STSM2616

Portfolio

Sample Distribution Theory and Inference

Franco Deacon
2025813755

Table of Contents

AI Declaration

2025 813 755

Franco Deacon,

I declare that each and every use of AI in this portfolio has been referenced. I understand that unreferenced use of AI counts as plagerism, and there are very strict UFS regulations against plagiarism.

Moment Generating Functions

Research Process

- I consulted my Rice textbook for the definition of an MGF and its related properties.
- I asked ChatGPT what the use cases and applications of MGF are.
- I read the Wikipedia on moment generating functions, where the series expansion of MGF's is explained.
- Since MGF was covered in STSM1624, I consulted my old class notes on MGF.
- I tried some exercises I did in my STSM1624 Workbook.
- I did some exercises 81 and 82 from (Rice, 2007).
- I tried some new exercises from source .

Definition

The moment-generating function (mgf) of a random variable X is $M(t) = E(e^{tX})$ if the expectation is defined. In the discrete case,

$$M(t) = \sum_x e^{tx} p(x)$$

And in the continuous case,

$$M(t) = \int_{-\infty}^{\infty} e^{tx} p(x)$$

Moment Generating functions are used to generate Raw Moments $E[X^n]$. We can use these Raw Moments to find the Central Moments $E[(X-\mu)^n]$, which gives is key insights about the distribution of the data.

The *n*th moment of *X*, $E[X^n]$ can be obtained by differentiating $M_X(t),\ n$ times with respect to *t* and evaluating at *t*=0:

$$E[X^n] = \frac{d^n}{dt^n} M_{x(t)|_{t=0}}$$

My Understanding of the MGF

The moment generating function $M(t) = E(e^{tX})$, generates moments by successive differentiation and evaluating at $t = 0$. The first derivative $M'(0)$ gives the first raw moment, which is the mean. The second central moment, the variance, can by calculated from the raw moments $M''(0) - [M'(0)]^2$. If we standardize the third central moment by σ^3, we find the skewness. Skewness is a measurement of asymmetry around the mean. If we standardize the fourth moment by σ^4, we find the kurtosis. Kurtosis measures how peaked the distribution is. I can more specifically say that kurtosis measures the heaviness of the distribution's tails, relative to the tails of a normal distribution.

The MGF can greatly simplify calculations with sums of independent random variables. The MGF can proof two distributions as equal through the uniqueness property, if they have the same MGF at t = 0. The MGF is often useful or required for the proof of statistical theorems, such as the continuity theorem, the central limit theorem and the proof that $\bar{X}$ and S^2 are identically distributed.

Properties of the Moment Generating Function

PROPERTY **A**

If the moment-generating function exists for t in an open interval containing zero, it uniquely determines the probability distribution. ■

FIGURE 1 (RICE, 2007, P. 155)

This Property is also known as the Uniqueness property. It states that if two random variables X and Y have the same MGF values of t in an open interval containing zero, X and Y follows the same distribution function. This property is an essential part of the continuity theorem.

PROPERTY **B**

If the moment-generating function exists in an open interval containing zero, then $M^{(r)}(0) = E(X^r)$. ■

FIGURE 2 (RICE, 2007, P. 155)

Property B shows how raw moments are derived from the MGF, by evaluating at $t = 0$. The first moment $r = 1$ is the mean $E(X)$.

PROPERTY **C**

If X has the mgf $M_X(t)$ and $Y = a + bX$, then Y has the mgf $M_Y(t) = e^{at} M_X(bt)$.

Proof

$$\begin{aligned} M_Y(t) &= E(e^{tY}) \\ &= E(e^{at+btX}) \\ &= E(e^{at} e^{btX}) \\ &= e^{at} E(e^{btX}) \\ &= e^{at} M_X(bt) \end{aligned}$$ ■

FIGURE 3 (RICE, 2007, P. 158)

PROPERTY **D**

If X and Y are independent random variables with mgf's M_X and M_Y and $Z = X + Y$, then $M_Z(t) = M_X(t)M_Y(t)$ on the common interval where both mgf's exist.

Proof

$$\begin{aligned} M_Z(t) &= E(e^{tZ}) \\ &= E(e^{tX+tY}) \\ &= E(e^{tX}e^{tY}) \end{aligned}$$

From the assumption of independence,

$$\begin{aligned} M_Z(t) &= E(e^{tX})E(e^{tY}) \\ &= M_X(t)M_Y(t) \end{aligned}$$

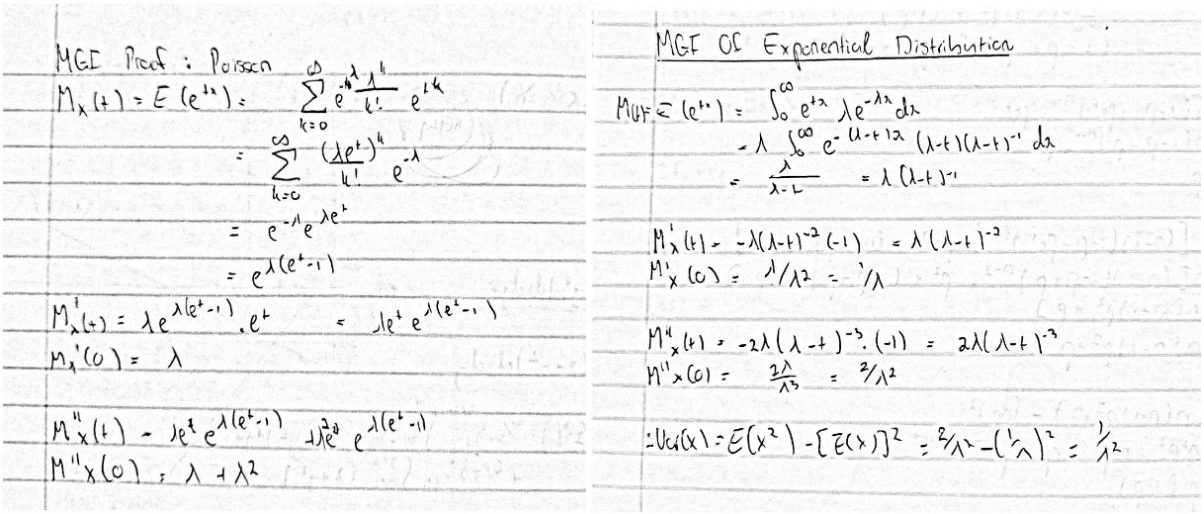
■

FIGURE 4 (RICE, 2007, P. 159)

Property D is particularly useful since it greatly simplifies the analysis of convolutions of random variables. It allows us to calculate the MGF of independent random variables, without the need of compute the convolution or transformation of the variables.

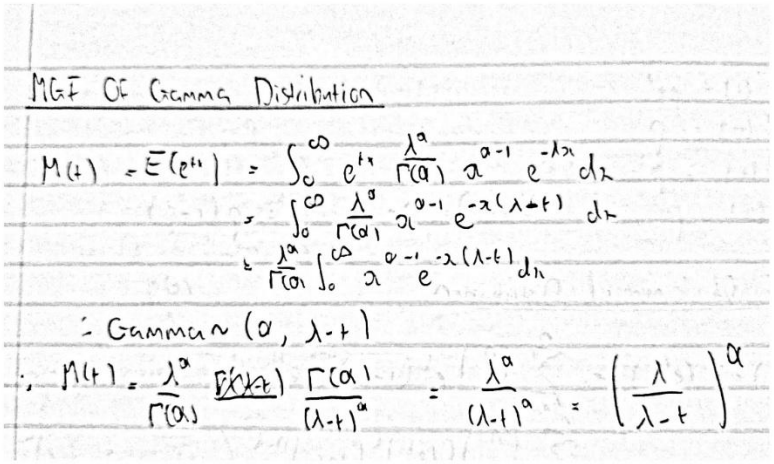
Proofs of common Moment Generating Functions

Poison and Exponential Distribution



Here are some examples of MGF proofs. On the left: The poison distribution, as shown in example A (Rice, 2007, p. 156). On the left is the exponential distribution, this proof uses some neat manipulation to simplify the integration, it was show to us by Prof. Adrehette Verster, from the University of the Free State.

Gamma Distribution

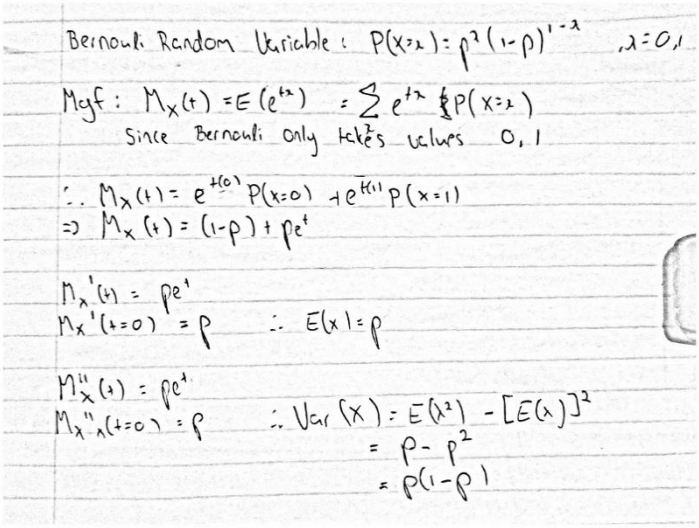


The Gamma distribution is my favourite and so is it's MGF proof. When substituting in e^{tx} and using properties of exponents, one can notice that the integral resembles that of a Gamma function, and with some manipulation you will get a Gamma function with parameters α and λ -t. The moment generated function can the easily be solved by using properties of the Gamma Function rather than struggling with a difficult integral.

Exercises

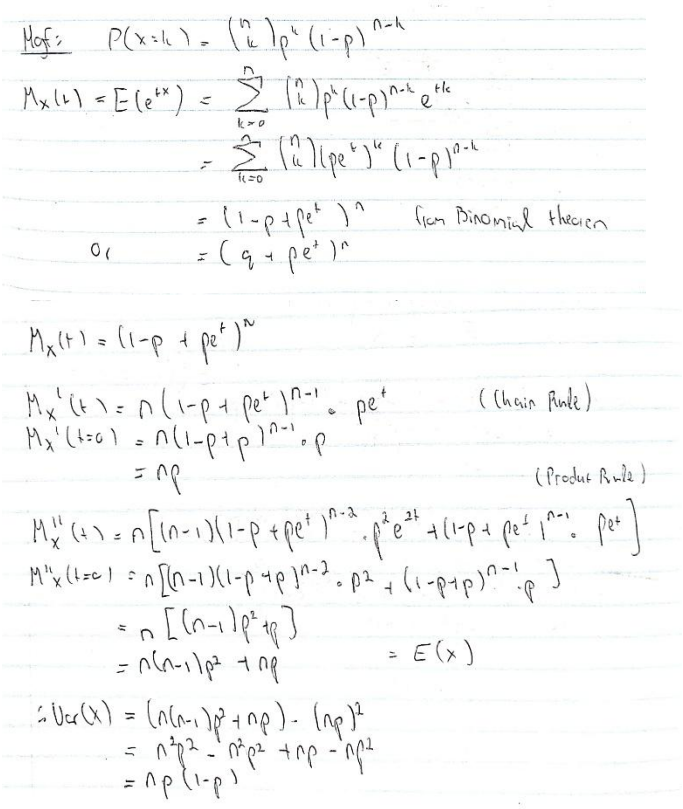
81. Find the moment-generating function of a Bernoulli random variable, and use it to find the mean, variance, and third moment.

FIGURE 5 (RICE, 2007, P. 173)



82. Use the result of Problem 81 to find the mgf of a binomial random variable and its mean and variance.

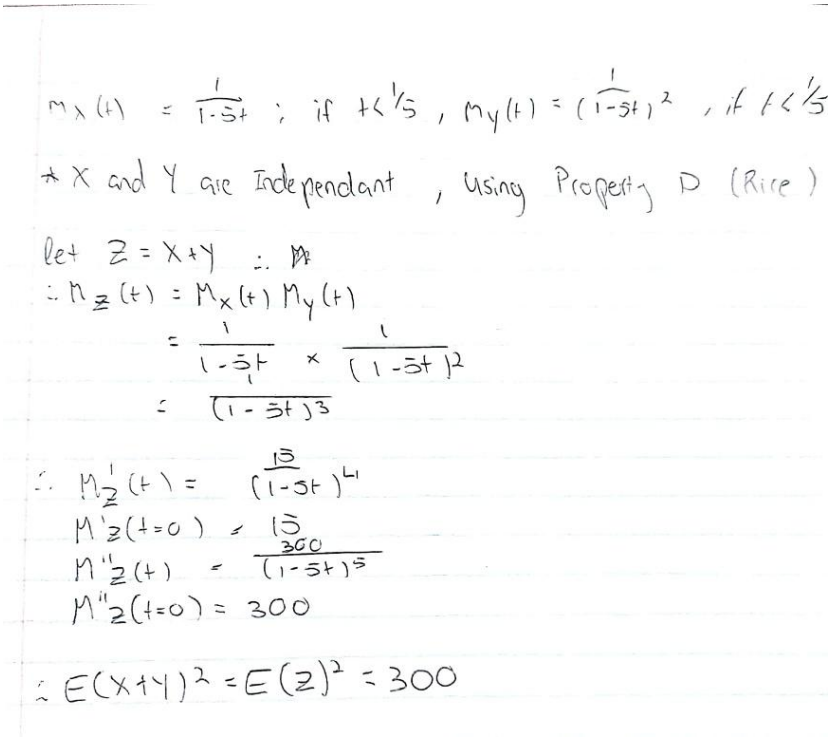
FIGURE 6 (RICE, 2007, P. 173)



Even though the chain rule gets a bit long, I find using MGF as a much easier way to prove the expected value and variance of the binomial distribution.

Exercise 13.4. Let X and Y be two independent random variables with respective moment generating functions $m_x(t) = \frac{1}{1-5t},\ if\ t < \frac{1}{5}, m_{y(t)} = \frac{1}{(1-5t)^2}, if\ t < \frac{1}{5}$

Find $E(X+Y)^2$



The use of property D of independent random variables made solving this problem very easy. It would have taken much longer and much harder if convolution of the variables were used.

Reflection

I now understand moment generating functions much better than I did when I first encountered it in STSM1626. I since learned MGF's are much more than an alternative method to find a mean, it is an important tool or necessity for many of the important statistical proofs, like the central limit theorem.

Moment generating functions is a bit difficult to understand, since it does not generate moments in time like I once thought, contrary to its name. Rather it is used to derive the raw moments at $t = 0$. The raw moments can be used to derive the central moments. I do believe I have a much better understanding now and is able to explain that understanding.

The proofs and properties from this section are easy to prove, even though I found it difficult in STSM1626.

I misunderstood the raw moments and thought that the third and fourth moments are the measurements of skewness and kurtosis. However, these raw moments must first be used in some calculations and formulae before it will give the measurements of skewness and kurtosis.

Limit Theorems

Index of exercises and examples

Markov's Inequality

i. Example from (Tsun, 2020)
ii. Chapter 4, Question 41 (Rice, 2007, p. 169)

Chebyshev's Inequality

i. Chapter 4, Question 33 (Rice, 2007, p. 169)
ii. Chapter 4, Question 42 (Rice, 2007, p. 169)
iii. Question 5.31 (George Casella, 2001, p. 260) (Found in CLT section)

Law of Large Numbers

i. Example, python simulation of dice rolls
ii. Exercise from the University of Notre Dame lecture notes
iii. Question 5.32(a) (George Casella, 2001, p. 261)
iv. Chapter 5, Question 7 (Rice, 2007, p. 188)

Monte Carlo Integration

i. Chapter 5, Question 19 (Rice, 2007, p. 190)
ii. Assignment 1

Convergence in Distribution

i. Example 7.7, Probability Course

Continuity Theorem

i. Chapter 5, Question 5 (Rice, 2007, p. 188)

Central Limit Theorem

i. ChatGPT combined CLT and Monte Carlo Integration problem
ii. Example from (Tsun, 2020)
iii. Chapter 5, Question 25 (Rice, 2007, p. 191)
iv. Question 5.31 (George Casella, 2001, p. 200)
v. Chapter 5, Question 17 (Rice, 2007, p. 189)

Markov's Inequality

Research Process

- I asked both ChatGPT4o and DeepSeek to define and explain Markov's Inequality at various levels of complexity.
- I consulted my (Rice, 2007) and (Tsun, 2020) Textbooks for the formal definition of Markov's Inequality.
- I read the Wikipedia article on Markov's inequality, where I learned that Markov's inequality can also be used to upper bound the expectation of a non-negative random variable in terms of its distribution function.
- I researched and attempted the proof of the inequality.
- I completed an exercise from my (Rice, 2007) textbook.

Definition

THEOREM A *Markov's Inequality*

If X is a random variable with $P(X \geq 0) = 1$ and for which $E(X)$ exists, then $P(X \geq t) \leq E(X)/t$.

Proof

We will prove this for the discrete case; the continuous case is entirely analogous.

$$E(X) = \sum_x xp(x)$$

$$= \sum_{x<t} xp(x) + \sum_{x \geq t} xp(x)$$

All the terms in the sums are nonnegative because X takes on only nonnegative values. Thus

$$E(X) \geq \sum_{x \geq t} xp(x)$$

$$\geq \sum_{x \geq t} tp(x) = tP(X \geq t)$$ ■

FIGURE 7 (RICE, 2007, P. 121)

My understanding of Markov's Inequality

Markov's inequality gives us an upper bound of probability. It tells us how likely it is for a random variable to take on a value much larger than its mean.

It is useful for when we don't have the distribution of a variable, and only it's expected value/ mean is known.

It describes the ratio between the mean/ expected value of the non-negative value X, and the actual size of X.

Proof:

Proof for Markov's Inequality (Continuous Case):

$$E(X) = \int_0^\infty x f_X(x)\,dx \qquad \text{(Because } X \geq 0\text{)}$$
$$= \int_0^t x f_X(x)\,dx + \int_t^\infty x f_X(x)\,dx \qquad \text{(Split at some } 0 \leq k \leq \infty\text{)}$$
$$\geq \int_t^\infty x f_X(x)\,dx \qquad \left(\int_0^t x f_X(x)\,dx \geq 0, \text{ because } t \geq 0, x \geq 0 \text{ and } f(x) \geq 0\right)$$
$$\geq \int_t^\infty t f_X(x)\,dx \qquad \text{(Because } x \geq t \text{ in integral)}$$
$$= t \int_t^\infty f_X(x)\,dx$$
$$= t\,\mathbb{P}(X \geq t)$$

Exercises and Examples:

i. Simple example I found from (Tsun, 2020)

The following example demonstrates how to use Markov's inequality, and how loose it can be in some cases.

Example(s)

A coin is weighted so that its probability of landing on heads is 20%, independently of other flips. Suppose the coin is flipped 20 times. Use Markov's inequality to bound the probability it lands on heads at least 16 times.

Solution We actually do know this distribution; the number of heads is $X \sim \text{Bin}(n = 20, p = 0.2)$. Thus, $\mathbb{E}[X] = np = 20 \cdot 0.2 = 4$. By Markov's inequality:

$$\mathbb{P}(X \geq 16) \leq \frac{\mathbb{E}[X]}{16} = \frac{4}{16} = \frac{1}{4}$$

Let's compare this to the actual probability that this happens:

$$\mathbb{P}(X \geq 16) = \sum_{k=16}^{20} \binom{20}{k} 0.2^k \cdot 0.8^{20-k} \approx 1.38 \cdot 10^{-8}$$

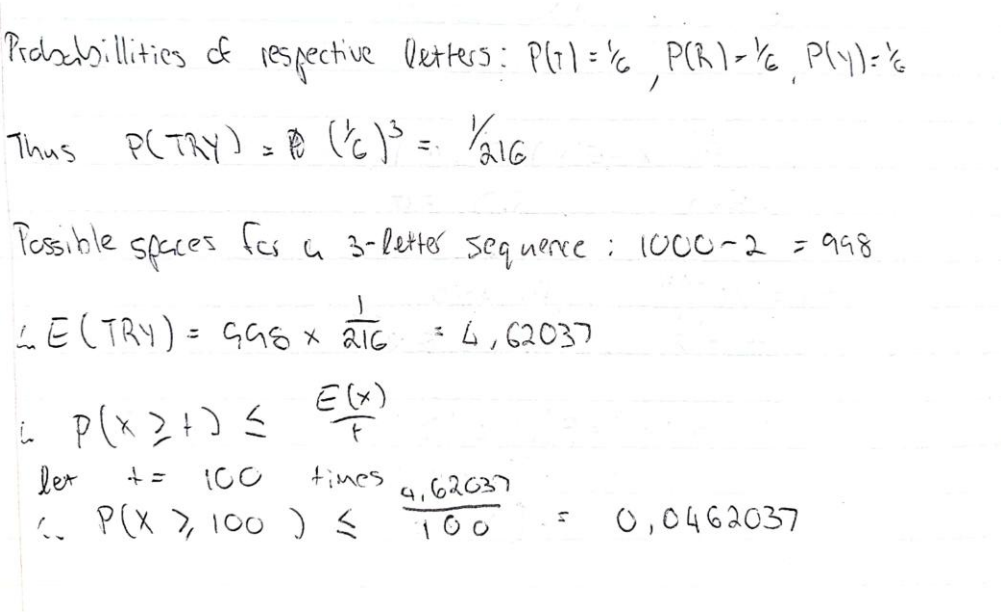
This is not a good bound, since we only assume to know the expected value. Again, we knew the exact distribution, but chose not to use any of that information (the variance, the PMF, etc.). □

This example shows that Markov's inequality can be inaccurate compared to the actual probability and that Markov's inequality is not that useful when the exact distribution of the variable is known.

40. A child types the letters Q, W, E, R, T, Y, randomly producing 1000 letters in all. What is the expected number of times that the sequence QQQQ appears, counting overlaps?

41. Continuing with the previous problem, how many times would we expect the word "TRY" to appear? Would we be surprised if it occurred 100 times? (*Hint:* Consider Markov's inequality.)

FIGURE 8 (RICE, 2007, P. 169)



Reflection

I found Markov's inequality easy to understand intuitively, as explained under my understanding. I am confident that I can explain it well.

Markov's inequality is the simplest at probably least used of the limit theorems, but I learned that it does not make it insignificant, since the inequality can be used when only the mean is known. Markov's inequality requires no further parameters or conditions, like the other limit theorems. Additionally, Markov's inequality is used to prove Chebyshev's inequality.

This section challenged me to sharpen my mathematical skills on working with inequalities. I have struggled with proofs involving inequalities in previous mathematics modules. Because Markov's inequality is relatively simple, it is a good starting point before working with more challenging inequalities and proofs.

Chebyshev's inequality

Research Process

- I asked both ChatGPt and DeepSeek to explain Chebyshev's inequality to me at various levels of complexity.
- I watched the following YouTube video by qncubed3 to familiarize myself with the inequality, he provided an explanation for the proof and a simple example.
- I read the Wikipedia article on Chebyshev's inequality, where I learned that Markov's inequality is sometimes referred as Chebyshev's first inequality.
- I looked up the formal definition of the inequality from my textbooks (Rice, 2007), (Tsun, 2020)
- I familiarised myself with the proof from (Tsun, 2020)
- I consulted by (George Casella, 2001) textbook, where I found some more exercises.
- I completed some exercises from my (Rice, 2007) textbook.

Definition

THEOREM C *Chebyshev's Inequality*

Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$,

$$P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}$$

Proof

Let $Y = (X - \mu)^2$. Then $E(Y) = \sigma^2$, and the result follows by applying Markov's inequality to Y. ■

FIGURE 9 (RICE, 2007, P. 133)

Chebyshev's Inequality can also be rewritten in the form:

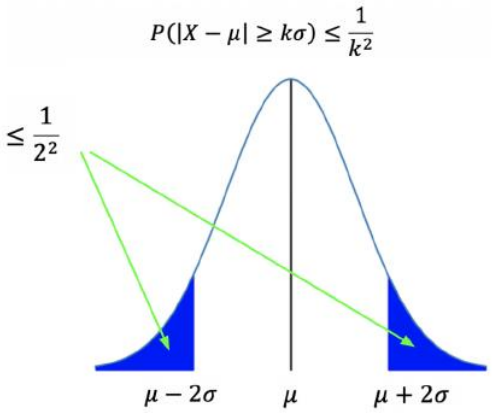


FIGURE 10 (TSUN, 2020) ILLUSTRATION OF THE PROBABILITY BOUNDS

My Understanding of Chebyshev's Inequality

Chebyshev's inequality gives us an upper bound on extreme deviation from the mean. It includes the parameter of the variance and is thus more accurate than Markov's inequality.

It tells us that the probability mass is concentrated around the mean, for example, if $k = 2$, then at least 75% of the data lies within 2 standard deviations from the mean

It tells us that the probability of X being more than k standard deviations from the mean is at most the upper bound of $\frac{1}{k^2}$

Proof

Proof of Chebyshev's Inequality. X is a random variable, so $(X - \mathbb{E}[X])^2$ is a **non-negative** random variable. Hence, we can apply Markov's inequality.

$$\begin{aligned}
\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \alpha\right) &= \mathbb{P}\left((X - \mathbb{E}[X])^2 \geq \alpha^2\right) && \text{[square both sides]} \\
&\leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{\alpha^2} && \text{[Markov's inequality]} \\
&= \frac{\mathsf{Var}(X)}{\alpha^2} && \text{[def of variance]}
\end{aligned}$$

□

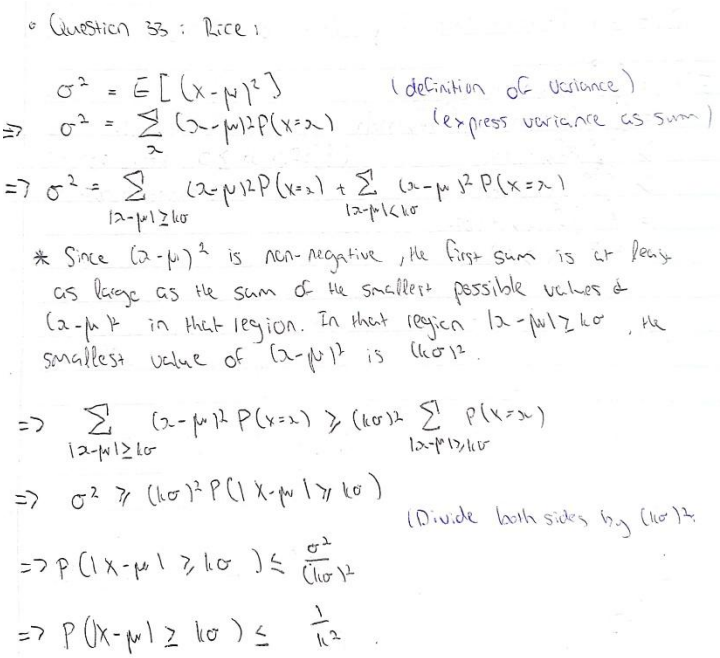
FIGURE 11 (TSUN, 2020)

The proof for Chebyshev's inequality uses Markov's inequality, yet Chebyshev's inequality was published first in 1867. Markov's inequality is more general than Chebyshev's inequality and was developed by Andrei Markov in the 1880s or 1890s.

Exercises

33. Prove Chebyshev's inequality in the discrete case.

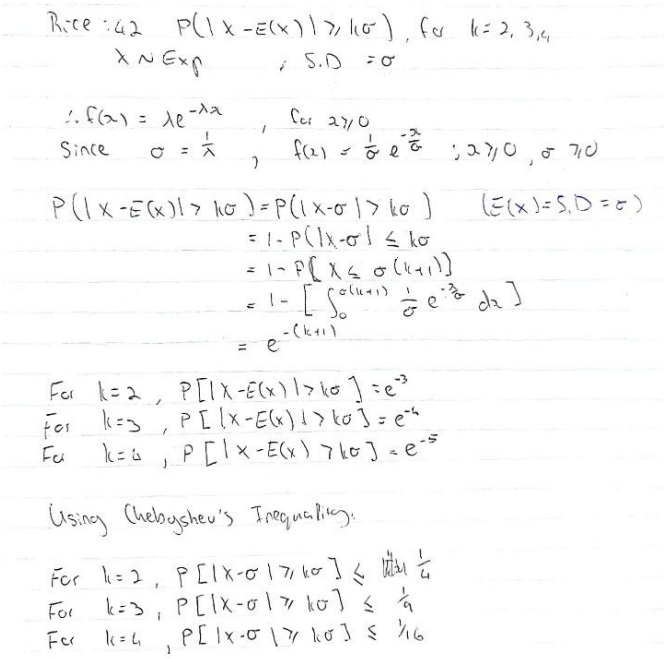
FIGURE 12 (RICE, 2007, P. 169)



I found the proof using the discrete case n bit challenging. I consulted DeepSeek, which helped me to set up the inequality between the sums and provided good reasoning, which I included.

42. Let X be an exponential random variable with standard deviation σ. Find $P(|X - E(X)| > k\sigma)$ for $k = 2, 3, 4$, and compare the results to the bounds from Chebyshev's inequality.

FIGURE 13 (RICE, 2007, P. 169)



Another great exercise, comparing the accuracy of Chebyshev's inequality to normal approximation using the Central Limit Theorem, can be found under the exercise heading in the Central Limit Theorem section.

Reflection

Because of my understanding of Markov's inequality, I was fast to understand Chebyshev's inequality. I found the discrete proof challenging, but I was much more confident when attempting question 42 from Rice and Question 5.31 from (George Casella, 2001) which I attempted much later than the first questions.

I am also confident that I am able to explain the inequality and its intuition.

I was surprised to learn that Chebyshev's inequality was published before Markov's inequality. I read it somewhere in a textbook and it sparked my interest whereafter I did some research on the history of Chebyshev and Markov.

Law of Large numbers

Research Process

- I watched a YouTube video by NStatum, as shown below
- I read the Wikipedia article on the Law of large numbers, where I learned that the Cauchy distribution will not converge in probability.
- I consulted (Tsun, 2020) for the definition of the Weak Law of Large Numbers
- I learned the proof from davis maths, as shown below
- I consulted my (Rice, 2007) textbook on the Law of Large Numbers
- I went through the lecture notes of the University of Notre Dame and attempted one of the exercises.
- I consulted my (George Casella, 2001) textbook, where I found an exercise on convergence in probability.
- I wrote a script in Python to visually illustrate the convergence effect of the Law of Large numbers.
- I asked Grok 3 to help me graph the results from my Python simulation.

Definition

THEOREM A *Law of Large Numbers*

Let $X_1, X_2, \ldots, X_i \ldots$ be a sequence of independent random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then, for any $\varepsilon > 0$,

$$P(|\overline{X}_n - \mu| > \varepsilon) \to 0 \qquad \text{as } n \to \infty$$

Proof

We first find $E(\overline{X}_n)$ and $\text{Var}(\overline{X}_n)$:

$$E(\overline{X}_n) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu$$

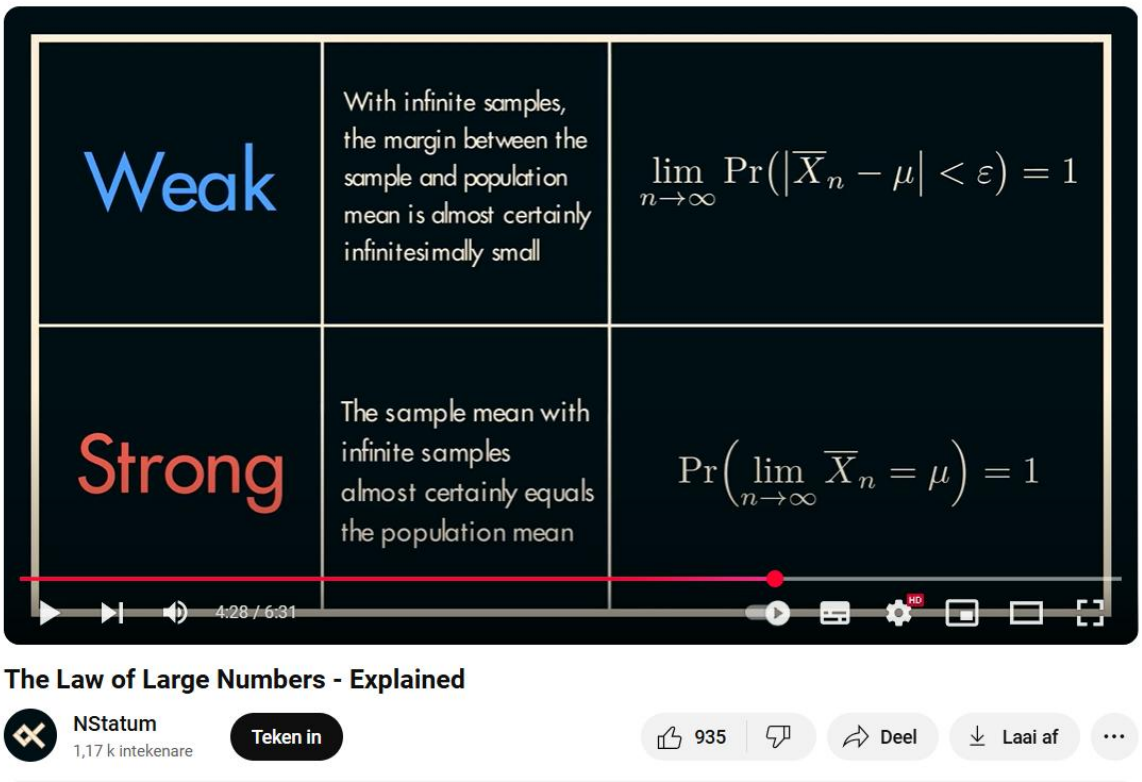
Since the X_i are independent,

$$\text{Var}(\overline{X}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{\sigma^2}{n}$$

The desired result now follows immediately from Chebyshev's inequality, which states that

$$P(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\text{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0, \qquad \text{as } n \to \infty \qquad \blacksquare$$

FIGURE 14 (RICE, 2007, P. 178)



This video from NStatum explains both the weak and strong law of large numbers. It also explained the meaning of convergence in probability. I enjoyed the visual illustrations, as the video made use of 3Blue1Brown's Manim Python package. I loved the simple explanation of the weak law: "With infinite samples, the margin between the sample and population mean is almost certainly infinitesimally small.

My understanding of the law of large numbers

The Weak Law of Large Numbers states that if the sample size n increases, the sample average of independent and identically distributed random variables, converges in probability to the expected value/mean of the distribution, if it exists. Or more simply, as n becomes large, the sample mean converges to the true mean.

In practise, it does not guarantee that the sample mean is equal to the true mean, like in the case of the gambler's fallacy, but that large deviations from the true mean become increasingly unlikely as n grows large.

Proof

Weak Law of Large Numbers: Let X_1, X_2,X_3, ... be a sequence of independent random variables with common distribution function. Set $\mu = E(X_j)$ and $\sigma^2 = \mathrm{Var}(X_j)$. As usual we define

$$S_n = X_1 + X_2 + \cdots + X_n$$

and let

$$S_n^* = \frac{S_n}{n} - \mu.$$

We apply Chebyshev's inequality to the random variable S_n^*. A by now routine calculation gives

$$E(S_n^*) = 0 \text{ and } \mathrm{Var}(S_n^*) = \frac{\sigma^2}{n}.$$

Then Chebyshev (1) says that for every $\varepsilon > 0$

$$\mathrm{P}\left(|S_n^*| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathrm{Var}(S_n^*).$$

Writing this out explicitly:

$$\mathrm{P}\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\frac{\sigma^2}{n}.$$

Thus for every $\varepsilon > 0$, as $n \to \infty$

$$\mathrm{P}\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \to 0.$$

FIGURE 15 PROOF FROM UC DAVIS MATHS

The law of large number's proof uses Chebyshev's inequality, which itself uses Markov's inequality. The inequalities are interrelated to each other, with an increasing degree of confidence/accuracy.

Law of Large Numbers Simulation

Law of Large Numbers.py × Manim.py

```python
import random
import matplotlib.pyplot as plt

# Function to simulate dice rolls and calculate running averages
def simulate_dice_rolls(num_rolls):  1 usage
    rolls = []
    averages = []

    # Simulate dice rolls and calculate cumulative average
    for i in range(1, num_rolls + 1):
        roll = random.randint( a: 1,  b: 6)
        rolls.append(roll)
        current_avg = sum(rolls) / i
        averages.append(current_avg)

    return averages

# Set up the simulation parameters
max_rolls = int(input("Enter maximum number of simulated rolls: "))
expected_value = 3.5

# Run the simulation
results = simulate_dice_rolls(max_rolls)

# Create the plot
plt.figure(figsize=(12, 6))
plt.plot( *args: range(1, max_rolls + 1), results, 'b-', label='Running Average')
plt.axhline(y=expected_value, color='r', linestyle='--', label='Expected Value (3.5)')
plt.grid( visible: True, alpha=0.3)

# Customize the plot
plt.title( label: 'Law of Large Numbers: Dice Roll Averages', fontsize=14)
plt.xlabel( xlabel: 'Number of Rolls', fontsize=12)
plt.ylabel( ylabel: 'Average Value', fontsize=12)
plt.legend()
plt.ylim( *args: 1, 6)

plt.show()
```

Here is a python script I wrote in PyCharm, with some help from Grok3. This script simulates a 6-side dice roll, up to a defined number of times. It graphs the running average of the dice rolls, as well as the expected value.

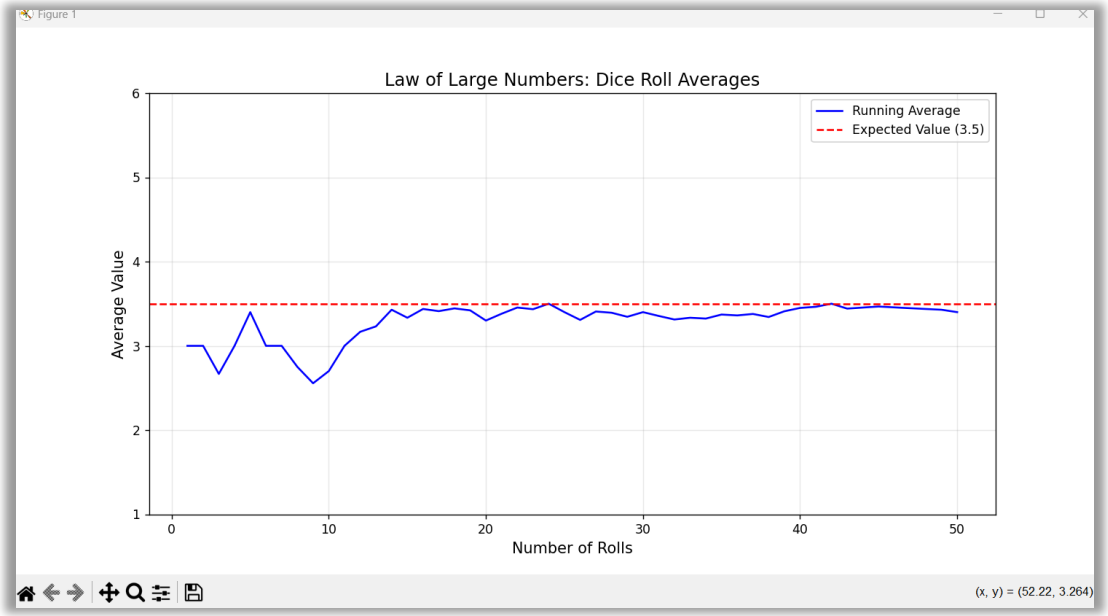


FIGURE 16 SIMULATION OF 50 DICE ROLLS

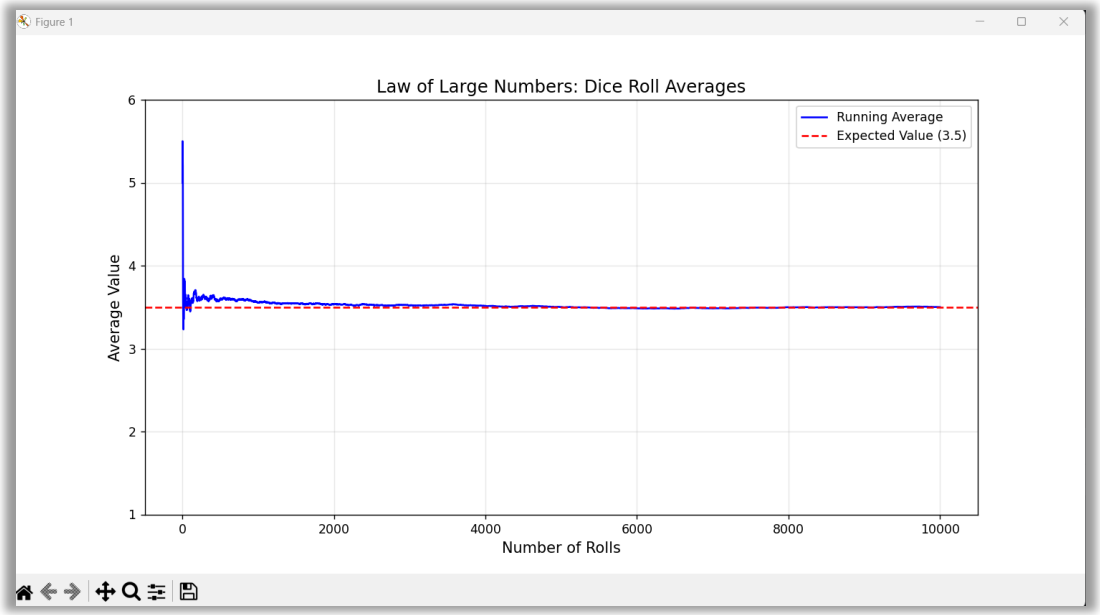


FIGURE 17 SIMULATION OF 10 000 DICE ROLLS

From Figure 16 and 17, it is apparent that the running average of the dice rolls, converges to the mean, illustrating the effect of the weak law of large numbers.

Exercises

Weak law of large numbers: For every $\varepsilon > 0$,

$$\Pr(|M_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Example: I roll s dice 1,000 times. How sure can I be that the average of the rolls is between 3 and 4?

$\Pr(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$

$\mu = 3.5$
$\sigma^2 = 2.92$
$n = 1000$
$\varepsilon = 0.5$

$\Pr(|\bar{X}_{1000} - 3.5| \geq 0.5) \leq \frac{2.92}{1000 \cdot 0.5^2} = 0.01168$

Therefore I can be 98,83 % Sure of an average between 3 and 4

I found this exercise from lecture notes from the University of Notre Dame. The law is manipulated to what they choose to call the effective version. Using this formula, one can estimate the certainty that the mean will be within a certain range. In this case, a 98.83% certainty that the mean of a thousand dice rolls, will fall between 3 and 4.

5.32 Let $X_1, X_2, \ldots$ be a sequence of random variables that converges in probability to a constant a. Assume that $P(X_i > 0) = 1$ for all i.

(a) Verify that the sequences defined by $Y_i = \sqrt{X_i}$ and $Y_i' = a/X_i$ converge in probability.

FIGURE 18 (GEORGE CASELLA, 2001, P. 261)

5.32 (a) Convergence in probability

For any $\varepsilon > 0$,

$P(|\sqrt{X_n} - \sqrt{a}| \leq \varepsilon) = P(|\sqrt{X_n} - \sqrt{a}||\sqrt{X_n} + \sqrt{a}| > \varepsilon|\sqrt{X_n} + \sqrt{a}|)$

$= P(|X_n - a| > \varepsilon|\sqrt{X_n} + \sqrt{a}|)$

$\leq P(|X_n - a| > \varepsilon\sqrt{a}) \to 0$

as $n \to \infty$, since $X_n \to a$, in probability.

Thus $\sqrt{X_n} \to \sqrt{a}$ in probability.

(b) for any $\varepsilon > 0$,

$P(|\frac{a}{X_n} - 1| \leq \varepsilon) = P(\frac{a}{1+\varepsilon} \leq X_n \leq \frac{a}{1-\varepsilon})$.

$= P(a - \frac{a\varepsilon}{1+\varepsilon} \leq X_n \leq a + \frac{a\varepsilon}{1-\varepsilon})$

$\geq P(a - \frac{a\varepsilon}{1+\varepsilon} \leq X_n \leq a + \frac{a\varepsilon}{1-\varepsilon})$ $(a + \frac{a}{1+\varepsilon} < a + \frac{a\varepsilon}{1-\varepsilon})$

$= P(|X_n - a| \leq \frac{a\varepsilon}{1+\varepsilon}) \to 1$

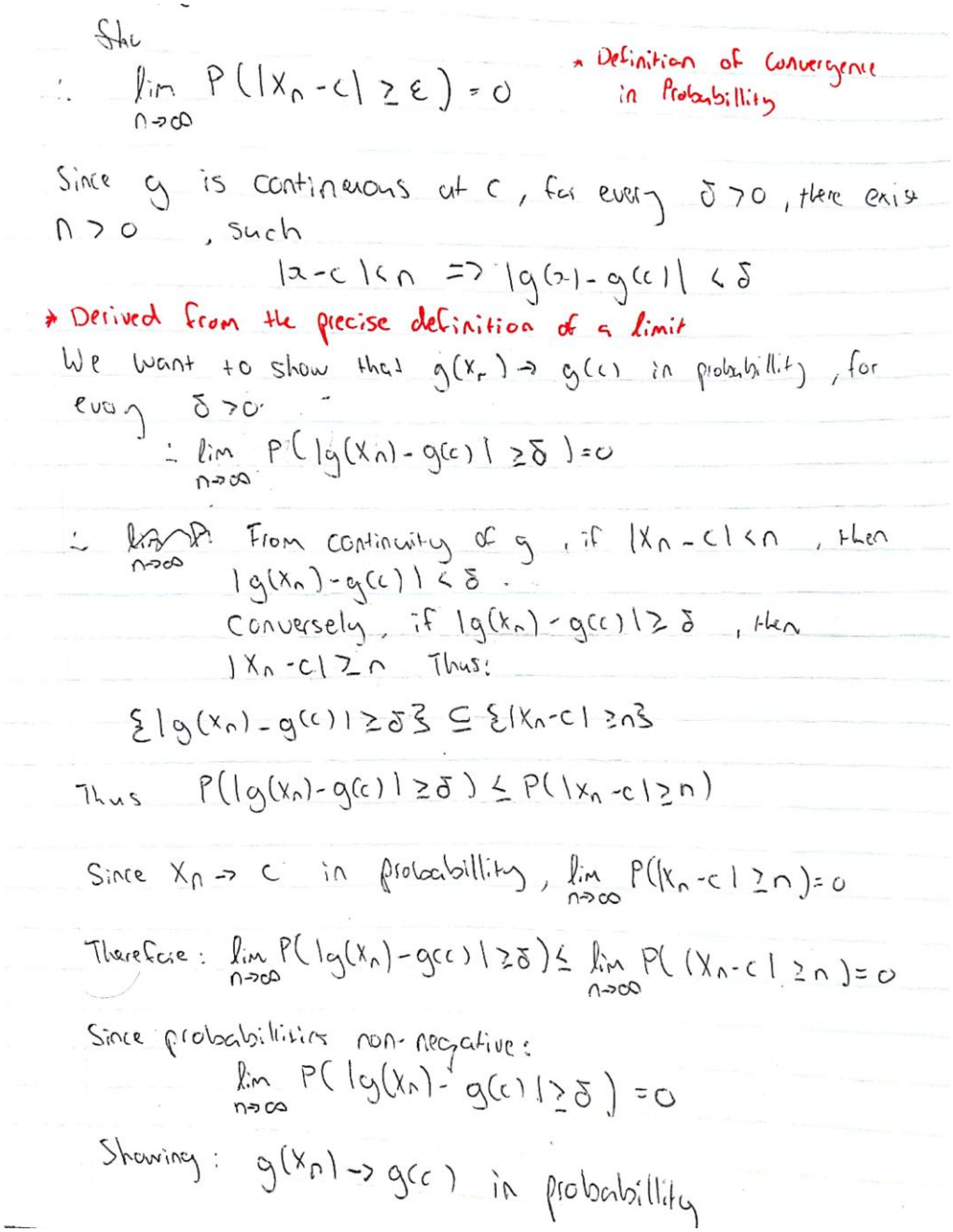
as $n \to \infty$, since $X_n \to a$, in probability.

Thus $\frac{a}{X_n} \to 1$ in probability

I really suck with epsilon delta proofs, for this example, I had to look up the solution manual online

7. Show that if $X_n \to c$ in probability and if g is a continuous function, then $g(X_n) \to g(c)$ in probability.

FIGURE 19 (RICE, 2007, P. 188)



The Convergence in probability can be proved by using the precise definition of a limit (Epsilon, Delta proof) and the definition of continuity. I found this proof rather difficult, and had to search the solution online as well as an explanation from ChatGPT

Reflection

I found the Law of Large Numbers as the most fun and interesting topic up to date. I had lots of fun during my research and designing a simulation to illustrate the LLN.

I learned and practiced some programming skills, particularly in python, which I have only self-studied.

I have a firm understanding of the LLN, and it is easy to explain it intuitively. It is also easy to give and explain examples of the LLN and convergence in probability.

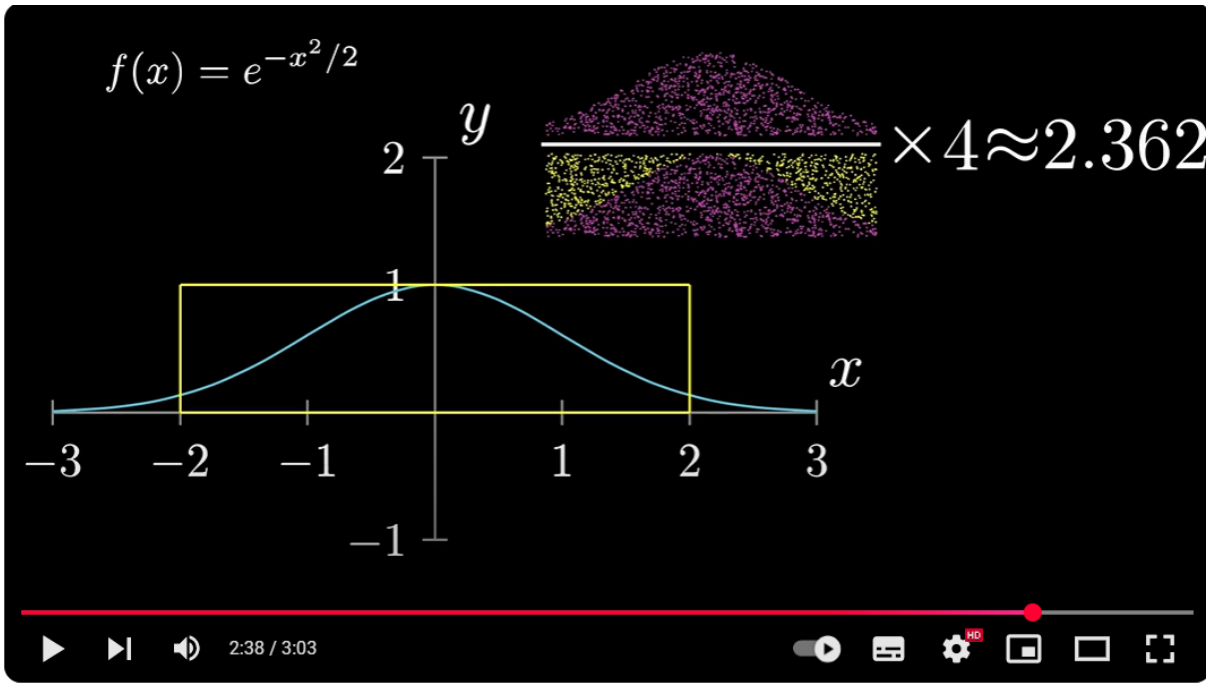
I did however find some of the exercises really challenging. I do understand but struggle to apply limit proof using the precise definition of a limit (epsilon delta proofs). Luckily generative AI can explain the steps of the solution, often omitted or poorly explained in the solution manuals.

