Hypothesis Testing

Chi-Squared Test & Shapiro Wilk Test

Math & Stats Tutorial
Day 5

Prabesh Dhakal
13 May 2019

PLAN FOR TODAY

1. Review: Quartiles, IQR, Outliers, Distribution
2. Density Distribution
3. p-Value
4. Shapiro-Wilk Test
5. Chi-square Test
6. R-Studio Session

REVIEW: QUARTILES

Quartiles are the values that divide a list of numbers into 4 equal parts *(quarters)*

Steps Involved

1. Put the list of number in order
2. Cut the list into 4 equal parts
3. If the values lie between two numbers from the observation, take an average of the two values

REVIEW: QUARTILES EXAMPLE (1)

Quartiles **for odd number of observations**

Observations: 5, 7, 4, 4, 6, 2, 8

1. Put them in order: 2, 4, 4, 5, 6, 7, 8
2. Cut the list into quarters:

2, 4, 4, 5, 6, 7, 8

Q1 Lower Quartile

Q2 = Median

Q3 Upper Quartile

REVIEW: QUARTILES EXAMPLE (21)

Quartiles **for even number of observations**

Observations: 4, 6, 3, 1, 80, 6, 3, 7, 5, 8

1. Put them in order: 1, 3, 3, 4, 5, 6, 6, 7, 8, 80
2. Cut the list into quarters:

1, 3, 3, 4, 5, | 6, 6, 7, 8, 80

Q1 Lower Quartile

Q2 = Median

Q3 Upper Quartile

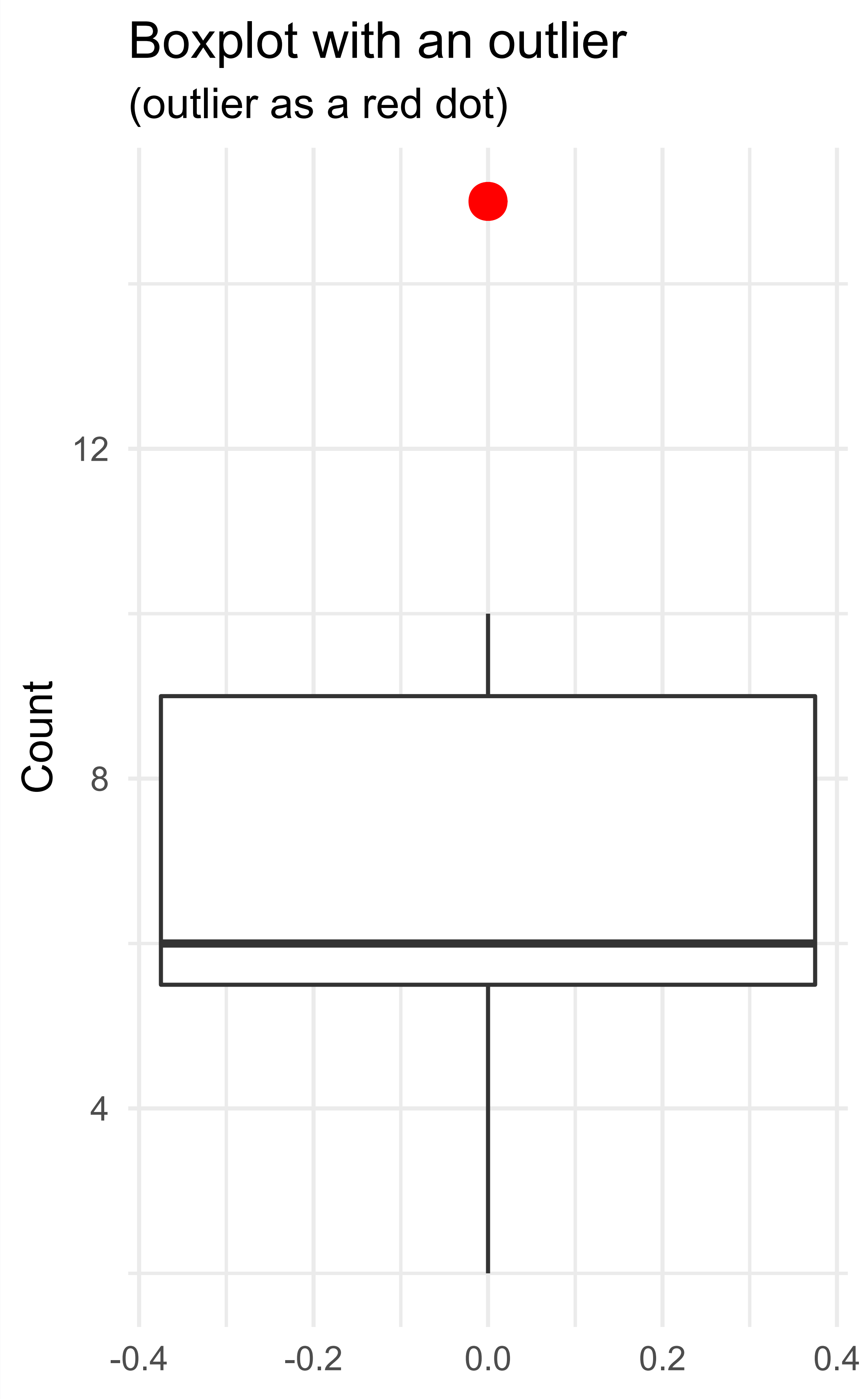
In this case, median is between 5 and 6: $\textbf{median} = (5+6)/2 = 5.5$

REVIEW: OUTLIERS

Outliers in a data set are observations that are:

> ***Either***: Lower than ($Q_1 - 1.5 * IQR$)
>
> ***Or***: Higher than ($Q_3 + 1.5 * IQR$)

The whiskers of the box plot end at the values ($Q_1 - 1.5 * IQR$) and ($Q_3 + 1.5 * IQR$).



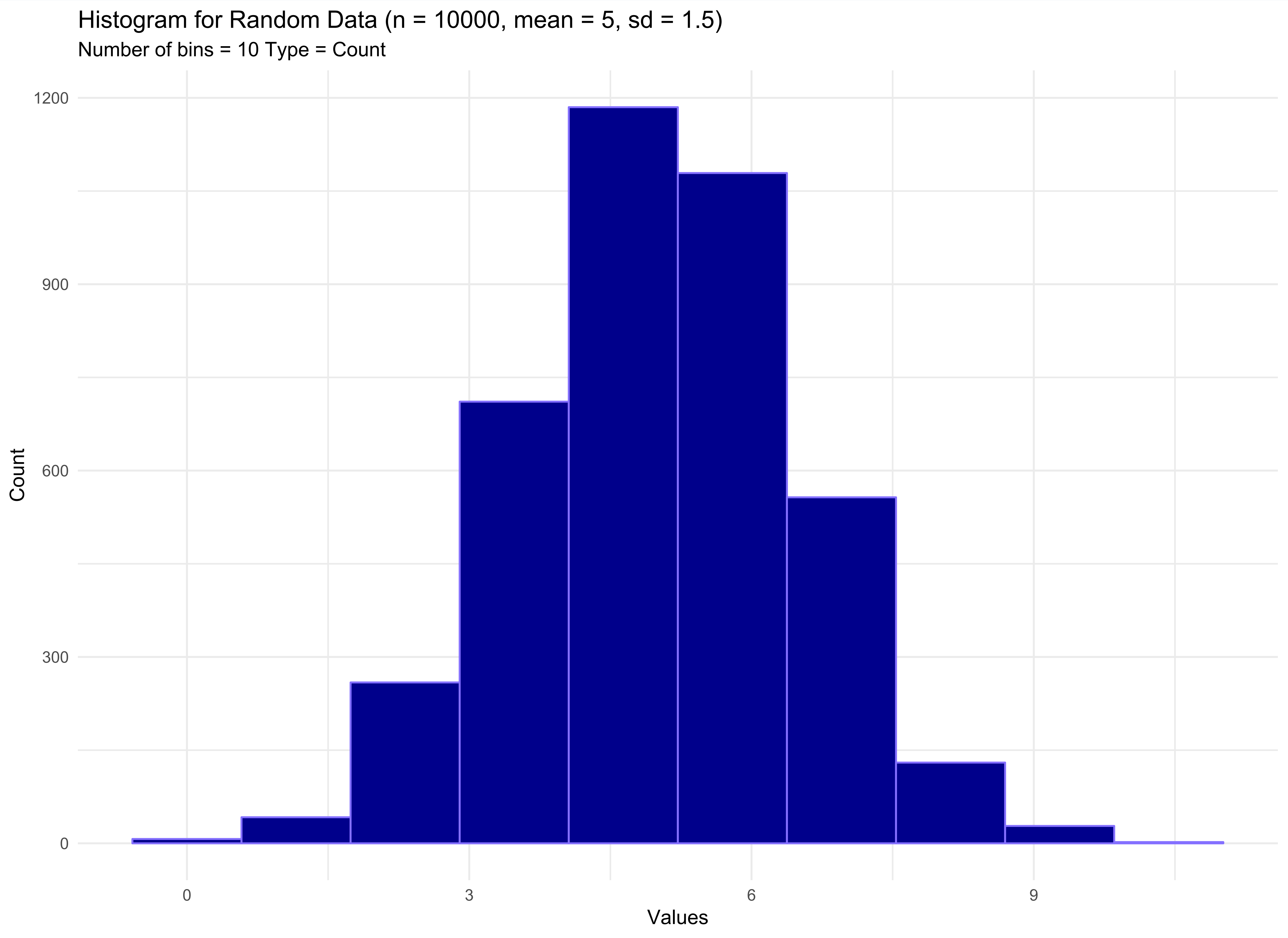
REVIEW: DISTRIBUTIONS (1)

