Session 3: Summarising probability distributions and density functions

Susan Thomas

http://www.igidr.ac.in/˜susant

susant@mayin.org

IGIDR

Bombay

Recap

- Discrete and continuous random variables
- Probability distributions
  A table of all the discrete values the RV can take, and it's associated probability.
- Probability density functions
  A function mapping a values that the RV can take and the probability in a ϵ region around the value.
- Cumulative distributions and density functions

Goals of this session

1. Expectation of RVs
2. Expectation of functions of RVs
3. Moment generating functions
4. Describing data

Expectations of RVs

Expectation of RVs

- E(x)
  (Also called “Average, arithmetic mean, mean, expected value, a measure of location, a measure of central tendency”)
- For a discrete RV:

$$\mathrm{E}(x) = \sum_{i=1}^{n} x_i \mathrm{Pr}(x_i)$$

  where Pr(x) is the probability distribution of x.
- For a continuous RV:

$$\mathrm{E}(x) = \int_{x=-\infty}^{x=\infty} x f(x) dx$$

Example: expectation of discrete variables

- Bernoulli RV: x = 0, 1; Pr(0) = p, Pr(1) = 1 - p

$$E(x) \;=\; 0 * p + 1 * (1-p) = 1-p$$

- Binomial RV: x = 0 . . . n, Pr(x) = $\begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}$

$$E(x) \;=\; \sum_{i=1}^{n} x_i \begin{pmatrix} n \\ x_i \end{pmatrix} p^{x_i} (1-p)^{n-x_i}$$

Testing concepts: expectation of a binary variable

Binary variable x, has the following PD:

x	Pr(x)
2	0.3
5	0.7

Questions:

1. What is E(x)?

Testing concepts: expectation of a binary variable

Binary variable x, has the following PD:

x	Pr(x)	Pr(x)
2	0.3	0.6
5	0.7	3.5

Questions:

1. What is E(x)? 4.1

Testing concepts: expectation of a discrete variable

RV x, can take the following discrete values, each with equal probability:

x	-1	2	5	7	10	11	12	15	20	30

Questions:

1. What is E(x)?

Testing concepts: expectation of a discrete variable

RV x, can take the following discrete values, each with equal probability:

x	-1	2	5	7	10	11	12	15	20	30
x*Pr(x)	-0.1	0.2	0.5	0.7	1.0	1.1	1.2	1.5	2.0	3.0

Questions:

1. What is E(x)?

Testing concepts: expectation of a discrete variable

RV x, can take the following discrete values, each with equal probability:

x	-1	2	5	7	10	11	12	15	20	30
x*Pr(x)	-0.1	0.2	0.5	0.7	1.0	1.1	1.2	1.5	2.0	3.0

Questions:

1. What is E(x)? 11.1

Testing concepts: expectation of a binomial variable

The binomial variable x comes from a distribution with $n = 5$ and $p = 0.2$. What is the expected value of x?

x	Pr(x)
0	0.3277
1	0.4096
2	0.2048
3	0.0512
4	0.0064
5	0.0003
E(x)	

Testing concepts: expectation of a binomial variable

The binomial variable x comes from a distribution with $n = 5$ and $p = 0.2$. What is the expected value of x?

x	Pr(x)	x*Pr(x)
0	0.3277	0
1	0.4096	0.4096
2	0.2048	0.4096
3	0.0512	0.1536
4	0.0064	0.0256
5	0.0003	0.0015
E(x)		0.9999

Example: expectation of continuous variables

- Uniform Continuous RV: x = [L,U]; Pr$(x_i) = p = 1/(U-L)$

$$
\begin{aligned}
E(x) &= \int_L^U \frac{x}{U-L}\, d(x) \\
&= \frac{1}{U-L} \int_L^U x \, d(x) = \frac{1}{U-L} \left(\frac{x^2}{2} \right)_L^U \\
&= \frac{U+L}{2}
\end{aligned}
$$