I can see how the LLN is relevant, and how it not understanding it can lead to disastrous consequences like the case of the gambler's fallacy.

Monte Carlo Integration

Research Process

- I watched a YouTube video by *Ethan Smith*, shown below.
- I asked Grok 3 and DeepSeek to explain Monte Carlo Integration to me, with some practical examples.
- I consulted my (Rice, 2007) textbook for the definition of Monte Carlo Integration
- I attempted question 19 from (Rice, 2007) and coded the question in both Python and R
- I watched a YouTube video by Younes Lab, which explained how to set up a Mote Carlos Simulation in Python.
- I read the course notes from Louis Aslett on Monte Carlo integration and how to apply it in statistical programming in R.



Monte Carlo Integration

Ethan Smith
200 intekenare
Teken in
913
Deel
Laai af

This video by Ethan Smith visually illustrates that one can estimate the area of the shape/ curve by using fraction of points that landed inside compared to the total points you threw, and multiplying that fraction by the area of the large rectangle.

Definition

We want to evaluate an integral:

$$area = I = \int_a^b g(x)dx$$
$$= (b-a)\int_a^b \frac{1}{b-a} g(x)dx$$

But we know that the density function of a uniform random variable between a and b is $f(x) = \frac{1}{b-a}$. So,

$$I = (b-a)\int_a^b f(x)g(x)dx = (b-a)E\big(g(X)\big)$$

with the expectation over x.

[This is the clue! We are going to define our estimator $\hat{I} = (b-a)\frac{1}{n}\sum_{i=1}^{n} g(X)$]

In some proofs, $\frac{1}{b-a}$ is replaced with any candidate density $p(x)$, and a new function $w(x) = \frac{g(x)}{p(x)}$ is created, $area = \int_a^b \left[\frac{g(x)}{p(x)}\right] p(x)dx = \int_a^b w(x)p(x)dx$. The integral just becomes the expected value of $w(x)$, with the expectation over x.

So, we can basically see that the area under a curve $g(x)$ is the average of the areas of the rectangles formed by the heights ($(g(x)$) under that curve and the length of the rectangle bases $(b-a)$.

- LLN says that $\bar{X}_n \to E(X) = \mu$ as $n \to \infty$.
- LLN says that $\frac{1}{n}\sum_{i=1}^{n} g(x_i) = \overline{g(X)}_n \to E\big(g(X)\big) = \int_a^b f(x)g(x)dx$ as $n \to \infty$.

This means that if we define, $\hat{I} = (b-a)\frac{1}{n}\sum_{i=1}^{n} g(x_i)$, then this must tend to the area I as $n \to \infty$.

Class notes by Prof. Michael von Maltitz from 25 to 26 February 2025

My understanding of Monte Carlo Integration

Monte Carlo Integration is an approximation technique for evaluating definite integrals. The process samples random points and then use their averages to estimate the integral.

I like to break the process down into the following three steps:

1. Select a large number of random points, uniformly, within the domain over which the integral is being evaluated.
2. Calculate the function values at these sampled points and average them.
3. Multiply the average function value by the total size of the domain of the integral. ("size" could be expressed as the area or volume depending on the dimension).

Monte Carlo integration is particularly useful in solving high-dimensional integrals, where traditional methods of integration might be impractical or impossible. Though it is a bit beyond the scope of this module.

Monte Carlo Integration for a multivariate Function

For a function $f: R^n \rightarrow R$ over a domain *D* with volume V(D), the Monte Carlo estimate of the integral is:

$$I \approx V(D) \cdot \frac{1}{N} \sum_{i=1}^{N} f(x_i), \qquad x_i \sim Uniform(D)$$

Where x_i are independent random samples drawn uniformly from D.

Exercises

19. a. Use the Monte Carlo method with $n = 100$ and $n = 1000$ to estimate $\int_0^1 \cos(2\pi x)\, dx$. Compare the estimates to the exact answer.

b. Use Monte Carlo to evaluate $\int_0^1 \cos(2\pi x^2)\, dx$. Can you find the exact answer?

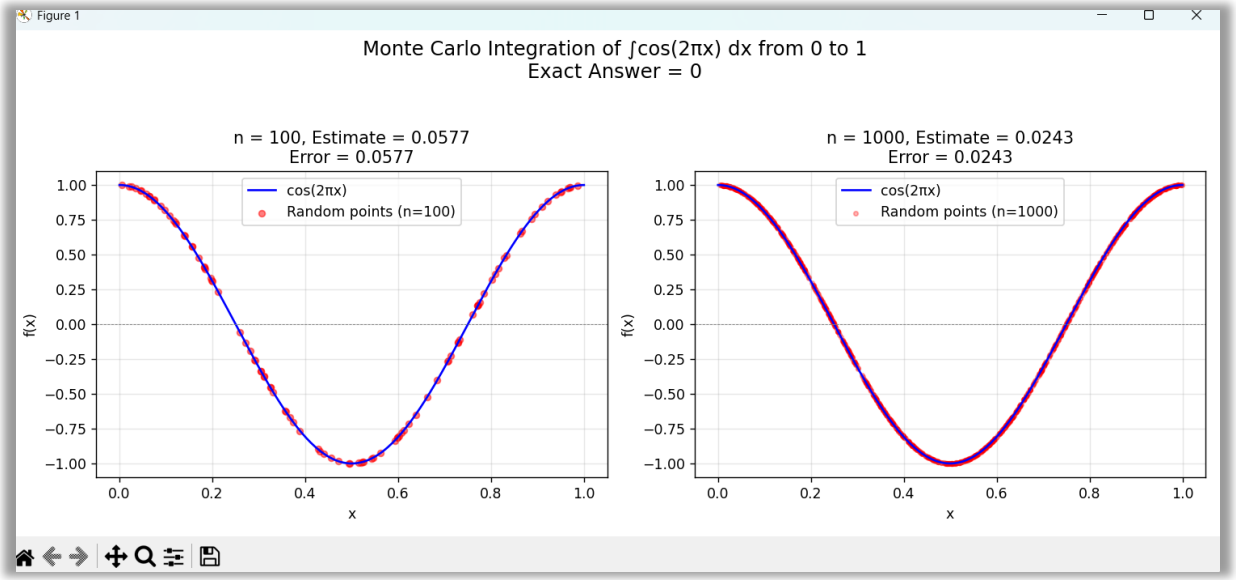
FIGURE 20 (RICE, 2007, P. 190)

Monte Carlo Rice.R | Monte Carlo.py

```python
import numpy as np

# Rice Chapter 6, Question 19 A
def f(x):
    return np.cos(2 * np.pi * x)

def monte_carlo_integral(n):
    x = np.random.uniform(0, 1, n)
    estimate = np.mean(f(x))
    return estimate

n100 = 100
estimate_100 = monte_carlo_integral(n100)
print(f"Estimate with n = {n100}: {estimate_100}")

n1000 = 1000
estimate_1000 = monte_carlo_integral(n1000)
print(f"Estimate with n = {n1000}: {estimate_1000}")

exact = 0
print(f"Exact answer: {exact}")

error_100 = abs(exact - estimate_100)
error_1000 = abs(exact - estimate_1000)
print(f"Error with n = {n100}: {error_100}")
print(f"Error with n = {n1000}: {error_1000}")
```

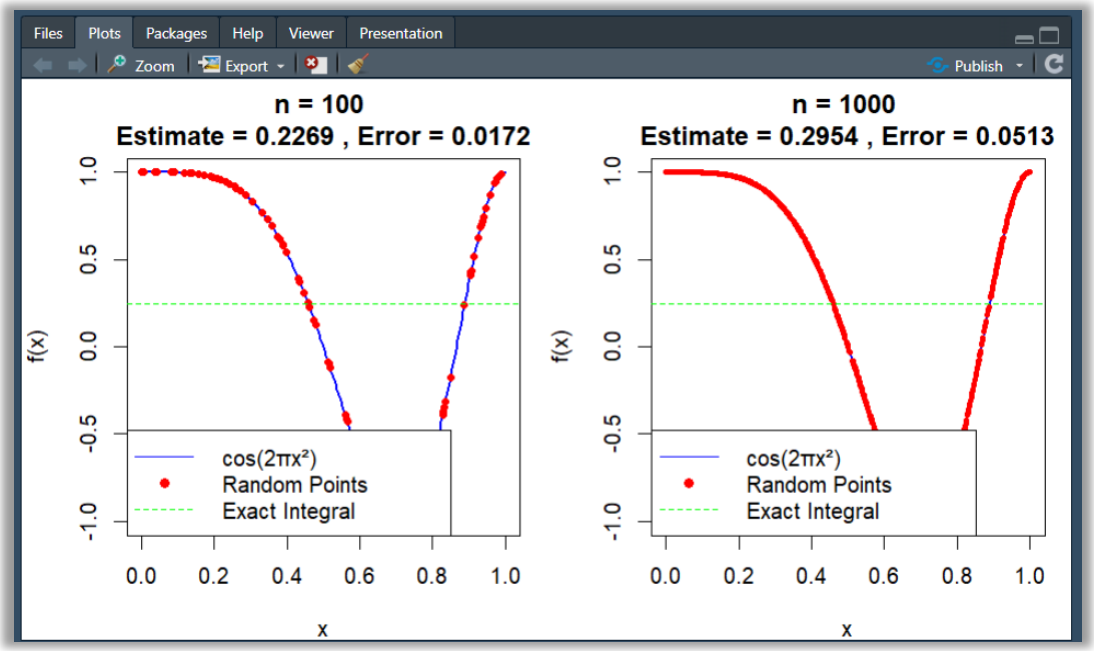
```
C:\Users\franc\PythonProject46\.venv\Scripts\python.exe "C:\Users\franc\Pytho
Estimate with n = 100: 0.025838871887985553
Estimate with n = 1000: -0.0026603757661540113
Exact answer: 0
Error with n = 100: 0.025838871887985553
Error with n = 1000: 0.0026603757661540113
```

I asked Grok 3 to rewrite the code to plot the simulation. I will not include those code. We know that the integral of $\cos(2\pi x)$ from 0 to 1 is 0. We can see that as the number of estimations (n) grows, the estimate approaches 0 and the error approaches 0.



```r

# Rice Chapter 6, Question 19 B
f <- function(x) {
  cos(2 * pi * x^2)
}

monte_carlo_integral <- function(n) {
  x <- runif(n, min: 0, max: 1)
  estimate <- mean(f(x))
  return(estimate)
}

n100 <- 100
estimate_100 <- monte_carlo_integral(n100)

n1000 <- 1000
estimate_1000 <- monte_carlo_integral(n1000)

cat("Estimate with n =", n100, ":", estimate_100, "\n")
cat("Estimate with n =", n1000, ":", estimate_1000, "\n")
```

Question 19 B was Run on R. I used a plugin to write R script in PyCharm



I asked Grok 3 to rewrite the Code to visualise the Graph in R Studio, I will not include the code, but here are the graphs.

Reflection

This section involved lots of programming and I had lots of fun. It was good to practise some statistical programming skills, and I did some examples both in Python and in R.

This section also sparked my interest, since Monte Carlo Integration is closely related to Monte Carlo Simulation, often used in Actuarial Sciences.

Its application in multivariate functions also sparked my interest, though it is beyond the scope of this module, and I did not include an example.

I found it hard to explain at first, but I formulated a good explanation in my understanding section.

Assignment 1

Question:

It is well known that the number π can be calculated numerically as the integral:

$$\pi = \int_0^1 \frac{4}{1+x^2} dx$$

Let $f(x) = \frac{4}{1+x^2}$ *for* $x \in [0,1]$ *and zero elsewhere*

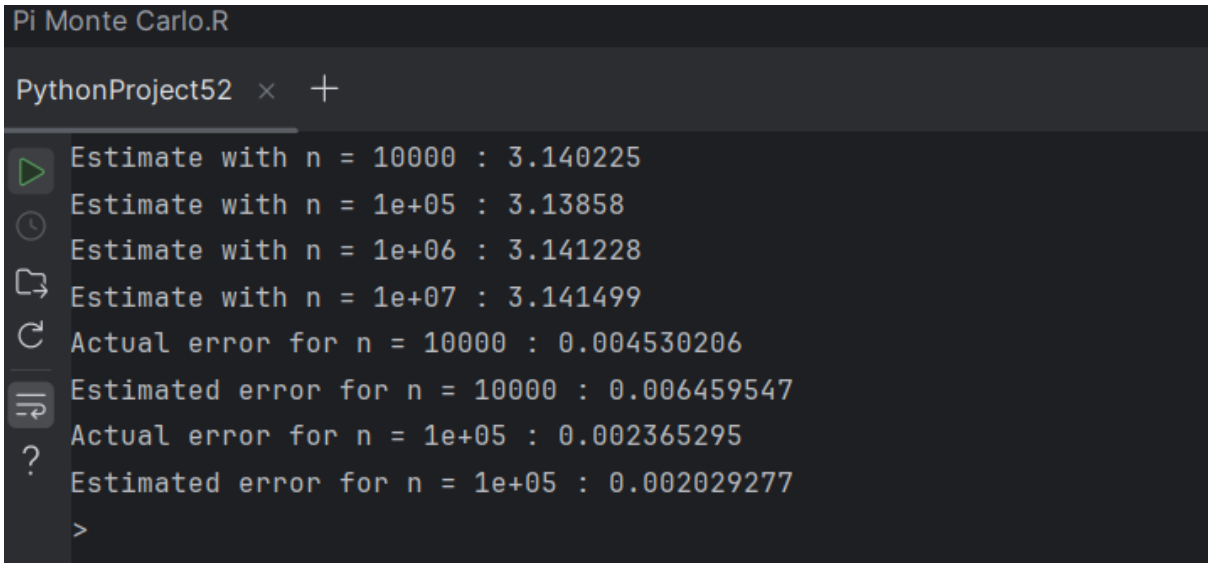
Use ordinary Monte Carlo integration to approximate the integral $I(f)$ *numerically. Do this for several "sample sizes"* n*, for example* $n = 10^5, 10^6, 10^7, \ldots$ *Perform an error estimate pretending that the real value of π is unknown and compare it with the actual error calculated using the real value of π, for* $n = 10^4, 10^5$.

Solution:

Pi Monte Carlo.R | Pi2.R | df | df_plot

```r
f <- function(x)  {
  4 / (1 + x^2)
}

monte_carlo_integral <- function(n) {
  x <- runif(n, min: 0, max: 1) # limits of integration x~U[0,1]
  estimate <- mean(f(x))
  return(estimate)
}

#Function to estimate the standard error
standard_error <- function(n) {
  x <- runif(n, min: 0, max: 1)
  fx_values <- f(x)
  sample_variance <- var(fx_values)
  se <- sqrt(sample_variance / n)
  return(se)
}

n1 <- 10^4
n2 <- 10^5
n3 <- 10^6
n4 <- 10^7
comparison_values <- c(10^4, 10^5)

estimates <- sapply(comparison_values[1:2], monte_carlo_integral)
actual_errors <- abs(estimates - pi)
std_errors <- sapply(comparison_values[1:2], standard_error)

```

```r
cat("Estimate with n =", n1 ,":", monte_carlo_integral(n1), "\n")
cat("Estimate with n =", n2 ,":", monte_carlo_integral(n2), "\n")
cat("Estimate with n =", n3 ,":", monte_carlo_integral(n3), "\n")
cat("Estimate with n =", n4 ,":", monte_carlo_integral(n4), "\n")

cat("Actual error for n =", comparison_values[1], ":", actual_errors[1], "\n")
cat("Estimated error for n =", comparison_values[1], ":", std_errors[1], "\n")

cat("Actual error for n =", comparison_values[2], ":", actual_errors[2], "\n")
cat("Estimated error for n =", comparison_values[2], ":", std_errors[2], "\n")
```



I used R and worked in PyCharm, using an R plugin to solve this problem.

For the Monte Carlo integration, I used two functions, firstly "f" to define $\int_0^1 \frac{4}{1+x^2}\,\mathrm{dx}$. The second function "monte_carlo_integral", contains the estimator, $\hat{I} = (b-a)\frac{1}{n}\sum_{i=1}^{n} f(X)$, where a = 0 and b =1, n is the input value for the number of estimations.

The third function "standard error", calculates the standard error from the formula: $SE = \frac{\sigma}{\sqrt{n}}$, or in this case $SE = \sqrt{\frac{\sigma^2}{n}}$, the actual error is simply calculated as the absolute difference between the estimated and known value of π.

From the estimates we can see the increase in accuracy as n *increases, from accurate to 2 decimal places for* $\mathrm{n} = 10^4$ *to 3 decimal places for* $\mathrm{n} = 10^6$ *and about 4 decimal places for* $\mathrm{n} = 10^7$

The estimated error for $\mathrm{n} = 10^4$ *was interestingly double that of the estimated error, while the actual error and estimated error was identical to 3 decimal places for* $\mathrm{n} = 10^5$*, suggesting that the accuracy of the estimated error increases, as* n *increases.*

Why I chose this problem?

Monte Carlo integration is the section of the module I find most interesting. Monte Carlo integration forms the basis of Monte Carlo simulations, which is quite an important concept in actuarial sciences, which is my major. The integral used to approximate pi, is not difficult to solve using normal calculus methods, but using the same structure of code, one can approximate much, much harder integrals. I chose the approximation of pi, not because it is difficult, but because it serves as an excellent base to explain and illustrate Monte Carlo integration, as you can see the decimal values creeping closer to the true value of π. This example is also good for comparing the estimated standard error to the actual error It was interesting to see the increase in accuracy of the estimated error, as the sample size increases. And finally, this exercise was a good opportunity to sharpen my statistical programming skills in R.

What I learned?

I learned a lot about statistical programming by attempting this exercise. Because the sample sizes were so large, I ran into some numerical stability problems. For example, I tried to define a vector with sample sizes 10^4 *to* 10^7*, but it ended up so large that I was unable to perform certain calculations on it, without troubles. I learned that I could use Welford's method, to increase numerical stability for the larger samples, but it was not necessary to implement in this solution. I learned about the estimated standard error, and it was interesting to see its relationship with the actual error. Monte Carlo integration taught me how to tie together my knowledge of calculus, mathematical statistics and statistical programming, to solve problems in mathematics, statistics, data science and actuarial sciences.*

Gamblers Fallacy

Research Process

- I browsed through the Wikipedia page of the Gambler's Fallacy
- I read an article by The Decision Lab on the Gamblers Fallacy

Definition

In general, Gambler's fallacy refers to our belief that the probability of a random event occurring in the future is influenced by the history of that type of event occurring.

In statistics, it is the mistaken belief that if a particular event has occurred more frequently than usual in the past, it is less likely to happen in the future (or vice versa), despite the events being independent and having fixed probabilities.

Interestingly, I learned that the gambler's fallacy significantly influences consumer behaviour. The most common example is collectables in packaging. For example, if there is a 20% chance that a box of cereal contains a collectable figurine, buying 5 boxes of cereal would not guarantee that you would get the figurine.

Example

The most famous example of gambler's fallacy took place at the roulette tables of a Monte Carlo casino in 1913. For the last 10 spins of the roulette wheel, the ball had landed on black.

Because the gamblers thought a red was long overdue, they started betting against black. But the ball kept on landing on black. As the trend continued, the gamblers became more and more convinced that the next turn would land on red. The crowds and wagers increased-- and so did their losses.

It was only after 26 consecutive blacks that the ball finally landed on red, and the streak came to an end. By this time, the losses were staggering. The casino had made a fortune. This became known as the "Monte Carlo fallacy," which is synonymous with gambler's fallacy.[9]

I found this description of the Monte Carlo Fallacy from The Decision Lab. *Even though 26 consecutive blacks are incredibly unlikely, it is independent from the next spin which still has a probability of 0.5*

The gambler's fallacy and AI

AI-powered tools could help us mitigate gambler's fallacy and make better decisions. For instance, Dr. Lance B. Eliot explores the idea of using AI to help judges avoid the gambler's fallacy—an issue we discussed earlier where judges are more likely to grant asylum to refugees after denying the previous applicant and vice versa.[16] Dr. Eliot suggests that AI could be used to uncover and flag concerning patterns in judicial decisions and help humans avoid fallacies that get in the way of fair, objective decision-making. But this opens another can of worms. AI systems are trained on data from historical legal databases, meaning they often inherit the biases reflected in previous court cases. It would be wrong to assume that AI could help solve the failings of our human reasoning, but this emerging technology can certainly be used as a tool to help us make better decisions in the face of ever-present biases like the gambler's fallacy.

I found this interesting idea whilst researching the The Decision Lab *article. The Gambler's Fallacy is so engrained into human psychology, it often influences decisions in our daily and professional lives. Exploring the idea of using AI to keep us objective in our decision making, seems like an idea worth exploring. That being said, large language models are being trained on our often bias data and may self-inherit that bias.*

Reflection

This was a very short, but interesting section. I did not expect to learn about human psychology in a statistics module, but I understand how often statistical knowledge or misapplication of statistical knowledge influences our behaviour.

I also understand and can explain how the gambler's fallacy is caused by a misinterpretation of the Law of Large Numbers.

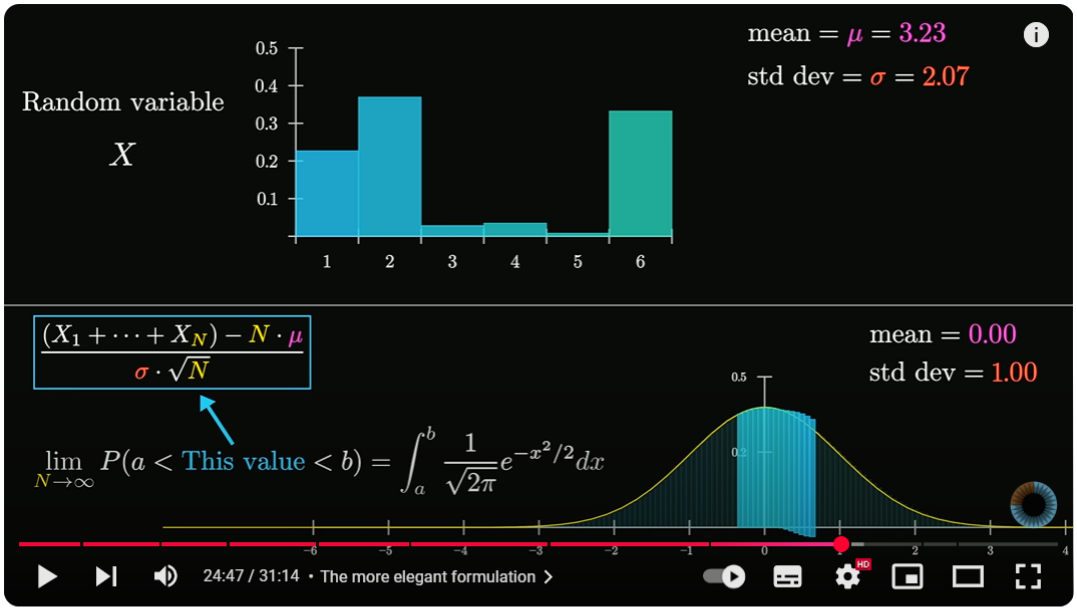
I have a good understanding of the gambler's fallacy and can explain it with a multitude of good examples.

I found the article by Dr. Lance B. Eliot interesting; AI can help us be unbiased, yet AI itself can be bias.

Central Limit Theorem

Research Process

- I watched a YouTube video by 3Blue1Brown on the Central Limit Theorem
- I prompted Grok 3 to do a deep search and explain the central limit theorem to me
- I read the Wikipedia article on the Central Limit Theorem
- I consulted (Rice, 2007) for the definition and proof of the Central Limit Theorem
- I consulted (Rice, 2007) for the definition of the Convergence in Distribution Theorem and Continuity Theorem. Thereafter, I explained each in my own words.
- I searched for some CLT exercises in my (George Casella, 2001) textbook.
- I read an article by Charles Di Renzo on Medium about the Central Limit theorem, confidence intervals and hypothesis testing
- I prompted ChatGPT to give me a good CLT and Monte Carlo exercise, which I completed in R Markdown.
- I consulted my (George Casella, 2001) textbook on the CLT and attempted one of its exercises.
- I read the from probability course article on convergence in distribution.



But what is the Central Limit Theorem?

3Blue1Brown
7,03 m intekenare
Ingeteken
71 k
Deel
Laai af

This video by 3Blue1Brown does an excellent job in simplifying and explaining the various components of the Central Limit Theorem, with clear visual illustrations. The video explains the normal distribution, and how its shape is determined by the parameters mean and variance. It explains the concept of confidence intervals. It explains how the convolutions of distributions, eventually forms a normal distribution. It explained the 3 underlying assumptions of the theorem: the variables must be independent, identically distributed and have finite variance.

My understanding of the Central Limit Theorem and its significance

The Central Limit Theorem revolutionized our understanding of statistics. The theorem could explain observations over a hundred years earlier, namely the occurrence of bell-shaped distributions in large sample sizes. This distribution was then proven as the normal distribution.

The Central Limit Theorem states that as the sample size grows (approaching infinity), the sampling distribution approaches a normal distribution, regardless of the population's original distribution. It has three underlying assumptions, the variables must be independent, identically distributed and have finite variance.

This is significant, because it allows make inferences and analyse data, when their distribution is unknown or non-normal.

Many observations made in nature will follow a normal distribution by the central limit theorem, because the observations themselves the sums of random variables. A simple example would be the distribution of heights in a population, which is well known to be normally distributed. The height of an individual in the population is itself the sum of a large amount of underlying factors, like genetics, diet, environment etcetera.

Convergence In Distribution

DEFINITION

Let $X_1, X_2, \ldots$ be a sequence of random variables with cumulative distribution functions $F_1, F_2, \ldots$, and let X be a random variable with distribution function F. We say that X_n converges in distribution to X if

$$\lim_{n \to \infty} F_n(x) = F(x)$$

at every point at which F is continuous. ■

FIGURE 21 (RICE, 2007, P. 181)

The Theorem of Convergence in Distribution explains that the probability distributions of the sequence (the CDF of all X_n 's) get ever closer to the distribution of X (converges to the CDF of X), as n goes to infinity. This theorem is off course relevant in the context of the central limit theorem, where there would be convergence in distribution to the normal distribution.

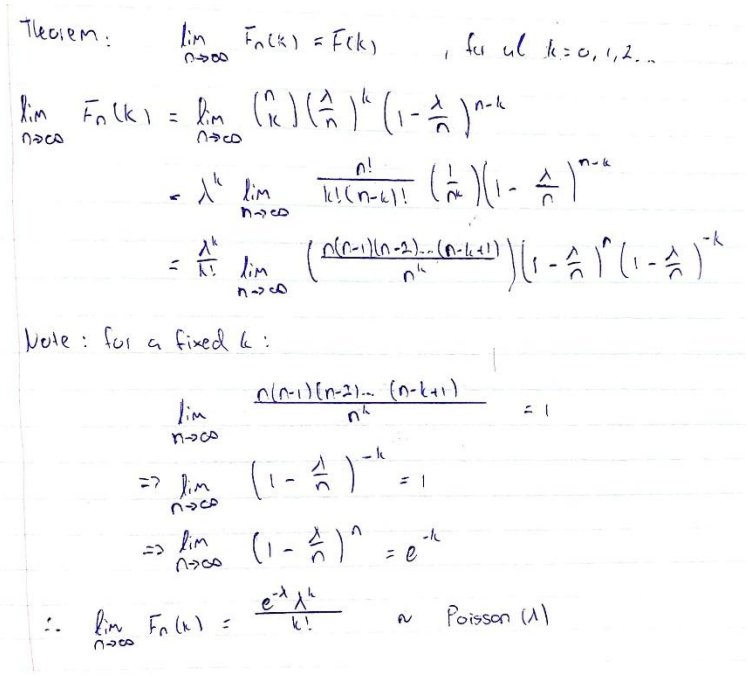
Example 7.7

Let $X_1, X_2, X_3, \cdots$ be a sequence of random variable such that

$$X_n \sim Binomial\left(n, \frac{\lambda}{n}\right), \qquad \text{for } n \in \mathbb{N}, n > \lambda,$$

where $\lambda > 0$ is a constant. Show that X_n converges in distribution to $Poisson(\lambda)$.

FIGURE 22 EXAMPLE FROM PROBABILITY COURSE



Continuity Theorem

THEOREM A *Continuity Theorem*

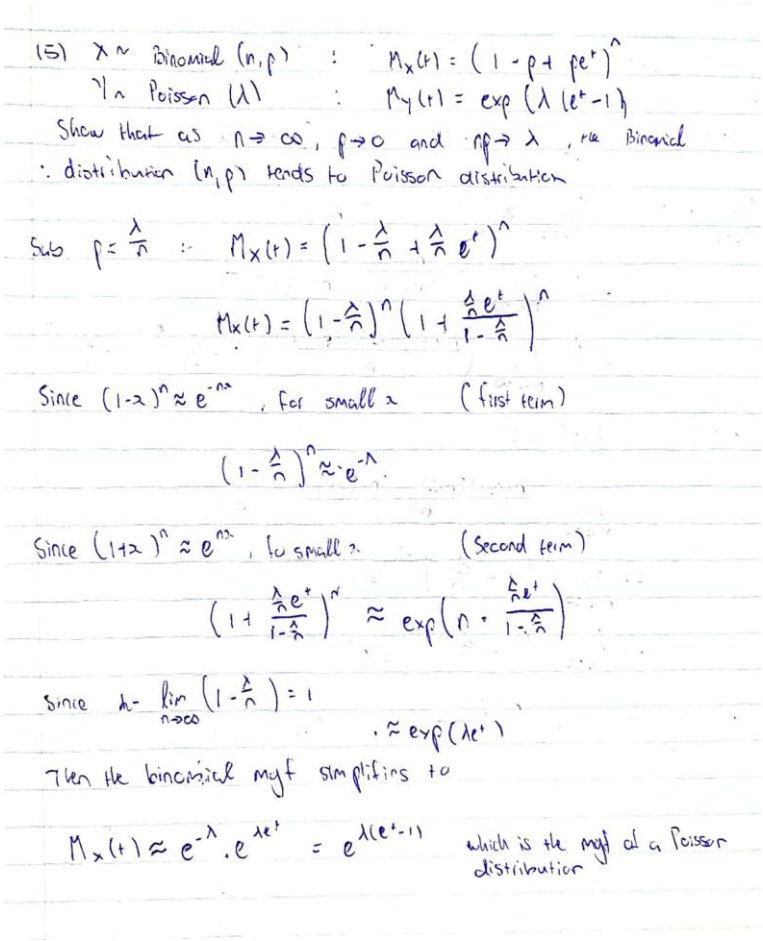
Let F_n be a sequence of cumulative distribution functions with the corresponding moment-generating function M_n. Let F be a cumulative distribution function with the moment-generating function M. If $M_n(t) \to M(t)$ for all t in an open interval containing zero, then $F_n(x) \to F(x)$ at all continuity points of F. ■

FIGURE 23 (RICE, 2007, P. 181)

The Continuity Theorem establishes that the moment generating function convergence, implies distribution function convergence, if the limit function is indeed a valid MGF. This theorem ties with the uniqueness property of MGF, property A (Rice, 2007, p. 155).

5. Using moment-generating functions, show that as $n \to \infty$, $p \to 0$, and $np \to \lambda$, the binomial distribution with parameters n and p tends to the Poisson distribution.

FIGURE 24 (RICE, 2007, P. 188)



This example is a proof of the Poisson limit theorem, which states that a binomial distribution will converge to a Poisson distribution, as the trails become infinitely large

Proof

THEOREM B *Central Limit Theorem*

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having mean 0 and variance σ^2 and the common distribution function F and moment-generating function M defined in a neighborhood of zero. Let

$$S_n = \sum_{i=1}^{n} X_i$$

Then

$$\lim_{n\to\infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \le x\right) = \Phi(x), \qquad -\infty < x < \infty$$

Proof

Let $Z_n = S_n/(\sigma\sqrt{n})$. We will show that the mgf of Z_n tends to the mgf of the standard normal distribution. Since S_n is a sum of independent random variables,

$$M_{S_n}(t) = [M(t)]^n$$

and

$$M_{Z_n}(t) = \left[M\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

$M(s)$ has a Taylor series expansion about zero:

$$M(s) = M(0) + sM'(0) + \tfrac{1}{2}s^2M''(0) + \varepsilon_s$$

where $\varepsilon_s/s^2 \to 0$ as $s \to 0$. Since $E(X) = 0$, $M'(0) = 0$, and $M''(0) = \sigma^2$. As $n \to \infty$, $t/(\sigma\sqrt{n}) \to 0$, and

$$M\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \frac{1}{2}\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + \varepsilon_n$$

where $\varepsilon_n/(t^2/(n\sigma^2)) \to 0$ as $n \to \infty$. We thus have

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + \varepsilon_n\right)^n$$

It can be shown that if $a_n \to a$, then

$$\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^a$$

From this result, it follows that

$$M_{Z_n}(t) \to e^{t^2/2} \quad \text{as } n \to \infty$$

where $\exp(t^2/2)$ is the mgf of the standard normal distribution, as was to be shown. ■

Breakdown of The Central Limit Proof:

Step 1: Understanding the Setup

We have a sequence of independant, identically distributed random variables: $X_1, X_2, \ldots$

For each X_i, $E(X_i) = 0$, $\text{Var}(X_i) = \sigma^2$ and a common MGF, $M_{X_i}(t)$, which exists in a neighborhood of $t = 0$.

Then the sum of X_i: $S_n = \sum_{i=1}^{n} X_i$

Which can be standardized to $Z_n = \frac{S_n}{\sigma\sqrt{n}}$

We want to show that as $n \to \infty$, the distribution of Z_n approaches the standard Normal distribution.

$$P(Z_n \leq x) \to \Phi(x)$$

where $\Phi(x)$ is the CDF of the standard normal

Step 2: Computing the MGF of S_n

Since S_n is a sum of independant random variables, the MGF is given by: $M_{S_n}(t) = [M(t)]^n$

Since $Z_n = \frac{S_n}{(\sigma\sqrt{n})}$, we obtain its MGF:

$$M_{Z_n}(t) = M_{S_n}\left(\frac{t}{\sigma\sqrt{n}}\right)$$

Substituting $M_{S_n}(t)$: $M_{Z_n}(t) = \left[M\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$

Step 3: Taylor Expansion of $M(s)$

Since the MGF, $M(s) = E(e^{sX})$ exists in a neighborhood of $s=0$, we can perform a Taylor series expansion around $s = 0$:

$$M(s) = 1 + sE(X) + \frac{s^2}{2}E(X^2) + \epsilon_s$$

where ϵ_s is the remainder term satisfying $\frac{\epsilon_s}{s^2} \to 0$, as $s \to 0$. Given $E(X) = 0$ and $E(X^2) = \sigma^2$, the expansion simplifies to:

$$M(s) = 1 + \frac{s^2}{2}\sigma^2 + \epsilon_s$$

We need $M(s)$ evaluated at $s = \frac{t}{\sigma\sqrt{n}}$, since this is the argument used in $M_{Z_n}(t)$. Substituting $s = \frac{t}{\sigma\sqrt{n}}$:

$$M\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \frac{1}{2}\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + \epsilon_{\frac{t}{\sigma\sqrt{n}}}$$

$$= 1 + \frac{1}{2}\sigma^2 \cdot \frac{t^2}{\sigma^2 n} + \epsilon_{\frac{t}{\sigma\sqrt{n}}}$$

$$= 1 + \frac{t^2}{2n} + \epsilon_{\frac{t}{\sigma\sqrt{n}}}$$

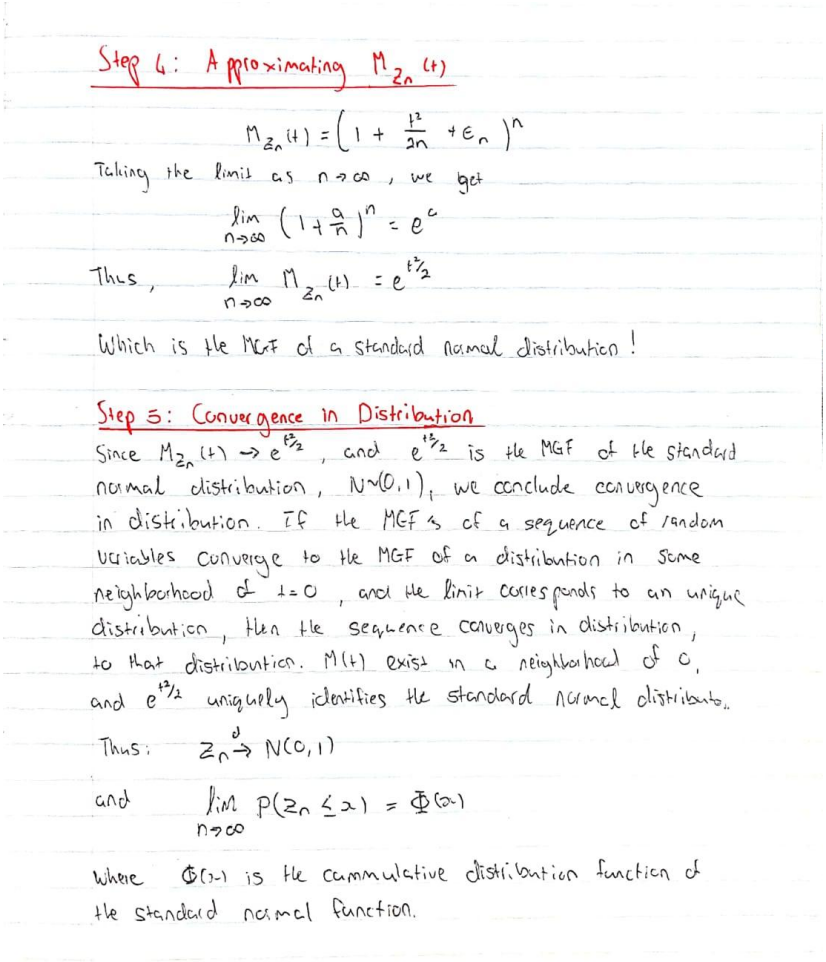
Denote the remainer as $\epsilon_n = \epsilon_{\frac{t}{\sigma\sqrt{n}}}$, Since $\frac{t}{\sigma\sqrt{n}} \to 0$, as $n \to \infty$ the property of the remainder gives

$$\epsilon_n \cdot \frac{\sigma^2 n}{t^2} \to 0$$

The Central Limit Theorem Proof is built upon 3 underlying assumptions:

- The variables $X_1, X_2, \ldots$ must be independent. I am aware that there are variations of the CLT that relax this condition, allowing some weakly dependant variables.
- The variables $X_1, X_2, \ldots$ must be identically distributed. There is also a version of the CLT that relaxes this condition, known as Lyapunov CLT
- The variables $X_1, X_2, \ldots$ must have a finite variance, σ^2.

Back to Table of Contents



Step 4: Approximating $M_{Z_n}(t)$

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + \epsilon_n\right)^n$$

Taking the limit as $n \to \infty$, we get

$$\lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n = e^a$$

Thus, $$\lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2}$$

Which is the MGF of a standard normal distribution!

Step 5: Convergence in Distribution

Since $M_{Z_n}(t) \to e^{t^2/2}$, and $e^{t^2/2}$ is the MGF of the standard normal distribution, $N(0,1)$, we conclude convergence in distribution. If the MGFs of a sequence of random variables converge to the MGF of a distribution in some neighborhood of $t=0$, and the limit corresponds to an unique distribution, then the sequence converges in distribution, to that distribution. $M(t)$ exist in a neighborhood of 0, and $e^{t^2/2}$ uniquely identifies the standard normal distribution.

Thus: $Z_n \xrightarrow{d} N(0,1)$

and $$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x)$$

Where $\Phi(x)$ is the cummulative distribution function of the standard normal function.

I prompted ChatGPT 4.5 to help me break down the individual steps of the CLT. I made some modifications to its response so that it better aligns with the steps given in Rice. I found that writing out a comprehensive proof really helped me to understand it and I can now explain it a much better and more concisely.

Exercises

A factory produces light bulbs, and the **lifespan (in hours) of a single light bulb** follows an **unknown distribution** with a **minimum of 500 hours** and a **maximum of 1500 hours**.

1. **Estimate the mean and variance** of the lifespan of a single light bulb, assuming a **uniform distribution** over [500,1500].
2. Suppose a **random sample of 64 light bulbs** is taken. Use the **Central Limit Theorem** to approximate the probability that the **sample average lifespan** exceeds 1100 hours.
3. **Verify your result with a Monte Carlo simulation** in R using 10,000 samples of size 64.

Solution: (made in R Markdown)

Question 1

Let $X \sim U(500, 1500)$, Then $E(X) = \frac{500+1500}{2} = 1000$ And $Var(x) = \frac{(1500-500)^2}{12} = \frac{250000}{3}$

Question 2

The sample mean $\bar{X}$ of n i.i.d uniform random variables follows approximately a normal distribution by CLT, when n is large. In this case n = 64 and the X_i's are i.i.d with finite mean and variance.

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\bar{X} \approx N\left(1000, \frac{83333.33}{64}\right)$$

$$\bar{X} \approx N(1000, 1302.083)$$

$$\sigma_{\bar{X}} = \sqrt{1302.08}\ 36.09$$

Thus, by CLT $\bar{X} \approx N(1000, 36.09^2)$

Question 3

We standardize $(\bar{X})$ using the Z-score formula:

Z = $\frac{\bar{X}-E[X]}{\sigma_{\bar{X}}}$

Substituting values with $(\sigma_{\bar{X}} \approx 36.09)$:

Z = $\frac{1100-1000}{36.09} = \frac{100}{36.09} \approx 2.77$

Now, we compute $\left(P(Z > 2.77)\right)$ using standard normal tables:

$$P(Z > 2.77) = 1 - P(Z \leq 2.77)$$

$$P(Z \leq 2.77) \approx 0.997$$

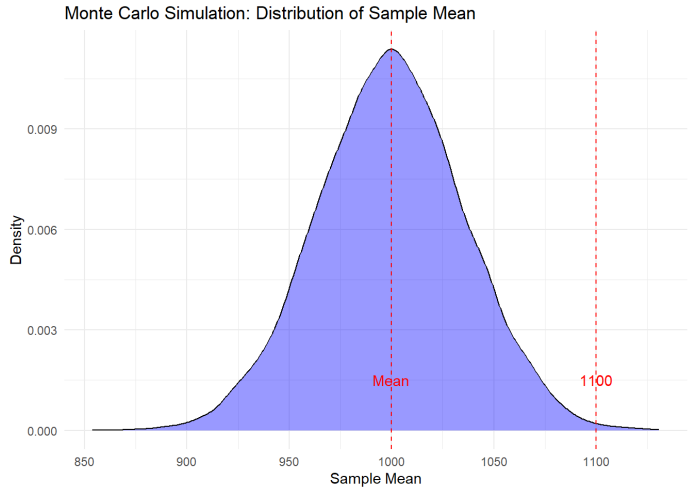
$P(Z > 2.77)$= 1 - 0.9972 = 0.0028

So, the probability that the sample mean exceeds 1100 hours is approximately:

$P(\bar{X} > 1100) \approx 0.0028$

Question 4

```
set.seed(123)
a <- 500
b <- 1500
n <- 64
num_simulations <- 10000

sample_means <- replicate(num_simulations, mean(runif(n, min =a, max =b)))

prob <- mean(sample_means > 1100)

cat("Monte Carlo estimate of P(X-bar > 1100", prob, "\n")

## Monte Carlo estimate of P(X-bar > 1100 0.0024

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.4.2

df <- data.frame(sample_means)
ggplot(df, aes(x = sample_means)) +
  geom_density(fill = "blue", alpha = 0.4) +
  geom_vline(xintercept = c(1000, 1100), linetype = "dashed", color = "red
") +
  annotate("text", x = c(1000, 1100), y = 0.0015, label = c("Mean", "1100"
), color = "red") +
  labs(title = "Monte Carlo Simulation: Distribution of Sample Mean",
       x = "Sample Mean", y = "Density") +
  theme_minimal()
```

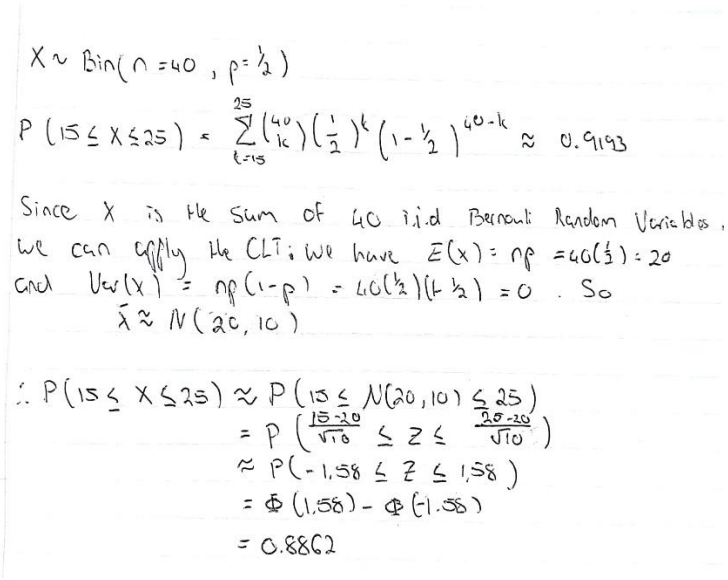


Using R markdown to solve this problem was rather tedious with all the LaTex. This problem shows how a uniform distribution converges to a normal distribution with the CLT. All code was written myself.

Example(s)

Let's consider the example of flipping a fair coin 40 times independently. What's the probability of getting between 15 to 25 heads? First compute this exactly and then give an approximation using the CLT.

FIGURE 25 (TSUN, 2020)



Here, a normal approximation was good, but not as precise as calculating the probability with a Binomial or Bernoulli distribution.

25. Let X be a continuous random variable with density function $f(x) = \frac{3}{2}x^2, -1 \le x \le 1$. Sketch this density function. Use the central limit theorem to sketch

the approximate density function of $S = X_1 + \cdots + X_{50}$, where the X_i are independent random variables with density f. Similarly, sketch the approximate density functions of $S/50$ and $S/\sqrt{50}$. For each sketch, label at least three points on the horizontal axis.

FIGURE 26 (RICE, 2007, P. 191)

Question 25, Rice

$f(x) = \frac{3}{2}x^2, \quad -1 \le x \le 1$

$S = X_1 + X_2 + \dots + X_{50}$

By CLT $\therefore S \sim N(n\bar{E}(x), n\mathrm{Var}(x)), \quad n = 50$

$\therefore E(x) = \int_{-1}^{1} x \frac{3}{2}x^2 dx$

$E(x) = \frac{3}{2}\int_{-1}^{1} x^3 dx$

$E(x) = 0$

$E(x^2) = \int_{-1}^{1} x^2 \cdot \frac{3}{2}x^2 dx$

$= \frac{3}{2}\int_{-1}^{1} x^4 dx$

$= \frac{3}{2}\left[\frac{x^5}{5}\right]_{-1}^{1}$

$= \frac{3}{5}$

$\mathrm{Var}(x) = \frac{3}{5}$

$\therefore \mathrm{Var}(S) = 50 \times \frac{3}{5} = 30$

So, $S \sim N(0, 30)$, Since S approximately normal, from CLT,

$f_S(x) \approx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$, Substitute $\sigma^2 = 30$:

$f_S(x) \approx \frac{1}{\sqrt{60\pi}} e^{-\frac{x^2}{60}}$

Approximate Density of $S/50$

$E\left[\frac{S}{50}\right] = \frac{E(S)}{50} = 0$

$\mathrm{Var}\left(\frac{S}{50}\right) = \frac{\mathrm{Var}(S)}{50^2} = \frac{30}{2500} = \frac{3}{250}$

Thus $\frac{S}{50} \sim N\left(0, \frac{3}{250}\right)$

with approximate density $f_{\frac{S}{50}}(x) \approx \frac{5\sqrt{15}}{3\sqrt{\pi}} e^{-\frac{300x^2}{3}}$

Approximate Density of $S/\sqrt{50}$

$E\left(\frac{S}{\sqrt{50}}\right) = \frac{E(S)}{\sqrt{50}} = 0$

$\mathrm{Var}\left(\frac{S}{\sqrt{50}}\right) = \frac{\mathrm{Var}(S)}{50} = \frac{30}{50} = 0.6$

Thus $\frac{S}{\sqrt{50}} \sim N(0, 0.6)$

with approximate density $f_{\frac{S}{\sqrt{50}}}(x) \approx \frac{\sqrt{15}}{3\sqrt{2\pi}} e^{-\frac{3x^2}{6}}$

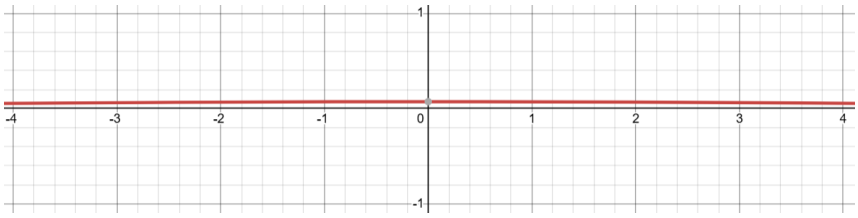


FIGURE 27 $f_S(x)$

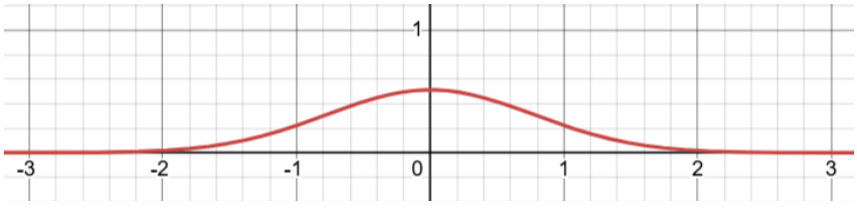


FIGURE 29 $f_{\frac{S}{\sqrt{50}}}(x)$

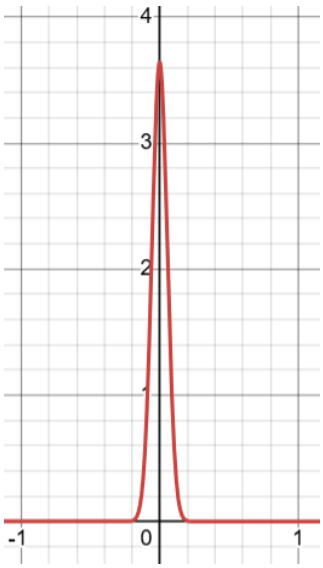
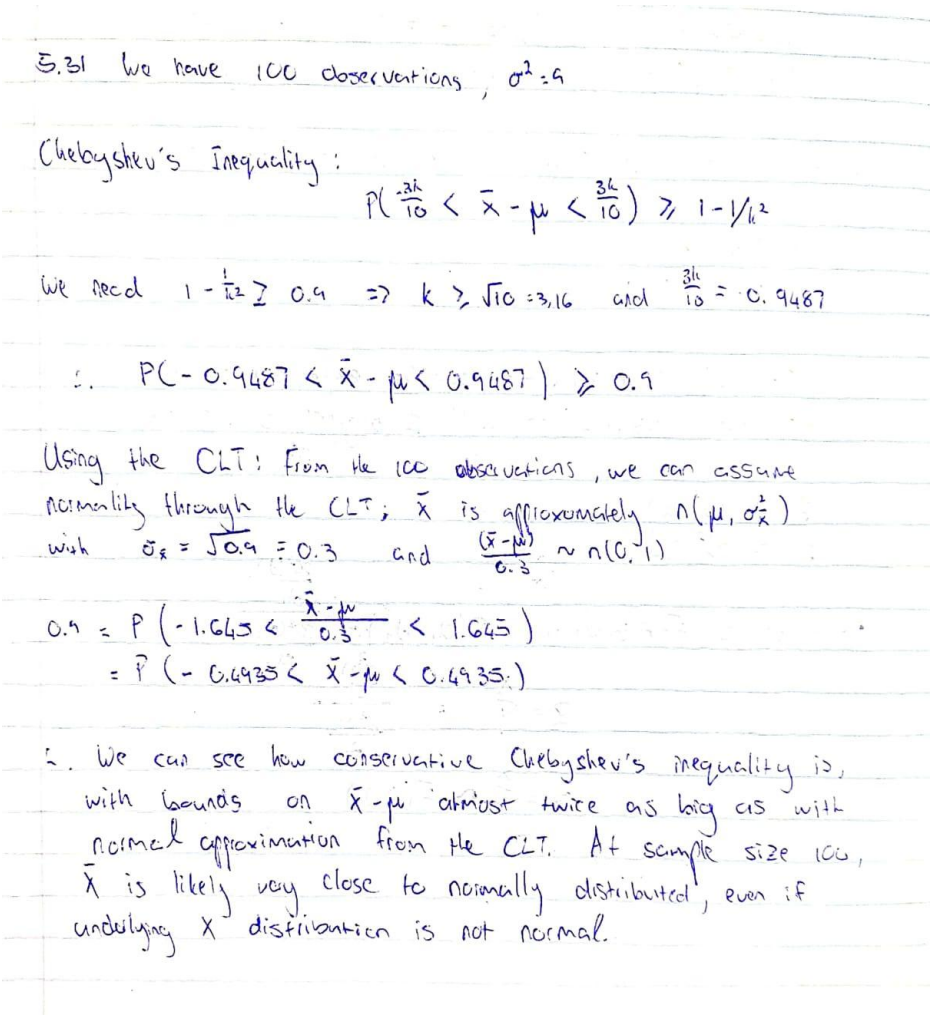


FIGURE 28 $f_{\frac{S}{50}}(x)$

5.31 Suppose $\bar{X}$ is the mean of 100 observations from a population with mean μ and variance $\sigma^2 = 9$. Find limits between which $\bar{X} - \mu$ will lie with probability at least .90. Use both Chebychev's Inequality and the Central Limit Theorem, and comment on each.

