MATH 10

INTRODUCTORY STATISTICS

Tommy Khoo

Your friendly neighbourhood graduate student.

Tentative Plans For Midterm Exam

- Week 5 - 26 April (Thursday), 1 hour 45 minutes.
- In class, here in this room.

- Tentatively looking at 6 questions, 15 minutes per question.
- Will confirm details next week.

- This is a fast paced course. If you feel that you're lagging behind, I can review key points and work through sample problems during office hours (or by appointment).

Week 3

- **Chapter 5 – Probability**
- **Chapter 7 – Normal Distribution**

- **Chapter 9 – Sampling Distributions** ← **today's lecture**

Sampling distributions of the mean and p. Difference between means. Central Limit Theorem.

- **Chapter 8 – Advanced Graphs** ← **today's lecture**

Chapter 5, Section 13 – Base Rates

Sample exam problem with Bayes Theorem

Police officers in a town are stopping drivers at random and subjecting them to a breathalyzer test that detects alcohol intoxication.

During the operation, there were 5000 drivers on the road. One of those drivers were chosen at random, say Tommy, and he tested positive (>_< !!!).

Chapter 5, Section 13 – Base Rates

Sample exam problem with Bayes Theorem

Police officers in a town are stopping drivers at random and subjecting them to a breathalyzer test that detects alcohol intoxication.

During the operation, there were 5000 drivers on the road. One of those drivers were chosen at random, say Tommy, and he tested positive (>_< !!!).

The police were using breathalyzers that detects drunkenness with 100% probability of IF the driver is drunk. (positive result)

If the driver is sober, it has a 99% probability of reporting that the driver is NOT drunk. (negative result)

Chapter 5, Section 13 – Base Rates

During the operation, there were 5000 drivers on the road. One of those drivers were chosen at random, say Tommy, and he tested positive (>_< !!!).

The police were using breathalyzers that detects drunkenness with 100% probability of IF the driver is drunk. (positive result)

If the driver is sober, it has a 99% probability of reporting that the driver is NOT drunk. (negative result)

1. The police wants to deport Tommy because they think the probability that he is driving drunk is 99%. Is their claim true or false based on the information so far? Explain. (2 pts)

Chapter 5, Section 13 – Base Rates

The police were using breathalyzers that detects drunkenness with 100% probability of IF the driver is drunk. (positive result) If the driver is sober, it has a 99% probability of reporting that the driver is NOT drunk. (negative result)

Now you are told that out of the 5000 drivers on the road, 5 of them are drunk, while the rest are sober.

2. If all 4995 sober drivers are given the breathalyzer test, approximately how many would get a positive result? (1 pt)

3. The police was picked a driver at random from the 5000. What is the probability that he or she is sober? You do not have to simplify your answer. (1 pt)

Chapter 5, Section 13 – Base Rates

The police were using breathalyzers that detects drunkenness with 100% probability of IF the driver is drunk. (positive result) If the driver is sober, it has a 99% probability of reporting that the driver is NOT drunk. (negative result)

Now you are told that out of the 5000 drivers on the road, 5 of them are drunk, while the rest are sober.

