the last decade or so that claimed explanatory power for stock market returns. One of the common features of these studies is the use of a large or massive sample size, accompanied by small effect sizes: see Kim (2019). What matters is whether lay or professional investors take such findings seriously in their investment decision-making. As Warren Buffet says in his 1984 interview, they are "not what businessman thinks about when buying businesses, as a stock is a piece of business."[xxiii] He further adds that "to a man with a hammer everything looks like a nail," referring to the researchers who mindlessly utilize statistical methods to produce meaningless studies measuring statistically significant correlations between the factors and stock returns.

Statistical insignificance.

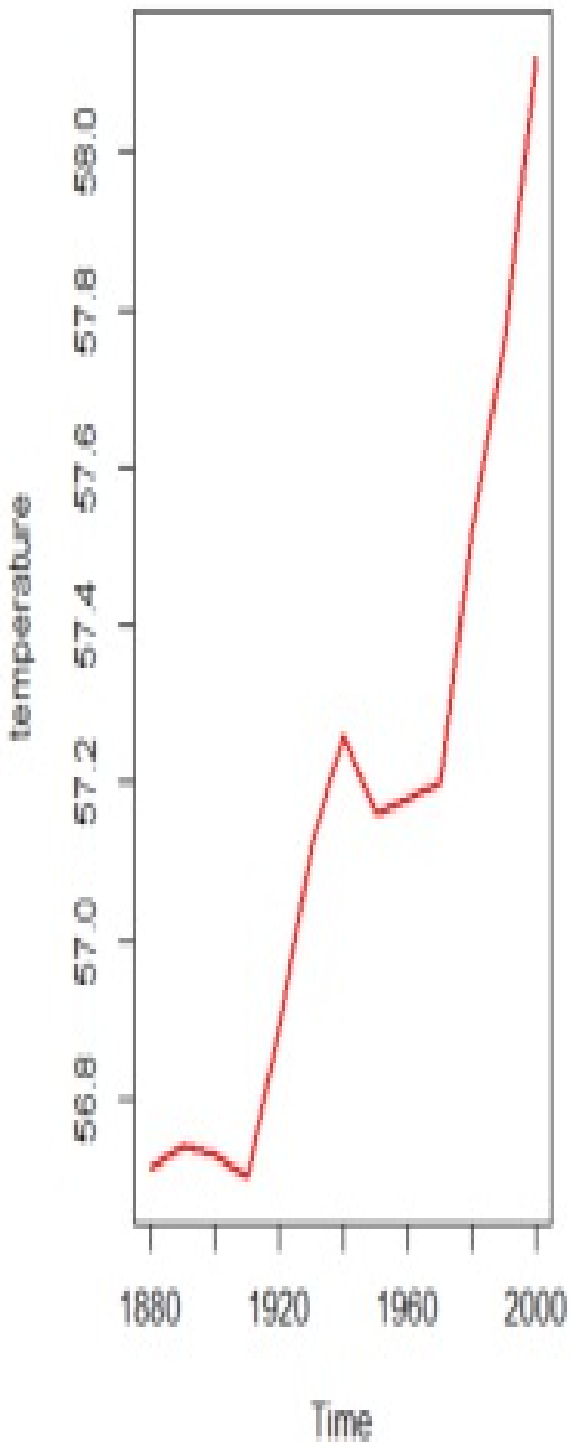
The statement from the American Statistical Association says, “The widespread use of ‘statistical significance’ (generally interpreted as ‘$p \leq 0.05$’ as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.” Without achieving statistical significance, researchers find it almost impossible to sell their results because, with statistical insignificance, it is difficult to convince journal editors and referees to accept their papers for publication. It is most unlikely that such findings (even if it is pathbreaking) see the light of day. This has created the serious problem of publication bias or the file drawer problem, where only the studies with statistically significant results are published in academic journals. Many meta-analytic studies report evidence of publication bias. A survey in empirical finance reports that 98% of the published papers come with statistical significance (Kim and Ji, 2014). More serious is the file drawer problem, where potentially important research findings are not published simply because they have not achieved statistical significance.