FIGURE 30 (GEORGE CASELLA, 2001, P. 260)



5.31 We have 100 observations, $\sigma^2 = 9$

Chebyshev's Inequality:

$$P\left(\frac{-3k}{10} < \bar{x} - \mu < \frac{3k}{10}\right) \geq 1 - 1/k^2$$

We need $1 - \frac{1}{k^2} \geq 0.9 \Rightarrow k \geq \sqrt{10} = 3.16$ and $\frac{3k}{10} = 0.9487$

$$\therefore P(-0.9487 < \bar{x} - \mu < 0.9487) \geq 0.9$$

Using the CLT: From the 100 observations, we can assume normality through the CLT; $\bar{x}$ is approximately $n(\mu, \sigma_{\bar{x}}^2)$ with $\sigma_{\bar{x}} = \sqrt{0.9} = 0.3$ and $\frac{(\bar{x} - \mu)}{0.3} \sim n(0,1)$

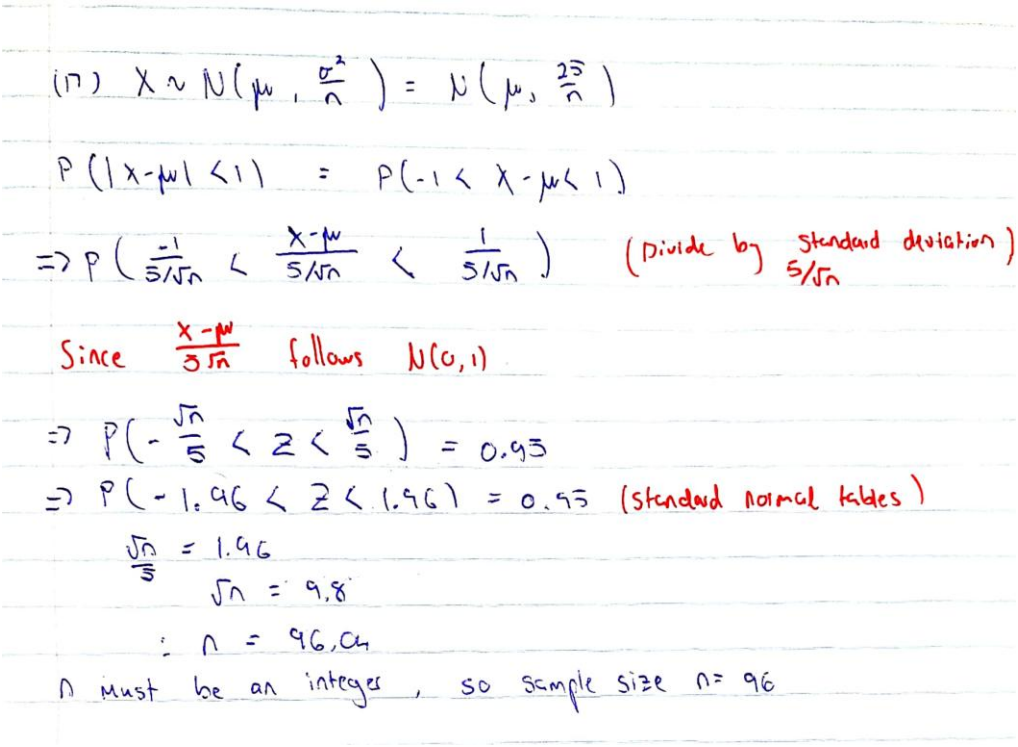
$$0.9 = P\left(-1.645 < \frac{\bar{x} - \mu}{0.3} < 1.645\right)$$
$$= P(-0.4935 < \bar{x} - \mu < 0.4935)$$

∴ We can see how conservative Chebyshev's inequality is, with bounds on $\bar{x} - \mu$ almost twice as big as with normal approximation from the CLT. At sample size 100, $\bar{x}$ is likely very close to normally distributed, even if underlying X distribution is not normal.

Here we can see how normal approximation through the CLT is much mor accurate than Chebyshev's inequality.

17. Suppose that a measurement has mean μ and variance $\sigma^2 = 25$. Let $\overline{X}$ be the average of n such independent measurements. How large should n be so that $P(|\overline{X} - \mu| < 1) = .95$?

FIGURE 31 (RICE, 2007, P. 189)



Reflection

Finally, we have one limit theorem to rule them all: the Central Limit Theorem. I find the idea of the CLT crazy, even the probabilities of random errors observed in nature and astrology can be defined by a singular distribution.

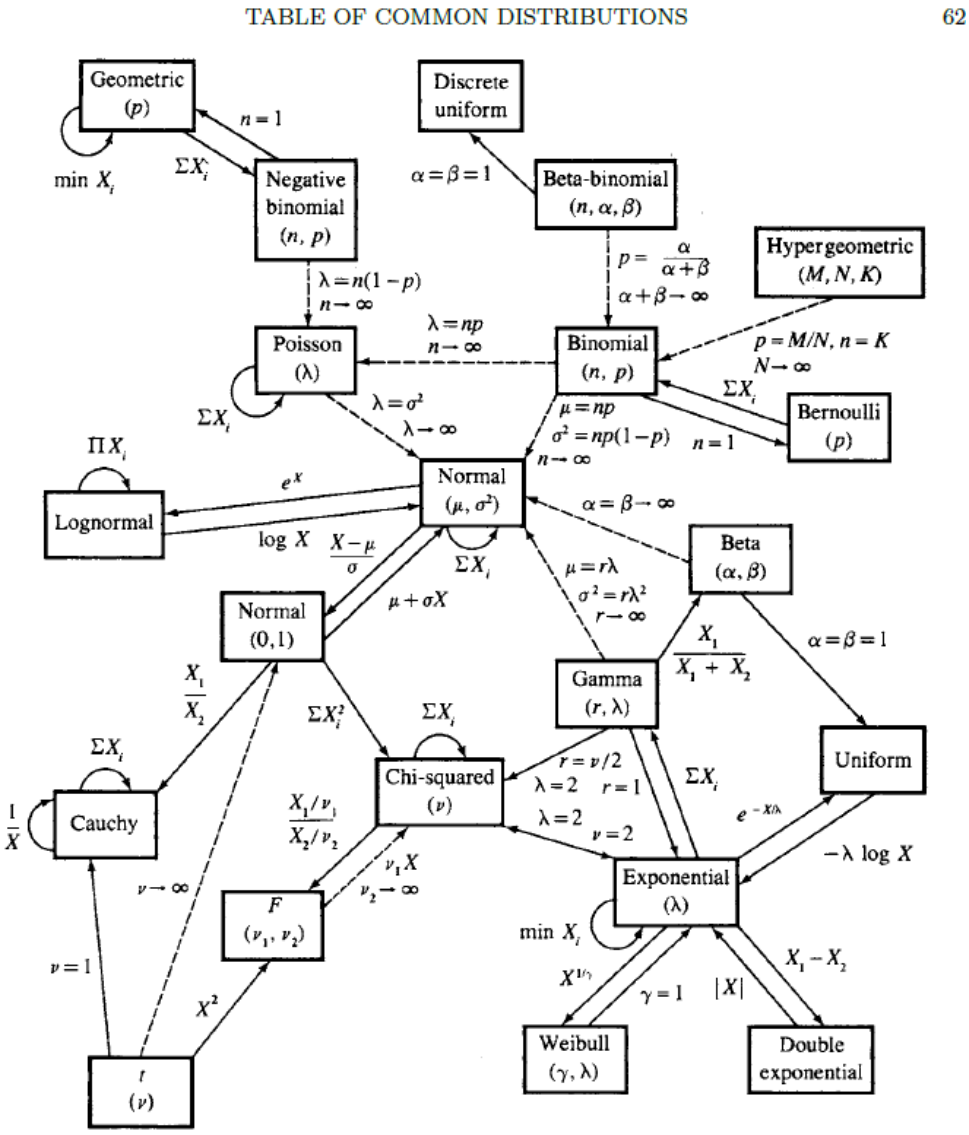
I understand and can explain the CLT well, including it's profound significance.

Limit theorems, and most importantly, the Central Limit Theorem is one of the 4 methods we use to make statistical inferences.

I spent a lot of time searching for exercises on the CLT, since most of them a garbage, including those in Rice.

I wasted some time doing bad exercises, which I chose to remove from my portfolio. I also chose to do one of the examples in R markdown, since it was what we were learning in STSM2634 at the time, and I wanted to kill two birds with one stone. It proved much more time consuming to use LaTex instead of equation editor, but on the plus side, I did learn to use LaTex.

Distributions Derived from the Normal Distribution



Relationships among common distributions. Solid lines represent transformations an special cases, dashed lines represent limits. Adapted from Leemis (1986).

FIGURE 32 (GEORGE CASELLA, 2001, P. 627)

From this web of distributions, we can clearly see the normal distribution in the centre, we can see how the standard normal and gamma distributions, are connected to the chi-squared distribution. How the F distribution is connected to the t and chi-squared distributions. And how the t distribution is related to the standard normal and F distribution.

Index of exercises and examples

Chi-Squared Distribution

T-Distribution

F Distribution

Sample mean and sample variance

Chi-Square Distribution

Research Process

- I watched a YouTube video by Equitable Equations to introduce myself to the distribution. He briefly explained the right skewness and how the chi-square distribution, will also approach a normal density as n increases, from the CLT. He briefly touched on using chi-squared distributions in R
- I watched a YouTube video by Very Normal, which explained the Chi-Squared test, including application in R.
- I read the Wikipedia article on the Chi-Square Distribution, where I learned that it is a special case of the Gamma distribution and the univariate Wishart distribution.
- I consulted my (Rice, 2007) textbook for the definition, related properties and proof of the Chi-Square Distribution.
- I read the Wikipedia article on Hypothesis test, to help me understand the underlying concepts of statistical test, in this case to help me with the Chi-Squared tests.
- I read the Wikipedia article on the Chi-Squared Goodness of Fit Test.
- I searched for examples of goodness of fit test online and found a tutorial form the University of Leicester

Definition

DEFINITION

If Z is a standard normal random variable, the distribution of $U = Z^2$ is called the chi-square distribution with 1 degree of freedom. ■

FIGURE 33 (RICE, 2007, P. 192)

DEFINITION

If $U_1, U_2, \ldots, U_n$ are independent chi-square random variables with 1 degree of freedom, the distribution of $V = U_1 + U_2 + \cdots + U_n$ is called the *chi-square distribution with n degrees of freedom* and is denoted by χ^2_n. ■

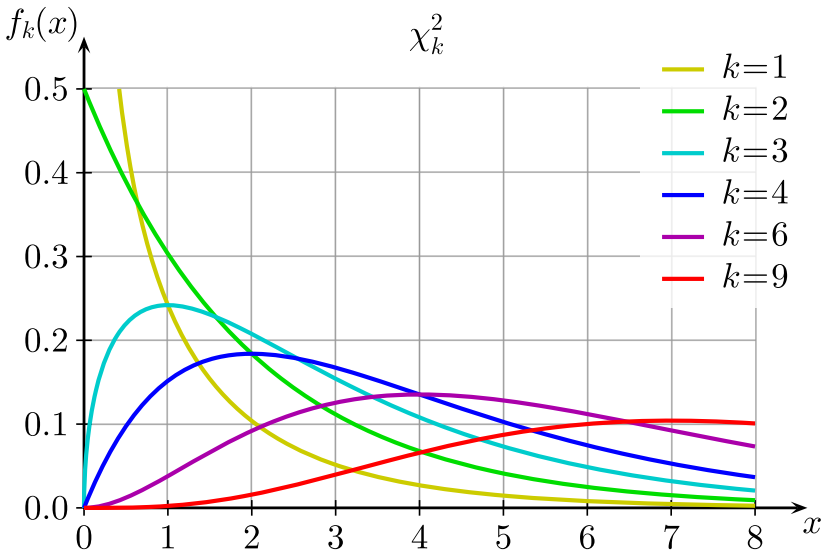
FIGURE 34 (RICE, 2007, P. 192)

The Probability Density Function for a Chi-Squared Distribution with n degrees of freedom, is a gamma distribution with $\alpha = \frac{n}{2}$ and $\lambda = \frac{1}{2}$:

$$f(x,n) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \qquad x \geq 0$$

And its moment-generating function is:

$$M(t) = (1-2t)^{-\frac{n}{2}} \qquad for\ t < \frac{1}{2}$$



Here is the pdf of the chi-squared distribution with various degrees of freedom. (Wikipedia)

My understanding of the Chi-Squared Distribution

The chi-squared distribution is a continuous probability distribution, with n degrees of freedom, that arises from the sum of the squares of n independent standard normal variables. The distribution is therefore controlled by the single parameter n.

The shape is skewed to the right but becomes more symmetric as n increases. As $n \to \infty$ the distribution approaches normality by the central limit theorem.

With mean = n

And variance = 2n

The chi-squared distribution is used in the Goodness-of-Fit Test, used to test whether observed categorical data, match an expected distribution of that data.

Chi Squared Goodness of Fit Test

The Chi-Squared Goodness of Fit Test compares the observed frequencies from the sample data to the expected frequencies based on a hypothesis. The test statistic, denoted as χ^2

, measures how much the observed data deviates from the expected. If the deviation is small, your data fits the expected distribution and if the deviation is large, your data forms a different distribution from the hypothesis. Pearson's chi-square test uses a measure of goodness of fit which is the sum of differences between observed and expected outcome frequencies, each squared and divided by the expectation:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where:

- O_i = Observed frequency for category i (bin i)
- E_i = Expected frequency for category i (bin i)

The expected frequency is calculated by:

$$E_i = \big(F(Y_u) - F(Y_l)\big)N$$

Where:

- F = the cumulative distribution function for the probability distribution being tested
- Y_u = the upper limit for bin i
- Y_l = the lower limit for bin i
- N = the sample size

You then compare this χ^2 value to the chi-square distribution table, based on your degrees of freedom (df) and significance level (α), to decide whether to reject the null hypothesis (H_0)

Where:

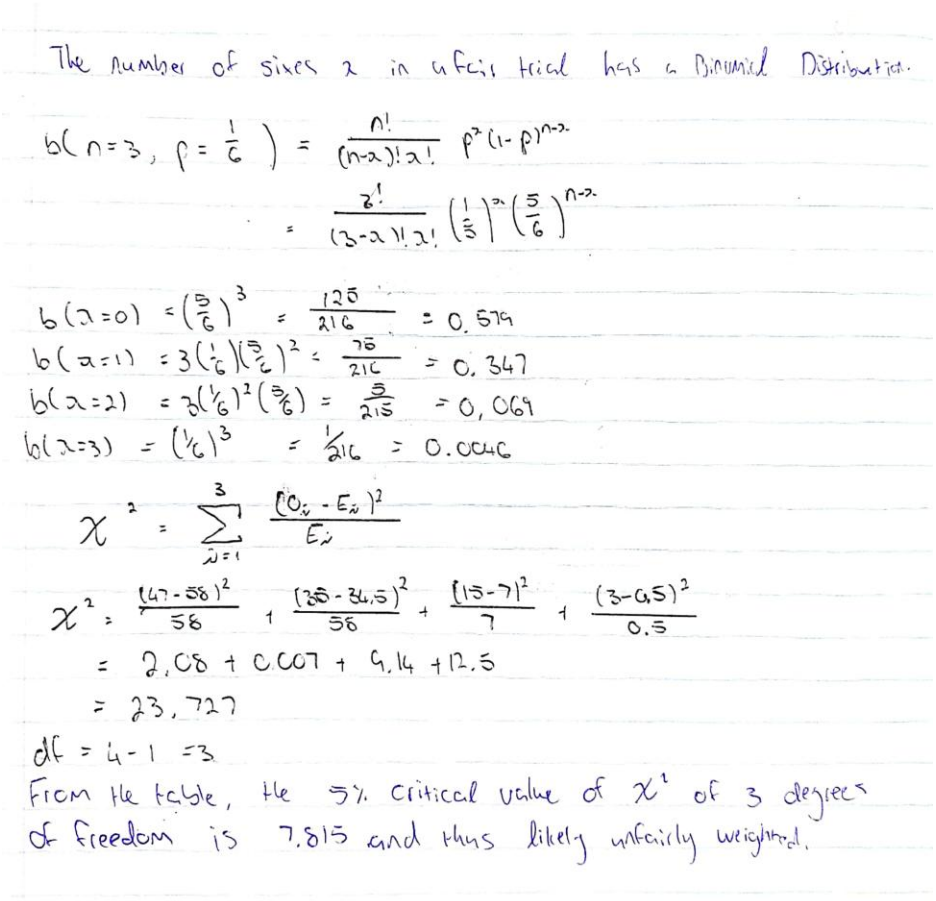
- Null Hypothesis (H_0): The observed distribution matches the expected distribution

Example: University of Leicester

A gambler plays a game that involves throwing 3 dice in a succession of trials. His winnings are directly proportional to the number of sixes recorded. If the dice are fair, what is the probability distribution that governs the outcome of each throw? The frequencies of the sixes observed in 100 trials are recorded, together with their expected values, in the following table:

Number of sixes	Expected Count	Observed Count
0	58	47
1	34.5	35
2	7	15
3	0.5	3

You are asked to assess whether it is likely that the dice have been unfairly weighted, using a chi-square test of goodness of fit.



The dice are likely unfairly weighted, since the observed frequencies deviate significantly from the expected value.

```r
observed_counts <- c(47, 35, 15, 3)
expected_counts <- c(58, 34.5, 7, 0.5)

chi_square_test <- chisq.test(x = observed_counts,
                              p = expected_counts / sum(expected_counts))
```

The test could easily be done in R, using the chisq.test() function. The 2 inputs needed are the observed counts and the expected counts, which could be saved as a vector of even a matrix.

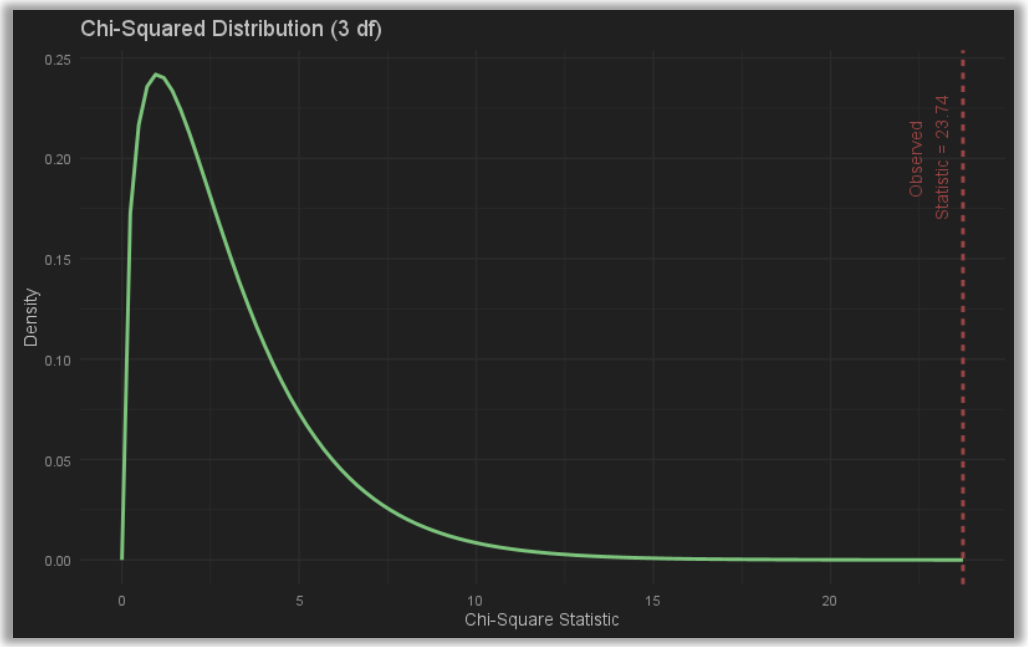


FIGURE 35

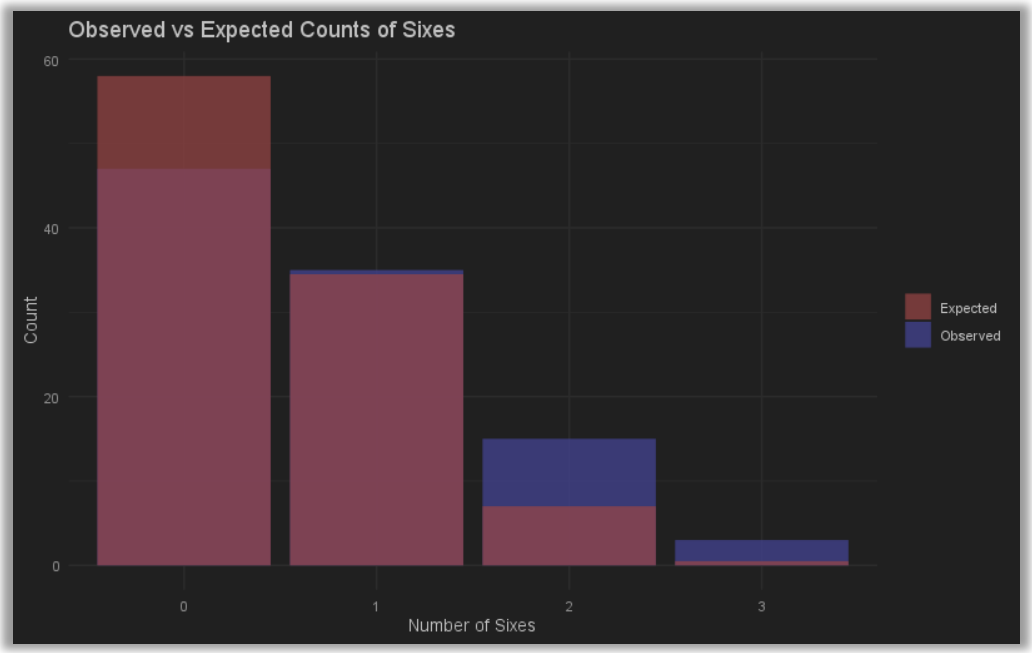


FIGURE 36

Using ggplot2 in R, I created the two graphs. From Figure 35 we see how the observed statistic deviates significantly from the mean of the chi-squared distribution with 3 degrees of freedom. We can say that the data from the observed dice throws "failed" the goodness of fit test, suggesting that the dice is unfairly weighted. Figure 36 shows how the observed number of sixes, 2 and 3 were much higher in occurrence than expected.

Exercises

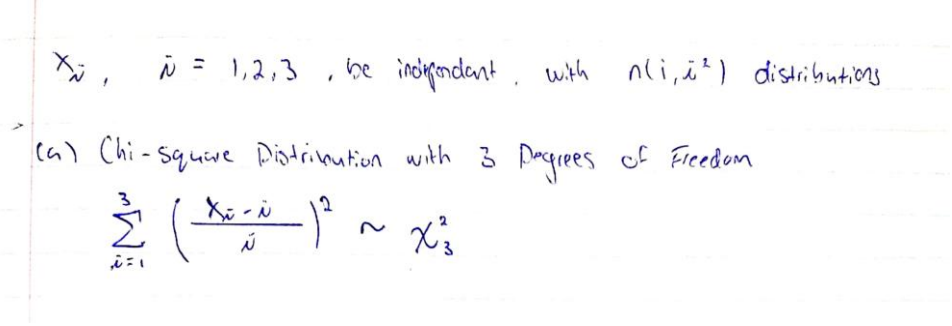
5.16 Let $X_i, i = 1, 2, 3$, be independent with $\mathrm{n}(i, i^2)$ distributions. For each of the following situations, use the X_is to construct a statistic with the indicated distribution.

(a) chi squared with 3 degrees of freedom

(b) t distribution with 2 degrees of freedom

(c) F distribution with 1 and 2 degrees of freedom

FIGURE 37 (GEORGE CASELLA, 2001, P. 258)

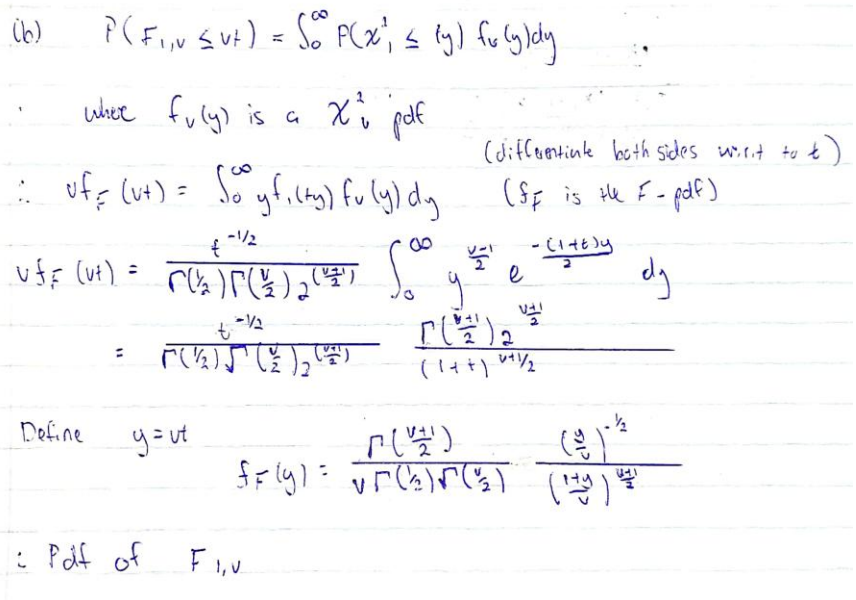


(b) A similar formula holds for the F distribution; that is, it can be written as a mixture of chi squareds. If $F_{1,\nu}$ is an F random variable with 1 and ν degrees of freedom, then we can write

$$P(F_{1,\nu} \le \nu t) = \int_0^\infty P\left(\chi_1^2 \le ty\right) f_\nu(y)\, dy,$$

where $f_\nu(y)$ is a χ^2_ν pdf. Use the Fundamental Theorem of Calculus to obtain an integral expression for the pdf of $F_{1,\nu}$, and show that the integral equals the pdf.

FIGURE 38 (GEORGE CASELLA, 2001, P. 259)



This interesting question shows the relationship between the chi-squared distribution and the F distribution The F distribution can be written as a combination of chi-squares. The gamma was once again useful in solving the integral.

Reflection

I felt a genuine sense of excitement when I learnt a new probability distribution. I felt the same way in STSM1614, when I was first introduced to distribution theory. It was then when I developed the interest and curiosity in statistics that I have now. It is exciting to know that I have still more to learn about probability distributions.

At first, I did not truly grasp what degrees of freedom where, but after some research I do. After grasping degrees of freedom, I found it much easier to understand and explain the Chi-Squared distribution. Furthermore, understanding its relationship to the standard normal distribution was key to understanding the distribution.

I had lots of fun researching and doing the goodness of fit test. It was a good opportunity to practise graphing in R using GGplot2, which we recently learned in STSM2634.

In this section, I did not include any exercises from Rice, however many of the Rice questions in the t-distribution and F-distribution section uses the Chi-Squared distribution in some manner. I found some fun and interesting exercises from another textbook.

T Distribution

Research Process

- I read the Wikipedia article on the t distribution, where I leaned that it is an generalisation of the normal distribution, controlled by the single parameter n, and has heavier tails than the normal distribution.
- I consulted my (Rice, 2007) textbook for the definition of, and related proofs of the t distribution.
- I derived the proof from *The Book of Statistical Proofs* and watched the YouTube video by *Computation Empire* which explained the proof in detail. However, I learned in class that adding the proof would be redundant, so I decided to remove it from my portfolio.
- I watched a YouTube video by Very Normal, titled " Why so we teach everyone the t test", which explained the t test, its relevance and prevalence.
- I watched a YouTube video by Very Normal, titled "The Essential Guide to Hypothesis Testing", which explained everything I needed to know about hypothesis testing, to do t-testing.
- I watched the full YouTube series by Very Normal, with videos about the one-sample t-test, two-sample t-test and How to do a t-test in R.
- I read an article on JMP Statistical Discovery on the t test, it explained the difference between the one-sample, two-sample and paired t test.
- I read the Wikipedia article on the t-test, and subsequently the Wikipedia on the Welch's t-test, in preparation for my t-test example.
- I downloaded a dataset from kaggle, about student alcohol consumption, which I used in my assignment 2.

Definition

DEFINITION

If $Z \sim N(0, 1)$ and $U \sim \chi^2_n$ and Z and U are independent, then the distribution of $Z/\sqrt{U/n}$ is called the t **distribution** with n degrees of freedom. ■

FIGURE 39 (RICE, 2007, P. 193)

PROPOSITION A

The density function of the t distribution with n degrees of freedom is

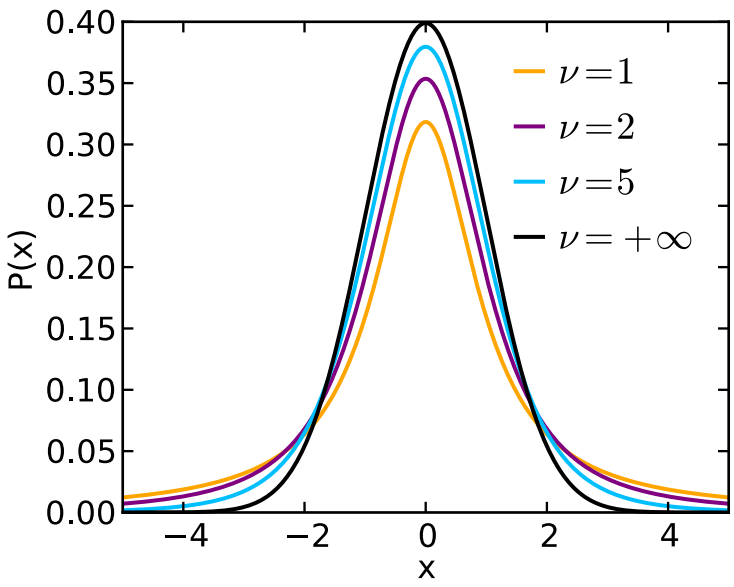
$$f(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

Proof

This is proved by a standard method. The density function of $\sqrt{U/n}$ is straightforward to obtain, and the density function of the quotient of two independent random variables was derived in Section 3.6.1. The details of the proof are left as an end-of-chapter problem. ■

FIGURE 40 (RICE, 2007, P. 193)

From the density function of Proposition A, $f(t) = f(-t)$, so the t −distribution is symmetric about zero. As the number of degrees of freedom approaches infinity, the t distribution tends to the standard normal distribution, for more than 20 or 30 degrees of freedom, the distributions are very close.



Here is the probability density function of the t-distribution, with various degrees of freedom. (Wikipedia)

My Understanding of the t-distribution

The t-distribution is a continuous probability distribution, controlled by a single parameter n, where n is the degrees of freedom. The t-distribution is used for estimating the mean of a normally distributed population when the sample size is small and the population standard deviation is unknown.

The distribution is shaped like the normal distribution, symmetric and bell shaped, but with thicker tails. As $n \to \infty$, the t-distribution converges to a standard normal distribution by central limit theorem.

Mean = 0 $(for\ n \geq 1)$

Variance $= \frac{n}{n-2}$ $(for\ n > 2; undefined\ for\ n \leq 2)$

The t-distribution is used in the t-test, which test if the difference between the response of two groups is statistically significant or not.

The T-Test

Student's *t*-test is a statistical test used to test whether the difference between the response of two groups is statistically significant or not. It is any in which the test statistic follows a student's t-distribution under the null hypothesis. (Wikipedia). The t-test is a way to determine if the means of two groups are statistically significant, and not just by chance.

One-Sample t-test

This test compares the mean of a single group to a known or hypothesized population mean, to test whether the mean differs significantly.

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Where:
$\bar{x}$ = sample mean
s = sample standard deviation
n = sample size
Degrees of freedom = $n - 1$
Sample means is assumed to be normal

Two-sample t-test (Independent t-test)

This test compares the means of two distinct groups, assuming the observations made are independent. If the variances of the populations are unequal, Welch's t-test is used, which is further discussed in assignment 2.

Paired t-test (Dependent samples t-test)

This test compares the means from the same group at different times or under different conditions, where observations are paired. This test looks at the mean difference between the paired measurements.

Assignment 2

Problem

I found a dataset on kaggle with different metrics and data on 395 high-school students, which I converted to an excel file. The data included grades, sex, relationship status, whether they receive educational support and even their workday and weekend alcohol consumption.

I wanted to test and see whether educational support, will lead to an increase in the final grade of the students. A two-sample t test would be appropriate, I formed two groups, students with educational support, and students without educational support (variable schoolsup: yes/no from dataset). I need to compare the mean final grade between the two groups. (variable G3 from dataset)

Why is it appropriate to perform a t-test? Since the groups (schoolsup = yes and schoolsup = no) are mutually exclusive, the groups are independent of each other.

Firstly, I need to define the Hypothesis:

Null Hypothesis (H_0): The mean final grades of students with and without extra educational support are equal ($\mu_1 = \mu_2$).

Alternative Hypothesis (H_1): The mean final grades of students with and without extra educational support are different ($\mu_1 \neq \mu_2$).

Secondly, I need to determine whether the equal variance t test, or Welch's t-test would be more appropriate. If the variances are equal, I will use the equal variance t-test / pooled-variance t-test, I can determine this by using a F-test, discussed in the next section. If the p-value is large, $p > 0.05$, I will use the equal variance test. If the variances are not equal, and the p-value of the F-test is small, $p < 0.05$, I must use the Welch's t-test.

The Welch t-test differs from the standard t-test, and is given by:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{s^2_{\overline{X_1}} + s^2_{\overline{X_2}}}} \quad , \quad s_{\overline{X}} = \frac{s_i}{\sqrt{N_i}}$$

Where $\overline{X_i}$ and $s_{\overline{X_i}}$ are the i^{th} sample mean and its standard error, with s_i denoting the corrected sample standard deviation, and sample size N_i. Unlike the Student's t-test, the denominator is not based on the pooled variance estimate. (Wikipedia)

Since the date sample is too large to conduct the test by hand, I will be using R to solve this problem, I can simply differentiate between the Welch and pooled variance t-test, by setting var.equal = FALSE, in the t.test() function, for the Wech t-test and setting the var.equal = TRUE, for the pooled variance t-test.

Chi-Square goodness of fit.R | t test.R | data

```r
library(readxl)
data <- read_excel( path: "C:/Users/franc/Downloads/student-mat fixed.xlsx")
df <- data.frame(data)

group_yes <- df$G3[df$schoolsup == "yes"]
group_no <- df$G3[df$schoolsup == "no"]

var_yes <- var(group_yes)
var_no <- var(group_no)

# Perform variance test
var_test_p <- var.test(group_yes, group_no)$p.value

# Decide which t-test to use
if (var_test_p < 0.05) {
  t_test_result <- t.test(G3 ~ schoolsup, data = df, var.equal = FALSE)  # Welch's t-test
} else {
  t_test_result <- t.test(G3 ~ schoolsup, data = df, var.equal = TRUE)   # Equal variance t-test
}

print(t_test_result)
```

From the imported excel data, I extracted the groups with and without academic support. I used a variance or F-test on the two groups, from which I extracted the p-value. If the p-value is small, P < 0.05, the if statement will perform a Welch's t-test, suppose this was not the case, or I used another dataset and find P > 0.05 the statement would perform a pooled variance t-test

```
    Welch Two Sample t-test

data:  G3 by schoolsup
t = 2.3705, df = 97.126, p-value = 0.01974
alternative hypothesis: true difference in means between group no and group yes is
 not equal to 0
95 percent confidence interval:
 0.1838414 2.0755065
sample estimates:
 mean in group no mean in group yes
        10.561047          9.431373
```

Interpretation of Results:

T statistic = 2.3705, thus the mean of the group with no educational support is higher. The larger the absolute value of the t-statistic the stronger the case against the null hypothesis.

P-Value = 0.01974, the p-value indicates the probability of observing a t-statistic as extreme as extreme as 2.3705 under the null hypothesis. Since 0.01974 is smaller than the common 0.05 significance level, I can reject the null hypothesis at the 5% significance level. This means that

there is statistically significant evidence that the mean final grades differ between the group of students with and without extra educational support

Conclusion:

Students without extra educational support, receive a higher mean final grade (10.56) versus students with extra educational support, who received a mean final grade of (9.43). The difference of 1.13 points is statistically significant.

One might have suspected that students with extra educational support would receive higher grades, however extra support is often provided to poor performing students, who already have worse grades due to other factors, like absences, learning difficulties or external challenges.

Based on the feedback from the assignment 2 draft, I decided to show all the mathematical steps. Because my sample size is so large, I calculated the mean and variance in excel:

Group	Sample Size n	Mean $\bar{x}$	Variance s^2
schoolsup = yes	$n_1 = 39$	$\overline{x_1} = 10.28$	$s_1^2 = 15.34$
schoolsup = no	$n_2 = 151$	$\overline{x_2} = 12.06$	$s_2^2 = 12.63$

Substituting these values into Welch's t-test

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

$$t = \frac{10.28 - 12.06}{\sqrt{\frac{15.34}{39} + \frac{12.63}{151}}}$$

$$t = -2.577$$

Compute the degrees of freedom (Welch-Satterthwaite Equation)

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

$$df = \frac{\left(\frac{15.34}{39} + \frac{12.63}{151}\right)^2}{\frac{\left(\frac{15.34}{39}\right)^2}{39 - 1} + \frac{\left(\frac{12.63}{151}\right)^2}{151 - 1}}$$

$$df = 55{,}25 \approx 55$$

Since this is a two tailed test, $P(|T| > 2.577)$

Using a p-value calculator, I found $p = 0.01267$. Since the significance level is $\alpha = 0.05$, I will reject the null hypothesis.

The t statistic and p-value does differ a bit between the manual test and coded test, since there were 395 samples, some rounding or handling mistakes could easily be made. I will consider the R test as more accurate. Nonetheless, both tests will reject the null hypothesis.

Why I found this problem interesting?

I found this problem interesting and fun, since I did not simply find this problem online. I had to find a sufficiently large dataset, formulate an appropriate hypothesis test, determine whether the variables are independent and if a t-test would be appropriate. Then I must determine whether to use a pooled variance test or Welch t-test. Then finally I can do the t-test and interpretate the results.

The conclusion from this problem was interesting, I thought that students with academic support would have done better than those without. However, this was not the case, as explained in my conclusion.

What I learned from this Problem?

Since I formulated and solved the problem myself, it was a great exercise to improve my problem-solving skills. At first, I thought it would be just a straightforward two-sample t-test, however, I had to learn and use the F-test to determine if the variances between the groups are the same or different. Secondly, I learned I needed to use the Welch t-test; I had to do some research about it. I learned how to implement both t-test and F-tests in R. I had to learn how to interoperate the t and p-values of the t-test, to derive a conclusion.

This problem challenged much more domains than I anticipated, hypothesis testing, F-tests, t-tests, variations of the t-test and statistical programming.

Exercises

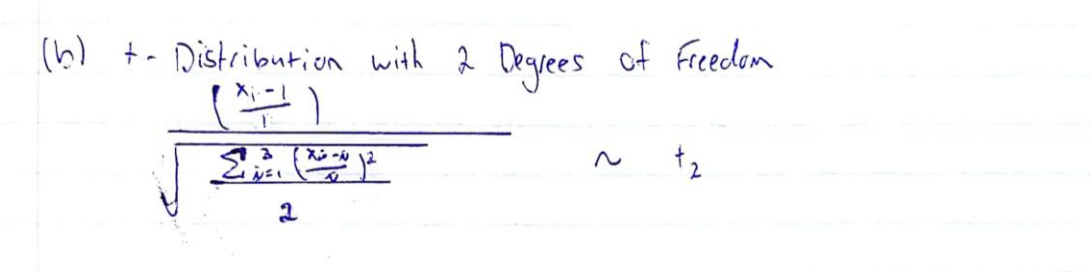
5.16 Let $X_i, i = 1, 2, 3$, be independent with $\mathrm{n}(i, i^2)$ distributions. For each of the following situations, use the X_is to construct a statistic with the indicated distribution.

(a) chi squared with 3 degrees of freedom

(b) t distribution with 2 degrees of freedom

(c) F distribution with 1 and 2 degrees of freedom

FIGURE 41 (GEORGE CASELLA, 2001, P. 258)



5.20 (a) We can see that the t distribution is a mixture of normals using the following argument:

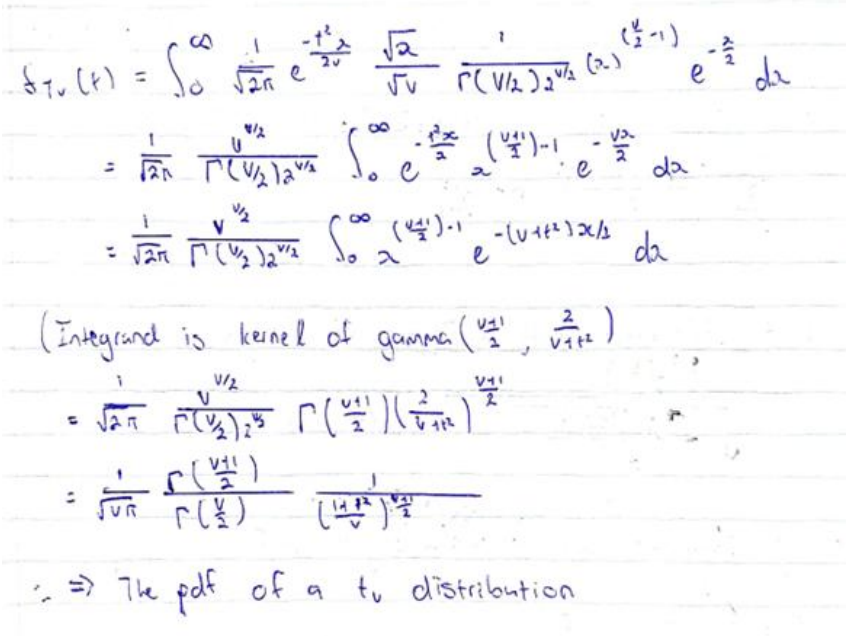
$$P(T_\nu \le t) = P\left(\frac{Z}{\sqrt{\chi^2_\nu/\nu}} \le t\right) = \int_0^\infty P\left(Z \le t\sqrt{x}/\sqrt{\nu}\right) P\left(\chi^2_\nu = x\right)\, dx,$$

where T_ν is a t random variable with ν degrees of freedom. Using the Fundamental Theorem of Calculus and interpreting $P(\chi^2_\nu = \nu x)$ as a pdf, we obtain

$$f_{T_\nu}(t) = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2 x/2\nu} \frac{\sqrt{x}}{\sqrt{\nu}} \frac{1}{\Gamma(\nu/2)2^{\nu/2}} (x)^{(\nu/2)-1} e^{-x/2}\, dx,$$

a scale mixture of normals. Verify this formula by direct integration.

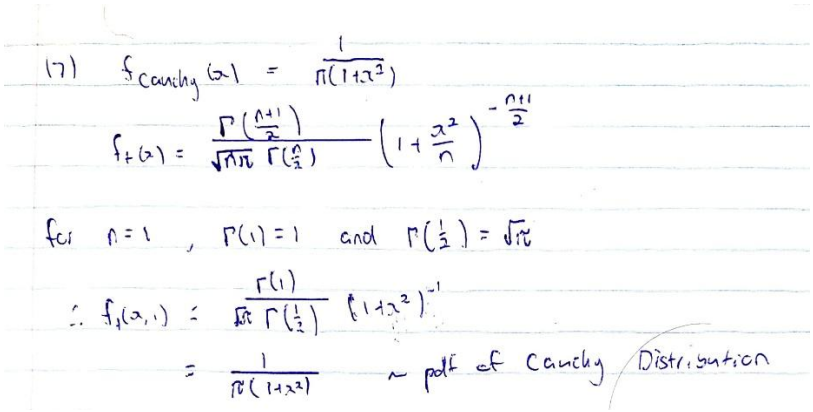
FIGURE 42 (GEORGE CASELLA, 2001, P. 259)



What an interesting question, showing the relationship between the normal distribution and the t-distribution. Once again, the gamma was a lifesaver, making the otherwise improper integral easily solvable.

7. Show that the Cauchy distribution and the t distribution with 1 degree of freedom are the same.

FIGURE 43 (RICE, 2007, P. 198)



Reflection

I spent a lot of time doing research on this section. In the process, I discovered a brilliant and relatively new YouTube channel: Very Normal.

I have a good understanding of the student's t-distribution, its relationship to the normal distribution and its appropriate use. I understand and can explain when it would be appropriate to use the t-distribution versus the normal distribution.

I had lots of fun doing t-test on an interesting dataset I found. I planned to do a standard two-sample t-test, but the variances of my two samples turned out not equal. I learned that a Welch's t-test would be appropriate and spent some time doing research on it and using it on assignment 2. I learned a lot more in this section and in my assignment than I thought I would have.

Again, most of my exercises were from another textbook (George Casella, 2001), however Rice question 6 also includes the t-distribution, question 6 can be found in the F-distribution section.

I did attempt the proof for the t-distribution and spent some time explaining it, however I learned in class that it would be redundant and removed it from my portfolio. However, I do not regret spending time on the proof, since it helped me understand the distribution much better.

F-Distribution

Research Process

- I read the Wikipedia article on the F-Distribution, where I learned that there exist a Central F-Distribution and a Noncentral F-Distribution. I will be looking at the Central F-Distribution.
- I consulted my (Rice, 2007) textbook for the properties and definition of the F-Distribution.
- I consulted my (George Casella, 2001) textbook, where I found the best explanation of the F distribution.
- I watched a YouTube video by Very Normal, explaining ANOVA and the F Test, he briefly showed how to use ANOVA in R.
- I prompted Grok3 to define and explain both ANOVA and the F-test.
- I read an article on datacamp, which explained the one-way ANOVA test, with examples.

Definition

DEFINITION

Let U and V be independent chi-square random variables with m and n degrees of freedom, respectively. The distribution of

$$W = \frac{U/m}{V/n}$$

is called the F distribution with m and n degrees of freedom and is denoted by $F_{m,n}$. ■

FIGURE 44 (RICE, 2007, P. 194)

PROPOSITION B

The density function of W is given by

$$f(w) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n}w\right)^{-(m+n)/2}, \qquad w \geq 0$$

Proof

W is the ratio of two independent random variables, and its density follows from the results given in Section 3.6.1. ■

FIGURE 45 (RICE, 2007, P. 194)

Example 5.3.5 (Variance ratio distribution) Let $X_1, \ldots, X_n$ be a random sample from a $n(\mu_X, \sigma_X^2)$ population, and let $Y_1, \ldots, Y_m$ be a random sample from an independent $n(\mu_Y, \sigma_Y^2)$ population. If we were interested in comparing the variability of the populations, one quantity of interest would be the ratio σ_X^2/σ_Y^2. Information about this ratio is contained in S_X^2/S_Y^2, the ratio of sample variances. The F distribution allows us to compare these quantities by giving us a distribution of

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}. \tag{5.3.8}$$

Examination of (5.3.8) shows us how the F distribution is derived. The ratios S_X^2/σ_X^2 and S_Y^2/σ_Y^2 are each scaled chi squared variates, and they are independent. ‖

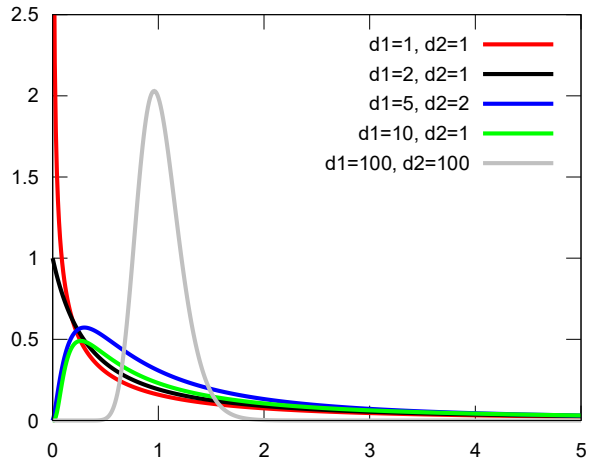
Definition 5.3.6 Let $X_1, \ldots, X_n$ be a random sample from a $n(\mu_X, \sigma_X^2)$ population, and let $Y_1, \ldots, Y_m$ be a random sample from an independent $n(\mu_Y, \sigma_Y^2)$ population. The random variable $F = (S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ has *Snedecor's F distribution with $n-1$ and $m-1$ degrees of freedom*. Equivalently, the random variable F has the F distribution with p and q degrees of freedom if it has pdf

$$f_F(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2}\frac{x^{(p/2)-1}}{[1+(p/q)x]^{(p+q)/2}}, \quad 0 < x < \infty. \tag{5.3.9}$$

The F distribution can be derived in a more general setting than is done here. A variance ratio may have an F distribution even if the parent populations are not normal. Kelker (1970) has shown that as long as the parent populations have a certain type of symmetry (*spherical symmetry*), then the variance ratio will have an F distribution.

FIGURE 46 (GEORGE CASELLA, 2001, P. 224)

(George Casella, 2001) *did a much better job in explaining the distribution than* (Rice, 2007). *The F Distribution arises when you take the ratio of two variances (two independent Chi-Square variables divided by their degrees of freedom). In practise, this comes up in the F-test, later discussed in this section.*



Here is the pdf of the F-distribution with various degrees of freedom. (Wikipedia)

My Understanding of the F distribution

The F distribution is a continuous probability distribution, arising from the ratio of two independent chi-squared distributions, each divided by their degrees of freedom, m and n.

Those degrees of freedom m and n, are the parameters that control the shape of the distribution. The distribution is right-skewed and non-zero.

For $n > 2$,the mean of an F-distributed random variable: $E(w) = \frac{n}{n-2}$

For $n > 4$, the variance of an F-distributed random variable: $Var(w) = \frac{2(2n-2)}{(n-2)^2(n-4)}$

The F-distribution is an important component of the ANOVA test (analysis of variance), used to test if multiple population means are equal. The F-test is used to compare the variances of two populations.

