3.1 Data collection

Learning goals

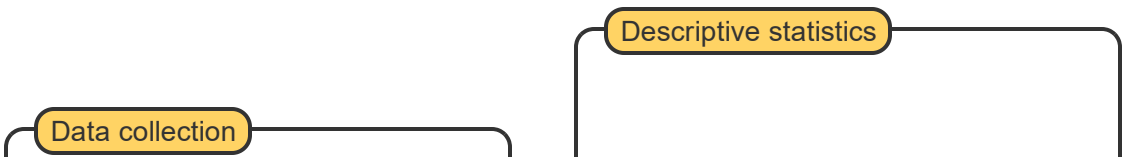
- Explain the role of statistics in data science.
- Define and compare populations and samples.
- Define and compare sampling methods.
- Define and compare an observational study and an experiment.
- Use Python to obtain a random sample from a dataset.

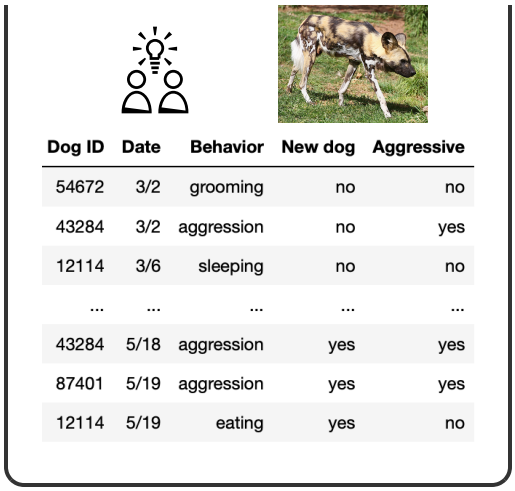
Statistics for data science

The field of statistics provides techniques for collecting, analyzing, and gaining insights from data. Data science relies on statistics to make data driven decisions and solve problems. Ex: A data scientist uses a statistical model to predict sales based on possible marketing strategies. Statistical techniques are used throughout the data science lifecycle.

- Sampling methods are used to efficiently and effectively collect data and reduce bias.
- Descriptive statistics provide ways to explore the observed data through visualizations and numerical summaries.
- Inferential statistics, like modeling and estimation, allow for conclusions to be drawn about the population from the observed sample of data. Probability is the foundation of inferential statistics.
- Understanding the statistical techniques of a data science project ensures results are appropriately interpreted and communicated.

PARTICIPATION ACTIVITY 3.1.1: Statistics in a data science project.





Animation content:

Animation describes the statistics components involved in a data science project working with a zookeeper.

- Step 1: Describes data science project. Two figures are shown collaborating to represent the zookeeper and data scientist along with a picture of a painted dog.
- Step 2: Describes data collection. A dataframe is shown with the following features: dog ID, date, behavior, whether or not the new dog had been introduced, and whether or not the behavior was considered aggressive. Several rows of the data are displayed.
- Step 3: Describes descriptive statistics. The aggressive feature from the dataset is summarized by a bar graph and numbers. The bar graph shows the proportion of aggressive behaviors observed before and after the new dog was introduced. The before group has a proportion of 0.128 aggressive behaviors and the after group has a proportion of 0.224 aggressive behaviors. The observed difference of 0.096 appears under the graph.
- Step 4: Describes inferential statistics. Output from a hypothesis test is given. The observed difference is 0.096, the test statistic is -1.993, and the p-value is 0.046.
- Step 5: Two figures are shown communicating representing the zookeeper and data scientist. A speech bubble is coming from the data scientist saying "Association, not causation".

Animation captions:

1. A data scientist collaborates on a project with a zookeeper to study behavior changes in wild African painted dogs when a new dog is introduced into the pack.

2. An observational study is designed to collect behavior data on a random sample of dogs in the pack before and after the new dog is introduced.
3. Graphical and numerical summaries of the data show the proportion of aggressive behaviors increased by 0.096 (0.128 to 0.224) after the new dog was introduced.
4. The observed increase was found to be statistically greater than 0 which indicates aggres behavior increases in the pack when a new dog is introduced.
5. The data scientist can report to the zookeeper that the data indicate introducing a new dog to the pack is associated with an increase in aggressive behavior.

PARTICIPATION ACTIVITY 3.1.2: Statistics and data science.

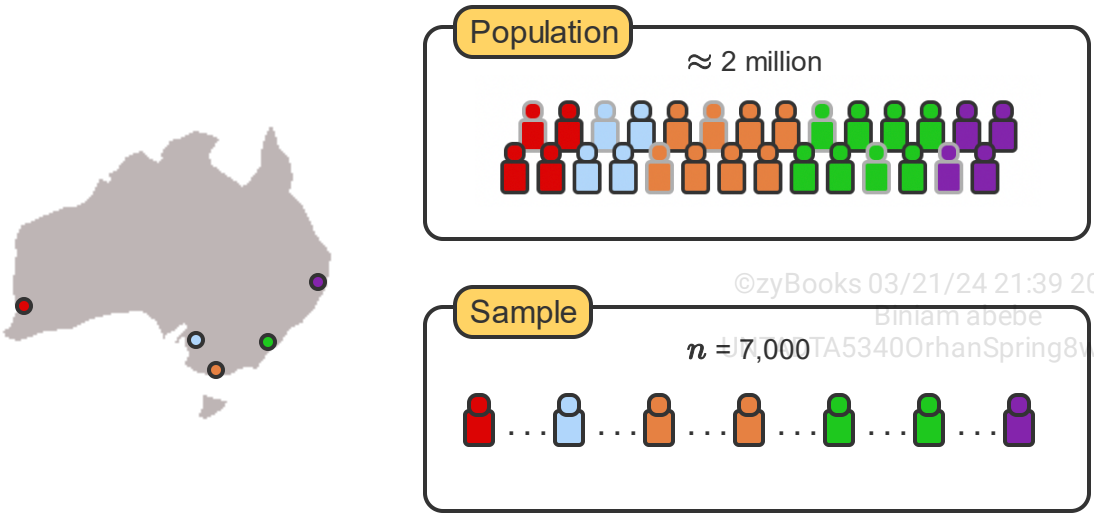
1) Is the modeling data step the only step of the data science lifecycle that involves statistics?
   - Yes
   - No

2) Which task would require a data scientist to have an understanding of probability?
   - Filtering a dataset to include only data from the current year.
   - Using software to create a scatter plot of temperature and electricity consumption.
   - Making predictions from a regression model.

Sampling

A common data science project goal is to analyze data to understand a population. A ***population*** is the entire set of all individuals, items, or events of interest. An ***observational unit*** is an individual, item, or event of the population on which data can be recorded. Because constraints on resources can make obtaining data on the entire population impractical, many projects use data from a sample. A ***sample*** is the subset of observational units from the population from which data will be collected. The number of observational units in the sample is denoted by n.

PARTICIPATION ACTIVITY 3.1.3: Passenger satisfaction in Australia.



Animation content:

The animation describes how a sample from a population can be selected.

- Step 1: A map of Australia with the 5 largest cities marked by black dots is shown.
- Step 2: The population of approximately 2 million adult train passengers from Australia is represented by several people figures all white with black outlines.
- Step 3: Seven randomly selected people figures turn gray in the population box to indicate they have been selected in the sample. A box labeled sample appears with 7 people figures from the population who represent the n=7,000 passengers to be surveyed.
- Step 4: The dots representing the 5 cities on the map all turn a unique color. Each figure in the population turns a color to match one of the 5 city colors, all 5 colors are represented in the population.Each figure in the sample also turns a color to match a city, with each city represented by at least one figure in the sample.

Final static image: Outline of Australia with the 5 cities highlighted by dots, each dot a different color. A box labeled population includes several people figures representing the population of approximately 2 million adult passengers. People figures are colored to match the 5 city colors indicating individuals in the population were from all 5 cities in Australia. A box labeled sample includes 7 people figures to represent the n=7,000 sample size. The people are colored to match the city colors, with each color represented at least once in the sample, to indicate individuals in the sample came from all 5 cities.

Animation captions:

1. A project to understand train passengers' satisfaction with train usage in the five largest cities in Australia will use data collected from surveying passengers.
2. Because of logistics and cost, all adult train passengers from the five cities in Australia cannot be surveyed.
3. A sampling method is used to select the 7,000 passengers to survey.
4. The sampling method incorporated random sampling from each city to obtain a representative sample and reduce possible bias from the sample selection.

PARTICIPATION ACTIVITY 3.1.4: Passenger satisfaction in Australia.

Match each description from the train satisfaction project with the appropriate term.

If unable to drag and drop, refresh the page.

Population **Sample** **Observational unit**

	The 7,000 adult train passengers surveyed
	An adult train passenger
	All adult train passengers from the five cities in Australia

Reset

Sampling methods

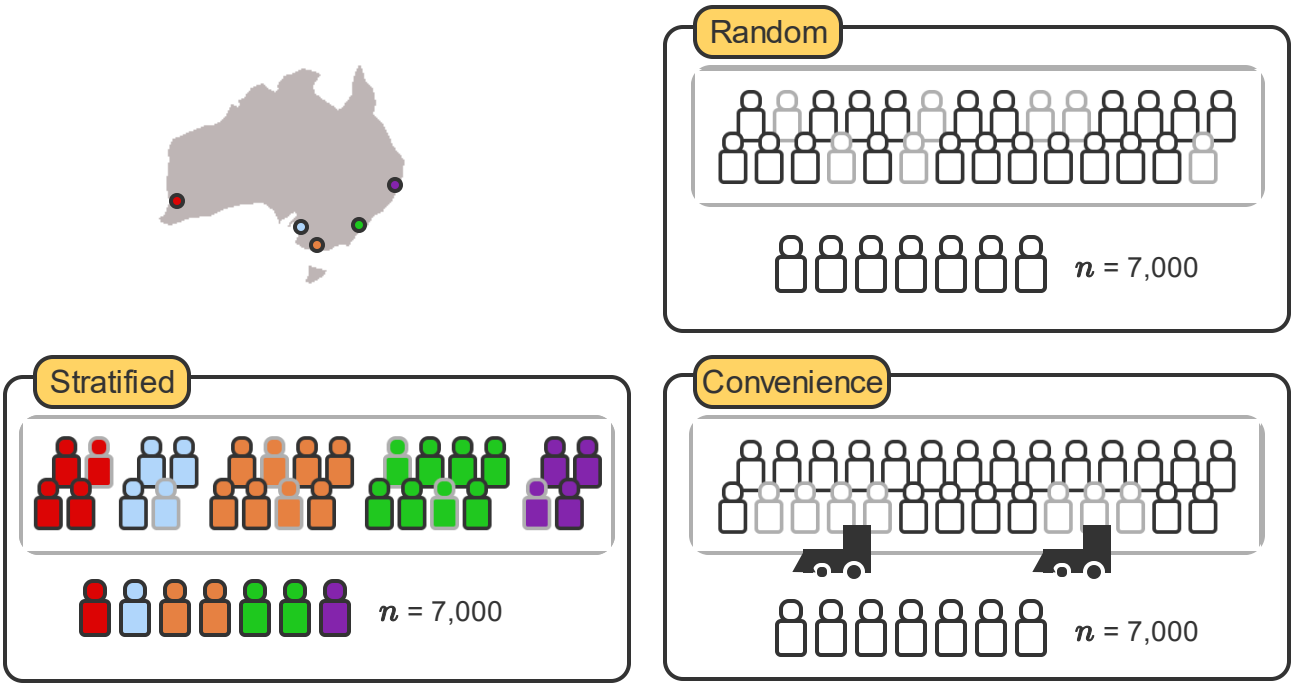
A ***sampling method*** is a process by which observational units are selected from the population to be included in the sample. Ideally, the observational units in the sample are representative of the population. Common sampling methods include the following:

- In ***random sampling***, observational units are selected at random from the population in which each subset of n units is equally likely.
- In ***stratified sampling***, the population is first divided into groups, called strata, based on a meaningful feature, then observational units are selected from each stratum. The strata feature is typically related to the primary features of interest in the study, and thus, all strata are sampled from.
- In ***cluster sampling***, the population is first divided into groups, called clusters, based on a feature, then a random sample of some clusters is selected. The cluster feature is typically not

related to the primary features of interest in the study, and thus, not all clusters are sampled from.

- In ***systematic sampling***, from a randomly selected starting point of the population, every k^{th} observational unit is selected. Typically the value of k is selected to be the population size divided by the desired sample size, n.
- In ***convenience sampling***, observational units from the population that are easier to include a selected.

PARTICIPATION ACTIVITY 3.1.5: Sampling methods for the passenger satisfaction project.



Animation content:

Animation shows how different sampling methods could have been used to select a sample from the population to survey in the Australian train passenger satisfaction project.

- Step 1: An outline of Australia with the 5 cities marked by black dots appears.
- Step 2: A box labeled random appears and several people figures, who all look the same, appear and fly from the map of Australia into the random box to represent the population. Seven randomly selected people figures in the population turn gey and are moved to represent the n=7,000 selected in the sample.
- Step 3: A box labeled stratified appears. The 5 city dots on the map each turn a unique color. Several people figures with colors the same as the 5 cities appear and fly from the map of

Australia into the stratified box to represent the population, grouped by city. Seven randomly selected people figures, at least one from each city, turn gray in the population and are moved to represent the n=7,000 selected in the sample.

- Step 4: a box labeled convenience appears and several people figures, who all look the sa appear and fly from the map of Australia into the convenience box to represent the population. Next, two tain images appear and the 7 people figures closest to the trains turn gray in the population and are moved to represent the n=1,000 selected in the sample.

Animation captions:

1. To select a sample of train passengers from the population of all adult passengers, any of the common sampling methods could be used. Each sampling method has benefits and costs.
2. In random sampling, passengers are selected at random from a list of all passengers in the five cities. Random sampling reduces the potential for sampling bias.
3. Stratified sampling ensures adequate representation from each city. Passengers are first divided into groups, or strata, based on city. Then from each strata, passengers are selected at random.
4. Selecting passengers waiting in the train stations uses convenience sampling. This method is easy and quick, but the sample is not likely representative of all train passengers.

PARTICIPATION ACTIVITY 3.1.6: Sampling methods.

A middle school (5th-8th grade) would like to survey students about school lunches. Match each sample selection description with the appropriate sampling method.

If unable to drag and drop, refresh the page.

Convenience **Cluster** **Stratified** **Systematic** **Random**

	As students go through the lunch line, survey the 10th student and then select every 25th student to survey.
	Group students by classroom, then randomly select 10 classrooms. Survey all students in the 10 selected classrooms.

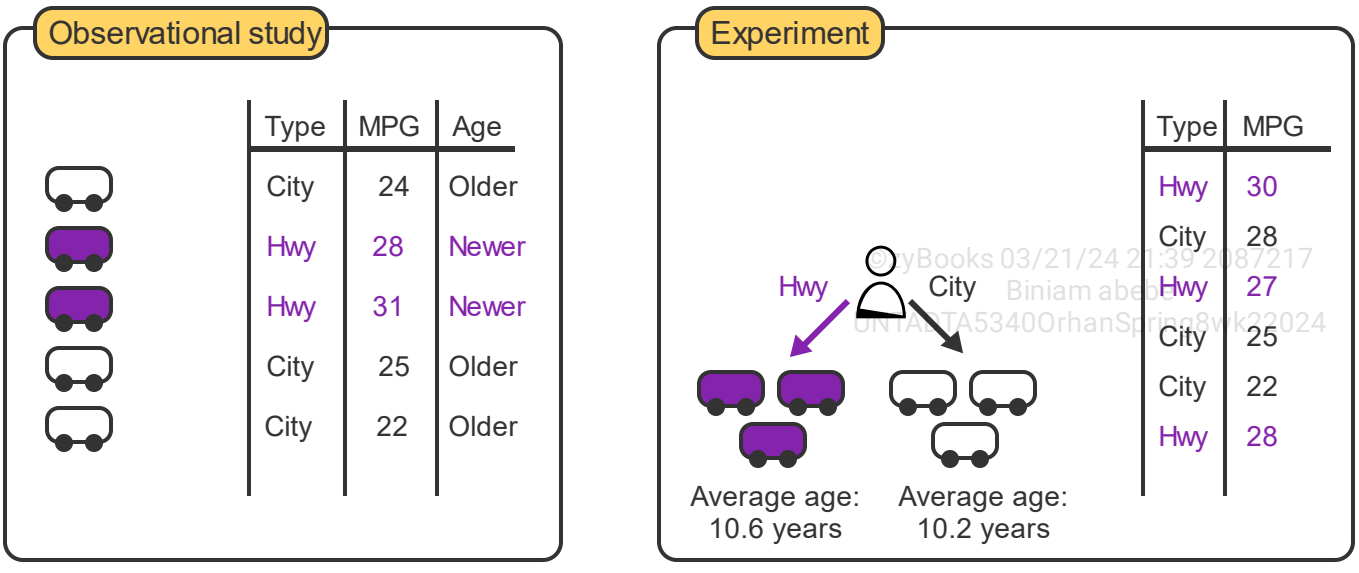
	From all 5th-8th grade students, randomly select 200 students to survey.
	Survey the first 200 students who arrive at school.
	Group students by grade, then randomly select 50 students from each of the four grades to survey.

Reset

Observational studies and experiments

Data can be collected either through an observational study or an experiment. In an ***observational study***, data is collected by recording the responses as they occur without any direct influence on the observed data. Ex: Collecting data on the features of homes listed for sale. In an ***experiment***, treatments are first assigned to observational units and then responses are recorded. Ex: An A/B test is conducted by randomly assigning participants to view one of two possible web page layouts then collecting data on webpage clicks for each layout. With random assignment of treatments to observational units, ***causal conclusions***, or concluding the treatments are likely to be the cause of the observed responses, can be made.

PARTICIPATION ACTIVITY 3.1.7: Collecting fuel efficiency data.



Animation content:

Animation shows the difference between an observational study and an experiment.

- Step 1: A box labeled Observational study appears along with a data table with column headings car, type, and MPG. A car appears and the first row of data is filled in for this car with values 1, city, and 24. Next, a second car appears and the second row of data is filled in for this car with values 2, Highway, and 28. Then three more cars appear and the next three rows of data are filled in the table.

- Step 2: An age column is added to the data table with two possible values, older and newer. The cars with city-type driving are all older and the cars with highway-type driving are all newer.

- Step 3: A box labeled Experiment appears with a data table with the same column headings: car, type, and mpg. Next 6 cars and a person figure appear. One car moves by the person and then moves under a highway heading. A second car moves by the person and then moves under a city heading. Next, the remaining 4 cars move down by the person and then get split under the two headings, so there are 3 cars in the highway group and 3 cars in the city group. After cars have been grouped, the data table fills in with values from the 6 cars.

- Step 4: The highway group of cars is labeled with an average age of 10.6 years and the city group of cars is labeled with an average age of 10.2 years.

Animation captions:

1. Data collected in an observational study are recorded as they occur, without prior influence. Ex: Sampling car owners and asking their car's efficiency in miles per gallon (MPG) and type of driving (highway or city).
2. The study does not control the type of driving for each car, so any relationships between MPG and type of driving are only associations. Other features, like age, may influence the observed relationships.
3. Data collected in an experiment are directly influenced by treatment assignment prior to data collection. Ex: One type of driving (treatment) is randomly assigned to each car, then the MPG is calculated and recorded.
4. By assigning treatments to cars, other related features, like age, should be balanced between the highway and city groups allowing for causal conclusions to be made.

PARTICIPATION ACTIVITY 3.1.8: Observational studies and experiments.

1) Count and size data for trout and salamanders were collected from two types of forest areas: clear-cut and old growth. Data has been collected yearly 1987-2019. This data was collected through an _____.

   - experiment
   - observational study

2) Collecting data on happiness and money by giving a five-question survey to individuals in a downtown area would be collecting data through an _____.

