Session 4: Samples and sampling distributions

Susan Thomas
http://www.igidr.ac.in/˜susant
susant@mayin.org

IGIDR
Bombay

Recap

- Probability and principles
- Random variables
- Probability distributions and densities
- Parameters of probability distributions and densities
- Data descriptors

The aim of this session

1. Samples and populations
2. Parameters vs. statistics
3. Sampling distribution of statistics
4. Central Limit Theorem
5. Estimating the population mean

Samples and populations

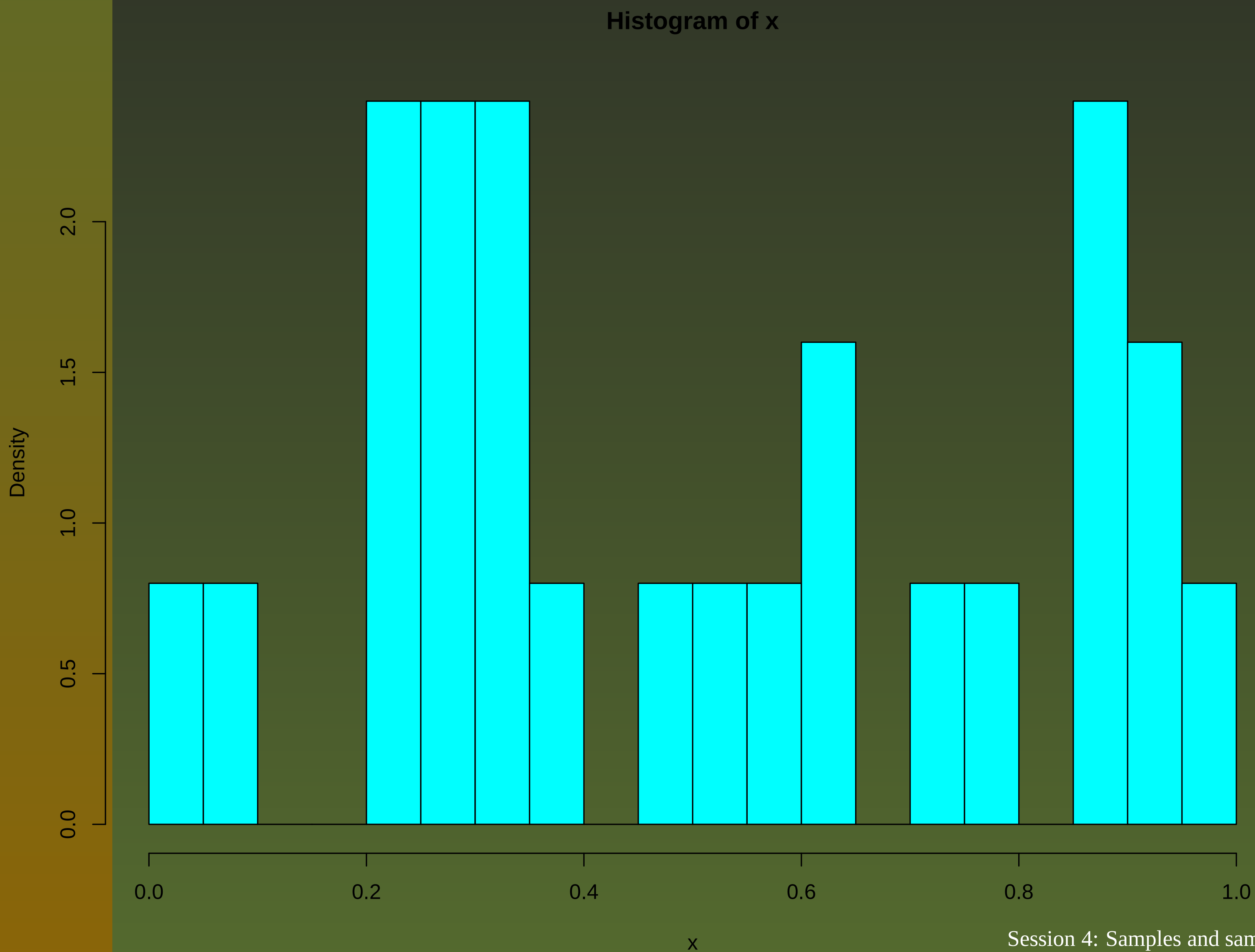
Samples

- A sample is a subset of the possible values for a RV.
- All the possible values is called the population.
- Samples are available; populations are not.
- Samples are used to infer characteristics of the population.
- A good sample is *representative* of the population, in terms of
  1. The values
  2. The frequency of the values