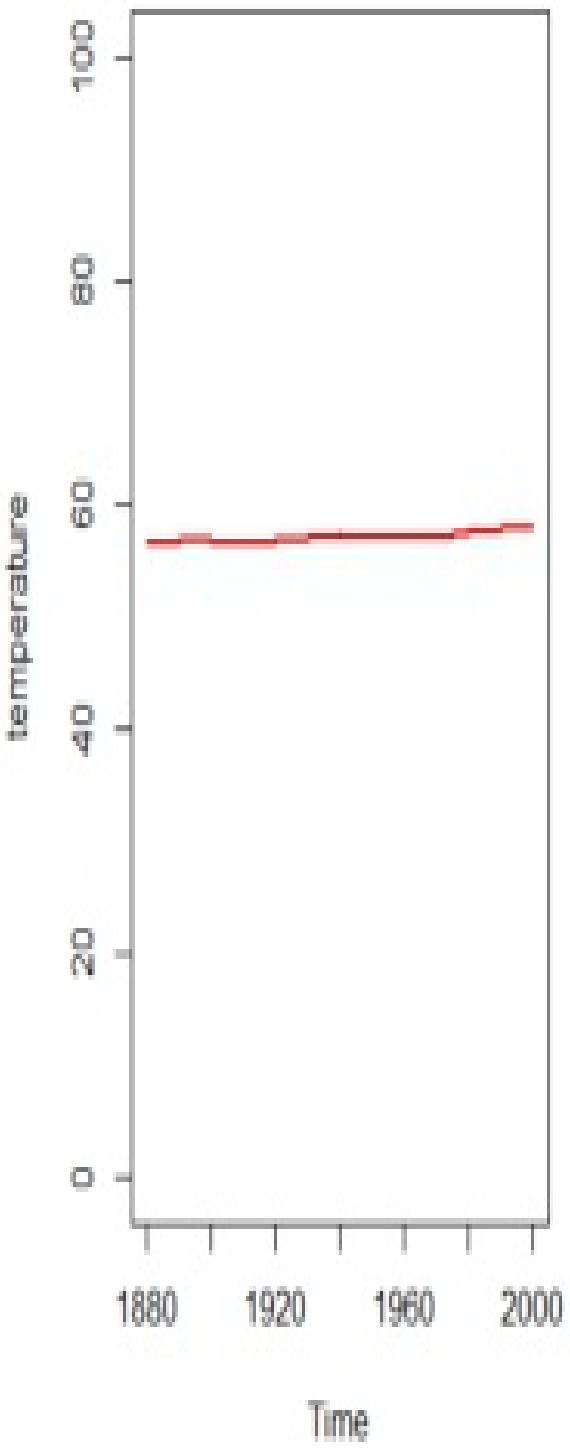
Statistical significance depends heavily on the sample size, as we have seen above. Other factors that affect statistical significance include the degree of variability associated, measurement errors, and effect size. Statistical insignificance, as well as statistical significance, should be evaluated by combining all available information, not solely on the magnitude of the Z-statistics or p-value. It should be considered in association with the plausibility of the hypothesis, the soundness of the research design, and effect size. Statistically insignificant results can be more informative than those statistically significant, as Abadie (2020) has shown from the Bayesian standpoint, who argues that statistically insignificant studies should also be reported and discussed in the same manner as those statistically significant.

Misleading visualization.

Data visualization is an important and powerful tool of descriptive statistics. A well-designed plot makes a strong visual impression and can tell a lot more about the data than a table full of numbers. However, this descriptive method can also be made deceptive, usually by manipulating the axis of the plot. The time plots in Figure 17 present the global mean temperature in Fahrenheit. The two are identical except that the Y-axis is scaled differently.[xxiv] The first one has the Y-axis closed with the minimum and maximum temperature values, while the second one has the Y-axis from 0 to 100. The effect of changing the scale of the Y-axis is dramatic, as in Figure 17. If you are a believer in climate change, the graph you want to see is on the left, but if you are a climate change antagonist, the one on the right will meet your eyes. Figure 18 presents a similar example where the difference of 4.6% looks deceptively massive. This plot is about what would happen to the top marginal tax rate in the U.S. if the tax cut implemented by President Bush expired in 2010. A properly scaled chart is also presented in Figure 18 for comparison.