Example: expectation of continuous variables

- Normal RV: $x = [-\infty, \infty]$; $\Pr(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(x-\mu/\sigma)^2}$

$$
\begin{aligned}
E(x) &= \int_{-\infty}^{\infty} x\, f(x)\, d(x) \\
&= \frac{e^{1/2\sigma^2}}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} x\, e^{-\frac{1}{2}(x-\mu)^2}\, d(x) \\
\text{Set } y &= x - \mu \\
E(x) &= C \int_{-\infty}^{\infty} y\, e^{-\frac{1}{2}y^2}\, d(y) - \mu C \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\, d(y) \\
&= C \int_{-\infty}^{\infty} y\, e^{-\frac{1}{2}y^2}\, d(y) - \mu \int_{-\infty}^{\infty} f(x)\, d(x)
\end{aligned}
$$

Example: expectation of normal distribution

- First integral on the RHS:

$$\begin{aligned} \int y\, e^{-\frac{1}{2}y^2}\, d(y) &= \left(-e^{-\frac{y^2}{2}}\right)_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

- Then the expectation becomes:

$$\begin{aligned} E(x) &= 0 + \mu \int_{-\infty}^{\infty} f(x)\, d(x) \\ E(x) &= \mu \end{aligned}$$

Testing concepts: Expectation of a uniform RV

- Uniform Continuous RV: $x = [0,10]$. What is $E(x)$?

Testing concepts: Expectation of a uniform RV

- Uniform Continuous RV: x = [0,10]. What is E(x)?

$$
\begin{aligned}
E(x) &= \int_0^{10} x\, f(x)\, d(x) \\
&= \int_0^{10} x\, \frac{1}{10}\, d(x) \\
&= \frac{1}{10} \int_0^{10} x\, d(x) \\
&= \frac{1}{10} \left(\frac{x^2}{2} \right)_0^{10} \\
&= 5
\end{aligned}
$$

Testing concepts: Expectation of a constant

- x always takes a constant value, 5. What is E(x)?

Testing concepts: Expectation of a constant

- x always takes a constant value, 5. What is E(x)?
- Since it is not a random variable, the expectation is the value itself.

$$E(\text{constant}) = \text{constant} = 5$$

Expectations of functions of RVs

Expectations of functions of discrete RVs

- Function $g(x)$ of a random variable x is a random variable.
- g(x) has a probability density that is calculated from Pr(x).
- **For any x which is a discrete RV with a known probabilitiy distribution, Pr(x), the expectation of any function $g()$ of x is calculated as:**

$$E(g(x)) = \sum_{\text{min}}^{\text{max}} g(x)Pr(x)$$

Example: E(x^2) for a binary variable

- Bernoulli RV: x = 0, 1, Pr(x) = p, (1-p)
- g(x) = x^2 = 0, 1, Pr(g(x)) = p, (1-p)
- E(g(x)) = 0*p + 1*(1-p) = (1-p)

Example: E(x^2) for a discrete variable

- Discrete RV: x = $x_1, x_2, x_3, \ldots$, Pr(x)
- g(x) = x^2 = $x_1^2, x_2^2, x_3^2, \ldots$
- E(g(x)) = $\sum_{\text{min}}^{\text{max}} x^2 Pr(x)$

Testing concepts: binary variable

RV x is binary with the following probability distribution:

x	Pr(x)
2	0.3
5	0.7

Questions:

1. What is E(x^2)?

Testing concepts: binary variable

RV x is binary with the following probability distribution:

x	Pr(x)	x^2	x^2*Pr(x)
2	0.3	4	1.2
5	0.7	25	17.5

Questions:

1. What is E(x^2)? 18.7

Testing concepts: discrete variable

RV, x, can take the following discrete values with uniform probability:

x	-1	2	5	7	10	11	12	15	20	30

Questions:

1. What is E(2x + 5)?

Testing concepts: discrete variable

RV, x, can take the following discrete values with uniform probability:

x	-1	2	5	7	10	11	12	15	20	30
2x + 5	3	9	15	19	25	27	29	35	45	65
x*Pr(x)	0.3	0.9	1.5	1.9	2.5	2.7	2.9	3.5	4.5	6.5

Questions:

1. What is E(2x + 5)?

Testing concepts: discrete variable

RV, x, can take the following discrete values with uniform probability:

x	-1	2	5	7	10	11	12	15	20	30
2x + 5	3	9	15	19	25	27	29	35	45	65
x*Pr(x)	0.3	0.9	1.5	1.9	2.5	2.7	2.9	3.5	4.5	6.5