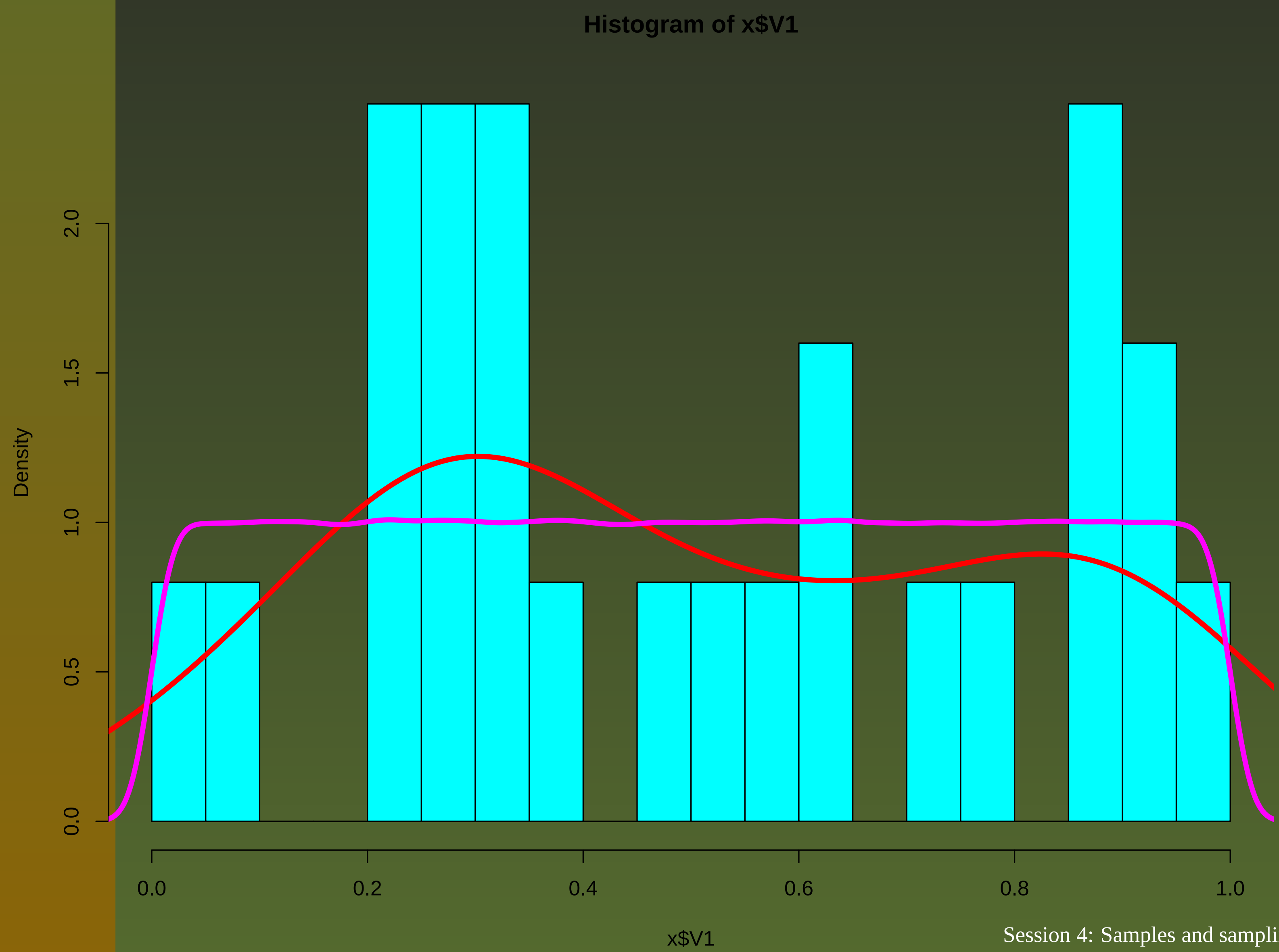
Example from a uniform distribution, u

- The population is U(0,1).
- u = Sample of size 25
- In **R**, the command
  `u = runif(25)`
  generates a sample of 25 draws from U(0,1)

