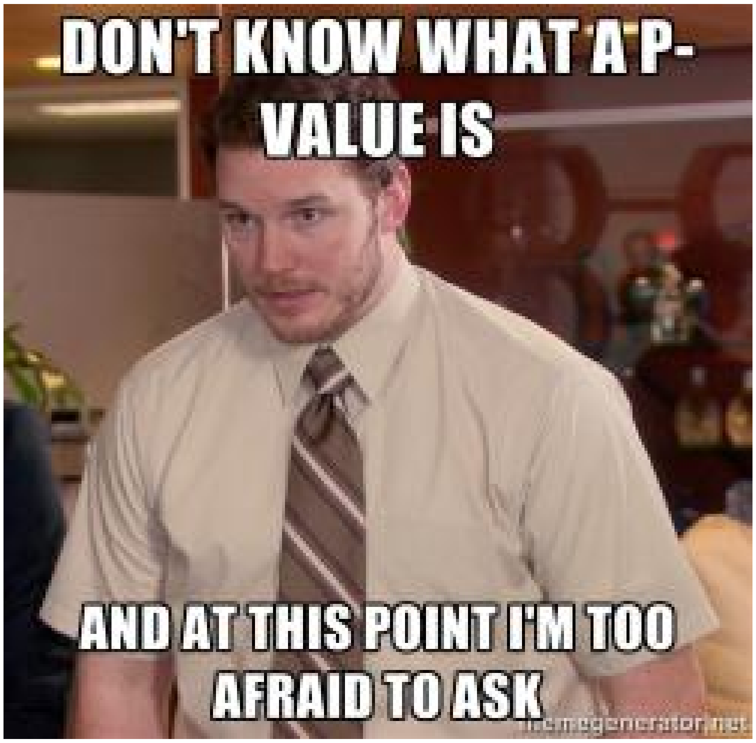
STA258 Master Notes

Haris Aljic

2023-11-24


DON'T KNOW WHAT A P-VALUE IS
AND AT THIS POINT I'M TOO AFRAID TO ASK
memegenerator.net

Contents

Module 1 - "Exploring Categorical Data"

- Categorical variables represent groups/categories like colour/gender/etc
- They can be nominal (non-ordered) or ordinal (ordered)
  - This is a "scale of measure" for the data
- They're graphical displays are most often bar graphs and pie charts

Frequency Distributions

- A "distribution" is a variables *pattern of variation*
- A frequency distribution orders a set of scores from highest to lowest
- Can be modeled as either a table or graph, but regardless, the *same two elements* are presented
  - Set of categories that make up the original measurement scale
  - A record of the frequency, or number of individuals in each category
- So a frequency distribution presents a **picture** of how individual scores/observations are distributed on the measurement scale

Describing Categorical Data

- Frequency (counts)
- Relative frequency (proportion)
  - Count of category divided by total count
- Percentage (proportion times 100)

This information is often *derived from* or *given* in a Frequency Distribution Table (and graphically displayed as either a bar graph or pie chart)

Consider the Titanic Example

```
titanic <- read.csv("Titanic.csv")
attach(titanic)
str(titanic)
```

```
## 'data.frame':    2201 obs. of  4 variables:
##  $ Survived: chr  "Dead" "Dead" "Dead" "Dead" ...
##  $ Age     : chr  "Child" "Child" "Child" "Child" ...
##  $ Sex     : chr  "Male" "Male" "Male" "Male" ...
##  $ Class   : chr  "Third" "Third" "Third" "Third" ...
```

We can see two types of distribution tables here for the "Survived" variable:

```
# See the Frequency Distribution Table for the Survived variable
addmargins(table(Survived))
```

```
## Survived
## Alive  Dead   Sum
##   711  1490  2201
```

This table tells us precisely HOW MANY (frequency/count) people survived or not

```r
# See the RELATIVE Frequency Distribution Table for the Survived variable
addmargins(prop.table(table(Survived)))
```

```
## Survived
##    Alive     Dead      Sum
## 0.323035 0.676965 1.000000
```

This table tells us the relative frequency (proportion) of survivors to the total population

We can extend this further to say that ≈ 32% survived, for example.

Variable Roles

A variable can be classified as **Response (outcome, dependent)**, or **Explanatory (predictor, independent)**

A study on two variable where one is a **response variable**, the other **explanatory**, we observe the outcomes of the response *depends* on the values of the explanatory

Contingency Tables and Marginal Distribution Tables

Consider the Ticket Classes for the Titanic dataset

```
## Class
##   Crew  First Second  Third  Total
##    885    325    285    706   2201
```

And define the contingency table as follows:

```r
Contingency.Table <- table(Survived, Class)
addmargins(Contingency.Table)
```

```
##         Class
## Survived Crew First Second Third  Sum
##    Alive  212   203    118   178  711
##    Dead   673   122    167   528 1490
##    Sum    885   325    285   706 2201
```

(Note this table is still classified as 2x4)

Notice how for any single cell, three different percentages exist.

