

MATH 10

INTRODUCTORY STATISTICS

Tommy Khoo

Your friendly neighbourhood graduate student.

It is Time for Homework...Again! (´•ω•`)

- Please hand in your homework.
- Second homework + data will be posted on the website, under the homework tab. And also sent out via email.
- **30%** weekly homework. Each homework might have different points assigned but carry the same weight.
- **Your other homework:** read and understand the relevant chapters in the textbook.

Week 3

- **Chapter 5 – Probability**

← today's lecture

Probability, gambler's fallacy, permutations and combinations, binomial distribution, Bayes' theorem.

- **Chapter 7 – Normal Distribution**

← today's lecture

What is the normal distribution? Areas under the curve, standard normal, normal approximation to binomial.

- **Chapter 8 – Advanced Graphs**

- **Chapter 9 – Sampling Distributions**

Chapter 5, Some Example Exam Questions

Question 1 : We have a population of size N . Let p be the independent probability of a person in the population developing a disease. *Answer the following questions in terms of N and p .*

1. What is the probability of a person in the population NOT developing the disease? (1 pt)

Chapter 5, Some Example Exam Questions

Question 1 : We have a population of size N . Let p be the independent probability of a person in the population developing a disease. *Answer the following questions in terms of N and p .*

1. What is the probability of a person in the population NOT developing the disease? *(1 pt)*
2. If $N = 2$, what is the probability that no one develops the disease? If your answer involves N , please replace it by the number 2. *(1 pt)*

Chapter 5, Some Example Exam Questions

Question 1 : We have a population of size N . Let p be the independent probability of a person in the population developing a disease. *Answer the following questions in terms of N and p .*

1. What is the probability of a person in the population NOT developing the disease? *(1 pt)*
2. If $N = 2$, what is the probability that no one develops the disease? If your answer involves N , please replace it by the number 2. *(1 pt)*
3. For a general size N , what is the probability that no one develops the disease? *(1 pt)*

Chapter 5, Some Example Exam Questions

Question 1 : We have a population of size N . Let p be the independent probability of a person in the population developing a disease. *Answer the following questions in terms of N and p .*

1. What is the probability that a person in the population NOT developing the disease? (1 pt)
2. If $N = 2$, what is the probability that no one develops the disease? If your answer involves N , please replace it by the number 2. (1 pt)
3. For a general size N , what is the probability that no one develops the disease? (1 pt)
4. What is the probability that, for a general size N , at least one person develops the disease? (1 pt)

• Tricks: $P(A \text{ and } B) = P(A)P(B)$, $P(A) = 1 - P(\text{not } A)$.

Chapter 5, Some Example Exam Questions

- **Question 2** : suppose 3 fair dice are rolled independently. Let their outcomes be D_1, D_2, D_3 . *Please simplify your answer as much as possible.*
 1. Suppose that $D_1 = 5$ and $D_2 = 6$, what is the probability that D_3 is a different outcome? (1 pt)

Chapter 5, Some Example Exam Questions

- **Question 2** : suppose 3 fair dice are rolled independently. Let their outcomes be D_1, D_2, D_3 . *Please simplify your answer as much as possible.*
 1. Suppose that $D_1 = 5$ and $D_2 = 6$, what is the probability that D_3 is a different outcome? (1 pt)
 2. In general, what is the probability that D_3 is different from D_1 and D_2 **GIVEN** that $D_1 \neq D_2$? (1 pt)

Chapter 5, Some Example Exam Questions

- **Question 2** : suppose 3 fair dice are rolled independently. Let their outcomes be D_1, D_2, D_3 . *Please simplify your answer as much as possible.*
 1. Suppose that $D_1 = 5$ and $D_2 = 6$, what is the probability that D_3 is a different outcome? (1 pt)
 2. In general, what is the probability that D_3 is different from D_1 and D_2 **GIVEN** that $D_1 \neq D_2$? (1 pt)
 3. In general, what is the probability that none of the rolls had the same outcome? (2 pts)

Chapter 5, Some Example Exam Questions

- **Question 2** : suppose 3 fair dice are rolled independently. Let their outcomes be D_1, D_2, D_3 . *Please simplify your answer as much as possible.*
 1. Suppose that $D_1 = 5$ and $D_2 = 6$, what is the probability that D_3 is a different outcome? (1 pt)
 2. In general, what is the probability that D_3 is different from D_1 and D_2 **GIVEN** that $D_1 \neq D_2$? (1 pt)
 3. In general, what is the probability that none of the rolls had the same outcome? (2 pts)
 4. What is the probability that at least two of the rolls had the same outcome? (1 pt)