Analysis of Variance (ANOVA) and the F-test

(I asked grok3 to define the ANOVA and F-Test, I compared its response to its sources which included, datacamp and Wikipedia, and I found no inaccuracies)

ANOVA

ANOVA is a statistical method used to test whether there are significant differences between the means of three or more independent groups on a continuous dependant variable. It generalises the t-test to multiple groups by analysing the variance within the and between the groups to determine if the observed differences in means are statistically significant.

The F-Test

The F-test is a statistical test used to compare variances, and in the context of ANOVA, it tests whether the between-group variance is significantly larger than the within-group variance, indicating that the group means are not all equal.

Null Hypothesis (H_0): All group means are equal $\mu_1 = \mu_2 = \cdots = \mu_k$

Alternative Hypothesis (H_1): At least one group mean is different.

F Statistic: The F statistic is the ratio of between-group variation to within group variation, a higher F value suggest greater differences between groups means relative to random variation. For a one-way ANOVA, the F statistic is defined as:

$$F = \frac{Between\ group\ variance}{within\ group\ variance}$$

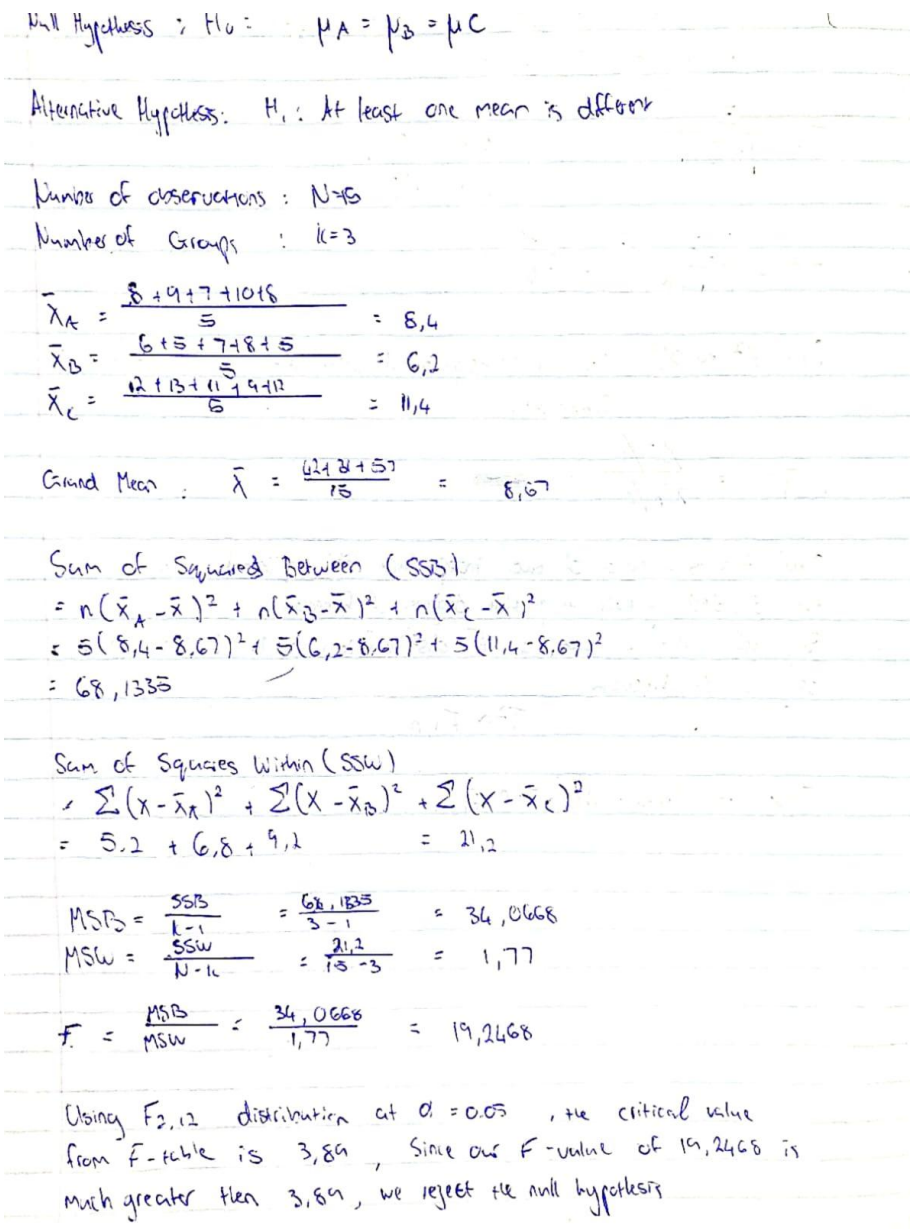
$$F = \frac{Mean\ square\ sum\ between\ groups}{Mean\ square\ sum\ within\ groups}$$

P-value: The p-value determines it the difference between the group means are statistically significant. If the p-value is lower than a predetermined threshold, commonly set at 0.05, the null hypothesis is rejected, and it can be concluded that at least one group has a significantly different mean.

Simple Example:

Suppose financial analyst wish to compare the mean returns of three different mutual funds (A, B and C). Each fund has 5 randomly selected returns in %.

Fund	Returns (%)
A	8, 9, 7, 10, 8
B	6, 5, 7, 8, 5
C	12, 13, 11, 9, 12



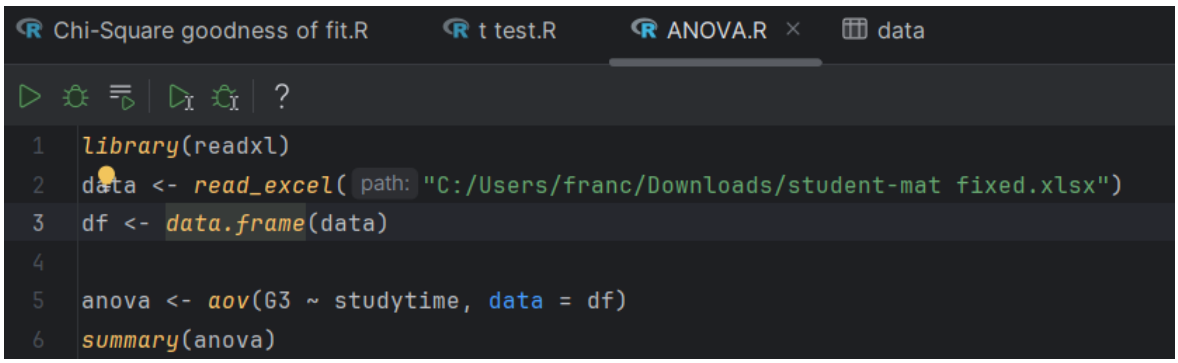
It was good to attempt this problem mathematically and by hand, as to learn each of the steps involved in the ANOVA test. However, even with only 15 observations, the sums became tedious to do with hand and cumbersome to type into a calculator. Luckily ANOVA test could easily be done on large samples, by using R.

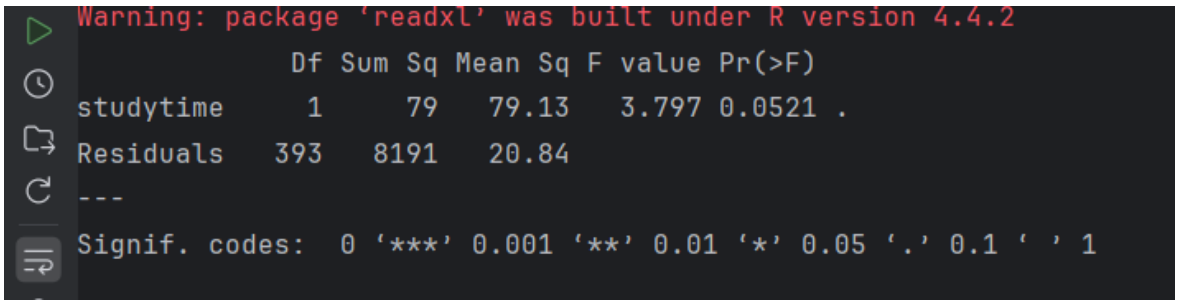
Coded Example:

Using the same dataset used for the t-test. I wanted to test and see whether study time levels had a significant impact on final grades. There were four groups in the studytime category, denoted 1 to 4, with 4 being the most amount of time spent studying. I am unsure what the unit measurement of study time is, but nonetheless, we can do a one-way ANOVA.

We define the null hypothesis: There are no significant difference in mean final grades (G3) between different study time levels.

We define the alternative hypothesis: At least one study time category has a significantly different mean grade from the other.





Interpretation and Conclusion:

The F-value is 0.0521, which is slightly above 0.05, however, it is not quite statistically significant at the 5% level. There may be a very weak effect.

At 5% significance level, we fail to reject the null hypothesis, Study time did not have a significant on the student's final grade.

Of course, I know that study time is not a perfect measure of effective study, But I did expect to see a higher F-value.

Exercises

6. Show that if $T \sim t_n$, then $T^2 \sim F_{1,n}$.

FIGURE 47 (RICE, 2007, P. 198)

$T \sim t_n$: A random variable T, follows a t-distribution, with n degrees of freedom

$$T = \frac{Z}{\sqrt{V/n}} \qquad Z \sim N(0,1), \quad V \sim \chi^2_n$$

$$T^2 = \left(\frac{Z}{\sqrt{V/n}}\right)^2 = \frac{Z^2}{V/n}$$

Since $Z^2 \sim \chi^2_1$ is a chi-squared distribution with 1 degree of freedom.

$$T^2 = \frac{\chi^2_1/1}{\chi^2_n/n}$$

Since this is a ratio of two independent chi-squared variables, each divided by their respective degrees of freedom.
By def, this follows an F-distribution with 1 and n degrees of freedom

$$T^2 \sim F_{1,n}$$

Thus if $T \sim t_n$, then $T^2 \sim F_{1,n}$

5.16 Let $X_i, i = 1, 2, 3$, be independent with $n(i, i^2)$ distributions. For each of the following situations, use the X_is to construct a statistic with the indicated distribution.

(a) chi squared with 3 degrees of freedom
(b) t distribution with 2 degrees of freedom
(c) F distribution with 1 and 2 degrees of freedom

FIGURE 48 (GEORGE CASELLA, 2001, P. 258)

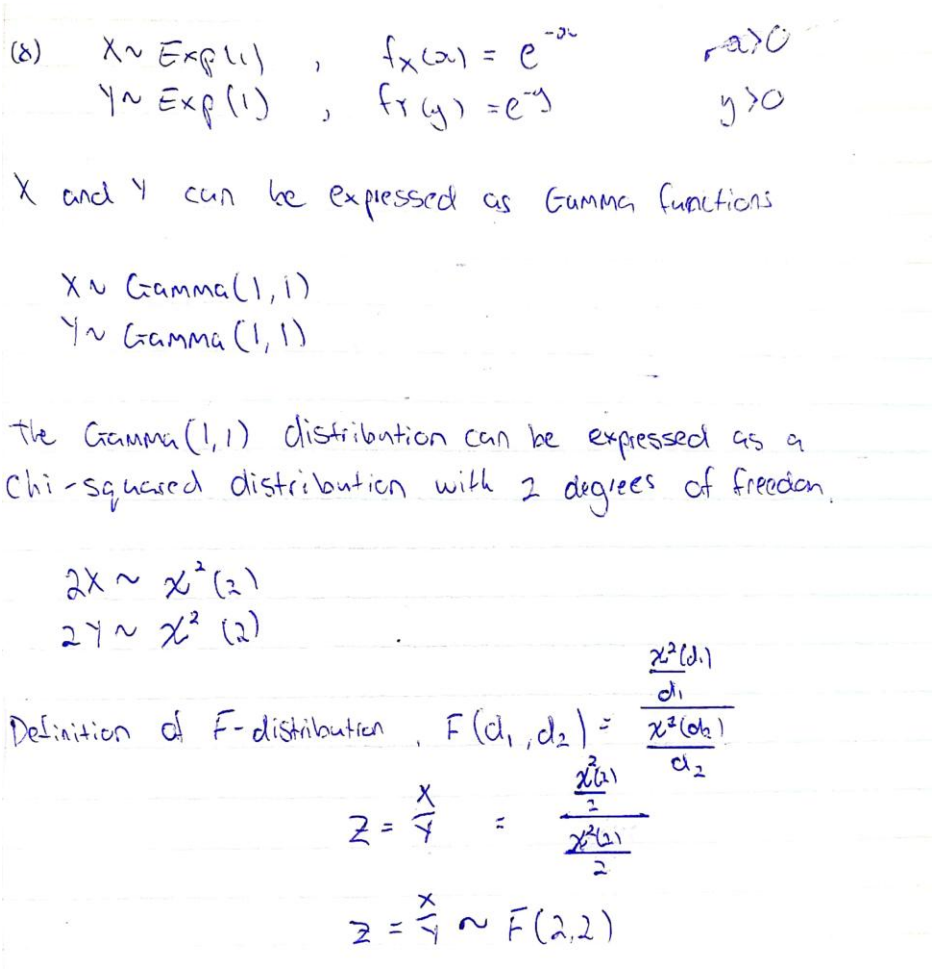
(c) F-Distribution with 1 and 2 degrees of freedom

$$F = \frac{\chi^2_1/1}{\chi^2_2/2} \sim F(1,2)$$

$$F = \frac{\left(\frac{X_1 - 1}{1}\right)^2/1}{\left[\sum_{i=2}^{3}\left(\frac{X_i - i}{i}\right)^2\right]/2} \sim F(1,2)$$

8. Show that if X and Y are independent exponential random variables with $\lambda = 1$, then X/Y follows an F distribution. Also, identify the degrees of freedom.

FIGURE 49 (RICE, 2007, P. 198)



It was cool going from an exponential distribution to a gamma distribution, to a chi-squared distribution, to a F-distribution. Once again, the gamma distribution seems to be the key to all my problems.

Reflection

While I was explaining the F-distribution to my girlfriend, she asked if I discovered it (my initial is F), to which I proclaimed yes. I then spent some 10 minutes impersonating Ronald Fisher, explaining my discovery to her. This joke showed me that I so understand the distribution and could explain it, but it also showed that I used the F test and ANOVA test interchangeably, the ANOVA is a method for comparing group means, while the F-test is a statistical test comparing variances.

Sample Mean and Sample Variance

Research Process

- I consulted my (Rice, 2007) textbook for all the relevant theorems and proofs.
- I read the Wikipedia article on the sample mean, where I learned that the sample mean can be expressed as a vector or matrix.
- I read sample variance article on StatLect, which explained that the sample variance is firstly the variance of the sampled observations, but secondly, an estimator of the variance of the population from which the observations have been drawn.

Definition

Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

$\bar{X}$ and S^2 are the sample mean and sample variance respectively. $\bar{X}$ is a linear combination of independent normal random variables, it is normally distributed with

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

THEOREM A

The random variable $\overline{X}$ and the vector of random variables $(X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X})$ are independent.

Proof

At the level of this course, it is difficult to give a proof that provides sufficient insight into why this result is true; a rigorous proof essentially depends on geometric properties of the multivariate normal distribution, which this book does not cover. We present a proof based on moment-generating functions; in particular, we will show that the joint moment-generating function

$$M(s, t_1, \ldots, t_n) = E\{\exp[s\overline{X} + t_1(X_1 - \overline{X}) + \cdots + t_n(X_n - \overline{X})]\}$$

factors into the product of two moment-generating functions—one of $\overline{X}$ and the other of $(X_1 - \overline{X}), \ldots, (X_n - \overline{X})$. The factoring implies (Section 4.5) that the random variables are independent of each other and is accomplished through some algebraic trickery. First we observe that since

$$\sum_{i=1}^{n} t_i(X_i - \overline{X}) = \sum_{i=1}^{n} t_i X_i - n\overline{X}\,\bar{t}$$

then

$$s\overline{X} + \sum_{i=1}^{n} t_i(X_i - \overline{X}) = \sum_{i=1}^{n} \left[\frac{s}{n} + (t_i - \bar{t})\right] X_i$$

$$= \sum_{i=1}^{n} a_i X_i$$

where

$$a_i = \frac{s}{n} + (t_i - \bar{t})$$

Furthermore, we observe that

$$\sum_{i=1}^{n} a_i = s$$

$$\sum_{i=1}^{n} a_i^2 = \frac{s^2}{n} + \sum_{i=1}^{n} (t_i - \bar{t})^2$$

Now we have

$$M(s, t_1, \ldots, t_n) = M_{X_1 \cdots X_n}(a_1, \ldots, a_n)$$

and since the X_i are independent normal random variables, we have

$$M(s, t_1, \ldots, t_n) = \prod_{i=1}^{n} M_{X_i}(a_i)$$

$$= \prod_{i=1}^{n} \exp\left(\mu a_i + \frac{\sigma^2}{2} a_i^2\right)$$

$$= \exp\left(\mu \sum_{i=1}^{n} a_i + \frac{\sigma^2}{2} \sum_{i=1}^{n} a_i^2\right)$$

$$= \exp\left[\mu s + \frac{\sigma^2}{2}\left(\frac{s^2}{n}\right) + \frac{\sigma^2}{2} \sum_{i=1}^{n} (t_i - \bar{t})^2\right]$$

$$= \exp\left(\mu s + \frac{\sigma^2}{2n} s^2\right) \exp\left[\frac{\sigma^2}{2} \sum_{i=1}^{n} (t_i - \bar{t})^2\right]$$

The first factor is the mgf of $\overline{X}$. Since the mgf of the vector $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$ can be obtained by setting $s = 0$ in M, the second factor is this mgf. ■

FIGURE 50 (RICE, 2007, P. 196)

Interpretation and Explanation of Theorem A

This might be the first time I was unable to find an easier proof online, than the proof found in Rice. Most other proofs use transformations, which I will soon learn in mathematics. However, for now using moment generating functions is the best way for me to understand.

The proof needs to show that $\bar{X}$ and the vector of deviation $(X_1 - \bar{X}, \dots, X_n - \bar{X})$ are independent. The strategy used in Rice, is to show that their joint mgf factors into the product of their individual mgfs, which implies independence, from the properties of moment generating functions.

By rewriting the exponent, the joint mgf is:

$$M(s, t_1, \dots, t_n) = E\left[\exp\left(s\bar{X} + \sum_{i=1}^{n} t_i(X_i - \bar{X})\right)\right]$$

The key step is rewriting the exponent as a linear combination of the $X_i's$:

$$\sum_{i=1}^{n}(\frac{s}{n} + (t_i - \bar{t}))X_i, where\ \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

This simplifies the problem to working with $\sum a_i X_i, where\ a_i = \frac{s}{n} + (t_i - \bar{t})$

Since the X_i are identically independent normal, their joint mgf is the product of individual mgf's:

$$M(s, t_1, \dots, t_n) = \prod_{i=1}^{n} \exp\left(\mu a_i + \frac{\sigma^2}{2}a_i^2\right) = \exp\left(\mu \sum a_i + \frac{\sigma^2}{2}\sum a_i^2\right)$$

Substituting the properties of a_i:

$$M = \exp\left(\mu s + \frac{\sigma^2}{2n}s^2\right) \cdot \exp\frac{\sigma^2}{2}\sum(t_i - \bar{t})^2$$

This factors into

$$\exp\left(\mu s + \frac{\sigma^2}{2n}s^2\right) \cdot \exp\left(\frac{\sigma^2}{2}(t_i - \bar{t})^2\right)$$

The factorization confirms that $\bar{X}$ and $(X_1 - \bar{X}, \dots, X_n - \bar{X})$ are independent, since their joint mgf is the product of their marginal mgf's. The deviations' mgf does not depend on s, and $\bar{X}$'s mgf does not depend on t_i , showing their independence.

COROLLARY A

$\overline{X}$ and S^2 are independently distributed.

Proof

This follows immediately since S^2 is a function of the vector $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$, which is independent of $\overline{X}$. ■

FIGURE 51 (RICE, 2007, P. 197)

THEOREM B

The distribution of $(n-1)S^2/\sigma^2$ is the chi-square distribution with $n-1$ degrees of freedom.

Proof

We first note that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi_n^2$$

Also,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}[(X_i-\overline{X})+(\overline{X}-\mu)]^2$$

Expanding the square and using the fact that $\sum_{i=1}^{n}(X_i-\overline{X})=0$, we obtain

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2+\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

This is a relation of the form $W = U + V$. Since U and V are independent by Corollary A, $M_W(t) = M_U(t)M_V(t)$. W and V both follow chi-square distributions, so

$$\begin{aligned}M_U(t)&=\frac{M_W(t)}{M_V(t)}\\&=\frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}}\\&=(1-2t)^{-(n-1)/2}\end{aligned}$$

The last expression is the mgf of a random variable with a χ^2_{n-1} distribution. ■

FIGURE 52 (RICE, 2007, P. 197)

Interpretation and Explanation of Theorem B

Since each term $\frac{(X_i-\mu)}{\sigma}$ is a standard normal variable $N(0,1)$ and the sum of squares of n independent standard normal follows χ_n^i

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi_n^2$$

The key trick is to rewrite the sum of the squares around μ using the sample mean $\bar{X}$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]\text{^}2$$

After expanding and using the property $\sum(X_i-\bar{X})=0$ and simplifying to

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}[(X_i-\bar{X})^2+\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2=U+V$$

Using Corollary A, $\bar{X}$ and S^2 are independent, and thus:

$V=\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$ (a function of $\bar{X}$)

$U=\frac{(n-1)S^2}{\sigma^2}$ (a function of S^2)

Since $\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right),\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\sim N(0,1),so\ V\sim\chi_1^2$

The original sum $W=\frac{1}{\sigma^2}\sum(X_i-\mu)^2\sim\chi_n^2$

Since $W=U+V$ and U,V are independent:

$$M_U(t)=(1-2t)^{\frac{1-n}{2}}$$

Which is the mgf of a χ_{n-1}^2 distribution

Important note: The loss of 1 degree of freedom (from n to $n-1$) accounts for estimating μ with $\bar{X}$

COROLLARY B

Let $\overline{X}$ and S^2 be as given at the beginning of this section. Then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Proof

We simply express the given ratio in a different form:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\left(\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)}{\sqrt{S^2/\sigma^2}}$$

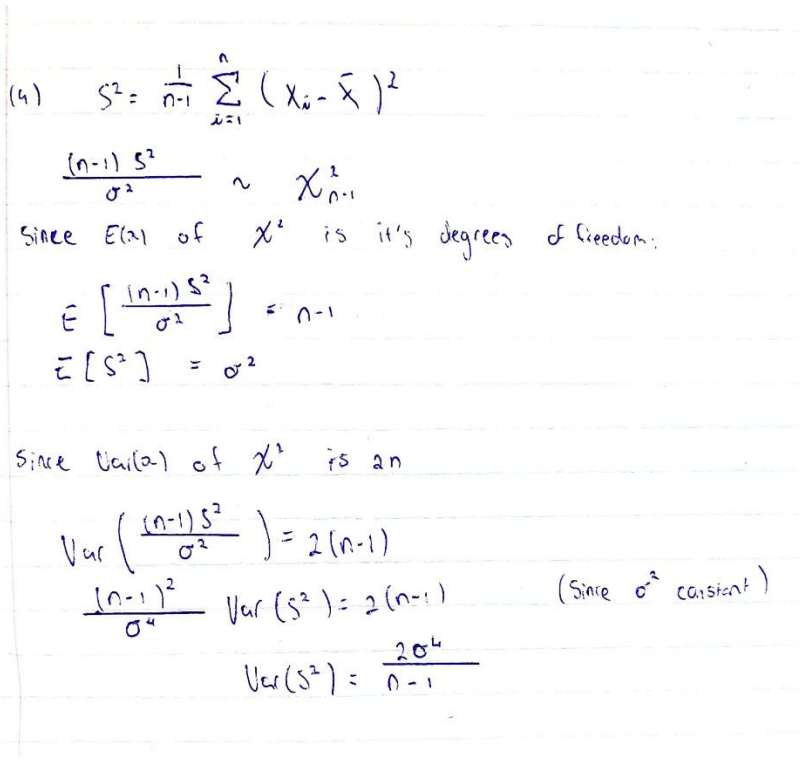
The latter is the ratio of an $N(0, 1)$ random variable to the square root of an independent random variable with a χ^2_{n-1} distribution divided by its degrees of freedom. Thus, from the definition in Section 6.2, the ratio follows a t distribution with $n - 1$ degrees of freedom. ■

FIGURE 53 (RICE, 2007, P. 198)

Exercises

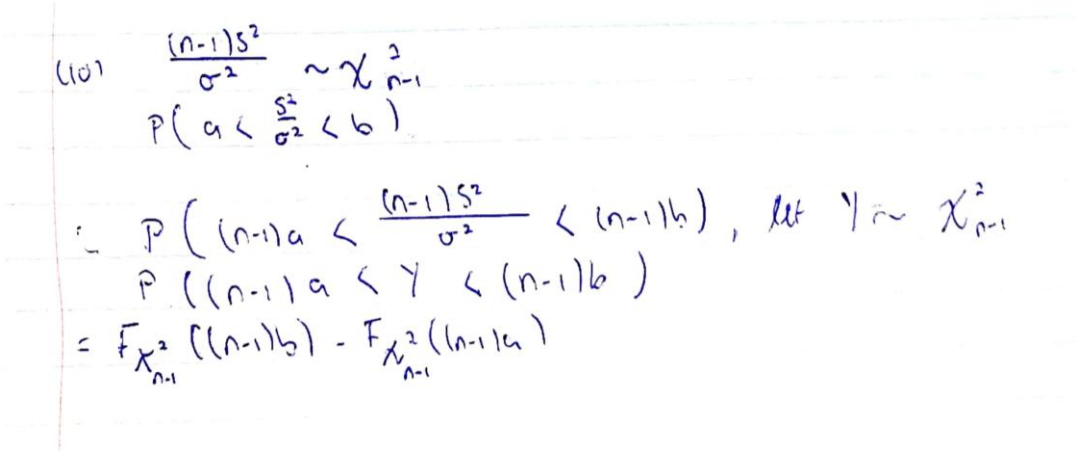
9. Find the mean and variance of S^2, where S^2 is as in Section 6.3.

FIGURE 54 (RICE, 2007, P. 198)



10. Show how to use the chi-square distribution to calculate $P(a < S^2/\sigma^2 < b)$.

FIGURE 55 (RICE, 2007, P. 198)



Reflection

At first, I completely overlooked section 6.3. Luckily, I realised my mistake, since this section turned out to be extremely important, defining the sample mean and variance and the important property that they are identically distributed.

I understand this section and understand why this section is extremely important in sampling distribution. However, I am yet to see when theorem B might be used.

The proofs from this section were rather difficult; after quickly browsing online, I decided to stick to the proofs in Rice and to dissect and explain them better.

Typing out the proofs did sharpen my skills with equation editor.

Survey Sampling

Index of examples and exercises

Simple Random Sampling

I. Chapter 7, Question 4 (Rice, 2007, p. 240)
II. Chapter 7, Question 6 (Rice, 2007, p. 240)
III. Chapter 7, Question 7 (Rice, 2007, p. 240)
IV. Chapter 7, Question 58 (Rice, 2007, p. 250) (Found in Stratified Random Sampling section, in conjunction with question 59)

The Normal Approximation to the Sampling Distribution of $\bar{X}$

I. Chapter 7, Question 8 (Rice, 2007, p. 240)
II. Chapter 7, Question 9 (Rice, 2007, p. 240)
III. Chapter 7, Question 22 (Rice, 2007, p. 242)

Stratified Random Sampling

I. Chapter 7, Question 59 (Rice, 2007, p. 250)
II. Assignment 3

Methods of Allocation

I. Chapter 7, Question 60 (Rice, 2007, p. 250)
II. Chater 7, Question 63 (Rice, 2007, p. 251)

Simple Random Sampling

Research Process

- I read an article on Investopedia on simple random sampling, which defined it as "A subset of a statistical population in which each member of the subset has an equal probability of being chosen and is meant to be an unbiased representation of a group"
- I prompted a Grok 3 deep search to help me understand simple random sampling, as well as its related properties such as Lemma A.
- I prompted ChatGPT to help define Lemma A, its comment is shown under the Lemma A comment box.
- I consulted my (Rice, 2007) textbook for the relevant theorems and definitions

LEMMA A

Denote the distinct values assumed by the population members by $\zeta_1, \zeta_2, \ldots, \zeta_m$, and denote the number of population members that have the value ζ_j by n_j, $j = 1, 2, \ldots, m$. Then X_i is a discrete random variable with probability mass function

$$P(X_i = \zeta_j) = \frac{n_j}{N}$$

Also,

$$E(X_i) = \mu$$
$$\text{Var}(X_i) = \sigma^2$$

Proof

The only possible values that X_i can assume are $\zeta_1, \zeta_2, \ldots, \zeta_m$. Since each member of the population is equally likely to be the ith member of the sample, the probability that X_i assumes the value ζ_j is thus n_j/N. The expected value of the random variable X_i is then

$$E(X_i) = \sum_{j=1}^{m} \zeta_j P(X_i = \zeta_j) = \frac{1}{N} \sum_{j=1}^{m} n_j \zeta_j = \mu$$

The last equation follows because n_j population members have the value ζ_j and the sum is thus equal to the sum of the values of all the population members. Finally,

$$\begin{aligned} \text{Var}(X_i) &= E\left(X_i^2\right) - [E(X_i)]^2 \\ &= \frac{1}{N} \sum_{j=1}^{m} n_j \zeta_j^2 - \mu^2 \\ &= \sigma^2 \end{aligned}$$

Here we have used the fact that $\sum_{i=1}^{N} x_i^2 = \sum_{j=1}^{m} n_j \zeta_j^2$ and the identity for the population variance derived in Section 7.2. ■

FIGURE 56 (RICE, 2007, P. 205)

I found it a bit difficult to explain the lemma in words, however with some help from ChatGPT 4, it can be described as follow: "Lemma A shows that when sampling from a finite population with repeated values, the random variable representing the value of any randomly chosen member follows a discrete distribution based on the frequency of each unique value". So, despite the repetition, the expected value and variance a variable still align exactly with the population mean and variance, allowing us to make inferences about the population.

LEMMA B

For simple random sampling without replacement,

$$\text{Cov}(X_i, X_j) = -\sigma^2/(N-1) \qquad \text{if } i \neq j$$

Using the identity for covariance established at the beginning of Section 4.3,

$$\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$$

and

$$\begin{aligned} E(X_i X_j) &= \sum_{k=1}^{m}\sum_{l=1}^{m} \zeta_k \zeta_l P(X_i = \zeta_k \text{ and } X_j = \zeta_l) \\ &= \sum_{k=1}^{m} \zeta_k P(X_i = \zeta_k) \sum_{l=1}^{m} \zeta_l P(X_j = \zeta_l | X_i = \zeta_k) \end{aligned}$$

from the multiplication law for conditional probability. Now,

$$P(X_j = \zeta_l | X_i = \zeta_k) = \begin{cases} n_l/(N-1), & \text{if } k \neq l \\ (n_l - 1)/(N-1), & \text{if } k = l \end{cases}$$

Now if we express

$$\begin{aligned} \sum_{l=1}^{m} \zeta_l P(X_j = \zeta_l | X_i = \zeta_k) &= \sum_{l \neq k} \zeta_l \frac{n_l}{N-1} + \zeta_k \frac{n_k - 1}{N-1} \\ &= \sum_{l=1}^{m} \zeta_l \frac{n_l}{N-1} - \zeta_k \frac{1}{N-1} \end{aligned}$$

the expression for $E(X_i X_j)$ becomes

$$\begin{aligned} \sum_{k=1}^{m} \zeta_k \frac{n_k}{N} \left(\sum_{l=1}^{m} \zeta_l \frac{n_l}{N-1} - \frac{\zeta_k}{N-1} \right) &= \frac{1}{N(N-1)} \left(\tau^2 - \sum_{k=1}^{m} \zeta_k^2 n_k \right) \\ &= \frac{\tau^2}{N(N-1)} - \frac{1}{N(N-1)} \sum_{k=1}^{m} \zeta_k^2 n_k \\ &= \frac{N\mu^2}{N-1} - \frac{1}{N-1}(\mu^2 + \sigma^2) \\ &= \mu^2 - \frac{\sigma^2}{N-1} \end{aligned}$$

Finally, subtracting $E(X_i)E(X_j) = \mu^2$ from the last equation, we have

$$\text{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

for $i \neq j$. ■

FIGURE 57 (RICE, 2007, P. 207)

Lemma B explains an important principle of simple random sampling without replacement: there exists a negative covariance between any two sample observations X_i and X_j, then $i \neq j$. I can therefore define the covariance as: $Cov(X_i, X_j) = -\frac{\sigma^2}{N-1}$.

Since we are sampling without replacement, there would be a dependence between the observations. If I sampled a value of X_i that is higher than the population mean, the remaining values available for sampling would hance a slightly lower average, making X_j more likely to be

below the mean. And hence we have a negative covariance. The converse is also true if X_i is below the mean.

Sample observations are therefore not independent when sampling without replacement. However, as the population size increases, the covariance would approach zero and the dependence becomes negligible.

THEOREM B

With simple random sampling,

$$\begin{aligned}\operatorname{Var}(\overline{X}) &= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\\ &= \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right)\end{aligned}$$

Proof

From Corollary A of Section 4.3,

$$\begin{aligned}\operatorname{Var}(\overline{X}) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}(X_i, X_j)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\operatorname{Cov}(X_i, X_j)\\ &= \frac{\sigma^2}{n} - \frac{1}{n^2}n(n-1)\frac{\sigma^2}{N-1}\end{aligned}$$

After some algebra, this gives the desired result. ■

FIGURE 58 (RICE, 2007, P. 207)

I guess it is up to me to do the remaining algebra of the proof:

$$\begin{aligned}Var(\overline{X}) &= \frac{\sigma^2}{n} - \frac{1}{n^2}n(n-1)\left(\frac{\sigma^2}{N-1}\right)\\ &= \frac{\sigma^2}{n} - \frac{(n-1)\sigma^2}{n(N-1)}\\ &= \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right)\end{aligned}$$

Finite Population Correction Factor (FPC)

The FPC adjust the standard error of an estimate, to account for the fact that the population is finite.

$$standard\ error: \sigma_{\bar{X}} \approx \frac{\sigma}{\sqrt{n}}$$

$$FPC = \sqrt{1 - \frac{n-1}{N-1}} = \sqrt{\frac{N-n}{N-1}}$$

Standard error of the sample mean with finite population correction (FPC):

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}}$$

From here we can prove Corollary B (Rice, 2007, p. 210)

$$Var(T) = N^2 Var(\bar{X}) = N^2 \sigma_{\bar{X}}, \quad (Since\ T = N\bar{X})$$

$$Var(T) = N^2 \left(\frac{\sigma^2}{n}\right)\left(\frac{N-n}{N-1}\right)$$

Estimation of Population Variance

$$\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

THEOREM A

With simple random sampling,

$$E(\hat{\sigma}^2) = \sigma^2 \left(\frac{n-1}{n}\right) \frac{N}{N-1}$$

Proof

Expanding the square and proceeding as in the identity for the population variance in Section 7.2, we find

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

Thus,

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - E(\overline{X}^2)$$

Now, we know that

$$\begin{aligned} E\left(X_i^2\right) &= \text{Var}(X_i) + [E(X_i)]^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

Similarly, from Theorems A and B of Section 7.3.1,

$$\begin{aligned} E(\overline{X}^2) &= \text{Var}(\overline{X}) + [E(\overline{X})]^2 \\ &= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) + \mu^2 \end{aligned}$$

Substituting these expressions for $E(X_i^2)$ and $E(\overline{X}^2)$ in the preceding equation for $E(\hat{\sigma}^2)$ gives the desired result. ■

FIGURE 59 (RICE, 2007, P. 211)

The theorem shows and addresses the bias of $\hat{\sigma}^2$ that leads to the underestimation of the population variance σ^2. It shows that the bias is dependent on both n and N. Completing the algebra left out in Rice, I have:

$$E(\bar{X}^2) = Var(\bar{X}) + [E(\bar{X})]^2 = \frac{\sigma^2}{n}\left(\frac{n-1}{N-1}\right) + \mu^2$$

Substituting these results into the expectation:

$$E(\hat{\sigma}^2) = (\sigma^2 + \mu^2) - \left[\frac{\sigma^2}{n}\left(\frac{n-1}{N-1}\right) + \mu^2\right]$$

$$E(\hat{\sigma}^2) = \sigma^2\left[1 - \frac{1}{n}\left(\frac{n-1}{N-1}\right) + \mu^2\right]$$

$$E(\hat{\sigma}^2) = \sigma^2\left(\frac{n-1}{n} \cdot \frac{N}{N-1}\right)$$

Because $\frac{n-1}{n} \cdot \frac{N}{N-1} < 1$, this means that $E(\hat{\sigma}^2) < \sigma^2$, causing an underestimation of the population variance. We can resolve the bias by multiplying $\hat{\sigma}^2$ by the reciprocal of the correction factor; $\frac{n(N-1)}{(n-1)N}$, then we finally have the unbiased population variance estimator:

$$s^2 = \frac{1}{n-1}\left(1-\frac{n}{N}\right)^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

However, when N is large to n, the factor will be approximately 1 and the bias will be negligible.

COROLLARY A

An unbiased estimate of $\text{Var}(\overline{X})$ is

$$s^2_{\overline{X}} = \frac{\hat{\sigma}^2}{n}\left(\frac{n}{n-1}\right)\left(\frac{N-1}{N}\right)\left(\frac{N-n}{N-1}\right)$$
$$= \frac{s^2}{n}\left(1-\frac{n}{N}\right)$$

where

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

Proof

Since

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

an unbiased estimate of $\text{Var}(\overline{X})$ may be obtained by substituting in an unbiased estimate of σ^2. Algebra then yields the desired result. ■

FIGURE 60 (RICE, 2007, P. 212)

COROLLARY B

An unbiased estimate of $\text{Var}(\hat{p})$ is

$$s^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)$$ ■

FIGURE 61 (RICE, 2007, P. 212)

My Understanding of Simple Random Sampling

Simple Random Sampling randomly selects variables from a finite population, without replacement. These randomly chosen variables follow a discrete distribution based on the frequency of each unique value.

There exists a negative covariance between the samples X_i and X_j. If a large value is chosen from a small sample, the sample mean of the remaining values would be reduced. This effect becomes negligible as the sample size increases.

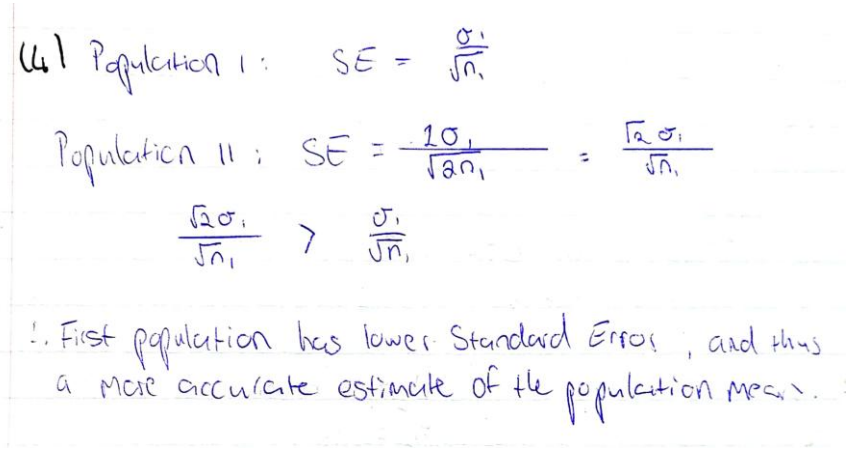
The Finite Population Correction Factor adjusts the standard error, accounting for the fact that the population is finite. We assume that the sample is small and that the sample is relatively large when compared to the population

Because $\widehat{\sigma^2}$ is biased, we need to derive the unbiased population variance estimator.

Exercises

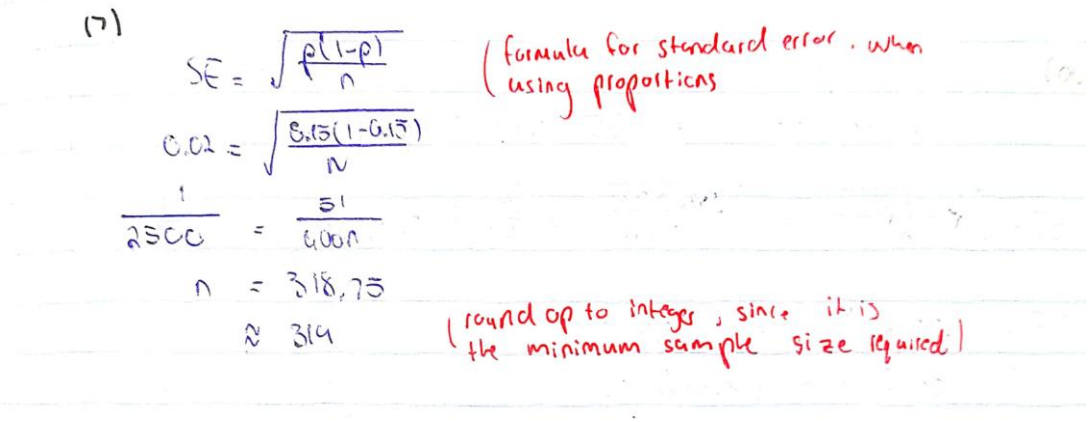
4. Two populations are surveyed with simple random samples. A sample of size n_1 is used for population I, which has a population standard deviation σ_1; a sample of size $n_2 = 2n_1$ is used for population II, which has a population standard deviation $\sigma_2 = 2\sigma_1$. Ignoring finite population corrections, in which of the two samples would you expect the estimate of the population mean to be more accurate?

FIGURE 62 (RICE, 2007, P. 240)



7. Suppose that a simple random sample is used to estimate the proportion of families in a certain area that are living below the poverty level. If this proportion is roughly .15, what sample size is necessary so that the standard error of the estimate is .02?

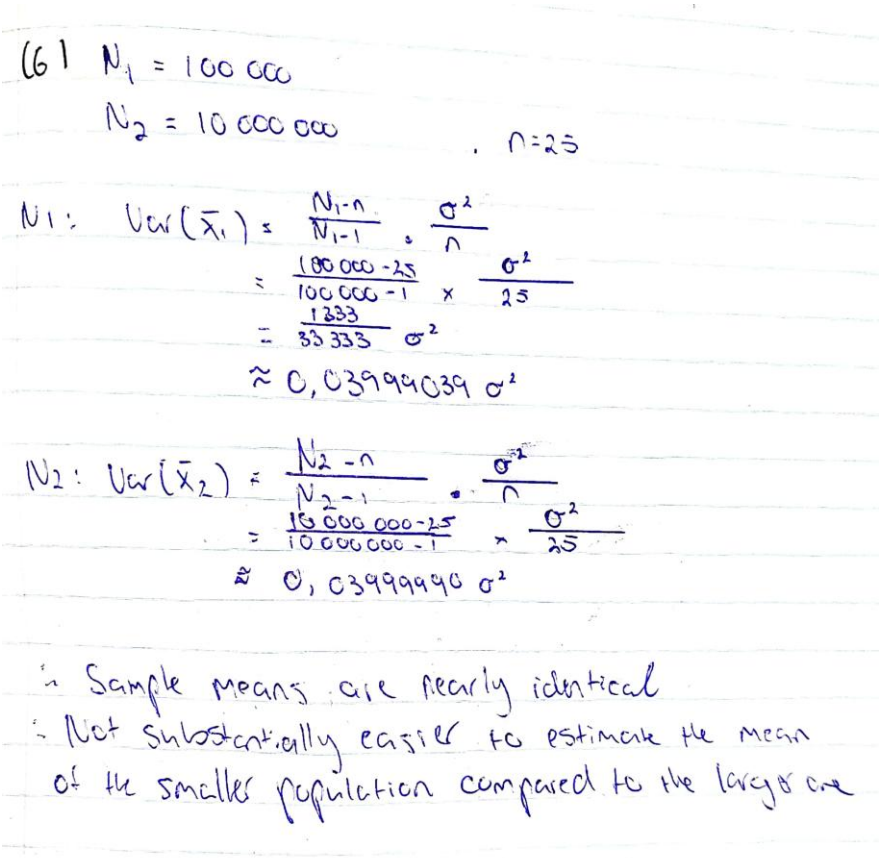
FIGURE 63 (RICE, 2007, P. 240)



Calculating the minimum sample needed to optimize the error, is probably a backwards approach to reducing sapling uncertainty, as mentioned in class, yet such strategies are often used, as explained by the statistician who spoke to us in class. In clinical studies, the minimum "effective" sample size is often used to reduce cost, or to reduce the number of participants exposed to an experimental drug, with unknown side effects.

6. Suppose that two populations have equal population variances but are of different sizes: $N_1 = 100{,}000$ and $N_2 = 10{,}000{,}000$. Compare the variances of the sample means for a sample of size $n = 25$. Is it substantially easier to estimate the mean of the smaller population?

FIGURE 64 (RICE, 2007, P. 240)



Reflection

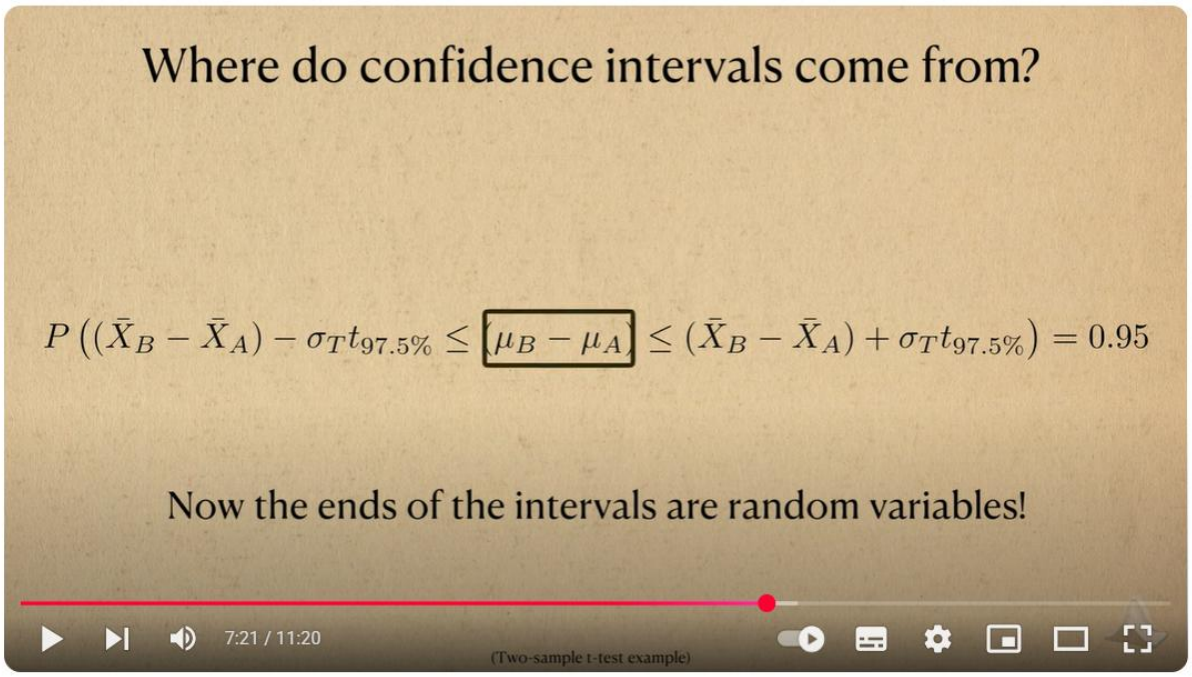
This was an interesting section since there is still much nuance to the simplest form of sampling. I understand that using finite population correction would be redundant in most cases and normal approximation would be just as effective and perhaps simpler.

But finite population correction has it's very important and specific use cases. Since I study Actuarial Sciences, I might find myself in a niche field of general insurance. For example, insuring luxury watches. It that case, the sample of clients is very small, with potentially very expensive payouts, and large variations in the valuation of the watches. The sample size would be large relative to the population. In such cases, finite population correction and other properties and approaches of small samples is extremely relevant and even a field of study on its own.

The Normal Approximation to the Sampling Distribution of $\overline{X}$

Research Process

- I read the Wikipedia article on confidence intervals, where I learned that the most common methods of constructing confidence intervals is using a bootstrap or the central limit theorem.
- I consulted my (Rice, 2007) textbook for the relevant definition and proof.
- I watched a YouTube video by Very Normal, which explained the confidence intervals



Explaining Confidence Intervals and The Critical Region | VNT #6

Very Normal
79,6 k intekenare
Ingeteken
629
Deel
Laai af

I learned a lot from this video, there was a lot more to confidence intervals than I initially thought. He explained that we can say that we are 95% "confident" that the interval we calculate will contain the population difference. There exists a close relationship between the tolerance for Type 1 errors and the confidence for a confidence interval. He also explained that if the value suggested by the null hypothesis is contained within the confidence interval, then it suggests that the data could possibly come from a world where the null hypothesis is true.

Theory

Since I have already learned about the sample mean and central limit theorem, now I can apply the central limit theorem to estimate the probability around the mean. We want to know the probability of the sample mean $\bar{X}_n$ being within some error margin δ of the true mean μ:

$$P|\bar{X}_n - \mu| \leq \delta$$

Rewriting and using the Central Limit Theorem

$$P(-\delta \leq \bar{X}_n - \mu \leq \delta) = \mathrm{P}\left(-\frac{\delta}{\sigma_{\bar{X}}} \leq \frac{\bar{X}_n - \mu}{\sigma_{\bar{X}}} \leq \frac{\delta}{\sigma_{\bar{X}}}\right)$$

$$\approx \Phi\left(\frac{\delta}{\sigma_{\bar{X}}}\right) - \Phi\left(-\frac{\delta}{\sigma_{\bar{X}}}\right)$$

Since $\Phi(-z) = 1 - \Phi(z)$

$$= \Phi\left(\frac{\delta}{\sigma_{\bar{X}}}\right) - \left(1 - \Phi\left(\frac{\delta}{\sigma_{\bar{X}}}\right)\right)$$

$$= 2\Phi\left(\frac{\delta}{\sigma_{\bar{X}}}\right) - 1$$

This gives us an approximate probability that our sample mean is within the margin δ of the population mean μ. The higher δ, or the lower $\frac{\sigma}{\frac{\sigma}{\sqrt{n}}}$, then the higher the probability of our sample mean being close to the true mean.

The Confidence interval

A Confidence interval for a population parameter θ, is a random interval, calculated from the sample, that contains θ with some specified probability.

The 95% confidence is the most prevalent of such intervals. It is the 95% probability of the population mean being within a specified interval and is therefore an important tool for statistical inference.

If the coverage probability is $1 - \alpha$, the interval is called a $100(1 - \alpha)\%$ confidence interval.