Questions:

1. What is E(2x + 5)? 27.2

Expectations of functions of continuous RVs

For any x which is a continuous RV with a known probabilitiy density, f(x), the expectation of any function $g()$ of x is calculated as:

$$E(g(x)) = \int_{\min}^{\max} g(x)f(x)dx$$

Example: E(x^2) for a continuous uniform variable

- Uniform Continuous RV: x = [L,U]; f$(x_i) = p = 1/(U - L)$
- g(x) = x^2

$$\begin{aligned} E(x^2) &= \int_L^U \frac{x^2}{U-L}\, d(x) \\ &= \frac{1}{U-L}\int_L^U x^2\, d(x) = \frac{1}{U-L}\left(\frac{x^3}{3}\right)_L^U \\ &= \frac{U^3 - L^3}{3 * (U-L)} \end{aligned}$$

- L=0, U=10; Pr(x_i) = 1/10, E(x^2) = 1000/30 = 33.3333

The variance of a distribution

- Examine, g(x) = $[x - E(x)]^2$
- The expectation of the squared value of the RV away from it's mean is the variance of the distribution. It is often denoted as either var(x) or σ^2.
- It is also called the "dispersion, dispersion around the mean, second moment around the mean".
- The square root of the variance ($\sqrt{\sigma^2}$) is called the standard deviation of the distribution.

Link between variance and mean

- The variance is linked with the mean as follows:

$$\begin{aligned}\sigma^2 &= E([x - E(x)]^2) \\ &= E(x^2 - 2xE(x) + E(x)^2) \\ E(2xE(x)) &= 2E(x)E(x) = 2E(x)^2 \\ \sigma^2 &= E(x^2) - E(x)^2\end{aligned}$$

Moment generating functions

Generalising the mean and the variance

- For any distribution, there can be a series of "moments" calculated as follows:

$$
\begin{aligned}
(\text{Discrete}) E(x^i) &= \sum_{\text{min}}^{\text{max}} x^i Pr(x) \\
(\text{Continuous}) E(x^i) &= \int_{-\infty}^{\infty} x \; f(x) \; d(x)
\end{aligned}
$$

- Each moment describes a feature of the distribution.

The unique moments of a distribution

- The moments are functions of the parameters of the distribution.
- For example, the bernoulli distribution had E(x) = E(x^2) = (1-p), the probability of success.
- Thus, every distribution has as many unique moments as parameters.
  The remainder of the moments can be expressed as functions of the parameters.
- For example, every moment of the normal distribution can be expressed as a function of the first two moments, the mean (μ) and the variance (σ^2).

Using what we’ve learnt so far

Describing data concisely

- One of the uses of the concepts of statistics to better describe data.
- The questions we try to answer are:
  1. What is the likely probability distribution/density?
  2. What are the parameters for the distribution/density?
- There are two kinds of tools: visual and numerical.

Visual tools

Graphical tools

- Eyeballing the data: is it continous or discrete?
- Most popularly used graphical tool: histograms or frequency distribution plots.
- Density plots: smoothed versions of histograms for continuous RVs.