   - experiment
   - observational study

3) Data for a project come from 30 randomly selected apple trees for which each tree was assigned one time to be pruned: early-, mid-, or late-season. Then the final apple yield for each tree was recorded. The data were collected from an experiment because the _____ .

   - 30 apple trees were randomly selected
   - pruning times were assigned to trees
   - final yield was collected for each pruned tree

4) A customer relations director needs to know which of three email messaging strategies the company currently uses causes the highest customer satisfaction score. Which data collection strategy should be used to best meet the director's needs?

- Randomly assign one of the three messaging strategies to a sample of current customers and then collect customer satisfaction data.
- Collect a random sample of prior customer satisfaction data from customers who received each of the three messaging strategies.

Random sampling in Python

The `pandas` method `DataFrame.sample(n=None, frac=None, replace=False, random_state=None)` returns a random sample of items from a dataframe. The parameters `n` or `frac` specify the number, or fraction, of items to be returned in the sample. The `replace` parameter specifies whether sampling is done with (`True`) or without (`False`) replacement. The `random_state` parameter optionally sets the random number generator seed for reproducible sampling. The rest of the parameters can be found in the documentation for sample.

Random sampling in Python.

Full screen

The dataset below lists the elements present in each painting for a selection of Bob Ross paintings. The code randomly samples from the dataset.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Note the name of the first painting in the sample, then run the sampling code cell again and notice a different sample was returned.
- Modify the code to return a sample of 10 paintings, sampling with replacement.

In [1]:
```
# Import pandas package
import pandas as pd
```

In [2]:
```
# Load the dataset
# Dataset contains 403 paintings (rows) and 69 features (columns)
paintings = pd.read_csv('bob_ross.csv')

# View dataframe
# Dataframes with more than 60 rows only show the first 5 and last 5
paintings
```


Out[2]:

	EPISODE	TITLE	APPLE_FRAME	AURORA_BOREALIS	BARN	BEACH	BOA
0	S01E01	"A WALK IN THE WOODS"	0	0	0	0	
1	S01E02	"MT. MCKINLEY"	0	0	0	0	
2	S01E03	"EBONY SUNSET"	0	0	0	0	
3	S01E04	"WINTER MIST"	0	0	0	0	
4	S01E05	"QUIET STREAM"	0	0	0	0	
...	...	...	...	...	...	...	...
398	S31E09	"EVERGREEN VALLEY"	0	0	0	0	
399	S31E10	"BALMY	0	0	0	1	

PARTICIPATION ACTIVITY 3.1.9: Random sampling in Python.

Consider the example above and the documentation for the `sample` method.

1) Select the code that will return a sample of 50 paintings in which the same painting could appear more than once in the sample.

- `paintings.sample(frac=0.50, replace=True)`
- `paintings.sample(n=50, replace=True)`
- `paintings.sample(n=50, replace=False)`

2) `paintings.sample(frac=0.10, replace=False)` will return a sample of _____ paintings.

- 0.10
- 10
- 40

3) Assuming no other changes are made to `paintings.sample(n=20, replace=False)`, specifying _____ will return the same exact sample of paintings every time the code is run.

- `n=1`
- `random_state=7`
- `replace=True`

CHALLENGE ACTIVITY 3.1.1: Data collection.

537150.4174434.qx3zqy7

Start

Data scientists working for an online retailer are interested in the version of the website resul purchases. 10,000 unique visits to the website are tracked over a single week. Each time a c website, one of two versions are loaded: Version A or Version B. The data scientists track the versions resulting in a purchase.

Identify each of the following.

Observational unit:

Pick

Sample size:

Ex: 1000

Population:

Pick

1 2 3

Check Next

3.2 Descriptive statistics

Learning goals

- Define descriptive statistics, including measures of center, spread, position, and shape.
- Explain which descriptive statistics are influenced by extreme values, and which are not.
- Use a histogram to identify the shape of a distribution.
- Explain how a distribution's shape relates to descriptive statistics.
- Use Python to calculate descriptive statistics for features in a dataset.

Descriptive statistics

A ***feature***, or variable, is a characteristic that can be measured or observed on an observational unit. ***Descriptive statistics*** are methods to summarize and describe a feature's important characteristics.

Ex: The median price for homes in an area is a meaningful way to numerically summarize the price feature from data about homes.

An effective way to begin understanding a feature is to examine the feature's distribution. The ***distribution*** of a feature is the possible values the feature can take on and a measure of how often each value occurs. Visualizing a feature's distribution with a graph gives insights into the distribution's shape. For numerical features:

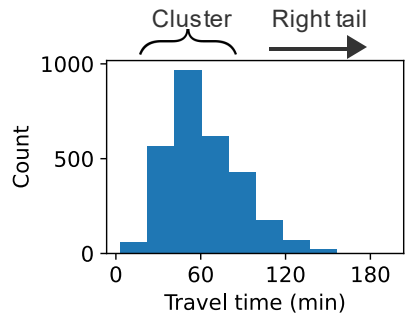
- A ***cluster*** is a distinct group of neighboring values in a distribution that occur noticeably more often than the values on either side of the group.
- The ***tails*** of a distribution are the end values of the distribution. The left tail refers to the lowest values of the distribution, and the right tail refers to the highest values of the distribution.

PARTICIPATION ACTIVITY 3.2.1: Summarizing Australian train passenger survey data.

Dataframe

	ID	License	Cost	TotalTime	Satisfaction
0	126	0	5.0	95	1
1	131	1	3.6	60	3
2	142	1	2.5	80	4
3	145	1	5.5	43	3
4	168	1	2.3	33	3
...	...	...	...	...	...
2922	22598	1	0.0	68	4
2923	22620	1	0.0	50	4
2924	22642	1	3.4	26	2
2925	23003	1	1.2	35	4
2926	23014	1	0.0	21	4

Visual summary



Numerical summary

count	2927.00
mean	62.60
std	25.83
min	3.00
25%	44.50
50%	59.00
75%	78.00
max	194.00

Animation content:

Animation shows an example of how descriptive statistics can be used to summarize a feature's distribution.

- Step 1: A dataframe appears with several rows of responses (instances) and column headings ID, license, cost, total time, and satisfaction.

- Step 2: The total time column in the dataframe is highlighted and a histogram of the total time feature appears.

- Step 3: The following numerical summaries appear: count 2927, mean 62.60, std 25.83, min 3, 25% 44.5, 50% 59, 75% 78, and max 194. The mean and std are highlighted.

Animation captions:

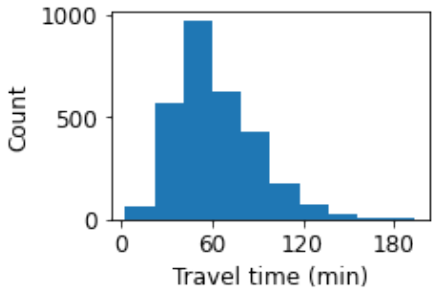
1. A dataset consists of survey responses from train passengers in Australia. Extracting meaningful information can be overwhelming and hard when viewing all 2,927 responses.
2. A histogram of the total travel time feature helps visualize the feature's distribution and shape. The distribution has one cluster and a longer right tail.
3. Descriptive statistics for the total travel time feature numerically summarize and describe the feature's distribution. The average is 62.6 minutes with a standard deviation of 25.83 minutes.

Data source: User satisfaction with train fares: A comparative analysis in five Australian cities.[1]

PARTICIPATION ACTIVITY 3.2.2: Describing the total travel time feature.

Consider the descriptive statistics and histogram below for the total travel time feature from the animation.

Count	Mean	Std	Min	25%	50%	75%	Max
2927	62.60	25.83	3.00	44.50	59.00	78.00	194.00



1) The total travel time feature's distribution takes on values between _____ minutes.

- 0 and 180
- 4 and 969
- 3 and 194

2) The total travel time feature's distribution has one cluster of data from about _____ minutes.

- 3 to 194
- 22 to 98
- 120 to 180

3) Would a time of 180 minutes (3 hours) be considered an unusual response for the total travel time feature?

- Yes, because the distribution of travel time shows few responses around 180 minutes.
- Yes, because 180 is close to the longest reported travel time.
- No, because the distribution of travel time shows many responses around 180 minutes.

Types of features

The two main types of features are categorical and numerical. Descriptive statistics for categorical features include category counts and proportions and are not included in this section. The focus of this section is on descriptive statistics for numerical features.

Measures of center

Two common methods to describe the center of a numerical feature's distribution are the mean and the median. The ***mean***, or average, of a numerical feature is the sum of all values divided by the total number of values. The ***median*** of a numerical feature is the middle value of the ordered data.

PARTICIPATION ACTIVITY 3.2.3: Comparing measures of center.

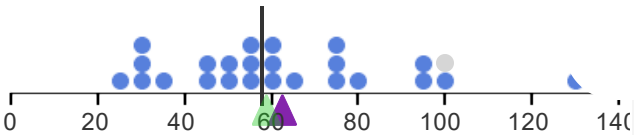
Dataframe

No driver's license (License = 0) and Extremely dissatisfied (Satisfaction = 0)

	ID	License	Cost	TotalTime	Satisfaction
99	2446	0	17.00	74	0
247	7034	0	6.60	93	0
291	7607	0	2.50	65	0
300	7748	0	0.00	28	0
337	9215	0	0.25	100	0
...	...	...	...	...	...
1845	1973	0	6.55	73	0
2002	8666	0	2.65	130	0
2089	11205	0	3.00	55	0
2119	13507	0	4.00	50	0
2747	16642	0	3.40	26	0

24 rows × 5 columns

Visual summary



Descriptive statistics

Mean = 61.04
Median = 58

Modified data

Mean = 59.79
Median = 58

Animation content:

Animation compares the mean and the median with an example.

- Step 1: A dataframe appears showing several rows and with column headings ID, license, cost, total time, and satisfaction. The dataframe only includes the 24 responses with no driver's license (License=0) and who were extremely dissatisfied (satisfaction=0). The total time column is highlighted and a dotplot appears as a visual summary of the total time feature.

- Step 2: A triangle appears on the scale of the dot plot and slides along the scale from 60 to 90 to 40. While the triangle is sliding the dotplot tilts as if the triangle was the balance point. The triangle comes to a rest at the distribution's mean and the dotplot returns to a level position. In a box labeled descriptive statistics the text mean = 61.04 appears.

- Step 3: A box highlights the smallest 12 values and another box highlights the largest 12 values so that all values are highlighted. A line is placed between the two boxes to represent the median. In the descriptive statistics box the text median = 58 appears.

- Step 4: The data is modified, the largest response on the dotplot at 130 moves to a value of 100. Step 5: The mean triangle on the dotplot moves to the left, almost to the median line, and

a box within the descriptive statistics labeled modified data appears and includes the text mean=59.79 and median=58.

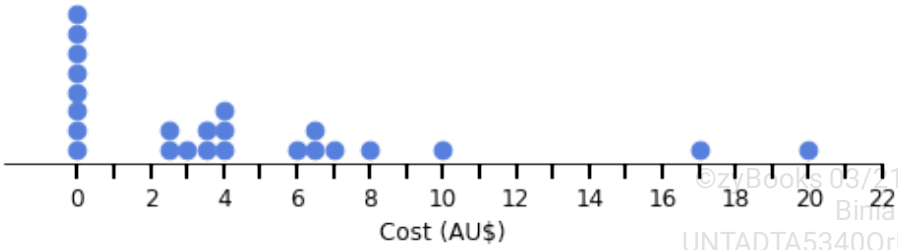
Animation captions:

1. 24 respondents did not have a driver's license and were "extremely dissatisfied" with the tr fare. A dot plot of the total travel time feature visualizes the distribution for this subset of responses.
2. The mean of a distribution can be thought of as a balance point. The average reported total travel time for this subset of responses is 61.04 minutes.
3. The median of a distribution splits the ordered data in half with the same number of responses greater than and less than the median. The median reported total travel time for this subset of responses is 58 minutes.
4. To show the effect of an extreme value, the response with a total travel time of 130 minutes is changed to 100 minutes.
5. For the modified data, the distribution's mean decreases to 59.79 minutes, but the median is still 58 minutes. The median is not influenced by extreme values, but the mean is influenced by extreme values.

PARTICIPATION ACTIVITY 3.2.4: Measures of center.

Consider the numerical summaries and dot plot below for the train fare cost for the subset of 24 respondents who did not have a driver's license and were "extremely dissatisfied" with the fare.

Count	Min	Max	Sum
24	0.00	20.00	108.95



1) The mean train fare cost is _____.

- $0
- $4.54
- $10

2) The median train fare cost is _____.

- $3.50
- $4.54
- $10

3) Suppose the response of $8 was instead $18. For the modified data, the mean would be _____ $4.54.

- less than
- equal to
- greater than

4) Suppose the response of $8 was instead $18. For the modified data, the median would be _____ $3.50.

- less than
- equal to
- greater than

Measures of spread

Measures of spread quantify how clustered or spread out the values of a numerical feature are in the distribution. Typically, small values for spread, near 0, indicate little to no variation in the feature's values, and large values for spread indicate large variation in the feature's values. Common measures of spread include the following:

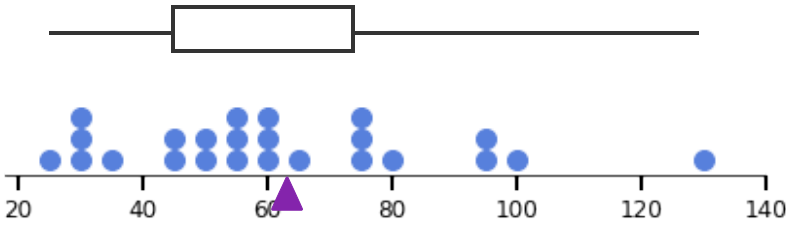
- The ***range*** is the distance from the minimum value to the maximum value of a numerical feature.
- The ***interquartile range*** (IQR) is the range of the middle 50% of the distribution of a numerical feature.
- The ***variance*** is the average squared distance a numerical feature's values lie from the distribution's mean. The equation for the variance of a sample is: $s^2 = \frac{\sum_{i=1}^{n} (\text{value}_i - \text{mean})^2}{n-1}$.
- The ***standard deviation*** is the square root of the variance and describes how far a numerical feature's values lie, on average, from the distribution's mean.

Calculating variance

The equation to calculate the variance of a sample is the sum of the squared distances from the mean divided by $n - 1$. Dividing by $n - 1$, rather than dividing by n, gives a

better estimate of the population variance from sample data. When calculating the variance of population data, the sum of the squared distances from the mean is divided by n. However, data science projects primarily use sample data, so the sample variance equation will almost always be used.

PARTICIPATION ACTIVITY 3.2.5: Measures of spread.



Travel time (min.)

Measures of spread

Range: $130 - 26 = 104$ minutes

IQR: $74.25 - 45 = 29.25$ minutes

Variance: $\frac{(-35.04)^2 + (-33.04)^2 + \cdots + (38.96)^2 + (68.96)^2}{24 - 1} \approx 669.87$

Standard deviation: $\sqrt{669.87} \approx 25.88$ minutes

Animation content:

Animation compares different measures of spread with an example.

- Step 1: A dotplot of the total travel time feature appears along with a box labeled measures of spread with the names of the measures listed: range, IQR, variance, and standard deviation.

- Step 2: The smallest and the largest values on the dotplot are highlighted and then a line is drawn above the dotplot from the largest to the smallest value to represent the range. The range calculation, 130-26=104, is shown.

- Step 3: A box appears above the dotplot that captures the middle 50% of the dots to repre the IQR. The IQR calculation, 74.5-45=29.25, is shown.

- Step 4: A triangle appears on the dotplot at the distribution's mean of 61.04. The smallest two values and the largest two values are highlighted and arrows drawn from the values to the mean as examples of finding how far values are from the mean. The numbers -35.04, -33.04, 38.96 and 68.96 are added to the equation from these four values to show the variance calculation involves finding the distance values lie from the mean. The equation for variance is shown as $\frac{(-35.04)^2+(-33.04)^2+...+(38.96)^2+(68.96)^2}{24-1} \approx 669.87$.

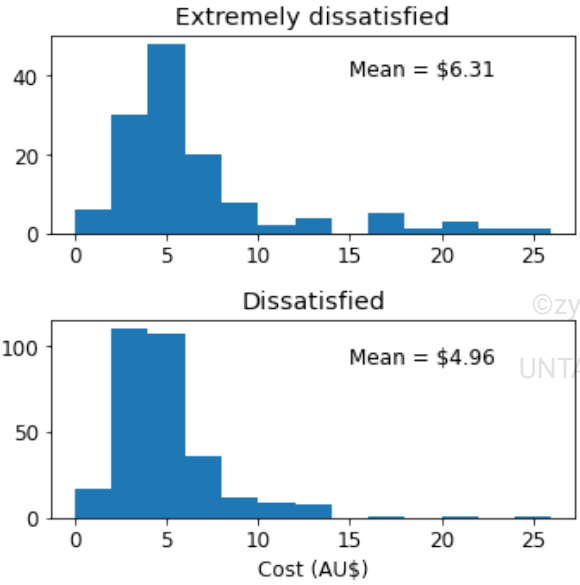
- Step 5: The standard deviation calculation, $\sqrt{(669.87)} \approx 25.88$, is shown.

Animation captions:

1. Measures of spread can be found to numerically describe the variation in total travel time for the 24 respondents who were "extremely dissatisfied" and did not have a driver's license.
2. For these respondents, the shortest travel time is 26 minutes and the longest is 130 minutes, giving a range of 104 minutes.
3. The middle 50% of responses range from 45 to 74.25 minutes. Thus, the IQR is 29.25 minutes indicating responses in the middle of the distribution all lie within about half an hour of each other.
4. Finding the average squared distance of responses from the distribution's mean gives a sample variance of 669.87.
5. The standard deviation is the square root of the variance. On average, responses in the distribution of travel time lie about $\sqrt{669.87} \approx 25.88$ minutes from the mean.

PARTICIPATION ACTIVITY 3.2.6: Measures of spread.

Consider the two histograms below of the distributions of train fare cost for respondents who paid a fare more than $0 and reported they were either "extremely dissatisfied" or "dissatisfied" with fare cost.



1) The range of the cost feature for respondents who were extremely dissatisfied is _____ the range of the cost feature for respondents who were dissatisfied.

- less than
- about equal to
- greater than

2) The variance of the cost feature for respondents who were extremely dissatisfied is _____ the variance of the cost feature for respondents who were dissatisfied.

- less than
- about equal to
- greater than

3) The measures of spread (range, IQR, variance, and standard deviation) are *never* _____.

- negative
- equal to 0
- positive

4) Suppose a response of $50 is reported instead of $25 in the extremely dissatisfied distribution. This change would _____ the distribution's IQR.

- decrease
- not affect
- increase

5) Suppose a response of $50 is reported instead of $25 in the extremely dissatisfied distribution. This change would _____ the distribution's variance.