For $0 \leq \alpha \leq 1$, let $z(\alpha)$ be a number such that the area under the standard normal density function to the right o $z(\alpha)$ is α. Note that the symmetry of the standard normal density function about zero implies that $z(1 - \alpha) = -z(\alpha)$. If Z follows a standard normal distribution, then, by definition of $z(\alpha)$,

$$P\left(-z\left(\frac{\alpha}{2}\right) \leq Z \leq z\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha$$

From the central limit theorem, $\frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$ has approximately a standard normal distribution, so:

$$P\left(-z\left(\frac{\alpha}{2}\right) \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq z\left(\frac{\alpha}{2}\right)\right) \approx 1 - \alpha$$

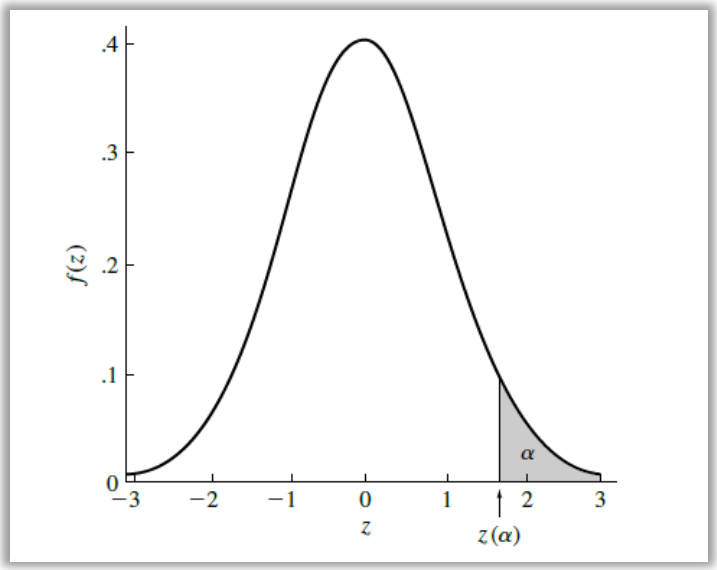


FIGURE 65 FIGURE 7.3 (RICE, 2007, P. 217)

My understanding of the Normal Approximation to the Sampling Distribution of $\bar{X}$

The normal approximation to sampling distribution, use of the normal distribution to approximate the sampling distribution. This is done through the central limit theorem and thus have the same underlying conditions; the random variables must be independently and identically distributed, and the sample size must be sufficiently large (an infinite population)

Normal approximation gives us access to one of our most important tools for making inferences, namely the confidence intervals. It is the percentage probability or "confidence" that the interval we calculate will contain the population parameter that we are interested in.

We can set the bounds of these intervals to the "confidence" of our choosing, 95% is most used confidence level. Setting tighter bounds (higher confidence levels) effectively reduces the tolerance for Type 1 errors, due to their close relationship. A Type 1 error could be described as a "false positive" and occurs when a statistical test incorrectly rejects a true null hypothesis. It is typically represented by a significance level α.

Exercises

8. A sample of size $n = 100$ is taken from a population that has a proportion $p = 1/5$.

a. Find δ such that $P(|\hat{p} - p| \geq \delta) = 0.025$.

b. If, in the sample, $\hat{p} = 0.25$, will the 95% confidence interval for p contain the true value of p?

FIGURE 66 (RICE, 2007, P. 240)

(8)(a) $P(|\hat{p} - p| \geq \delta) = 0.025$, $n = 100$, $p = \frac{1}{5} = 0.2$

$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.2(1-0.2)}{100}} = 0.04$

(8)(a) $P(|\hat{p} - p| \geq \delta) = 0.025$, $n = 100$, $p = \frac{1}{5} = 0.2$

(since $\hat{p}$ is approximately normal by CLT :)

$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.2(1-0.2)}{100}} = 0.04$

The critical value for a total probability of 0.025 in both tails (0.0125 in each tail) is 2.24 using the standard normal tables

$\delta = z_* \cdot SE = 2.24 \times 0.04 = 0.0896$

(b) For sample portion $\hat{p} = 0.25$, the 95% confidence interval is calculated using the SE based on $\hat{p}$:

$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.25 \times 0.75}{100}} = \frac{\sqrt{3}}{40} \approx 0.0433$

$ME = 1.96 \times \frac{\sqrt{3}}{40} = 0.08487$

∴ The confidence interval is :

$0.25 \pm 0.08487 \Rightarrow (0.1651 ; 0.3349)$

∴ Since the true proportion $p = 0.2$ is within the interval, the answer is YES

9. In a simple random sample of 1,500 voters, 55% said they planned to vote for a particular proposition, and 45% said they planned to vote against it. The estimated margin of victory for the proposition is thus 10%. What is the standard error of this estimated margin? What is an approximate 95% confidence interval for the margin?

FIGURE 67 (RICE, 2007, P. 240)

(4) Margin of victory $= 0{,}55 - 0{,}45 = 0{,}10$

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2) + 2\hat{p}_1\hat{p}_2}{n}}$$

$$= \sqrt{\frac{4\hat{p}_1(1-\hat{p}_1)}{n}}$$

since $n_1 = n_2 = 1500$

$\hat{p}_2 = 1 - \hat{p}_1$

$$= \sqrt{\frac{4(0.55)(1-0.55)}{1500}}$$

$= 0{,}02569 \approx 2{,}57\%$

95% Confidence interval for the margin:

95% CI $\therefore z = 1{,}96$

$\Rightarrow (0.1) \pm (1{,}96)(0{,}02569)$

$= 0{,}049647 \; ; \; 0{,}150352$

$\approx 4{,}96\%$ to $15{,}04\%$

22. An investigator quantifies her uncertainty about the estimate of a population mean by reporting $\overline{X} \pm s_{\overline{X}}$. What size confidence interval is this?

FIGURE 68 (RICE, 2007, P. 242)

(22) $\overline{X} \pm S_{\overline{X}}$

And I know $S_{\overline{X}}$ is equal to one standard error. Therefore a confidence interval centered at $\overline{X}$, with a margin of error equal to one standard error.

Since the confidence level is based on the standard normal distribution, I can find the z-score of 1 $S_{\overline{X}}$

$\therefore$ Confidence level = 68,27%

This problem was really easy, but it does show how the standard error notation can be used to describe confidence intervals. Setting up confidence intervals about the standard error is not uncommon.

Reflection

This was a short yet very important section. It was a relief from all the proofs in the previous section and the ones up ahead.

I have a firm understanding of confidence intervals due to prior exposure in STSM2634 statistical programming and a firm understanding of the central limit theorem.

Stratified Random Sampling

Research Process

- I read an article by Adam Hayes on investopedia on stratified random sampling, where I learned that it is a method of sampling that involves the division of a population into smaller sub-groups known as strata.
- I read the Wikipedia article on Stratified Sampling, where I learned that the process of stratification is both collectively exhaustive and mutually exclusive.
- I prompted DeepSeek R1 a comprehensive overview of stratified random sampling and its related properties.
- I consulted my (Rice, 2007) textbook for the relevant theorems and proofs.

Definition

Stratified random sampling divides a heterogeneous population into homogeneous subgroups (strata) based on shared characteristics (e.g., income, region, age). Each stratum l has:

I. Population size: N_l
II. Stratum mean and Stratum variance: μ_l and σ_l^2
III. Weight: $W_l = \frac{N_l}{N}$, where $N = N_1 + N_2 + \cdots + N_L$ is the total population

The total population mean μ is a weighted average sum of stratum means:

$$\mu = \sum_{l=1}^{L} W_l \mu_l$$

To estimate μ, we take a simple random sample of size n_l from each stratum and compute the stratum sample mean $\overline{X_l}$. Then the stratified estimator:

$$\overline{X_s} = \sum_{l=i}^{L} W_l \overline{X_l}$$

Stratified random sampling can be a powerful tool for reducing estimation error and therefore the sampling uncertainty in heterogeneous populations, which are just populations with different subgroups within them that share common traits. Stratified sampling will outperform simple random sampling in accuracy and reliability.

My Understanding of Stratified Random Sampling

The process of stratification divides the population into distinct sub-groups called strata, based on shared characteristics. The strata are both mutually exclusive and collectively exhaustive; meaning that each individual belongs to one and only one stratum.

Stratified sampling works under the assumption that the population is infinitely large, and we can only take a limited amount of samples.

Random individuals are sampled from within each stratum. Samples can be drawn equally across strata, or proportionally or optimally, as I will discuss in the following section.

From theorem A, I know that the stratified estimate of the population mean $\overline{X_s}$ is unbiased. Stratified sampling is more precise when compared to simple random sampling.

THEOREM A

The stratified estimate, $\overline{X}_s$, of the population mean is unbiased.

Proof

$$E(\overline{X}_s) = \sum_{l=1}^{L} W_l E(\overline{X}_l)$$
$$= \frac{1}{N} \sum_{l=1}^{L} N_l \mu_l$$
$$= \mu \quad \blacksquare$$

FIGURE 69 (RICE, 2007, P. 229)

Theorem A states that $\overline{X_s}$ is an unbiased estimator of the population mean μ. We assume that the samples from different strata are independent and that each stratum's sample mean $\overline{X}_l$ is calculated from simple random sampling within that stratum.

Then the stratified estimator $\overline{X_s}$ would "inherit" the unbiasedness from the unbiasedness of individual stratum means $(E(\overline{X}_l) = \mu)$ and weighting these stratum means by their population proportions $\left(W_l = \frac{N_l}{N}\right)$.

THEOREM B

The variance of the stratified sample mean is given by

$$\text{Var}(\overline{X}_s) = \sum_{l=1}^{L} W_l^2 \left(\frac{1}{n_l}\right)\left(1 - \frac{n_l - 1}{N_l - 1}\right)\sigma_l^2$$

Proof

Since the $\overline{X}_l$ are independent,

$$\text{Var}(\overline{X}_s) = \sum_{l=1}^{L} W_l^2 \text{Var}(\overline{X}_l)$$

From Theorem B of Section 7.3.1, we have

$$\text{Var}(\overline{X}_l) = \frac{1}{n_l}\left(1 - \frac{n_l - 1}{N_l - 1}\right)\sigma_l^2$$

Therefore, the desired result follows. ■

FIGURE 70 (RICE, 2007, P. 229)

From Theorem B, I can derive some important aspects: higher variance within a stratum will lead to a higher overall variance. Larger samples would reduce the variance within each stratum. (This property is leveraged in optimal allocation to reduce the overall variance) The Finite Population Correction adjust for sampling without replacement, and it becomes negligible if N_l becomes large relative to n_l. Larger strata, with higher W_l , would dominate the variance. Their Influence are amplified by their weights being squared.

THEOREM C

With stratified random sampling, the difference between the variance of the estimate of the population mean based on proportional allocation and the variance of that estimate based on optimal allocation is, ignoring the finite population correction,

$$\text{Var}(\overline{X}_{sp}) - \text{Var}(\overline{X}_{so}) = \frac{1}{n}\sum_{l=1}^{L} W_l(\sigma_l - \bar{\sigma})^2$$

where

$$\bar{\sigma} = \sum_{l=1}^{L} W_l\sigma_l$$

Proof

$$\text{Var}(\overline{X}_{sp}) - \text{Var}(\overline{X}_{so}) = \frac{1}{n}\left[\sum_{l=1}^{L} W_l\sigma_l^2 - \left(\sum_{l=1}^{L} W_l\sigma_l\right)^2\right]$$

The term within the large brackets equals $\sum_{l=1}^{L} W_l(\sigma_l - \bar{\sigma})^2$, which may be verified by expanding the square and collecting terms. ■

FIGURE 71 (RICE, 2007, P. 235)

$$Var(\bar{X}_{sp}) - Var(\bar{X}_{so}) = \frac{1}{n}\sum_{l=1}^{L} W_l\sigma_l^2 - \frac{1}{n}\left(\sum_{l=1}^{L} W_l\sigma_l\right)^2$$

$$= \frac{1}{n}\left[\sum_{l=1}^{L} W_l\sigma_l^2 - \bar{\sigma}^2\right]$$

$$= \frac{1}{n}\left[\sum_{l=1}^{L} W_l\sigma_l^2 - 2\bar{\sigma}^2 + \bar{\sigma}^2\right]$$

$$= \frac{1}{n}\left[\sum_{l=1}^{L} W_l\sigma_l^2 - 2\bar{\sigma}\sum_{l=1}^{L} W_l\sigma_l + \bar{\sigma}^2\sum_{l=1}^{L} W_l\right]$$

$$= \frac{1}{n}\left[\sum_{l=1}^{L} W_l\left(\sigma_l^2 - 2\bar{\sigma}\sigma_l + \bar{\sigma}^2\right)\right]$$

$$= \frac{1}{n}\left[\sum_{l=1}^{L} W_l\left(\sigma_l - \bar{\sigma}\right)^2\right]$$

Here is a more thorough breakdown of the proof as explained by Prof. Michael Von Maltitz in the module course guide. He also explained that it is always a non-negative number, meaning that the variance from optimal allocation is at least as good as the variance from proportional allocation, if not better. Optimal allocation will be discussed in more detail in the following section.

Exercises

58. (Computer Exercise) Construct a population consisting of the integers from 1 to 100. Simulate the sampling distribution of the sample mean of a sample of size 12 by drawing 100 samples of size 12 and making a histogram of the results.

59. (Computer Exercise) Continuing with Problem 58, divide the population into two strata of equal size, allocate six observations per stratum, and simulate the distribution of the stratified estimate of the population mean. Do the same thing with four strata. Compare the results to each other and to the results of Problem 58.

FIGURE 72 (RICE, 2007, P. 250)

```r
library(ggplot2)
library(tidyr)
set.seed(123)

# Question 58: Simple Random sampling
population <- 1:100 # population of integers 1 to 100
pop_mean <- mean(population)

sample_means <- 0
for (i in 1:100){
  sample_data <- sample(population, size = 12, replace = FALSE)
  sample_means[i] <- mean(sample_data)
}

plot1_data <- data.frame(mean = sample_means)
srs_mean <- mean(sample_means)
srs_sd <- sd(sample_means)

#Question 59: Stratified Sampling
# Using 2 Stata
strata2 <- list(
  stratum1 = 1:50,
  stratum2 = 51:100
)

# Allocate 6 observations to each stratum
strata2_means <- numeric(100)
for (i in 1:100){
  sample1 <- sample(strata2$stratum1, size = 6, replace = FALSE)
  sample2 <- sample(strata2$stratum2, size = 6, replace = FALSE)
  ms1 <- mean(sample1)
  ms2 <- mean(sample2)
  strata2_means[i] <- 0.5 * ms1 + 0.5 * ms2
}
```

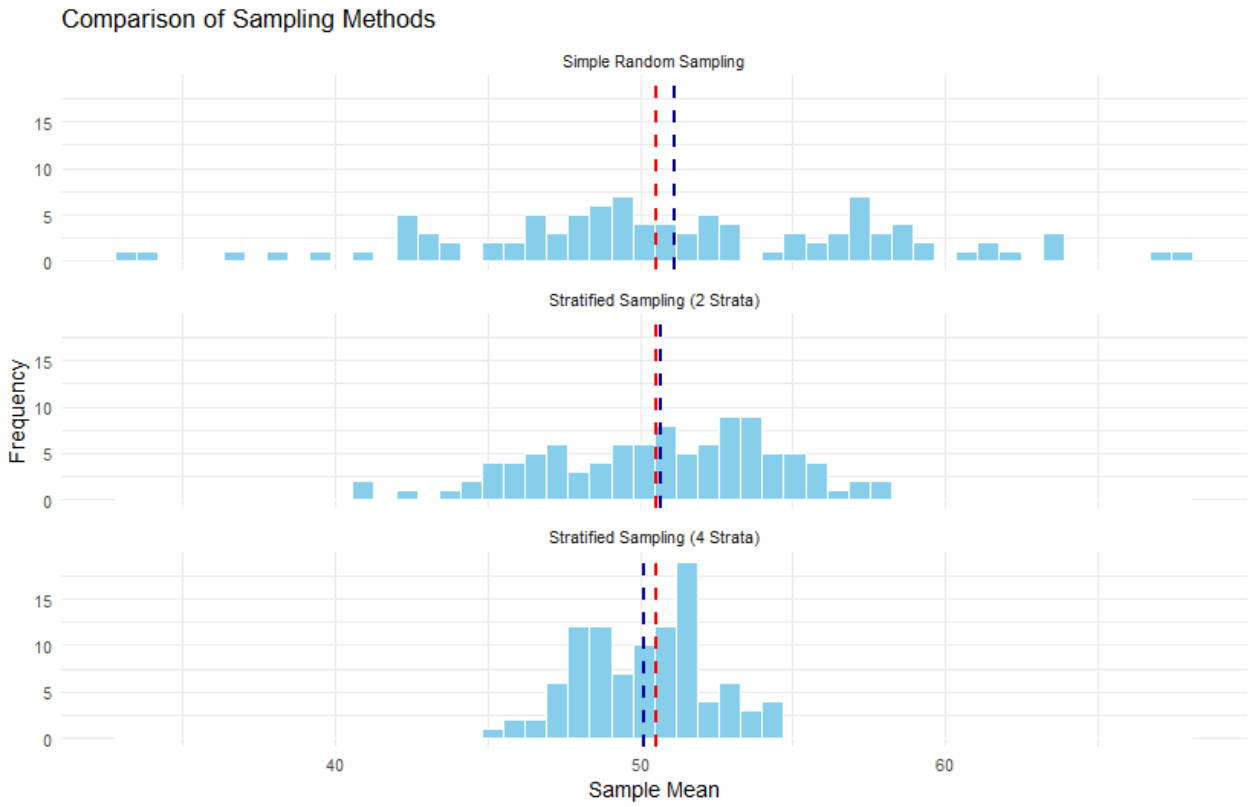
```r
plot2_data <- data.frame(mean = strata2_means)
strata2_mean <- mean(strata2_means)
strata2_sd <- sd(strata2_means)


# Using 4 Strata
strata4 <- list(
  stratum1 = 1:25,
  stratum2 = 26:50,
  stratum3 = 51:75,
  stratum4 = 76:100
)

# Allocate 3 Observations per stata
strata4_means <- numeric(100)
for (i in 1:100) {
  sample1 <- sample(strata4$stratum1, size = 3, replace = FALSE)
  sample2 <- sample(strata4$stratum2, size = 3, replace = FALSE)
  sample3 <- sample(strata4$stratum3, size = 3, replace = FALSE)
  sample4 <- sample(strata4$stratum4, size = 3, replace = FALSE)
  ms1 <- mean(sample1)
  ms2 <- mean(sample2)
  ms3 <- mean(sample3)
  ms4 <- mean(sample4)
  strata4_means[i] <- 0.25 * ms1 + 0.25 * ms2 + 0.25 * ms3 + 0.25 * ms4
}

plot3_data <- data.frame(mean = strata4_means)
strata4_mean <- mean(strata4_means)
strata4_sd <- sd(strata4_means)

```

```r
plot3_data <- data.frame(mean = strata4_means)
strata4_mean <- mean(strata4_means)
strata4_sd <- sd(strata4_means)

results <- data.frame(
  Method = c("Simple Random Sampling",
             "Stratified Sampling (2 Strata)",
             "Stratified Sampling (4 Strata)"),
  Mean = c(srs_mean, strata2_mean, strata4_mean),
  Standard_Deviation = c(srs_sd, strata2_sd, strata4_sd)
)
print(results)

# I need to create a data frame for all the data,
# so I can create a single facated plot, for fair comparison
combined_data <- data.frame(
  Mean = c(sample_means, strata2_means, strata4_means),
  Method = factor(
    rep(c("Simple Random Sampling", "Stratified Sampling (2 Strata)",
          "Stratified Sampling (4 Strata)"), each = 100),
    levels = c("Simple Random Sampling", "Stratified Sampling (2 Strata)",
               "Stratified Sampling (4 Strata)")
  )
)

# Data frame for method means, so I can graph their intercepts
method_means <- data.frame(
  Method = c("Simple Random Sampling", "Stratified Sampling (2 Strata)",
             "Stratified Sampling (4 Strata)"),
  Mean = c(srs_mean, strata2_mean, strata4_mean)
)

```

```r
histograms <- ggplot(combined_data, aes(x = Mean)) +
  geom_histogram(bins = 50, fill = "skyblue", color = "white") +
  geom_vline(xintercept = pop_mean, color = "red",
             linetype = "dashed", size = 1) +
  geom_vline(data = method_means, aes(xintercept = Mean),
             color = "darkblue", linetype = "dashed", size = 1) +
  facet_wrap(~ Method, ncol = 1) +
  labs(title = "Comparison of Sampling Methods",
       x = "Sample Mean",
       y = "Frequency") +
  theme_minimal()
histograms
```

Ctrl+L to chat, Ctrl+K to generate

```
                        Method     Mean Standard_Deviation
1         Simple Random Sampling 51.14833           6.865088
2 Stratified Sampling (2 Strata) 50.65583           3.769295
3 Stratified Sampling (4 Strata) 50.14583           2.068156
>
```



I had the opportunity to have my portfolio reviewed a second time, whereafter, I decided to redo this question. I wrote the code myself, but there were much room for improvement. I had three different plots, all with different scales, which made fair comparison impossible. I have now completed my module on statistical programming and have learned a lot since I first attempted this question, including how to facet wrap graphs.

We can now observe how the data become more concentrated around the mean when we use stratified random sampling and becomes even more concentrated if we increase the strata.

The blue line shows the strata mean and the read line shows the population mean for comparison.

Reflection

I enjoyed this section very much and I am curious to see how proportional and optimal allocation would work. I had lots of fun attempting the coded exercises 58 and 59 from Rice. The code turned out quite a bit longer than I expected, but luckily, it was not too complicated.

I do believe that I have a firm understanding of stratified random sampling and can compare it, for example, to simple random sampling.

When there is some information available about the composition of the population and we can identify heterogeneous subgroups within our population, stratified random sampling can be a powerful method to reduce sampling uncertainty.

Methods of Allocation

Research Process

- I read an article on ScienceDirect on proportional allocation, and learned that it was introduced by Bowley in 1926.
- I read an article on ScienceDirect on optimal allocation.
- I asked ChatGPT to differentiate between and give the proper use cases of both optimal allocation and proportional allocation.
- I Consulted section 14.8 of my (James Steward, 2020) Calculus textbook on the method of Lagrange Multipliers.
- I consulted my (Rice, 2007) textbook on all relevant theorems and definitions.
- I found a pdf of some lecture notes from CalTech.edu comparing Neyman allocation to proportional allocation.

Optimal Allocation

Optimal Allocation refers to the allocation of sample sizes across strata in a way that minimizes the variance of estimated population parameters. It aims to provide more precise parameter estimates compared to simple random sampling or proportional allocation, by sampling strata in proportion to their standard deviations and population fractions.

In this Section I will be looking at Neyman allocation specifically.

This is an AI generated definition based on: Encyclopaedia of Social Measurement, 2005, That I found in the ScienceDirect article on Optimal Allocation. Using AI for articles could be a problem if future models are trained on these articles, however using AI to define or summarise from a specific text, is very effective and reduce the risk of the AI model making mistakes.

THEOREM A

The sample sizes $n_1, \ldots, n_L$ that minimize $\text{Var}(\overline{X}_s)$ subject to the constraint $n_1 + \cdots + n_L = n$ are given by

$$n_l = n \frac{W_l \sigma_l}{\sum_{k=1}^{L} W_k \sigma_k}$$

where $l = 1, \ldots, L$.

FIGURE 73 (RICE, 2007, P. 232)

Proof

We introduce a Lagrange multiplier, and we must then minimize

$$L(n_1, \ldots, n_L, \lambda) = \sum_{l=1}^{L} \frac{W_l^2 \sigma_l^2}{n_l} + \lambda \left(\sum_{l=1}^{L} n_l - n \right)$$

For $l = 1, \ldots, L$, we have

$$\frac{\partial L}{\partial n_l} = -\frac{W_l^2 \sigma_l^2}{n_l^2} + \lambda$$

Setting these partial derivatives equal to zero, we have the system of equations

$$n_l = \frac{W_l \sigma_l}{\sqrt{\lambda}}$$

for $l = 1, \ldots, L$. To determine λ, we first sum these equations over l:

$$n = \frac{1}{\sqrt{\lambda}} \sum_{l=1}^{L} W_l \sigma_l$$

Thus,

$$\frac{1}{\sqrt{\lambda}} = \frac{n}{\sum_{l=1}^{L} W_l \sigma_l}$$

and

$$n_l = n \frac{W_l \sigma_l}{\sum_{l=1}^{L} W_l \sigma_l}$$

which proves the theorem. ■

FIGURE 74 (RICE, 2007, P. 233)

Theorem A explains the Neyman allocation, which is an optimal allocation method used to minimize variance by leveraging stratified random sampling.

This is done by minimizing the stratified sample variance ($\mathrm{Var}(\overline{X_s})$), by subjugating it to the constraint $\sum_{l=1}^{L} n_l = n$. This was done with the use of a Lagrange multiplier. Lagrange multiplier was not yet covered in Vector Analysis, so I consulted my (James Steward, 2020, p. 1021) calculus textbook, where I learned that the Lagrange multiplier λ follows from $\nabla f(x_0, y_0, z_0) = \lambda \nabla g(x_0, y_0, z_0)$. This essentially allocates more samples with larger weights W_l or larger variances σ_l. As explained in Rice, if W_l is large, the stratum contains a large fraction of the population and if σ_l is large, the population values in the stratum are quite variable. Thus, heavier sampling would be needed to reduce the standard error of those strata and therefore provide a better estimation of the mean.

This method of optimal allocation could be a more efficient estimation of the population mean, by prioritising the strata that have the largest contribution to the variability of the population, or the population composition. Even though this method is mathematically justified, it can open the door to some forms of "data wrangling" or at least inappropriate implementation, as

mentioned in class. The optimal allocation may for example oversample small, but highly variable strata, therefore underrepresenting the much larger and more stable strata.

COROLLARY A

Denoting by $\overline{X}_{so}$, the stratified estimate using the optimal allocations as given in Theorem A and neglecting the finite population correction,

$$\operatorname{Var}(\overline{X}_{so}) = \frac{\left(\sum_{l=1}^{L} W_l \sigma_l\right)^2}{n}$$ ■

FIGURE 75 (RICE, 2007, P. 233)

Proportional Allocation

Proportional allocation is an intuitive way of reducing sampling uncertainty. For example, some pre-election polls would survey the same proportion of people from a province, as the proportion of that province of the country's population. If 10 million people live a province within a country with a population of 100 million. A poll that takes 100 samples would choose to select 10 random samples from within that province.

We which to maintain the populations natural composition within the sample; using the same sampling fraction in each stratum

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_l}{N_l}$$

Which holds if $n_1 = n\left(\frac{N_l}{N}\right) = nW_l$

For $l = 1, \ldots, L$. This method is called proportional allocation. The estimate of the population mean based on proportional allocation is

$$\begin{aligned}\overline{X_{sp}} &= \sum_{l=1}^{L} W_l \overline{X}_l \\ &= \sum_{l=1}^{L} W_l \left(\frac{1}{n_l}\right) \sum_{i=1}^{n_l} X_{il} \\ &= \frac{1}{n} \sum_{l=1}^{L} \sum_{i=1}^{n_l} X_{il}\end{aligned}$$

Since $\frac{W_l}{n_l} = \frac{1}{n}$. This estimate is simply the unweighted mean of the sample values

THEOREM **B**

With stratified sampling based on proportional allocation, ignoring the finite population correction,

$$\mathrm{Var}(\overline{X}_{sp}) = \frac{1}{n}\sum_{l=1}^{L} W_l\sigma_l^2$$

FIGURE 76 (RICE, 2007, P. 234)

Proof

From Theorem B of Section 7.5.2, we have

$$\begin{aligned}\mathrm{Var}(\overline{X}_{sp}) &= \sum_{l=1}^{L} W_l^2\mathrm{Var}(\overline{X}_l)\\ &= \sum_{l=1}^{L} W_l^2\frac{\sigma_l^2}{n_l}\end{aligned}$$

Using $n_l = nW_l$, the result follows. ■

FIGURE 77 (RICE, 2007, P. 235)

THEOREM C

With stratified random sampling, the difference between the variance of the estimate of the population mean based on proportional allocation and the variance of that estimate based on optimal allocation is, ignoring the finite population correction,

$$\mathrm{Var}(\overline{X}_{sp}) - \mathrm{Var}(\overline{X}_{so}) = \frac{1}{n}\sum_{l=1}^{L} W_l(\sigma_l - \bar{\sigma})^2$$

where

$$\bar{\sigma} = \sum_{l=1}^{L} W_l\sigma_l$$

Proof

$$\mathrm{Var}(\overline{X}_{sp}) - \mathrm{Var}(\overline{X}_{so}) = \frac{1}{n}\left[\sum_{l=1}^{L} W_l\sigma_l^2 - \left(\sum_{l=1}^{L} W_l\sigma_l\right)^2\right]$$

The term within the large brackets equals $\sum_{l=1}^{L} W_l(\sigma_l - \bar{\sigma})^2$, which may be verified by expanding the square and collecting terms. ■

FIGURE 78 (RICE, 2007, P. 235)

Theorem C compares that if the strata variances are all equal, both proportional allocation and optimal allocation would yield the same results. However, if these strata variances are higher, optimal allocation would yield the lower variance and would thus be more applicable.
Despite this property, proportional allocation is much more prevalent and often the better option over optimal allocation. Proportional allocation is simpler and does not require all the stratum standard deviations to be known. Proportional allocation is fairer in the sense that it ensures that each group is represented in the same proportion as in their population.

My understanding of the different methods of Allocation

Neyman allocation, which is an optimal allocation method, seeks to minimize the variance of the estimated population parameters. Strata with larger weights and variances are sampled more heavily. Larger sample sizes are allocation to these strata, to reduce their sampling error.

Optimal allocation can be the most precise sampling method and does produce the smallest estimate of the population variance. However, it does require all stratum standard deviations to be known. Optimal allocation might not always be a fair representation of data, since a smaller,

highly variable data might be sampled more heavily, therefore underrepresenting the much larger and more stable data.

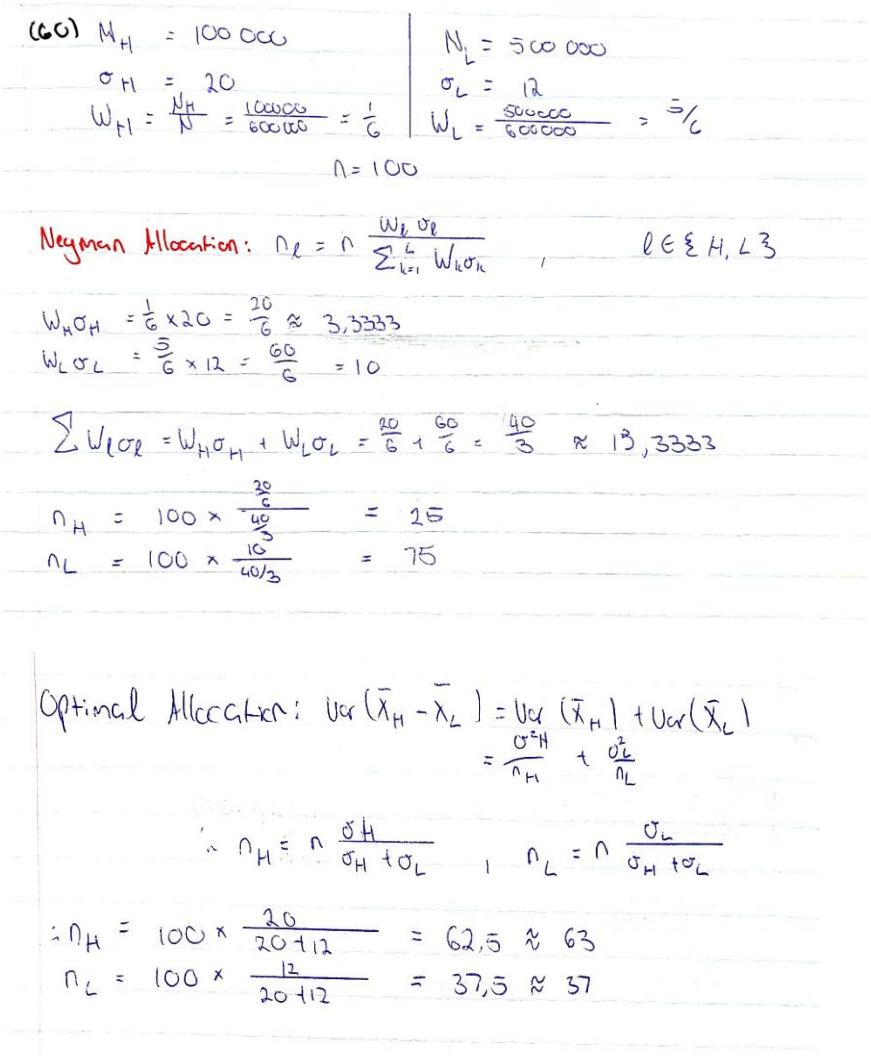
Proportional allocation is also a stratified random sampling method, where the sample size of each stratum is proportional to the stratum's size in the population. When there are some information available about out population's composition, proportional allocation is one of our most powerful tools of reducing uncertainty when sampling.

Exercises

60. A population consists of two strata, H and L, of sizes 100,000 and 500,000 and standard deviations 20 and 12, respectively. A stratified sample of size 100 is to be taken.

a. Find the optimal allocation for estimating the population mean.
b. Find the optimal allocation for estimating the difference of the means of the strata, $\mu_H - \mu_L$.

FIGURE 79 (RICE, 2007, P. 250)



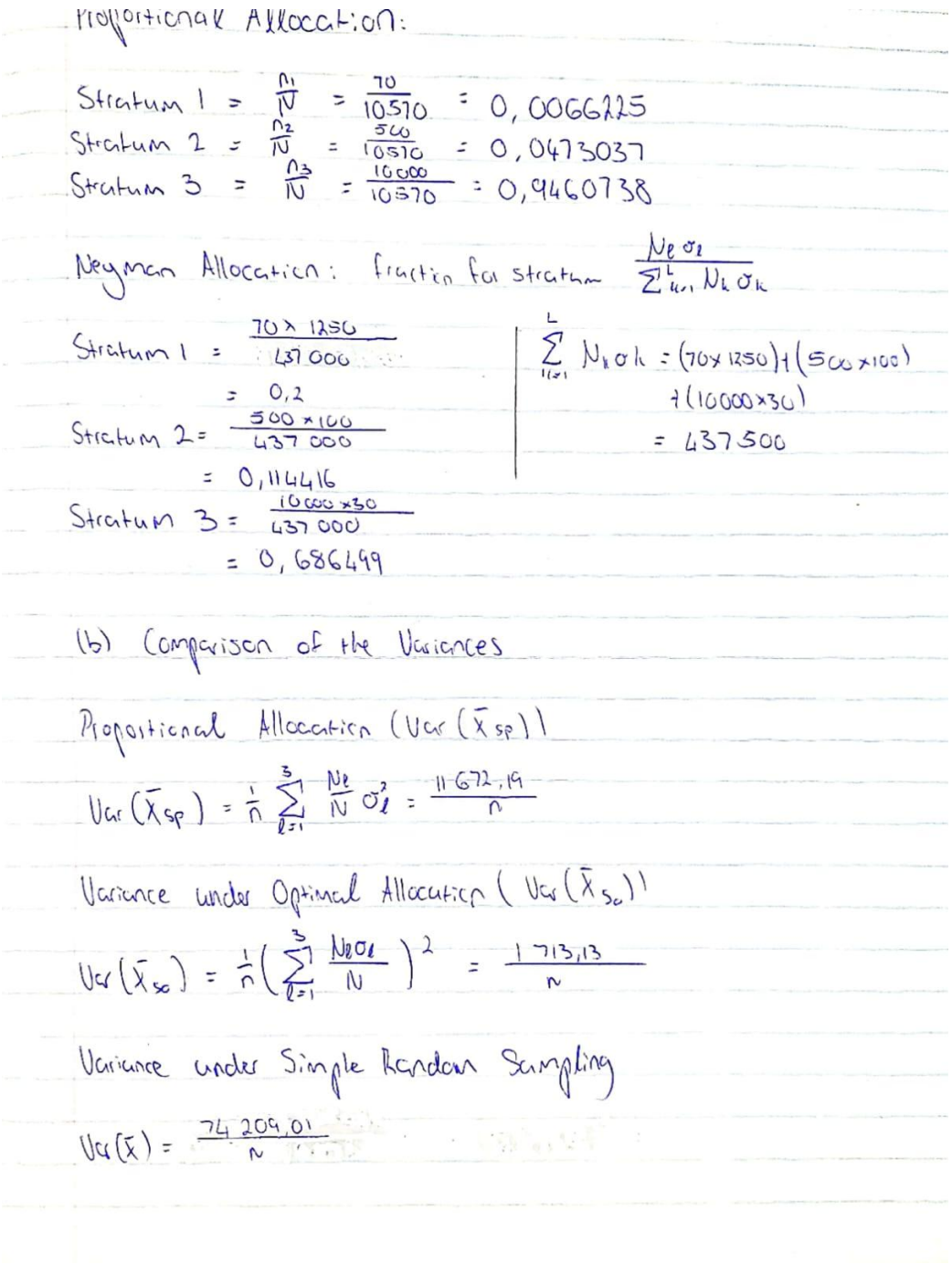
64. The value of an inventory is to be estimated by sampling. The items are stratified by book value in the following way:

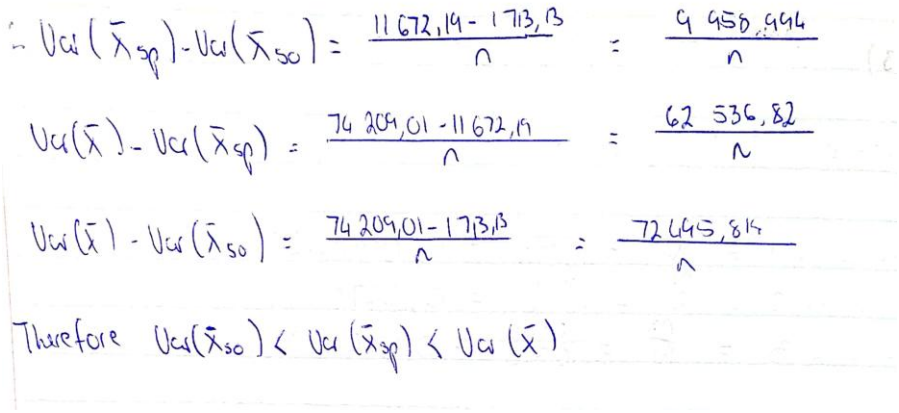
Stratum	N_l	μ_l	σ_l
\$1000 +	70	3000	1250
\$200–1000	500	500	100
\$1–200	10,000	90	30

a. What should the relative sampling fraction in each stratum be for proportional and for optimal allocation? Ignore the finite population correction.

b. How do the variances under each type of allocation compare to each other and to the variance under simple random sampling?

FIGURE 80 (RICE, 2007, P. 251)





This problem could be considered "garbage" in the sense that the stratum means and variances are given. However, it is an effective comparison showing the difference in variance sizes of the different sampling methods. For optimal allocation, it is required for the stratum standard deviation to be known. As expected, the smallest variance was given by optimal allocation, followed by proportional allocation then simple random sampling.

Reflection

I found this section very interesting, since the methods discussed is more widely applicable and used than finite population corrections.

I can compare the different methods of allocation and when they would be appropriate or not and explain their respective advantages and disadvantages. I can explain their difference in relative accuracy.

Neyman allocation is probably the most complicated method I encountered in this chapter, and it can be quite tedious to solve even small samples by hand. I know there are a package in R that makes optimal allocation much easier, but I have not used it yet.

Optimal allocation was more prevalent than I first anticipated, particularly in the study of consumer behaviour. It can be observed in many industries that the consumer behaviour, like spending and preference, are more variable in younger generations than in older generations. There could many reasons for this observation, like younger people being more likely to follow trends. But regardless of the underlying reasons, using optimal allocation to more heavily sample younger consumers, can reduce the overall variance of estimates.

Proportional allocation is incredibly useful and prevalent. It is more widely used than other methods discussed in this chapter, like Neyman allocation and even simple random sampling with finite population correction.

Assignment 3

Problem

Using the palmer penguins dataset, compare the estimated variance and 95% confidence intervals of simple random sampling with and without finite population correction of the penguin body mass, using 25 and 100 samples. Compare the estimated, variance and 95% confidence intervals using proportional stratified random sampling according to species using sample sizes of 25 and 100.

Important Assumptions

The penguin's dataset contains 344 samples from three species of penguins living on three islands. We know that these populations are finite since there can only live a finite number of penguins on each island. For the purpose of this assignment, I will set the "population" as the dataset total of 344 members, from which I can perform random sampling. I am aware that the dataset itself is a sample from the actual population of penguins living on those islands.

Solution

```r
library(dplyr)
library(palmerpenguins)
set.seed( seed: 123)
library(ggplot2)

data(penguins)
data1 <- na.omit(penguins) # Since there are some missing values
data1$Species <- as.factor(data1$species) # add a column for specis as stratum

ggplot(data1, aes(x = body_mass_g, color = species, fill = species)) +
  geom_density(alpha = 0.3) +
  labs(title = "Density Plot of Penguin Body Mass according to Species",
       x = "Body Mass (g)", y = "Density") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1") +
  scale_color_brewer(palette = "Set1") +
  theme_minimal()

N <- nrow(data1) # Population Size
n <- 25 # Sample Size
```

```r
# SRS sample
srs_indices <- sample(1:N, size = n, replace = FALSE)
srs_sample <- data1[srs_indices, ]
x_srs <- srs_sample$body_mass_g

# Sample Statistics
x_bar_srs <- mean(x_srs)
s2_srs <- var(x_srs)

# Finite Population Correction Factor
fpc_factor <- 1 - (n / N)
Var_hat1 <- s2_srs / n               # Variance without FPC
Var_hat2 <- Var_hat1 * fpc_factor     # Variance with FPC
```

```r
# Standard Errors
se1 <- sqrt(Var_hat1)
se2 <- sqrt(Var_hat2)

# 95% Confidence Intervals
z_value <- qnorm( p: 0.975)

ci_lower1 <- x_bar_srs - z_value * se1
ci_upper1 <- x_bar_srs + z_value * se1
ci_width1 <- ci_upper1 - ci_lower1

ci_lower2 <- x_bar_srs - z_value * se2
ci_upper2 <- x_bar_srs + z_value * se2
ci_width2 <- ci_upper2 - ci_lower2
```

The sample mean:

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$= \frac{1}{25}\sum_{i=1}^{25} X_i$$

$$= 4275$$

Sample Variance:

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

$$= \frac{1}{25-1}\sum_{i=1}^{25}(X_i - 4275)^2$$

$$= 698{,}802.1$$

Variance of Sample Mean Estimator without Finite Population Correction

$$\widehat{Var}(\overline{X_n}) = s^2_{\overline{X_1}} = \frac{s_n^2}{n}$$

$$= \frac{698802.1}{25}$$

$$= 27{,}953.084$$

Variance of Sample Mean Estimator with Finite Population Correction

$$\widehat{Var}(\overline{X_n}) = s^2_{\overline{X_2}} = \left(\frac{N-n}{N-1}\right)\cdot\frac{s_n^2}{n}$$

$$= \left(\frac{344-25}{344-1}\right)\left(\frac{698802.1}{25}\right)$$

$$= 25{,}996.25$$

Standard Errors

$$SE_1 = \sqrt{s^2_{\overline{X_1}}} = \sqrt{27{,}953.084} = 167.1918$$

$$SE_2 = \sqrt{s^2_{\overline{X_2}}} = \sqrt{25{,}920.68} = 160{,}999$$

95% Confidence Intervals for μ

I am using normal approximation for consistency between sample sizes, however at a sample size or 25, using a t distribution would arguably have been more appropriate. At 100 samples, normal approximation is appropriate.

$$z_{0.975} = 1.96$$

Without Finite Population Correction:

$$\overline{X_n} \pm z_{0.975} \cdot SE_1$$

$$4275 \pm 1.96 \cdot 167.1918 = (3947.316, 4602.684)$$

With Finite Population Correction:

$$\overline{X_n} \pm z_{0.975} \cdot SE_2$$

$$4275 \pm 1.96 \cdot 160{,}9996 = (3959.856, 4590.144)$$

Stratified Random Sampling:

Since there are three species of penguins, with different proportions of the population, proportional stratified sampling would be appropriate on this dataset. I provided the code I had already written for the simple random sampling to ChatGPT and asked it, “Using the code provided, help me set up the stratums according to species, for proportional random sampling, do not modify any other parts of the code.” Lines 62 to 65 and 83 to 95 was generated by ChatGPT.

```r
# Stratified Random Sampling (Proportional to Species)
# Population in each Species
N_l <- table(data1$Species)  # Total in each stratum
strata_names <- names(N_l)
n_l <- round(n * N_l / N)        # Proportional sample size per stratum

stratified_sample <- data1 %>%
  group_by(Species) %>%
  group_modify(~ .x[sample(1:nrow(.x), size = n_l[.y$Species], replace = FALSE), ]) %>%
  ungroup()

x_strat <- stratified_sample$body_mass_g

# Sample mean and variance
x_bar_strat <- mean(x_strat)
s2_strat <- var(x_strat)
```

```r
stratum_stats <- stratified_sample %>%
  group_by(Species) %>%
  summarise(
    N_l = nrow(data1[data1$Species == unique(Species), ]), # true stratum size
    n_l = n(),             # sample size within stratum
    xbar_l = mean(body_mass_g),
    s2_l = var(body_mass_g)
  ) %>%
  mutate(
    weight = N_l / N,
    fpc_l = (N_l - n_l) / (N_l - 1),  # FPC for each stratum
    var_l = (weight^2) * (s2_l / n_l) * fpc_l  # Apply FPC to each stratum
  )
```

```r
# Stratified Sample Mean (weighted mean)
x_bar_strat <- sum(stratum_stats$weight * stratum_stats$xbar_l)

# Variance estimator for stratified sampling (without FPC)
var_strat <- sum(stratum_stats$var_l)

# Standard error
se_strat <- sqrt(var_strat)

# 95% Confidence interval for stratified sample mean
ci_strat <- c(x_bar_strat - z_value * se_strat,
              x_bar_strat + z_value * se_strat)
```

```r
x_bar_strat <- sum(stratum_stats$weight * stratum_stats$xbar_l)
var_strat_no_fpc <- sum(stratum_stats$var_l_no_fpc)
var_strat_fpc <- sum(stratum_stats$var_l_fpc)

se_strat_no_fpc <- sqrt(var_strat_no_fpc)
se_strat_fpc <- sqrt(var_strat_fpc)

z_value <- qnorm( p: 0.975)
ci_strat_no_fpc <- c(x_bar_strat - z_value * se_strat_no_fpc,
                     x_bar_strat + z_value * se_strat_no_fpc)

ci_strat_fpc <- c(x_bar_strat - z_value * se_strat_fpc,
                  x_bar_strat + z_value * se_strat_fpc)
```

Stratified Sample Mean

$$\overline{X_s} = \sum_{l=1}^{L} W_l \overline{X_l}$$

$$\overline{X_s} = \sum_{l=1}^{3} \frac{25}{344} \overline{X_l}$$

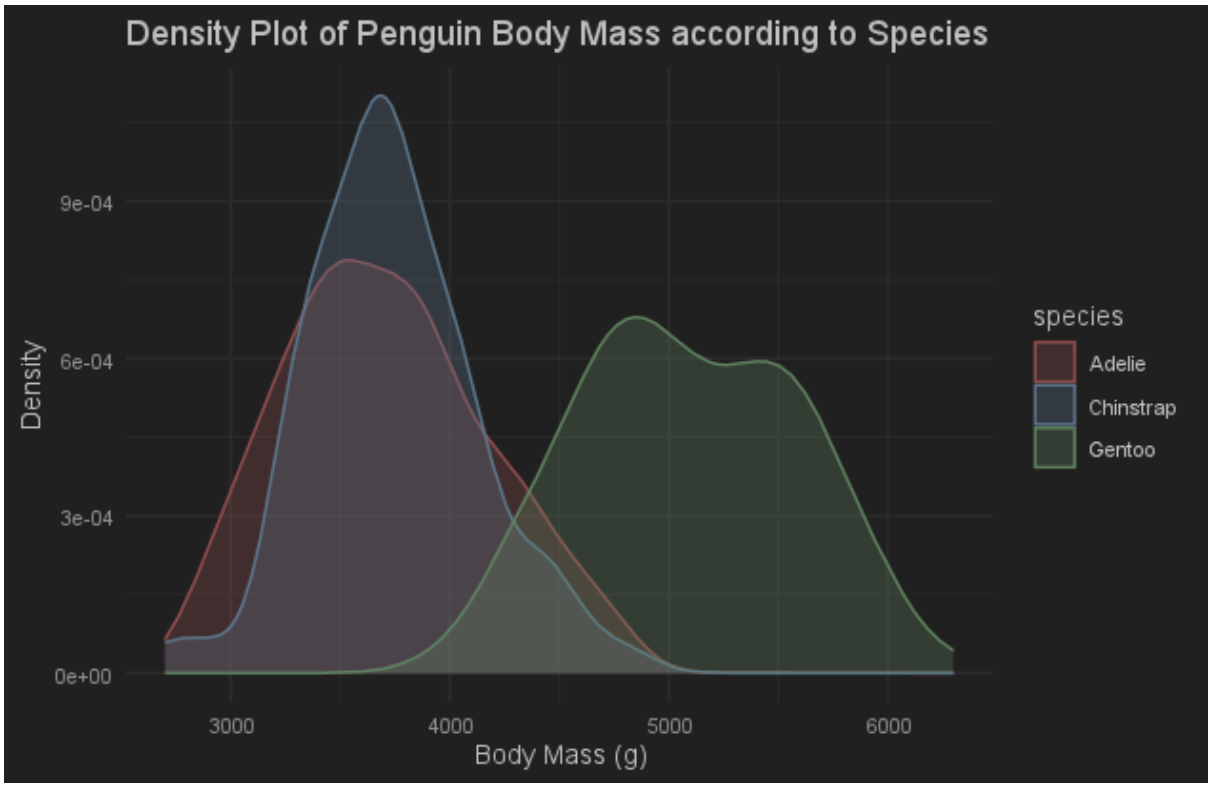
$$= 4179.923$$

Variance of $\overline{X_s}$

$$Var(\overline{X_s}) = \sum_{l=1}^{L} (W_l^2 \cdot \frac{s_l^2}{n_l}) \left(\frac{N_l - n_l}{N_l - 1}\right)$$

$$Var(\overline{X_s}) = 11{,}078.84$$

Interpretation of Results



Method	Sample Mean (g)	Estimated Variance	95% CI (Lower, Upper)	CI Width (g)
Simple Random Sampling (no FPC), 25 samples	4275	27,952.08	(3947.316, 4602.684)	655.3679
Simple Random Sampling (no FPC), 100 samples	4271.75	6748.82	(4110.737, 4432.763)	322.0268
Simple Random Sampling (with FPC), 25 samples	4275	25,931.08	(3959.382, 4590.618)	631.2356
Simple Random Sampling (with FPC), 100 samples	4271.75	4736.371	(4136.863, 4406.637)	269.7745
Stratified Sampling (Proportional to Species), 25 samples	4179.923	11,078.84	(3973.625, 4386.221)	427.01
Stratified Sampling (Proportional to Species), 100 samples	4176.165	1429.439	(4102.439, 4250.267)	148.2043

There was a more meaningful difference between simple random sampling with and without finite population correction than I anticipated. The confidence intervals were 3.68% more narrow using FPC with 25 samples and 16.18% more narrow at 100 samples. These differences are 25,08g and 52.66g respectively. These differences are notable, but not that large considering the mean weight of the penguins are around 4.2kg.

The estimated variance when using simple random sampling with FPC was considerably smaller than the estimated variance without FPC, 7.51% smaller at 25 samples and 42.918% smaller at 100 samples.

We can see how proportional stratified allocation is much more accurate compared to simple random sampling, providing us with the lowest estimate of the variance, 1429.439, and the narrowest confidence interval width, 148.2043.

We can also note the drastic increase in accuracy from 100 samples when compared to 25 samples. The estimated variance of simple random sampling without FPC was reduced by 75.86% from 27,952.08 to 6748.82. The estimated variance of simple random sampling with FPC was reduced by 81,73% from 25,931.08 to 4736.371. The estimated variance of proportional stratified random sampling was reduced by 87.1% from 11,078.84 to 1429.439.

Why did the effect of Finite Population become more significant as the sample size increased?