Histograms

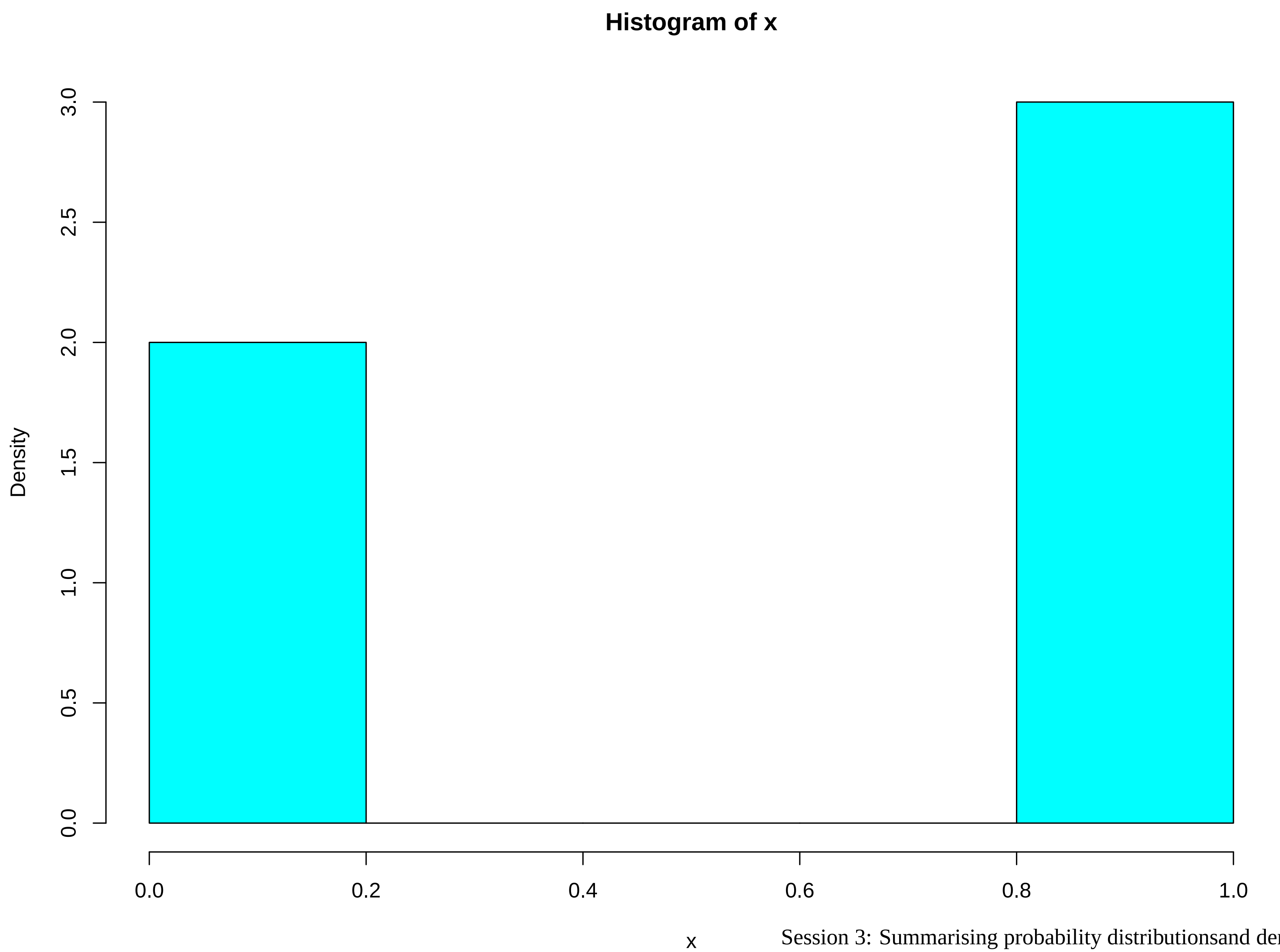
- The histogram is the plot of the unique values in a sample and the frequency with which they are observed.
- RV Value on the x-axis, frequency (with which the value occur in the data) on the y-axis.
- Histograms for the discrete case is easy: have to rework the goal a little for the continuous case.

Example 1: Discrete RV

1. The data is a set of 20 values.
2. $x = 0\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1$
3. It looks discrete. It looks binary.
4. Frequency table:

x	Freq
0	13
1	7

Histogram of x



Guessing the PD

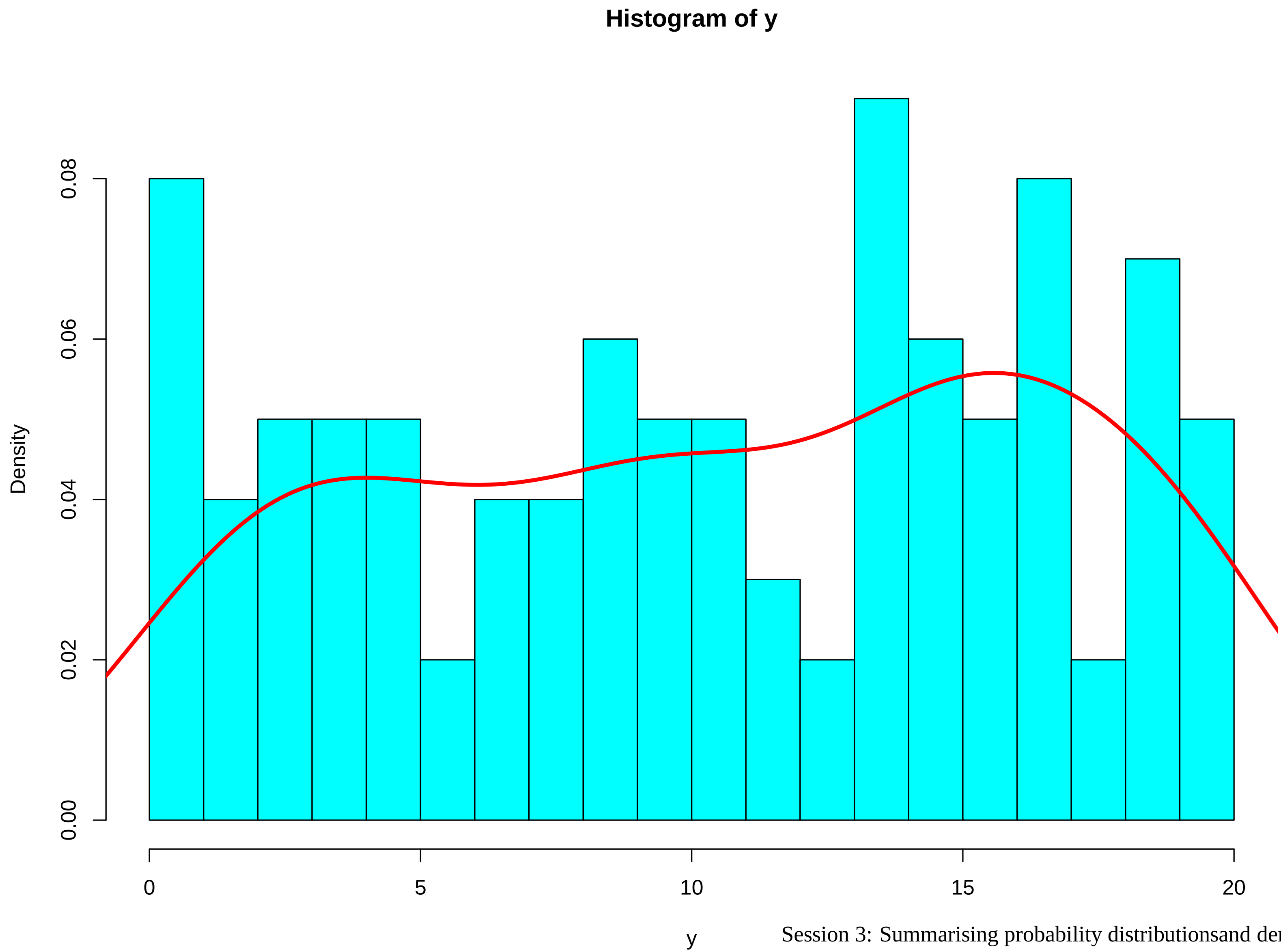
- Binary distribution – Bernoulli?
- $p = 0.45$

Example 2: Discrete RV

1. Sample = 100 values
2. (The first 13 values)
   $y = 11 \;\; 9 \;\; 6 \;\; 13 \;\; 15 \;\; 18 \;\; 17 \;\; 11 \;\; 8 \;\; 10 \;\; 2 \;\; 0 \;\; 12$

y	Freq	y	Freq	y	Freq	y	Freq
0	4	6	5	11	7	16	5
1	7	7	6	12	5	17	7
2	4	8	9	13	4	18	4
3	1	9	5	14	3	19	0
4	5	10	5	15	4	20	3
5	7						