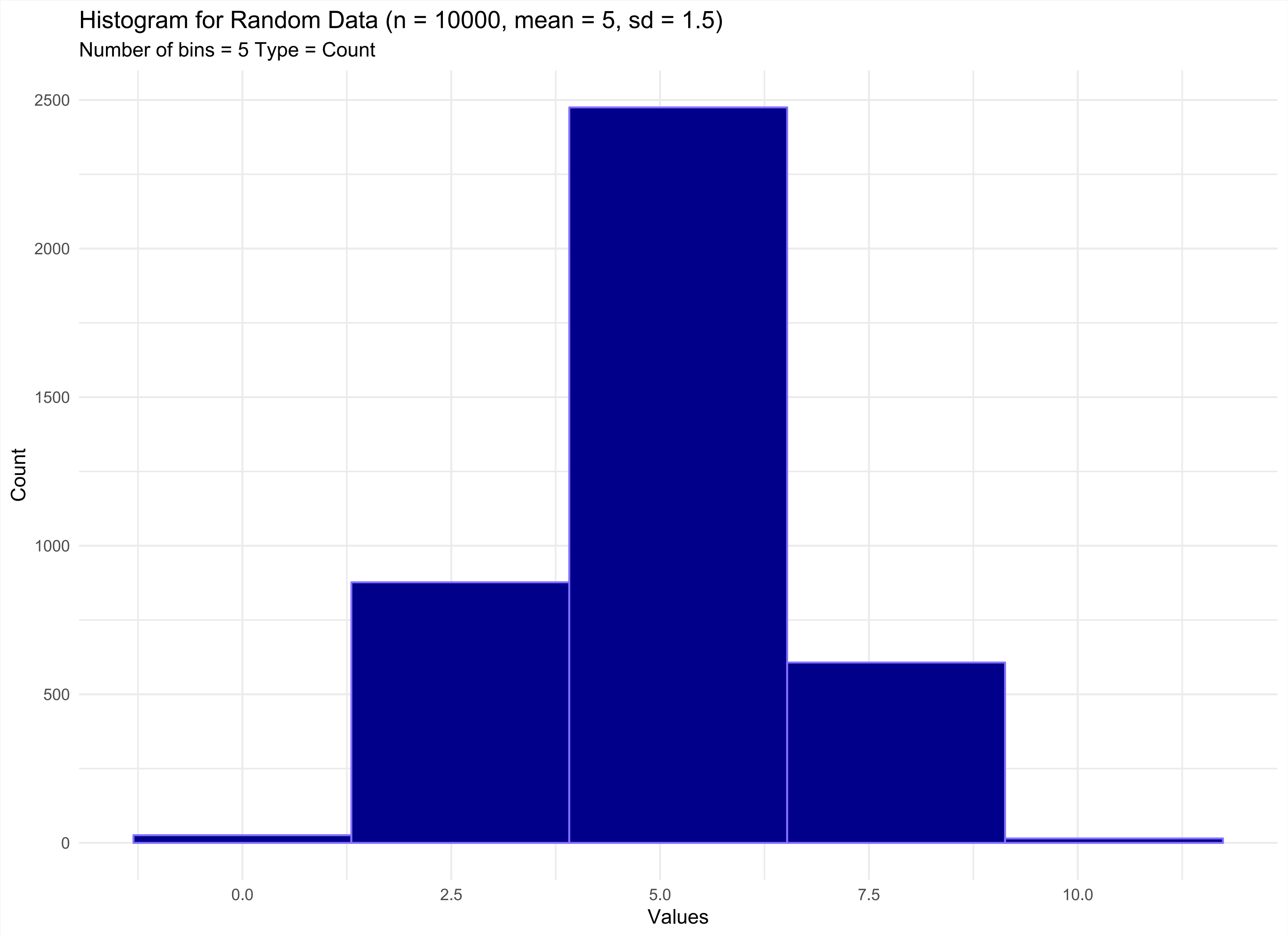
There are many ways a data set might be distributed.

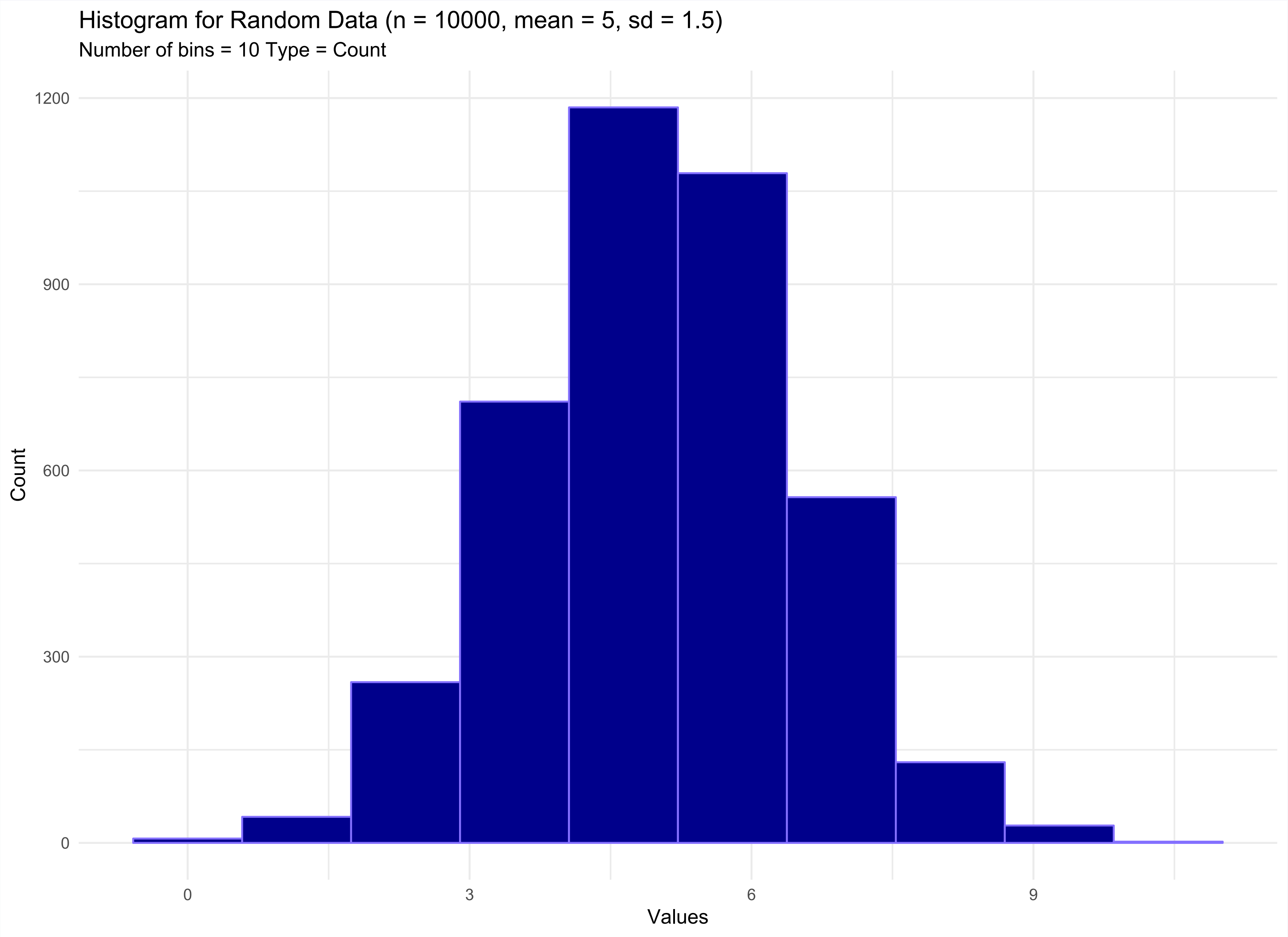
However, the nature of distribution in a given industry/topic tends to be similar.