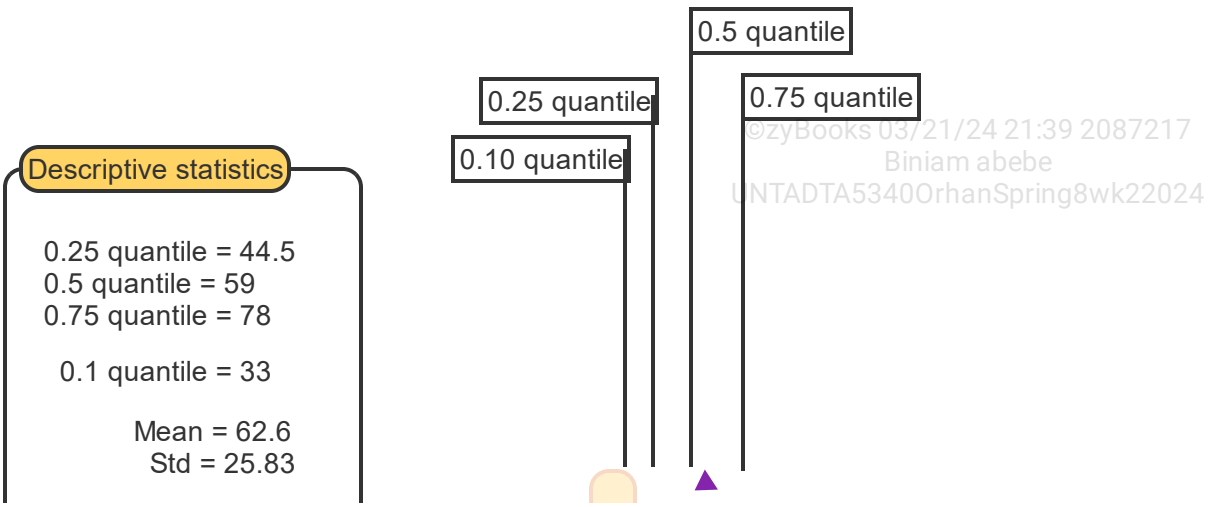
- decrease
- not affect
- increase

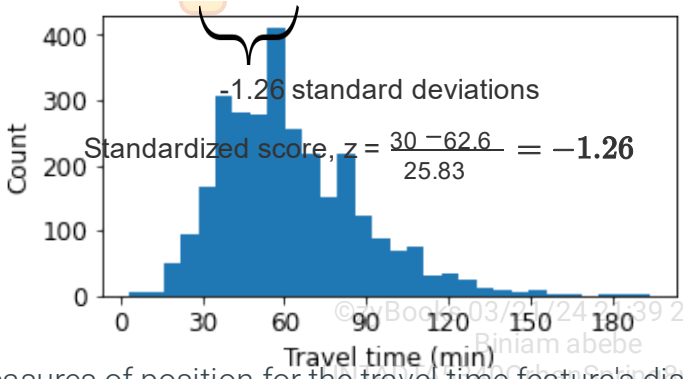
Measures of position

Measures of position indicate where a value is located in a numerical feature's distribution. A ***quantile*** is a value for which a specified proportion of the distribution falls at or below the value. Ex: The median is the 0.5 quantile since 0.5, or half, of the distribution falls at or below the median.

Alternatively, a value's location in the distribution can also be measured relative to the mean. A ***standardized score***, or z-score, describes how many standard deviations—and in which direction—a value lies from the distribution's mean. The equation to find a value's standardized score is $z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$.

PARTICIPATION ACTIVITY 3.2.7: Measures of position.





Animation content:

Animation calculates and compares measures of position for the travel time feature's distribution.

- Step 1: The histogram of total travel time appears with the 0.25, 0.5 and 0.75 quantiles labeled on the histogram. Under Descriptive statistics, the corresponding quantile values are given: 0.25 quantile = 44.5, 0.5 quantile = 59, and 0.75 quantile = 78.
- Step2: A box highlighting the distribution from the minimum to the 0.25 quantile is shown, then the box grows to highlight 50% of the distribution to the 0.5 quantile, the box grows again to highlight 75% of the distribution to the 0.75 quantile.
- Step 3: The value of 30 is highlighted on the x-axis, a line at 33 appears and is labeled 0.1 quantile. The 0.1 quantile = 33 is added under descriptive statistics.
- Step 4: The mean=62.6 and std=25.83 are added under descriptive statistics. A calculation for the standardized score, $z = \frac{30-62.6}{25.83} = -1.26$ appears.
- Step 5: The distance from the mean at 62.6 to the value of 30 is marked on the histogram and labeled as -1.26 standard deviations.

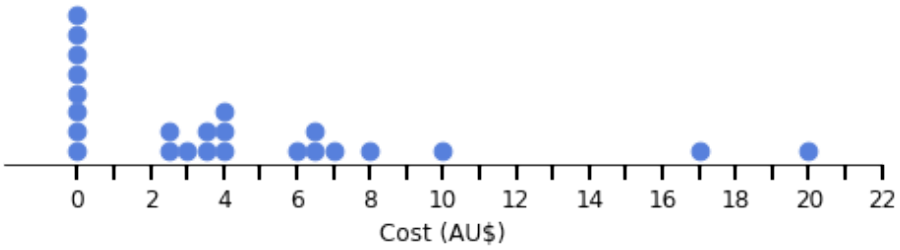
Animation captions:

1. The 0.25, 0.5, and 0.75 quantiles, or quartiles, divide a distribution into four parts, each with the same number of responses.
2. The shortest 25% of travel times are at most 44.5 minutes, half of the reported travel times are 59 minutes or less, and 75% of travel times are at most 78 minutes.
3. Other quantiles can also be meaningful. Ex: Is 30 minutes a short travel time? The 0.10 quantile is 33 minutes, so 30 minutes would be in the shortest 10% of travel times.
4. The location of a 30-minute travel time in the distribution can also be described by the value's standardized score, $z = \frac{30-62.6}{25.83} = -1.26$.
5. Because the standardized score of -1.26 is negative, a 30-minute travel time is 1.26 standard deviations below the distribution's mean.

PARTICIPATION ACTIVITY 3.2.8: Measures of position.

Consider the numerical summaries and dotplot below for the train fare cost for the subset of 24 respondents who did not have a driver's license and were "extremely dissatisfied" with the fare. Responses are rounded to the nearest half dollar in the dot plot.

Count	Mean	Standard deviation
24	4.54	5.24



1) The median cost, or 0.50 quantile, of the distribution is _____.

(#.##)

Check Show answer

2) The first quartile, or 0.25 quantile, of the cost distribution is _____.

(#.##)

Check Show answer

3) The third quartile, or 0.75 quantile, of the cost distribution is _____.

(#.##)

Check Show answer

4) The interquartile range (IQR) of the cost distribution is _____.

(#.##)

Check Show answer

5) The standardized score of the response at $10 is _____.

(#.##)

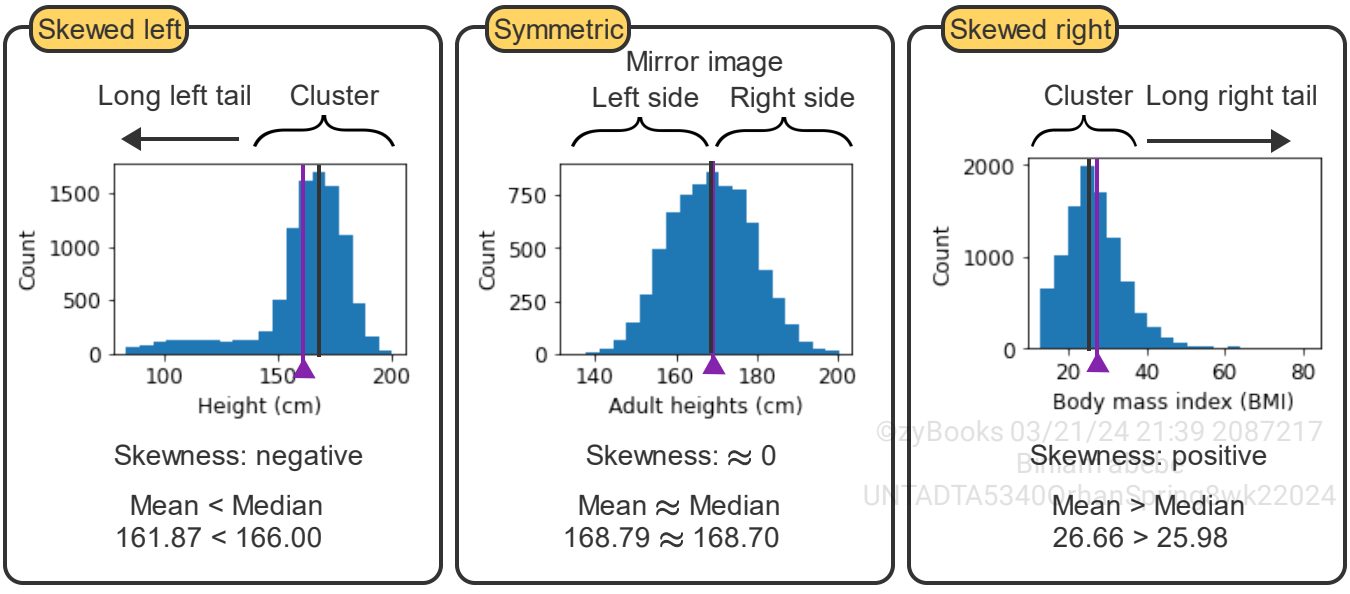
Check Show answer

Measures of shape

A distribution's shape can be described from a visual representation of the data as well as numerically through measures of shape. A distribution is ***symmetric*** when the right side of center and the left side of center are roughly mirror images. A distribution is ***skewed*** if the distribution's values extend farther to one side of the distribution's center. Two common measures of shape are skewness and kurtosis.

- ***Skewness*** is a measure of the amount and direction of skew, or departure from symmetry. Positive values indicate the distribution has an extended right tail and negative values indicate the distribution has an extended left tail.
- ***Kurtosis*** is a measure of tail heaviness. Larger values of kurtosis indicate a greater presence of extreme values in the distribution.

PARTICIPATION ACTIVITY 3.2.9: Comparing distribution shapes in terms of skewness.



Animation content:

Animation compares the shapes of different distributions.

- Step 1: Histogram of height (cm) feature appears.
- Step 2: Histogram of height shows a cluster of data on the right side of the distribution and long left tail. Distribution is labeled as skewed left and that skewness would be negative.
- Step 3: Histogram of adult heights (cm) appears.
- Step 4: Histogram of adult heights shows mirror images for the right and left sides of the distribution's center. Distribution is labeled as symmetric and that skewness would be approximately 0.
- Step 5: Histogram of body mass index (BMI) appears and shows a cluster of data on the left side of the distribution and a long right tail. Distribution is labeled as skewed right and that skewness would be positive
- Step 6: The mean and median are labeled all three distributions. For the symmetric distribution of adult heights, the mean of 168.79 is approximately equal to the median of 168.70. For the skewed left distribution of heights, the mean 161.87 is less than the median 166. For the skewed right distribution of BMI the mean 26.66 is greater than the median 25.98.

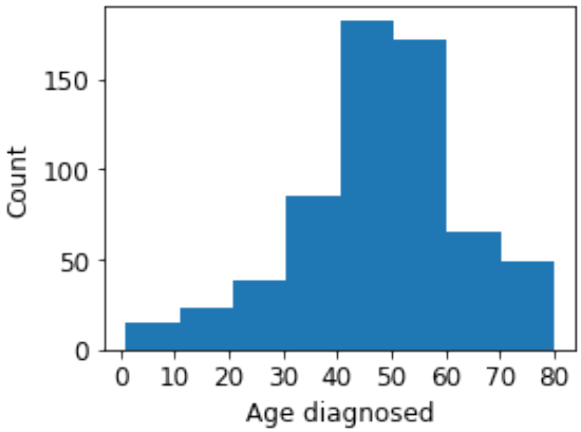
Animation captions:

1. The National Health and Nutrition Examination Survey (NHANES) surveys Americans about their health and nutrition. A histogram visualizes the distribution of heights from the NHANES dataset.
2. The height distribution's shape is skewed left because of the longer left tail and the cluster of data is toward the right tail. A skewed left distribution would have a negative skewness value.
3. Heights in the dataset were recorded on individuals aged 2-80, which explains the skewed left distribution. Subsetting the dataset to individuals aged 20+ only gives the distribution of adult heights.
4. The adult height's distribution is symmetric because the distribution to the left of center and to the right of center are almost mirror images. A symmetric distribution's skewness value will be about 0.
5. Another feature in the dataset is body mass index (BMI). BMI's distribution is skewed right because of the longer right tail and cluster to the left. Skewed right distributions have positive skewness values.
6. In a symmetric distribution, the mean and median are about equal. However, in a skewed distribution, the mean is influenced by the extreme values and pulled in the direction of the longer tail.

Data source: Pruim, Randall. "NHANES: Data from the US National Health and Nutrition Examination Study," 2015.https://CRAN.R-project.org/package=NHANES.

PARTICIPATION ACTIVITY 3.2.10: Measures of shape.

Consider the histogram of the age first diagnosed with diabetes feature from the NHANES dataset.



1) The distribution of age first diagnosed with diabetes is _____.

- skewed left
- symmetric
- skewed right

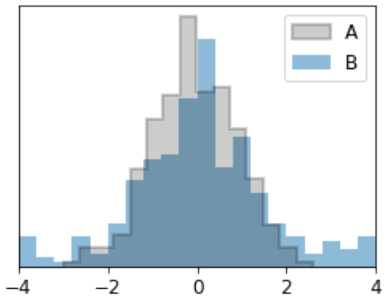
2) The age-diagnosed distribution's skewness value is _____.

- positive
- about 0
- negative

3) The age-diagnosed distribution's mean is expected to be _____ the distribution's median.

- about equal to
- greater than
- less than

4) The graph below visualizes the distributions of two features, labeled A and B, with overlaid histograms. Is distribution A's value of kurtosis less than distribution B's value of kurtosis?



- Yes, because distribution A has fewer values in the tails than distribution B.
- Yes, because distribution A has more values around the distribution's center than distribution B.
- No, because both distribution A and distribution B are symmetric.

Descriptive statistics in Python

The `pandas` DataFrame class has methods that compute descriptive statistics. Methods to compute the descriptive statistics discussed in this section are described in the table below. Additional methods and further information about parameters can be found in the DataFrame documentation.

Table 3.2.1: Pandas descriptive statistics methods.

Method	Parameters	Description
`DataFrame.mean()` `DataFrame.median()`	`axis=None` `skipna=True`	Returns the mean or median of the values over the requested axis. `skipna=True` excludes NA/null values.
`DataFrame.var()` `DataFrame.std()`	`axis=None` `skipna=True`	Returns the unbiased sample variance

	`ddof=1`	(divides by $n-1$) or standard deviation of the values over the requested axis. The divisor used is $n-$`ddof`, where n represents the number of non-NA/null values.
`DataFrame.min()` `DataFrame.max()`	`axis=None` `skipna=True`	Returns the minimum or maximum of the values over the requested axis.

`DataFrame.quantile()`	`q=0.5`
	`axis=None`
	`interpolation='linear'`
		Returns the value of the given quantile(s), `q`, over the requested axis. `interpolation` specifies the method to determine a quantile when the quantile lies between two values.

`DataFrame.skew()`	`axis=None` `skipna=True`	Returns the skewness of the values over the requested axis.
`DataFrame.kurtosis()`	`axis=None` `skipna=True`	Returns the kurtosis of the values over the requested axis. Computes Fisher's definition of kurtosis where a normal distribution has 0 kurtosis.
`DataFrame.describe()`	`percentiles=None`	Returns descriptive statistics. For numerical features, results include the count, mean, standard deviation, minimum, maximum, 0.25 quantile, 0.50 quantile or median, and 0.75 quantile. The returned percentiles can

		be modified with `percentiles`.

Descriptive statistics in Python.

Full screen

National Health and Nutrition Examination Survey (NHANES) is conducted every year to survey Americans about their health and nutrition. The dataset includes both physical characteristics and behaviors, such as exercise and eating habits. The code below calculates descriptive statistics of features from the NHANES dataset.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code to create a histogram and calculate descriptive statistics for the `Length` feature. The `Length` feature is the length, in centimeters, measured lying down for respondents aged 0-3.

©zyBooks 03/21/24 21:39 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

In [1]:
```
# Import pandas and matplotlib packages
import pandas as pd
import matplotlib.pyplot as plt  # used to create a histogram
```

In [2]:
```
# Load the dataset
nhanes = pd.read_csv('nhanes.csv')

# View dataset (first/last 5 rows and the first/last 10 columns)
nhanes
```


Out[2]:

	ID	SurveyYr	Gender	Age	AgeDecade	AgeMonths	Race1	Race3	Educa
0	51624	2009_10	male	34	30-39	409.0	White	NaN	Sc
1	51624	2009_10	male	34	30-39	409.0	White	NaN	Sc
2	51624	2009_10	male	34	30-39	409.0	White	NaN	Sc
3	51625	2009_10	male	4	0-9	49.0	Other	NaN	
4	51630	2009_10	female	49	40-49	596.0	White	NaN	S Col
...	...	...	...	...	...	...	...	...	
9995	71909	2011_12	male	28	20-29	NaN	Mexican	Mexican	9 - G
9996	71910	2011_12	female	0	0-9	5.0	White	White	
9997	71911	2011_12	male	27	20-29	NaN	Mexican	Mexican	Col

PARTICIPATION ACTIVITY 3.2.11: Descriptive statistics in Python.

Consider the example above and the documentation for the descriptive statistics methods.

1) The count statistic returned by `describe()` calculates the number of non-NA values. What is the number of non-NA values for the `Length` feature?

Check Show answer

2) Type the code to return the same statistic as `nhanes['Length'].quantile(q=0.5)` but using a different pandas method.

```
nhanes['Length'].
```

Check Show answer

3) To calculate the variance of a population, the sum of squared distances from the mean would be divided by n rather than $n - 1$. Insert the correct `ddof` parameter value to return the variance of a population.

```
DataFrame.var(ddof= )
```

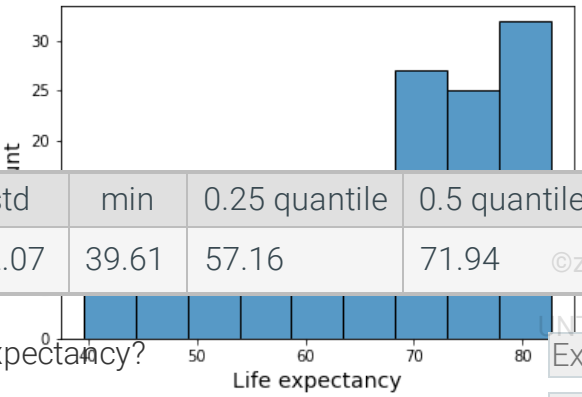
Check Show answer

CHALLENGE ACTIVITY 3.2.1: Descriptive statistics.

537150.4174434.qx3zqy7

Start

The histogram provided shows the distribution of life expectancy for 142 countries in 2007.



count	mean	std	min	0.25 quantile	0.5 quantile	0.75 quantile	max
142	67.01	12.07	39.61	57.16	71.94	76.41	82.6

What is the mean life expectancy? Ex: 5.00

What is the median life expectancy?

What is the shape of the distribution? Pick

1 2 3 4

Check Next

CHALLENGE ACTIVITY 3.2.2: Descriptive statistics using pandas.

537150.4174434.qx3zqy7

Start

This dataset contains a summary of real estate listings for 20 different cities in Texas during June 2

Using a descriptive statistics method, calculate the mean number of homes listed for sale ("lis

The code contains all imports, loads the dataset, and prints the mean.

main.py txhousing.csv

```
# Import packages and functions
import pandas as pd

housing = pd.read_csv('txhousing.csv')

meanHomes = # Your code goes here

print('Mean:', meanHomes)
```

1 2

Check Next level

(*1) Paramita P., Zheng Z., Haque M.M., Washington S., Hyland P. (2018) "User satisfaction with train fares: A comparative analysis in five Australian cities." *PLoS ONE* 13(6): e0199449. https://doi.org/10.1371/journal.pone.0199449

3.3 Probability

Learning goals

- Define probability.
- Express events and probabilities of interest using probability notation.
- List the basic rules of probability.
- Define conditional probability.
- Define Bayes' rule.
- Solve probability problems using the rules of probability.

