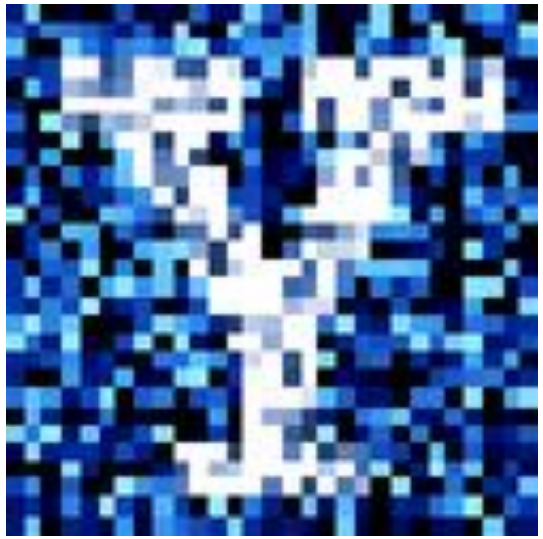
YData: Introduction to Data Science



Lecture 22: Examples

Overview

Python basics

Tables

Histograms

Probability

Testing hypotheses

Announcements

Exam information

- 50 minutes
- On paper
- In class on the 18th
- We will provide you with the function sheet that is on Canvas
  - No other "cheat sheet" allowed
- Just bring some pencils/pens
  - And press hard so we can read your writing when it is scanned

If you have accommodations and have not spoken with me yet, please contact me ASAP

Python basics

Python basics

Function example

```
def spread(values):
    return max(values) - min(values)
```

For loop example

```
animals = make_array("cat", "dog", "bat")
for creature in animals:
        print(creature)
```

Conditional statements

Conditional statements control the sequence of computations that are performed in a program

We use the keyword if to begin a conditional statement to only execute lines of code if a particular condition is met.

We can use elif to test additional conditions

We can use an else statement to run code if none of the if or elif conditions have been met.

```
num = 5
if num == 1:
    print("Monday")
elif num == 2:
    print("Tuesday")
elif num == 3:
    print("Wednesday")
elif num == 4:
    print("Thursday")
elif num == 5:
    print("Friday")
elif num == 6:
    print("Saturday")
elif num == 7:
    print("Sunday")
else:
    print("Invalid input")
```

Exercise

Write a function `sum_up_to()`

- Takes a positive number k as an input argument
- Adds numbers together from 1 to k
- If k is negative, return the string "invalid input fool"

```
def sum_up_to(k):
    if k < 0:
        return "invalid input fool"

    total = 0
    for i in np.arange(k + 1):
        total = total + i
    return total
```

Exercise

Write a function `sum_up_to()`

- Takes a positive number k as an input argument
- Adds numbers together from 1 to k
- If k is negative, return the string "invalid input fool"

```
def sum_up_to(k):
    if k < 0:
        return "invalid input fool"

    return np.sum(np.arange(k + 1))
```

Tables

Table operations

See the list of methods…

Examples: SF employee salaries

```
sf = Table.read_table('san_francisco_2015.csv')
```

1. What is the Mayor's total compensation?

2. What proportion of employees make more than $100,000 in total compensation?

Let's explore this in Jupyter!

```
Table()
Table.read_table(filename)
tbl.num_rows
tbl.num_columns
tbl.labels
tbl.with_column(name, values)
tbl.with_columns(n1, v1, n2, v2…)
tbl.column(column_name_or_index)
tbl.select(col1, col2, …)
tbl.drop(col1, col2, …)
tbl.relabeled(old_label, new_label)
tbl.take(row_indices)
tbl.sort(column_name_or_index)
```