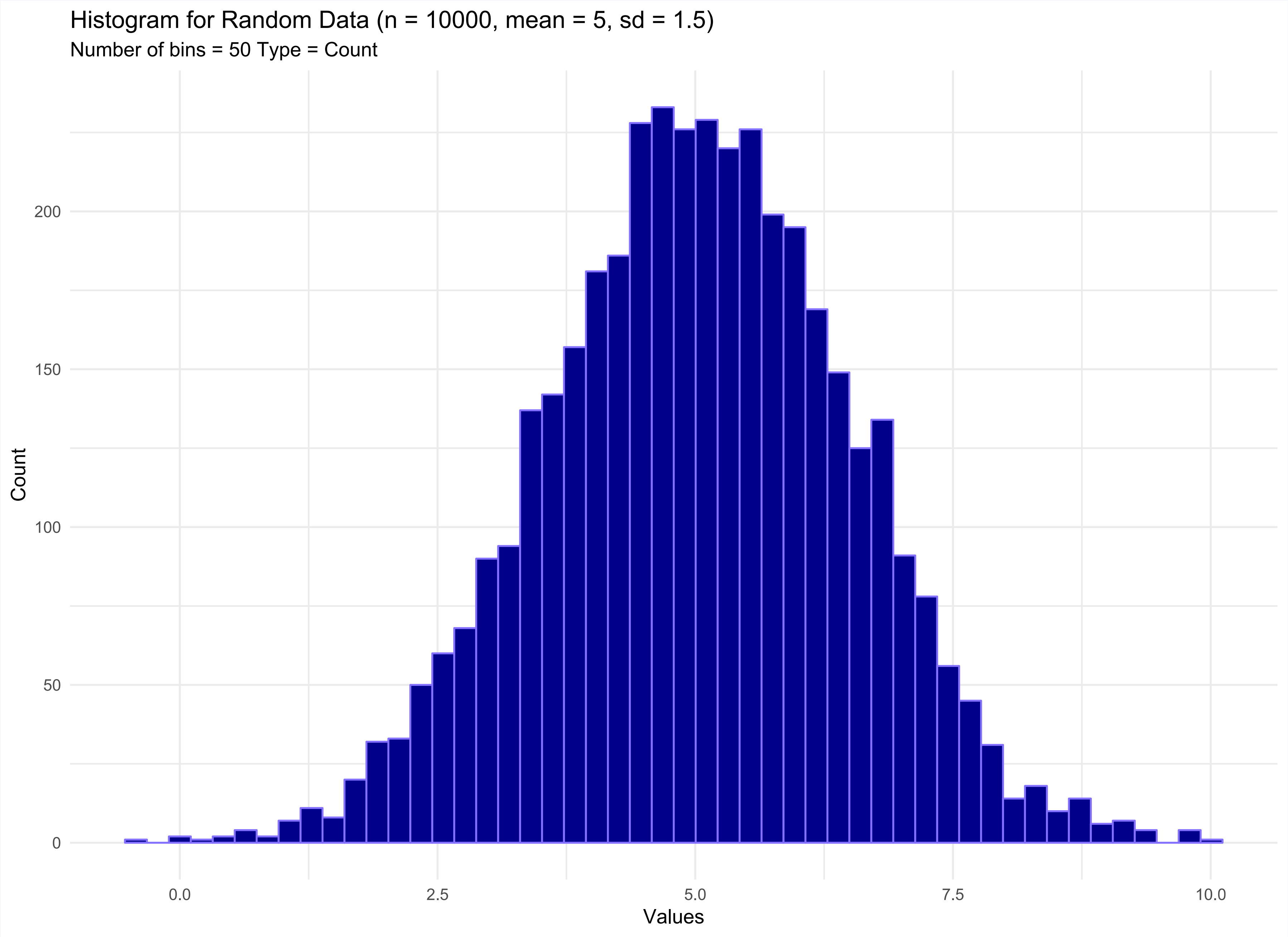
Distribution of data can be represented with scatter plots, bar plots, histograms, box plots, density plots (represented by curves), etc.

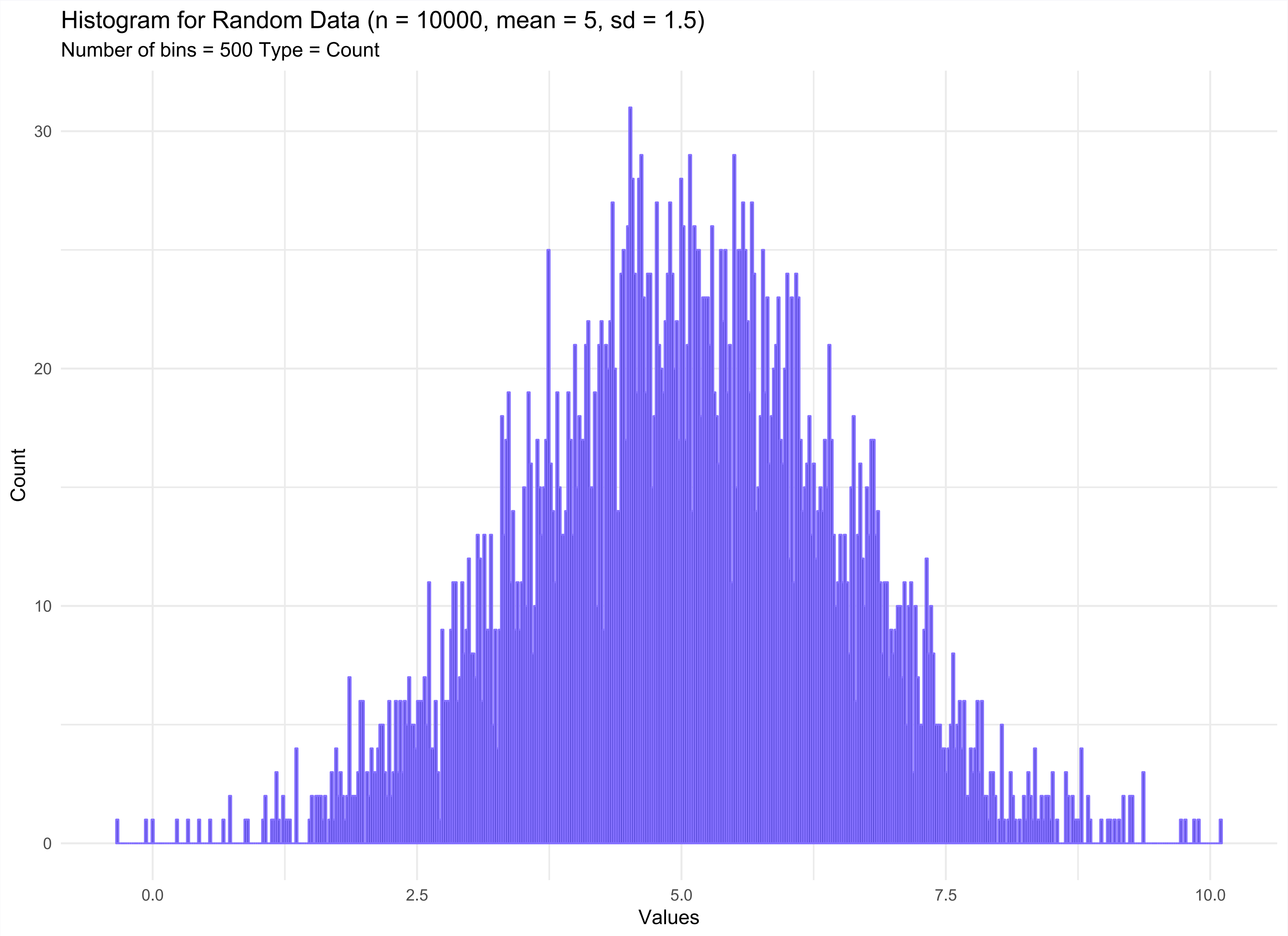
REVIEW: DISTRIBUTIONS (2)



