STAT1201 - Semester One 2022

Lecture 4 - Probability Distributions and Sampling Distributions

Dr. Wasanthi Thenuwara

08/12/2022

1

Lecture 4 - Probability Distributions and Sampling Distributions

In this lecture, you will practice

- Binomial Distribution
- Normal Distribution
- Sampling distribution of the sample mean
- Probability calculations using the distribution of the sample mean

2

Binomial Distribution

- In lecture 3, we focused on discrete random variables and defined the discrete probability distribution. Binomial distribution is an important discrete probability distribution.
- We use the concept of Bernoulli trial to describe the Binomial distribution.
- A Bernoulli trial is a random process with only two possible outcomes. These outcomes are usually labelled "success" and "failure".

Define P(Success) = *p.* This is a constant.

Consider a series of independent Bernoulli trials and count the number of successes.

Let X be the number of successes from *n* number of independent Bernoulli trials and P(Success) = *p*.

Then we call X has a Binomial distribution with parameters *n* and *p*. Mathematically represent,

$$X \sim Binom(n, p)$$

3

Binomial Distribution

Poll Question 1

Suppose you toss a coin three times. Let X be the number of heads. What is the probability distribution of X?

1. $X \sim Binom(3, 0.5)$
2. $X \sim Binom(2, 0.5)$
3. $X \sim Binom(3, 0.25)$
4. $X \sim Binom(2, 0.25)$

4

Binomial Distribution

Suppose you toss a coin three times. Let X be the number of heads. The probability distribution function of X can be written in a table as follows.

X	P(X=x)
0	0.125
1	0.375
2	0.375
3	0.125

The general formula to find the probabilities from a Binomial distribution is;
P(X = x) = $\binom{n}{x} p^x (1-p)^{n-x}$

Use this formula to calculate the probabilities in the table above.

5

Binomial Distribution

Example

The length of lizards living in one island in Australia has mean length of 50cm. Based on the previous research it is known that 60% of these lizards length is above the mean length. A random sample of 5 lizards is selected. What is the probability that

a) Exactly 2 lizards' length is above the mean length?
b) Less than 2 lizards' length is above the mean length?
c) At least 2 lizards' length is above the mean length?

6

Binomial Distribution

Example solution

Let X be the number of lizards whose length is above the mean

X = 0, 1, 2, 3, 4, 5

Then X ~ Binom(5, 0.6)

a) P(X=2) = $\binom{5}{2}0.6^2(0.4)^{5-2}$

```
P(X=2) = 0.2304
```

We will use dbinom() in R to find this probability

```
dbinom(x=2, size=5, prob=0.6)
```

b) P($X < 2$) = P(X = 0) + P(X = 1)

```
P(X= < 2) = 0.08704
sum(dbinom(x=0:1, size=5, prob=0.6))
```

c) Do by yourself (Homework)

7

Binomial Distribution

Poll Question 2

The length of lizards living in one island in Australia has mean length of 50cm. Based on the previous research it is known that 60% of these lizards' length is above the mean length. A random sample of 5 lizards is selected. What is the probability that at least 2 but no more than 4 lizards' length is above the mean length?

1. 0.3174
2. 0.5760
3. 0.8352
4. 0.6826

8

Binomial Distribution - Mean and Standard Deviation of X

X ~Binom(n, p)

Mean = E(X) = np

Var(X) = $np(1-p)$

sd(X) = $\sqrt{np(1-p)}$

9

Normal Distribution (also called as Gaussian Distribution)

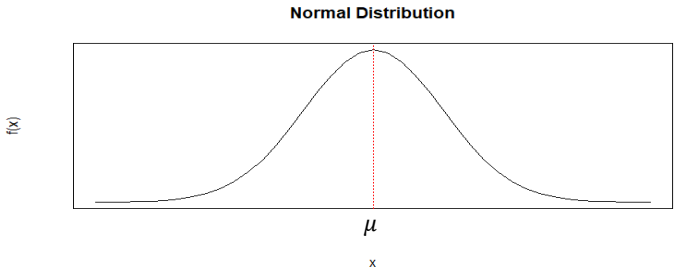
In Lecture 3, we focused on continuous random variables and continuous probability distributions.

- Normal distribution is the most famous continuous probability distribution.
- Two parameters are used to describe a Normal distribution. They are mean (μ) and standard deviation (σ).
- Let X be a continuous random variable. If X has a Normal distribution, we write X ~Normal (μ, σ).
- Bell shaped and symmetrical about the mean (μ).

10

Normal Distribution

- Location is determined by the mean (μ)
- Spread is determined by the standard deviation (σ)
- the random variable X has an infinite theoretical range from - ∞ to + ∞



11

Normal Distribution

Probability calculations

From lecture 3, you know that we need to consider the area under the continuous probability density functions to calculate the probabilities.

There is a rough rule to calculate the areas under Normal density curve.

- the area within 1 standard deviation of the mean is 68%
- the area within 2 standard deviation of the mean is 95%
- the area within 3 standard deviation of the mean is 97%

The area under the Normal density curve is 1.