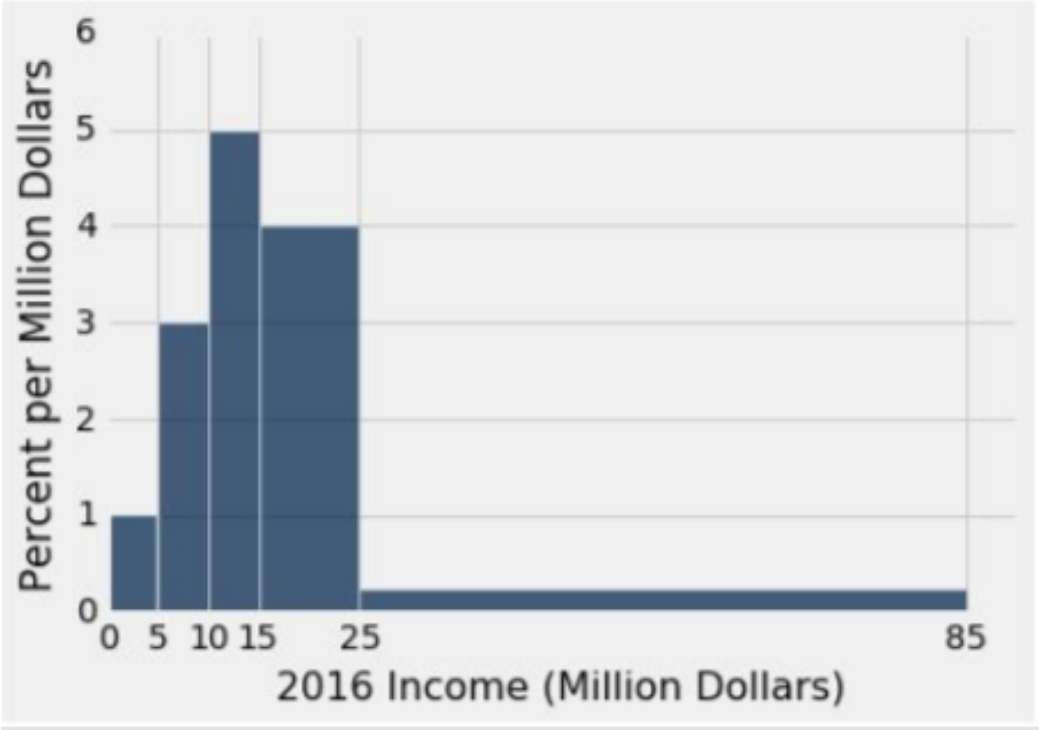
Histograms

Using the density scale

The **area** of bin in a normalized histogram is the percentage of values in a particular bin

Questions:

- What percent of values are between 0 and 5?
  - A: 1 * 5 = 5%
- What percent of values are between 10 and 25?
  - A: 5 * 5 + 10 * 4 = 65%



Probability

Probability rules

Probability models assigns values between 0 and 1 to random events

Multiplication rule: the chance that two events A and B both happen is:

P(A happens) x P(B happens given that A has happened)

Addition rule: If event A can happen in exactly one of two ways, then:

P(A) = P(first way) + P(second way)

Probability exercises

Marbles: G, G, G, G, R, R, R, B, B, Y

Draw 4 at random

P(no G) = ?

If with replacement:

(6/10)*(6/10)*(6/10)*(6/10)

If without replacement:

(6/10)*(5/9)*(4/8)*(3/7)

P(all G) = ?

If with replacement:

(4/10)*(4/10)*(4/10)*(4/10)

If without replacement:

(4/10)*(3/9)*(2/8)*(1/7)

Probability exercises

Marbles: G, G, G, G, R, R, R, B, B, Y

Draw 4 at random with replacement

P(at least one G) = ?

1 - (6/10)*(6/10)*(6/10)*(6/10)

Testing hypotheses

Start by stating the hypotheses

Figure out the viewpoint the question wants to test, and formulate:

- Null hypothesis: Completely specified chance model under which you can simulate data
- Alternative hypothesis: Viewpoint comes from the question
- Test statistic: to help you choose one viewpoint

Say what kind of values of the statistic will make you lean towards each alternative

Testing hypotheses for categorical data

Example: Mendel's model

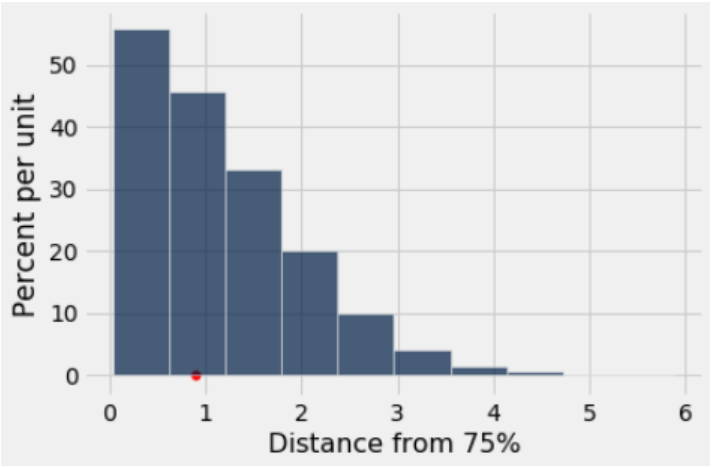
Null: Each pea plant has 75% chance of being purple flowering

Alternative: The model isn't good.

