STA220H1: The Practice of Statistics I

Elena Tuzhilina

January 24, 2023

Please turn on your videos :)

Figure 1: [picture source]

Feedback

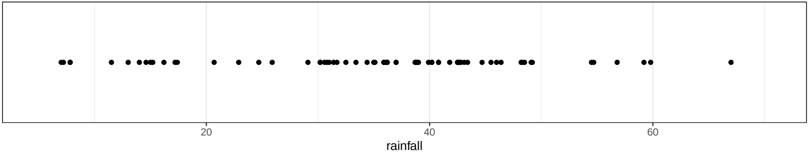
1. Use this form to provide your feedback for STA220 class. It is anonymous.
2. February 7 (in-person): 1 hour of review + 20 mins break + 80 mins test. The review will be recorded via Zoom and posted. The exam will be proctored.

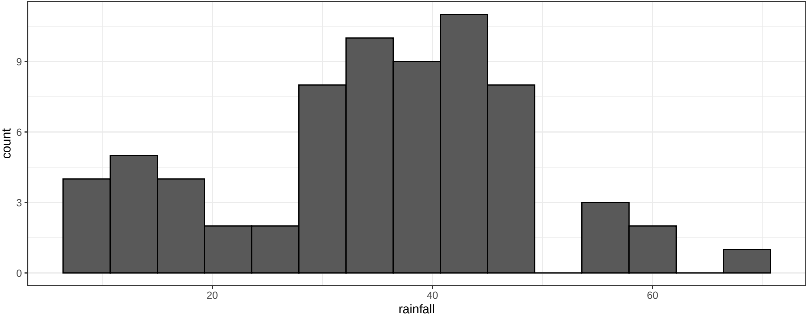
Agenda for today

- Recap: histogram, standard deviation, scatterplot and correlation
- Summarizing relationship between two variables: categorical vs categorical
- Introduction to probability

Recap: histogram

- **Histogram** is used for visualizing the **distibution** of a quantitative variable





Recap: standard deviation

- Measures the **spread** of a quantitative variable

$$variance = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2$$

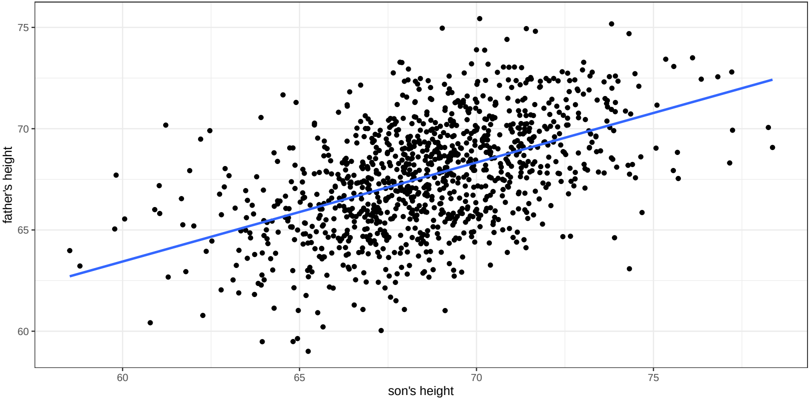
$$standard\ deviation = \sqrt{variance} = s_x$$

Can s_x be negative?

What happens to s_x when we multiply $x_1, \ldots, x_n$ by 2?

Recap: scatterplot

- **Scatterplot** is used for visualizing the relationship between to quantitative variables



Recap: covariance

- **Covariance** measures the relationship trend between two quantitative variables
- Positive covariance $\Rightarrow$ the variables tend to both increase together (on average!). Negative covariance $\Rightarrow$ one variable tends to increase when the other decreases (on average!)

$$covariance = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = cov_{xy}$$

What happens to cov_{xy} when we multiply $x_1, \ldots, x_n$ by 2?

What happens to cov_{xy} when we also multiply $y_1, \ldots, y_n$ by 3?

Recap: correlation

- **Correlation** is the scaled form of covariance
- Correlation value is between -1 and 1
- If there is a perfect linear relationship, e.g. $y_i = a \cdot x_i + b$, then correlation is 1 (if $a > 0$) or -1 (if $a < 0$)

$$correlation = \frac{cov_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = r_{xy}$$

What happens to r_{xy} when we multiply $x_1, \ldots, x_n$ by 2?

What happens to r_{xy} when we also multiply $y_1, \ldots, y_n$ by 3?