12

Normal Distribution

Poll Question 3

In a Normal probability density curve what is the area left to the mean?

1. 0.25
2. 0.5
3. 0.75
4. 1

What is the area right to the mean?

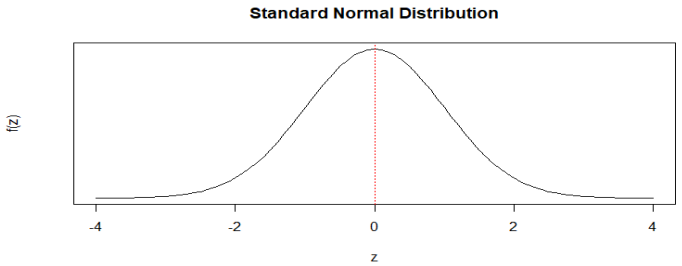
13

Normal Distribution - Probability calculations

There is no simple formula for working out areas under the Normal density curve. We can transform the Normal distribution to a Standard Normal distribution.

X ~ Normal(μ,σ)

Then Z = $\frac{X-\mu}{\sigma}$ and Z ~Normal(0, 1)



14

Normal Distribution - Probability calculations

Example

The body length of lizards living in one island in Australia follows a Normal distribution with a mean of 50cm and a standard deviation of 8cm.

a) What is the probability that a randomly selected lizard has a body length is less than 45cm?

b) What is the probability that a randomly selected lizard has a body length is between 45cm and 55cm?

c) Suppose that you have a 20% chance of finding a lizard less than a certain length. What is the maximum length of lizard you could expect to find?

15

Normal Distribution - Probability calculations

Example Solution (Please watch the lecture recording).

16

Normal Distribution - Probability calculations

Poll Question 4

The body length of lizards living in one island in Australia follows a Normal distribution with a mean of 50cm and a standard deviation of 8cm.

What is the probability that a randomly selected lizard has a body length longer than 58cm?

1. 0.1587
2. 0.8413
3. 0
4. 1

17

Sampling Distribution of the Sample Mean

What is sampling distribution of the Mean?

The distribution of all possible sample means using the same sample size, selected from a population.

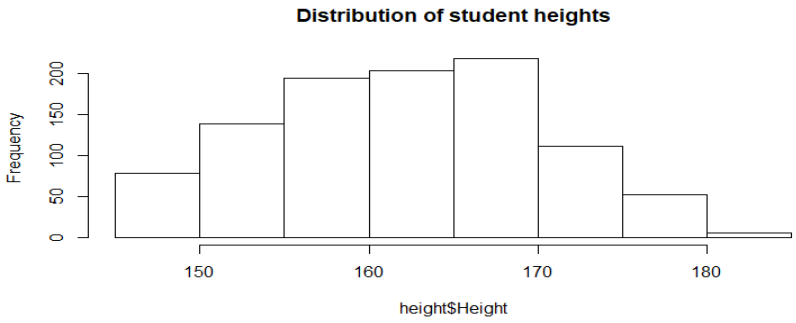
Suppose that our population is the STAT1201 students. Assume that 1000 students have enrolled in 2022. That is N=1000. We measure the heights of all 1000 students in cm and record them in "M4StudentHeights.csv" file. We can calculate the population mean (μ) and population standard deviation (σ) of height.

```
Population mean =  162.1504 cm
Population Standard Deviation =  8.147348 cm
```

18

Sampling Distribution of the Sample Mean

The distribution of student heights are as follows.



19

Sampling Distribution of the Sample Mean

Now we take a random sample of size 4 (i.e. n=4) and calculate the sample mean ($\overline{X}$) and sample standard deviation (s) of this sample.

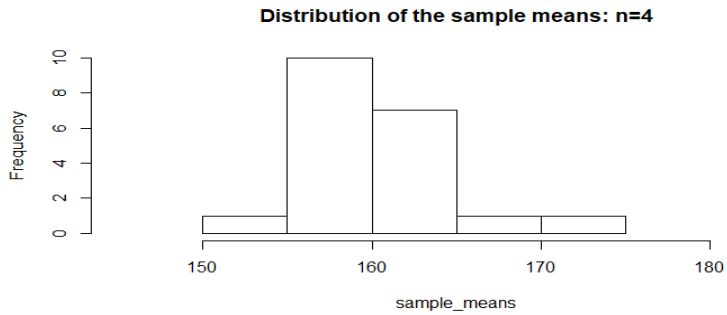
```
Sample mean =  159.975
Sample Standard deviation =  9.30927
```

Now we take 20 samples, each of size 4 from the same student height population and calculate means and standard deviations for each sample. The following histogram shows the distribution of the sample means.

Also, we can treat sample means ($\overline{X}$) as a random variable and calculate the mean and the standard deviation of that 20 sample means.

20

Sampling Distribution of the Sample Mean




```
Mean of the sample means = 161.95
Standard Deviation of the sample means = 4.619
```

Compare this with the population mean (μ) and population standard deviation (σ).

21

Sampling Distribution of the Sample Mean

Now we take 20 samples, each of size 16 from the same student height population.

```
Mean of the sample means = 162.55
Standard Deviation of the sample means = 2.095
```

Compare this with the population mean (μ) and population standard deviation (σ).