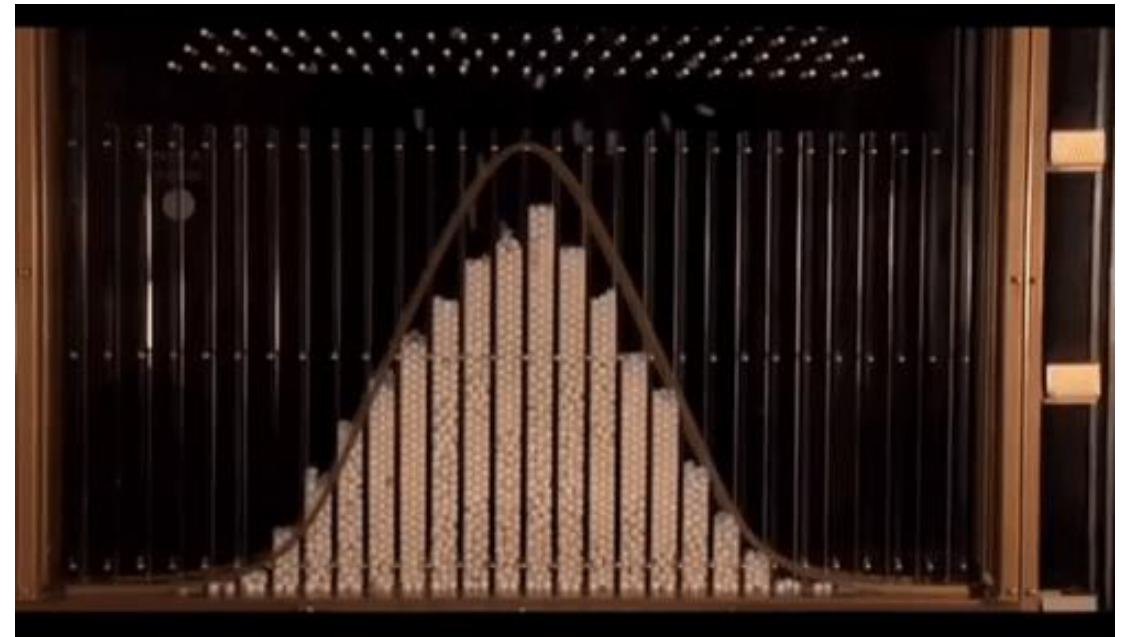
Chapter 5, Section 8 – The Binomial Distribution

- n trials, with probability p of success. → we say n, p are parameters
- If you repeat the experiment: “n trials” many times, count the number of successes each time and plot a histogram of the results, you get a **binomial distribution**.

- P(k success in those n trials) =

$$\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

- Combination: $nCk = \frac{n!}{k!(n-k)!}$



Galton board and video by Index Fund Advisors, Inc.

Chapter 5, Section 8 – The Binomial Distribution

- Some math associated with the binomial distribution.
- Cumulative probabilities.
- Textbook's example: toss a fair coin 12 times. What is the probability that we get between 0 to 3 heads?
- Answer: $P(0 \text{ heads}) + P(1 \text{ heads}) + P(2 \text{ heads}) + P(3 \text{ heads})$
- You can add them up because these events are mutually exclusive!
- (Population) Mean and Variance