Test statistic: | percent purple in sample - 75|

P-value direction: to the right

Mendel's peas



statistic: absolute distance from 75%

Example: Alameda County jury panels

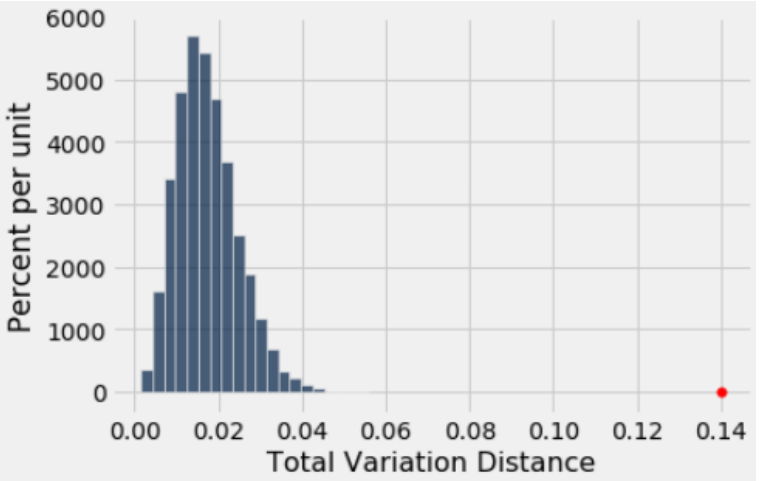
Null: The Alameda County jury panels were drawn at random from the specified distribution of eligible jurors

Alternative: The panels were not drawn at random from the specified distribution.

Test statistic: TVD

P-value direction: to the right

Alameda county jury panels



statistic: TVD between distributions

Testing hypotheses for numerical data

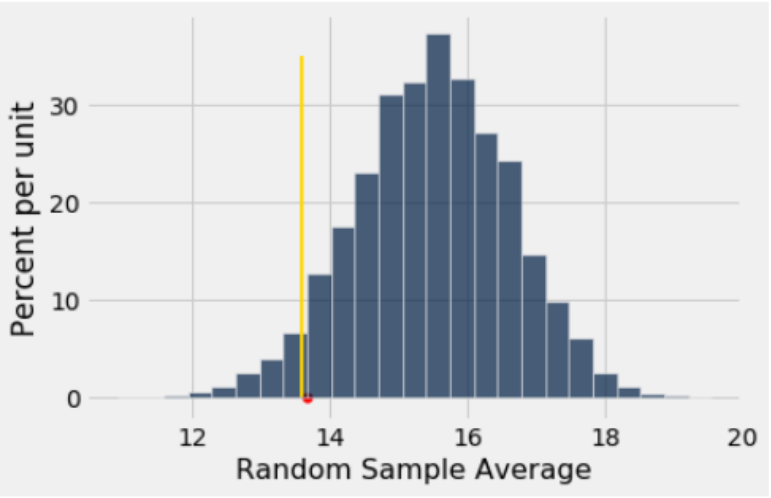
Example: Graduate student instructors

Null: Section 3 scores are like a sample drawn at random without replacement from the whole class.

Alternative: The Section 3 average is too low for the section to be a random sample from the class.

Test statistic: Section 3 average

P-value direction: to the left



Testing hypotheses for comparing two samples

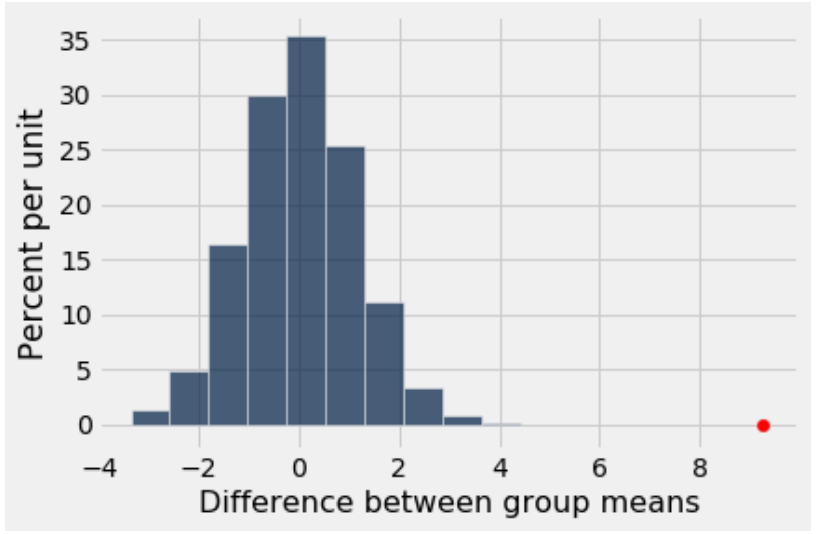
Birthweights

Null: In the population, the distributions of the birth weights of the babies in the two groups are the same.