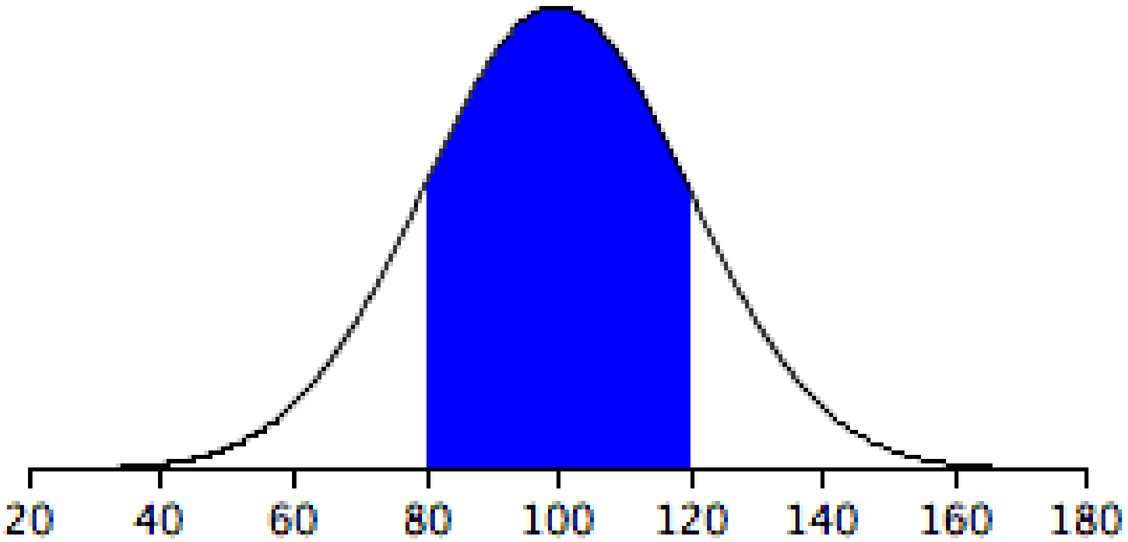
4. Using Bayes Theorem, write down a numerical fraction for the probability that a randomly selected driver is not drunk, given that he or she got a positive result. (4 pts)

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 $[a, b]$.

Normal distribution
mean = 100
standard deviation = 20

Blue area = 0.68 or 68%

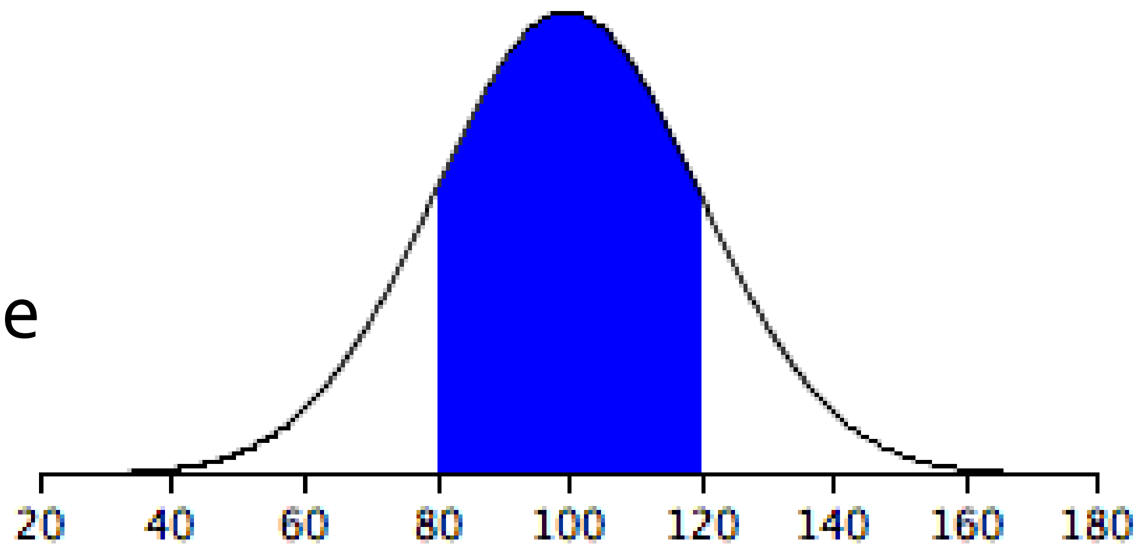


Chapter 7, Section 4 – Areas Under Normal Distributions

- Take any interval $[a, b]$.
- The area under the curve, within this interval, is the probability that a normally distributed variable has value in $[a, b]$.
- Alternatively, for an approximately normal set of data, the normal curve tells us the proportion of data with values within [a,b].