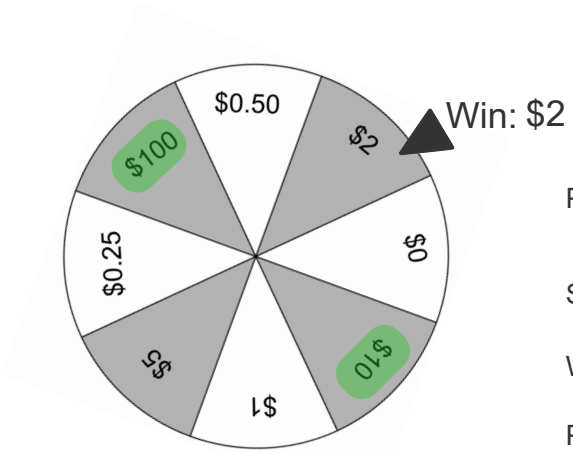
Probability

Probability theory provides a framework to quantify uncertainty. An understanding of probability allows for measures of uncertainty to be constructed, which is essential for drawing statistical inferences about the population beyond the sample data. ***Probability*** is a measure of likelihood

between 0 and 1. Probabilities near 0 correspond to unlikely events and probabilities near 1 correspond to almost certain events. The following terms and notation provide an introduction to probability.

- A ***random process*** is an action or process which results in an outcome determined by chanc
- An ***outcome*** is one possible result from a random process.
- The ***sample space*** is the set of all possible outcomes of a random process and denoted as S.
- An ***event*** is an outcome or collection of outcomes from a sample space. Events are typically denoted by A, B, or C.
- The ***probability of an event A***, denoted $P(A)$, is measured as the number of outcomes in A divided by the total number of equally likely outcomes in the sample space, S. The probability of an event can also be understood as the long-run proportion of times event A occurs if the random process is observed many times.

PARTICIPATION ACTIVITY 3.3.1: Prize wheel probability example.



Possible outcomes: \$0, \$0.25, \$0.50, \$1,, \$2, \$5, \$10, \$100

Sample space: S = {0, 0.25, 0.50, 1, 2, 5, 10, 100}

Win at least \$10: A = {10, 100}

Probability of A: $P(A) = \frac{2}{8} = 0.25$

Animation content:

Animation shows an example of a random process and the key probability terms.

- Step 1: A prize wheel is shown with 8 equally sized spaces, each labeled with a dollar amount. Dollar amounts are \$0, \$0.25, \$0.50, \$1, \$2, \$5, \$10, and \$100.
- Step 2: The 8 possible winning amounts are highlighted on the wheel and the possible outcomes: \$0, \$0.25, \$0.50, \$1, \$2, \$5, \$10, and \$100 are listed. The sample space, S = {0, 0.25, 0.50, 1, 2, 5, 10, 100}, is also shown.

- Step 3: The wheel space labeled \$10 and the wheel space labeled \$100 are highlighted and the text, win at least \$10: A={10, 100}, appears.
- Step 4: The probability calculation, probability of A: $P(A) = \frac{2}{8} = 0.25$, is shown.

Animation captions:

1. A prize wheel has eight equally sized spaces. The wheel is spun and the amount won is the amount on the indicated winning space. A spin of the wheel is a random process.
2. The possible prize amounts, or eight possible outcomes, are \$0, \$0.25, 0.50, \$1, \$2, \$5, \$10, and \$100. Thus, the sample space is S = {0, 0.25, 0.50, 1, 2, 5 10, 100}.
3. Define event A as winning at least \$10. Two outcomes from the sample space are in event A, \$10 and \$100.
4. The probability of event A occurring on any spin is $P(A) = \frac{2}{8} = 0.25$, or that if the wheel is spun many, many times, the long-run proportion of times \$10 or \$100 is won is 0.25.

PARTICIPATION ACTIVITY 3.3.2: Probability.

The following table gives the distribution of US household size (number of people in the household) based on the 2020 Current Population Survey.

Size	1	2	3	4	5	6	7+
Proportion	0.29	0.35	0.15	0.12	0.06	0.02	0.01

1) A US household is randomly selected and the household size is recorded. Is randomly selecting a household a random process?

- Yes, because the household selected was obtained by chance.
- No, because the distribution of household size is known.

2) A US household is randomly selected and the household size is recorded. Identify the sample space.

- S = {0.29, 0.35, 0.15, 0.12, 0.06, 0.02, 0.01}
- S = {1, 2, 3, 4, 5, 6, 7+}

3) Define event A as randomly selecting a household with a household size of 2. The long-run proportion of times event A would occur, $P(A)$, is _____.

- 0.35
- 35

4) Define event A as randomly selecting a household with a household size of 2, and define event B as randomly selecting a household with a household size of 1. Event _____ is more likely to occur.

- A
- B

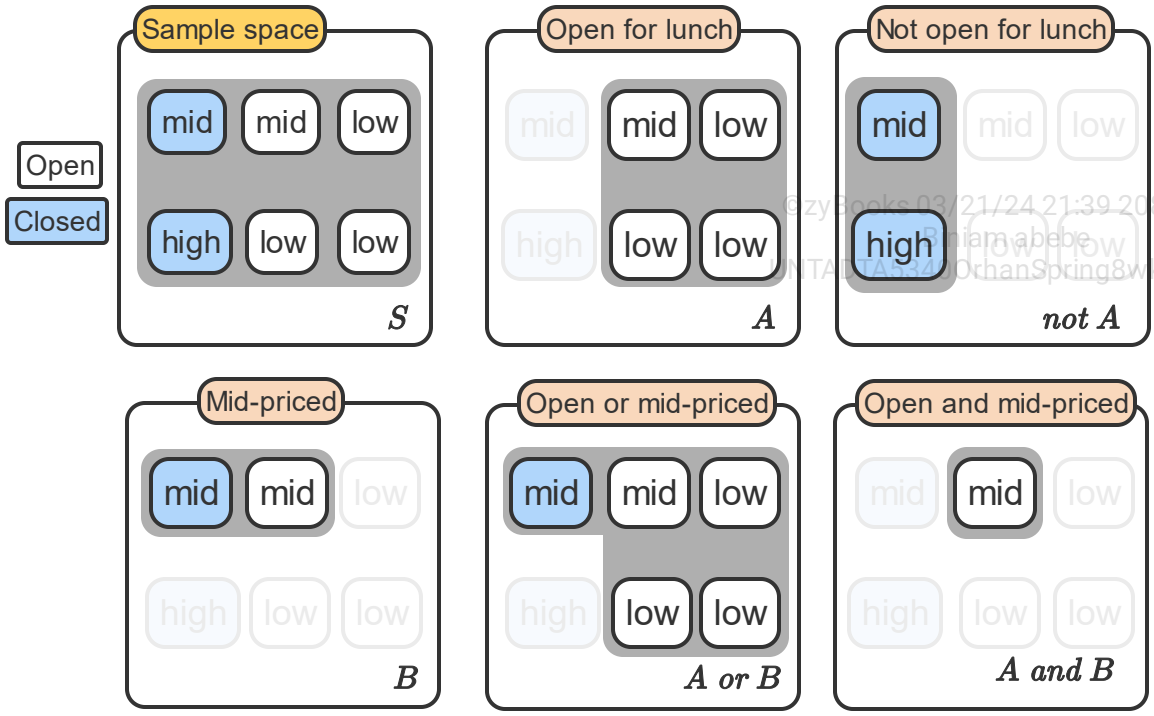
Event operations

Sometimes new events need to be constructed from known events. Ex: The event that a randomly selected household has a size of 4 or more and the household income is $75,000 or more. The following event operations define new events from known events.

- The ***complement*** of event A, denoted $not\ A$, is the event consisting of all outcomes in the sample space S that are not in event A.
- The ***union*** of two events, A and B, denoted $A\ or\ B$, is the event consisting of all outcomes in A or B, including outcomes in both A and B.
- The ***intersection*** of two events, A and B, denoted $A\ and\ B$, is the event consisting of only the outcomes in both A and B.

Two events, A and B, are considered ***disjoint***, or ***mutually exclusive***, if the two events have no outcomes in common. In other words, the event $A\ and\ B$ is empty or contains no outcomes.

PARTICIPATION ACTIVITY 3.3.3: Event operations.



Animation content:

Animation shows events constructed from a sample space using event operations.

- Step 1: The sample space consists of six restaurants. Three restaurants are low-priced, two restaurants are mid-priced, and one restaurant is high-priced. Four restaurants are open for lunch: the three low-priced restaurants and one mid-priced restaurant. Two restaurants are closed for lunch: one mid-priced restaurant and the high-priced restaurant.
- Step 2: Event A, or open for lunch, highlights the four restaurants from the sample space that are open for lunch.
- Step 3: Event $notA$, or not open for lunch, highlights the two restaurants from the sample space that are closed for lunch.
- Step 4: Event B, or mid-priced, highlights the two restaurants from the sample space that are mid-priced.
- Step 5: Event $AorB$, or open for lunch or mid-priced, highlights the four restaurants that are open for lunch, and the two restaurants that are mid-priced, including the one restaurant that is open for lunch and mid-priced. A total of five restaurants from the sample space are

highlighted. Step 6: Event $AandB$, or open for lunch and mid-priced, highlights the one restaurant that is open for lunch and mid-priced, which is in both event A and event B.

Animation captions:

1. One restaurant will be randomly selected from six options. Each restaurant has a price lev (low, mid, and high) and is either open or closed for lunch.
2. Define event A as selecting a restaurant that is open for lunch. Four restaurants, or outcomes of the sample space, are open for lunch and included in A.
3. The complement of event A is selecting a restaurant that is not open for lunch. The event $not\ A$ contains the two outcomes of the sample space which are not in A.
4. Define event B as selecting a mid-priced restaurant. B consists of two outcomes from the sample space.
5. The event $A\ or\ B$ is selecting a restaurant that is either open for lunch or is mid-priced, including the one restaurant that is both open for lunch and mid-priced.
6. The event $A\ and\ B$ is selecting a restaurant that is open for lunch and mid-priced. Because $A\ and\ B$ contains one outcome, events A and B are not disjoint.

PARTICIPATION ACTIVITY 3.3.4: Event operations.

Consider the animation example of randomly selecting a restaurant from the six options listed in the table below.

Restaurant	1	2	3	4	5	6
Price	low	low	low	mid	mid	high
Lunch	open	open	open	open	closed	closed

Define the events:

- A as randomly selecting a restaurant that is open for lunch
- B as randomly selecting a mid-priced restaurant
- C as randomly selecting a high-priced restaurant

1) Event C contains _____ outcome(s).

- ○ 1
- ○ 2
- ○ 4

2) The union of events B and C is randomly selecting a restaurant that is _____.

- either mid-priced or high-priced
- mid-priced and high-priced
- open for lunch or high-priced

3) The complement of the event $B \text{ or } C$ contains _____ .

- no outcomes
- restaurants 1, 2, and 3
- restaurants 4, 5, and 6

4) The event of randomly selecting a restaurant that is open for lunch and high-priced can be expressed as _____.

- $A \text{ and } B$
- $A \text{ or } C$
- $A \text{ and } C$

5) Are events A and C disjoint?

- No, because $A \text{ or } C$ contains five outcomes.
- Yes, because $A \text{ or } C$ contains five outcomes.
- Yes, because $A \text{ and } C$ contains no outcomes.

Probability rules

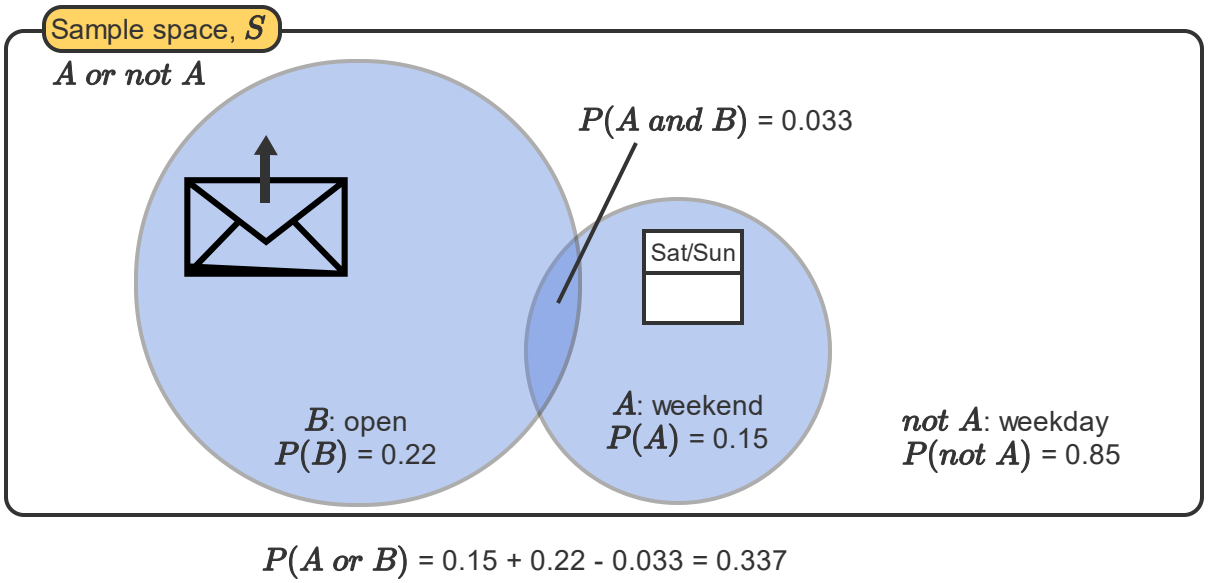
The three axioms, or fundamental properties, of probability are:

- The probability of any event is non-negative, $P(A) \geq 0$.
- The probability of the sample space is $P(S) = 1$.
- If A and B are disjoint events, $P(A \text{ or } B) = P(A) + P(B)$.

Additional probability rules can be derived from the three axioms.The following rules are useful for finding the probability of a complement of an event, the union of any two events, and the intersection of two independent events.Two events are considered ***independent*** if knowing one event has occurred does not affect the probability of the other event.

- The probability of the complement of event A can be found from the probability of event A, $P(not\ A) = 1 - P(A)$.
- The probability of the union of any two events is $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$.
- The probability of the intersection of independent events is $P(A\ and\ B) = P(A) * P(B)$

PARTICIPATION ACTIVITY 3.3.5: Probability rules.



$P(A\ or\ B)$ = 0.15 + 0.22 - 0.033 = 0.337

Animation content:

Animation shows examples of finding probabilities using the probability rules.

- Step 1: A large box is labeled as the sample space, S. A small circle inside the sample space is labeled event A, weekend, with $P(A) = 0.15$.
- Step 2: The area outside event A circle is highlighted and labeled $notA$, weekday, with $P(notA) = 0.85$. The sample space is also labeled as $A\ or\ notA$.
- Step 3: Another circle appears inside the sample space and labeled event B, open, with $P(B) = 0.33$. The circles for events A and B overlap a small amount.
- Step 5: The area where the circles for events A and B overlap is labeled as $P(AandB) = 0.033$.

- Step 6: The equation for the union of A and B is shown: $P(AorB) = 0.15 + 0.22 - 0.033 = 0.337$.

Animation captions:

1. 15% of marketing emails are sent over the weekend.
2. The probability a marketing email is sent on a weekday is $P(not\ A) = 1 - 0.15 = 0.85$. A and $not\ A$ are disjoint events and their union makes up the sample space, $S = A\ or\ not\ A$
3. Regardless of the day the email is sent, the probability of a recipient opening a marketing email is 0.22.
4. If opening the email is independent of when the email is sent, the probability a marketing email is sent on the weekend and is opened is $P(A\ and\ B) = P(A) * P(B) = 0.15 * 0.22 = 0.033$.
5. The probability a marketing email is sent on the weekend or is opened is $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$ because $P(A) + P(B)$ includes $P(A\ and\ B)$ twice.

PARTICIPATION ACTIVITY 3.3.6: Probability rules.

The following table gives the distribution of US household size (number of people in the household) based on the 2020 Current Population Survey.

Size	1	2	3	4	5	6	7+
Proportion	0.29	0.35	0.15	0.12	0.06	0.02	0.01

Define event A as randomly selecting a household of size 1 and event B as randomly selecting a household of size 5 or more.

1) Find the probability of randomly selecting a household with a size of more than 1, $P(not\ A)$.

(#.##)

Check Show answer

2) Find the probability of randomly selecting a household with a size of 1 or more than 1, $P(A \text{ or not } A)$.

(#.##)

Check Show answer

3) Find the probability of randomly selecting a household with a size of 5 or more, $P(B)$.

(#.##)

Check Show answer

4) Find the probability of randomly selecting a household with size 1 or size of 5 or more, $P(A \text{ or } B)$.

(#.##)

Check Show answer

5) One household will be randomly selected from all households, and then a second household will be randomly selected from all households. Find the probability that both selected households are of size 1.

(#.##)

Check Show answer

Conditional probability

The probability of an event occurring can also be determined under the condition of knowing another event has occurred. Ex: Determining the probability an email is spam given the email contains an HTML link. A ***conditional probability*** is a measure of the likelihood of one event occurring, given

another event has occurred. The conditional probability of event A given event B has occurred, denoted as $P(A|B)$, is $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$.

For independent events, A and B, recall $P(A \text{ and } B) = P(A) * P(B)$. Thus, for independent events, the conditional probabilities are $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

PARTICIPATION ACTIVITY 3.3.7: Conditional probability.

	Not covered	Covered	Row total
Northeast	1,068	16,456	17,524
Midwest	1,460	19,588	21,048
South	4,528	33,141	37,669
West	2, 028	21,731	23,759
Column total	9,084	90,916	100,000

Events

A: covered B: Northeast

Probabilities

$$P(A) = \frac{90916}{100000} \approx 0.909$$

$$P(B) = \frac{17524}{100000} \approx 0.175$$

$$P(A \text{ and } B) = \frac{16456}{100000} \approx 0.165$$

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{0.165}{0.175} \approx 0.943$$

$$P(A|B) = \frac{16456}{17524} \approx 0.939$$

Animation content:

Animation shows example probability calculations, including conditional probability.

- Step 1: A table appears showing the distribution of regions (Northeast, Midwest, South, and West) and whether or not they have health insurance coverage for 100,000 individuals.
- Step 2: A box labeled events appears with event A, Covered, and event B, Northeast.
- Step 3: A box labeled probabilities appears, $P(A) = \frac{90916}{100000} \approx 0.909$ is shown with the values for the total number of individuals covered, 90916, and the total number of individuals in the sample space, 100,000, highlighted in the table. $P(B) = \frac{17524}{100000} \approx 0.175$ is also shown with the values for the total number of individuals in the Northeast, 17524, and the total number of individuals in the sample space, 100,000, highlighted in the table.
- Step 4: $P(AandB) = \frac{16456}{100000} \approx 0.165$ is shown with the values for the total number of individuals covered and in the Northeast, 16,456, and the total number of individuals in the

sample space, 100,000, highlighted in the table.

- Step 5: The equation $P(A|B) = \frac{P(AandB)}{P(B)} = \frac{0.165}{0.175} \approx 0.943$ is shown.
- Step 6:The highlight that was on the total number of individuals in the sample space, 100 moves to highlight the total number of individuals in the Northeast, 17524. The total number of individuals covered and in the Northeast, 16456, is also highlighted and the equation $P(A|B) = \frac{16456}{17524} \approx 0.939$ is shown.
- Step 7: Because $P(A) \approx 0.909$ and $P(A|B) \approx 0.94$, knowing the selected individual lives in the Northeast slightly increases the probability the individual has health insurance

.

Animation captions:

1. From a database of 100,000 adults, the region where the individual lives, and whether or not the individual has health insurance coverage, is summarized in a table.
2. An individual will be randomly selected from the database. Define A as selecting an individual with health insurance coverage, and define B as selecting an individual who lives in the Northeast.
3. The probability of selecting an individual who is covered is $P(A) \approx 0.909$, because 90,916 of the 100,000 individuals are covered. Similarly, $P(B) \approx 0.175$
4. $P(A \text{ and } B)$ can be found using the table without assuming events A and B are independent. The probability of selecting an individual who lives in the Northeast and with coverage is 0.165.
5. The probability of selecting an individual with coverage, given the individual lives in the Northeast, can be found using the conditional probability equation $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$.
6. The conditional probability $P(A|B)$ can also be thought of as restricting the sample space from S to only B. The difference between the two values for $P(A|B)$, 0.943 and 0.939, is due to rounding.
7. Because $P(A) \approx 0.909 \neq P(A|B) \approx 0.94$, A and B are not independent. Knowing the selected individual lives in the Northeast slightly increases the probability the individual has health insurance.