I believe that there are two reasons for this observation, firstly because it depends on the proportion of the population being sampled. The FPC factor is $\sqrt{\frac{N-n}{N-1}}$, where N is the population size and n the sample size. As n increases, $\frac{N-n}{N-1}$ decreases, thereby reducing the standard error. The more you sample from a finite population, the less variability remains in your data not yet sampled. This reduces the sampling uncertainty. The correction is minimal for small n, but for larger n, the effect would be substantial. Sampling 100 out of the 344 penguins (29%), introduces a much stronger dependence between the draws than sampling only 25 (7%), so the FPC would adjust the variance more. The more of the population you observe, the less remaining randomness there is, hence the growing FPC effect.

The second reason is because one of the species of penguins, Gentoo, are much heavier than the other two species and heavier than the mean of all the species. The sample mean and its variance would heavily depend on how many heavy Gentoo penguins are sampled. The Gentoo Penguins would skew the overall distribution upward. If more Gentoo penguins are sampled, which might randomly happen in a larger sample, the body mass distribution would tighten and reduce the uncorrected variance and magnify the practical effect of the FPC.

Why I found this problem interesting?

I was very surprised to see how the effect of finite population correction increased as I increased my sample size increased. I fully expected to see the difference become negligible. Proportional stratified random sampling had a very powerful effect on reducing the estimated sample variance, compared to simple random sampling. This effect was much larger than I had anticipated.

It was very interesting and a bit challenging to explain and provide reasons for my observations. It was interesting to investigate the effect that the heavier Gentoo penguins might have had on my observations.

What I learned from this Problem?

I learned a lot from this problem, how to calculate the estimated variances and confidence intervals of SRS with and without FPC and proportional stratified random sampling, both mathematically and using R.

I learned how to identify the best method of sampling, because the population is finite, random sampling with FPC would be appropriate. And since the population is divided into different species, stratified random sampling, and specifically proportional allocation would be a good approach to sampling.

I sharpened my programming skills and R and learn how to implement AI assistance. Like how to ask AI to provide certain lines of functions of code, while remaining in control and not allowing AI to write the code completely by itself. Modern development studios, like Cursor, with direct access to the API keys of major AI models, makes this process much easier. You can request a review of you code by your preferred model and can accept or reject suggested changes by the AI in real time.

Estimation of Parameters and Fitting of Probability Distributions

Index of Exercises and Examples

Parameter Estimation

I. Chapter 8, Question 50a (Rice, 2007, p. 323)

Methods of Allocation and Bootstrapping

I. Coded solution for Example A, Chapter 8.4 (Rice, 2007, p. 261)
II. Coded solution for Example C, Chapter 8.4 (Rice, 2007, p. 263)
III. Chapter 8, Question 16a (Rice, 2007, p. 319)
IV. Assignment 4

Method of Maximum Likelihood

I. Chapter 8, Question 16 b and c (Rice, 2007, p. 319)
II. Chapter 8, Question 50 b and c (Rice, 2007, p. 323)
III. Assignment 4

Maximum Likelihood Estimates of Multinomial Cell probabilities.

I. Chapter 8, Question 55a (Rice, 2007, p. 324)

Cramér-Rao Lower Bound

I. Question 7.48 a (George Casella, 2001, p. 364)

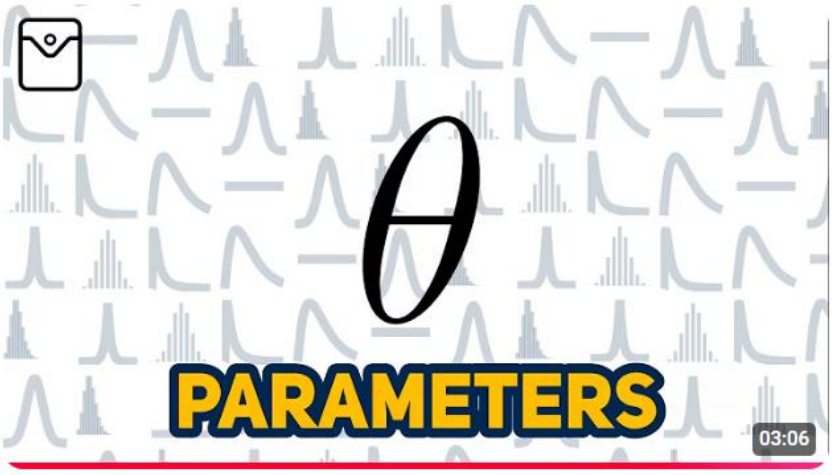
Sufficiency

I. Chapter 8, Question 75 (Rice, 2007, p. 328)
II. Chapter 8, Question 16d (Rice, 2007, p. 319)

Parameter Estimation

Research Process

- I watched a YouTube video by Very Normal from his pocket-stats playlist, that briefly defined and explained statistical parameters.



What are parameters?
10 k kyke • 1 jaar gelede
Very Normal
A brief video on what **parameters** are
Beginner Oorsig

I loved Very Normal's pocket-stats, series. It provides short and concise explanations of all the basic statistical concepts. He explained that there are two ways of understanding parameters. Firstly, it could be seen as a turning knob, that tunes the distribution. Secondly, it could be seen as a value of interest. In most circumstances, the true population parameters are unknown.

Exercises

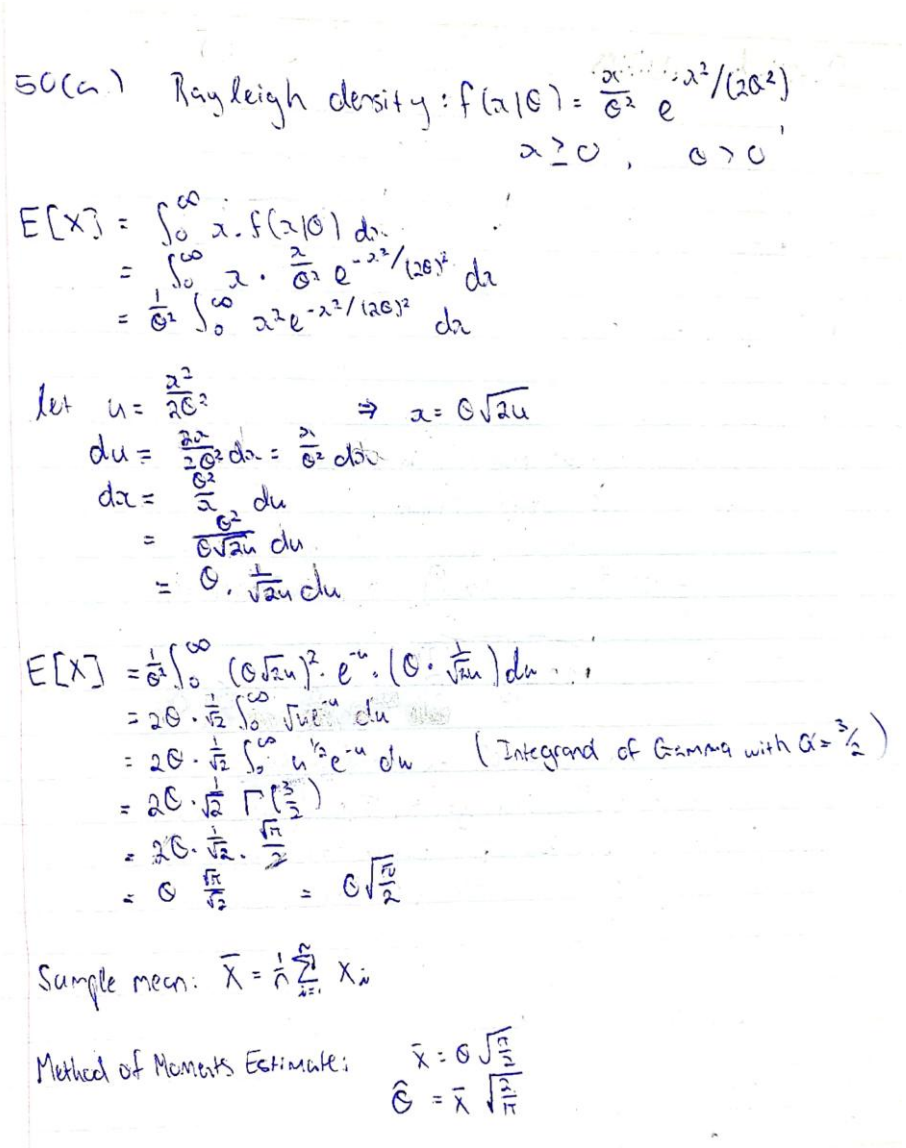
50. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a Rayleigh distribution with parameter $\theta > 0$:

$$f(x|\theta) = \frac{x}{\theta^2} e^{-x^2/(2\theta^2)}, \qquad x \geq 0$$

(This is an alternative parametrization of that of Example A in Section 3.6.2.)

a. Find the method of moments estimate of θ.
b. Find the mle of θ.
c. Find the asymptotic variance of the mle.

FIGURE 81 (RICE, 2007, P. 323)

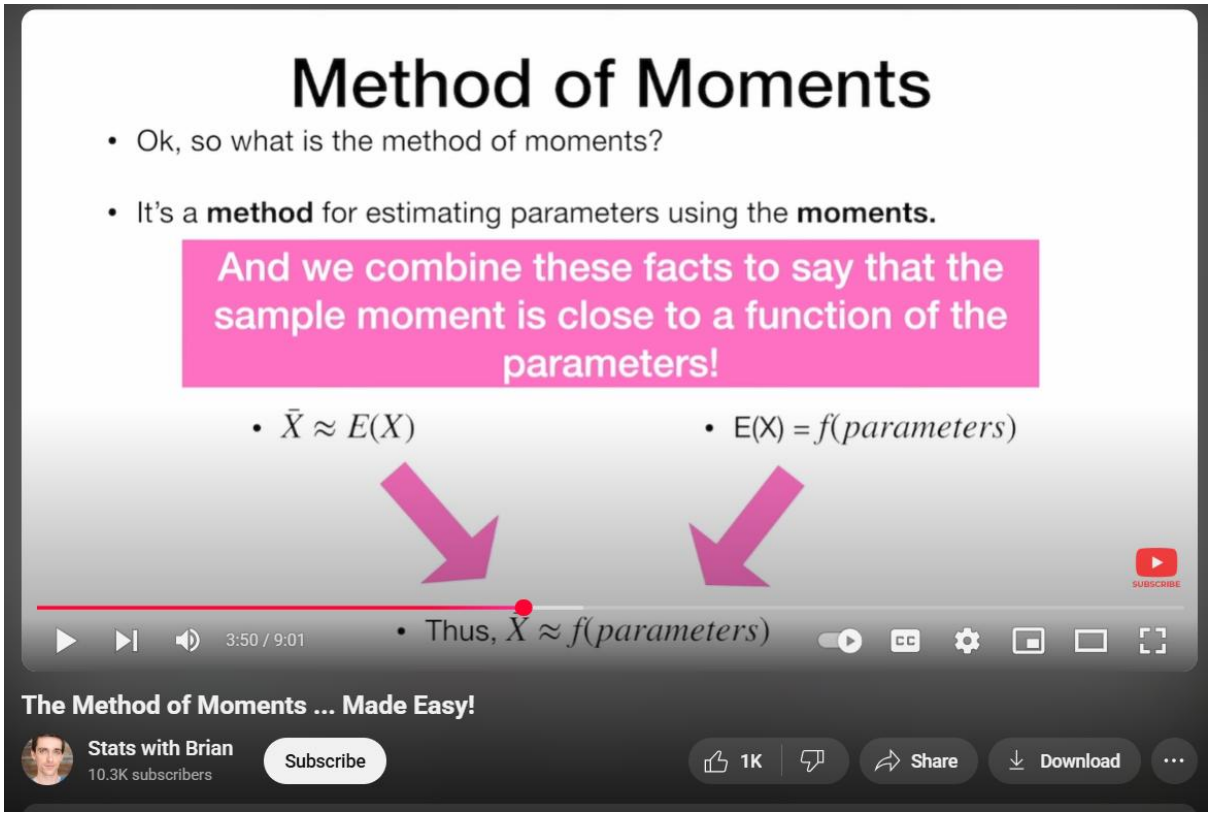


This is my first time working with the Rayleigh distribution and I derived the expected value myself.

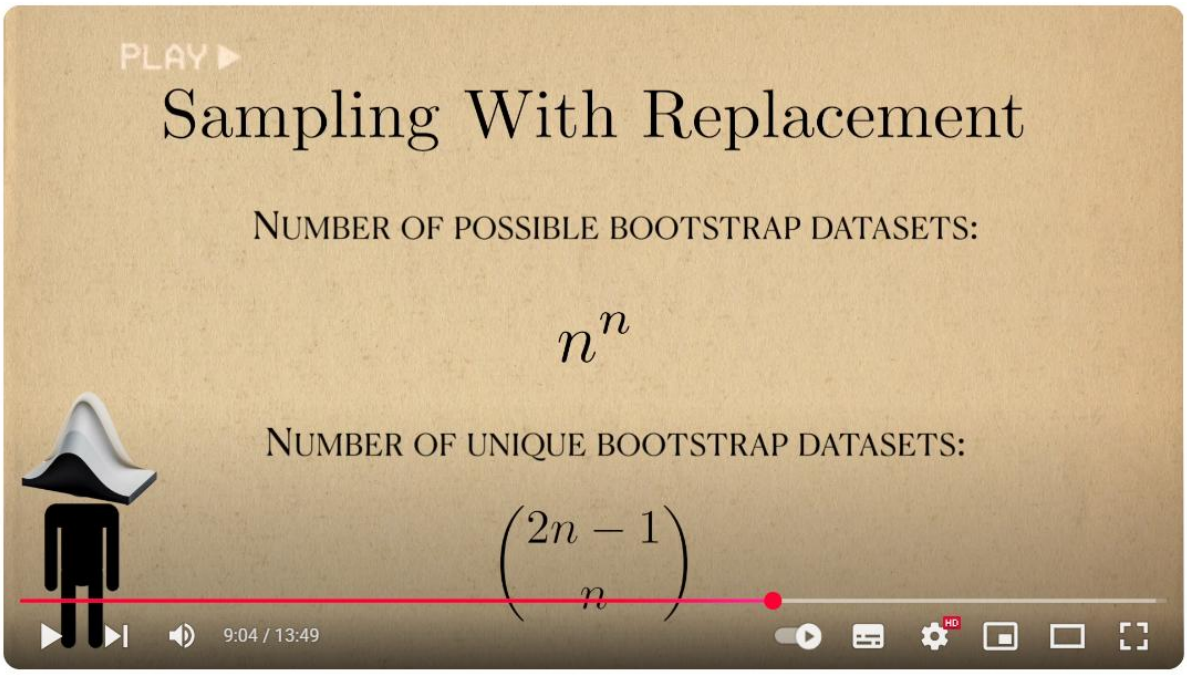
The Method of Moments and Bootstrapping

Research Process

- I watched a YouTube video by Stats with Brian on the Method of Moments.
- I read through the lecture notes of PennState University on bootstrapping.
- I watched a YouTube video by Very Normal on bootstrapping.
- I read an article by Josef Waples on DataCamp which explained the difference between parametric and non-parametric bootstapping.



Brian Explains how the method of moments is used to estimate the population parameters, by matching the sample moments to the corresponding theoretical moments by using a function of parameters. He also explained the relationship between the population moments and the parameters, the population moments depend on the parameters of the probability distribution.



Statistical Inception: The Bootstrap (#SoME3)

Very Normal
97,9 k intekenare
Ingeteken
1,8 k
Deel
Laai af

I learned a lot from this video by Very Normal, including how the number of unique bootstrap datasets can be calculated. He explained how the proof of the bootstrap is quite complicated, and one cannot simply prove the bootstrap estimator equal to the sampling distribution directly.

Definition

The kth moment of a probability law is defined as

$$\mu_k = E(X^k)$$

where X is a random variable following that probability law (of course, this is defined only if the expectation exists). If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from that distribution, the kth sample moment is defined as

$$\widehat{\mu_k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

We can view $\widehat{\mu_k}$ as an estimate of μ_k. The method of moments estimates parameters by finding expressions for them in terms of the lowest possible order moments and then substituting sample moments into the expressions. Suppose, for example, that we wish to

estimate two parameters, θ_1 and θ_2. If θ_1 And θ_2 can be expressed in terms of the first two moments as

$$\theta_1 = f_1(\mu_1, \mu_2)$$

$$\theta_2 = f_2(\mu_1, \mu_2)$$

Then the method of moment estimates are

$$\widehat{\theta_1} = f_1(\widehat{\mu_1}, \widehat{\mu_2})$$

$$\widehat{\theta_2} = f_2(\widehat{\mu_1}, \widehat{\mu_2})$$

The construction of a method of moments estimate involves three basic steps:

1. Calculate low order moments, finding expressions for the moments in terms of the parameters. Typically, the number of low order moments needed will be the same as the number of parameters. We need to compute the theoretical moments $(\mu_1, \mu_2, \dots)$ as functions of the parameters.

2. Invert the expressions found in the preceding step, finding new expressions for the parameters in terms of the moments. We need to solve the parameters in terms of the moments $(\theta_1 = f_1(\mu_1, \mu_2))$

3. Insert the sample moments into the expressions obtained in the second step, thus obtaining estimates of the parameters in terms of the sample moments. We need to replace the theoretical moments with the sample moments to obtain the estimators $(\widehat{\theta_1} = f_1(\widehat{\mu_1}, \widehat{\mu_2}))$

Moments are quantitative measures that describe the shape of a probability distribution. Michael explained in class that the first moment is a measurement of the position or central location of our distribution on the number line. The second moment is a measurement of spread or scale of the distribution around the mean.

Bootstrap

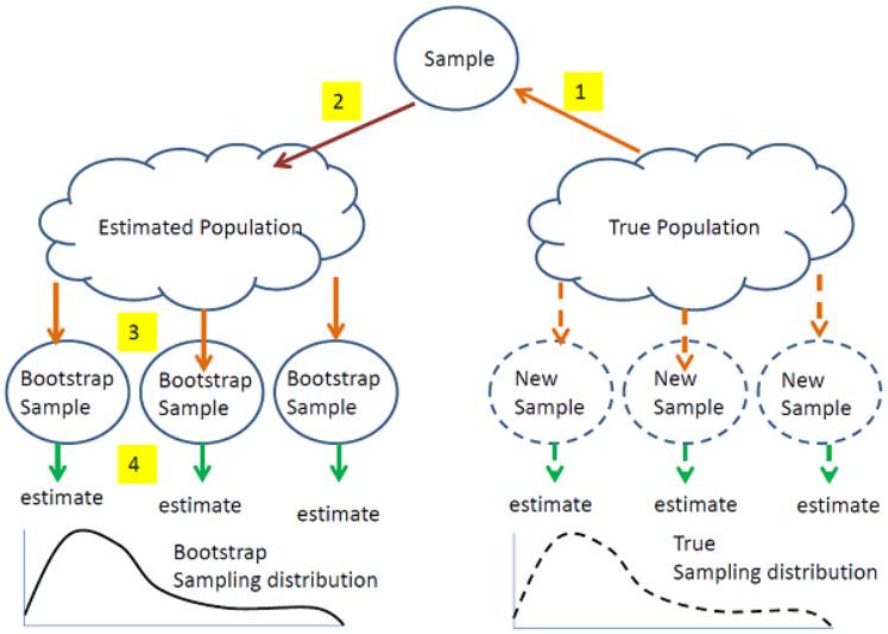


FIGURE 82 ILLUSTRATION FROM PENN STATE LECTURE NOTES

This illustration from the STAT555 lecture notes from the PennState Eberly College of Science, is a great way of illustrating bootstrapping. The observed quantities are denoted by solid curves and unobserved quantities by dashed curves.

Bootstrapping essentially replicates the process we would follow, under perfect circumstances, to form a sampling distribution. We would take as much and as large new samples from our true population, as our hearts desire, and accumulate all the values into the true sampling distribution.

Bootstrapping replicates this process from only a single sample, by allowing for resampling. We can take as many samples as we like, with replacement, the same size as our sample, from that sample, and use their estimates to form a sampling distribution.

DEFINITION

Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is said to be consistent in probability if $\hat{\theta}_n$ converges in probability to θ as n approaches infinity; that is, for any $\epsilon > 0$,

$$P(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \qquad \text{as } n \to \infty$$ ■

FIGURE 83 (RICE, 2007, P. 266)

This definition is already familiar, since it follows from the weak law of large numbers. It has the same underlying conditions and requires the functions relating the estimates to the sample moments, to be continuous. Then the sample moments would converge in probability to the population moments and the estimates would converge to the parameters.

My Understanding of the Method of Moments and Bootstrapping

The **Method of Moments** is a technique to estimate parameters. We solve these unknown parameters by equating sample moments, to their corresponding population moments.

Moments are measurements that describes the shape of a probability distribution. The first moment is a measurement of position or central location on the number line, while the second moment is measurement of spread or scale of the distribution around the mean.

Bootstrap is a resampling method that estimates the sampling distribution of a statistic, by repeatedly resampling from the original data, with replacement.

What is the "Magic" behind the Bootstrap?

The Bootstrap treats our sample as a "population". This allows us to take the same approach to sampling from our population, that we would have, in a perfect world. With a bootstrap we can take as many samples as we can realistically compute from our "population" (which is in fact only the single sample), this can be done on demand and vey fast, via simulation. And all of this only at a fraction of the cost of taking another sample from our true population. This work because our sample should be representative of our population if drawn randomly.

I must differentiate between two distinct methods of the Bootstrap, the Non-Parametric (Classic) Bootstrap and the Parametric Bootstrap.

The **Non-Parametric Bootstrap** is quite simple, we resample directly from the original sample, size n , with replacement, creating new samples of size n.

The **Parametric Bootstrap** assumes a specific parametric distribution. Therefore, we must find the parameter estimates from our original sample. We then sample from a distribution fitted to these estimates, rather than from our sample itself.

The Non-Parametric Bootstrap has the major advantage that we do not have to assume any particular distribution. However, since we are sampling from our original data, rather than from a distribution, we cannot account for values outside of our original data, that may occur on the

tails of a distribution. Therefore, if we can confidently fit our sample to a distribution, Parametric Bootstrapping would be more efficient.

Example A: Poisson Distribution

The first moment for the Poisson Distribution is the parameter $\lambda = E(X)$. The first sample moment is

$$\widehat{\mu_1} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Which is, therefore, the method of moments estimate of λ: $\hat{\lambda} = \bar{X}$.

As a concrete example, let us consider a study done by the National Institute of Science and Technology (Steel et al. 1980). Asbestos fibers on filters were counted as part of a project to develop measurement standards for asbestos concentration. Asbestos dissolved in water was spread on a filter, and 3-mm diameter punches were taken from the filter and mounted on a transmission electron microscope. An operator counted the number of fibers in each of 23 grid squares, yielding the following counts:

31	29	19	18	31	28
34	27	34	30	16	18
26	27	27	18	24	22
28	24	21	17	24	

The Poisson distribution would be a plausible model for describing the variability from grid square to grid square in this situation and could be used to characterize the inherent variability in future measurements. The method of moments estimate of λ is simply the arithmetic mean of the counts listed above, these or $\hat{\lambda} = 24.9$

A standard statistical technique for addressing this question is to derive the sampling distribution of the estimate or an approximation to that distribution. The statistical model stipulates that the individual counts X_i are independent Poisson random variables with parameter λ_0. Letting $S = \sum X_i$, the parameter estimate $\hat{\lambda} = \frac{S}{n}$ is a random variable, the distribution of which is called the sampling distribution. The distribution of the sum of independent Poisson random variables is Poisson distributed, so the distribution of S is Poisson $(n\lambda_0)$. Thus, the probability mass function of $\hat{\lambda}$ is:

$$P(\hat{\lambda} = v)$$

$$= P(\frac{1}{n}\sum_{i=1}^{n} X_i = v)$$

$$= P(s = nv)$$

$$= \frac{(n\lambda_0)^{nv} e^{-n\lambda_0}}{(nv)!}$$

For v such that nv is a nonnegative integer.

Since S is Poisson, its mean and variance are both $n\lambda_0$, so

$$E(\hat{\lambda}) = \frac{1}{n}E(S) = \lambda_0$$

$$Var(\hat{\lambda}) = \frac{1}{n^2}Var(S) = \frac{\lambda_0}{n}$$

Michael covered this example in class and explained that $\bar{X} \sim Poisson$, $S \sim Poisson$, but as $n\lambda_0$ becomes large, the approximate distribution will become approximately normal. The estimator $\frac{1}{n}\sum_{i=1}^{n} X_i$, will also be approximately normally distributed.

Because $E(\hat{\lambda}) = \lambda_0$, we say the estimate is unbiased: the sampling distribution is cantered at λ_0. The standard error of $\hat{\lambda}$ is:

$$\sigma_{\hat{\lambda}} = \sqrt{\frac{\lambda_0}{n}}$$

We don't know the sampling distribution of or the standard error of $\hat{\lambda}$, since it is dependent on λ_0, which is unknown. We can however calculate the estimated standard error of $\hat{\lambda}$ as:

$$s_{\hat{\lambda}} = \sqrt{\frac{\hat{\lambda}}{n}}$$

And

$$s_{\hat{\lambda}} = \sqrt{\frac{24.9}{23}} = 1.04$$

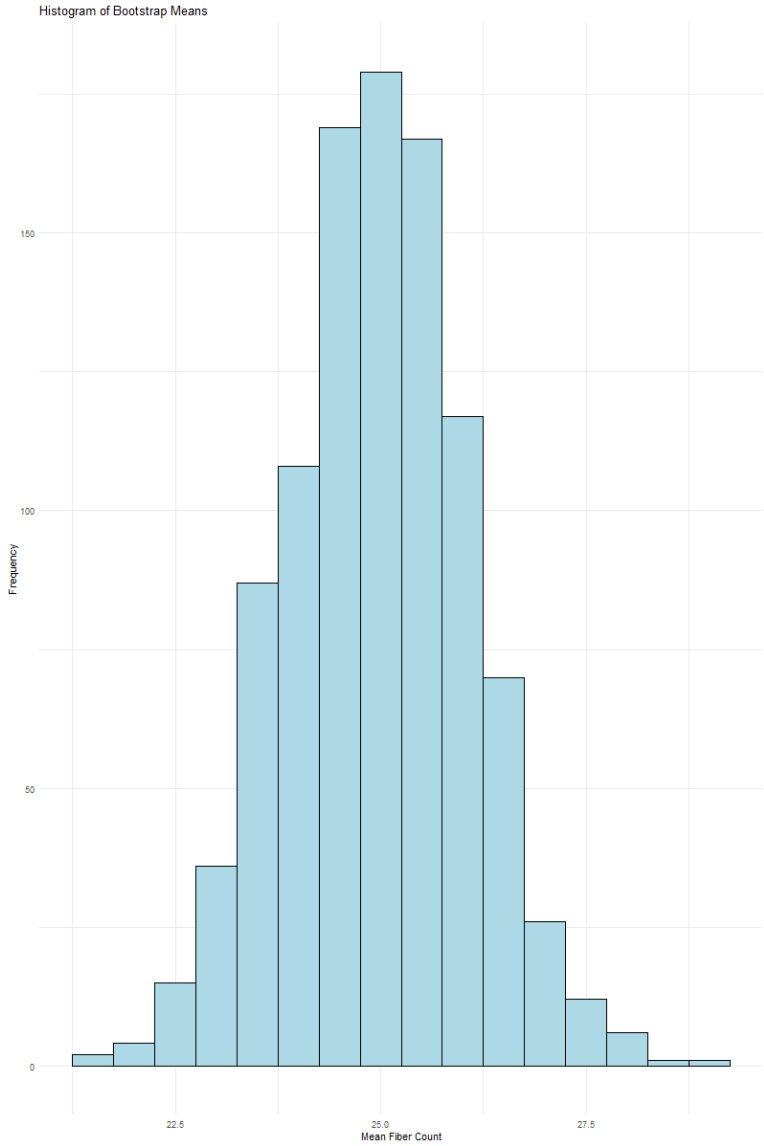
Using this example, I set up a bootstrap of 1000 samples, using R, and found the standard error as 1.102

Poisson Bootstrap.R

```r
fibers <- c(31, 29, 19, 18, 31, 28,
            34, 27, 34, 30, 16, 18,
            26, 27, 27, 18, 24, 22,
            28, 24, 21, 17, 24)


N <- 1000

bootstrap_means <- numeric(N)
n <- length(fibers) # I am not hardcoding the lenght of the sample

for (i in 1:N) {
  bootstrap_sample <- sample(fibers, size = n, replace = TRUE)
  bootstrap_means[i] <- mean(bootstrap_sample)
}

p1 <- ggplot(data.frame(bootstrap_means), aes(x = bootstrap_means)) +
  geom_histogram(binwidth = 0.5, fill = "lightblue", color = "black") +
  theme_minimal() +
  labs(title = "Histogram of Bootstrap Means",
       x = "Mean Fiber Count",
       y = "Frequency")
print(p1)

conf_interval <- quantile(bootstrap_means, probs = c(0.025, 0.975))
print("95% Confidence Interval for the mean:")
print(conf_interval)

bootstrap_se <- sd(bootstrap_means)
print(paste("Bootstrap Standard Error of the Mean:", round(bootstrap_se, 3)))
```

```
package 'ggplot2' was built under R version 4.4.2
> source("c:\\Users\\franc\\PythonProject65\\Poisson Bootstrap.R", encoding = $
[1] "95% Confidence Interval for the mean:"
[1] "95% Confidence Interval for the mean:"
    2.5%    97.5%
    2.5%    97.5%
22.86957 27.13043
[1] "Bootstrap Standard Error of the Mean: 1.102"
[1] "Bootstrap Standard Error of the Mean: 1.102"
>
```

This is an example of a non-parametric bootstrap. The code was written in R in Cursor Studio.



Example C

The first two moments or the gamma distribution are (Derived from the moment generating function)

$$\mu_1 = \frac{\alpha}{\lambda}$$

$$\mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2}$$

To apply the method of moments, we must express α and λ in terms of μ_1 and μ_2. From the second equation

$$\mu_2 = \mu_1^2 + \frac{\mu_1}{\lambda}$$

Or

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2}$$

Also, from the equation for the first moments given here,

$$\alpha = \lambda\mu_1 = \frac{\mu_1^2}{\mu_2 - \mu_1^2} \qquad ,\left(\frac{estimate\ of\ mean}{estimate\ of\ variance}\right)$$

The method of moment estimates are, since $\widehat{\sigma^2} = \widehat{\mu_2} - \widehat{\mu_1^2}$,

$$\hat{\lambda} = \frac{\overline{X}}{\widehat{\sigma^2}}$$

And

$$\hat{\alpha} = \frac{\overline{X^2}}{\widehat{\sigma^2}}$$

For this example, let us consider the fit of amounts of precipitation during 227 storms in Illinois from 1960 to 1964 to a gamma distribution. For these data, $\bar{X} = .224$ and $\widehat{\sigma^2} = .1338$ and therefor $\hat{\alpha} = .375$ and $\hat{\lambda} = 1.674$.

Since it would be difficult to derive the exact forms of the sampling distributions of $\hat{\lambda}$ and $\hat{\alpha}$, because they are each rather complicated functions of the sample values $X_1, X_2, \ldots, X_n$. However, the problem can be approached by simulation (Bootstrap). Imagine if we knew the true values λ_0 and α_0. We could generate many samples of size $n = 277$ from the gamma distribution with these parameter values, and from each of these samples we could calculate estimates of λ and α. The only problem with this idea is that it requires knowing the true parameter values. So, we substitute our estimates of λ and α for the true values; that is, we draw many, many samples of size $n = 277$ from a gamma distribution with parameters $\alpha = .375$ and $\lambda = 1.674$.

The variability shown by the histograms can be summarized by calculating the standard deviations of the 1000 estimates, thus providing estimated standard errors of $\hat{\alpha}$ and $\hat{\lambda}$. To be precise, if the 1000 estimates of α are denoted by $\alpha_i^*, i = 1,2, \ldots, 1000$, $\hat{\alpha}$ and $\hat{\lambda}$.

$$s_{\hat{\alpha}} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\alpha_i^* - \bar{\alpha})\text{^}2}$$

Where $\bar{\alpha}$ is the mean of the 1000 values. The results of this calculation and the corresponding one for $\hat{\lambda}$ are $s_{\hat{\lambda}} = .34$ and $s_{\hat{\alpha}} = .06$

This process is known as a parametric bootstrap. We are not sampling from our observed data, but rather from a known distribution. We sample from a gamma distribution based on our estimates for α and λ. After programming the bootstrap for this example in R, I found $s_{\hat{\lambda}} = .3356$ and $s_{\hat{\alpha}} = .0626$

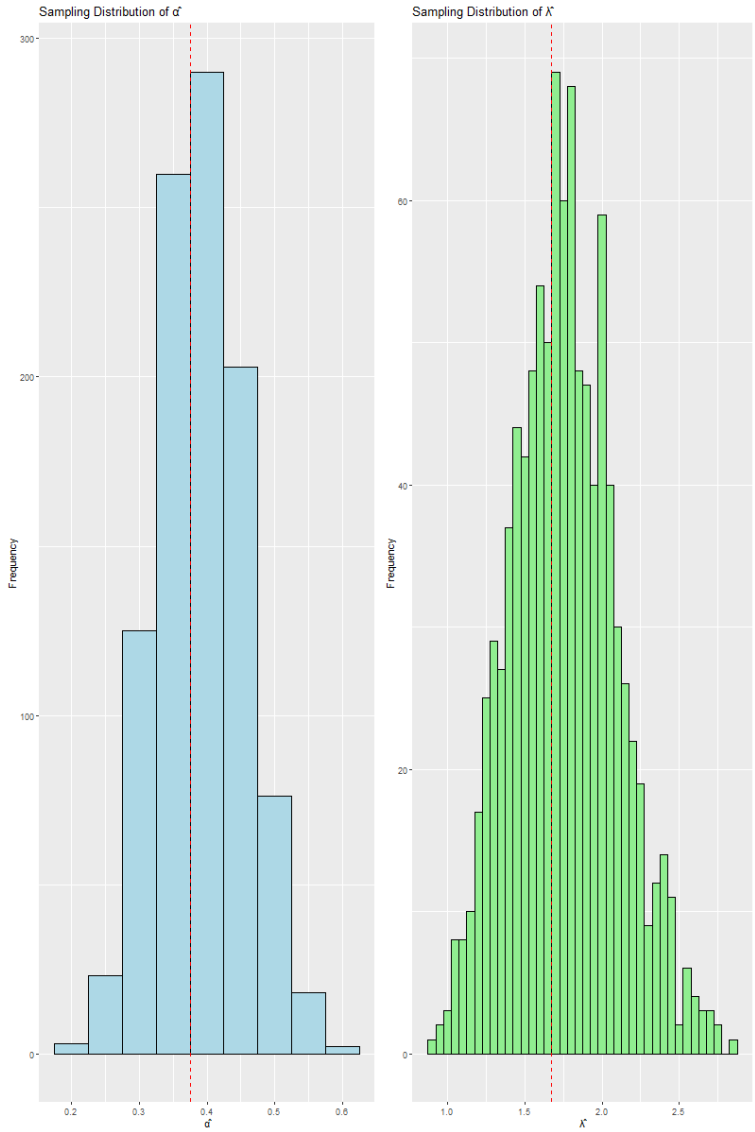
Gamma Bootstrap.R

```r
set.seed(123)

alpha_true <- 0.375
lambda_true <- 1.674
n <- 227
n_sim <- 1000

gamma_moment_estimator <- function(x) {
  x_bar <- mean(x)
  s2 <- var(x)
  alpha_hat <- x_bar^2 / s2
  lambda_hat <- x_bar / s2
  return(c(alpha_hat, lambda_hat))
}

alpha_hats <- numeric(n_sim)
lambda_hats <- numeric(n_sim)

for (i in 1:n_sim) {
  sample_data <- rgamma(n, shape = alpha_true, rate = lambda_true)
  estimates <- gamma_moment_estimator(sample_data)
  alpha_hats[i] <- estimates[1]
  lambda_hats[i] <- estimates[2]
}

se_alpha <- sqrt(mean((alpha_hats - mean(alpha_hats))^2))
se_lambda <- sqrt(mean((lambda_hats - mean(lambda_hats))^2))
```

```r
library(ggplot2)
library(gridExtra)

p1 <- ggplot(data.frame(alpha_hats), aes(x = alpha_hats)) +
  geom_histogram(binwidth = 0.05, fill = "lightblue", color = "black") +
  geom_vline(xintercept = alpha_true, color = "red", linetype = "dashed") +
  labs(title = "Sampling Distribution of α", x = "α", y = "Frequency")

p2 <- ggplot(data.frame(lambda_hats), aes(x = lambda_hats)) +
  geom_histogram(binwidth = 0.05, fill = "lightgreen", color = "black") +
  geom_vline(xintercept = lambda_true, color = "red", linetype = "dashed") +
  labs(title = "Sampling Distribution of λ", x = "λ", y = "Frequency")

cat("Estimated standard error of α̂:", se_alpha, "\n")
cat("Estimated standard error of λ̂:", se_lambda, "\n")
cat("Mean of α̂estimates:", mean(alpha_hats), "\n")
cat("Mean of λ̂estimates:", mean(lambda_hats), "\n")

grid.arrange(p1, p2, nrow = 1)
```

```
> source("c:\\Users\\franc\\PythonProject64\\Gamma Bootstrap.R", encoding = "U$
Estimated standard error of α̂: 0.06258315
Estimated standard error of λ̂: 0.3356083
Mean of α̂estimates: 0.3906373
Mean of λ̂estimates: 1.758075
```

Though not explicitly AI generated, this code was written in Curser, a coding studio with integrated AI text prediction and optimisation recommendations. All code was myself and I implemented some of the syntax/ optimisation recommendations. For example, if I type lambda^hat it automatically suggests $\hat{\lambda}$, and I don't have to bother with LyTex codes. This is an example of a parametric bootstrap.


Sampling Distribution of $\hat{\alpha}$
Frequency
300
200
100
0
0.2
0.3
0.4
0.5
0.6
$\hat{\alpha}$
Sampling Distribution of $\hat{\lambda}$
Frequency
60
40
20
0
1.0
1.5
2.0
2.5
$\hat{\lambda}$

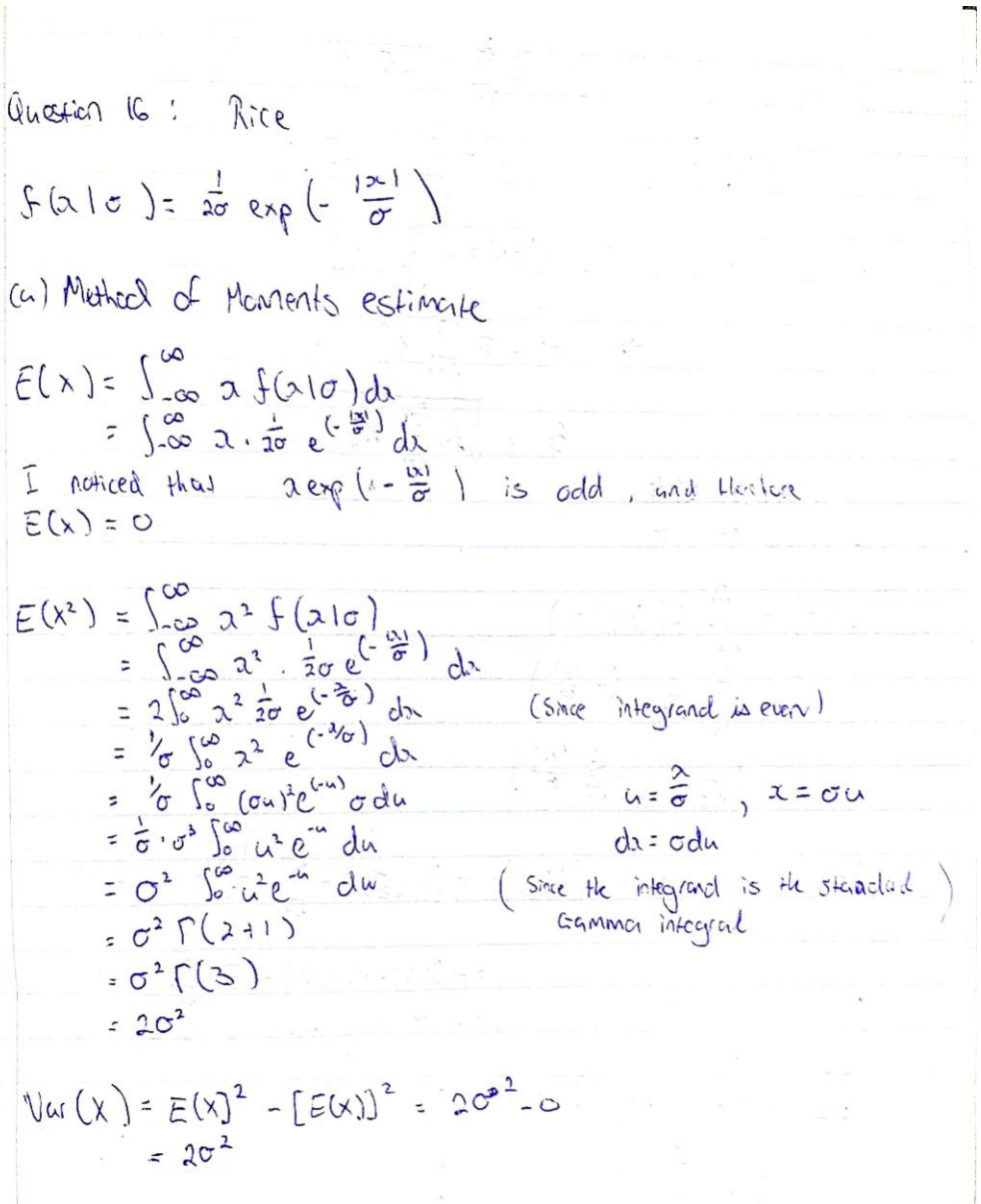
Exercises

16. Consider an i.i.d. sample of random variables with density function

$$f(x|\sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right)$$

a. Find the method of moments estimate of σ.
b. Find the maximum likelihood estimate of σ.
c. Find the asymptotic variance of the mle.
d. Find a sufficient statistic for σ.

FIGURE 84 (RICE, 2007, P. 319)



Question 16 : Rice

$f(x|\sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right)$

(a) Method of Moments estimate

$E(x) = \int_{-\infty}^{\infty} x f(x|\sigma)\,dx$

$= \int_{-\infty}^{\infty} x \cdot \frac{1}{2\sigma} e^{\left(-\frac{|x|}{\sigma}\right)} dx$

I noticed that $x \exp\left(-\frac{|x|}{\sigma}\right)$ is odd, and therefore

$E(x) = 0$

$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x|\sigma)$

$= \int_{-\infty}^{\infty} x^2 \cdot \frac{1}{2\sigma} e^{\left(-\frac{|x|}{\sigma}\right)} dx$

$= 2\int_{0}^{\infty} x^2 \frac{1}{2\sigma} e^{\left(-\frac{x}{\sigma}\right)} dx$ (Since integrand is even)

$= \frac{1}{\sigma} \int_{0}^{\infty} x^2 e^{\left(-\frac{x}{\sigma}\right)} dx$

$= \frac{1}{\sigma} \int_{0}^{\infty} (\sigma u)^2 e^{(-u)} \sigma\,du$ $u = \frac{x}{\sigma}$, $x = \sigma u$

$= \frac{1}{\sigma} \cdot \sigma^3 \int_{0}^{\infty} u^2 e^{-u}\,du$ $dx = \sigma\,du$

$= \sigma^2 \int_{0}^{\infty} u^2 e^{-u}\,du$ (Since the integrand is the standard Gamma integral)

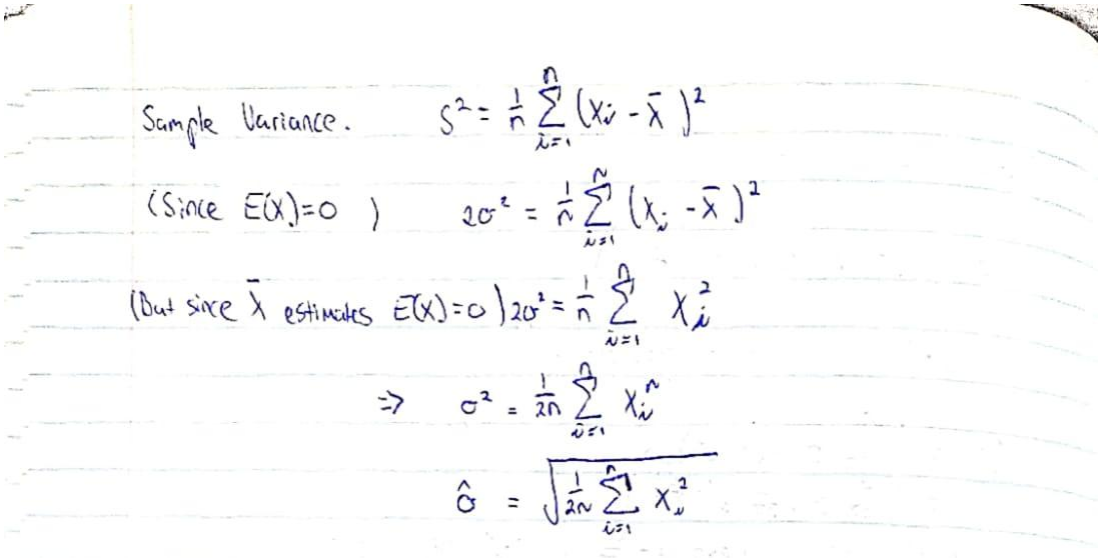
$= \sigma^2 \Gamma(2+1)$

$= \sigma^2 \Gamma(3)$

$= 2\sigma^2$

$Var(X) = E(X)^2 - [E(X)]^2 = 2\sigma^2 - 0$

$= 2\sigma^2$



When integrating to calculate moments, keeping an eye out for properties of odd/even functions can save you a lot of work. I always like to use the standard gamma integral as an easy shortcut to solve integrals such as these

Reflection

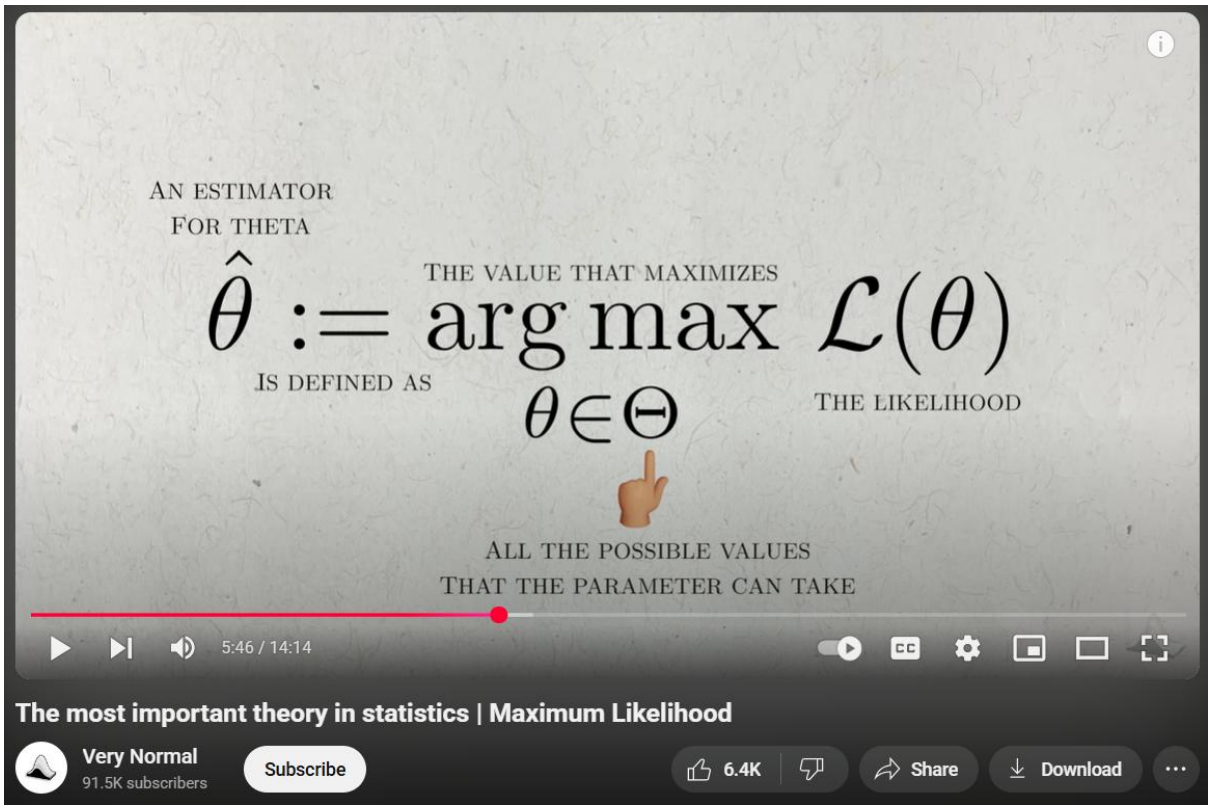
This was probably my favourite section this far. I found the Bootstrap fascinating and spend a lot of time trying to understand why it works like “magic”. I believe I do understand, and have explained it in my summary.

This section involved a lot of coding, which I enjoyed a lot. At the time of writing, I am running a free trial of Cursor studio, and it is honestly crazy. It uses the API keys to all major AI models, allowing for text predictions, and the AI to make changes directly to your code files. It is probably not the best to use if you are trying to learn code. But once you are already a proficient coder, it can increase your efficiency exponentially. You can focus on how you want to approach and solve the problem and lay out the structure. You don’t have to bother with syntax, or even tasks like filling out plots or vectors, you just check it is good and press tab.

The Method of Maximum Likelihood

Research Process

- I consulted my (Rice, 2007) textbook for all relevant theory, theorems and proofs.
- I watched a YouTube video by Very Normal about the method of maximum likelihood.
- I read an article on gregorygunderson.com about the proof for asymptotic normality.
- I read the Wikipedia article about Fisher Information.
- I read an article on StatLec.com about Slutsky's theorem.
- I read the MIT OpenCourseWare lecture notes on maximum likelihood of large sample theory.
- I read through MIT OpenCourseWare lecture 3, on Properties of MLE: consistency, asymptotic normality. Fisher information.



This was a great video by Very Normal. He explained how the MLE is consistent estimator, because it gets closer to the true parameter, as the sample size grows to infinity. Consistency is an asymptotic property.
Another Interesting thing I learned from this video was, that the summary of different regression models in R, like the proportional hazard, logistic regression and mixed defect model, uses MLE the obtain the t, or z-values.

Definition

Suppose that random variables $X_1, \ldots, X_n$ have a joint density or frequency function $f(x_1, x_2, \ldots., x_n|\theta)$. Given observed values $X_i = x_i$, where $i = 1, \ldots, n$, the likelihood of θ as a function of $x_1, x_2, \ldots, x_n$ is defined as

$$lik(\theta) = f(x_1, x_2, \ldots., x_n|\theta)$$

If the distribution is discrete, so that f is a frequency function, the likelihood function gives the probability of observing the given data as a function of the parameter θ. The maximum likelihood estimate (mle) of θ is that value of θ that maximizes the likelihood – that is, makes the observed data "most probable" or "most likely". This paragraph from (Rice, 2007, p. 267) could be described more simply. The maximum likelihood estimation (MLE) finds the parameter values that makes our observed data, most probable or "likely" to occur.

If the X_i are assumed to be i.i.d., their joint density is the product of the marginal densities, and the likelihood is

$$lik(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

Rather than maximizing the likelihood itself, it is usually easier to maximize its natural logarithm (which is equivalent since the logarithm is a monotonic function). For an i.i.d. sample, the log likelihood is

$$l(\theta) = \sum_{i=1}^{n} \log [f(X_i|\theta)]$$

Large Sample Theory for Maximum Likelihood Estimates

THEOREM A

Under appropriate smoothness conditions on f, the mle from an i.i.d. sample is consistent.