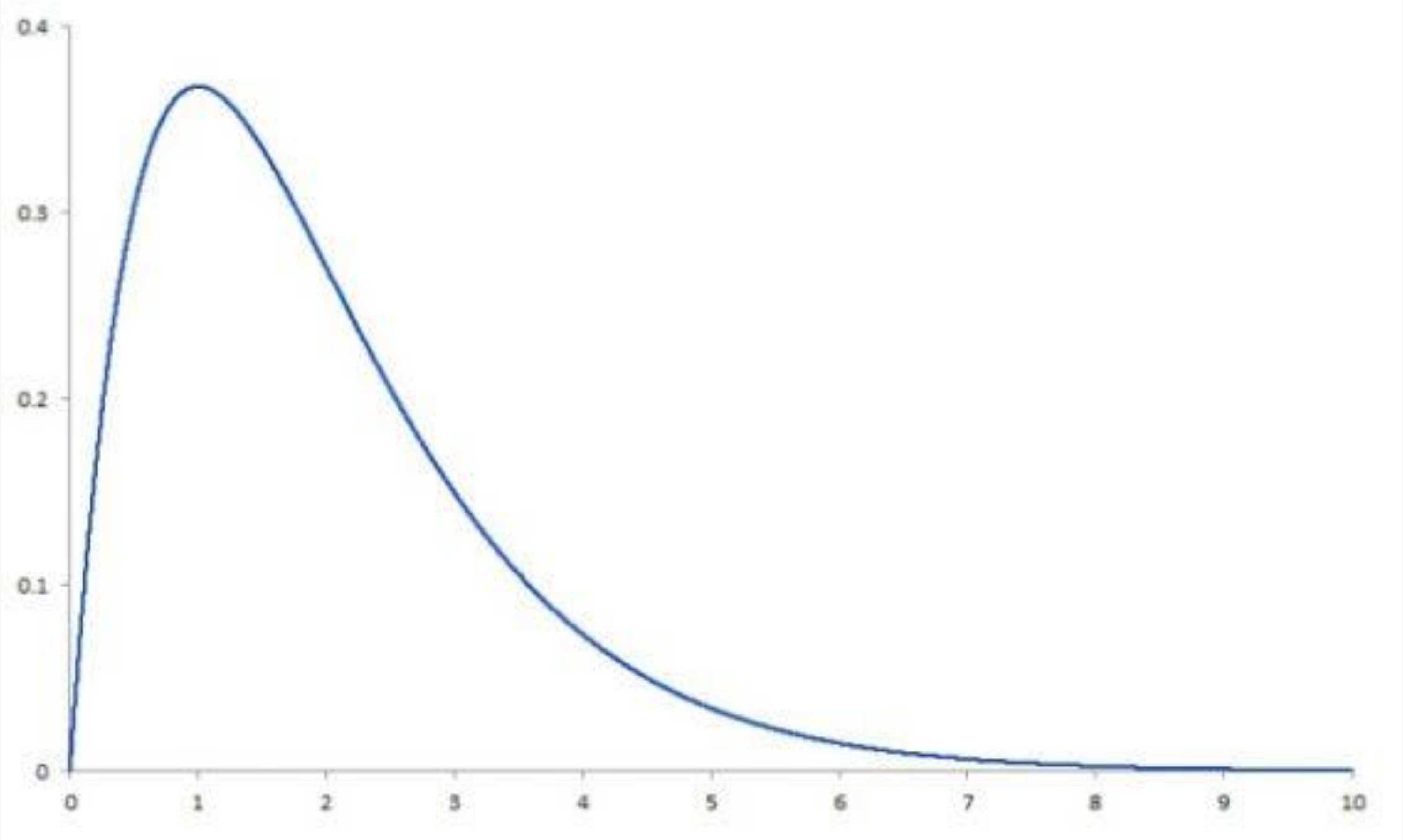

Histogram for Random Data (n = 10000, mean = 5, sd = 1.5)
Number of bins = 5 Type = Count
Count
Values
0
500
1000
1500
2000
2500
0.0
2.5
5.0
7.5
10.0


Histogram for Random Data (n = 10000, mean = 5, sd = 1.5)
Number of bins = 10 Type = Count
Count
Values
0
300
600
900
1200
0
3
6
9


Histogram for Random Data (n = 10000, mean = 5, sd = 1.5)
Number of bins = 50 Type = Count
Count
Values
0
50
100
150
200
0.0
2.5
5.0
7.5
10.0

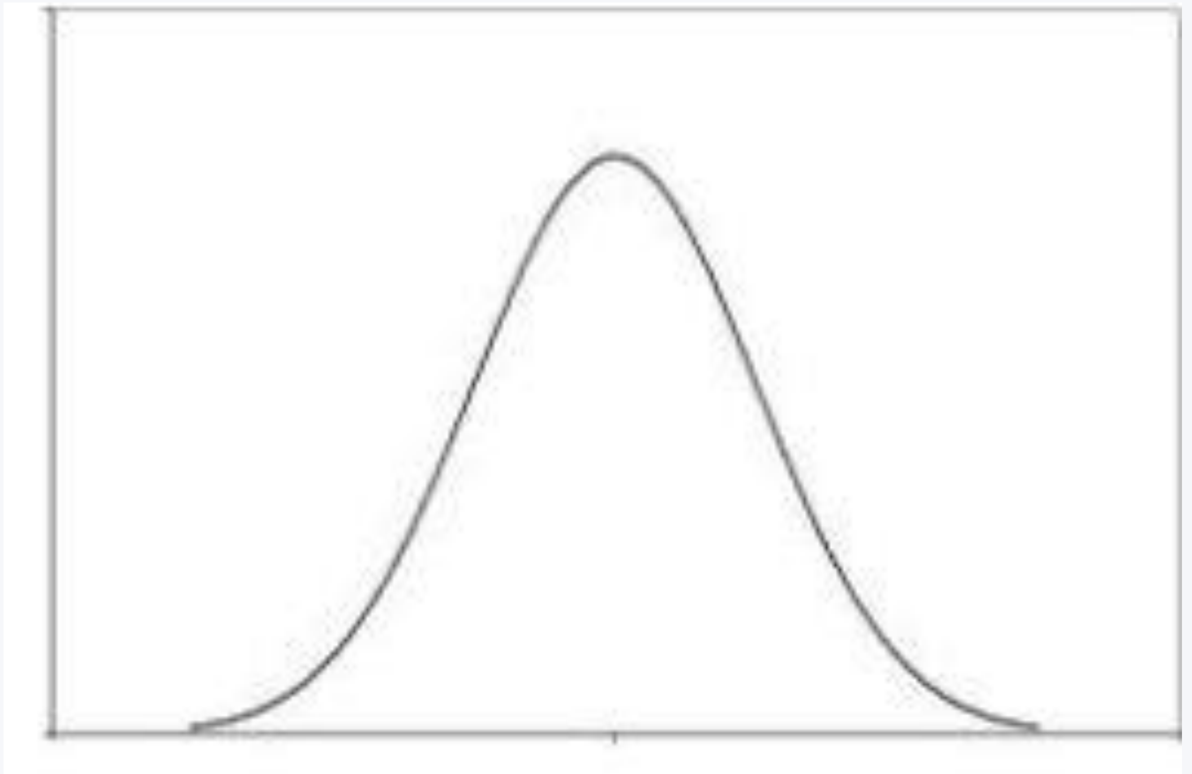

Histogram for Random Data (n = 10000, mean = 5, sd = 1.5)
Number of bins = 500 Type = Count
Count
Values
0
10
20
30
0.0
2.5
5.0
7.5
10.0