22

Sampling Distribution of the Sample Mean

Now we take 20 samples, each of size 25 from the same student height population.

```
Mean of the sample means = 162.53
Standard Deviation of the sample means = 1.521
```

Compare this with the population mean (μ) and population standard deviation (σ).

23

Sampling Distribution of the Sample Mean

Now increase the sample size to 100.

```
Mean of the sample means = 162.36
Standard Deviation of the sample means = 0.780
```

Compare this with the population mean (μ) and population standard deviation (σ).

24

Sampling Distribution of the Sample Mean

Sample Size	Mean of the sample means	SD of the sample sample means
4	161.95	4.619
16	162.55	2.095
25	162.53	1.521
100	162.36	0.78
Population mean	162.1504	
Population SD		8.8147348

You observe that the mean of the sample means are very closer to the population mean. The standard deviation of the sample means are becoming smaller as the sample size increases. The ratio of the standard deviation of the sample means to the population standard deviation is $\frac{1}{\sqrt{n}}$.

The distribution of $\overline{X}$ can be summarised as follows.

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

$sd(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

25

Sampling Distribution of the Sample Mean

We can use algebra to prove the formulas we have induced from the simulations we did.

Suppose $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with mean, μ and variance, σ^2.

Then we can show that

$E(\overline{X}) = \mu$ and $Var(\overline{X}) = \frac{\sigma^2}{n}$

Hence

$sd(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

Try by yourself using the concepts learnt in the previous lecture.

26

Sampling Distribution of the Sample Mean

- If the population is normally distributed, the sampling distribution of sample means ($\overline{X}$) is normally distributed.
- The population variable of interest may NOT be normally distributed (eg: skewed), or the shape of the population variable distribution is unknown.

Under what conditions will the sampling distribution of the sample mean be normally distributed?

Central Limit Theorem

As the sample size increases, the sampling distribution of sample means ($\overline{X}$) becomes approximately normally distributed regardless of the shape of population variable distribution.

We will use this theorem in subsequent lectures; calculation of confidence intervals and hypothesis testing.

27

Sampling Distribution of the Sample Mean

Poll Question 5

The body length of lizards living in one island in Australia follows a Normal distribution with a mean of 50cm and a standard deviation of 8cm. You are taking 10 samples of each size 4 from this population. The mean and the standard deviation of the sample means approximately are

1. 12.5cm and 4cm respectively
2. 50cm and 1cm respectively
3. 5cm and 4cm respectively
4. 50cm and 4cm respectively

28

Summary

If X ~ Normal(μ, σ)

$Z = \frac{X-\mu}{\sigma}$ and Z ~ Normal(0, 1)

The distribution of sample means ($\overline{X}$)

$\bar{X} \sim Normal(\mu, \frac{\sigma}{\sqrt{n}})$

Then $Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ and Z ~ Normal(0, 1)

Even the population is not normally distributed, for sufficiently large *n*, From the Central Limit Theorem,

$\bar{X} = Normal(\mu, \frac{\sigma}{\sqrt{n}})$

Then $Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ and Z ~ Normal(0, 1)

29

Sampling Distribution of the Sample Mean

Applications

The body length of lizards living in one island in Australia follows a Normal distribution with a mean of 50cm and a standard deviation of 8cm. Suppose you are taking a random sample of 4 lizards from this population. What is the probability that average length of lizards is between 52cm and 54cm? (We will discuss this in the class)

30

Sampling Distribution of the Sample Mean

Poll Question 6

Consider the example we just discussed. Now suppose that you take a random sample of 16 lizards. Without doing any calculations the probability of average length of lizards between 52cm and 54cm will be

1. smaller than that of taking a sample of size 4
2. larger than that of taking a sample of size 4
3. stay the same
4. cannot determine

31

Sampling Distribution of the Sample Proportions

The sample proportion ($\hat{p}$)

Again, consider our height of STAT1201 students. Now define *p* as the population proportion of students whose height is less than or equal to 155cm.

Following similar steps like we did for the sampling distribution of the sample mean we can find the distribution of the sample proportion.

Let $\hat{p} = \frac{x}{n}$

where x is the number of students in the sample whose height is less than or equal to 155cm.

If all possible samples of size *n* are selected from a binomial distribution, We can show that

$E(\widehat{p}) = p$

$sd(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}$

32

Sampling Distribution of the Sample Proportions

Provided that *n* is large such that $np > 5$ and $n(1-p) > 5$, we can show that

$$\hat{p} \sim Normal\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Then

$Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ and Z ~Normal(0, 1)

We will use this in subsequent lectures.

33

Next …

Reminders:

Quizzes 3 and 4 are closed at 3:00 pm on Monday, 12 Dec.

Lecture 5 – Statistical Inference

Tuesday, 13 Dec 2022 at 12:00 via Zoom

(818 1453 7986)

Lecture 6 – Ethical Case Studies by Dr. Julian Lamont

Tuesday, 13 Dec 2022 at 12:00 via Zoom

(818 1453 7986)

34