Histogram of u



Histogram of u



Summary statistics

	Population	Sample
Mean	1.000	0.504
Sigma	0.289	0.294
1^{st} Quartile	0.250	0.275
3^{rd} Quartile	0.750	0.792

Second sample from U(0,1) $u2$

- The population is U(0,1).
- $u2$ = Sample of size 25
- In **R**, the command
  $u2$ `= runif(25)`
  generates a second sample of 25 draws from U(0,1)

Summary statistics

	Population	Sample1	Sample2
Mean	1.000	0.504	0.422
Sigma	0.289	0.294	0.203
1^{st} Quartile	0.250	0.275	0.311
3^{rd} Quartile	0.750	0.792	0.534

Third sample from U(0,1) $u3$

- The population is U(0,1).
- $u2$ = Sample of size 2500
- In **R**, the command
  $u2$ `= runif(2500)`
  generates a second sample of 2500 draws from U(0,1)

Summary statistics

	Population	Sample1	Sample2	Sample3
Mean	0.500	0.504	0.422	0.494
Sigma	0.289	0.294	0.203	0.289
1^{st} Quartile	0.250	0.275	0.311	0.237
3^{rd} Quartile	0.750	0.792	0.534	0.741

Statistics vs. Parameters

- A numerical descriptive measure of a population is called a parameter.

  For example, the bernoulli distribution has one parameter p, and the normal distribution has two parameters μ, σ.

- A quantity calculated from a sample set of observations of the RV is called a statistic.
- Parameters of a given distribution are **constant**.
- Statistics calculated for different samples from the same distribution are different – they are **random variables**!

  For example, the mean of a sample is a RV is a random variable.

Sampling distributions

Sampling distribution of sample statistics

- Like all RVs, sample statistics have probability distributions – these are called sampling distributions.
- The sampling distribution is the relative frequency distribution, theoretically generated by
  1. Repeatedly taking random samples (of size n) of the RV,
  2. Calculating the statistic for each sample
- Each sample used in generating the sampling distribution has to have the **same number of observations**.
- Therefore, the sampling distribution is generated for (a) a given statistic and (b) for a given sample size.

Examples of the sampling distribution of sample means, $\bar{x}$

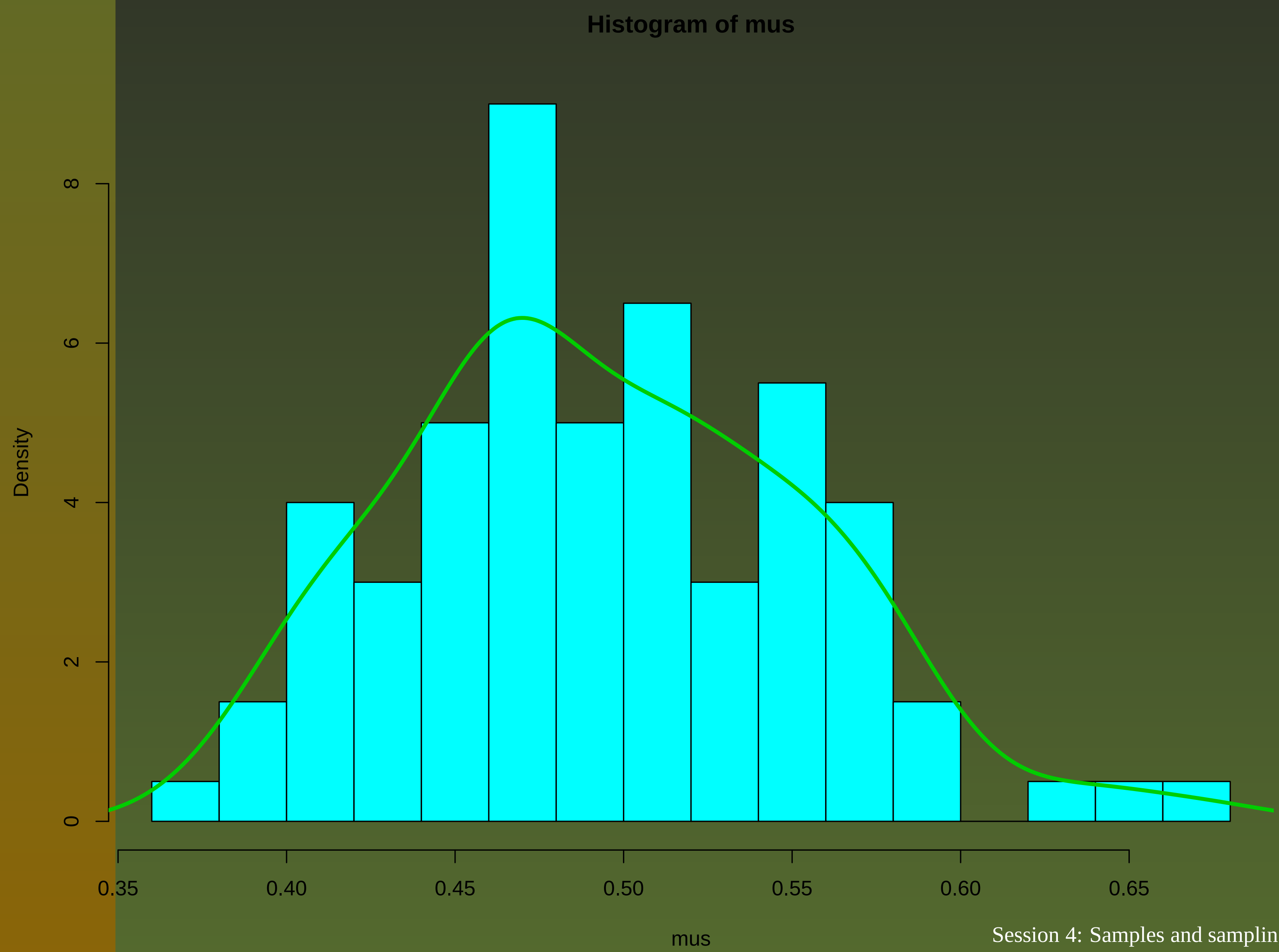
Example1: $n = 25$, U(0,1)