REVIEW: SKEW OF DISTRIBUTIONS



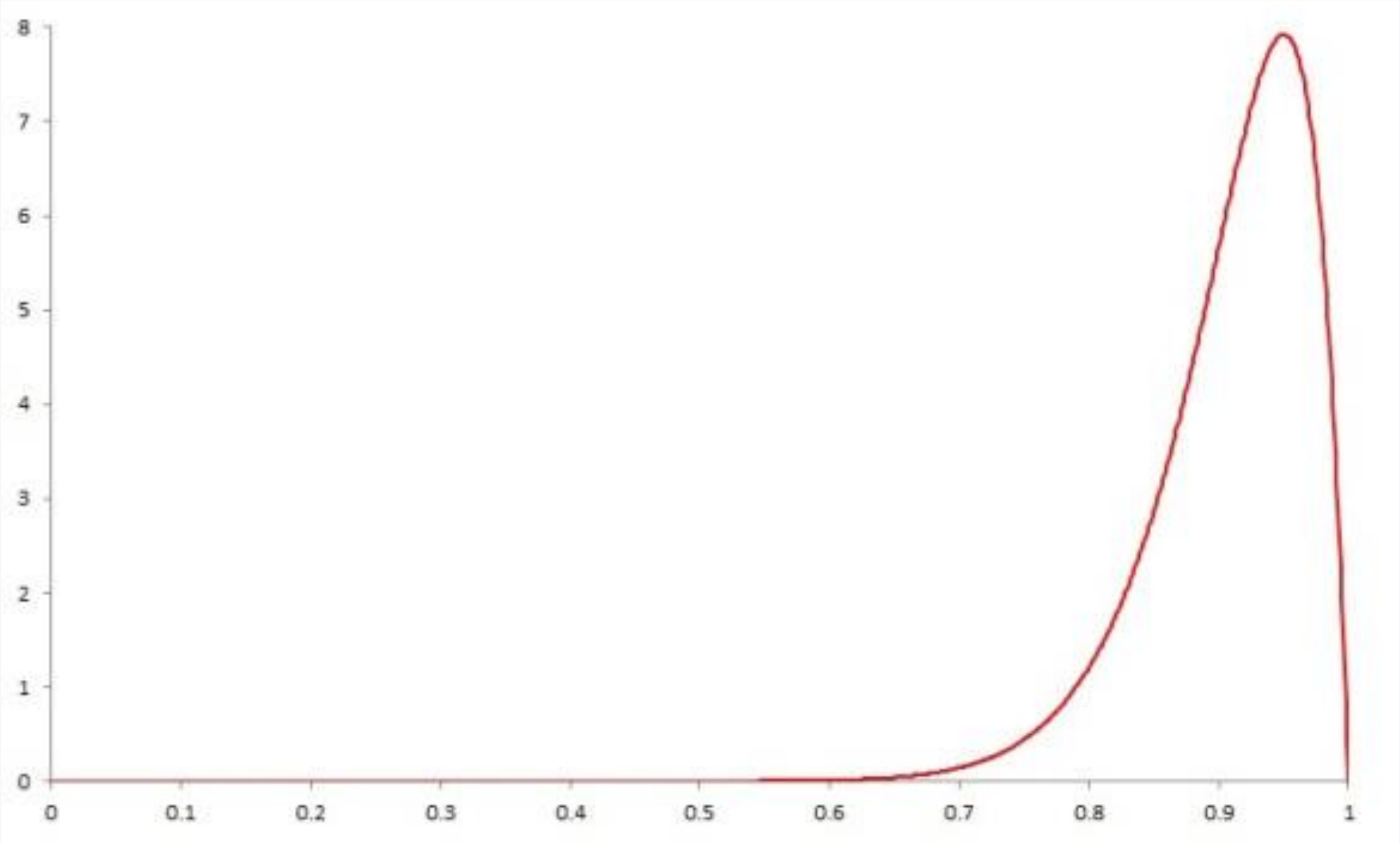
Right Skew

mode < median < mean



Symmetric

mean = median = mode



Left Skew

mean < median < mode

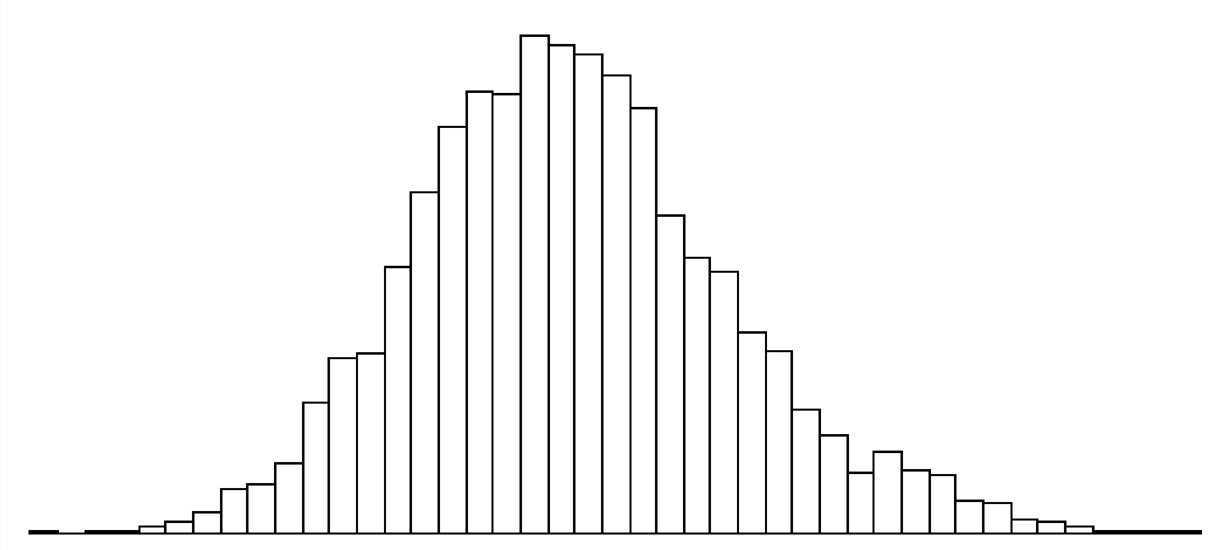
CLASS EXERCISE - 1

TASK:

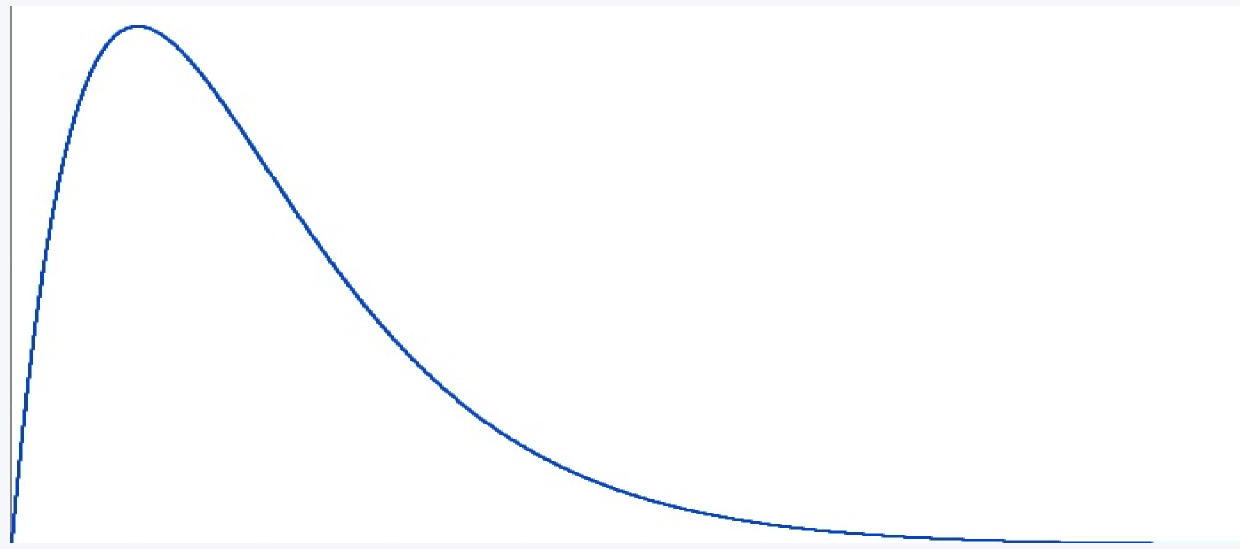
Following charts were created based on different data sets.