Data summary: categorical vs categorical variables

- **Numerical summary** is very limited - frequencies and relative frequencies
- Use **plots** - barplot

Data summary: categorical vs categorical variables

Data set: provides information on the fate of 891 passengers on the fatal maiden voyage of the ocean liner “Titanic”, summarized according to economic status (class), sex, age and survival.

PassengerId	Sex	Age	Class	Survived
1	male	22	3	No
2	female	38	1	Yes
3	female	26	3	Yes
4	female	35	1	Yes
5	male	35	3	No
6	male	NA	3	No
7	male	54	1	No
8	male	2	3	No
9	female	27	3	Yes
10	female	14	2	Yes
11	female	4	3	Yes
12	female	58	1	Yes
13	male	20	3	No
14	male	39	3	No
15	female	14	3	No

Numerical summary: joint distribution

Is it true that rich people (e.g. 1st class passengers) survived more often that poor people (e.g. 3rd class passengers)?

```
table(titanic.data$Class)
```

```
##
##   1   2   3
## 216 184 491
```

```
table(titanic.data$Survived)
```

```
##
##  No Yes
## 549 342
```

Numerical summary: joint distribution

- **Joint distribution** is the frequency/relative frequency of observations for a combination of two variables

```
tab = table(titanic.data$Class, titanic.data$Survived)
tab
```

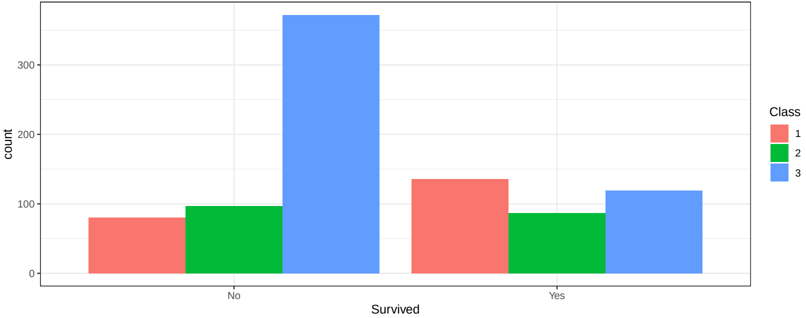
```
##
##      No Yes
##   1  80 136
##   2  97  87
##   3 372 119
```

```
ptab = prop.table(tab)
ptab
```

```
##
##             No        Yes
##   1 0.08978676 0.15263749
##   2 0.10886644 0.09764310
##   3 0.41750842 0.13355780
```

Plots: barplot

- There are many 3rd class passengers that did not survive
- But it is hard to compare as there were many people who did not survive



Numerical summary: marginal distribution

- **Marginal distribution** is the frequency/relative frequency of only one variable

```
addmargins(tab)
```

```
##
##          No Yes Sum
##   1      80 136 216
##   2      97  87 184
##   3     372 119 491
##   Sum   549 342 891
```

Numerical summary: conditional distribution

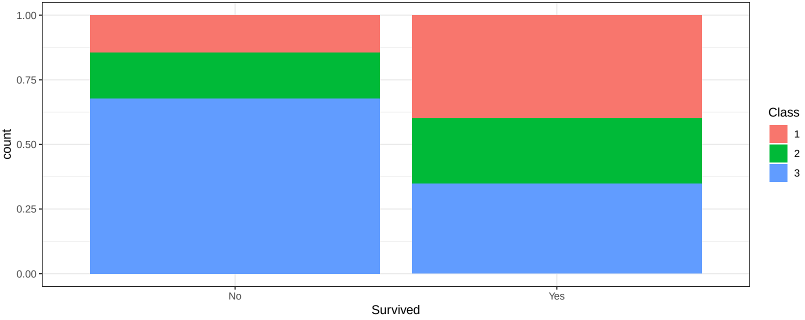
- **Conditional distribution** is the distribution of one variable within a fixed value of a second value
- Comparing conditional distributions for each cetegory can tell if there is any relationship between two variables

```
##
##        No Yes
##   1    80 136
##   2    97  87
##   3   372 119
##   Sum 549 342
```

```
##
##              No       Yes
##   1   0.1457195 0.3976608
##   2   0.1766849 0.2543860
##   3   0.6775956 0.3479532
##   Sum 1.0000000 1.0000000
```

Plots: stacked barplot

- Two variables are **independent** if conditional distribution of one variable is the same for all values of the other variable



Exercise

Find conditional distribution of Sex and Survived variables. Do you think there is any relationship?

```
##
##            No Yes
##   female   81 233
##   male    468 109
```

Randomness

We encounter randomness every day!