Proof

The following is merely a sketch of the proof. Consider maximizing

$$\frac{1}{n}l(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta)$$

As n tends to infinity, the law of large numbers implies that

$$\frac{1}{n}l(\theta) \to E \log f(X|\theta)$$
$$= \int \log f(x|\theta) f(x|\theta_0)\, dx$$

It is thus plausible that for large n, the θ that maximizes $l(\theta)$ should be close to the θ that maximizes $E \log f(X|\theta)$. (An involved argument is necessary to establish this.) To maximize $E \log f(X|\theta)$, we consider its derivative:

$$\frac{\partial}{\partial\theta}\int \log f(x|\theta) f(x|\theta_0)\, dx = \int \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)} f(x|\theta_0)\, dx$$

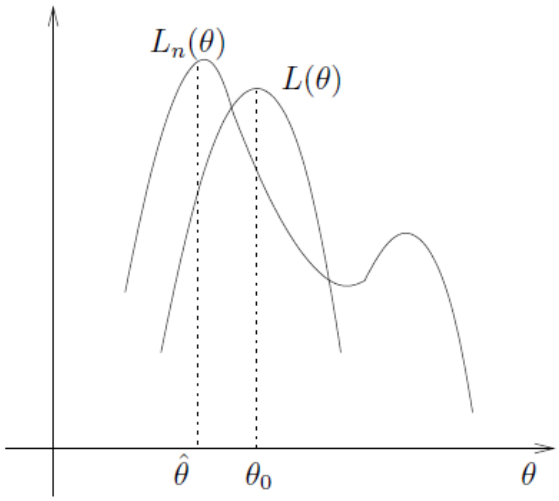
If $\theta = \theta_0$, this equation becomes

$$\int \frac{\partial}{\partial\theta} f(x|\theta_0)\, dx = \frac{\partial}{\partial\theta}\int f(x|\theta_0)\, dx = \frac{\partial}{\partial\theta}(1) = 0$$

which shows that θ_0 is a stationary point and hopefully a maximum. Note that we have interchanged differentiation and integration and that the assumption of smoothness on f must be strong enough to justify this. ■

FIGURE 85 (RICE, 2007, P. 276)

Theorem A is the consistency property of MLE. An estimate $\hat{\theta}$ is consistent if $\hat{\theta}$ converges in probability, as the sample tends to infinity. θ_0 is the true, but unknown parameter of distribution of the sample.



We can observe that θ_0 is indeed a stationary point of $L(\theta)$

FIGURE 86 MIT OPENCOURSEWARE

In Figure 83, we have:

1. $\hat{\theta}$ is the maximiser of $L_n(\theta)$ (by definition).
2. θ_0 is the maximiser of $L(\theta)$ (by Lemma).
3. $\forall\theta$ we have $L_n(\theta) \to L(\theta)$ by LLN

Therefore, since two functions L_n and L are getting closer, the points of maximum should also get closer which exactly means that $\hat{\theta} \to \theta_0$.

LEMMA A

Define $I(\theta)$ by

$$I(\theta) = E\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2$$

Under appropriate smoothness conditions on f, $I(\theta)$ may also be expressed as

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$$

Proof

First, we observe that since $\int f(x|\theta)\,dx = 1$,

$$\frac{\partial}{\partial\theta}\int f(x|\theta)\,dx = 0$$

Combining this with the identity

$$\frac{\partial}{\partial\theta}f(x|\theta) = \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)$$

we have

$$0 = \frac{\partial}{\partial\theta}\int f(x|\theta)\,dx = \int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\,dx$$

where we have interchanged differentiation and integration (some assumptions must be made in order to do this). Taking second derivatives of the preceding expressions, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\,dx \\ &= \int \left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] f(x|\theta)\,dx + \int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]^2 f(x|\theta)\,dx \end{aligned}$$

From this, the desired result follows. ■

THEOREM B

Under smoothness conditions on f, the probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0)$ tends to a standard normal distribution.

Proof

The following is merely a sketch of the proof; the details of the argument are beyond the scope of this book. From a Taylor series expansion,

$$0 = l'(\hat{\theta}) \approx l'(\theta_0) + (\hat{\theta}-\theta_0)l''(\theta_0)$$

$$(\hat{\theta}-\theta_0) \approx \frac{-l'(\theta_0)}{l''(\theta_0)}$$

$$n^{1/2}(\hat{\theta}-\theta_0) \approx \frac{-n^{-1/2}l'(\theta_0)}{n^{-1}l''(\theta_0)}$$

First, we consider the numerator of this last expression. Its expectation is

$$\begin{aligned} E[n^{-1/2}l'(\theta_0)] &= n^{-1/2}\sum_{i=1}^{n} E\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\right] \\ &= 0 \end{aligned}$$

as in Theorem A. Its variance is

$$\begin{aligned}\text{Var}[n^{-1/2}l'(\theta_0)] &= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\right]^2 \\ &= I(\theta_0)\end{aligned}$$

Next, we consider the denominator:

$$\frac{1}{n}l''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x_i|\theta_0)$$

By the law of large numbers, the latter expression converges to

$$E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta_0)\right] = -I(\theta_0)$$

from Lemma A.

We thus have

$$n^{1/2}(\hat{\theta}-\theta_0) \approx \frac{n^{-1/2}l'(\theta_0)}{I(\theta_0)}$$

Therefore,

$$E[n^{1/2}(\hat{\theta}-\theta_0)] \approx 0$$

Furthermore,

$$\begin{aligned}\text{Var}[n^{1/2}(\hat{\theta}-\theta_0)] &\approx \frac{I(\theta_0)}{I^2(\theta_0)} \\ &= \frac{1}{I(\theta_0)}\end{aligned}$$

and thus

$$\text{Var}(\hat{\theta}-\theta_0) \approx \frac{1}{nI(\theta_0)}$$

The central limit theorem may be applied to $l'(\theta_0)$, which is a sum of i.i.d. random variables:

$$l'(\theta_0) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta_0}\log f(X_i|\theta) \quad \blacksquare$$

Asymptotic normality: Assume $\widehat{\theta_N} \xrightarrow{p} \theta_0$, with $\theta_0 \in \Theta$ and that other regularity conditions hold. Then

$$\sqrt{N}(\widehat{\theta_N} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$$

Where $I(\theta_0)$ is the Fisher Information

Large parts of this proof were derived from gregorygunderson.com

Asymptotic Normality

If I am working with a statistical model $\mathbb{P}_\theta$ and a random variable $X \sim \mathbb{P}_{\theta_0}$, where θ_0 is the true parameter value. Then the Maximum Likelihood estimate is given by $\hat{\theta}_N$, with sample size of N.

Let $X = (X_1, \ldots, X_N)$ be a finite sample, where $x \sim \mathbb{P}_{\theta_0}$, with $\theta_0 \in \Theta$ being the true, unknown parameter.

If asymptotic normality holds, then efficiency falls out, since it immediately implies

$$\hat{\theta}_N \xrightarrow{d} N(\theta_0, I_N(\theta_0)^{-1})$$

where $I_N(\theta)$ is the Fisher information for X and $I(\theta)$ is the Fisher Information for a single $X_n \in X$. Therefore $I_N(\theta) = NI(\theta)$, (Provided the data are i.i.d)

Fisher Information: A Method of Measuring the amount of information that an observable random variable X, carries about an unknown parameter θ. It quantifies the sensitivity of the likelihood, (or log-likelihood) to changes in θ; The sharper the likelihood peak around the true value, the more information you have about θ.

Here I have defined the normalized log-likelihood function and its first and second derivatives, with respect to θ

$$L_N(\theta) = \frac{1}{N} \log f_X(x,\theta)$$

$$L_N'(\theta) = \frac{\partial}{\partial\theta}\left(\frac{1}{N} \log f_X(x,\theta)\right)$$

$$L''N(\theta) = \frac{\partial^2}{\partial\theta^2}\left(\frac{1}{N} \log f_X(x,\theta)\right)$$

From the definition of MLE, the MLE is the maximum of the log likelihood function and therefore,

$$\hat{\theta}_N = \underset{\theta\in\Theta}{\operatorname{argmax}}\, f_X(x,\theta) \Rightarrow L_N'(\hat{\theta}_N) = 0$$

Mean Value theorem: Let f be a continuous function on the closed interval $[a,b]$ and differentiable on the open interval. Then there exists a point $c \in (a,b)$ such that

$$f'(c) = \frac{f(a) - f(b)}{a-b}$$

Where $f = L'N$, $a = \hat{\theta}$ and $b = \theta_0$, Then for some point $c = \tilde{\theta} \in (\hat{\theta}_N, \theta_0)$, we have

$$L'_N(\hat{\theta}_N) = L_N'(\theta_0) + L''_N(\tilde{\theta})(\hat{\theta}_N - \theta_0)$$

From definition $L'_N(\hat{\theta}_N) = 0$ (since the maximum likelihood occurs at a stationary point where the first derivative is zero)

Where $L_N' = \frac{L_N'(\tilde{\theta}) - L_N'(\theta_0)}{\tilde{\theta} - \theta_0}$ *from the mean value theorem*

$$0 = L'_N(\theta_0) + L''_N(\tilde{\theta})(\hat{\theta}_N - \theta_0)$$

$$\hat{\theta} - \theta_0 = -\frac{L'_N(\theta_0)}{L''_N(\tilde{\theta})} \Rightarrow \sqrt{N}(\hat{\theta}_N - \theta_0) = -\frac{\sqrt{N}\, L'_N(\theta_0)}{L''_N(\tilde{\theta})}$$

* We can show that the denominator converges in distribution to a normal distribution from the Central Limit Theorem, and the denominator converges in probability to a constant value using the Weak Law of Large Numbers.

Slutsky's theorem: Concerns the convergence in distribution of the transformation of two sequences of random vectors, one converging in distribution and the other one converging in probability to a constant.

Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables

$X_n \xrightarrow{d} X$ (converges in distribution)

$Y_n \xrightarrow{p} c$ (converges in probability to constant c.

Then $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$

From Slutsky's theorem, for the numerator, by the linearity of differentiation and the log of products, we have:

$$\begin{aligned}
\sqrt{N}\, L'_N(\theta_0) &= \sqrt{N}\left(\frac{1}{N}\left[\frac{\partial}{\partial\theta}\log f_X(X, \theta_0)\right]\right) \\
&= \sqrt{N}\left(\frac{1}{N}\left[\frac{\partial}{\partial\theta}\log \prod_{n=1}^{N} f_X(x_n, \theta_0)\right]\right) \\
&= \sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N}\left[\frac{\partial}{\partial\theta}\log f_X(x_n, \theta_0)\right]\right) \\
&= \sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N}\left[\frac{\partial}{\partial\theta}\log f_X(x_n, \theta_0)\right] - E\left[\frac{\partial}{\partial\theta}\log f_X(x_i, \theta)\right]\right)
\end{aligned}$$

From the last line, we use the property that the expected value of the derivative of the log function is zero

$$E\left[\frac{\partial}{\partial\theta}\log f_x(x_1,\theta)\right]=0$$

Using the CLT: $\sqrt{N}\,L'_N(\theta_0)\xrightarrow{d}N\left(0,\,V\left[\frac{\partial}{\partial\theta}\log f_x(x_1,\theta_0)\right]\right)$

(The variance of the Fisher information for a single variable)

$$V\left[\frac{\partial}{\partial\theta}\log f_x(x_1,\theta)\right]=E\left[\left(\frac{\partial}{\partial\theta}\log f_x(x_1,\theta_0)\right)^2\right]-\left(E\left[\frac{\partial}{\partial\theta}\log f_x(x_1;\theta_0)\right]\right)^2$$
$$=I(\theta_0)$$

For the denominator, I can use the Weak Law of Large Numbers, for any θ,

$$L''_N(\theta)=\frac{1}{N}\left(\frac{\partial^2}{\partial\theta^2}\log f_x(X,\theta)\right)$$
$$=\frac{1}{N}\left(\frac{\partial^2}{\partial\theta^2}\log\prod_{n=1}^{N}f_x(x_n,\theta)\right)$$
$$=\frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial^2}{\partial\theta^2}\log f_x(x_n,\theta)\right)$$

(convergence in probability) $\xrightarrow{p}E\left[\frac{\partial^2}{\partial\theta^2}\log f_x(x_1;\theta)\right]$

* Note that $\bar{\theta}\in(\hat{\theta}_N,\theta_0)$ by construction, and I have assumed that $\hat{\theta}_N\xrightarrow{p}\theta_0$

Finally, $L''_N(\bar{\theta})\xrightarrow{p}E\left[\frac{\partial^2}{\partial\theta^2}\log f_x(x_1;\theta_0)\right]=-I(\theta_0)$

So $\sqrt{N}(\hat{\theta}_N-\theta_0)=-\frac{\sqrt{N}\,L'_N(\theta_0)}{L''_N(\bar{\theta})}$

$$Var\left(\sqrt{N}(\hat{\theta}_N-\theta_0)\right)=Var\left(-\frac{\sqrt{N}\,L'_N(\theta_0)}{L''_N(\bar{\theta})}\right)$$
$$=Var\left(-\frac{\sqrt{N}\,L'_N(\theta_0)}{-I(\theta_0)}\right)$$
$$=Var\left(\frac{\sqrt{N}\,L'_N(\theta_0)}{I(\theta_0)}\right)$$

(constant θ_0 have no influence on variance)

$$= \frac{1}{[I(\theta_0)]^2} \operatorname{Var}\left(\sqrt{N}\, L'_N(\theta_0)\right)$$

$$N \operatorname{Var}(\hat{\theta}_N) = \frac{1}{I(\theta_0)}$$

$$\operatorname{Var}(\hat{\theta}_N) = \frac{1}{N I(\theta_0)}$$

$$E\left(\sqrt{N}\,(\hat{\theta}_N - \theta_0)\right) = E\left(\frac{\sqrt{N}\, L'_N(\theta_0)}{L''_N(\tilde{\theta})}\right)$$

$$= \frac{1}{L''_N} E\left(\sqrt{N}\, L'_N(\theta_0)\right)$$

$$= 0$$

Therefore $E(\sqrt{N}\,\hat{\theta}_N) - E(\sqrt{N}\,\theta_0) = 0$

$$E(\hat{\theta}_N) = \theta_0$$

In Summary, I have shown that

$$\sqrt{N}\, L'_N(\theta_0) \xrightarrow{d} N(0, I(\theta_0))$$

$$L''_N(\tilde{\theta}) \xrightarrow{p} -I(\theta_0)$$

$$\hat{\theta}_N \sim N\left(\theta_0, \frac{1}{N I(\theta_0)}\right)$$

$$\widehat{\theta_N} \xrightarrow{d} N(\theta_0, I_N(\theta_0)^{-1}$$

The results prove the asymptotic normality of the maximum likelihood estimate. As $N \to \infty$.

Not only does the estimator converge to the unknown parameter, but it converges fast enough at rate $\frac{1}{\sqrt{n}}$

Maximum Likelihood Estimates of Multinomial Cell probabilities.

Suppose that $X_1, \dots, X_m$, the counts in cell 1, … m, follow a multinomial distribution with total count of n and cell probabilities $p_1, \dots, p_m$. The joint frequency function of $X_1, \dots, X_m$ is:

$$f(x_1, \dots, x_m | p_1, \dots, p_m) = \frac{n!}{\prod_{i=1}^{n!} x_i!} \cdot \prod_{i=1}^{m} p_i^{x_i}$$

From the joint frequency function, the log likelihood is:

$$l(p_1, \dots, p_m) = \log n! - \sum_{i=1}^{m} \log x_i! + \sum_{i=1}^{m} x_i \log p_i$$

After introducing a Lagrange multiplier and maximizing the likelihood subject to the constraint. We set the partial derivatives equal to zero, and find the following system of equations:

$$\widehat{p_j} = -\frac{x_j}{\lambda}, \qquad j = 1, \dots, m$$

Therefore,

$$\widehat{p_j} = \frac{x_j}{n}$$

However, the multinomial cell probabilities are often functions of other unknown parameters θ; that is, $p_i(\theta)$. In such cases, the log likelihood of θ is:

$$l(\theta) = \log n! - \sum_{i=1}^{m} \log x_i! + \sum_{i=1}^{m} x_i \log p_i(\theta)$$

Confidence Intervals for Maximum Likelihood Estimates

The maximum likelihood estimates of μ and σ^2 from an i.i.d. normal sample are

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})\text{^}2$$

To find the confidence interval for μ, we will use a t-distribution with $n-1$ degrees of freedom. Because our σ^2 is unknown, we must estimate it by S^2, where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})\text{^}2$. The standardised statistic $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$ therefore follows a t_{n-1} distribution. And we find our confidence intervals for μ as:

$$P\left(-t_{n-1}\left(\frac{\alpha}{2}\right) \leq \frac{\sqrt{n}(\bar{X}-\mu)}{S} \leq t_{n-1}\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha$$

$$\bar{X} \pm \frac{S}{\sqrt{n}} \cdot t_{n-1}\left(\frac{\alpha}{2}\right)$$

The interval is symmetric about $\bar{X}$. The interval is random, since it is dependent of $\bar{X}$ and S. The center is at random point $\bar{X}$ and the width is proportional to S, which is also random.

To find the confidence interval for σ^2, we will use a Chi-squared distribution with $n-1$ degrees of freedom. Since:

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

Let $\chi^2_m(\alpha)$ denote the point beyond which the chi-square distribution with m degrees of freedom have probability α. Then, by definition:

$$P\left(\chi^2_{n-1}\left(\frac{(1-\alpha)}{2}\right) \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{n-1}\left(\frac{\alpha}{2}\right)\right) = 1-\alpha$$

Therefore, a $100(1-\alpha)\%$ confidence interval for σ^2 is:

$$\left(\frac{n\hat{\sigma}^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{n\hat{\sigma}^2}{\chi^2_{n-1}\left(\frac{(1-\alpha)}{2}\right)}\right)$$

However, this interval is not symmetric about $\hat{\sigma}^2$

Approximate Confidence Intervals Using Large Sample Theory.

From the property of asymptotic normality of MLE's, we know that for large n, the distribution of $\sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0)$ is approximately normal (θ_0 is the "true" parameter). We can leverage the symmetry of the normal distribution:

$$P\left(-z\left(\frac{\alpha}{2}\right) \leq \sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0) \leq z\left(\frac{\alpha}{2}\right)\right) = 1-\alpha$$

Here, $z\left(\frac{\alpha}{2}\right)$ is the critical value such that $P\left(Z \geq z\left(\frac{\alpha}{2}\right)\right) = \frac{\alpha}{2}\ for\ Z \sim N(0,1)$.

Then the approximate $100(1-\alpha)\%$ confidence interval for θ is:

$$\hat{\theta} \pm z\left(\frac{\alpha}{2}\right) \cdot \frac{1}{\sqrt{nI(\widehat{\theta})}}$$

My Understanding of the Maximum Likelihood Estimate

The Maximum Likelihood Estimator (MLE) is our second method for estimating the parameters of a statistical model. It is our fourth major approach to create sample distributions learned it this module. The MLE finds the parameter values that maximizes the likelihood (make it most probable) of observing the given sample data.

The property of consistency states, that the parameter estimate would converge in probability to the true, but unknown population parameter, as the sample tends to infinity.

Another key property of the MLE is asymptotic normality. As the sample size tends to infinity, the MLE will converge to a normal distribution around the true parameter value.

For my preparation for my interview, I have summarised the steps of calculating the Maximum Likelihood Estimate as follows:

1. **Formulate the Likelihood Function $L(\theta|X)$**
   Write the joint pdf of the observed sample data as a function of unknown parameters.
2. **Determine the Log-Likelihood Function $l(\theta)$**
   Take the natural logarithm of the likelihood function.
3. **Maximize the Log-Likelihood**
   Take the partial derivatives of the log-likelihood w.r.t the parameter, set it equal to zero and solve the equation.

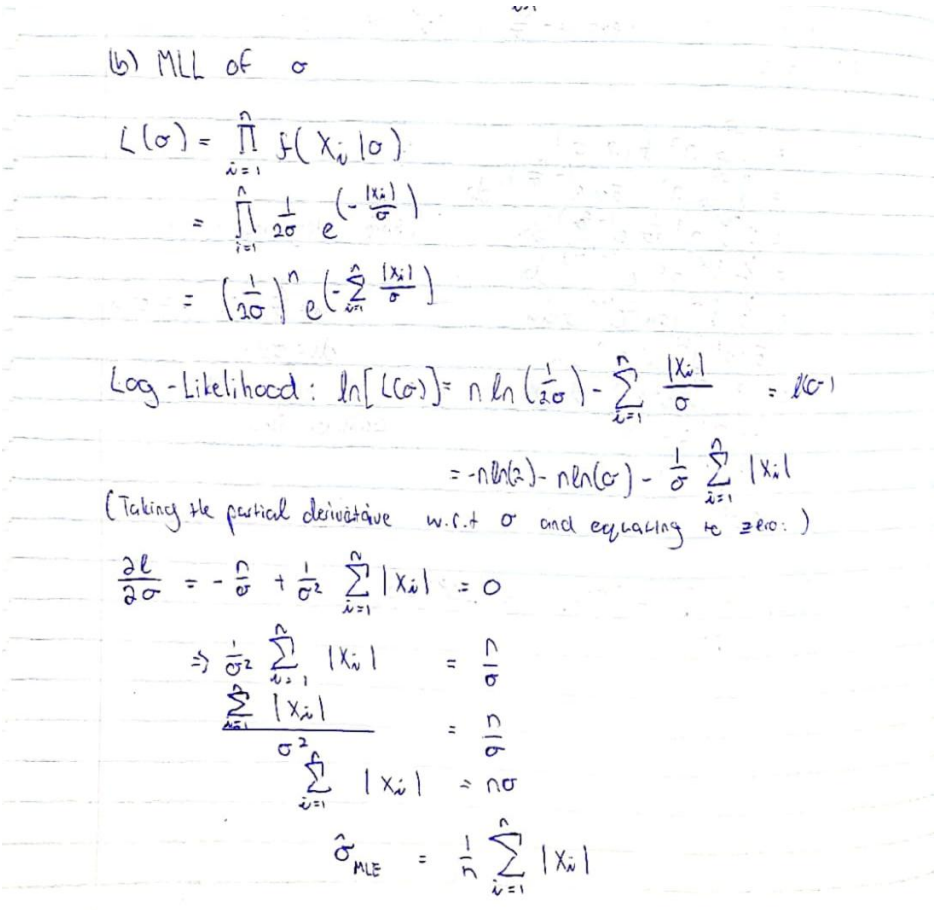
Exercises

16. Consider an i.i.d. sample of random variables with density function

$$f(x|\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|x|}{\sigma}\right)$$

a. Find the method of moments estimate of σ.
b. Find the maximum likelihood estimate of σ.
c. Find the asymptotic variance of the mle.
d. Find a sufficient statistic for σ.

FIGURE 87 (RICE, 2007, P. 319)



(b) MLE of σ

$$L(\sigma) = \prod_{i=1}^{n} f(x_i|\sigma)$$

$$= \prod_{i=1}^{n} \frac{1}{2\sigma} e^{\left(-\frac{|x_i|}{\sigma}\right)}$$

$$= \left(\frac{1}{2\sigma}\right)^n e^{\left(-\sum_{i=1}^{n}\frac{|x_i|}{\sigma}\right)}$$

Log-Likelihood: $\ln[L(\sigma)] = n\ln\left(\frac{1}{2\sigma}\right) - \sum_{i=1}^{n}\frac{|x_i|}{\sigma} = l(\sigma)$

$$= -n\ln(2) - n\ln(\sigma) - \frac{1}{\sigma}\sum_{i=1}^{n}|x_i|$$

(Taking the partial derivative w.r.t σ and equating to zero:)

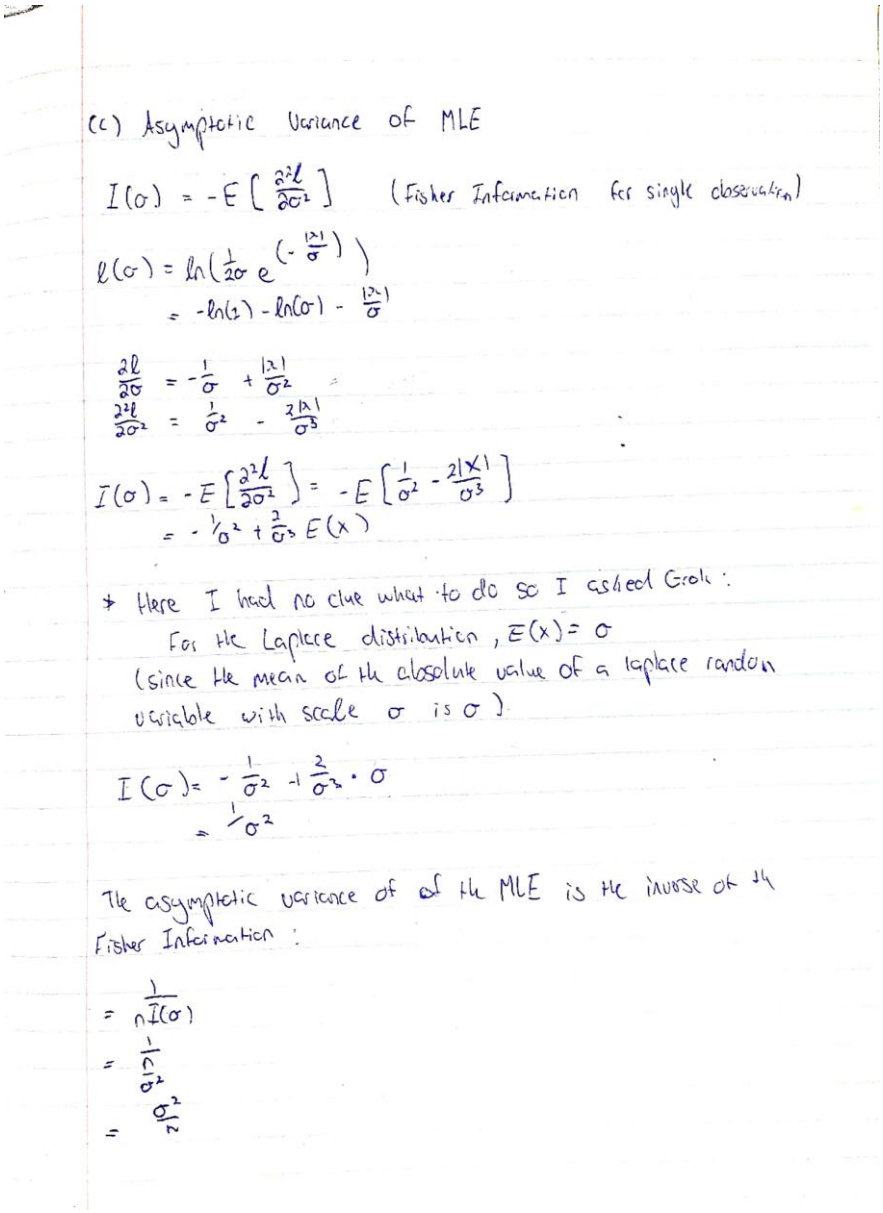
$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^2}\sum_{i=1}^{n}|x_i| = 0$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}|x_i| = \frac{n}{\sigma}$$

$$\frac{\sum_{i=1}^{n}|x_i|}{\sigma^2} = \frac{n}{\sigma}$$

$$\sum_{i=1}^{n}|x_i| = n\sigma$$

$$\hat{\sigma}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}|x_i|$$



(c) Asymptotic Variance of MLE

$I(\sigma) = -E\left[\frac{\partial^2 \ell}{\partial \sigma^2}\right]$ (Fisher Information for single observation)

$\ell(\sigma) = \ln\left(\frac{1}{2\sigma} e^{\left(-\frac{|x|}{\sigma}\right)}\right)$

$= -\ln(2) - \ln(\sigma) - \frac{|x|}{\sigma}$

$\frac{\partial \ell}{\partial \sigma} = -\frac{1}{\sigma} + \frac{|x|}{\sigma^2}$

$\frac{\partial^2 \ell}{\partial \sigma^2} = \frac{1}{\sigma^2} - \frac{2|x|}{\sigma^3}$

$I(\sigma) = -E\left[\frac{\partial^2 \ell}{\partial \sigma^2}\right] = -E\left[\frac{1}{\sigma^2} - \frac{2|X|}{\sigma^3}\right]$

$= -\frac{1}{\sigma^2} + \frac{2}{\sigma^3} E(X)$

* Here I had no clue what to do so I asked Grok:

For the Laplace distribution, $E(X) = \sigma$

(since the mean of the absolute value of a laplace random variable with scale σ is σ)

$I(\sigma) = -\frac{1}{\sigma^2} + \frac{2}{\sigma^3} \cdot \sigma$

$= \frac{1}{\sigma^2}$

The asymptotic variance of of the MLE is the inverse of the Fisher Information:

$= \frac{1}{n I(\sigma)}$

$= \frac{1}{\frac{n}{\sigma^2}}$

$= \frac{\sigma^2}{n}$

For Question 16c, I prompted Grok3 for some assistance. I have not taken linear algebra yet and are not that familiar with Laplace transformations.

50. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a Rayleigh distribution with parameter $\theta > 0$:

$$f(x|\theta) = \frac{x}{\theta^2} e^{-x^2/(2\theta^2)}, \qquad x \geq 0$$

(This is an alternative parametrization of that of Example A in Section 3.6.2.)

a. Find the method of moments estimate of θ.
b. Find the mle of θ.
c. Find the asymptotic variance of the mle.

FIGURE 88 (RICE, 2007, P. 323)

(b) MLE: $L(\theta) = \prod_{i=1}^{n} \frac{x_i}{\theta^2} e^{-x_i^2/(2\theta)^2}$

$= \left(\frac{1}{\theta^2}\right)^n \left(\prod_{i=1}^{n} x_i\right) e^{-\sum x_i^2/(2\theta)^2}$

$l(\theta) = -2n \log\theta + \sum_{i=1}^{n} \log x_i - \frac{1}{2\theta^2} \sum_{i=1}^{n} x_i^2$

$\frac{dl}{d\theta} = -\frac{2n}{\theta} + \frac{1}{\theta^3} \sum_{i=1}^{n} x_i^2 = 0$

$\frac{1}{\theta^3} \sum_{i=1}^{n} x_i^2 = \frac{2n}{\theta}$

$\therefore \hat{\theta}_{MLE} = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} x_i^2}$

(c) Asymptotic Variance:

$\frac{d^2l}{d\theta^2} = \frac{2n}{\theta^2} - \frac{3}{\theta^4} \sum_{i=1}^{n} x_i^2$

Using $E[x^2] = 2\theta^2$ (I don't want to derive that myself)

$I(\theta) = -E\left[\frac{d^2l}{d\theta^2}\right] = -\left(\frac{2n}{\theta^2} - \frac{3}{\theta^4} n E[x^2]\right)$

$= -\left(\frac{2n}{\theta^2} - \frac{3n \cdot 2\theta^2}{\theta^4}\right)$

$= -\left(\frac{2n}{\theta^2} - \frac{6n}{\theta^2}\right)$

$= \frac{4n}{\theta^2}$

Asymptotic variance of $\hat{\theta}_{MLE} = \frac{1}{I(\theta)} = \boxed{\frac{\theta^2}{4n}}$

The solution manual for this question has no intermediate steps, which I included.

55. For two factors—starchy or sugary, and green base leaf or white base leaf—the following counts for the progeny of self-fertilized heterozygotes were observed (Fisher 1958):

Type	Count
Starchy green	1997
Starchy white	906
Sugary green	904
Sugary white	32

According to genetic theory, the cell probabilities are $.25(2+\theta)$, $.25(1-\theta)$, $.25(1-\theta)$, and $.25\theta$, where $\theta(0 < \theta < 1)$ is a parameter related to the linkage of the factors.

a. Find the mle of θ and its asymptotic variance.
b. Form an approximate 95% confidence interval for θ based on part (a).
c. Use the bootstrap to find the approximate standard deviation of the mle and compare to the result of part (a).
d. Use the bootstrap to find an approximate 95% confidence interval and compare to part (b).

FIGURE 89 (RICE, 2007, P. 324)

Question 55 (a)

observations		Cell probabilities
n_1 (Starchy green)	= 1997	$P_1(\theta) = 0.25(2+\theta)$
n_2 (Starchy white)	= 906	$P_2(\theta) = 0.25(1-\theta)$
n_3 (Sugary Green)	= 904	$P_3(\theta) = 0.25(1-\theta)$
n_4 (Sugary white)	= 32	$P_4(\theta) = 0.25\,\theta$
N	= 3939	where $0 < \theta < 1$

$$\ell(\theta) = 1997 \log(2+\theta) + (906+904)\log(1-\theta) + 32\log(\theta)$$

Differentiating $\ell(\theta)$ w.r.t θ and equating to zero:

$$\frac{d\ell}{d\theta} = \frac{1997}{2+\theta} - \frac{1810}{1-\theta} + \frac{32}{\theta} = 0$$

Multiply by $\theta(2+\theta)(1-\theta)$:

$$1997\theta(1-\theta) - 1810\theta(2+\theta) + 32(2+\theta)(1-\theta) = 0$$
$$\Rightarrow 1997(\theta-\theta^2) - 1810(2\theta+\theta^2) + 32(2-\theta-\theta^2) = 0$$
$$\Rightarrow 1997\theta^2 - 3620\theta - 1810\theta^2 + 64 - 32\theta - 32\theta^2 = 0$$
$$\Rightarrow -3939\theta^2 - 1655\theta + 64 = 0$$

$$\hat{\theta} = \frac{-1655 \pm \sqrt{1655^2 - 4(3939)(-64)}}{2(3939)}$$

since $0 < \theta < 1$, I only take the positive root:

$$\hat{\theta}_{MLE} = 0.03565$$

Asymptotic variance of $\hat{\theta}$ ($[I(\theta)]^{-1}$)

$$I(\theta) = -E\left[\frac{d^2\ell}{d\theta^2}\right]$$
$$= \frac{1997}{(2+\theta)^2} + \frac{1810}{(1-\theta)^2} + \frac{32}{\theta^2}$$

Reflection

This was an interesting section with some particularly challenging proofs. I spend some considerable time trying to decipher them. I believe that I now have a pretty good understanding of the properties and their related proofs.

I was under a lot of pressure during the weeks I spent on this section, writing a lot of semester test throughout the week. I did not spend less time on this module, but the time I spent was at inconvenient times. I often found myself working on my portfolio in between classes and late at night. Regardless of the circumstances, I got it done and still enjoyed the process.

Efficiency and the Cramér-Rao Lower Bound

Research Process

- I consulted my (Rice, 2007) textbook for the definition, proof and related properties of the Cramér-Rao Lower Bound.
- I found the definition of the Cramér-Rao Inequality on the Glossary of Statistics.com. They defined it as follows: "Every unbiased estimator has a variance greater than or equal to a lower bound called the Cramer – Rao lower bound. If the variance of an unbiased estimator achieves the Cramer – Rao lower bound, then that estimator is a minimum variance unbiased estimator, or, simply, efficient estimator "
- I found the proof for the Cramer-Rao Inequality on gregorygunderson.com.
- I consulted my (George Casella, 2001) textbook for some additional exercises.
- I prompted Claude 4 Sonnet the explain the CRLB and its proof.

Definition

THEOREM A *Cramér-Rao Inequality*

Let $X_1, \ldots, X_n$ be i.i.d. with density function $f(x|\theta)$. Let $T = t(X_1, \ldots, X_n)$ be an unbiased estimate of θ. Then, under smoothness assumptions on $f(x|\theta)$,

$$\operatorname{Var}(T) \geq \frac{1}{nI(\theta)}$$

FIGURE 90 (RICE, 2007, P. 301)

Proof

Let

$$Z = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i|\theta)$$

$$= \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \theta} f(X_i|\theta)}{f(X_i|\theta)}$$

In Section 8.5.2, we showed that $E(Z) = 0$. Because the correlation coefficient of Z and T is less than or equal to 1 in absolute value

$$\operatorname{Cov}^2(Z, T) \leq \operatorname{Var}(Z)\operatorname{Var}(T)$$

It was also shown in Section 8.5.2 that

$$\operatorname{Var}\left[\frac{\partial}{\partial \theta} \log f(X|\theta)\right] = I(\theta)$$

Therefore,

$$\operatorname{Var}(Z) = nI(\theta)$$

The proof will be completed by showing that $\text{Cov}(Z, T) = 1$. Since Z has mean 0,

$$\text{Cov}(Z, T) = E(ZT)$$
$$= \int \cdots \int t(x_1, \ldots, x_n) \left[\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i|\theta)}{f(x_i|\theta)} \right] \prod_{j=1}^{n} f(x_j|\theta)\, dx_j$$

Noting that

$$\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i|\theta)}{f(x_i|\theta)} \prod_{j=1}^{n} f(x_j|\theta) = \frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(x_i|\theta)$$

we rewrite the expression for the covariance of Z and T as

$$\text{Cov}(Z, T) = \int \cdots \int t(x_1, \ldots, x_n) \frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(x_i|\theta)\, dx_i$$
$$= \frac{\partial}{\partial\theta} \int \cdots \int t(x_1, \ldots, x_n) \prod_{i=1}^{n} f(x_i|\theta)\, dx_i$$
$$= \frac{\partial}{\partial\theta} E(T) = \frac{\partial}{\partial\theta}(\theta) = 1$$

which proves the inequality. [Note the interchange of differentiation and integration that must be justified by the smoothness assumptions on $f(x|\theta)$.] ■

The proof in Rice is hard to follow, as we realized in class. Luckily, Gregory Gundersen tackled this proof, and Claude 4 Sonnet just released in the week of writing. Putting these resources to good use, I have a better breakdown and understanding of this proof.

For the first strep of our proof, we need to find the mean and variance of the score function (derivative of the log-likelihood).

For any parameter θ, the mean of the score function is:

$$\begin{aligned} E[\partial/\partial\theta \; log\, f(X|\theta)] &= \int (\partial/\partial\theta \; log\, f(x|\theta))\, f(x|\theta)\, dx \\ &= \int (\partial f(x|\theta)/\partial\theta)/f(x|\theta) \cdot f(x|\theta)\, dx \\ &= \int \partial f(x|\theta)/\partial\theta \, dx \\ &= \partial/\partial\theta \int (x|\theta)\, dx \\ &= \partial/\partial\theta \; (1) = 0 \end{aligned}$$

Variance of the score function:

$$Var(S(\theta)) = E[S(\theta)^2] - (E[S(\theta)])^2 = E[S(\theta)^2] = I(\theta)$$

For the second step, we need to find the covariance between the estimator and the score function. For and unbiased estimator T, where $E(T) = \theta$:

$$\begin{aligned} Cov(T, S(\theta)) &= E[T \cdot S(\theta)] - E[T] \cdot E[S(\theta)] \\ &= E[T \cdot S(\theta)] - \theta \cdot 0 \end{aligned}$$

$$= E[T \cdot S(\theta)]$$

We can compute $E[T \cdot S(\theta)]$:

$$E[T \cdot S(\theta)] = \int t(x) \cdot (\partial/\partial\theta \, log \, f(x|\theta)) \cdot f(x|\theta) \, dx$$
$$= \int t(x) \cdot \partial f(x|\theta)/\partial\theta \, dx$$

Under smoothness assumptions on $f(x|\theta)$, we can interchange differentiation and integration.

$$\frac{\partial}{\partial\theta} E[T] = \frac{\partial}{\partial\theta} \int t(x) f(x|\theta) dx$$
$$= \int t(x) \, \partial f(x|\theta)/\partial\theta \, dx$$

And since T is unbiased, $E[T] = \theta$:

$$\partial/\partial\theta \, E[T] = \partial\theta/\partial\theta = 1$$

Therefore

$$Cov(T, S(\theta)) = E[T \cdot S(\theta)] = 1$$

For the finale step, we need to apply the Cauchy-Schwarz Inequality. It states that for any random variables U and V:

$$|Cov(U,V)|^2 \leq Var(U) \cdot Var(V)$$

We need to apply the inequality to our estimator T and our score function $S(\theta)$:

$$|Cov(T, S(\theta))|^2 \leq Var(T) \cdot Var(S(\theta))$$
$$|1|^2 \leq Var(T) \cdot I(\theta)$$
$$1 \leq Var(T) \cdot I(\theta)$$

And Finally

$$Var(T) \geq \frac{1}{I(\theta)}$$

My understanding of Efficiency and the Cramér-Rao Lower Bound

The Cramér-Rao lower bound gives us the minimum variance that an unbiased estimator of θ can achieve. If an estimator is equal to the Cramér-Rao lower bound, it is said to be an efficient estimator. An efficient estimator is the best possible estimator, in terms of the minimum variance.

Maximum likelihood estimates are asymptotically efficient (under regularity conditions), since the asymptotic variance of MLE is equal to the CRLB. However, Rice warned that this may not be the case for finite sample sizes.

How does it relate to the bigger picture?

The Cramér-Rao lower bound is much more significant than just a variance bound; it is more like a cornerstone of statistical theory. The CRLB is the connection between efficiency and sufficiency in exponential families. (The MLE based on a sufficient statistic is efficient). The CRLB also has a close connection to the Rao-Blackwell theorem. Rao-Blackwellization can

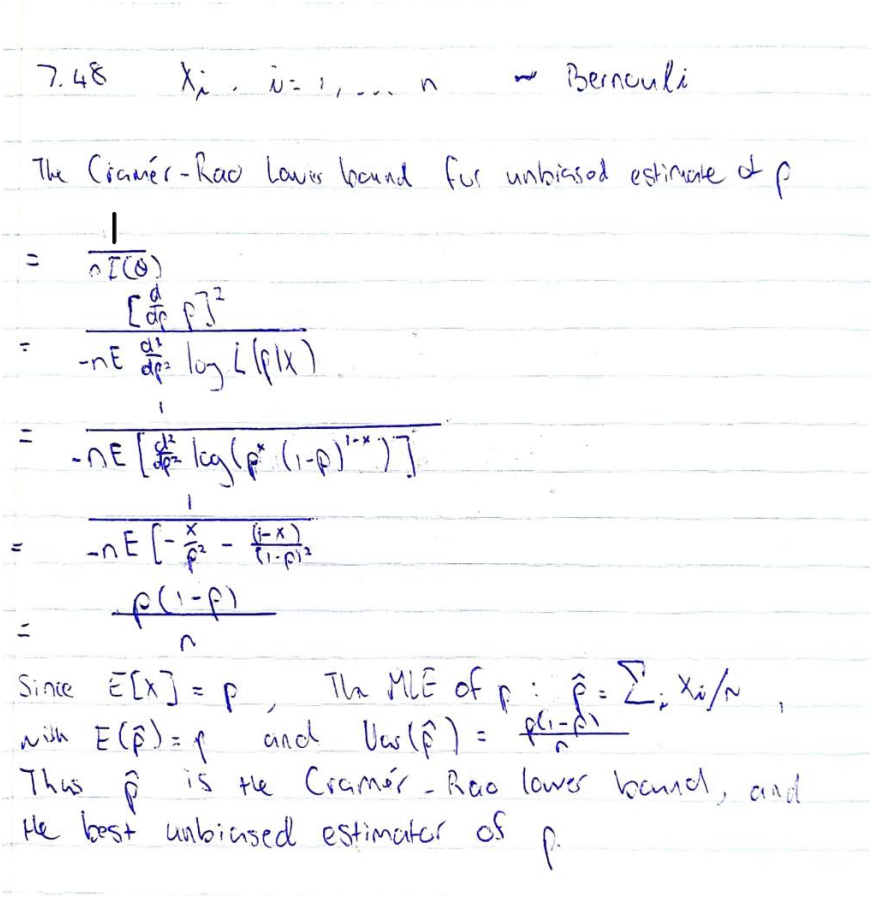
produce an estimator that attains the CRLB, making it efficient. The CRLB connects to information theory through Fisher Information (the denominator of the CRLB). This rabbit hole of relationships to the CRLB goes much deeper than the scope of this module.

Exercises

7.48 Suppose that $X_i, i = 1, \ldots, n$, are iid Bernoulli(p).

(a) Show that the variance of the MLE of p attains the Cramér–Rao Lower Bound.

FIGURE 91 (GEORGE CASELLA, 2001, P. 364)



Reflection

This time around, I had a much better experience, of proving the inequality. I used all available resources and technology, so I don't have to build the bridge myself.

I made an interesting observation about myself. I am not the most tidy and organised person. But I am very neat with my portfolio, I want it to be the finest and most professional portfolio there is.

Sufficiency

Research Process

- I consulted my (Rice, 2007) textbook for the definition, proof and related properties of the Sufficiency principle.
- I read an article by Will Fithian from the UC Berkeley Statistics Department. The article defined the principle of sufficiency and proved it in the discrete case.
- I prompted Claude 4 Sonnet to define sufficiency and explain the factorization theorem.

Definition

What is a statistic?

A statistic could be defined as a summary measure of our sample. It is calculated from our sample data and used to make inferences about our population.

DEFINITION

A statistic $T(X_1, \ldots, X_n)$ is said to be **sufficient** for θ if the conditional distribution of $X_1, \ldots, X_n$, given $T = t$, does not depend on θ for any value of t. ■

FIGURE 92 (RICE, 2007, P. 305)

Sufficient statistics allows us to focus on the essential aspects of our data to make inferences, while ignoring irrelevant detail or bulky information. A sufficient statistic $T(X)$ should carry all the information contained in our data, allowing us to make inferences about θ.

Factorization Theorem

THEOREM A

A necessary and sufficient condition for $T(X_1, \ldots, X_n)$ to be sufficient for a parameter θ is that the joint probability function (density function or frequency function) factors in the form

$$f(x_1, \ldots, x_n|\theta) = g[T(x_1, \ldots, x_n), \theta]h(x_1, \ldots, x_n)$$

FIGURE 93 (RICE, 2007, P. 306)

The easiest method to prove a statistic to be sufficient, would be to prove that the joint probability function $f(X|\theta)$, factorizes into a part that involves only θ and $T(X)$, so $g[T(x_1, \dots, x_n), \theta]$. And a part that only involves $h(x)$, $h(x_1, \dots, x_n)$.

Proof

We give a proof for the discrete case. (The proof for the general case is more subtle and requires regularity conditions, but the basic ideas are the same.) First, suppose that the frequency function factors as given in the theorem. To simplify notation, we will let $\mathbf{X}$ denote $(X_1, \dots, X_n)$ and $\mathbf{x}$ denote $(x_1, \dots, x_n)$. We have

$$\begin{aligned} P(T = t) &= \sum_{T(x)=t} P(\mathbf{X} = \mathbf{x}) \\ &= g(t, \theta) \sum_{T(x)=t} h(\mathbf{x}) \end{aligned}$$

Here the notation indicates that the sum is over all $\mathbf{x}$ such that $T(\mathbf{x}) = t$. We then have

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|T = t) &= \frac{P(\mathbf{X} = \mathbf{x}, T = t)}{P(T = t)} \\ &= \frac{h(\mathbf{x})}{\sum_{T(\mathbf{X})=t} h(\mathbf{x})} \end{aligned}$$

This conditional distribution does not depend on θ, as was to be shown.

To show that the conclusion holds in the other direction, suppose that the conditional distribution of $\mathbf{X}$ given T is independent of θ. Let

$$\begin{aligned} g(t, \theta) &= P(T = t|\theta) \\ h(\mathbf{x}) &= P(\mathbf{X} = \mathbf{x}|T = t) \end{aligned}$$

We then have

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|\theta) &= P(T = t|\theta) P(\mathbf{X} = \mathbf{x}|T = t) \\ &= g(t, \theta) h(\mathbf{x}) \end{aligned}$$

as was to be shown. ■

FIGURE 94 (RICE, 2007, P. 307)

My explanation of this proof is mostly based on class notes. This proof is bi-directional, It T is sufficient, then we can factor the joint density, and if we can factor the joint density, then T is sufficient.

For the first step, we define a statistic as sufficient for θ, if the conditional distribution of X given T does not depend on θ.

For the second step, we need want to factorise the joint density in $g(t,\theta)$ and $h(x)$, which would mean that t is sufficient for θ. Therefore, our goal is to show:

$$f(x|\theta) = g(t,\theta) \cdot h(x)$$

For step 2.1, Working in the forward direction. Suppose that we can factor, then the distribution of the sample given the statistic, is not dependent on θ. Therefore, we need to show that $P(X = x \mid T = t)$ does not depend on θ. We can find the marginal distribution of T:

$$\begin{aligned} P(T = t) &= \sum P(X = x) \quad (sum\ over\ all\ x\ such\ that\ T(x) = t) \\ &= \sum g(t,\theta) \cdot h(x) \quad (factorising) \\ &= g(t,\theta) \cdot \sum h(x) \quad (since\ g(t,\theta) is\ constant\ when\ t\ is\ fixed) \end{aligned}$$

And

$$\sum h(x) = H(t) \quad (sum\ over\ all\ x\ where\ T(x) = t)$$

So:

$$P(T = t) = g(t,\theta) \cdot H(t)$$

For step 2.2, we are now working in the reverse direction. Suppose the sample given the statistic does not depend on θ. (i.e. we are assuming sufficiency). We now have to compute $P(X = x \mid T = t)$:

$$= \frac{P(X = x, T = t)}{P(T = t)}$$

Since $T = T(X)$, when X = x and T = t, we must have that $T(x) = t$.

Therefore, $P(X = x, T = t) = P(X = x)$ if $T(x) = t$, and 0 otherwise.

$$\begin{aligned} P(X = x \mid T = t) &= P(X = x)/P(T = t) \\ &= [g(t,\theta) \cdot h(x)] \,/\, [g(t,\theta) \cdot H(t)] \\ &= h(x) \,/\, H(t) \\ &= h(x) \,/ \sum h(x) \quad (where\ sum\ is\ over\ T(x) = t) \end{aligned}$$

Now our final expression does not contain θ at all. The conditional distribution is now only dependent on $h(x)$ and $\sum h(x)$, neither of which involves θ.

COROLLARY A

If T is sufficient for θ, the maximum likelihood estimate is a function of T.

Proof

From Theorem A, the likelihood is $g(T, \theta)h(\mathbf{x})$, which depends on θ only through T. To maximize this quantity, we need only maximize $g(T, \theta)$. ■

FIGURE 95 (RICE, 2007, P. 309)

My Understanding of Sufficiency

I would define a statistic simply and intuitively, as a summary measure of the sample.