PARTICIPATION ACTIVITY 3.3.8: Conditional probability.

From a database of 100,000 adults, the distribution of the region where the individual lives and whether or not the individual has health insurance coverage is given in the following table.

	Not covered	Covered	Row total
Northeast	1,068	16,456	17,524
Midwest	1,460	19,588	21,048
South	4,528	33,141	37,669
West	2, 028	21,731	23,759
Column total	9,084	90,916	100,000

An individual is randomly selected from the database. Define event A as randomly selecting an individual from the South region and define event B as randomly selecting an individual with no health insurance coverage. Find the following probabilities.

1) $P(B)$

(#.###)

Check Show answer

2) $P(A \text{ and } B)$

(#.###)

Check Show answer

3) $P(A|B)$

(#.###)

Check Show answer

4) $P(B|A)$

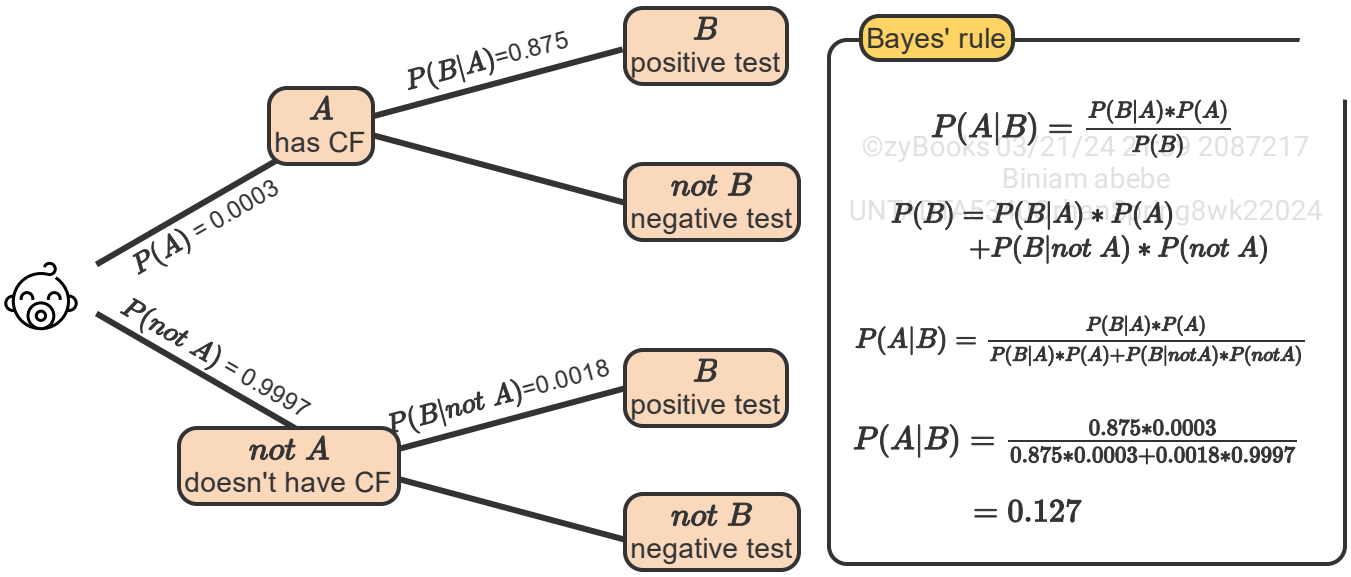
(#.###)

Check Show answer

Bayes' Rule

Sometimes the information needed to find the desired conditional probability, such as $P(A \text{ and } B)$, is not known. Bayes' rule provides an alternative way to find $P(A|B)$ from $P(B|A)$. ***Bayes' rule*** for finding the conditional probability of A given B is $P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$.

PARTICIPATION ACTIVITY 3.3.9: Bayes' rule.



Animation content:

Animation shows an example of when Bayes' rule is useful in calculation probabilities.

- Step 1: Event A, has CF, with $P(A) = 0.0003$ and Event $notA$, doesn't have CF, with $P(notA) = 0.9997$ appears.
- Step 2: Branching from event A, an event B, positive test, appears with P(B|A)=0.875.
- Step 3: Branching from event $notA$, another event B, positive test, appears with $P(B|notA) = 0.0018$.
- Step 4: A box labeled Bayes' rule appears with the equation \(P(A|B)=\frac{P(B|A)*P(A)}{P(B)}.
- Step 5: The equation $P(B) = P(B|A) * P(A) + P(B|notA) * P(notA)$ appears followed by an equation substituting this result in for the denominator from the first equation, \(P(A|B)=\frac{P(B|A)*P(A)}{P(B|A)*P(A)+P(B|not A)*P(not A)}.
- Step 6: The equation to calculation $P(A|B)$ is filled in with the known numbers: $P(A|B) = \frac{0.875*0.0003}{0.875*0.003+0.0018*0.9997} = 0.127$.

Animation captions:

1. A common application of Bayes' rule is in diagnostic testing. In a certain population, 1 in 3,000 newborns are born with cystic fibrosis (CF), a genetic disorder that causes problems with breathing and digestion.
2. An initial diagnostic screening test screens all newborns. For newborns with CF, a positive test result will occur 87.5% of the time.
3. For newborns without CF, the diagnostic test incorrectly shows a positive result 0.18% of ti time.
4. Suppose a test came back positive. What is the probability the newborn actually has CF? Bayes' rule provides a way to find $P(A|B)$ based on the known information.
5. The probability of a positive test result can be found from the conditional probabilities of a positive test given the newborn has or doesn't have CF times the probabilities of each event.
6. The probability a newborn with a positive screening test actually has CF is 0.127.

PARTICIPATION ACTIVITY 3.3.10: Bayes' rule.

The sales department for an online product reported the following information about potential customer contacts and purchases:

- 5% of all contacts end up purchasing the product.
- 10% of all contacts attend a demo webinar.
- 75% of the contacts who purchased the product attended a demo webinar.

Potential customers are randomly contacted. Define the events:

- A as a contact who ends up purchasing the product.
- B as a contact who attends a demo webinar.

1) $P(A)$ is ____.

- 0.05
- 0.10
- 5

2) 0.75 is ____ .

- $P(A \text{ and } B)$
- $P(A|B)$
- $P(B|A)$

3) Primary interest is in the probability that a randomly selected contact who agrees to attend the webinar purchases the product, or $P(A|B)$. Identify the correct equation to find $P(A|B)$.

- $\frac{0.05*0.10}{0.10} = 0.05$
- $\frac{0.75*0.05}{0.10} = 0.375$
- $\frac{0.75}{0.10} = 7.5$

4) Events A and B are not independent because _____.

- $P(A|B) = 0.375$ is not equal to $P(A) = 0.05$
- $P(A|B) = 0.375$ is not equal to $P(B|A) = 0.75$
- $P(A) = 0.05$ is not equal to $P(B) = 0.10$.

CHALLENGE ACTIVITY 3.3.1: Probability.

537150.4174434.qx3zqy7

Start

A clothing retailer uses a rewards program to track and reward customer spending. The reta estimates that 62% of rewards members have made an online purchase in the past year. 35 rewards members have made an in-store purchase, and 27% of rewards members have purc both online and in-store.

Define A as a rewards member making an online purchase and B as a rewards member mak in-store purchase.

What is $P(A)$? Ex: 0.02

What is $P(B)$?

What is $P(A \text{ and } B)$?

1 2 3

Check Next

3.4 Probability distributions

Learning goals

- Define a random variable.
- Categorize a random variable as discrete or continuous.
- Relate a probability distribution to a random variable.
- Explain characteristics of a Bernoulli(π) or a binomial(n, π) distribution.
- Explain characteristics of a normal(μ, σ) distribution.
- Compare a t-distribution is to the standard normal distribution.
- Use Python to graph probability distributions.
- Use Python to calculate probabilities for Bernoulli, binomial, normal, and t-distributions.

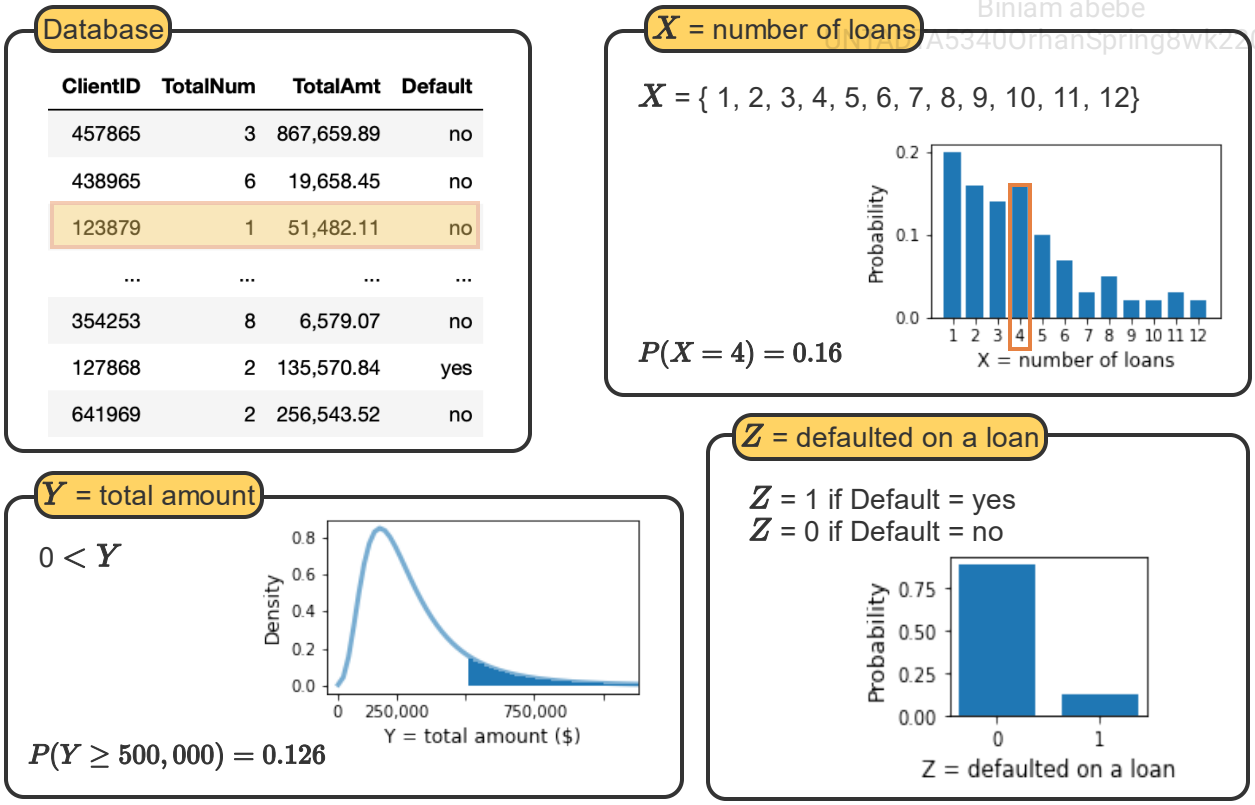
Random variables

A ***random variable*** defines numerical values for a random process's outcomes. Random variables are typically denoted by X, Y, or Z. Ex: Let X be the number of website visitors on a randomly selected day. A ***discrete random variable*** takes on a countable number of distinct values. Ex: Household size of a randomly selected household takes on values such as 1, 2, 4, or 9. A ***continuous random variable*** takes on all values within an interval. Ex: The body mass index of a randomly selected healthy adult typically takes on values between 18 and 25 such as 19.02 or 22.86.

A random variable's ***probability distribution*** gives the probability of occurrence for the random variable's possible numerical values. A probability distribution, like a feature's distribution, can be

visualized with graphs and described with numerical summaries. The statistical inference methods and modeling often rely on probability distributions.

PARTICIPATION ACTIVITY 3.4.1: Probability distributions for random variables.



Animation content:

Animation shows examples of random variables and their probability distributions.

- Step 1: A few entries of the bank's database of loan clients are shown.
- Step 2: Box with information for the random variable X= total # of loans is shown. Includes X = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}.
- Step 3: Graph of the probability distribution for X appears with the values of X on the horizontal axis and probabilities on the vertical axis; probabilities range from 0.02 to 0.20. The bar for X=4 is highlighted and has height 0.16 to show $P(X = 4)$=0.16.
- Step 4: Box with information for the random variable Y=total loan amount is shown. Includes the possible values of Y as (\0\lt Y\). Graph of the probability distribution for Y is shown with values of Y on the horizontal axis and density on the vertical axis. Distribution is shown by a

single cluster curve with the peak just below $250,000 and values from about 0 to just over $1 million. Area under the curve and above horizontal axis is shaded from $500,000 to maximum value to show $P(Y > 500,000)$=0.126.

- Step 5: Box with information for the random variable Z = defaulted on a loan is shown. Includes defining Z=0 if Default=no and Z=1 if Default=yes. Probability distribution for Z is shown with possible values of 0 and 1 on the horizontal axis and probability on the vertical axis. Z=0 has height 0.88 and Z=1 has height 0.12.

Animation captions:

1. A bank's database contains information on all the bank's loan clients. Consider the random process of randomly selecting a client from the database.
2. Define X as the total number of loans the selected client has had with the bank. All clients have taken out at least one loan and the most is 12 loans. Thus, X is a discrete random variable.
3. A graph of the probability distribution visualizes the probability for each possible value. Ex: The probability of selecting a client who has had four loans with the bank is $P(X = 4) = 0.16$.
4. Define Y as the total loan amount, which is a continuous random variable. The probability distribution is graphed with a density curve. The probability of an interval of values is the area under the curve.
5. Non-numerical outcomes can be described by a numerical random variable of interest. Ex: Let Z indicate whether or not the client has defaulted on a loan.

PARTICIPATION ACTIVITY 3.4.2: Random variables.

1) Consider randomly selecting a vehicle from a state's vehicle registration database. Which of the following defines a random variable?

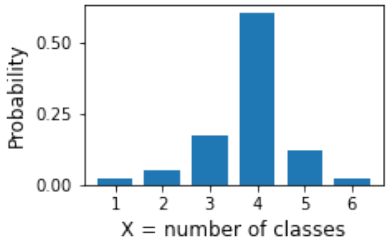
- X = the make of the vehicle. Ex: Toyota
- Y = the weight of the vehicle. Ex: 1.2 tons
- Z = the vehicle identification number. Ex: 2X1TL65348Z419438

2) Let X be the order wait time, such as 382.42 minutes, of a randomly selected drive-through order. X is _____.

- a continuous random variable
- a discrete random variable
- not a random variable

3) Let X be the number of classes a randomly selected college student is currently enrolled in at a particular college. The probability distribution of X for the college is given in the graph and table below. X is _____ .

X	1	2	3	4	5	6
Probability	0.03	0.05	0.17	0.60	0.12	0.02



- a continuous random variable
- a discrete random variable
- not a random variable

4) Consider the random variable and probability distribution of X, the number of classes a randomly selected college student is currently enrolled in, from question 3. The possible values of X are _____ .

- 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- all values between 1 and 6
- 1, 2, 3, 4, 5, 6

5) Consider the random variable and probability distribution of X, the number of classes a randomly selected college student is currently enrolled in, from question 3.The most likely value of X is ____ .

- 4
- 6
- 60

Bernoulli distribution

Certain probability distributions of random variables occur often enough to be given a name, and several named distributions are commonly used in data science. A discrete random variable in which the only possible values are 0 and 1 follows a ***Bernoulli distribution***. Bernoulli distributions typically describe non-numerical outcomes in which a value of 1 indicates a "success" and a value of 0 indicates a "failure". Ex: Let X describe whether a randomly selected student passed a class in which a 1 indicates the student passed and a 0 indicates the student failed.

The ***parameters*** of a probability distribution determine the distribution's shape and probabilities. The Bernoulli distribution is specified by one parameter, π, which is the probability of a "success" or 1.

PARTICIPATION ACTIVITY 3.4.3: Bernoulli distribution.

Database

ClientID	TotalNum	TotalAmt	Default
457865	3	867,659.89	no
438965	6	19,658.45	no
123879	1	51,482.11	no
...	...	...	...
354253	8	6,579.07	no
127868	2	135,570.84	yes
641969	2	256,543.52	no

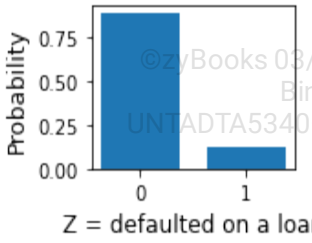
12%

Z = defaulted on a loan

Z = 1 if Default = yes
Z = 0 if Default = no

$Z \sim$ Bernoulli($\pi = 0.12$)

$\mu_Z = 0.12 \quad \sigma_Z^2 \approx 0.11$



$P(Z = 0)$ = 0.88

$P(Z = 1)$ = 0.12

Animation content:

Animation shows an example of a Bernoulli random variable and probability distribution.

- Step 1: A few rows of a bank's database of loan clients are shown with the column for default highlighted. A box appears for the random variable Z = defaulted on a loan. Z= 1 if Defau' yes and Z = 0 if Default = no.
- Step 2: The notation Z~Bernoulli(π) is shown.
- Step 3: In the database 12% of clients have defaulted, the 0.12 moves from the database into the notation Z~Bernoulli($\pi = 0.12$).
- Step 4: A bar graph of the probability distribution for Z is shown, Values of Z, 0 and 1, are on the x-axis and probability is on the y-axis. $P(Z = 0) = 0.88$, so 0.88 is the height of the Z = 0 bar. $P(Z = 1) = 0.12$, so 0.12 is the height of the Z = 0.12 bar.
- Step 5: Notation for the mean, $\mu_Z = 0.12$, and variance, $\sigma_Z^2 \approx 0.11$, is shown.

Animation captions:

1. Let Z be whether or not a randomly selected loan client has ever defaulted on a loan. $Z = 1$ if a client has defaulted on a loan and $Z = 0$ if the client has not defaulted on a loan.
2. Because the random variable Z is selecting one client and has only two possible outcomes, 0 or 1, Z follows a Bernoulli distribution.
3. 12% of all loan clients at the bank have defaulted on a loan, so $P(Z = 1)$ = 0.12 and $Z \sim$ Bernoulli(π = 0.12) distribution.
4. Discrete probability distributions can be visualized with a bar graph. The height of a value's bar is the probability of the value occuring.
5. Probability distributions are typically described numerically with the mean and variance. A Bernoulli(π) distribution has mean $\mu = \pi$ and variance $\sigma^2 = \pi * (1 - \pi)$.

PARTICIPATION ACTIVITY 3.4.4: Bernoulli distribution.

1) A dataframe has 100,000 instances and 20 features. Let X be the number of missing feature values for a randomly selected instance. Would X follow a Bernoulli distribution?
   - ○ No
   - ○ Yes

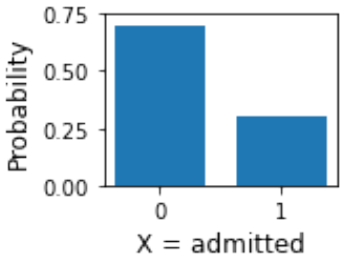
2) Flight delay times are negative when the flight departed early, 0 when the flight departed on time, or positive when the flight departed late. Let X be 1 if the delay time for a randomly selected flight is greater than 0 and X is 0 otherwise. Would X follow a Bernoulli distribution?