Histogram of y



Guessing the PD

- Discrete RV from 0 to 20
- Could be a uniform discrete PD

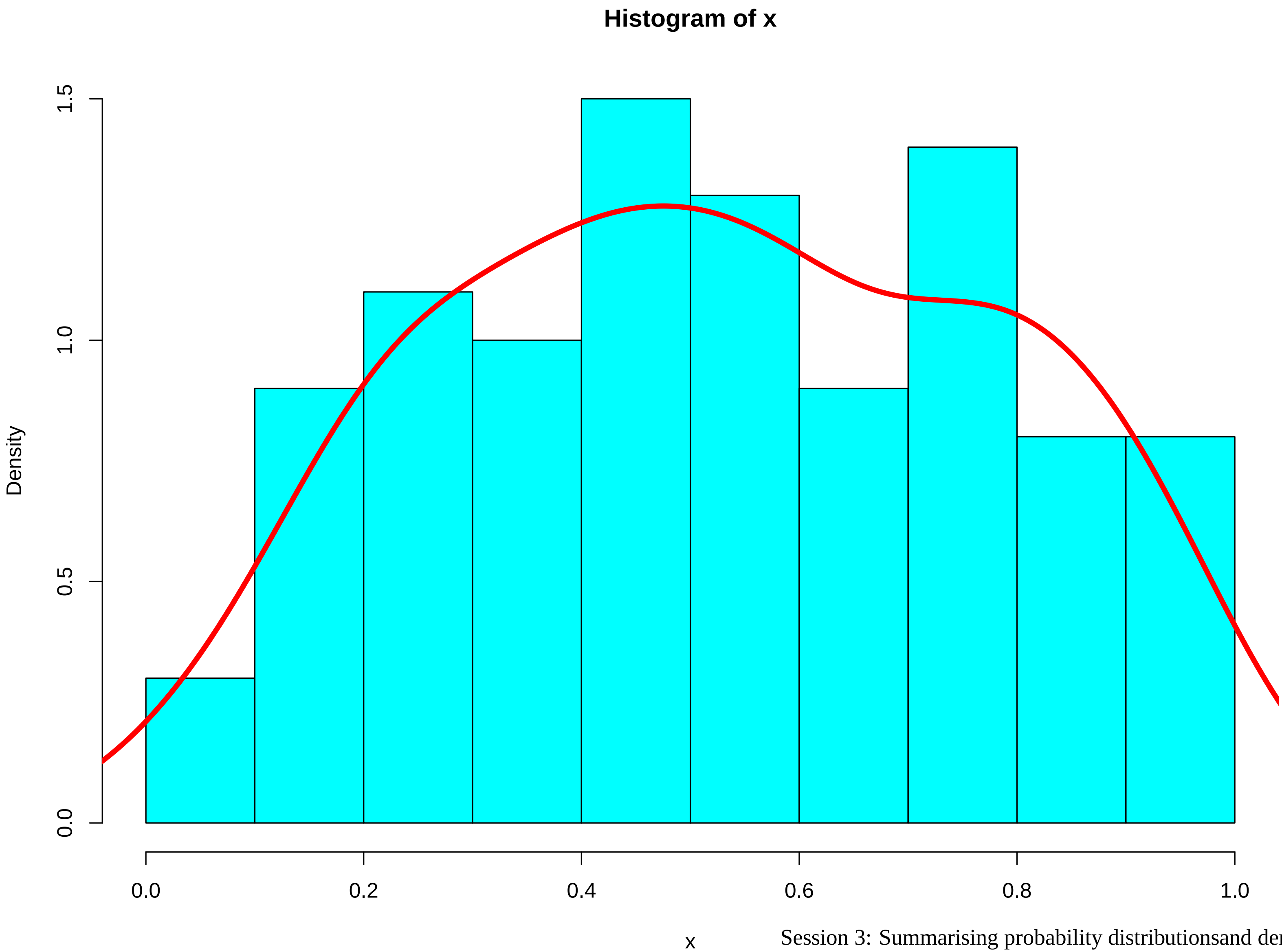
Example 4: Continuous RV

1. Sample = 100 values
2. $k =$

0.993941516	0.612212929	0.201375686
0.240819249	0.142533204	0.430064859
0.697499793	0.030674237	0.944907661
$\dots$	$\dots$	$\dots$