For example, take the number of First Class passengers that are alive (203):

- Row percentage: $203/711 = 28.6\%$ of survivors in first-class
- Column percentage: $203/325 = 62.5\%$ of first-class passengers survived
- Overall percentage: $203/2201 = 9.2\%$ of passengers were in first-class

Marginal Distributions and Joint Distributions

The sum for each row or for each column give totals. These are the *margins*. In the Ticket Class example, we have two variables - **Survived** (row) and **Ticket Class** (column)

Marginal distributions can be displayed as counts or percentages.

```
## Survived
##    Alive     Dead      Sum
## 0.323035 0.676965 1.000000
```

Similar to Marginal Distributions are Joint Distributions, where each cell of the dataset sums together to the overall total, like so:

```
##         Class
## Survived       Crew      First     Second      Third        Sum
##    Alive 0.09631985 0.09223080 0.05361199 0.08087233 0.32303498
##    Dead  0.30577010 0.05542935 0.07587460 0.23989096 0.67696502
##    Sum   0.40208996 0.14766015 0.12948660 0.32076329 1.00000000
```

Comparing Two Conditional Proportions (Difference and Ratio of Proportions)

The **Difference of Proportions** is simply subtracting one proportion from another to determine which has higher **percentage points**

For example, with the ticket class data set, the first-class passengers who survived (62%), is 37 **percentage points higher** than the third-class passengers (25%)

It is incredibly important to note that the percentage of first-class survivors is NOT 37% higher than the percentage of third-class survivors. It is 37 **percentage points** higher.

The **Ratio of Proportions**, however, tells us much larger or smaller one proportion is to another.

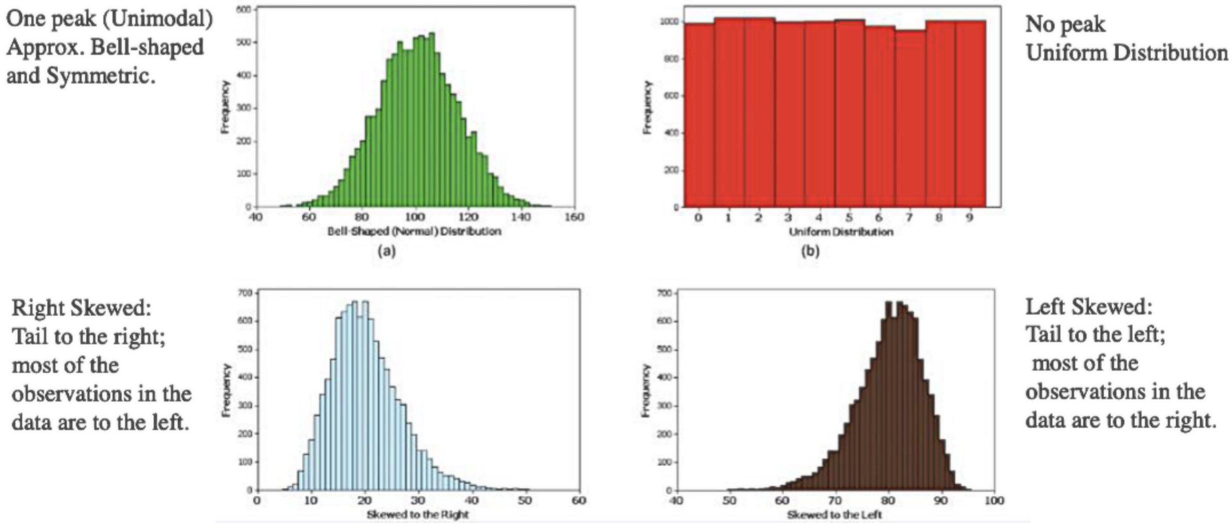
Following the same example from earlier, $0.62/0.25 = 2.48$ implies that the proportion of survived is 2.48 times larger for first-class passengers than for third-class (or that first-class passengers were 2.48 times more likely to have survived than third-class)

Interpreting Ratios

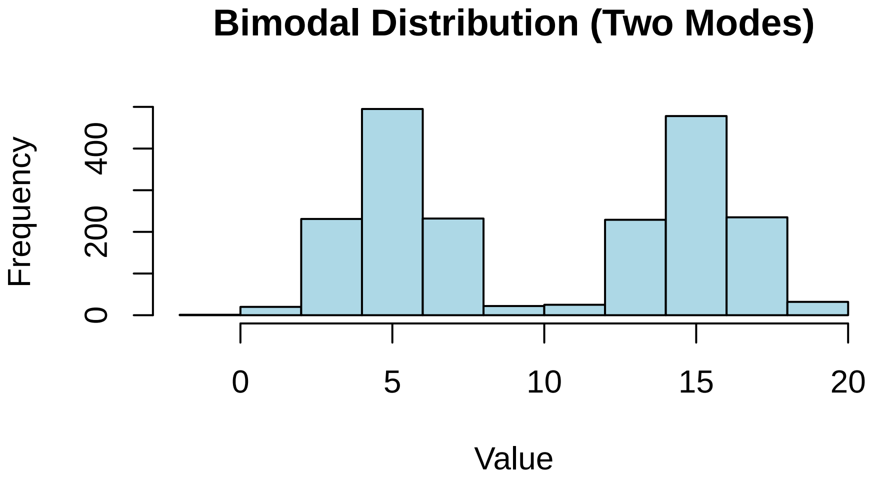
- When proportions are (nearly) identical, their difference is zero, ratio is one, and are said to have **no association** - The greater their difference, the stronger their association.
- Ratio of proportions is often preferred to the difference of proportions when the proportions themselves are both small.
- It is common to speak of *percent increases* when you have a ratio in between 1 and 2
  - For example, a ratio $\frac{a}{b} = 1.09$ implies that proportion a is 1.09 times larger than b, and hence is 9% higher.
- Conversely, a ratio below 1 implies a percent decrease - $\frac{b}{a} = 0.91 \implies 100\% - 91\% = 9\%$ decrease (note this is just the reciprocal of the earlier fraction)