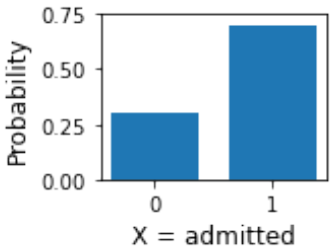
- No
- Yes

3) A graduate program admits 30% of applicants. Let X be 1 if a randomly selected applicant was admitted and 0 if the applicant was not admitted. X would follow a Bernoulli(π) distribution with π = _____ .

- 0.30
- 0.70

4) A graduate program admits 30% of applicants. Let X be 1 if a randomly selected applicant was admitted and 0 if the applicant was not admitted. Select the correct graph of the probability distribution for X.

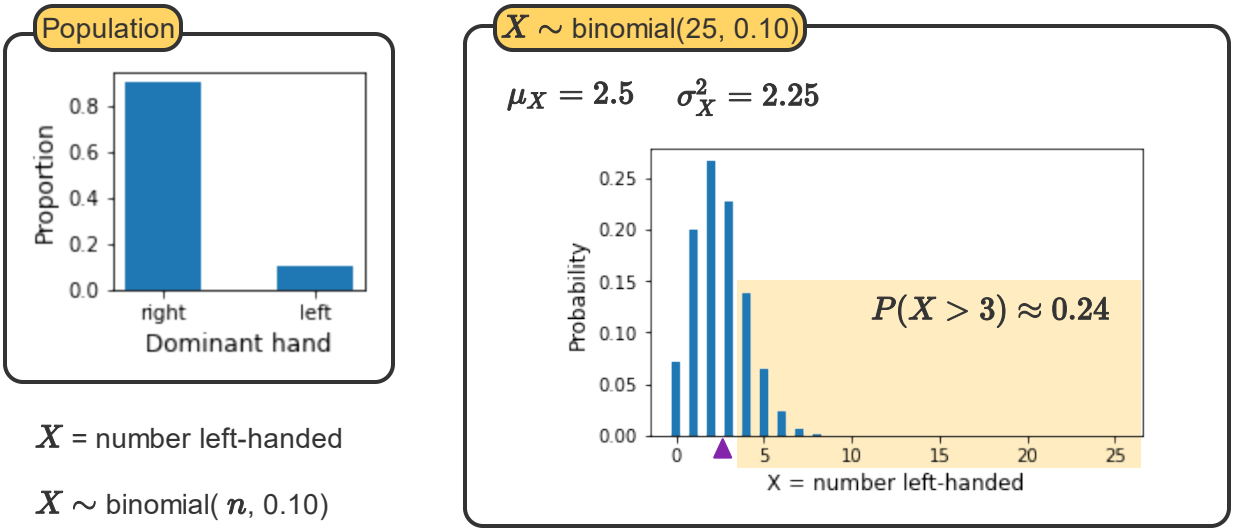




Binomial distribution

The outcome of a random process is usually observed more than once. Ex: Randomly select 25 students and record whether each student passed (1) or failed (0) a class. The total number of students who passed can be thought of as the sum of n = 25 independent Bernoulli(π) random variables. A random variable describing the number of "successes" from n independent observations of a random process in which the probability of a success is π follows a ***binomial distribution***. Th binomial distribution, binomial(n, π), is specified by two parameters: the number of outcomes observed, n, and the probability of a "success", π. Ex: Let Y be the total number of students who passed a class out of 25 randomly selected students. Then $Y \sim$ binomial(n = 25, π) with π as the probability of selecting a student who passed the class.

PARTICIPATION ACTIVITY 3.4.5: Binomial distribution.



Animation content:

Animation shows an example of a Binomial random variable.

- Step 1: A bar graph of the population distribution is shown for the dominant hand; the proportion of the population is the y-axis. The right-hand bar has a height of 0.90 and the left-hand bar has a height of 0.10. X = number left-handed and notation Z~binomial(n, 0.10) is shown.

- Step 2: A box with heading X~binomial(25, 0.100) is shown, inside the box is the probability distribution for X. Horizontal axis, X = number left-handed, is shown from 0 to 25, and the vertical axis is probability. Distribution has one main cluster and is right-skewed. The greatest probability is for X = 2, and after X = 8 bars are too small to notice.

- Step 3: A triangle indicating the mean appears on the on the graph at X = 2.5, notation for the mean, $\mu_X = 2.5$, and variance $\sigma_X^2 = 2.25$ is shown.

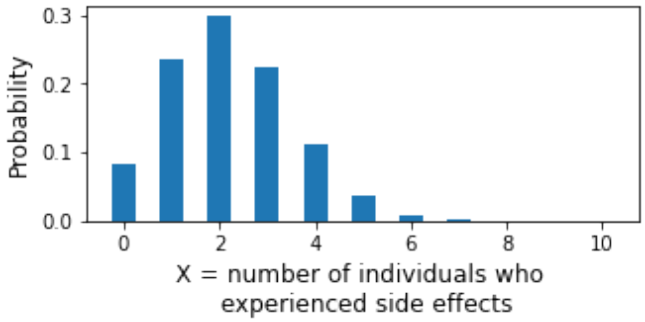
- Step 4: Area of the probability distribution from X = 4 to X = 25 is highlighted and notation $P(X > 3) \approx 0.24$ is shown.

Animation captions:

1. 10% of the population is left-handed. Let X be the number of left-handed individuals in a random sample of size n from the population, then $X \sim$ binomial(n, π = 0.10).
2. Suppose 25 individuals are randomly selected. The possible outcomes for X are the whole numbers 0 to 25, and the probability distribution can be visualized in a bar graph.
3. The mean of the binomial distribution is $\mu = n * \pi$ and the variance is $\sigma^2 = n * \pi * (1 - \pi)$. Ex: In random samples of size 25, expect an average of $\mu_X = 25 * 0.1 = 2.5$ individuals to be left-handed.
4. Probability distributions can provide useful insights. Ex: A classroom for 25 students has three left-handed desks. The probability that more than 3 of 25 students are left handed is $P(X > 3) \approx 0.24$.

PARTICIPATION ACTIVITY 3.4.6: Binomial distribution.

22% of individuals taking a particular medication experience substantial side effects. Let X be the number of individuals who experienced substantial side effects out of 10 randomly selected individuals who took the medication. X follows a binomial distribution, and the graph of the probability distribution for X is given below.



1) $X \sim$ binomial(n, π). Identify the value of n.

(##)

Check Show answer

2) $X \sim$ binomial(n, π). Identify the value of π.

(#.##)

Check Show answer

3) How many total possible values can the random variable X take on?

(##)

Check Show answer

4) Identify the probability of observing 4 of 10 individuals who experienced substantial side effects, $P(X = 4)$.

(#.##)

Check Show answer

5) What is the average number of total individuals who experienced substantial side effects, the mean of X, out of 10 randomly selected individuals who took the medication?

(#.#)

Check Show answer

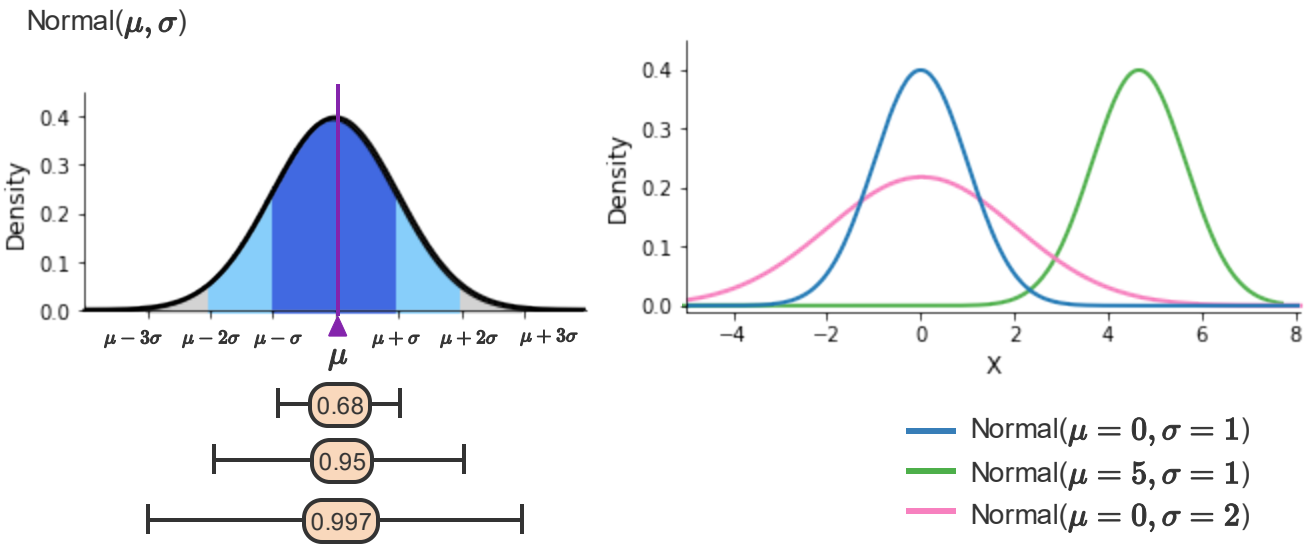
Normal distribution

Several common and named probability distributions also exist for continuous random variables. The ***normal distribution***, normal(μ, σ), is a symmetric, bell-shaped distribution and is determined by two parameters: the mean, μ, and the standard deviation, σ. The normal distribution tends to be a good approximation for physical measurements and test scores. Ex: The distribution of body temperature for a randomly selected adult approximately follows a normal distribution. In data

science, inference methods, such as hypothesis tests for proportions and regression analyses, typically use normally distributed random variables.

An important normal distribution is the standard normal distribution. The ***standard normal distribution*** is the normal distribution with mean μ = 0 and standard deviation σ = 1. A normally distributed random variable X can be standardized by subtracting the mean and dividing by the standard deviation, $Z = \frac{X-\mu}{\sigma}$.The standardized variable Z of z-scores follows the standard normal distribution,
$Z \sim$ normal(μ = 0,σ = 1).

PARTICIPATION ACTIVITY 3.4.7: Normal distribution.



Animation content:

Animation shows properties of normal probability distributions.

- Step 1: A graph with density on the vertical axis shows a symmetric, bell-shaped curve. Points X = a and X = b are labeled on the graph, the area under the curve above the interval [a,b] is shaded and labeled $P(a \leq X \leq b)$.

- Step 2: The shading and points disappear from the graph, the graph is labeled as normal(μ, σ). The center of the graph is labeled on the horizontal axis as mu and a triangle and vertical line are shown to illustrate the mean is the line of symmetry. Horizontal axis labels and ticks appear to denote the spread as one, two, and three standard deviations on both sides of the mean.

- Step 3: The curve is shaded between $\mu - \sigma$ and $\mu + \sigma$ and labeled as 0.68, then shaded between $\mu - 2\sigma$ and $\mu + 2\sigma$ and labeled as 0.95, finally shaded between $\mu - 3\sigma$ and $\mu + 3\sigma$ and labeled as 0.997.

- Step 4: Another graph appears with three normal density curves shown. The normal(0,1) cu is centered at X = 0 and extends from about X = -3 to X = 3. The normal(5,1) curve is centered at X = 5 and extends from about X = 2 to X = 8. So the same overall shape as the normal(0,1), but shifted to the right. The normal(0,2) curve is centered at X = 0 and extends from about X = -6 to X = 6. So, the normal(0,2) curve is centered at the same place as the normal(0,1), but more spread out.

Animation captions:

1. Density curves visualize continuous distributions, like the normal distribution. The area under the curve above an interval of values gives the probability of an observation falling within the interval.
2. Normal(μ, σ) distributions are bell-shaped and symmetric about the mean, μ. The spread of the distribution is specified by the standard deviation, σ.
3. For all normal distributions, about 68% of the distribution lies within one standard deviation of the mean, 95% within two standard deviations and, 99.7% within three standard deviations.
4. Changing the mean moves the location of the distribution. Increasing the standard deviation increases the distribution's spread.

PARTICIPATION ACTIVITY 3.4.8: Normal distributions.

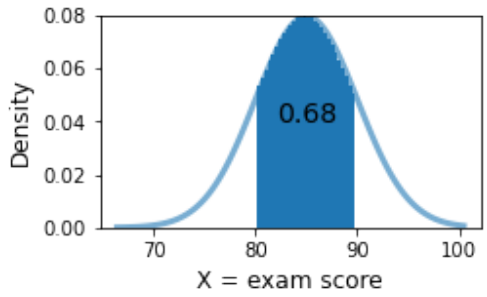
1) Normal distributions will _____ have a symmetric shaped distribution.
   - ○ never
   - ○ sometimes
   - ○ always

2) The standard normal distribution, normal(0,1), has a center _____ the center of the normal(10, 2) distribution.
   - ○ less than
   - ○ equal to
   - ○ greater than

3) The spread or variation of the normal(10,2) distribution is _____ the spread of the standard normal distribution, normal(0,1).

- less than
- equal to
- greater than

4) Let X be the exam score for a randomly selected student, and suppose $X \sim$ normal(85,5). The shaded area on the graph of the normal(85,5) probability distribution below shows _____ = 0.68.

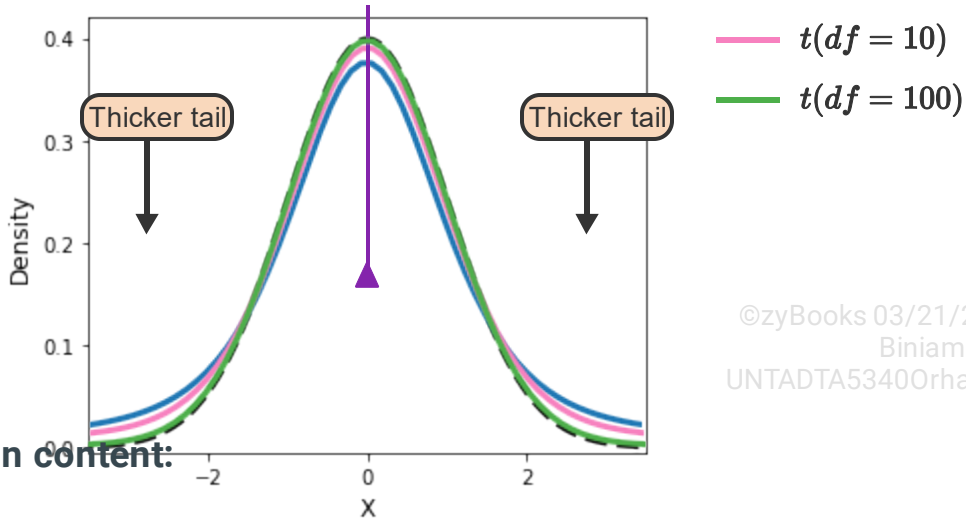


- $P(X = 80)$
- $P(X \geq 80)$
- $P(80 \leq X \leq 90)$

t-distribution

Sometimes a random variable may almost follow the standard normal distribution but has more variability.The t-distribution is a symmetric, bell-shaped curve, centered at 0, with a standard deviation greater than 1. The t-distribution, denoted as $t(df)$, is determined by one parameter, the degrees of freedom (df). The t-distribution has a similar shape and center as the standard normal distribution, but the t-distribution has a larger standard deviation.

PARTICIPATION ACTIVITY 3.4.9: t-distribution.



Animation content:

Animation shows several density curves for t-distributions and the standard normal distributi comparison.

- Step 1: Graph appears with X on the horizontal axis ranging from about -4 to 4 and den on the y-axis. The standard normal, normal(0,1) curve is graphed as a symmetric and be shaped distribution. The mean and center of the distribution is at 0 and marked by a tria and vertical line. The $t(df = 4)$ probability distribution is added to the graph. The distribution has a similar symmetric bell shape and is centered at 0, but the peak of the distribution is lower and the tails are thicker.

- Step 2: The t-distribution's thicker lower and upper tails are emphasized indicating more under the t-distribution curve in the tails compared to the standard normal.

- Step 3: The $t(df = 10)$ probability distribution is added to the graph and has the same overall shape and center as the $t(df = 4)$ and normal(0,1) but the tails are not as thick the $t(df = 4)$ distribution.

- Step 4: The $t(df = 100)$ distribution is added to the graph and is approximately equal t standard normal, normal(0,1), curve.

Animation captions:

1. The standard normal, normal(0,1), distribution is symmetric and bell-shaped with a mea 0. The t-distribution is also symmetric and bell-shaped with a mean of 0.
2. However, the t-distribution has more variability. Ex: The $t(4)$ has a standard deviation of about 1.41. Visually, $t(df)$ probability distributions have thicker tails.
3. The $t(df)$ distribution has standard deviation $\sigma = \sqrt{\frac{df}{df-2}}$ for df>2. Thus, t-distributions larger values of df have less variability but still more than the normal(0,1) distribution.
4. t probability distributions for large values of df are almost equal to the standard norma distribution. Ex: The $t(100)$ is approximately equal to the normal(0,1) distribution.

PARTICIPATION ACTIVITY 3.4.10: t-distribution.

1) The center of a t-distribution will be _____ the center for the standard normal distribution, normal(0,1).
   - less than
   - equal to
   - greater than

2) The spread or variability of a t-distribution will be _____ the variability of the standard normal distribution, normal(0,1).
   - less than
   - equal to
   - greater than

3) If $Z \sim$ normal(0,1), $P(Z \geq 2)$ = 0.067. Let $X \sim t(8)$. $P(X \geq 2)$ is _____ 0.067.
   - less than
   - equal to
   - greater than

Probability distribution graphs.

Use the drop-down menus to select a distribution and the corresponding parameters.

Distribution

binomial

Enter value or use - and +

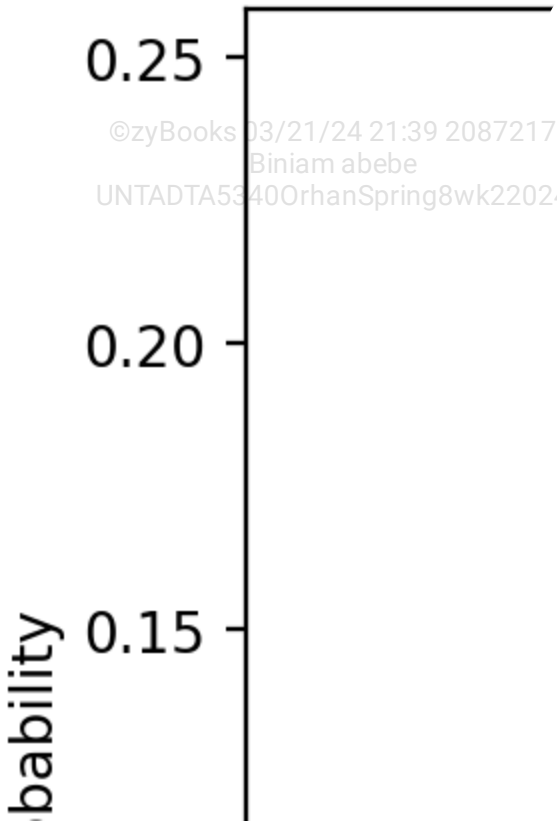
n

10

Probability

0.50

Add second distribution



Probability distributions in Python

The `SciPy` library contains a `stats` module which has various functions for probability distributions. Functions for the probability distributions discussed in this section are described in the table below. Additional functions and further information about parameters can be found in the SciPy stats documentation.

Table 3.4.1: SciPy functions for probability distributions.

Distribution	Functions	Parameters	Description
Bernoulli	`bernoulli.pmf(k, p)` `bernoulli.cdf(k, p)`	`p`= π sets the probability of a "success".	`bernoulli.pmf()` returns the probability $P(X = k)$, and the `bernoulli.cdf()`

			returns the probability $P(X \leq$ `k`$)$.
Binomial	`binom.pmf(k, n, p)` `binom.cdf(k, n, p)`	`n`$= n$ sets the number of observations. `p`$= \pi$ sets the probability of a "success".	`binomial.pmf()` returns the probability $P(X =$ `k`$)$, and the `binomial.cdf()` returns the probability $P(X \leq$ `k`$)$.
Normal	`norm.pdf(x, loc, scale)` `norm.cdf(x, loc, scale)`	`loc`$= \mu$ sets the mean and `scale`$= \sigma$ sets the standard deviation.	`norm.pdf()` returns the density curve's value at `x`, and `norm.cdf()` returns the probability $P(X \leq$ `x`$)$.
t	`t.pdf(x, df)` `t.cdf(x, df)`	`df`$= df$ sets the degrees of freedom for the distribution.	`t.pdf()` method returns the density curve's value at `x`, and `t.cdf()` returns the probability $P(X \leq$ `x`$)$.