3. Frequency table: each element has a frequency of one.

Histogram of k



Guessing the PD

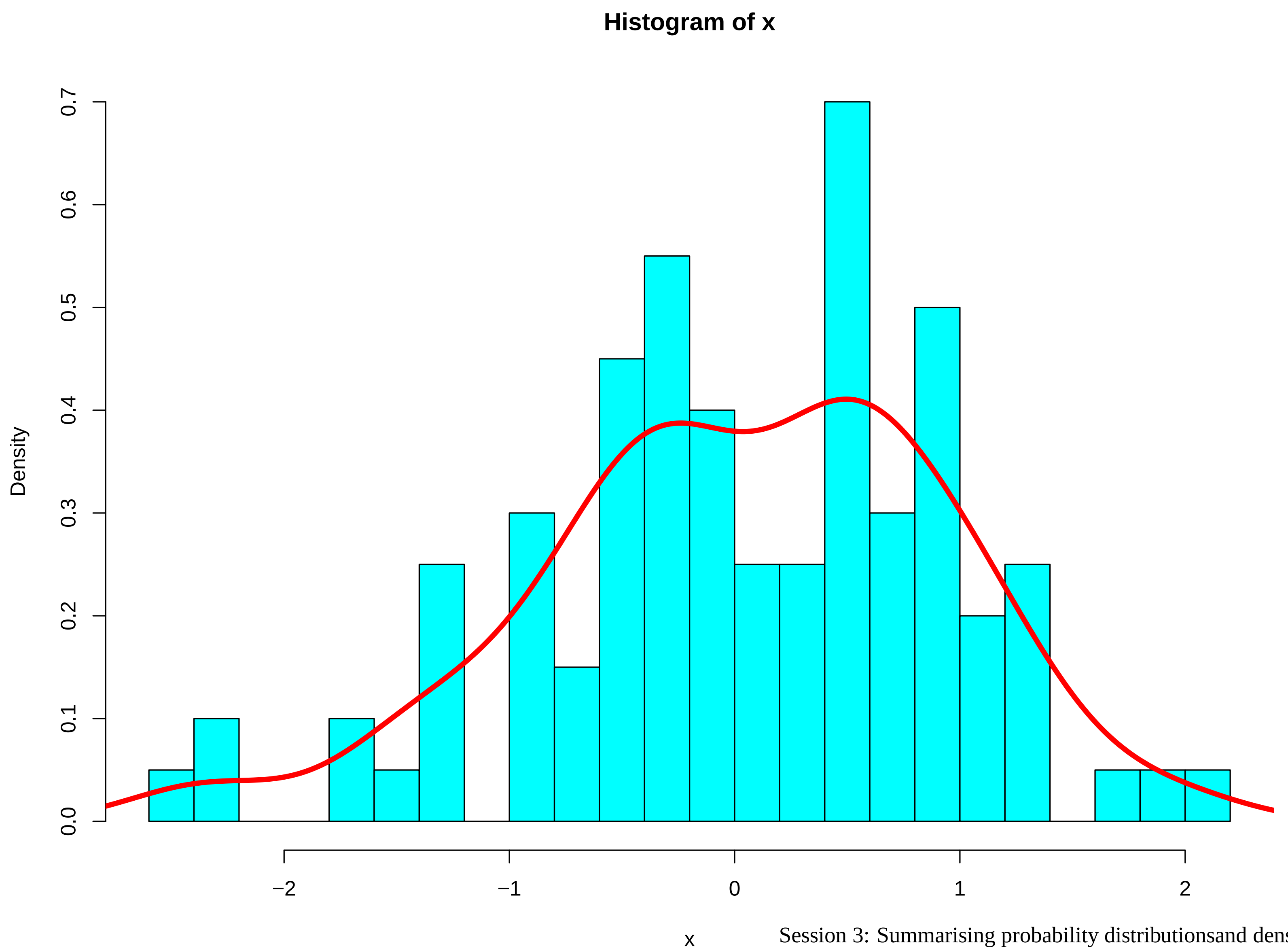
- RVs are continuous
- Could be a uniform distribution? (No negative values, data appears range bound.)

Example 5: Continuous RV

1. Sample = 100 values
2. $n =$

0.724773431	0.500281917	0.903696952
0.612282267	-0.185570961	0.247409823
-0.820567376	-1.413818678	1.368954272
...	...	...

Histogram of n



Guessing the PD

- Continuous RV
- Could be normally distributed?

Numerical tools

Numerical description for data

Statistical measures!

- Measure of location: Mean, mode, median
- Measure of dispersion: Variance, range, quartiles

Measures of location

- Mean: An expected value on a random draw from the dataset.
- Mode: The value that occurs with the maximum frequency.
  Easily interpreted for discrete variables.
  The mode for the continuous RV datasets is interpreted in terms of the “range/set” of values that are most often observed.
- Median: The value of the RV at which 50% of the dataset is observed.