Module 2 - Exploring Quantitaive Data

Shapes of Histograms



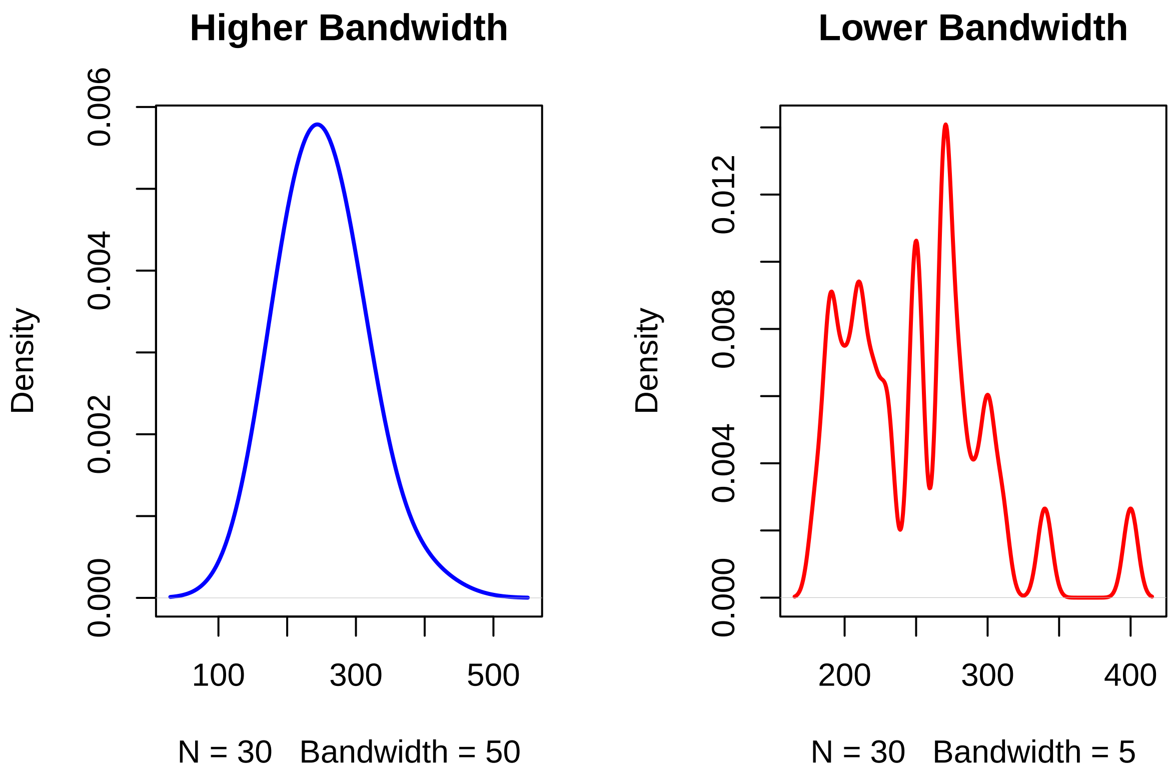
An easy way to remember, is that skewness corresponds to the tail. For example, if most of the observations are on the right side, that means there's a tail to the left, which means the distribution is **left-tailed**.



Density Plot

A **Density Plot** is a smooth, continuous function over an interval to visualize the distribution of data. The advantage of density plots over histograms is that they are unaffected by the number of bins (bars) used, and can have a shape or form with lower bin counts that histograms typically would not.

Another detail to consider when discussing density plots is their **bandwidth selection**. Higher bandwidth smooths out the curve, and will likely leaves out smaller variations. Lower bandwidth, however, shows the much finer details and fluctuations. Take a look below to see the difference bandwidth makes on the *same* dataset:



Centre and Central Tendency

The **centre** is defined to be a value in the data

Mean, Median, and Mode

Spread - Sample Variance (and Sample Standard Deviation)

Percentiles, Quartiles, and Interquartile Range

Z-Scores

Different Plots and how they relate to Normal Distributions

Module 3

Module 4

Module 5

Module 6

Module 7

Module 8

Module 9

Module 10

Module 11