- We will create 100 samples to generate the sampling distribution of the mean from U(0,1), $n = 25$
  1. Sample 1, `S1 = runif(25)`
  2. Mean 1, `m[1] = mean(S1)`
- Run 1. and 2. in a loop to get 100 values for the mean stored in `m`.

Analysing the generated data

- Visually:
  1. To get the histogram of the means:
     `hist(m,breaks=20,freq=FALSE,col=5)`
  2. To superimpose the kernel density plot over the histogram:
     `lines(density(m),col=2,lwd=4)`
- Numerically:
  1. To get the mean, 1^{st} and 3^{rd} quartile:
     `summary(m)`

Example1: histogram of $\bar{x}$ for U(0,1), $n = 25$



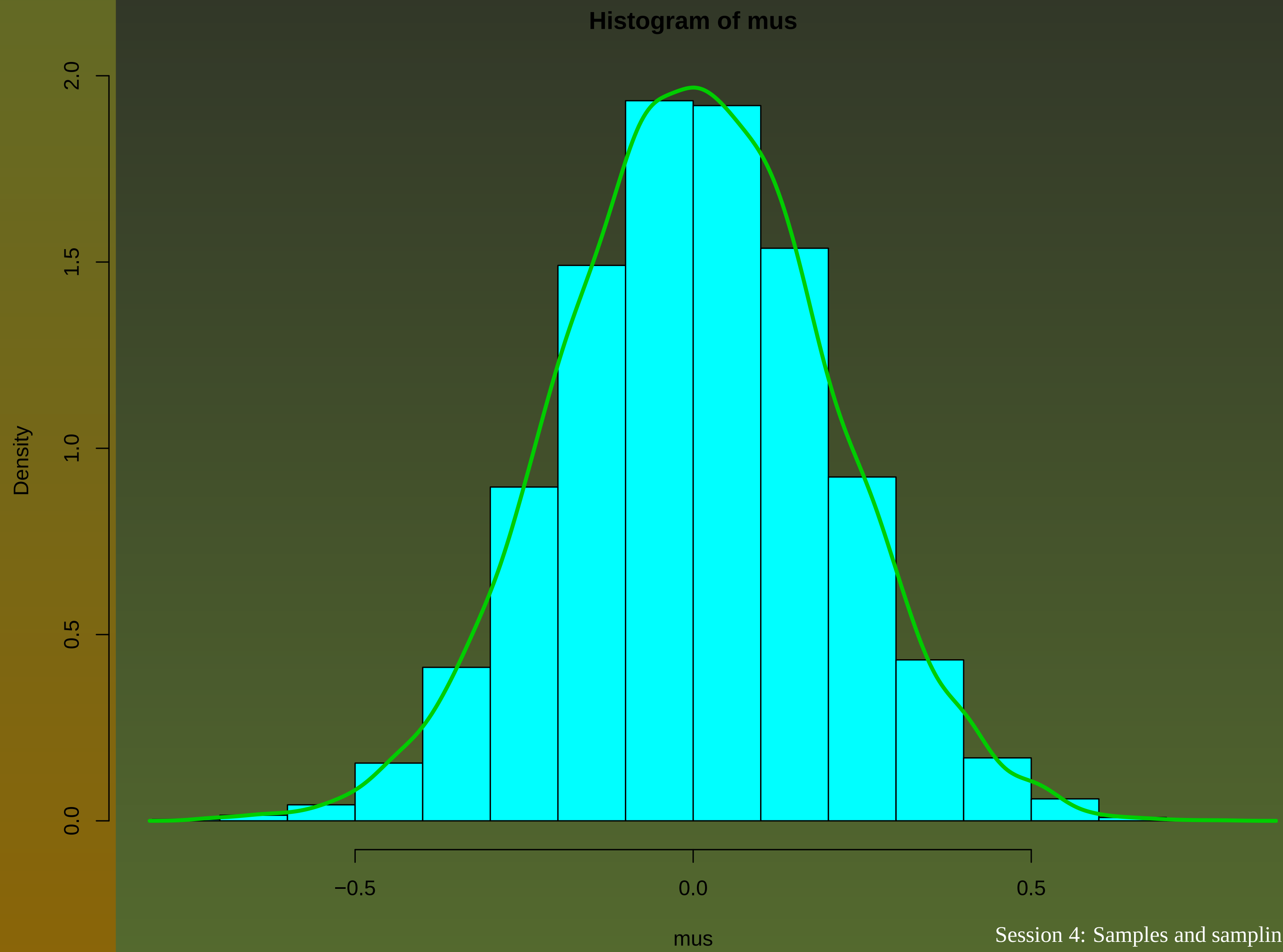
Example1: Summary statistics of $\bar{x}$