Figure 17. Global Mean Temperature in Fahrenheit.


temperature
56.8
57.0
57.2
57.4
57.6
57.8
58.0
1880
1920
1960
2000
Time


temperature
0
20
40
60
80
100
1880
1920
1960
2000
Time

Data Source[xxv]

Figure 18. An Example of Misleading Visualization

Data Source[xxvi]

Concluding Remarks

This book has been a review of statistical methods and concepts and their applications to real-world problems, with a focus on statistical thinking. The paradigm of statistical decision-making, labelled the null ritual, does not fully promote sound statistical thinking. Rather, it promotes mindless and mechanical decision-making. Our critiques of this null ritual include:

not fully evaluating statistical uncertainty (Type I and II error probabilities),

not fully considering the consequence of Type I and II errors,

an arbitrary choice of the level of significance, and

making a decision using test statistics or p-value only, not fully assessing the effect size or signal from data

While statistics is widely taught and practised, sound statistical thinking is not fully understood. Rather, statistical decisions are being made using the null ritual as a single decision rule in a mechanical and mindless way. Many statistical findings are being misinterpreted, and these misinterpretations of statistical results may lead to incorrect decisions that could be highly costly. In many fields of science, they have accumulated false stylized facts and replication crises, which makes questionable the integrity and credibility of scientific research. Recent statements from the American Statistical Association called for urgent actions to fix this crisis. This book has been written with these calls in mind.

The root of the problems and crises is the null ritual, which has become a deep habit of the generations of statistical researchers. As discussed, this ritual is not what the pioneers of modern statistics have taught us, but the null ritual somehow has put itself in the mainstream method of statistical decisions in modern days. As long as this ritual is being taught to our future generation of statistical researchers and adopted by contemporary researchers, a drastic change or improvement towards sound statistical

thinking is unlikely. Now is the time that textbook writers, university professors, and journal editors should take a range of actions. They should educate future decision-makers in business and government and young academic researchers with a new paradigm of statistical research.

The new paradigm should negate the null ritual and re-embrace the teaching of the pioneers, especially a decision-theoretic approach proposed by Neyman and Pearson. The new paradigm should also invite a new statistical decision rule suitable for the big data era because the current paradigm of statistical research is based on small-sample methods conceived and developed nearly 100 years ago, an era when the scale of the big data of modern times was simply unimaginable. The new paradigm should also propose a range of credible alternatives, so statistical researchers promote sound statistical thinking without adopting a single-hammer approach. Establishing this new paradigm should be the priority of the leaders of statistical thinkers of the big data era.

A.R. and Jae H. Kim, PhD

Before You Go...

_____

I would be so very grateful if you would take a few seconds and rate or review this book! Reviews – testimonials of your experience - are critical to an author's livelihood. While reviews are surprisingly hard to come by, they provide the life blood for me being able to stay in business and dedicate myself to the thing I love the most, writing.

If this book helped, touched, or spoke to you in any way, please leave me a review and give me your honest feedback.

Thank you so much for reading this book!

About the Authors

Albert Rutherford

We often have blind spots for the reasons that cause problems in our lives. We try to fix our issues based on assumptions, false analysis, and mistaken deductions. These create misunderstanding, anxiety, and frustration in our personal and work relationships.

Resist jumping to conclusions prematurely. Evaluate information correctly and consistently to make better decisions. Systems and critical thinking skills help you become proficient in collecting and assessing data, as well as creating impactful solutions in any context.

Albert Rutherford dedicated his entire life to find the best, evidence-based practices for optimal decision-making. His personal mantra is, "ask better questions to find more accurate answers and draw more profound insights."

In his free time, Rutherford likes to keep himself busy with one of his long-cherished dreams - becoming an author. In his free time, he loves spending time with his family, reading the newest science reports, fishing, and pretending he knows a thing or two about wine. He firmly believes in Benjamin Franklin's words, “An investment in knowledge always pays the best interest.”