- coin flip: heads or tails?
- weather: will it rain today?
- lottery: will by ticket win?
- driving: will I be late for a meeting?
- health: will the surgery be successful?

We don't know the exact outcome, but we know that there is a structure to how often different outcomes occur.

Probability

Probability is a number that describes the likelihood of some event occurring that ranges from zero (impossibility) to one (certainty).

What are the chances of

- getting heads in one coin flip? 50%
- snowing tomorrow? 80%
- of wining in Lotto Max? 0.000003 %
- dying from COVID-19? 3.4%
- sharing a birthday for two students in class? (more than 50% if more that 23 students)

Ways to compute probability

Theoretical - compute the probability directly based on our knowledge of the situation.

- If we flip a coin we have two possible outcomes
- Since the coin is "balanced", H and T will have equal probability
- The probability of heads is $1/2$ and tails is $1/2$

Here we made an assumption that the coin is balanced.

Ways to compute probability

Empirical - experiment many times and count how often each event happens. Interpret relative frequency of the different outcomes as the (approximation of) probability of each outcome.

- There were 272 rainy days in Toronto in the past two years
- The probability to rain in Toronto is 272/730

Here we describe the probability based on observed results.

Theoretical vs. empirical

Let's flip a coin 100, 1000, 10000 and 100000 times.

```
##
##  H  T
## 49 51
```

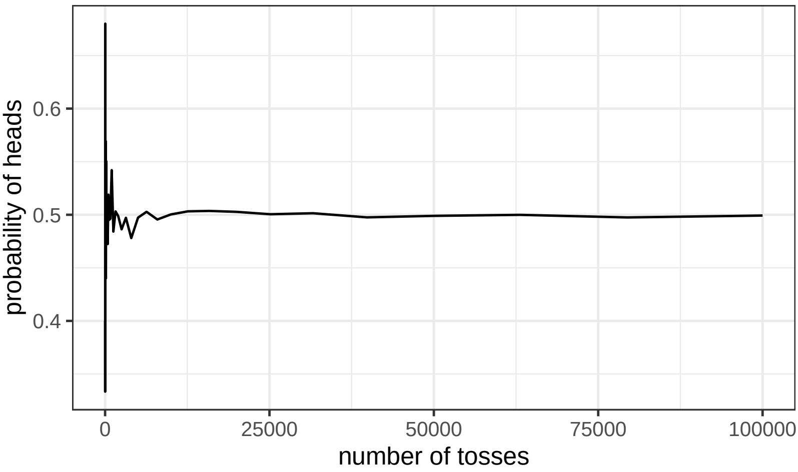
```
##
##   H   T
## 498 502
```

```
##
##    H    T
## 5007 4993
```

```
##
##     H     T
## 50033 49967
```

Theoretical vs. empirical

Theoretical probability shows what will happen in "long run"



Probability: vocabulary

Experiment - any activity that produces an outcome

- flipping a coin
- rolling a 6-sided die
- checking number of Facebook friends
- computing the weight of a new-born baby

Probability: vocabulary

Sample space - the collection of all possible outcomes of an experiment

$$S = \{O_1, O_2, \ldots, O_N\}$$

- flipping a coin
- rolling a 6-sided die
- checking number of Facebook friends
- computing the weight of a new-born baby

Probability: vocabulary

Event - a subset of the sample space (could be one or more of possible outcomes in the sample space)

$$A = \{O_1, O_3, O_{10}\}$$

Outcomes are also called elementary events

- flipping a coin and getting heads
- rolling a 6-sided die and getting an odd number
- having less than 100 friends on Facebook
- a baby weighting more than 4kg

Probability of outcomes

Probability cannot be negative

$$P(O_i) \geq 0$$

The total probability of all outcomes is one

$$\sum_{i=1}^{N} P(O_i) = 1$$

The probability of an outcome cannot be greater than one

$$P(O_i) \leq 1$$

Probability of outcomes

Very often we can assume that all outcomes in the sample space are equally likely

$$P(O_i) = \frac{1}{total\ number\ of\ outcomes}$$

- tossing a coin
- rolling a 6-sided die

Probability of events

To compute the probability of a complex event, add together the probabilities of the outcomes.