	$\bar{x}$
Mean	0.494
Standard deviation	0.061
1^{st} quartile	0.453
3^{rd} quartile	0.533

Example2: $n = 25$, N(0,1)

- We will create 100 samples to generate the sampling distribution of the mean from N(0,1), $n = 25$
  1. Sample 1, `S1 = rnorm(25)`
  2. Mean 1, `m[1] = mean(S1)`
- Run 1. and 2. in a loop to get 100 values for the mean stored in `m`.

Example2: histogram of $\bar{x}$ of N(0,1), $n = 25$



Example2: Summary statistics of $\bar{x}$

	$\bar{x}$
Mean	0.003
Standard deviation	0.199
1^{st} quartile	-0.131
3^{rd} quartile	0.136

Central Limit Theorem

Stating the CLT

- If the sample size is sufficiently large,
- the sample mean $\bar{x}$ has a sampling distribution that is approximately normal
- This is *irrespective* of the distribution of the underlying RV.

Properties of the sampling distribution of $\bar{x}$

If $\bar{x}$ is the mean of a sample of RVs from a population with mean parameter μ and standard deviation parameter σ, then:

1. The mean of the sampling distribution of $\bar{x}$ is called $\mu_{\bar{x}}$ and is equal to population μ.

$$\mu_{\bar{x}} = \mu$$

2. The standard deviation of the sampling distribution of $\bar{x}$ is called $\sigma_{\bar{x}}$ and is a fraction of the population σ, as:

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

Where n is the size of the sample.

Recap

1. Population
   - Parameters: μ, σ
   - This distribution generates the underlying RV.
2. Sample of size n:
   - Statistics: $\bar{x}, \bar{\sigma}$
   - This is a feature of one sample.
3. Distribution of sample statistics from samples of size n
   - Sampling distribution parameters: $\mu_{\bar{x}}, \sigma_{\bar{x}}$
   - This is a frequency distribution from a set of samples all the same size.

Testing concepts

A small-town newspaper reported that for families in their circulation area, the distribution of weekly expenses for food consumed away from home has an average of Rs.237.60 and a standard deviation of Rs.50.40. An economist randomly sampled 100 families for their outside-home food expenses for a week.

1. What is the distribution of the mean weekly outside-home food expenses for the 100 families?

2. What is the probability that the sample mean weekly expenses will be at least Rs.252?

Testing concepts

A small-town newspaper reported that for families in their circulation area, the distribution of weekly expenses for food consumed away from home has an average of Rs.237.60 and a standard deviation of Rs.50.40. An economist randomly sampled 100 families for their outside-home food expenses for a week.