Discrete distributions in Python.

Full screen

The code graphs probability distributions and calculates probabilities for Bernoulli and binomial distributions.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code to graph a Bernoulli(0.15) distribution.
- Modify the code to find $P(X = 15)$ if $X \sim$ binomial(100, 0.20)

In [1]:
```
# Import matplotlib package for graphing
# Import distributions from scipy.stats
import matplotlib.pyplot as plt
from scipy.stats import bernoulli
from scipy.stats import binom
```


In [2]:
```
# Let X~Bernoulli(0.35)

# Define possible values of X
x = [0, 1]

# Graph the probability distribution of X

# Get bar heights from Bernoulli distribution
plt.bar(x, height=bernoulli.pmf(k=x, p=0.35), width=0.75)
plt.ylabel('Probability', fontsize=12)
plt.xlabel("X", fontsize=12)
plt.xticks(ticks=x)  # add ticks for values of x
plt.show()
```

Continuous distributions in Python.

Full screen

The code graphs probability distributions and calculates probabilities for normal and t- distributions.

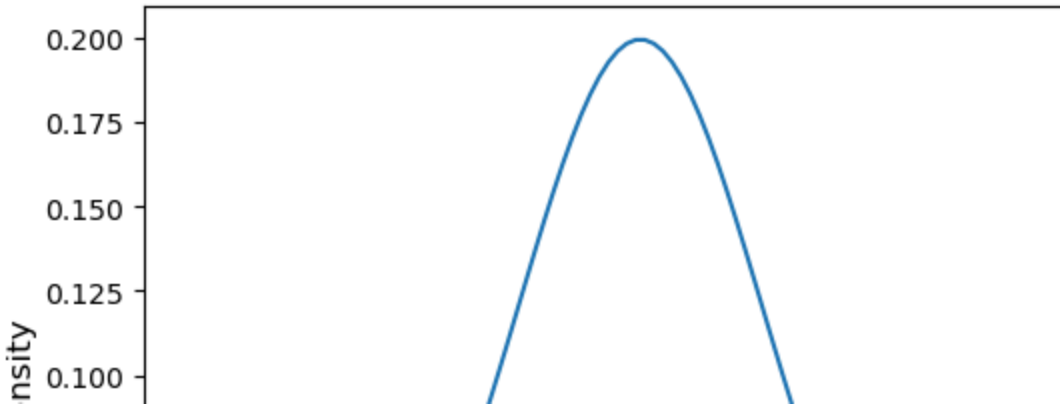
- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code to graph a normal(0,1) distribution.
- Modify the code to find $P(X \leq 1)$ if $X \sim$ normal(0, 1)
- Modify the code to find $P(X \leq 1)$ if $X \sim t(4)$. Compare $P(X \leq 1)$ for the two distributions.

In [1]:
```
# Import matplotlib and numpy packages
# Import distributions from scipy.stats
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm
from scipy.stats import t
```


In [2]:
```
# Let X~Normal(10,2)

# Define some values of X to use to graph the density curve
xvals = np.linspace(norm.ppf(0.0001, 10, 2), norm.ppf(0.9999, 10, 2)

# Graph the density curve for Normal(10,2)
plt.plot(xvals, norm.pdf(x=xvals, loc=10, scale=2))
plt.ylabel('Density', fontsize=12)
plt.xlabel("X", fontsize=12)
plt.show()
```



PARTICIPATION ACTIVITY 3.4.11: Probability distributions in Python.

1) A graph of the Bernoulli(0.80) distribution can also be generated from the binomial distribution using `binom.pmf(k=[0,1],_____, p=0.80)`.

- n=0.20
- n=80
- n=1

2) Let $X \sim$ binomial(50,0.50). The probability of X = 30, $P(X = 30)$, can be calculated using _____.

- `binom.cdf(k=30, n=50, p=0.50)`
- `binom.pmf(k=50, n=30, p=0.50)`
- `binom.pmf(k=30, n=50, p=0.50)`

3) Let $X \sim$ normal(120, 12), the code `norm.cdf(x=100,loc=120, scale=12)` calculates _____.

- $P(X \leq 100)$
- $P(X = 100)$
- $P(X \geq 100)$

CHALLENGE ACTIVITY 3.4.1: Normal distribution and z-scores.

537150.4174434.qx3zqy7

Start

A medical device manufacturer is producing ball bearings for hip replacements. The manufa process has natural variability, so not every ball bearing is exactly the same. A certain type c bearing has a mean diameter of 35 mm, with a standard deviation of 0.3 mm.

A recently manufactured ball bearing was selected at random for quality control testing. The randomly selected ball bearing has a diameter of 35.56 mm.

What is the ball bearing's z-score? Ex: 1.23

1 2

Check Next

CHALLENGE ACTIVITY 3.4.2: Calculating probabilities using SciPy.

537150.4174434.qx3zqy7

Start

Use Python methods to calculate the following probabilities for a binomial distribution with variable

- $P(X = 19)$
- $P(X \leq 19)$
- $P(X > 19)$

The code provided loads the packages and functions, reads variables `n` and `pi` from input, and prin probabilities.

```
# Import packages and functions
from scipy.stats import binom

# Set values for n and pi
n = int(input())
pi = float(input())

# Calculate probabilities
# Your code goes here
pEqual =
pLess =
pGreater =

# Print probabilities
print('P(X = 19) =', pEqual)
print('P(X <= 19) =', pLess)
print('P(X > 19) =', pGreater)
```

1 2

Check Next level

3.5 Inferential statistics

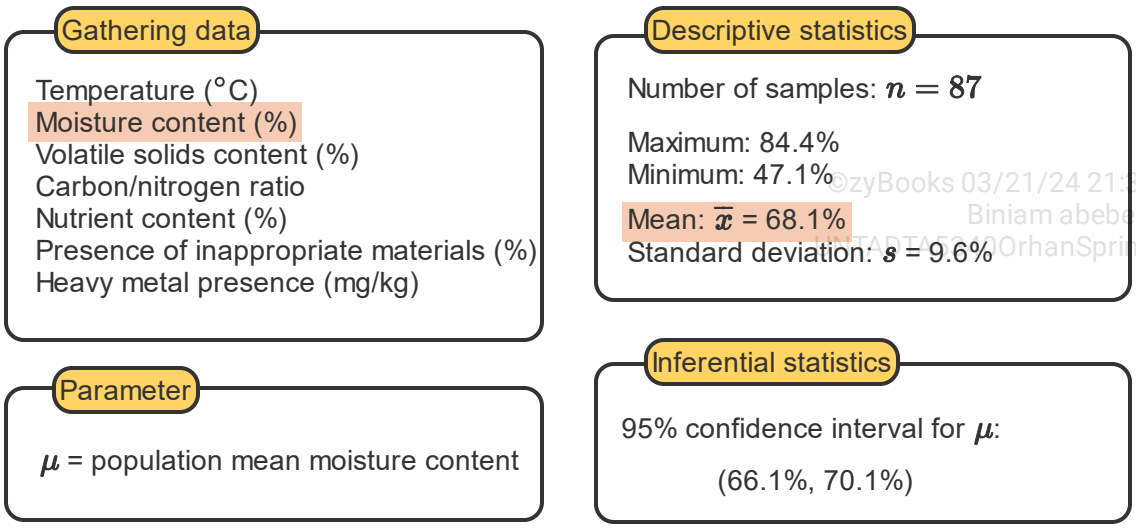
Learning goals

- Define a sampling distribution.
- Use a sampling distribution to determine if a sample statistic is unexpected.
- Define the Central Limit Theorem.
- List the steps of a hypothesis test for a population parameter.
- Define possible errors of a hypothesis test, and identify when an error may have been made.
- Interpret a confidence interval for a population parameter.

Inferential statistics

The questions and objectives of data science projects often need conclusions to be made beyond the scope of the collected sample data to a broader population. ***Inferential statistics*** are methods that result in conclusions and estimates about the population based on a sample. Testing claims about the population and estimating quantities of the population are the two primary inference methods. A numerical quantity of the population, such as the population mean or population proportion, is called a ***parameter***. Population parameters are usually unknown, and inferential statistics allow for generalizations to be made about the population based on the observed sample.

PARTICIPATION ACTIVITY 3.5.1: Inferential statistics for estimating compost moisture content.



Animation content:

Animation illustrates the information provided by both descriptive and inferential statistics using an example from a study about compost.

- Step 1: A box appears labeled Gathering data and contains a list of measurements recorded on the compost from each bin: temperature, moisture content, volatile solids content, carbon/nitrogen ratio, nutrient content, presence of inappropriate materials, and heavy metal presence. The moisture content measurement is highlighted.
- Step 2: A box appears labeled Descriptive statistics and contains various summaries of the observed data. The summaries include the number of samples n = 87, the maximum of 84.4%, the minimum of 47.1%, the mean of $\bar{x}$ = 68.1%, and the standard deviation of s = 9.6%. The sample mean is highlighted.
- Step 3: A box appears labeled Parameter and contains the population parameter, μ = population mean moisture content.
- Step 4: A box appears labeled Inferential statistics and contains the 95% confidence interval for μ is shown as (66.1%, 70.1%).

Animation captions:

1. A study examined the quality of compost produced by home compost bins in Galicia, Spain. One measure of compost quality is moisture content, or the percentage of water in the compost.
2. Descriptive statistics summarize the moisture of the compost from 87 randomly selected bins. The sample mean of $\bar{x}$ = 68.1% is higher than the desired maximum moisture content of 60%.
3. Inferential statistics allow the researchers to estimate the mean moisture content for the population of all similarly produced compost in Galicia, Spain, based on compost from the 87 sampled bins.
4. A 95% confidence interval for the population mean moisture content is (66.1%, 70.1%). Based on this analysis, a review of bin drainage and protection against heavy rains is recommended.

The efficiency of home composting programmes and compost quality. [1]

PARTICIPATION ACTIVITY 3.5.2: Descriptive and inferential statistics.

Match each component from the compost example with the appropriate term.

If unable to drag and drop, refresh the page.

Descriptive statistic | **Sample** | **Inferential statistic** | **Parameter** | **Population**

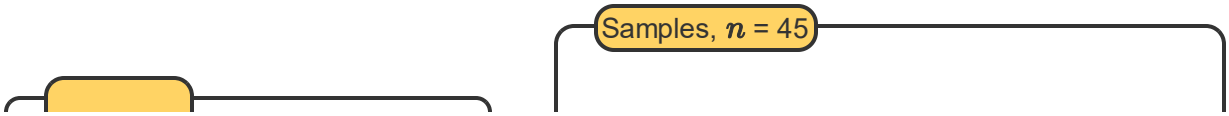
	The mean moisture content for all compost similarly produced by the home compost bins in Galicia Spain, μ
	The 87 randomly selected home compost bins
	The sample mean moisture content, $\overline{x} = 68.1$
	All home compost bins in Galicia, Spain
	The 95% confidence interval for the population mean moisture content, (66.1%, 70.1%)

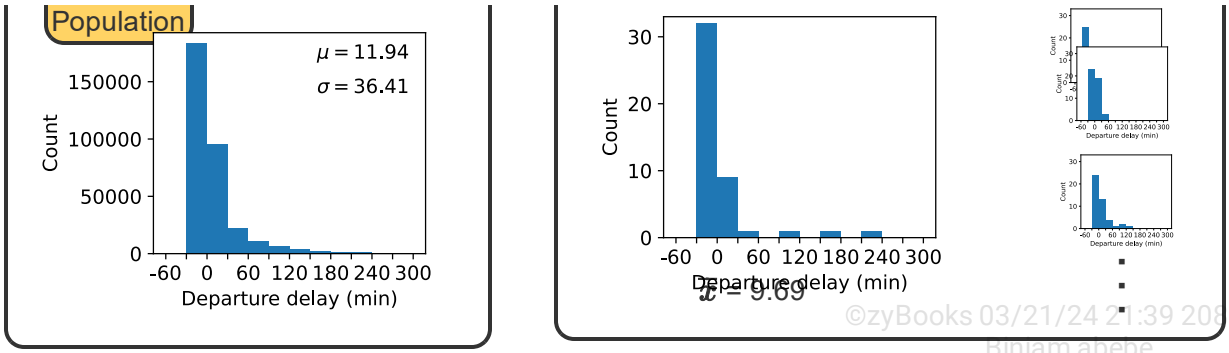
Reset

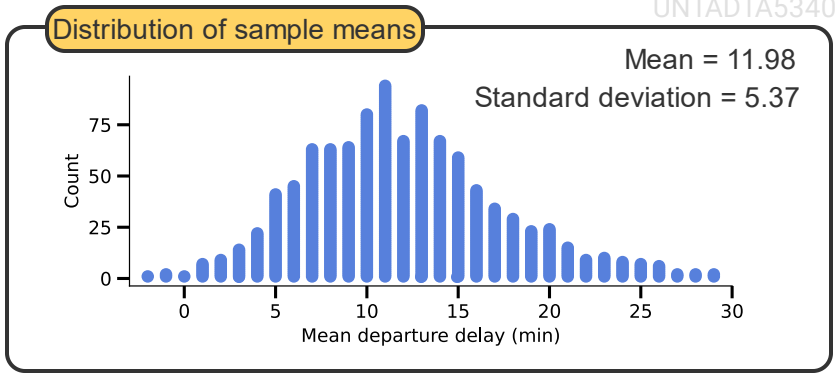
Sampling distributions

Sample statistics provide estimates of population parameters. Ex: The sample proportion of voters who plan to vote by mail is a statistic and estimates the parameter, or the population proportion, of all voters who plan to vote by mail. Because calculated statistics will vary from sample to sample due to natural sampling variability, statistics do not estimate parameters with 100% accuracy. The overall behavior of a statistic from repeated sampling is described by a sampling distribution.The ***sampling distribution*** of a statistic describes the statistic's possible values and a measure of how likely the values are to occur.

PARTICIPATION ACTIVITY 3.5.3: Sampling distribution of a sample mean.







Animation content:

Animation illustrates taking multiple random samples from a population, calculating the sample mean, and graphing the sample mean values to estimate the sampling distribution.

- Step 1: A box appears labeled Population and includes a histogram of the population of flight departure delay times. The horizontal axis is labeled Departure delay (min) and ranges from -60 to 300 minutes. The vertical axis is labeled count and ranges from 0 to about 180,000. The distribution is unimodal and very skewed right with most observations between -30 and 60 minutes. The population mean is μ = 11.94 minutes and the population standard deviation is σ = 36.41 minutes.

- Step 2: A box appears labeled Sample, n=45. A histogram of 45 randomly selected departure delay times is shown. The sample distribution has values between about -30 and 150 minutes and is slightly right skewed with most observations around 0 minutes. The sample mean $\bar{x}$ = 14.96. Another box appears labeled Distribution of sample means and shows a blank graph. The horizontal axis is labeled mean departure delay (min) with a range of -3 to 30 minutes and the vertical axis is labeled Count with a range of 0 to 85. The first dot appears on the graph to represent the first sample mean of $\bar{x}$ = 14.96.

- Step 3: The first sample fades away and a new histogram of a different 45 randomly selected departure delay times is shown. The distribution's values range from about -30 to 60 minutes with a sample mean of $\bar{x}$ = 3.58. A second dot appears on the distribution of sample means graph to represent the second sample's mean.

- Step 4: The second sample histogram fades and a third histogram of a different 45 randomly selected departure delay times is shown. The third sample mean is $\overline{x}$ = 13.09 and a third dot is added to the distribution of the sample means graph to represent the third sample's mean. The third histogram fades and three dots appear to indicate more samples of size 45 are taken. Then a final histogram of 45 randomly selected departure delay times appears with a mean of $\overline{x}$=9.69. A total of 1,000 dots are now shown on the distribution of sample means graph, including one for the final sample mean of 9.69 minutes. The distribution of sample means is unimodal and fairly symmetric with values ranging from about -3 to 30.

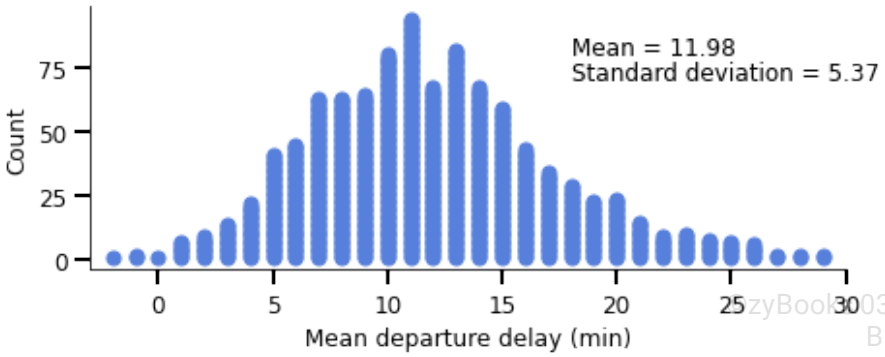
- Step 5: The distribution of sample means is shown to have a mean of 11.98 and a standard deviation of 5.37.

Animation captions:

1. The distribution of departure delay for all 327,911 flights that departed New York City airports with delay of less than 5 hours is skewed right with a mean of 11.94 minutes and standard deviation of 36.41 minutes.
2. A random sample of 45 flights is selected from this population. The distribution of departure delay for the sample is skewed right with a mean of 14.96 minutes.
3. A second random sample of 45 flights is selected from this population. The distribution of departure delay for the second sample has a mean of 3.58 minutes.
4. The process of selecting a sample of 45 flights from this population and finding the sample mean departure delay is repeated many times to estimate the sampling distribution of the sample mean.
5. Based on 1,000 random samples, the estimated sampling distribution of the sample mean is unimodal and fairly symmetric with a mean of 11.98 minutes and standard deviation of 5.37 minutes.

PARTICIPATION ACTIVITY 3.5.4: Sampling distribution of a sample mean.

Consider the estimated sampling distribution of the sample mean departure delay from the animation given below.



1) Each observation in the estimated sampling distribution represents _____.

- one departure delay time of a sample of 45
- the mean departure delay of a sample of 45

2) The estimated sampling distribution of the sample mean departure delay has a mean close to _____.

- μ = 11.94, the population mean departure delay
- $\overline{x}$ = 9.69, the sample mean from one sample of size 45

3) A random sample of 45 drawn from the population is _____ to have a mean departure delay of 10 minutes or less.

- likely
- unlikely

4) The graphed distribution is referred to as the estimated sampling distribution because the distribution graphs the means of _____ possible samples of size 45 drawn from the population.

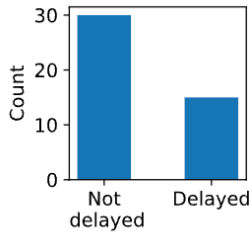
- all
- 1,000

PARTICIPATION ACTIVITY 3.5.5: Sampling distributions of counts and proportions.

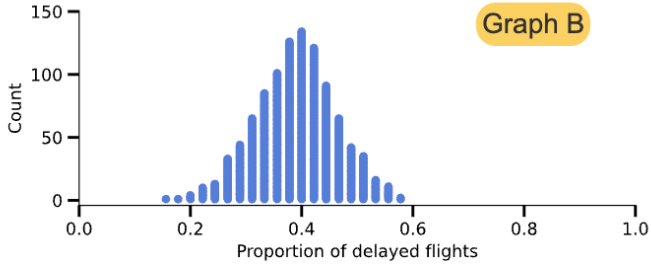
Flights with delay times greater than 0 minutes are considered delayed, and flights with delay times of 0 minutes or less are considered to have no delay. The sample count, or total number of delayed flights, and the sample proportion of delayed flights are found for random samples of 45 flights from the population.