Read more books from Albert Rutherford:

Advanced Thinking Skills

The Systems Thinker Series

Game Theory Series

Critical Thinking Skills

Jae H. Kim, PhD

Jae H. Kim is a freelance writer in econometrics, statistics, and data analysis. Since obtaining his PhD in econometrics in 1997, he has been a professor in major Australian universities until 2022. He has published more than 70 academic articles and book chapters in econometrics, empirical finance, economics, and applied statistics, which have attracted nearly 5000 citations to date. His articles are listed in the Google Scholar page (https://scholar.google.com/citations?user=zEs_RAgAAAAJ&hl=en).

Kim's research has made major contributions to testing for financial market efficiency, time series forecasting, and statistical inference. He is also an expert R programmer, as an author of five R packages (vrtest, OptSig, BootPR, VAR.etp, and GRS.test).

References

Abadie, A. (2020). Statistical nonsignificance in empirical economics. American Economic Review: Insights, 2(2), 193-208.

Annual Average Temperature History for Earth - Current Results. (n.d.). Retrieved October 13, 2022, from https://www.currentresults.com/Environment-Facts/changes-in-earth-temperature.php

Biau, David Jean, Brigitte M. Jolles, and Raphaël Porcher. "P value and the theory of hypothesis testing: an explanation for new researchers." Clinical Orthopaedics and Related Research® 468.3 (2010): 885-892.

Bouman, S., & Jacobsen, B. (2002). The Halloween indicator," Sell in May and go away": Another puzzle. American Economic Review, 92(5), 1618-1635.

Calling Bullshit: Data Reasoning in a Digital World. (n.d.). Retrieved October 13, 2022, from https://www.callingbullshit.org/index.html

Edmans, A., Garcia, D., & Norli, Ø. (2007). Sports sentiment and stock returns. The Journal of finance, 62(4), 1967-1998.

Gigerenzer, G. (2004). Mindless statistics. The Journal of Socio-Economics, 33(5), 587-606.

Harford, T. (2014). Big data: A big mistake? Significance, 11(5), 14-19.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected

returns. The Review of Financial Studies, 29(1), 5-68.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. The Journal of Finance, 72(4), 1399-1440.

Hayward, S. (2015, October 21). The Only Global Warming Chart You Need from Now On. Power Line. Retrieved October 13, 2022, from https://www.powerlineblog.com/archives/2015/10/the-only-global-warming-chart-you-need-from-now-on.php

Hirshleifer, D., & Shumway, T. (2003). Good day sunshine: Stock returns and the weather. The journal of Finance, 58(3), 1009-1032.

Kamstra, M. J., Kramer, L. A., & Levi, M. D. (2003). Winter blues: A SAD stock market cycle. American Economic Review, 93(1), 324-343.

Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big data and large sample size: a cautionary note on the potential for bias. Clinical and translational science, 7(4), 342-346.

Kim, J. H., & Ji, P. I. (2015). Significance testing in empirical finance: A critical review and assessment. Journal of Empirical Finance, 34, 1-14.

Kim, J. H. (2019). Tackling false positives in business research: A statistical toolbox with applications. Journal of Economic Surveys, 33(3), 862-895.

Melvin, R. A. L. (2020, July 29). More p-values, more problems – Perioperative Data Science. Retrieved October 13, 2022, from https://sites.uab.edu/periop-datascience/2020/07/29/more-p-values-mode-problems/

Mendes, E. (2018, April 3). Coffee and Cancer: What the Research Really Shows. American Cancer Society. Retrieved October 13, 2022, from https://www.cancer.org/latest-news/coffee-and-cancer-what-the-research-really-shows.html

Notopoulos, K. (2014, October 3). 13 Graphs That Are Clearly Lying. BuzzFeed News. Retrieved October 13, 2022,

from https://www.buzzfeednews.com/article/katienotopoulos/graphs-that-lied-to-us

statistical mean, median, mode and range. (2020, December 22). SearchDataCenter. Retrieved October 13, 2022, from https://www.techtarget.com/searchdatacenter/definition/statistical-mean-median-mode-and-range

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. Significance, 12(3), 30-32.