Normal distribution
mean = 100
standard deviation = 20

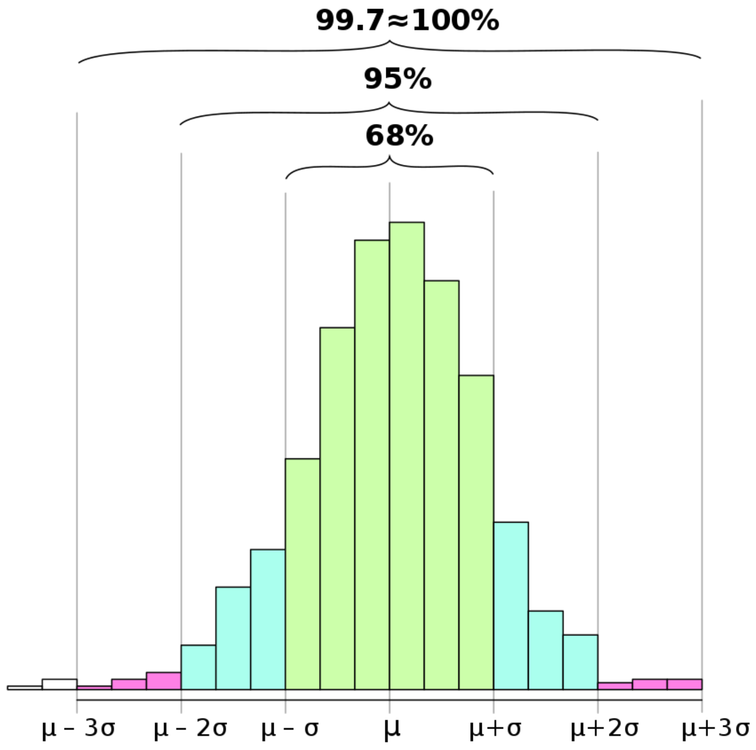
Blue area = 0.68 or 68%



Chapter 7, Section 4 – Areas Under Normal Distributions

- The "68 – 95 – 99.7" heuristic.
- You should know this for the exam.

- …or at remember enough to look it up a z-value table.



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

Chapter 7, Section 4 – Areas Under Normal Distributions

Example exam question:

- *The proportion of scores for 1,000 students in a class are well approximated by a normal distribution, with mean 50 and standard deviation 10.*

- *Approximately how many students scored 70 and above? (2 pts)*
- *Show your work. Illustrations are allowed.*

Chapter 7, Section 6 – Standard Normal

- Alternative way to solve this question.

- Let X be a normal distributed (random) variable with mean μ and variance σ^2.
- Apply the linear transformation: $Z = \frac{X-\mu}{\sigma}$.
- Now Z is standard normal.

→ normally distributed, mean 0, variance 1

Chapter 7, Section 6 – Standard Normal

- Apply the linear transformation: $Z = \frac{X - \mu}{\sigma}$.
- Now Z is standard normal. → normally distributed, mean 0, variance 1

- *The proportion of scores for 1,000 students in a class are well approximated by a normal distribution, with mean 50 and standard deviation 10.*
- *Approximately how many students scored 70 and above? (2 pts)*

- Will be used a lot for hypothesis testing later in the course.

Chapter 7, Section 4 – Areas Under Normal Distributions

Let's try a tougher sample exam question:

- *The weights for 500 bars of gold are well approximated by a normal distribution, with mean 100 grams and standard deviation 20 grams.*

- *Approximately how bars of gold weigh between 100 grams and 130 grams inclusive? (2 pts)*

Chapter 7, Section 4 – Areas Under Normal Distributions

- *The weights for 500 bars of gold are well approximated by a normal distribution, with mean 100 grams and standard deviation 20 grams.*
- *Approximately how bars of gold weigh between 100 grams and 130 grams inclusive? (2 pts)*

- *Two ways to think about this…*

1. *Area under curve as proportion of those 500 bars of gold.*
2. *Area under curve as the probability of getting a bar of gold within certain weights.*

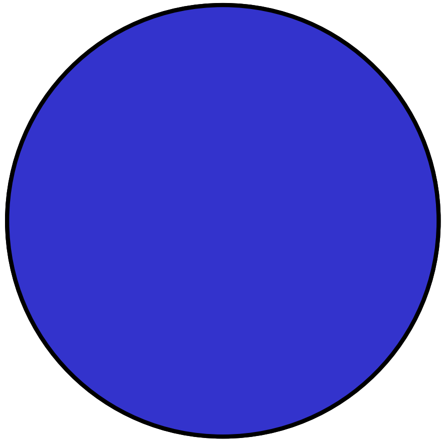
Break time!! \o/

12 minutes

- Break starts after I hand out the exercise.