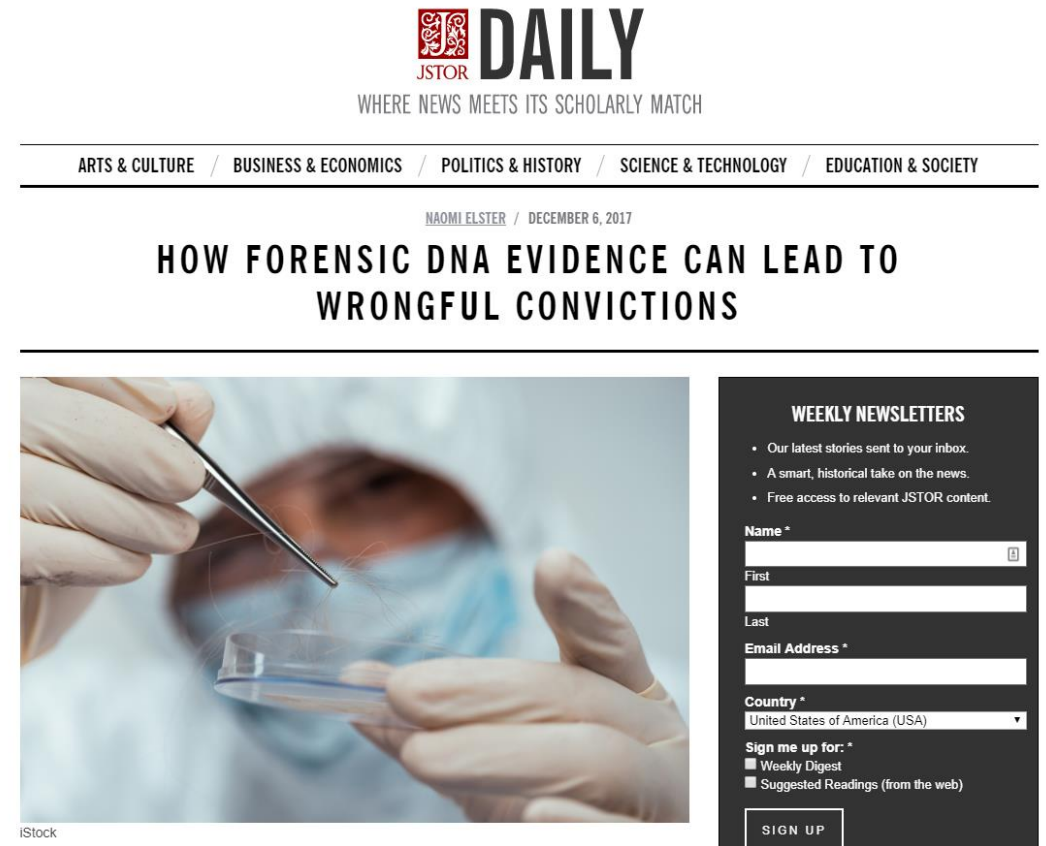
$$\mu = np \quad \sigma^2 = np(1 - p)$$

Public Service Announcement

- We are skipping these sections in Chapter 5.
- Section 10 – Poisson Distribution (*another limit of binomial*)
- Section 11 – Multinomial Distribution (*multiple types of success*)
- Section 12 – Hypergeometric Distribution (*drawn without replacement*)

Chapter 5, Section 13 – Base Rates

- Base rate = true proportion of a population having some condition, attribute or disease.
- The probability of positives that are false in tests depends heavily on the base rate.
- **Example:** a test for a disease is 99% accurate. You took the test and the result is positive. What is the probability that you actually have the disease?



JSTOR DAILY
WHERE NEWS MEETS ITS SCHOLARLY MATCH

ARTS & CULTURE / BUSINESS & ECONOMICS / POLITICS & HISTORY / SCIENCE & TECHNOLOGY / EDUCATION & SOCIETY

NAOMI ELSTER / DECEMBER 6, 2017

HOW FORENSIC DNA EVIDENCE CAN LEAD TO WRONGFUL CONVICTIONS

WEEKLY NEWSLETTERS

- Our latest stories sent to your inbox.
- A smart, historical take on the news.
- Free access to relevant JSTOR content.

Name *

First

Last

Email Address *

Country *
United States of America (USA)

Sign me up for: *

- Weekly Digest
- Suggested Readings (from the web)

Chapter 5, Section 13 – Base Rates

- **Example:** a test for a disease is 99% accurate. You took the test and the result is positive. What is the probability that you actually have the disease?
- The answer depends on the *base rate*, which is the proportion of people having the disease.
- Suppose 1 mil people are tested. 1% or 10,000 of them actually has the disease.
- Out of the 990,000 disease-free people, the test would produce 9,900 positives!
- Out of the 10,000 diseased people, the test would also produce 9,900 positives.
- Chance that a positive test result is correct = 50%!

Break Time!! \o/

- 10 minutes break starts after I have handed out the exercise.
- Question 1-5 should be “easy”.
- Question 6 is tough and not something you should expect during the exam.
- Question 6 was a job interview question (I am not joking ^_^).