Selvanathan, E. A., Selvanathan, S., and Keller, G. (2017), Business Statistics: Australia/New Zealand (7th ed.), South Melbourne, Victoria: Cengage Learning

Soyer, E., & Hogarth, R. M. (2012). The illusion of predictability: How regression statistics mislead experts. International Journal of Forecasting, 28(3), 695-711.

Study.com | Take Online Courses. Earn College Credit. Research Schools, Degrees & Careers. (n.d.). Retrieved October 13, 2022, from https://study.com/learn/lesson/mean-median-mode-range-measures-central-tendency.html

Squire, P. (1988). Why the 1936 Literary Digest poll failed. Public Opinion Quarterly, 52(1), 125-133.

Thayqua. (2018, January 16). Warren Buffett: How to Pick Stocks & Get Rich (1985). YouTube. Retrieved October 13, 2022, from https://www.youtube.com/watch?v=PEs5caq8QNs

The Basel Committee - overview. (2011, June 28). Retrieved October 13, 2022, from https://www.bis.org/bcbs/

Best, R. A. B. (2022, October 13). President: general election Polls.

FiveThirtyEight. Retrieved October 13, 2022, from https://projects.fivethirtyeight.com/polls/president-general/

What Are Clinical Trials and Studies? (n.d.). National Institute on Aging. Retrieved October 13, 2022, from https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies

What Is the GARCH Process? How It's Used in Different Forms. (2020, October 25). Investopedia. Retrieved October 13, 2022, from https://www.investopedia.com/terms/g/generalalizedautogregressivecon

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. The American Statistician, 70(2), 129-133.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". The American Statistician, 73(sup1), 1-19.

Wikipedia contributors. (2022a, October 3). Opinion poll. Wikipedia. Retrieved October 13, 2022, from https://en.wikipedia.org/wiki/Opinion_poll

Wikipedia contributors. (2022b, October 13). 1936 United States presidential election. Wikipedia. Retrieved October 13, 2022, from https://en.wikipedia.org/wiki/1936_United_States_presidential_election

Ziliak, S., & McCloskey, D. N. (2008). The cult of statistical significance: How the standard error costs us jobs, justice, and lives. University of Michigan Press.

Endnotes

---

[1] See, for an example, https://propertyupdate.com.au/whats-difference-average-median-house-prices/.

[2] https://onlinestatbook.com/2/calculators/normal_dist.html

[3] For more information, check: https://www.pnas.org/doi/10.1073/pnas.1313476110

[4] See Bias et al. (2010)

[5] Gigerenzer (2004)

[6] This example is taken from Selvanathan, E. A., Selvanathan, S., & Keller, G. (2016). Business Statistics: Australia New Zealand with Online Study Tools 12 Mo Nths. Cengage AU.

[7] See also Harford (2014) and Squire (1988) providing a detailed analysis.

---

[i] statistical mean, median, mode and range. (2020, December 22). SearchDataCenter. Retrieved October 13, 2022, from https://www.techtarget.com/searchdatacenter/definition/statistical-mean-median-mode-and-range

[ii] Study.com | Take Online Courses. Earn College Credit. Research Schools, Degrees & Careers. (n.d.). Retrieved October 13, 2022, from https://study.com/learn/lesson/mean-median-mode-range-measures-central-tendency.html

[iii] The P-value. N.A.

https://www.cisu.edu/faculty/gerstman/EpiInfo/pvalue.htm

https://www.sjsu.edu/faculty/gerstman/EpiInfo/pvalue.htm

[iv] Biau, David Jean, Brigitte M. Jolles, and Raphaël Porcher. "P value and the theory of hypothesis testing: an explanation for new researchers." Clinical Orthopaedics and Related Research® 468.3 (2010): 885-892.

[v] Gigerenzer, G. (2004). Mindless statistics. The Journal of Socio-Economics, 33(5), 587-606.

[vi] Ziliak, S., & McCloskey, D. N. (2008). The cult of statistical significance: How the standard error costs us jobs, justice, and lives. University of Michigan Press.