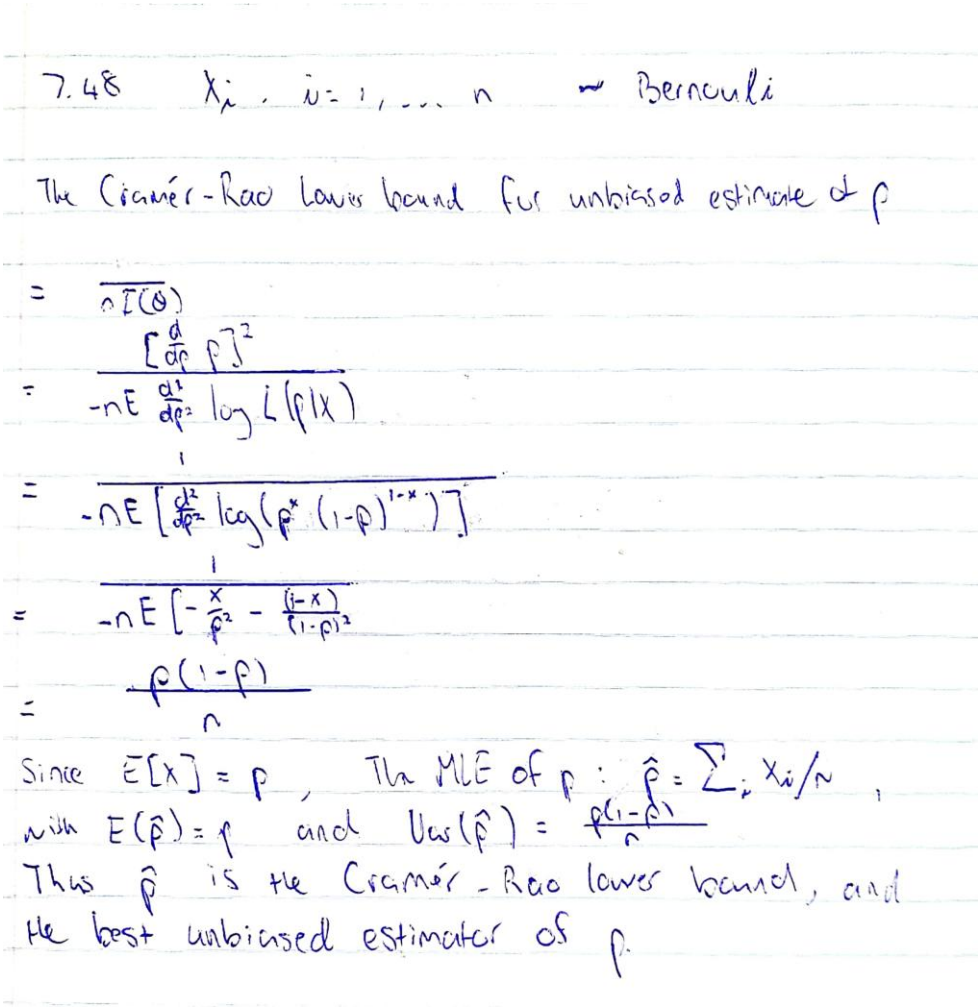
A statistic is sufficient for parameter θ, if the conditional distribution of the sample given the statistic, is not dependent on θ. A sufficient statistic should therefore summarise all the sample information relevant to our estimation of θ.

Another important property briefly mentioned in Rice is, that for exponential families, the sufficient statistic $T(X)$, achieves the Cramér-Rao Lower Bound. Therefore $T(X)$ is not only sufficient, but also efficient.

Exercises

75. Show that the gamma distribution belongs to the exponential family.

FIGURE 96 (RICE, 2007, P. 328)



I found the proof from the Purdue Department of Statistics, and asked ChatGPT to explain some of the steps to me.

16. Consider an i.i.d. sample of random variables with density function

$$f(x|\sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right)$$

a. Find the method of moments estimate of σ.
b. Find the maximum likelihood estimate of σ.
c. Find the asymptotic variance of the mle.
d. Find a sufficient statistic for σ.

FIGURE 97 (RICE, 2007, P. 319)

16. (d) $f(x|\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|x|}{\sigma}\right), \quad x \in \mathbb{R}, \quad \sigma > 0$

Variables are i.i.d, so joint density is:

$$f(x_n|\sigma) = \prod_{i=1}^{n} f(x_i|\sigma) = \prod_{i=1}^{n}\left(\frac{1}{2\sigma}\exp\left(-\frac{|x_i|}{\sigma}\right)\right)$$

$$f(x_n|\sigma) = \underbrace{\left(\frac{1}{2\sigma}\right)^n}_{\text{constant}}\exp\left(-\frac{1}{\sigma}\sum_{i=1}^{n}|x_i|\right)$$

from factorization theorem: $f(x_n|\sigma) = g[T(x_n)|\sigma]\,h(x_n)$

let $T(x) = \sum_{i=1}^{n}|x_i|$

$$\therefore f(x_1,\dots,x_n|\sigma) = \underbrace{\left(\frac{1}{2\sigma}\right)^n\exp\left(-\frac{1}{\sigma}T(x)\right)}_{g[T(x),\sigma]} \cdot \underbrace{1}_{h(x_1,\dots x_n)}$$

By the Factorization Theorem, $T(x) = \sum_{i=1}^{n}|x_i|$ is a sufficient statistic for σ

The rest of question 16 was done in earlier sections.

The Rao-Blackwell Theorem

Research Process

- I consulted my (Rice, 2007) textbook for the definition, proof and related properties of the Rao-Blackwell Theorem.
- I read an article on blackhc.net on the Rao-Blackwell Theorem, which explained it's proof.
- I read the article on the Rao-Blackwell Theorem and it's proof on gregorygundersen.com.
- I prompted Claude 4 Sonnet to explain the relationship between the Rao-Blackwell Theorem and the Cramer-Rao Lower Bound.

Definition

THEOREM A *Rao-Blackwell Theorem*

Let $\hat{\theta}$ be an estimator of θ with $E(\hat{\theta}^2) < \infty$ for all θ. Suppose that T is sufficient for θ, and let $\tilde{\theta} = E(\hat{\theta}|T)$. Then, for all θ,

$$E(\tilde{\theta} - \theta)^2 \leq E(\hat{\theta} - \theta)^2$$

The inequality is strict unless $\hat{\theta} = \tilde{\theta}$.

Proof

We first note that, from the property of iterated conditional expectation (Theorem A of Section 4.4.1),

$$E(\tilde{\theta}) = E[E(\hat{\theta}|T)] = E(\hat{\theta})$$

Therefore, to compare the mean squared error of the two estimators, we need only compare their variances. From Theorem B of Section 4.4.1, we have

$$\text{Var}(\hat{\theta}) = \text{Var}[E(\hat{\theta}|T)] + E[\text{Var}(\hat{\theta}|T)]$$

or

$$\text{Var}(\hat{\theta}) = \text{Var}(\tilde{\theta}) + E[\text{Var}(\hat{\theta}|T)]$$

Thus, $\text{Var}(\hat{\theta}) > \text{Var}(\tilde{\theta})$ unless $\text{Var}(\hat{\theta}|T) = 0$, which is the case only if $\hat{\theta}$ is a function of T, which would imply $\hat{\theta} = \tilde{\theta}$. ■

FIGURE 98 (RICE, 2007, P. 310)

My Understanding of the Rao-Blackwell Theorem and it's Proof

The Rao-Blackwell Theorem states that if we have an unbiased estimator for parameter θ, and a sufficient statistic $T(X)$ for the parameter θ, we can construct another estimator $\tilde{\theta}$, such that $\tilde{\theta}$ will still be unbiased and may have less variance. Therefor conditioning an estimator on a sufficient statistic never increases its mean squared error (variance) and typically reduces it.

Proof:

$$E(\tilde{\theta}) = E[E(\hat{\theta}|T)] = E(\hat{\theta})$$

Theorem A from Section 4.4.1 is the law of total expectation. The Rao-Blackwellization process preserves the bias structure of the original estimator. So, if $\hat{\theta}$ was unbiased, then $\tilde{\theta}$ would remain unbiased. And if $\hat{\theta}$ had a bias, then $\hat{\theta}$ would retain exactly the same bias.

$$Var(\hat{\theta}) = Var[E(\hat{\theta}|T)] + E[Var(\hat{\theta}|T)]$$

This is the conditional variance formula, then:

$$Var(\hat{\theta}) = Var(\tilde{\theta}) + E[Var(\hat{\theta}|T)]$$

Since $E[Var(\hat{\theta}|T)] \geq 0$, because the variance is always non-negative, we have:

$$Var(\hat{\theta}) \geq Var(\tilde{\theta})$$

The theorem relies on the fact that T is sufficient for θ. Sufficiency ensures that $E(\tilde{\theta})$ doesn't depend on θ, by the factorization theorem. Furthermore, it ensures that the conditional expectation captures all relevant information about θ contained within the data.

Reflection

How does the Rao-Blackwell improvement relate to the Cramér-Rao bound?

The Cramér-Rao bound provides us with a theoretical limit of precision, while the Roa-Blackwell theorem provides us with a method to improve our estimators. We can use the Rao-Blackwellization process to find an estimator as close as possible or even equal to that limit.

After spending well north of 200 hours on this module, and going the extra mile to grasp the content, I firmly believe that I understand the content, purpose and importance of this module.

Assignment 4

Description of Problem, and why I chose it

I wanted to put the different methods of parameter estimation on equal turf and compare them. I used the morley dataset, which contains 100 measurements of the speed of light (measured in km/s, with 299000 being subtracted). I used the Maximum likelihood estimate to find an estimate for the mean and find the asymptotic variance. I used the MLE parameters and run a parametric bootstrap with 1000 and 100000 samples. I also ran a non-parametric bootstrap on the data with 1000 and 100000 samples.

Solution

I can assume for now that the speed of light measurements are normally distributed. The errors of measurements in physics are often normally distributed. I will later test this assumption with a chi-squared goodness of fit tests. From the pdf of a $N(\mu, \sigma^2)$ distribution is:

$$f(X_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)$$

Then the likelihood function of the observations will be the product of the individual PDFs, (since they are i.i.d.):

$$L(\mu,\sigma^2|X) = \prod_{i=1}^{n} f(X_i|\mu,\sigma^2)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right)$$

Then the Log-Likelihood Function is:

$$l(\mu,\sigma^2|X) = \ln(L)$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

To find their maximum values, we need to take the partial derivatives with respect to μ and σ^2 and set them equal to zero:

$$\frac{\partial l}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(X_i-\mu)(-1) = 0$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0$$

$$\sum_{i=1}^{n}(X_i-n\mu) = 0$$

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$= \bar{X}$$

The maximum likelihood estimate for the mean is equal to the sample mean.

Let $\theta = \sigma^2$ and take the partial derivative with respect to θ.

$$l = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\theta) - \frac{1}{2\theta}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\frac{\partial l}{\partial \theta} = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Equating to zero and substituting μ with $\hat{\mu}_{MLE} = \bar{X}$:

$$-\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(X_i - \bar{X})\text{^}2 = 0$$

$$\frac{1}{2\theta^2}\sum_{i=1}^{n}(X_i - \bar{X})\text{^}2 = n/2\theta$$

$$\hat{\theta}_{MLE} = \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})\text{^}2$$

Therefore, the Maximum Likelihood Estimate for the variance, is the biased sample variance.

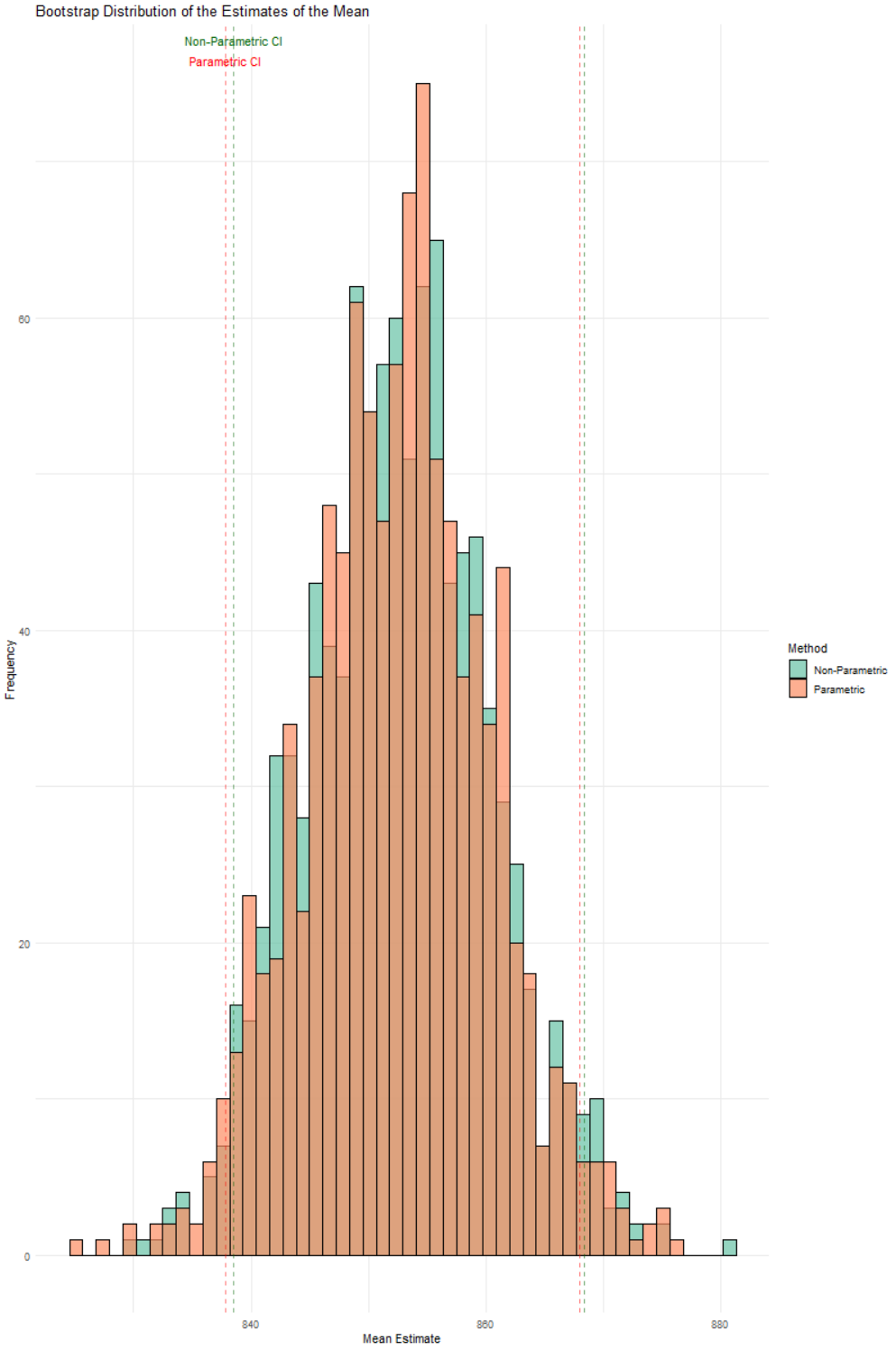
asssignmnet 4 stsm2616.R

```r
library(boot)
data("morley")

speed <- morley$Speed
n <- length(speed)

R <- 1000
set.seed(12345)

# Chi-squared test (I can assume normality, but its even better to test)
breaks <- seq(min(speed), max(speed), length.out = 10)
observed <- hist(speed, breaks = breaks, plot = FALSE)$counts
expected <- n * diff(pnorm(breaks, mean = mean(speed), sd = sd(speed)))
chi_test <- chisq.test(observed, p = expected/sum(expected))

# Since I showed that the MLE mean is equal to the sample mean
mean_mle <- mean(speed)

var_mle <- var(speed) # this is an unbiased estimator of the variance
asymptotic_var <- var_mle / n
se_mle <- sqrt(asymptotic_var)

```

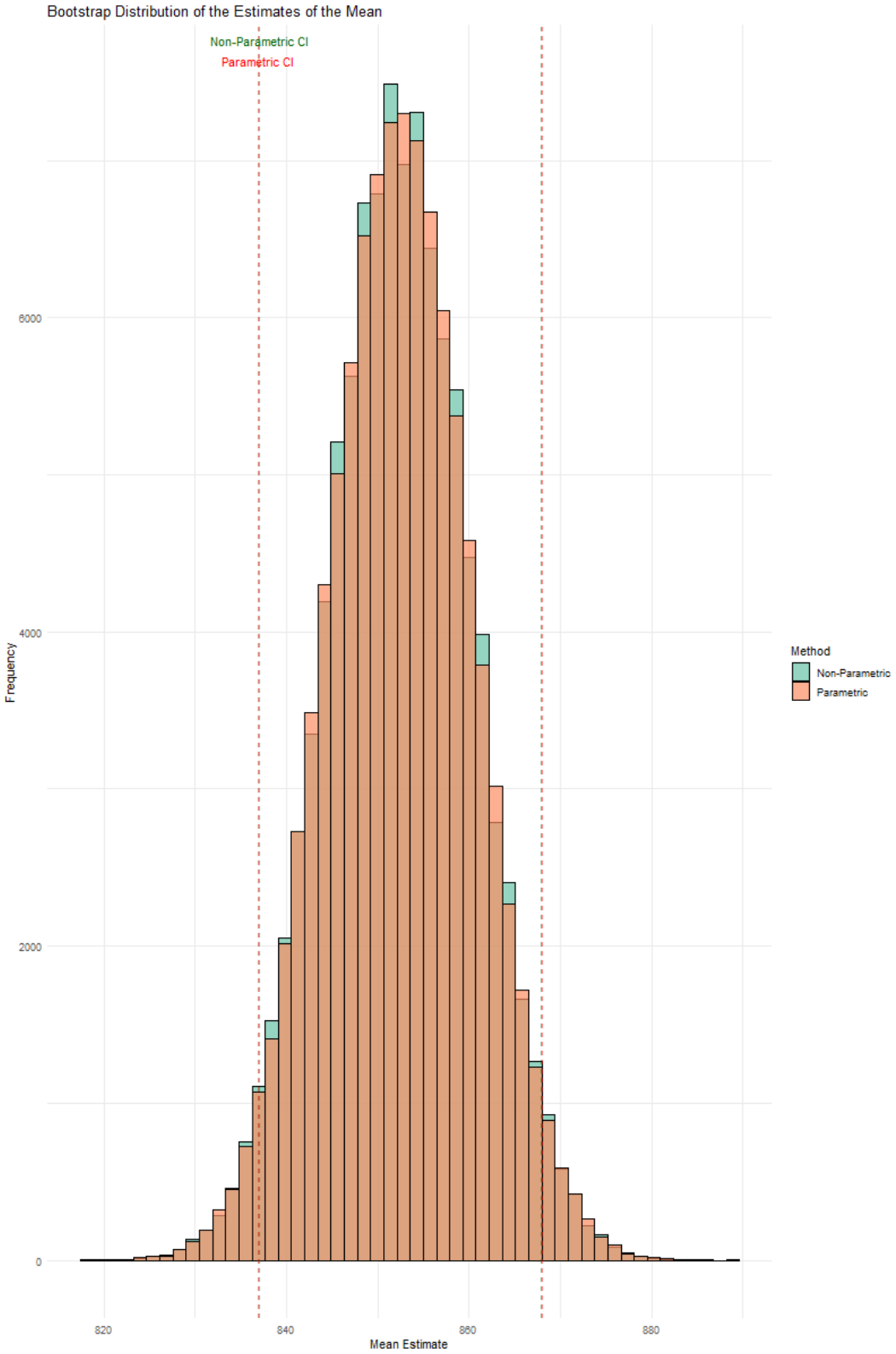
```r
# Non-parametric / Classical bootstrap
bootstrap_mean <- function(data, indices) {
  return(mean(data[indices]))
}

np_boot_result <- boot(data = speed, statistic = bootstrap_mean, R = R)

np_boot_means <- np_boot_result$t
np_boot_mean <- mean(np_boot_result$t)
np_boot_var <- var(np_boot_result$t)
np_boot_se <- sd(np_boot_result$t)
np_ci <- quantile(np_boot_means, c(0.025, 0.975))

# Parametric bootstrap
mu_hat <- mean_mle
sigma_hat <- sd(speed)
p_boot_means <- numeric(R)
p_boot_sds <- numeric(R)

for(i in 1:R){
  p_boot_sample <- rnorm(n, mean = mu_hat, sd = sigma_hat)
  p_boot_means[i] <- mean(p_boot_sample)
  p_boot_sds[i] <- sd(p_boot_sample)
}

p_boot_mean <- mean(p_boot_means)
p_boot_variance <- var(p_boot_means)
p_boot_se <- sd(p_boot_means)
p_ci <- quantile(p_boot_means, c(0.025, 0.975))

```

```r
library(ggplot2)

df_boot <- data.frame(
  Method = rep(c("Non-Parametric", "Parametric"), each = R),
  Mean = c(np_boot_means, p_boot_means)
)

p <- ggplot(df_boot, aes(x = Mean, fill = Method)) +
  geom_histogram(alpha = 0.7, bins = 50,
                 position = "identity", color = "black") +
  labs(title = "Bootstrap Distribution of the Estimates of the Mean",
       x = "Mean Estimate",
       y = "Frequency") +
  theme_minimal() +
  scale_fill_brewer(type = "qual", palette = 7) +
  geom_vline(xintercept = np_ci, color = "darkgreen",
             linetype = "dashed", alpha = 0.5) +
  geom_vline(xintercept = p_ci, color = "red",
             linetype = "dashed", alpha = 0.5) +
  annotate("text", x = np_ci[1], y = Inf,
           label = "Non-Parametric CI", vjust = 2, color = "darkgreen") +
  annotate("text", x = p_ci[1], y = Inf,
           label = "Parametric CI", vjust = 4, color = "red")
print(p)
```

All code was written myself using R in Cursor studio.



Back to Table of Contents



Notice how the confidence intervals of both methods of the bootstrap is nearly identical.

MLE Mean	852.4	
MLE Asymptotic Variance	62.42667	
MLE Standard Error	7.901055	
Number of Samples	1000	100 000
Non-Parametric Bootstrap Mean	852.6125	852.4164
Non-Parametric Bootstrap Variance	58.83889	62.37965
Non-Parametric Bootstrap Standard Error	7.670651	7.898079
Non-Parametric Bootstrap 95% confidence interval	(838.4975, 868.305)	(837, 867.9)
Parametric Bootstrap Mean	852.5198	852.4472
Parametric Bootstrap Variance	59.76608	62.16558
Parametric Bootstrap Standard Error	7.730853	7.884515
Parametric Bootstrap 95% Confidence interval	(837.8242, 867.9889)	(836.9654, 867.9999)

Interpretation of Results

The Chi-squared goodness of fit test for normality gave a p-value of 0.9484705, proving I was right with my assumption that the data must be normally distributed. I also observed the very distinct shape of the normal distribution with the non-parametric bootstrap at 100 000 samples Of course the parametric bootstrap will be normal since I am sampling from a normal distribution fitted to my parameters. Since the values of the parametric and non-parametric bootstrap are so close it further suggests that the sample data is normally distributed.

After putting all methods on equal turf, they all provided consistent, similar estimates, each proving their worth. Off course each method will have circumstances in which they will outshine the others. Non-parametric bootstrapping is more robust, since we do not have to assume any distribution at all. Parametric bootstrapping is theoretically more accurate than the non-parametric bootstrap but does require us to estimate the parameters.

For this example, there was little advantage for using a parametric bootstrap with MLE parameter estimates, over a simple and uncomplicated non-parametric bootstrap.

Why I found this problem interesting?

I found it interesting that the $\hat{\mu}_{MLE}$ for the normal distribution is just the sample mean, and $\hat{\theta}_{MLE}$ is the biased sample variance.

The bootstrap was the most interesting method I have learned. I was very curious to see whether there would be significant differences between the methods, if I used them in a problem that gives neither a large comparative advantage.

I found it very interesting that despite all the math and technologies used in the maximum likelihood estimate and parametric bootstrap. It showed no significant advantage over a simple classical bootstrap.

What I learned from this Problem?

I revisited the chi-squared goodness of fit test, done earlier much earlier and this module. It is the first time I have used it to help me solve an actual problem.

I learned that the maximum likelihood estimates of the parameters of the normal distribution is just the sample mean and biased sample variance.

I learned that, when you have a good fit to a distribution, the difference between a parametric and a non-parametric bootstrap is negligible.

This assignment was a good opportunity to sharpen my statistical coding skills in R, and I must say, I am getting nifty with my graphs.

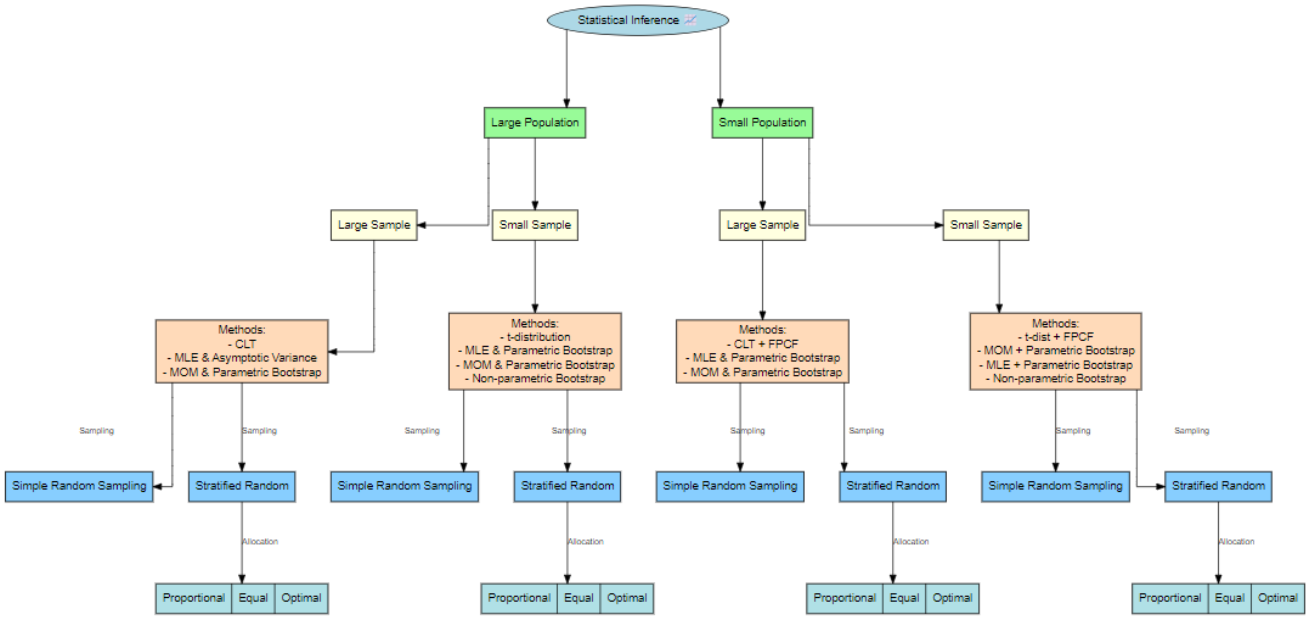
The Big Picture

The Core Questions for this Module:

How do we construct sampling distributions and make statistical inferences about our population, accounting for the fact that there exists sampling uncertainty?

Can we quantify, minimize and account for the potential variation in our sample's composition, compared to our target population?

My Map for the process of Statistical Inference



I recreated my summary from class activity 3 using R and some meticulous prompt engineering for Gemini 2.5.

Important notes

- Small samples may not always be approximately normally distributed, and therefore the t-distribution may not always be applicable.
- Non-Parametric Bootstrapping is extremely versatile and may just as well be used for larger samples.

My Bigger, Big Picture

The Dawn of the AI Revolution

The use of AI and its capabilities was a hot topic in class throughout this module. I have a deep interest in artificial intelligence and have done a lot of research throughout the year. I have some thoughts I like to share.

Despite the AI-craze and the possible bubble it created in the stock market, I truly believe that AI is "underhyped", and that people do not grasp the possibilities. I would compare this to the dot com bubble of the 90's (Cassidy, 2002). Every company rushed to have their own webpage, some thought of it as an interesting gimmick, but few understood how radically the internet would change society and the economy (Brynjolfsson & McAfee, 2014).

There were astronomically improvements in AI technology throughout 2025. AI become a much more powerful and useful tool for myself. The launch of DeepSeek challenged the dominance that OpenAI enjoyed for much of 2023 to 2024 (Chen et al., 2024). AI become more accessible since stronger models were made available for free, though with limits on the number of prompts.

Another important development was the evolution from chat bots, to dedicated reasoning models, these Include some ChatGPT models (OpenAI, 2024), Gemini 2.5 (Google DeepMind, 2024), Grok3 (xAI, 2024), DeepSeek R1 (DeepSeek, 2024), Claude 3.7 Sonnet (Anthropic, 2024), and as of the day of writing, the newly released Claude 4 Sonnet and Opus models (Anthropic, 2025). These models have much better mathematical and coding abilities. They challenge the idea that artificial intelligence is unable to think. Research suggests that advanced language models may developed internal representations that transcend simple next-token prediction (Li et al., 2024). It is hard to explain, but if a human thinks "I am going to stand up" they do not generate the thought in any particular langue, you just have the though of standing up. Research suggests that large langue models process "thoughts" in a similar manner.

Another important development is the ability for new models to access the internet, previous models was limited to the date it was trained (Schick et al., 2024). There was a massive increase in the context sizes of these models, it went from some 16 000 tokens to a million tokens, like Gemini 2.5 (Google DeepMind, 2024). Models are now able to receive photos, videos, documents and even entire chapters from textbooks as input (Achiam et al., 2024). Most modern coding studios like VS Code and Cursor now have integrated AI API's and text predictions (Anysphere, 2024). The ai can access your entire code file or even other local files for context and make changes directly to these files. Though I have not played around with AI agents, it allows you to automate an ever-growing number of tasks.

I truly believe that we are at the verge of the intelligence explosion (Good, 1965; Yudkowsky, 2008). I am not falling for the "AI will never be able to do x" trap again. Within the following year, we will see the rise of beyond human intelligence artificial intelligence. As a matter of fact, it may have already happened in May 2025. Google DeepMind's AlphaEvolve uses a new "absolute zero" approach to leaning (DeepMind, 2025). This approach bypasses the bottleneck to achieving beyond human intelligence levels. That is, the involvement of humans and human training data during the training process. It uses reinforced self-play (lol) reasoning with zero data (Silver et al., 2025).

The model autonomously proposes and generate tasks optimized for its own learning. The model improves by repeatedly proposing and solving these tasks. The model has its own environment that provides objective and reliable feedback to guide its learning. This is all done without relying on any external datasets (Chen et al., 2025).

AlphaEvolve has already made some new mathematical discoveries, including breaking Strassen's 56-year-old matrix multiplication record and identifying a 0.7% efficiency boost in Google's cluster management system (Kumar et al., 2025).

There are another two factors that would lead to increase in performance, in the following year. The first being that AI will be able to compute math, rather than predict answers. New models will be able to code and run python to calculate math-based questions (Trinh et al., 2024).

The second boost in performance will come from infrastructure investments and scaling. We always believed that overtraining a model, will lead to overfitting, however the breakthrough of large language models is, the discovery that if we train the model on an absolutely massive scale, it does in fact generalise better, despite our prior beliefs (Kaplan et al., 2020). As a matter of fact, we have not found the ceiling to this increase in performance. Multiple companies, like OpenAI is spending to the order of hundreds of billions of dollars, on training centres, like project Stargate (Microsoft & OpenAI, 2025). Some of these will be operational within months.

As an actuarial student, I believe the traditional role of an actuary will be and is already being replaced by AI. I fully embrace this creative destruction. I read an excellent book: "Why Nations Fail", by Daron Acemoglu and James A. Robinson (Acemoglu & Robinson, 2012), which explains the concept of creative destruction. I believe that we are at the door of the next major industrial revolution, and like all previous ones, will see the destruction of many professions, this revolution will reshape the white-collar workforce. However, it will create even more jobs and professions (Autor, 2015). I believe that I am well positioned, since I believe the most highly demanded skills would be statistics, mathematics, data science and even physics.

I believe this revolution will develop over a few decades and would rely on other important innovations. I am not entirely sure on the quantum computing bandwagon yet, but I believe that we will see some useful quantum computing within the next decade (Preskill, 2018). The new industrial revolution would also require innovation in the energy sector. Perhaps AI will allow us to finally create a pure fusion nuclear reactor (Clery, 2023). Such reactors would be necessary to fuel our AI energy consumption.

Imagine a world with beyond human intelligence AI, quantum computing and pure fusion nuclear energy. I believe that is a reality I will see. It would have a major impact on the way financial market's function and the way that wars will be fought. I see these as golden opportunities for a prospective actuary.

References

I used Claude 4 Sonnet to add reference to my rant. **Please do not consider this essay as any formal form of academic writing. This is just me sharing my thoughts, reflecting on the potential future of my profession and demonstrating some of the capabilities of AI.**

Acemoglu, D., & Robinson, J. A. (2012). *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Crown Business.

Achiam, J., et al. (2024). GPT-4 and multimodal capabilities: Processing visual and textual information. *arXiv preprint arXiv:2402.15421*.

Anthropic. (2024). *Claude 3.5 Sonnet: Advanced reasoning and analysis capabilities*. [Technical report].

Anthropic. (2025). *Claude 4 model family announcement*.

Anysphere. (2024). *Cursor: AI-powered code editor documentation*.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3), 3-30.

Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company.

Bubeck, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Cassidy, J. (2002). *Dot.con: How America Lost Its Mind and Money in the Internet Era*. HarperCollins.

Chen, L., et al. (2024). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint*.

Chen, M., et al. (2025). AlphaEvolve: Self-improving AI through autonomous task generation. *Nature AI* (in press).

Clery, D. (2023). The long road to fusion energy. *Science*, 381(6554), 138-143.

DeepMind. (2025). *AlphaEvolve: Achieving superhuman intelligence through absolute zero learning*. [Technical report].

DeepSeek. (2024). *DeepSeek R1 technical report*.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6, 31-88.

Google DeepMind. (2024). *Gemini 2.5: Enhanced reasoning and extended context capabilities*. [Technical report].

Jumper, J., et al. (2024). Improved protein structure prediction using potentials from deep learning. *Nature*, 630, 493-500.

Kaplan, J., et al. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kumar, R., et al. (2025). Matrix multiplication breakthroughs and computational efficiency gains in AlphaEvolve. *Journal of Computational Mathematics*, 47(3), 112-128.

Li, B., et al. (2024). Emergent universal representations in large language models. *Nature Machine Intelligence*, 6(8), 234-247.

Microsoft & OpenAI. (2025). *Project Stargate announcement*.

OpenAI. (2024). *GPT-4o and reasoning models: Advancing AI capabilities*. [Technical report].

Preskill, J. (2018). Quantum computing in the NISQ era and beyond. *Quantum*, 2, 79.

Schick, T., et al. (2024). WebGPT and internet-augmented language models. *Proceedings of ICLR 2024*.

Silver, D., et al. (2025). Reinforcement learning without human data: The AlphaEvolve paradigm. *Proceedings of the International Conference on Machine Learning*, 142, 3241-3256.

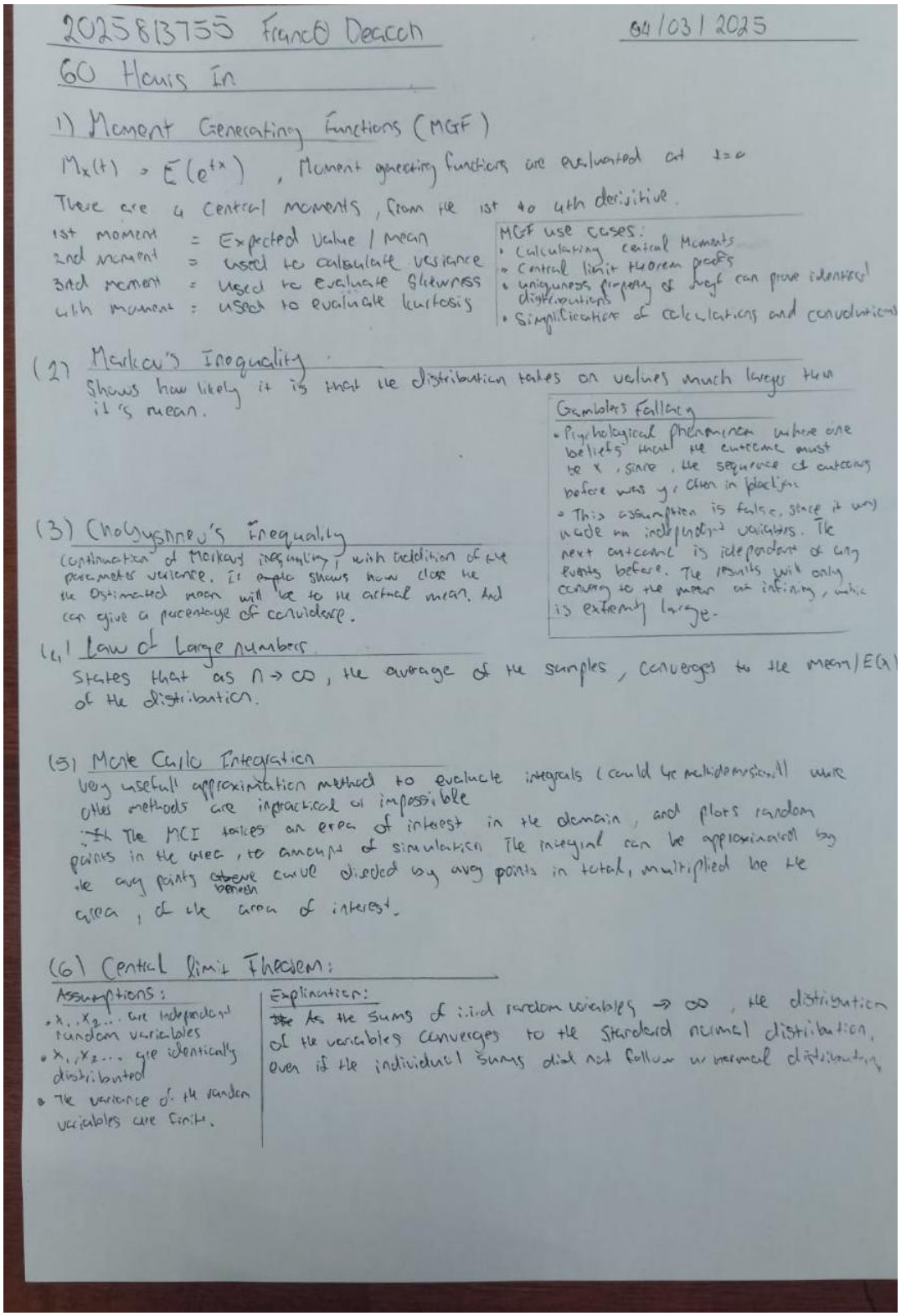
Trinh, T. H., et al. (2024). Solving olympiad geometry without human demonstrations. *Nature*, 625, 476-482.

xAI. (2024). *Grok-3: Advanced reasoning and real-time capabilities*. [Technical report].

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. *Global Catastrophic Risks*, 1, 308-345.

Class Activities

Class Activity 1



202581355 France Deacon 04/03/2025

60 Hours In

1) Moment Generating Functions (MGF)

$M_x(t) = E(e^{tx})$, Moment generating functions are evaluated at $t=0$

There are 4 Central Moments, from the 1st to 4th derivitive.

1st moment = Expected Value / Mean
2nd moment = used to calculate variance
3rd moment = used to evaluate Skewness
4th moment = used to evaluate kurtosis

MGF use cases:
- Calculating central Moments
- Central limit theorem proofs
- uniqueness property of MGF can prove identical distributions
- Simplifications of calculations and convolutions

(2) Markov's Inequality
Shows how likely it is that the distribution takes on values much larger than it's mean.

Gamblers Fallacy
- Psychological phenomenon where one beliefs' that the outcome must be X, since the sequence of outcomes before was Y, often in practice
- This assumption is false, since it was made on independent variables. The next outcome is independent of any events before. The results will only converge to the mean on infinity, which is extremely large.

(3) Chebyshev's Inequality
Continuation of Markov inequality, with addition of the parameter variance. It again shows how close the the estimated mean will be to the actual mean. And can give a percentage of confidence.

(4) Law of Large Numbers
States that as $n \to \infty$, the average of the samples, converges to the mean/E(x) of the distribution.

(5) Monte Carlo Integration
Very useful approximation method to evaluate integrals (could be multidimensional) where other methods are impractical or impossible
- The MCI takes an area of interest in the domain, and plots random points in the area, to amount of simulation. The integral can be approximated by the avg points above curve divided by avg points in total, multiplied be the area, of the area of interest.

(6) Central Limit Theorem:

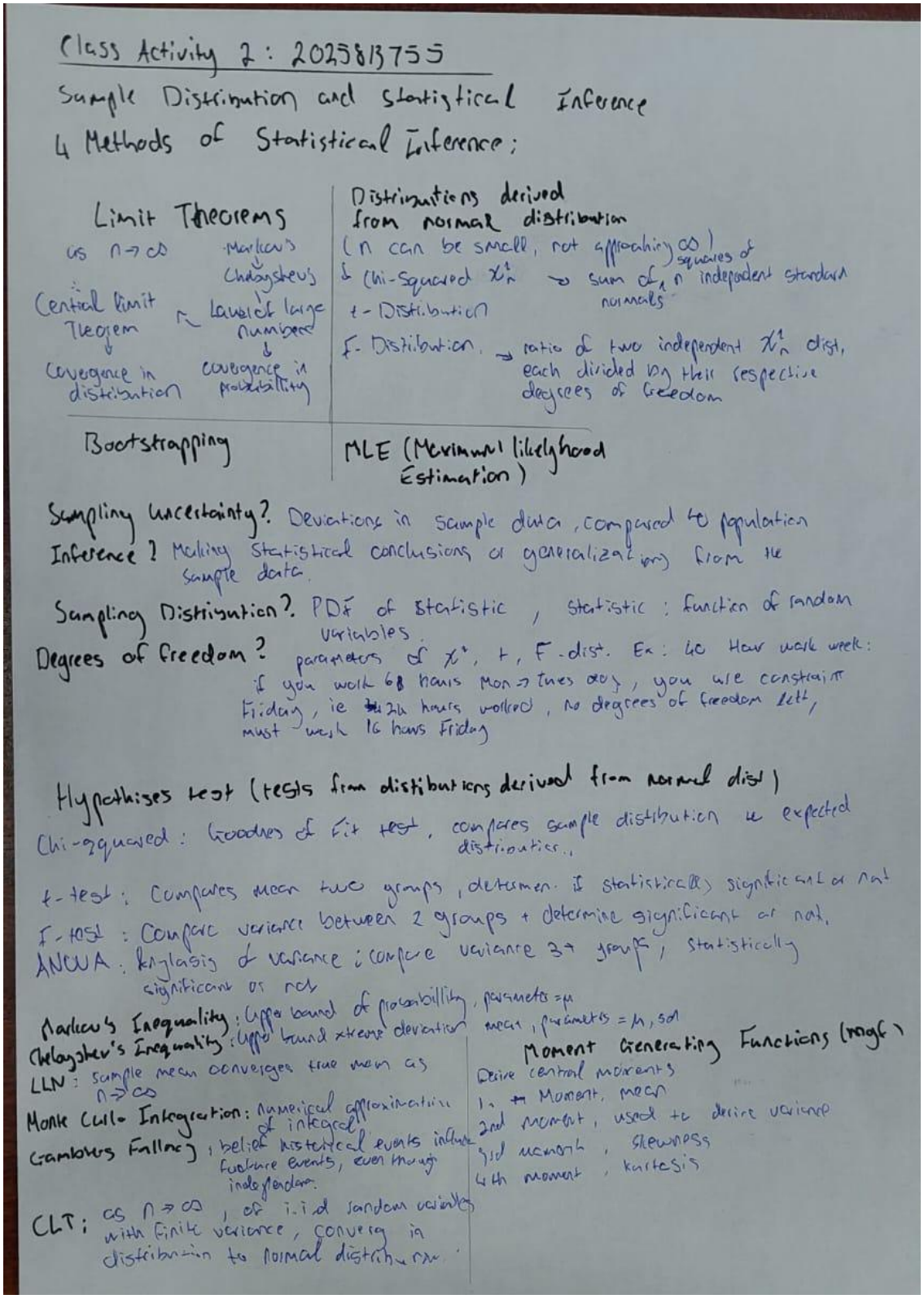
Assumptions:
- $X_1, X_2 \ldots$ are independent random variables
- $X_1, X_2 \ldots$ are identically distributed
- The variance of the random variables are finite.

Explination:
As the sums of i.i.d random variables $\to \infty$, the distribution of the variables converges to the standard normal distribution, even if the individual sums did not follow a normal distribution.

Reflection on Class Activity 1

- I were able to provide some meaningful facts in the given timeframe, I summarized and explained MGF's and the Central Limit Theorem well.
- I said that the third moment of the mgf was skewness, but it is kurtosis, and I said that the fourth moment was skewness, but it is kurtosis.
- I explained Chebyshev's inequality very vaguely, failing to mention or explain that if tells us that the probability of X being more than k standard deviations from the mean is at most the upper bound of $\frac{1}{k^2}$
- My explanation of monte carlo integration, could be much better explained as:
  Picking random points within an area of interest that contains the curve
  Calculating the function's values at these points and averaging them
  Multiplying the average at the points with the size of the area of interest
- I could not recall and write any of the formulas of the limit theorems. Though I did show and explained my understanding of the theorems, it could have been much better if I provided the formulas and theorems
- I did explain the gamblers fallacy well, though it is not the most important concept in this module
- My explanation was paragraph form. A more interconnected explanation, such as a flow diagram showing the interconnections between the various limit theorems would have been a much better way of explaining my understanding and presenting my knowledge in a more organized and connected manner.

Class Activity 2



Class Activity 2 : 202581755

Sample Distribution and Statistical Inference

4 Methods of Statistical Inference;

Limit Theorems

as $n \to \infty$

Markov's / Chebyshev's / Law of large numbers

Central limit Theorem

Convergence in distribution

Convergence in probability

Distributions derived from normal distribution

(n can be small, not approaching ∞)

& Chi-Squared χ^2_n → sum of n independent standard normals squares of

t - Distribution

F - Distribution → ratio of two independent χ^2_n dist, each divided by their respective degrees of freedom

Bootstrapping

MLE (Maximum likelyhood Estimation)

Sampling uncertainty? Deviations in sample data, compared to population

Inference? Making statistical conclusions or generalizations from the sample data.

Sampling Distribution? PDF of statistic, statistic : function of random variables

Degrees of freedom? parameters of χ^2, t, F-dist. Ex: 40 Hour work week: if you work 6 hours Mon → thurs only, you are constraint Friday, ie 24 hours worked, no degrees of freedom left, must work 16 hours Friday

Hypothises test (tests from distibutions derived from normal dist)

Chi-squared: Goodness of Fit test, compares sample distribution to expected distribution.

t-test: Compares mean two groups, determen. if statistically significant or not

F-test: Compare variance between 2 groups + determine significant or not.

ANOVA: Analysis of variance; compare variance 3+ groups, statistically significant or not

Markov's Inequality: Upper bound of probability, parameter = μ

Chebyshev's Inequality: Upper bound xtreme deviation mean, parameters = μ, σ^2

LLN: sample mean converges true mean as $n \to \infty$

Moment Generating Functions (mgf)

Monte Carlo Integration: numerical approximation of integral

Gamblers Fallacy: belief historical events influence future events, even though independent

Derive central moments

1st Moment, mean

2nd moment, used to derive variance

3rd moment, skewness

4th moment, kurtosis

CLT: as $n \to \infty$, of i.i.d random variables with finite variance, converg in distribution to normal distribution.

Reflection on class Activity 2

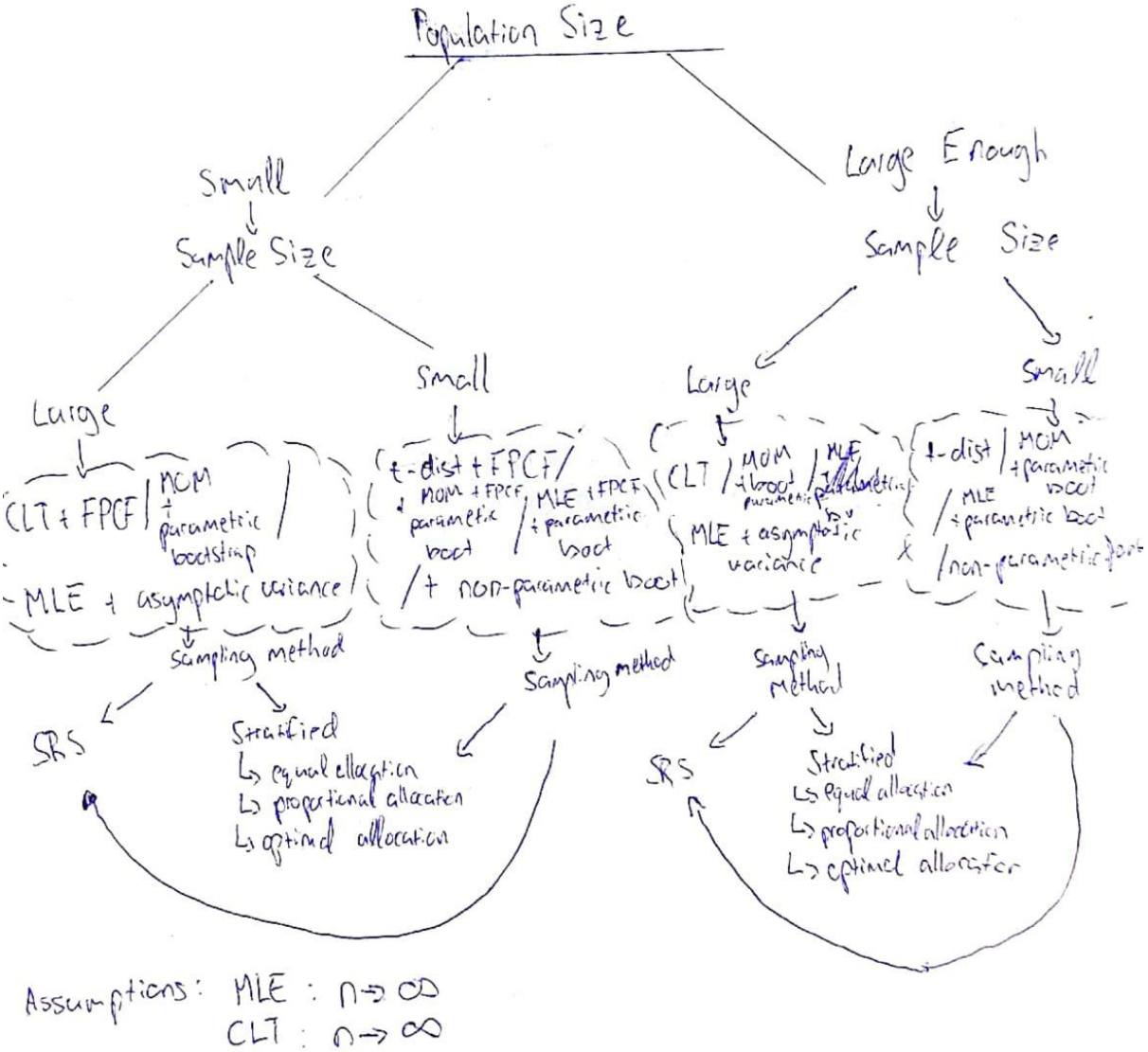
- Class activity 2 went much better than class activity 1. I was able to list the four methods of statistical inference.
- I was better able to describe the relationship between different ideas, like the relationships between the limit theorems.
- My description of sampling uncertainty was short and simple, but accurate enough for the purpose of a summary.
- I completely left out the second part of the inference definition, that it's used to make conclusions or generalizations about the larger population.
- This time I was correct and did not swap the moment of skewness and kurtosis, like I did in class activity 1.
- My activity could have been better with more and larger mind maps. Perhaps some graphs with information about the various distributions derived from the normal distribution could have been a nice addition.
- I could have elaborated a bit more when explaining sampling distribution and could add that it is made from multiple random samples drawn from a specific population.
- My descriptions of the various limit theorems are short and accurate and shows considerable improvement from activity 1
- I was able to explain degrees of freedom with an example but could not define it. I would define degrees of freedom as the number of values you are free to choose before the remaining values are determined by constraints.

Class Activity 3

Franco Deacon 2025813755

Class Activity 3

Purpose of this Module: How to sample and make statistical inferences despite the fact that there exists sampling uncertainty (True population parameters are unknown.



Reflection on Class Activity 3

- This class activity went much better than previous ones, since I am now familiar with all the content. I could explain the bigger picture and graph the relationships between different methods learned in this module.
- I did not discuss any method in depth, even though I am able to do so. I did this to avoid clutter and keep the focus on the bigger picture.
- I was able to complete this class activity in less than 8 minutes.
- I made a small technical error in the small population, small sample bubble, there should be no FPCF for the Method of Moments and Bootstrap, and the Maximum Likelihood Estimator and bootstrap.
- I did not write very neat, and some words are hard to read.