Chapter 5, Section 13 – Base Rates

- **False Positive** : tests result shows positive, but you don't actually have the disease.
- **Bayes Theorem**

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|ND)P(ND)}$$

- D = have disease. ND = no disease. T = tested positive for disease.

Sample exam question with real world impact.

"The prosecutor's fallacy"

- Suppose we have a DNA comparison test, and a sample of the murderer's DNA. Let "innocent" be the event that a person is not the murderer. Let "match" be the event that the test says the person's DNA is a match to the murderer.
- Suppose that $P(\text{match} \mid \text{innocent}) = 0.1\% = 0.001$.
- And that $P(\text{match} \mid \text{not innocent}) = 1$.
- Suppose you apply this test to 1000 suspects and found a person who matched. Would you conclude that the probability that this person has a 0.1% chance of being innocent? Explain. (2 pts)

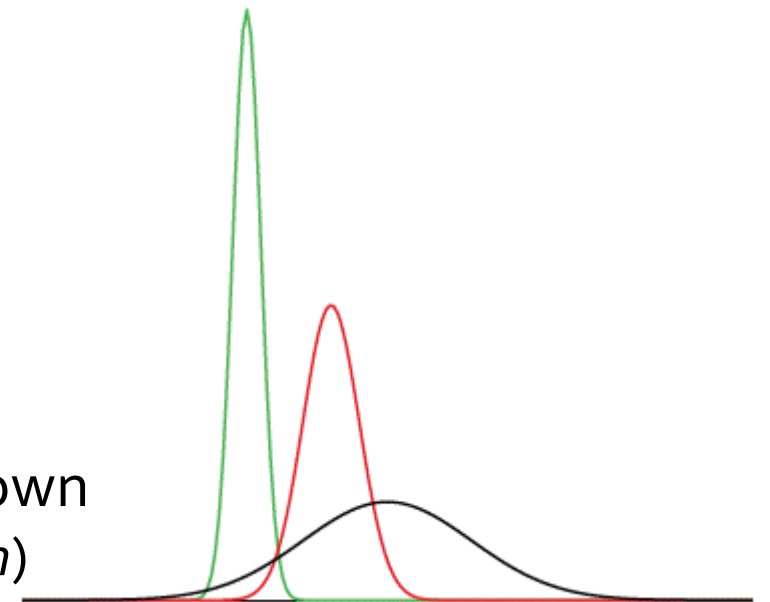
Chapter 7, Section 2 - Introduction

- The Normal distribution is also known as the Gaussian distribution.
- Was invented by Gauss as a model of measurement errors.

("The Evolution of the Normal Distribution" by Saul Stahl)

- It has two parameters: mean μ and variance σ^2 .
- Given these two parameters, you can draw the normal distribution as a curve.
- The horizontal axis goes from minus infinity to infinity.
- The value on the vertical axis is given by the function shown above. *(for illustration only, formula **NOT REQUIRED** for exam)*

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Chapter 7, Section 2 - Introduction

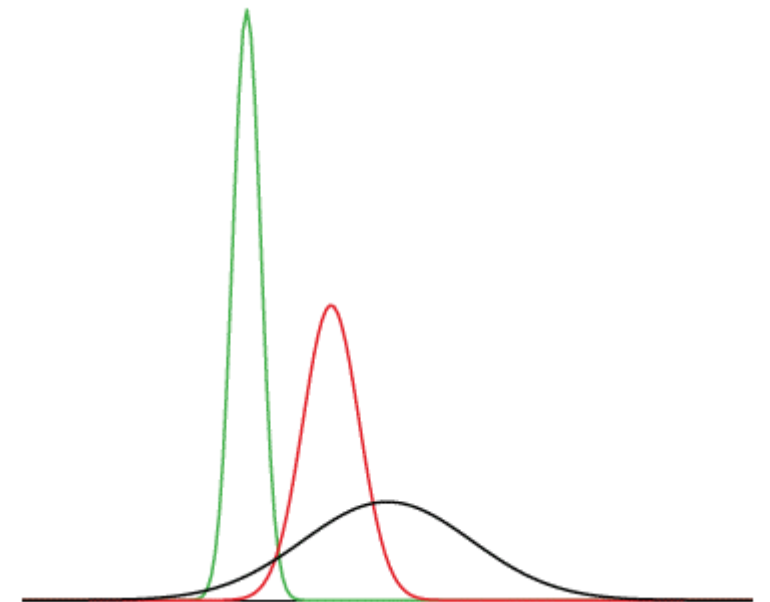
- **Some properties:**

- Symmetric around the mean.