- Circle is a timer that becomes blue. O_o →

(please ignore if it glitches)



Chapter 9, Section 2 - Introduction

- Inferential statistics.
- Have population, and a variable of interest, X.
- Take a simple random sample, calculate estimators for certain aspects of the population distribution of X.

- E.g. sample mean = estimator for population mean.
- E.g. estimator of the sample variance.

- We will now quantify how "good" these estimators are.

Chapter 9, Section 6 – Sampling Distribution of the Mean

- The sample mean, of a simple random sample A with size n, is a random variable.

- If you collect another simple random sample B with size n, it is likely to have a different sample mean.

- If $\bar{X}_n$ is a random variable that represents the mean of a sample of size n, then $\bar{X}_n$ has a distribution.

Chapter 9, Section 6 – Sampling Distribution of the Mean

- The distribution of $\bar{X}_n$ is the sampling distribution of the mean

(of a sample of size n).

- This distribution has mean $\mu_M = \mu$, where μ is the population mean.

- This distribution has variance $\sigma_M^2 = \frac{\sigma^2}{n}$, where σ^2 is the population variance.

Chapter 9, Section 6 – Sampling Distribution of the Mean

- Sampling Distribution of the Mean has,

$$\mu_M = \mu \qquad \sigma_M^2 = \frac{\sigma^2}{n}$$

- Standard error, $\sigma_M = \frac{\sigma}{\sqrt{n}}$.

Chapter 9, Section 6 – Sampling Distribution of the Mean

- Sampling Distribution of the Mean has,

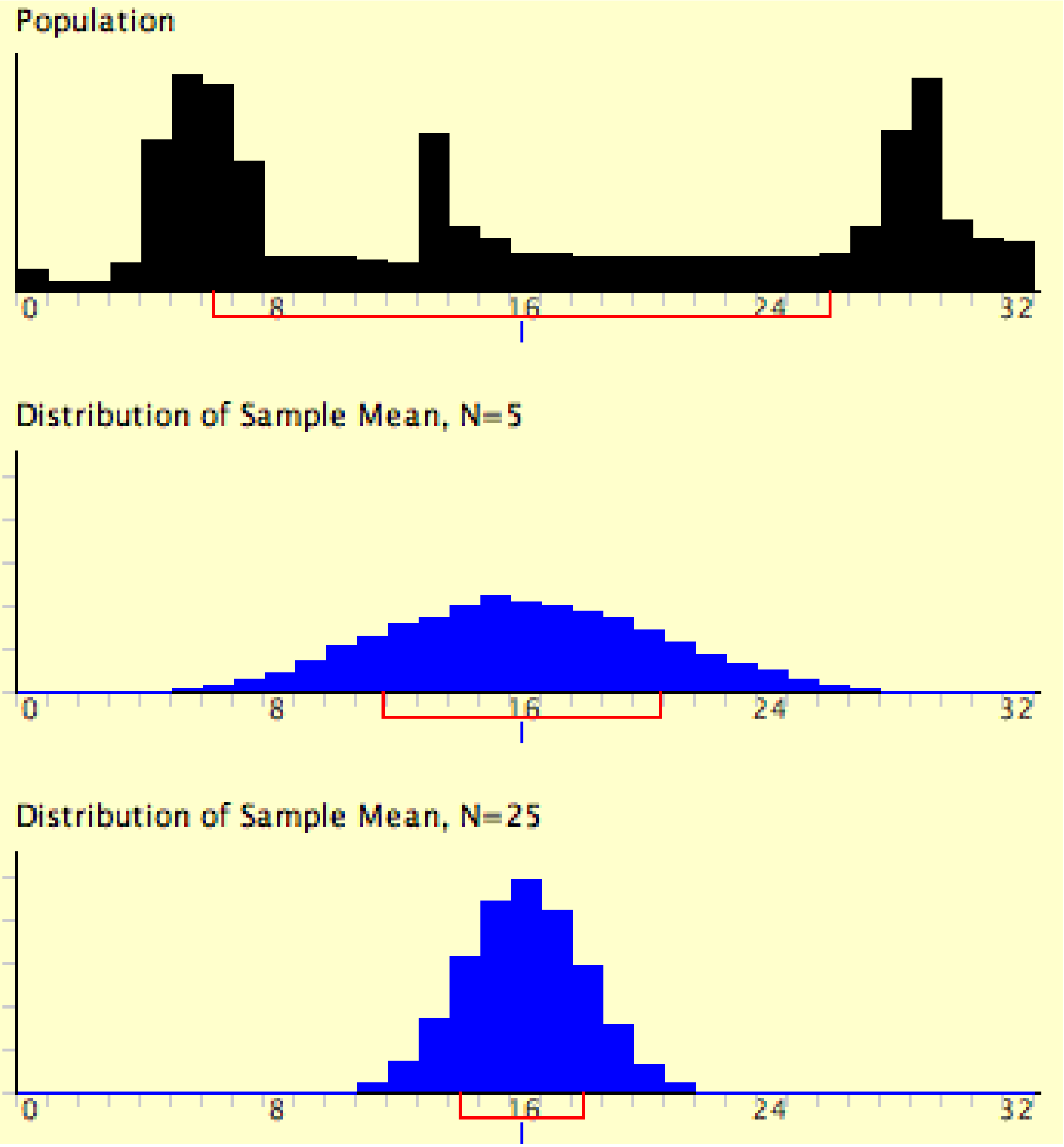
$$\mu_M = \mu \qquad \sigma_M^2 = \frac{\sigma^2}{n}$$