[vii] Richard W. Brislin (1980). "Cross-Cultural Research Methods: Strategies, Problems, Applications". In Irwin Altman; Amos Rapoport; Joachim F. Wohlwill (eds.). Environment and Culture. Springer. p. 73. ISBN 978-0-306-40367-5.

[viii] Calling Bullshit: Data Reasoning in a Digital World. (n.d.). Retrieved October 13, 2022, from https://www.callingbullshit.org/index.html

[ix] Leamer, Edward E. 1978. https://www.anderson.ucla.edu/faculty/edward.leamer/books/specification_sea

[x] Jae H. Kim (2020) Decision-Theoretic Hypothesis Testing: A Primer With R Package OptSig, The American Statistician, 74:4, 370-379, DOI: 10.1080/00031305.2020.1750484

[xi] Wikipedia contributors. (2022, October 3). Opinion poll. Wikipedia. Retrieved October 13, 2022, from https://en.wikipedia.org/wiki/Opinion_poll

[xii] Best, R. A. B. (2022, October 13). President: general election Polls. FiveThirtyEight. Retrieved October 13, 2022, from https://projects.fivethirtyeight.com/polls/president-general/

[xiii] What Are Clinical Trials and Studies? (n.d.). National Institute on Aging. Retrieved October 13, 2022, from https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies

https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies

[xiv] Lo, Andrew (2004). "The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective" (PDF). Journal of Portfolio Management. 5. 30: 15–29. doi:10.3905/jpm.2004.442611.

[xv] Harvey, Campbell R. and Liu, Yan, A Census of the Factor Zoo (February 25, 2019). Available at SSRN: https://ssrn.com/abstract=3341728 or http://dx.doi.org/10.2139/ssrn.3341728

[xvi] The Basel Committee - overview. (2011, June 28). Retrieved October 13, 2022, from https://www.bis.org/bcbs/

[xvii] What Is the GARCH Process? How It's Used in Different Forms. (2020, October 25). Investopedia. Retrieved October 13, 2022, from https://www.investopedia.com/terms/g/generalalizedautogregressivecondition

[xviii] Harford, T. (2014). Big data: A big mistake? Significance, 11(5), 14-19.

[xix] Wikipedia contributors. (2022b, October 13). 1936 United States presidential election. Wikipedia. Retrieved October 13, 2022, from https://en.wikipedia.org/wiki/1936_United_States_presidential_election

[xx] Squire, P. (1988). Why the 1936 Literary Digest poll failed. Public Opinion Quarterly, 52(1), 125-133.

[xxi] Mendes, E. (2018, April 3). Coffee and Cancer: What the Research Really Shows. American Cancer Society. Retrieved October 13, 2022, from https://www.cancer.org/latest-news/coffee-and-cancer-what-the-research-really-shows.html

[xxii]Melvin, R. A. L. (2020, July 29). More p-values, more problems – Perioperative Data Science. Retrieved October 13, 2022, from https://sites.uab.edu/periop-datascience/2020/07/29/more-p-values-mode-problems/

[xxiii] Thayqua. (2018, January 16). Warren Buffett: How to Pick Stocks &

Get Rich (1985). YouTube. Retrieved October 13, 2022, from

Get Rich (1985). YouTube. Retrieved October 13, 2022, from https://www.youtube.com/watch?v=PEs5caq8QNs

[xxiv] Hayward, S. (2015, October 21). The Only Global Warming Chart You Need from Now On. Power Line. Retrieved October 13, 2022, from https://www.powerlineblog.com/archives/2015/10/the-only-global-warming-chart-you-need-from-now-on.php

[xxv] Annual Average Temperature History for Earth - Current Results. (n.d.). Retrieved October 13, 2022, from https://www.currentresults.com/Environment-Facts/changes-in-earth-temperature.php

[xxvi] Notopoulos, K. (2014, October 3). 13 Graphs That Are Clearly Lying. BuzzFeed News. Retrieved October 13, 2022, from https://www.buzzfeednews.com/article/katienotopoulos/graphs-that-lied-to-us