1. What is the distribution of the mean weekly outside-home food expenses for the 100 families? It should be approximately normal distributed, with $\mu_{\bar{x}} = Rs.237.60$, and $\sigma = 50.40/\sqrt{100} = 5.50$
2. What is the probability that the sample mean weekly expenses will be at least Rs.252? We convert N(0,1) distribution as $z = 252 - 237.60/5.5 = 2.62$. Then Pr(average expenses > 252) = Pr(z > 2.62) $= 0.5 - 0.4956 = 0.0044\%$

Testing concepts: Comparing sampling distributions

Say $\bar{x}_{25}$ is the mean of a random sample of size 25 from a population of $\mu = 17, \sigma = 10$. Say $\bar{x}_{100}$ is the mean of a sample of size 100 selected from the same sample.

1. What is the sampling distribution of $\bar{x}_{25}$?

2. What is the sampling distribution of $\bar{x}_{100}$?

3. Which of the probabilities, P$(15 < \bar{x}_{25} < 19)$ or P$(15 < \bar{x}_{100} < 19)$, would you expect to be larger?

4. Calculate the probabilities in the third question.

Testing concepts: Comparing sampling distributions

Say $\bar{x}_{25}$ is the mean of a random sample of size 25 from a population of $\mu = 17, \sigma = 10$. Say $\bar{x}_{100}$ is the mean of a sample of size 100 selected from the same sample.

1. What is the sampling distribution of $\bar{x}_{25}$? Normal, $\mu_{\bar{x}_{25}} = 17, \sigma_{\bar{x}_{25}} = 2$
2. What is the sampling distribution of $\bar{x}_{100}$? Normal, $\mu_{\bar{x}_{25}} = 17, \sigma_{\bar{x}_{100}} = 1$
3. Which of the probabilities, P$(15 < \bar{x}_{25} < 19)$ or P$(15 < \bar{x}_{100} < 19)$, would you expect to be larger? The latter.
4. Calculate the probabilities in the third question. At $z = 1$, area=0.3413, $z = 2$, area=0.4772. Thus the first is 0.6826, the second is 0.9544.

Problems to be solved

Q1: Sampling distribution parameters

Suppose a random sample of $n = 100$ is selected from a population with μ, σ as follows. Find the values of $\mu_{\bar{x}}, \sigma_{\bar{x}}$

- $\mu = 10, \sigma = 20$
- $\mu = 20, \sigma = 10$
- $\mu = 50, \sigma = 300$
- $\mu = 100, \sigma = 200$

Q2: Sampling distribution probabilities

Suppose a random sample of $n = 225$ is selected from a population with $\mu = 70, \sigma = 30$. Find the following probabilities:

- $\Pr(\bar{x} > 72.5)$
- $\Pr(\bar{x} < 73.5)$
- $\Pr(69.1 < \bar{x} < 74.0)$
- $\Pr(\bar{x} < 65.5)$

Q3: Tobacco company research

Research by a tobacco company says that the relative freq. distribution of the tar content of a new low-tar cigarette has $\mu = 3.9$ mg of tar and $\sigma = 1.0$ mg. A sample of 100 low-tar cigs are selected from one-day's production and the tar content is measured:

- What is Pr(mean tar content of the sample) is greater than 4.15mg?
- Suppose that the sample mean tar content works out to be $\bar{x} = 4.18$ mg. Based on the first question, do you think the tobacco company may have understated the tar content?
- If the tobacco company's figures *are* correct, then rationalise the observed value of $\bar{x} = 4.18$ mg.

Q4: Mean of an exponential RV

The length of time between arrivals at a hospital clinic and the length of service are two RVs that are important in designing a clinic, and how many doctors/nurses are needed there. Suppose the relative freq. dist. of the interarrival time (between patients) has a mean of 4.1 minutes and $\sigma = 3.7$ minutes.

1. A sample of 20 interarrival times are selected and $\bar{x}$ the sample mean is calculated. What is the sampling distirbution of $\bar{x}$?
2. What is the Pr(the mean interarrival time) will be less than 2 minutes in this sample?
3. What is the Pr(the mean interarrival time) of the sample will exceed 6.5 minutes?
4. Would you expect $\bar{x}$ to be greater than 6.5 minutes? Explain.

References

- Chapter 7,