Given the charts, how would you describe the skew of the distribution?

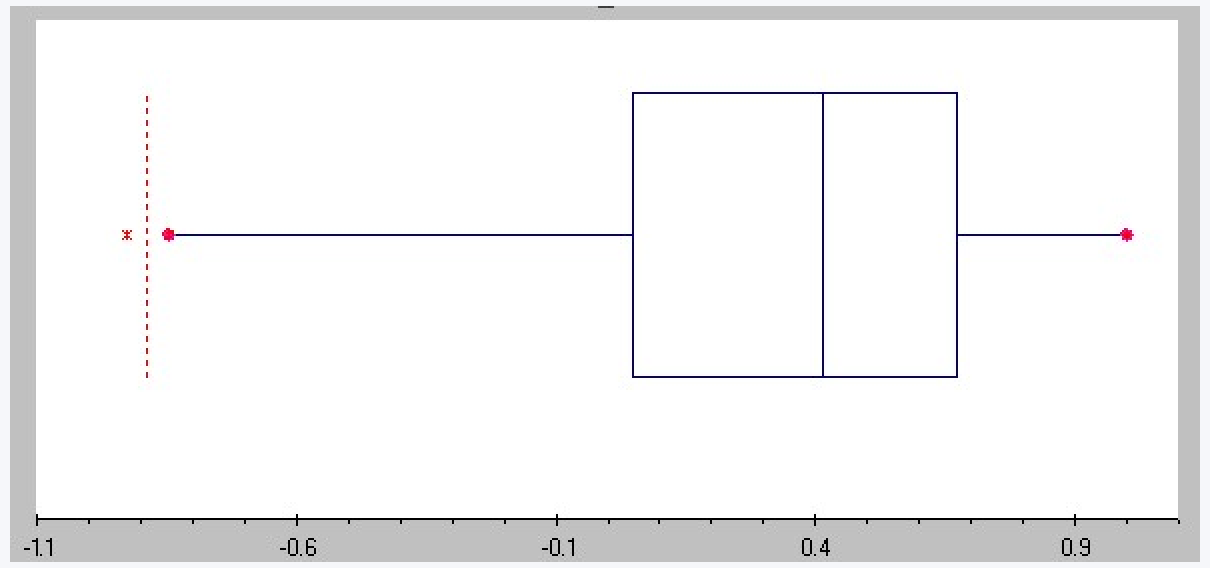
a.



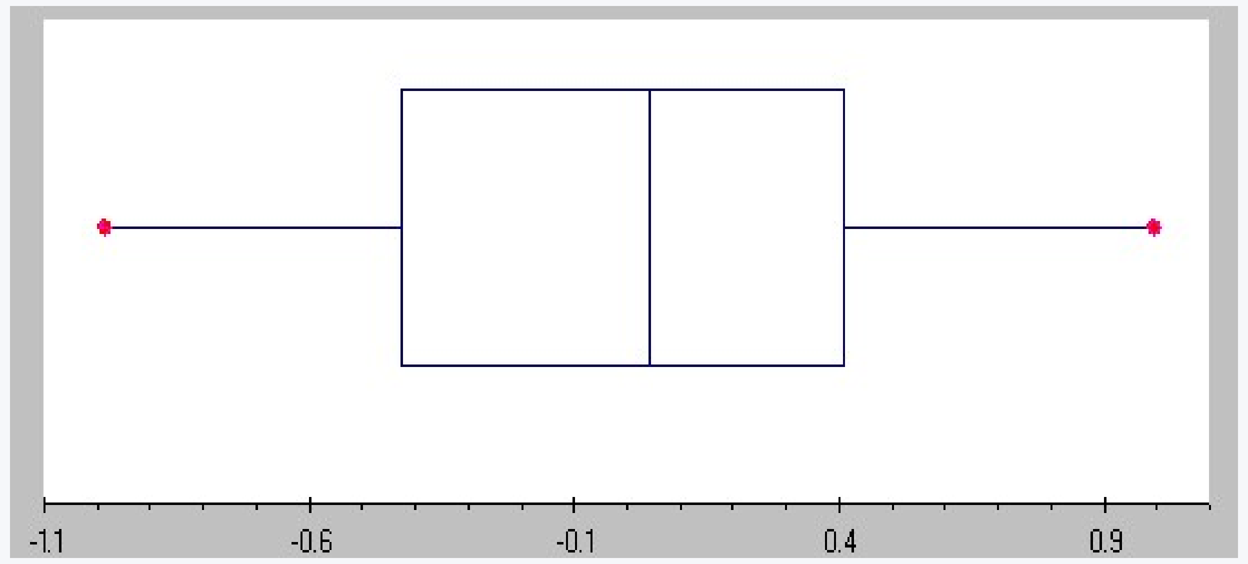
b.



c.



d.



SIDE REMARK

FREQUENCY VS RELATIVE FREQUENCY

Frequency is the raw count of your data.

- the number of times an event occurred

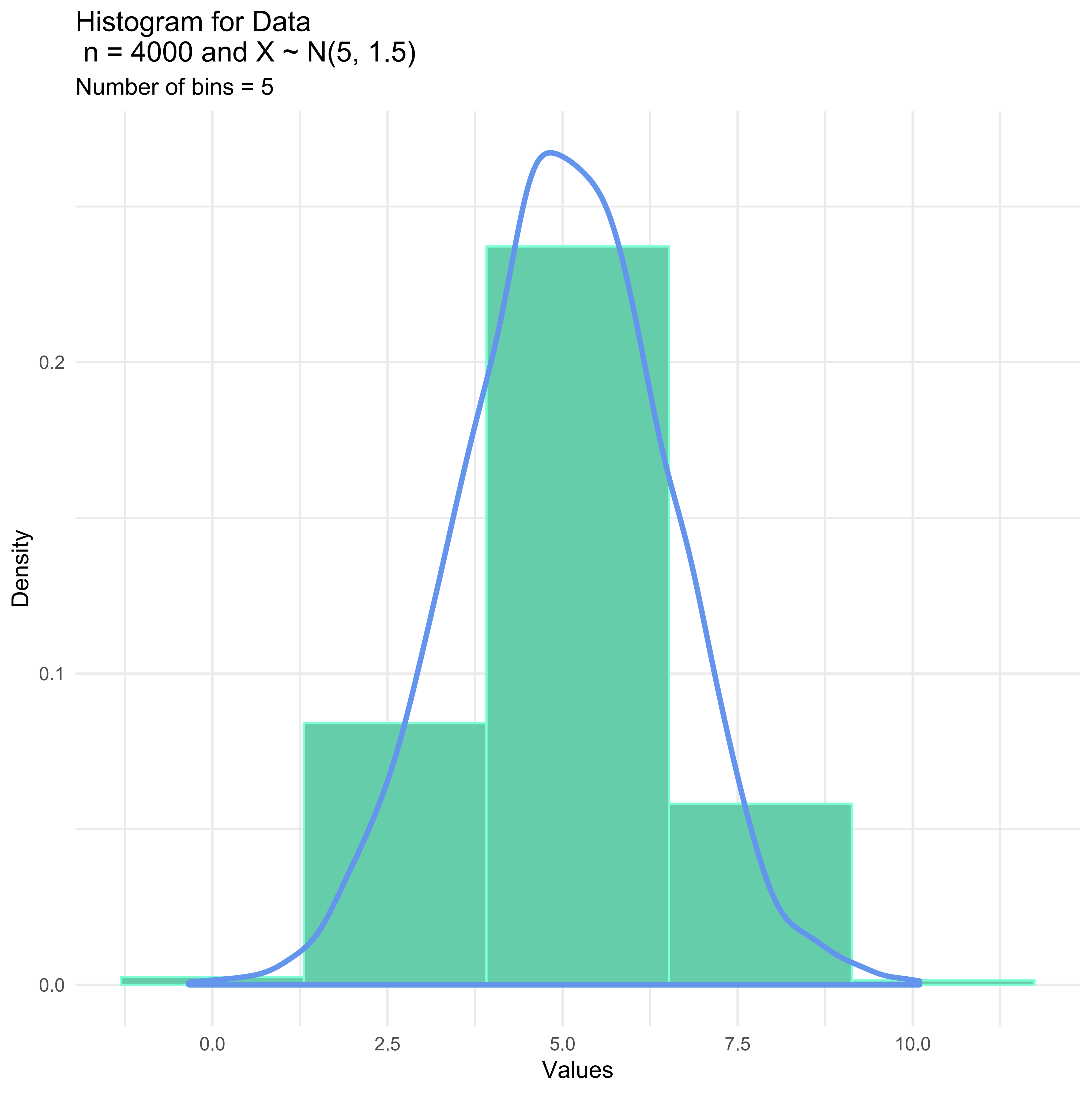
Relative Frequency is the absolute frequency normalized by the total number of events.