Alternative: In the population, the babies of the mothers who didn't smoke (B) were heavier, on average, than the babies of the smokers (A).

Test statistic: Group B sample average - Group A sample average

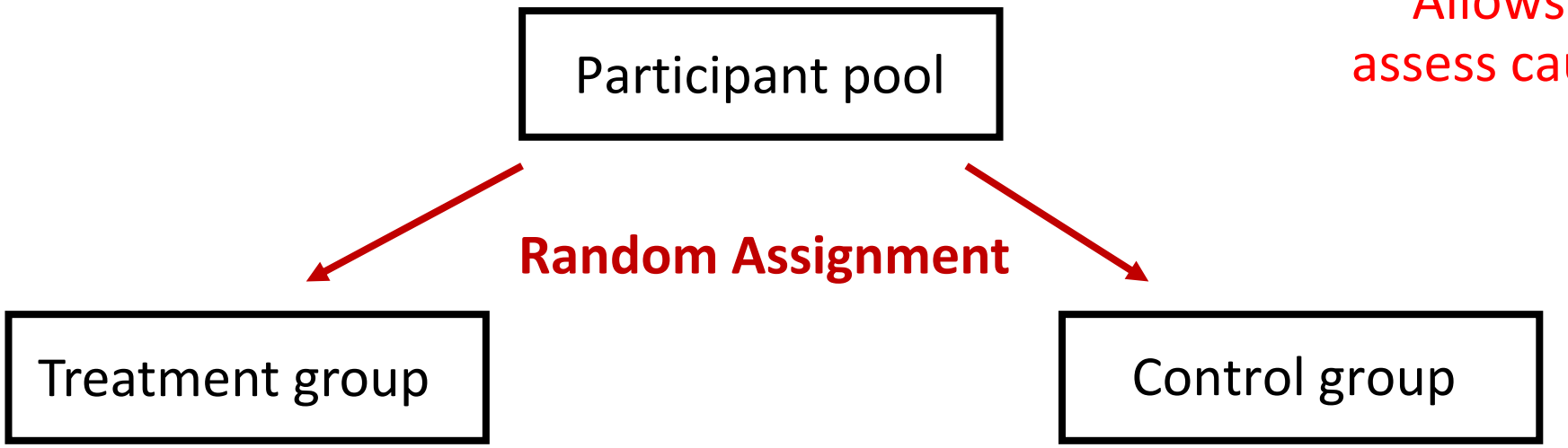
P-value direction: to the right



Randomized Controlled Trial

Take a group of participant and ***randomly assign***:

- Half to a *treatment group* where they get chocolate
- Half in a *control group* where they get a fake chocolate (placebo)
- See if there is more improvement in the treatment group compared to the control group



Allows us to assess causation!

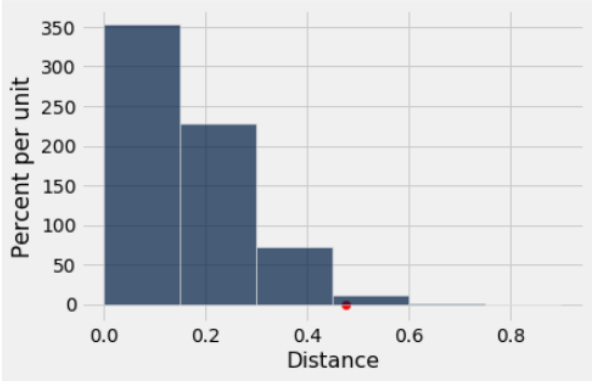
Randomized Controlled Trial

Null: The distribution of all the potential control scores is the same as the distribution of all the potential treatment scores.

Alternative: The distribution of all the potential control scores is different from the distribution of all the potential treatment scores.

Test statistic: | control group average - treatment group average |

P-value direction: to the right



GOOD LUCK in your EXAMS