(What does this say about the median and the mode?)

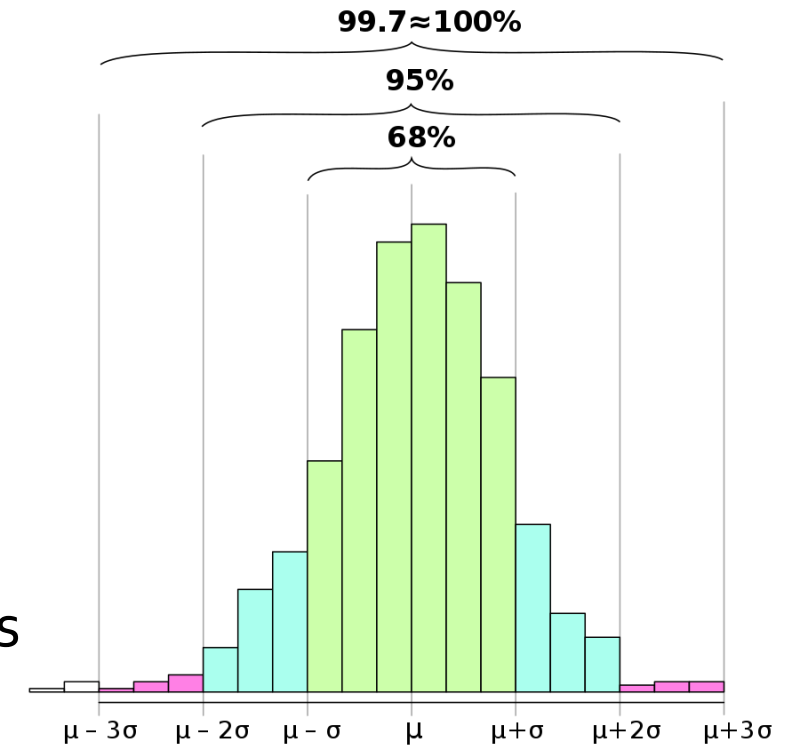
- Area under the normal curve is 1.
- Denser around the center, and less dense in the tails.
- Is a *continuous distribution* : gives you the probability of getting values within an interval. But the probability of getting a particular value is zero!

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Chapter 7, Section 4 – Areas Under Normal Distributions

- Take any interval $[a, b]$. The area under the curve, within this interval, is the probability that a normally distributed variable has value in $[a, b]$.
- The “68 – 95 – 99.7” rule.
- You should know this for the exam or at least remember enough to look it up in the “z-value table” that we will provide you with.
- For the exam, there are other games you can play with this concept.
- E.g. *The scores of 1,000 students in a class are normally distributed, with mean 50 and standard deviation 10.*
- *Approximately how many students scored 70 and above?*



The 68-95-99.7 rule in practice for an approximately normal histogram. Image from Wikipedia.

Chapter 7, Section 6- Standard Normal

- Special case: mean 0 and variance 1.
- Let X be a normal distributed (random) variable with mean μ and variance σ^2 .
- Apply the linear transformation: $Z = \frac{X - \mu}{\sigma}$. (why is this linear?)
- Linear transformations of normal variables are also normal.
- Using what we learned in the previous chapters:
 - What is the mean of the new variable Z ?
 - What is the variance of the new variable Z ?

Chapter 7, Section 7 – Normal Approximation to the Binomial Distribution.

- Remember what the Binomial distribution with parameters n and p is.
- If you thought the picture I shown previously looks like a normal curve, you are right!
- As n becomes large, and p is fixed, the binomial distribution becomes more and more like the normal distribution.
- For this course, we will tell you when you are **not supposed** to be using the normal approximation.



Galton board and video by Index Fund Advisors, Inc.

de Moivre–Laplace theorem (fun fact, not required for exam)

- Fix p , let n goes to infinity, then the binomial distribution goes to the normal distribution with mean np and variance $np(1 - p)$.
- Ok, actually, $\frac{X - np}{\sqrt{np(1-p)}}$ goes to standard normal.
- **Note:** the Poisson version has n goes to infinity, p goes to zero, while $np = c$ fixed.