- how often something happened divided by the total outcomes

Hobby	Freq.	Rel. Freq.
Music	45	0.31
Dance	32	0.22
Running	12	0.08
Skating	8	0.06
Reading	39	0.27
Salsa	7	0.05
Total:	**143**	**1**

Rel. Frequency Histogram: https://www.statisticshowto.datasciencecentral.com/relative-frequency-histogram-2/

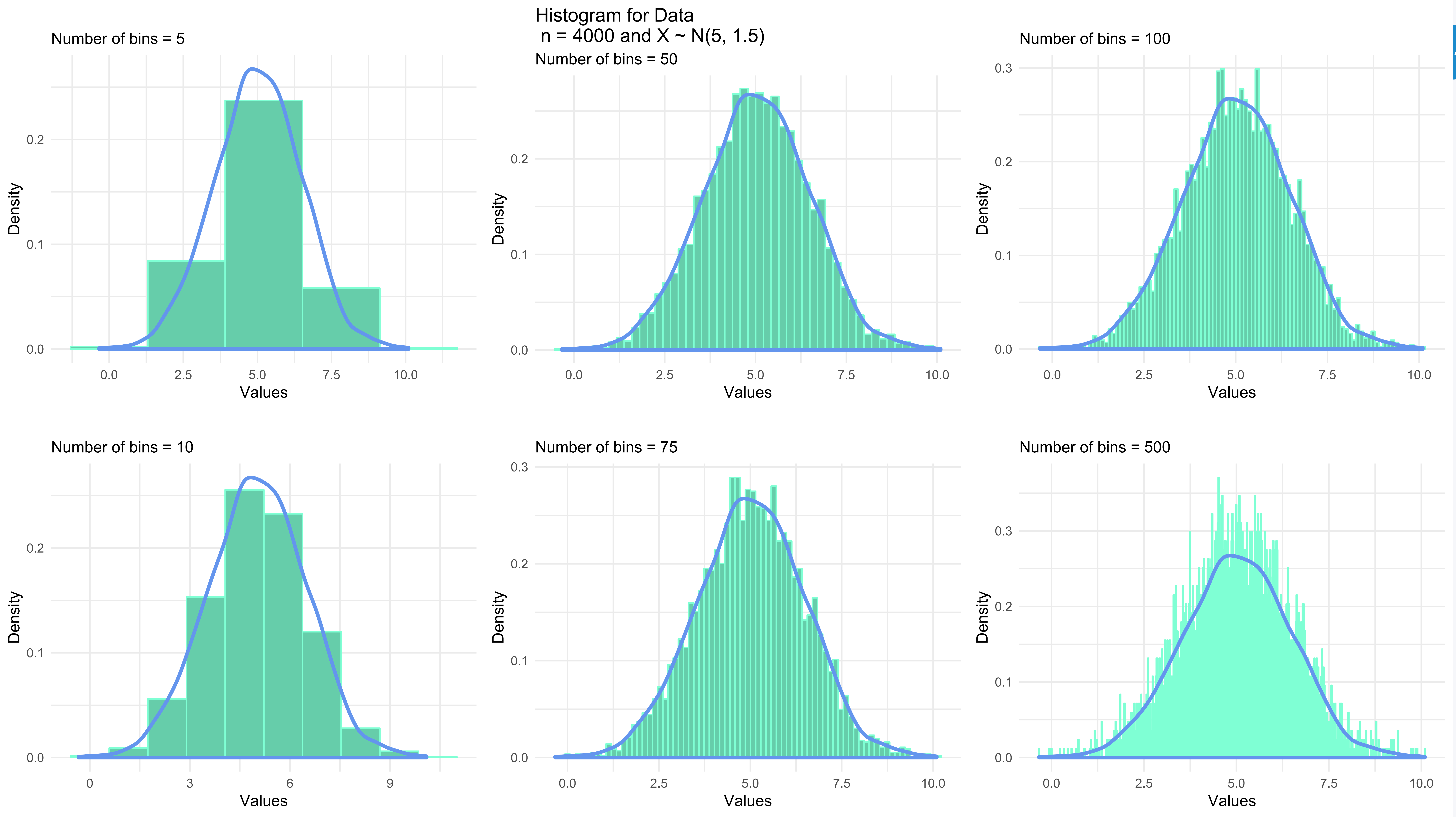
RELATIVE FREQUENCY (DENSITY) HISTOGRAMS



Blue line represents the **density** of the data

The histogram and blue lines do not match that well here.

What happens if the number of bins is increased?


Histogram for Data
n = 4000 and X ~ N(5, 1.5)
Number of bins = 5
Number of bins = 50
Number of bins = 100
Number of bins = 10
Number of bins = 75
Number of bins = 500
Density
Values
0.0
0.1
0.2
0.3
0.0
2.5
5.0
7.5
10.0
0
3
6
9

Hypothesis Testing

Process of Hypothesis Testing

Significance Level (α)

$\mathbf{p-value}$

PROCESS FOR HYPOTHESIS TESTING

Step 1	Specify H_0, H_1, and an acceptable level of $\boldsymbol{\alpha}$
Step 2	Define a sample-based test statistic and the rejection region for the specified H_0
Step 3	Collect the sample data and calculate the test statistic
Step 4	Decide to either reject or fail to reject H_0
Step 5	Interpret the results/make recommendation for action

SIGNIFICANCE LEVEL (α)

α denotes the probability of making a Type I Error (rejecting true H_0)

Significance level of a hypothesis test is the max. acceptable probability of rejecting a true null hypothesis.

The significance level (α) should be low so that the risk of incorrectly rejecting H_0 is minimized. (typically, $\alpha = 0.10\ or\ 0.05\ or\ 0.01$)

WHY CARE ABOUT TYPE I ERROR?

Type I Error (rejecting true H_0) 'generally' leads to more serious consequence

Identify the H_0 and H_1 of the following cases:

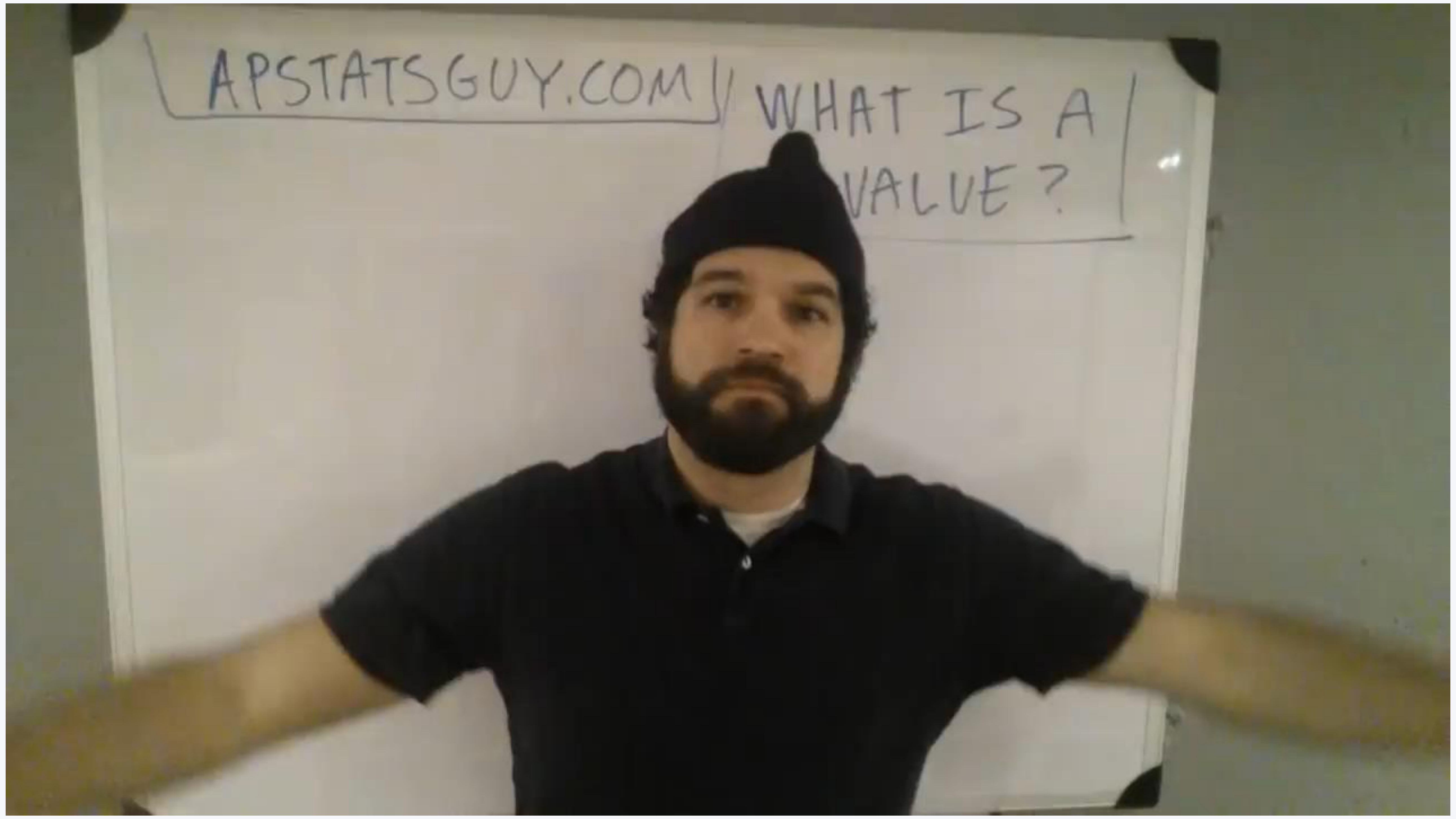
1. Bayer wants to test the toxicity (side effects) of a drug that it is testing.
2. Continental emits some pollutant during its manufacturing process. The German environmental agency, who is performing its periodic test, requires it to have a **mean** emission less than a threshold T.

What α value would you choose for these examples?

p-value

probability of you making the observations **if H_0 were true**

$$p - value = P(data \mid H_0 \; is \; true)$$



Video source: https://www.youtube.com/watch?v=-MKT3yLDkqk

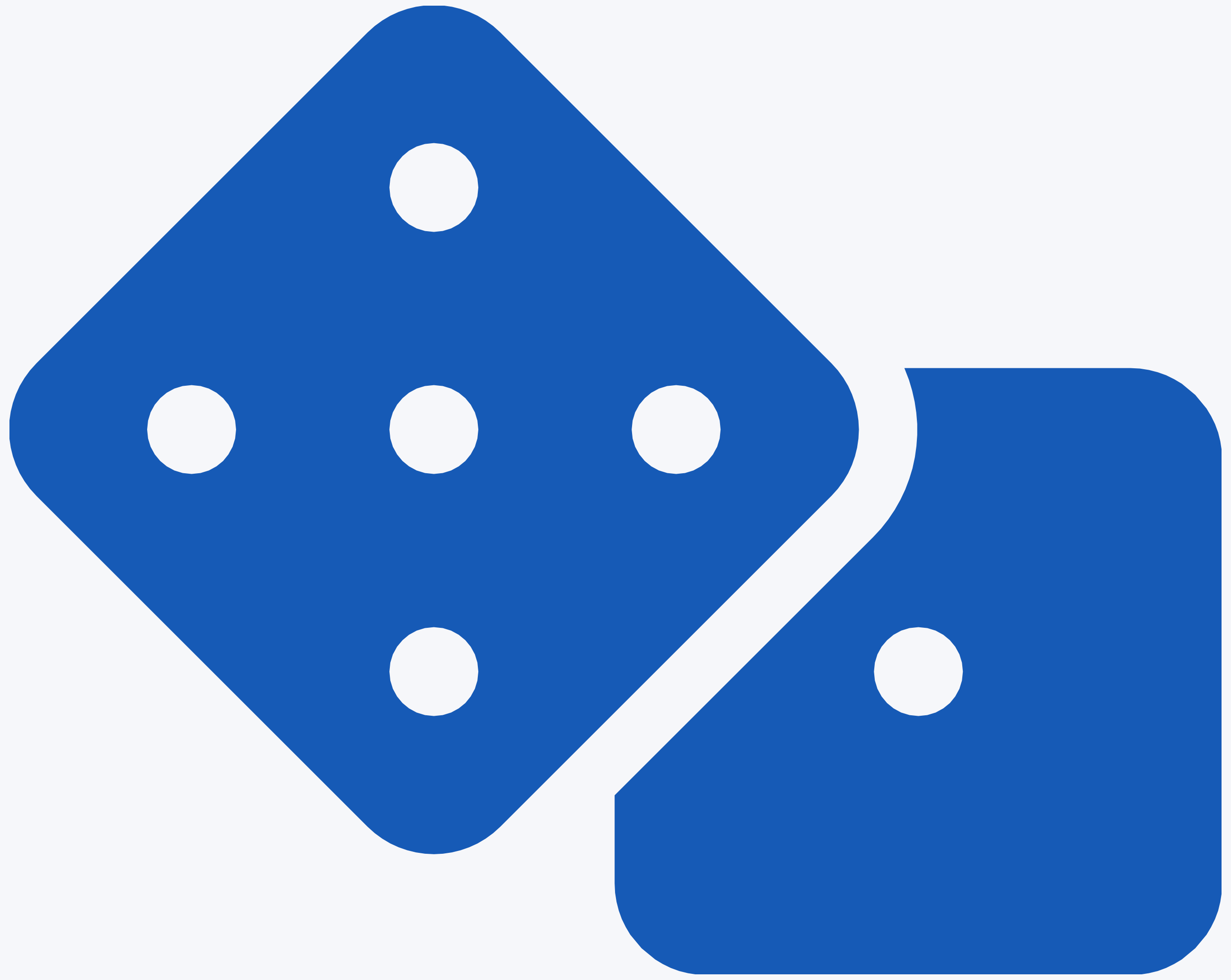
COMPARING (α) AND $p - value$

When $p - value \leq \alpha$, we reject H_0

- The result is statistically significant
- We are reasonably sure that there is something besides chance that gave us an observed sample

When $p - value > \alpha$, we fail to reject the H_0

- The result is not statistically significant.
- We are reasonably sure that our observed data can be observed by chance alone



Exercises

χ^2 $Test$ for a Multinomial Population (Example)

Suppose we had a genetic experiment where we hypothesize the 9:3:3:1 ratio of characteristics A, B, C and D. A sample of 160 offspring are observed and the actual frequencies are 82, 35, 29, and 14, respectively.

The hypotheses to be tested are:

$$H_0\!: p_1 = 9/16,\; p_2 = 3/16, p_3 = 3/16, p_1 = 1/16$$

$$H_1\!: the\ proportions\ differ\ from\ those\ specified$$

We reject H_0 if χ^2 statistic that we calculate exceeds the critical value from χ^2 distribution (from the table) at a given α and degree of freedom (k).

χ^2 TEST (Chi-squared Test) of Independence

χ2 Test is used to test how likely is it that an observed distribution is due to chance/randomness.

Hypotheses:

- H_0 = features are stochastically independent (patterns are random)
- H_1 = there is a statistically significant relationship

Test:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - Ei)^2}{E_i}$$

$Shapiro - Wilk\ Test$ for Normality

Tests whether a random sample $x_1, x_2, \dots x_n$ comes from a normal distribution.

The hypotheses to be tested are:

H_0: $the\ population\ is\ normally\ distributed$

H_1: $the\ proportions\ is\ not\ normally\ distributed$

You are testing AGAINST the assumption of Normality.

!!Caution!! : This test does not apply to large data sets.

R-Session

1. Read the data
2. Perform Chi-square Test of Independence
3. Perform Shapiro-Wilk’s Test

PLAN FOR NEXT WEEK

That's it for today! :-)

Next week, we are going to discuss:

1. Other tests
2. Correlation and Regression

If you want to reach me, mail me at:

`prabesh.dhakal@stud.leuphana.de`