- Standard error, $\sigma_M = \frac{\sigma}{\sqrt{n}}$.

- **Central Limit Theorem!!!** ᕕ(ᐛ)ᕗ

If the population has finite mean μ, and finite non-zero variance σ^2, then the sampling distribution of the mean becomes better approximated by a normal distribution N(μ, $\frac{\sigma^2}{n}$), as sample size n increases.

Chapter 9, Section 6 – Sampling Distribution of the Mean



- Central limit theorem works for **any** distribution with finite mean and finite non-zero variance.

Chapter 9, Section 7 – Difference Between Means

- Finally, we can use statistics to compare two populations.

- Suppose you have two simple random samples with size n_1 and n_2.
- Samples from population 1 and 2 respectively.

- Calculate their sample means M_1 and M_2.
- The difference has a sampling distribution with mean

$\mu_{M_1 - M_2} = \mu_1 - \mu_2$.

Chapter 9, Section 7 – Difference Between Means

- The difference has a sampling distribution with mean $\mu_{M_1-M_2} = \mu_1 - \mu_2$.
- And variance $\sigma^2_{M_1-M_2} = \sigma^2_{M_1} + \sigma^2_{M_1}$.

- $\sigma^2_{M_i} = \frac{\sigma^2}{n_i}$, which is variance of the sampling distribution of M_i.
- Since the sample means are independent (as random variables), the variance sum law was used to derive the variance.

- $\sigma^2_{M_1-M_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Chapter 9, Section 7 – Difference Between Means

- The difference has a sampling distribution with mean $\mu_{M_1 - M_2} = \mu_1 - \mu_2$.

- And variance $\sigma^2_{M_1 - M_2} = \sigma^2_{M_1} + \sigma^2_{M_1} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

- Standard error $\sigma_{M_1 - M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

- This becomes much easier if the sample sizes and population variances are equal.

Public Service Announcement

- We are skipping “Chapter 9, Section 8, Sampling Distribution of r”.
- This chapter is about the sampling distribution of the correlation coefficient.
- Not usually taught at Math 10 level.
- So we’re nuking it from orbit (it’s the only way to be sure).