Examples: Mean, $\bar{x}$

- Find the mean of the data: 5, 1, 6, 2, 4:

$$\bar{x} = \frac{\sum x}{n} = \frac{18}{5} = 3.6$$

Examples: Median

- Find the median of: 1, 7, 3, 1, 4, 5, 3.
- First step is to order the data: 1, 1, 3, 3, 4, 5, 7.
- The median is 3, the midway point, for an odd number of data.
- When the data has an even number of points, the median is calculated as the midpoint between the two choices.
- For a dataset: 9, 5, 7, 3, 1, 8, 4, 6, ordered as 1, 3, 4, 5, 6, 7, 8, 9, the median is 5.5.

Pros and cons of location measures

- Typically, all three measures tend to cluster together - the differences are not very large.
- **However** the mean is most sensitive to the presence of **outliers**.
  (For example, a day on which a trader places a buy limit order for 100 million shares of Reliance instead of a a thousand shares.)
- The median is less sensitive to the mean. It is not influenced by the value of the observations, just their number.
  Thus, it can be a more robust measure of location than the mean.

Measures of dispersion

- Range: The difference between the highest and the lowest value of the RV in the dataset.
  Example: In data, 3, 7, 2, 1, 8 the range = 8 - 1 = 7.
- Variance: The value of the RVs as differences from the average value, squared and summed up. It is denoted by $\sigma(x)^2$.

$$\sigma(x) = \frac{\sum x_i - \bar{x}}{(n-1)}$$

Example: $\bar{x} = 4.2, \sigma(x)^2 =$
$(-1.2^2 + 2.8^2 + -2.2^2 + -3.2^2 + 3.8^2)/4 = 9.7.$

Calculating the data dispersion using $\bar{x}, \sigma^2$

- Question: what is the range of values of the RV between which we can find 95% of the data?
- Answer:
  1. Upper range value = $\bar{x} + 1.96 * \sigma$
  2. Lower range value = $\bar{x} - 1.96 * \sigma$

Empirical rules

- $\bar{x} \pm \sigma = 85\%$ of the dataset
  The percentage will be larger for more skewed distributions. The percentage will be closer to 70% for distributions that are more symmetric.
- $\bar{x} \pm 2\sigma = 97\%$ of the dataset
- $\bar{x} \pm 3\sigma = 99\%$ of the dataset

Measures of dispersion: Percentiles, Quartiles

- Percentiles: Denoted as p^{th} percentile. The value of RV, x, such that p% of the dataset falls below the value x, and $(100 - p)$% is above.
- Quartiles: A set of three specific percentiles at the $25^{th}, 50^{th}, 75^{th}$ percentiles. They are the lower, median and upper quartile values.
  The median is the 2^{nd} quartile and the 50^{th} percentile.
- Inter-quartile range (IQR): The distance between the lower and the upper quartile values.

Problems to be solved

Q1: Transforming normally distributed variables

What is the impact of the function:

$$g(x) = \frac{x - \mu}{\sigma}$$

upon the behaviour of the probability density of a RV which comes from a normal distribution with mean μ and variance σ^2?

Q2: Constructing measures of location

The EPS for 20 companies collected from the 1985 Fortune 500 companies are (in USD):

0.75	4.65	3.54	1.85	2.92	5.23	3.75	2.80	3.27	0.72
6.58	1.35	6.28	9.11	1.72	2.75	1.96	4.40	2.01	1.12

1. Create a relative frequency distribution for this data
2. Calculate the mean, median and mode. Locate them on the frequency distribution.
3. Do these measures of location appear to locate the center of the data distribution?

Q3: Calculating measures of location and dispersion

Calculate the variance and standard deviations for the following datasets:

1. $n = 10, \sum x^2 = 331, \sum x = 50$
2. $n = 25, \sum x^2 = 163,456, \sum x = 2,000$
3. $n = 5, \sum x^2 = 26.46, \sum x = 11.5$

References

- Chapter 2, SHELDON ROSS. *Introduction to Probability Models*. Harcourt India Pvt. Ltd., 2001, 7th edition