If all outcomes in the sample space are equally likely then the probability of event A is

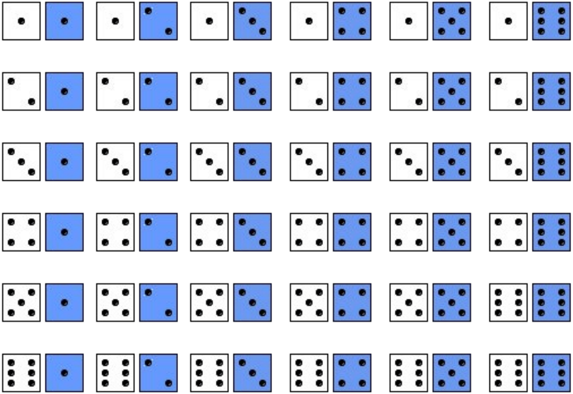
$$P(A) = \frac{number\ of\ outcomes\ favorable\ to\ event\ A}{total\ number\ of\ outcomes}$$

- rolling a 6-sided die and getting an odd number
- rolling two 6-sided dies and getting a number less than 4

Exercise

We roll two dice. What are outcomes? How many outcomes are there? What is the sample space?

Exercise



Exercise

What is the probability to get equal numbers when rolling two dice?

Exercise

	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Probability rules: subtraction

The **complement** of an event, $\bar{A}$ is the the event that A does not happen.

Probability of event A not happening is one minus the probability of the event happening.

$$P(\bar{A}) = 1 - P(A)$$

- rolling a 6-sided die and getting an odd number
- rolling a 6-sided die and getting an even number

Venn diagram: one event

Probability rules: multiplication

If events A and B are independent then

$$P(\textit{both } A \textit{ and } B \textit{ will occur}) = P(A) \cdot P(B) = P(A \cap B)$$

Exercise

What is the probability of rolling two dice and getting one on both roll?

	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Exercise

What is the probability rolling three dice and getting at least one six?

Probability rules: addition

To find the probability of either of two events occurring, add together the individual probabilities, then subtract the probability of both occurring together

$$P(either\ A\ or\ B\ occur) = P(A) + P(B) - P(A \cap B) = P(A \cup B)$$

Exercise

What is the probability of rolling two 6-sided dice and getting one on either roll?

	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Venn diagram: two events

Conditional probability

We often wish to determine the probability of some event given that some other event has occurred, which are known as **conditional probabilities**.

- The probability of a heart attack is

$$P(heart\ attack) = 0.1$$

- The probability of being physically active?

$$P(active) = 0.5$$

- What is the probability that a person has heart attack, given that he/she is physically active?

$$P(heart\ attack|active) = ?$$

Conditional probability

To compute the conditional probability of A given B we need to know the joint probability (that is, the probability of both A and B occurring) as well as the overall probability of B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional probability

You can use joint distribution table to find the conditional distribution.

```
##                 active not-active
## heart attack      0.01       0.09
## no heart attack   0.49       0.41
```

Independence

Statistical **independence** between events A and B means that knowing about A does not tell us anything about B.

That is, *the probability of A given some value of B is just the same as the overall probability of A*

$$P(A|B) = P(A)$$

This is equivalent to

$$P(A \cap B) = P(A) \cdot P(B)$$

Exercise

Are events *being active* and *heart attack* independent?

```
##                    active not-active
## heart attack        0.01       0.09
## no heart attack     0.49       0.41
```

Conditional probability: tree diagram

- The probability of getting COVID-19 is

$$P(covid) = 0.5$$

- Given that you have COVID-19, the probability of getting a positive test is

$$P(positive|covid) = 0.9$$

- Given that you do not have COVID-19, the probability of getting a positive test is

$$P(positive|no\ covid) = 0.01$$

Conditional probability: tree diagram

What is the probability to have COVID given that your test is positive?

$$P(covid|positive) = ?$$

Exercsie

Create the tree diagram using joint distribution table.

```
##                  active not-active
## heart attack       0.01       0.09
## no heart attack    0.49       0.41
```

Bayes' rule

In many cases, we know $P(A|B)$ but we really want to know $P(B|A)$.

In order to reverse a conditional probability, we can use **Bayes' rule**

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|\bar{B}) \cdot P(\bar{B})}$$

TO DO

1. Module 1. Summarizing Data: Relationships Between Variables and Module 2. Introduction to Probability: Events
2. Quiz 3 due Monday (January 30) @ 11:59 PM (EST)
3. Practice Problem Set 3