Match each graph with the appropriate description.

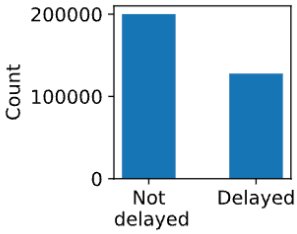
Graph A



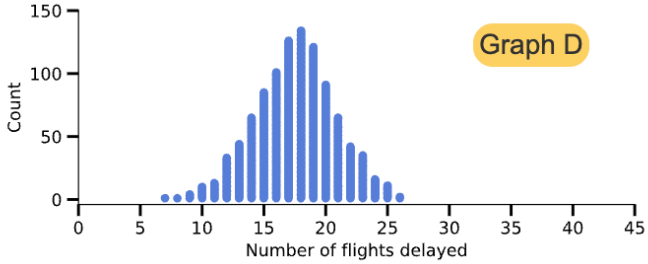
Graph B



Graph C



Graph D



If unable to drag and drop, refresh the page.

Graph A | Graph C | Graph B | Graph D

	Population distribution of whether or not the flight was delayed for all flights
	Estimated sampling distribution of the sample proportion of delayed flights based on 1,000 random samples of 45 flights
	Estimated sampling distribution of the sample count of delayed flights based on 1,000 random samples of 45 flights
	Sample distribution of whether or not the flight was delayed for

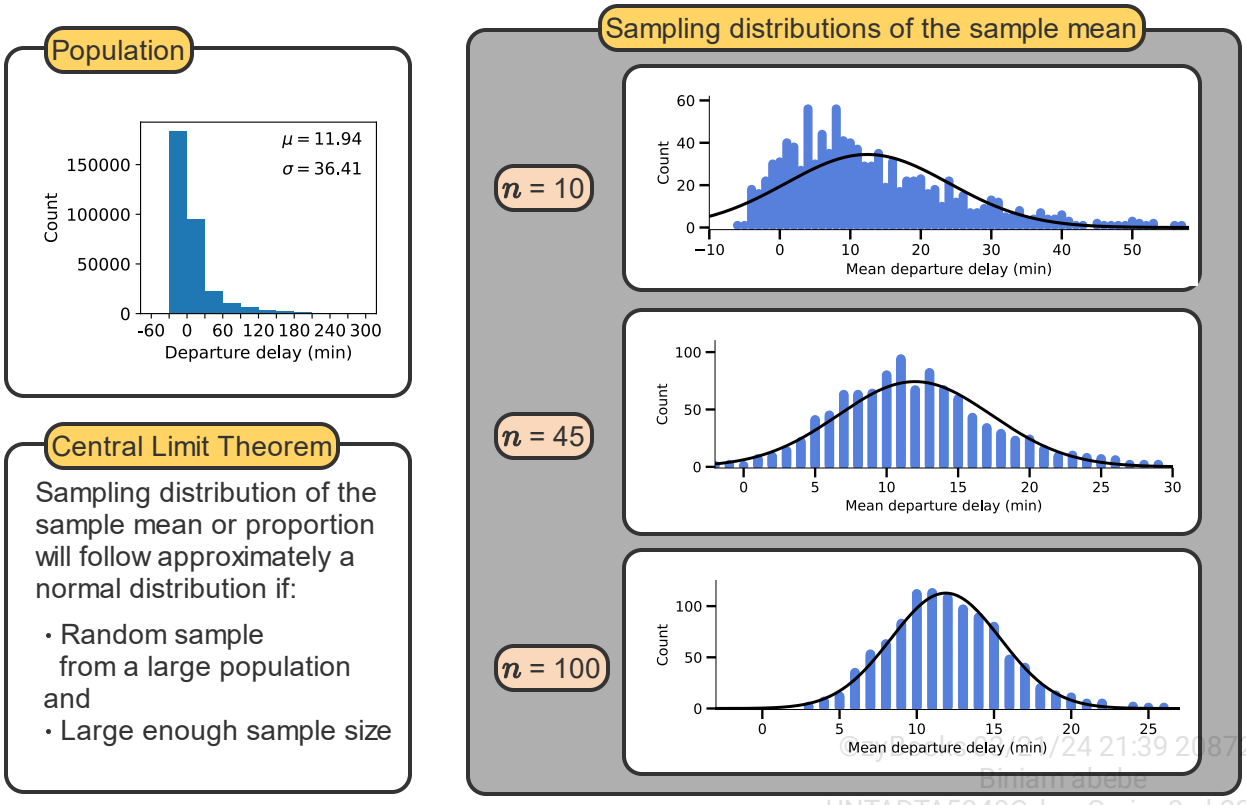
one random sample of 45 flights

Reset

Central Limit Theorem

The sampling distributions of many statistics of interest look similarly symmetric and bell-shaped. An important mathematical result, the Central Limit Theorem, specifies the conditions for which a statistic's sampling distribution will approximately follow a normal distribution. The ***Central Limit Theorem*** (CLT) states that if random samples of size n are drawn from a large population and n is large enough, then the sampling distribution of the sample mean will follow approximately a normal distribution. The CLT also applies to proportions since a proportion can be expressed as the mean of zeros and ones.

PARTICIPATION ACTIVITY 3.5.6: Central Limit Theorem.



Animation content:

Animation shows the sampling distributions for the sample mean for three different sample sizes: n=10, n=45, and n=100.

- Step 1: A box appears labeled Population and includes a histogram of the population of flight departure delay times. The horizontal axis is labeled Departure delay (min) and ranges from -60 to 300 minutes. The vertical axis is labeled count and ranges from 0 to about 180,000. The distribution is unimodal and very skewed right with most observations between -30 and 60 minutes. The population mean is μ = 12.64 minutes and the population standard deviation is σ = 40.21 minutes.

- Step 2: A box appears labeled Sampling distribution of the sample mean. A dot plot of 1,000 sample means of random samples of size n=10 is shown. The horizontal axis is labeled Mean departure delay (min) and ranges from 10 to 60 minutes. The vertical axis is labeled Count and ranges from 0 to 60. The distribution is skewed right with the cluster of observations between about 0 minutes and 15 minutes.

- Step 3: Two more dot plots appear for sample sizes of n=14 and n=100. The dotplot of 1,000 sample means of random samples of size n=45 is close to symmetric with values ranging from about -5 to 30 minutes. The dot plot of 1,000 sample means of random samples of size n=100 is close to symmetric with values ranging from about 3 to 26.

- Step 4: A box labeled Central Limit Theorem appears with the following text. Sampling distribution of the sample mean or proportion will follow approximately a normal distribution if: random sample from a large population and large enough sample size.

- Step 5: A normal distribution curve is overlaid on each of the three dot plots. The normal curve overlay does not fit the dot plot distribution for the n=10 graph, fits reasonably well for the n=45 graph and fits the closest for the n=100 graph.

Animation captions:

1. A population of departure delay times for all flights out of New York City airports is skewed right with a mean of μ=11.94. Consider the sample mean of a random sample of size n.
2. The sampling distribution of the sample mean captures the sampling variability of the sample mean. Ex: If 1,000 samples of size n=10 are drawn, a dot plot of the sample means estimates the sampling distribution.
3. The sampling distribution of the sample mean is dependent on the sampling situation. Ex: For samples of size n =45 or n =100, the shape and variability of the sample mean's distributions are different.
4. The Central Limit Theorem (CLT) is a useful mathematical result for knowing when the sampling distribution of the sample mean or proportion will follow approximately a normal distribution.

5. How large a sample needs to be will depend on the situation. Ex: For the sampling distribution of mean departure delay, as the sample size increases, the distribution more closely follows a normal distribution.

PARTICIPATION ACTIVITY 3.5.7: Central Limit Theorem.

1) If certain conditions are met, the Central Limit Theorem states the _____ will approximately follow a normal distribution.
   - ○ distribution of a sample
   - ○ population distribution
   - ○ sampling distribution of a sample mean

2) One condition needed for the Central Limit Theorem to apply is that the _____ is sufficiently large.
   - ○ sample size
   - ○ total number of samples
   - ○ the sample mean or proportion

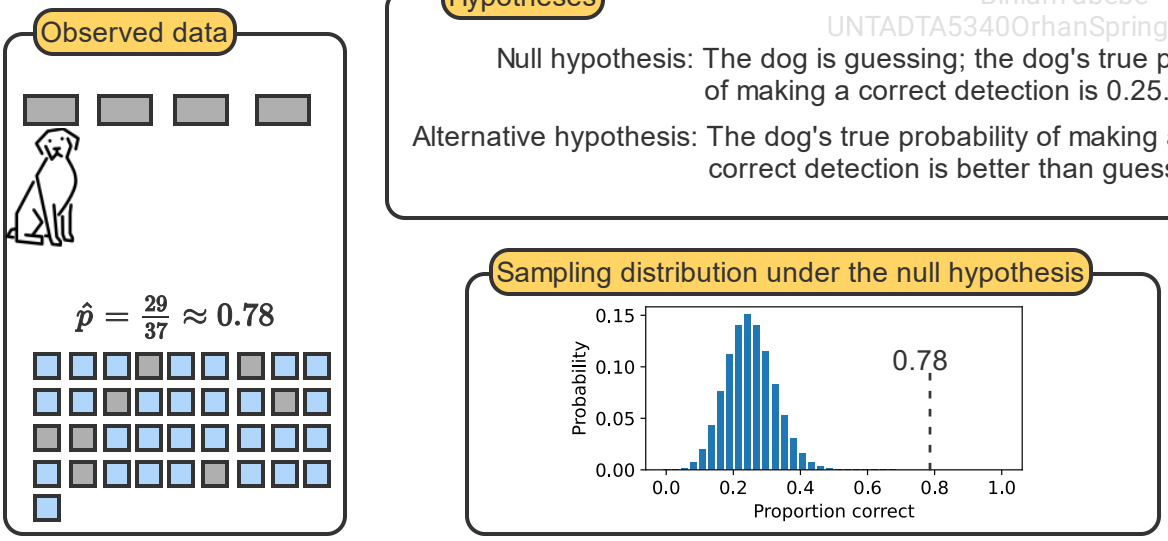
Hypothesis testing

A ***hypothesis test*** is a method for evaluating a claim, or hypothesis, about a population parameter by examining the statistical evidence against the claim based on a sample. The conclusion of a hypothesis test is a decision that the observed data either indicate the claim is plausible or support an alternative explanation. The following steps outline the general process of conducting a hypothesis test.

1. State null and alternative hypotheses about parameters. The ***null hypothesis***, H_0, is typically the by-chance or no-effect explanation, and the ***alternative hypothesis***, H_a, is typically the explanation of an effect, or difference.
2. Calculate a statistic of interest from the sample data that is used to evaluate the null hypothesis.
3. Determine the ***p-value***, or likelihood, of obtaining a statistic at least as extreme as the observed statistic when the null hypothesis is true.

4. Draw a conclusion about the null hypothesis based on the statistical evidence provided by the p-value.

PARTICIPATION ACTIVITY 3.5.8: Can a dog detect cancer by smell?



Animation content:

Animation illustrates using a hypothesis test to answer a research question about whether or not a dog can be trained to detect cancer by smell.

- Step 1: A box appears labeled Observed data with an icon of a dog and a question mark.
- Step 2: In the observed data box, four rectangles appear to represent the masks. The dog moves by each mask and then moves back by one of the masks. This mask turns blue to indicate a correct detection was made.
- Step 3: A blue square appears indicating the first trial a correct detection was made. Then the dog goes through another trial of smelling each mask and choosing one. A total of 37 boxes appear to represent the 37 total trials, and in 29 of the trials a correct detection was made. The sample proportion, $\hat{p} = \frac{29}{37} \approx 0.78$ is shown.
- Step 4: A box labeled Hypotheses appears with the following text. Null hypothesis: The dog is guessing; the dog's true probability of making a correct detection is 0.25. Alternative hypothesis: The dog's true probability of making a correct detection is better than guessing.

- Step 5: A box labeled Sampling distribution under the null hypothesis appears with a probability distribution graph for the sampling distribution of the sample proportion. The horizontal axis is labeled Proportion correct and ranges from 0 to 1. The vertical axis is labeled Probability and ranges from 0 to 0.15. The distribution has a symmetric cluster between 0.1 and 0.4, centered around 0.25. Proportions less than 0.05 or greater than 0.5 have almost zero probability of occurring.

- Step 6: The value of the sample proportion, 0.78, is marked on the sampling distribution graph. A proportion correct of 0.78 is in the right tail of the distribution, where proportions in the tail occur with almost 0 probability.

Animation captions:

1. Can a dog be trained to detect whether or not an individual has liver cancer by smelling the individual's breath?
2. A pilot study collected data on a trained dog. For each trial, the dog smelled four face masks. One mask had been worn by an individual with liver cancer and the other three were controls.
3. The dog made a correct detection for 29 of 37 trials for a sample proportion of $\hat{p} \approx 0.78$. A hypothesis test can help determine whether the data indicate the dog detected correctly better than by chance.
4. The null hypothesis states the dog is guessing and the probability of a correct guess is $\pi = 0.25$. The alternative is the explanation the dog did better than chance and might be able to detect liver cancer.
5. If the null hypothesis is correct and the dog is guessing, the sampling distribution of the sample proportion describes the possible values of the sample proportion and their likelihood.
6. The observed sample result is compared to the sampling distribution assuming the null hypothesis is true. The null hypothesis is rejected if a sample proportion of 0.78, or one more extreme, is unlikely.

The detection of hepatocellular carcinoma (HCC) from patients' breath using canine scent detection: a proof-of-concept study.
[2]

PARTICIPATION ACTIVITY 3.5.9: Hypothesis testing.

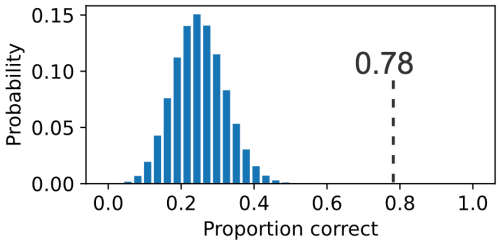
1) The null hypothesis states the dog is guessing and the dog's actual probability of making a correct detection, π, is 1/4 or 0.25. The null hypothesis stated in symbols is _____.

- $H_0 : \pi > 0.25$
- $H_0 : \pi = 0.25$
- $H_0 : \pi = 0.78$

2) The alternative hypothesis is that the dog does better than guessing and might be able to detect liver cancer. The alternative hypothesis stated in symbols is _____.

- $H_a : \pi > 0.25$
- $H_a : \pi \neq 0.25$
- $H_a : \pi = 0.78$

3) A graph of the sampling distribution of the sample proportion of correct detections out of 37 trials when the null hypothesis is true is given below. Based on this graph, observing a sample proportion of 0.78 or greater is _____ .



- likely
- unlikely

4) Based on the hypothesis test, the conclusion drawn is the observed data provide statistical evidence that _____ .

- the dog was likely only guessing
- the dog can actually smell cancer
- the dog's proportion of correct detection is more than only guessing

5) The Central Limit Theorem could apply to the sampling distribution of the sample proportion for this situation if the sample size is large enough. The large enough sample size condition _____ .

- is reasonably met because the sampling distribution of the sample proportion is symmetric and resembles a normal distribution
- is not reasonably met because the sampling distribution of the sample proportion is skewed right and does not resemble a normal distribution

Type I and type II errors

The decision from a hypothesis test is to either reject the null hypothesis or fail to reject the null hypothesis. The ***significance level***, α, of a hypothesis test is how small the p-value must be to conclude the data provide enough statistical evidence to reject the null hypothesis. The decision is to reject the null hypothesis if the p-value is less than or equal to α, and fail to reject the null hypothesis if the p-value is greater than α.

In reality, either the null hypothesis is true or the alternative hypothesis is true. Thus, the conclusion from a hypothesis test is either correct or incorrect. Ex: Suppose the conclusion from a hypothesis test is that the data support a population mean commute time of 25 minutes. If the mean commute time of the population is actually about 25 minutes, then a correct decision

is made. But, if the population mean commute time is actually 40 minutes, then an incorrect decision is made.

- A ***type I error*** is rejecting the null hypothesis in favor of the alternative when in reality the null hypothesis is true.
- A ***type II error*** is failing to reject the null hypothesis when in reality the alternative hypothesis is true.

PARTICIPATION ACTIVITY 3.5.10: Type I and type II errors.

Decision errors

Hypothesis test decision	Reality: H_0 true	Reality: H_a true
p-value $\leq \alpha$, Reject H_0	Type I error	Correct
p-value $> \alpha$, Fail to reject H_0	Correct	Type II error

Example

p-value $\approx 0 \leq \alpha = 0.05$ Reject $H_0 : \pi = 0.25$ Correct or Type I error

Animation content:

Animation illustrates using a hypothesis test to answer a research question about whether or not a dog can be trained to detect cancer by smell.

- Step 1: A box appears labeled Decision errors, the rows of the table represent the hypothesis test decisions, Reject H_0 if p-value $\leq \alpha$ and Fail to reject H_0 if p-value $> \alpha$.

- Step 2: The columns of the table represent reality where either H_0 is true or H_a is true. The cells of the table are filled in with either error or correct.

	H_0 true	H_a true
p-value $\leq \alpha$, Reject H_0	Error	Correct

p-value $>$ α, Fail to reject H_0	Correct	Error

- Step 3: The Reject H_0 row heading and H_0 true column heading are highlighted and the corresponding table cell with Error is changed to Type I error.
- Step 4: The Fail to reject H_0 row heading and H_a true column heading are highlighted and the corresponding table cell with Error is changed to Type II error. The table is now

	H_0 true	H_a true
p-value $\leq$ α, Reject H_0	Type I Error	Correct
p-value $>$ α, Fail to reject H_0	Correct	Type II Error

- Step 5: A box labeled Example appears along with the following text. P-value $\approx 0 \leq \alpha$ and Reject $H_0 : \pi = 0.25$
- Step 6: The text Correct or Type II error appears in the Example box.

Animation captions:

1. The decision from a hypothesis test is to either reject the null hypothesis or fail to reject the null hypothesis based on how the p-value compares to the significance level, α.
2. The hypothesis test decision is either correct or an error. In reality, either the null hypothesis is true or the alternative hypothesis is true.
3. A type I error occurs when the hypothesis test decision rejects a true null hypothesis. The probability of making a type I error is the significance level, α.
4. A type II error occurs when the hypothesis test decision fails to reject a false null hypothesis.
5. The study examining whether or not a dog can be trained to smell cancer resulted in a small p-value. Because the p-value is less than α = 0.05, the null hypothesis is rejected.
6. The decision is either correct or a type I error. A potential consequence of making a type I error is spending more resources when the dog does not do better than guessing.

PARTICIPATION ACTIVITY 3.5.11: Type I and type II errors.

A new app alerts users of severe weather in advance of the weather event using device location, weather information, and public postings. The app's development team collects data to conduct a hypothesis test on whether the proportion of accurate early alerts is at least 0.85 before releasing the app.

Match each description with the appropriate term.

If unable to drag and drop, refresh the page.

Type I error | **Type II error** | **Type I error consequence** | **Hypotheses** | **Type II error consequence**

	$H_0 : \pi = 0.85, H_a : \pi < 0.85$
	Spend additional time and money to improve the app's accuracy when the app already meets the desired level of accuracy and could be released.
	Decide based on the hypothesis test that the app's proportion of accurate alerts is 0.85, but in reality the proportion is less than 0.85.
	Decide based on the hypothesis test that the app's proportion of accurate alerts is less than 0.85, but in reality the proportion is 0.85.
	Releasing the app claiming accuracy of at least 85% and receiving poor reviews because the app is actually less accurate.

Reset

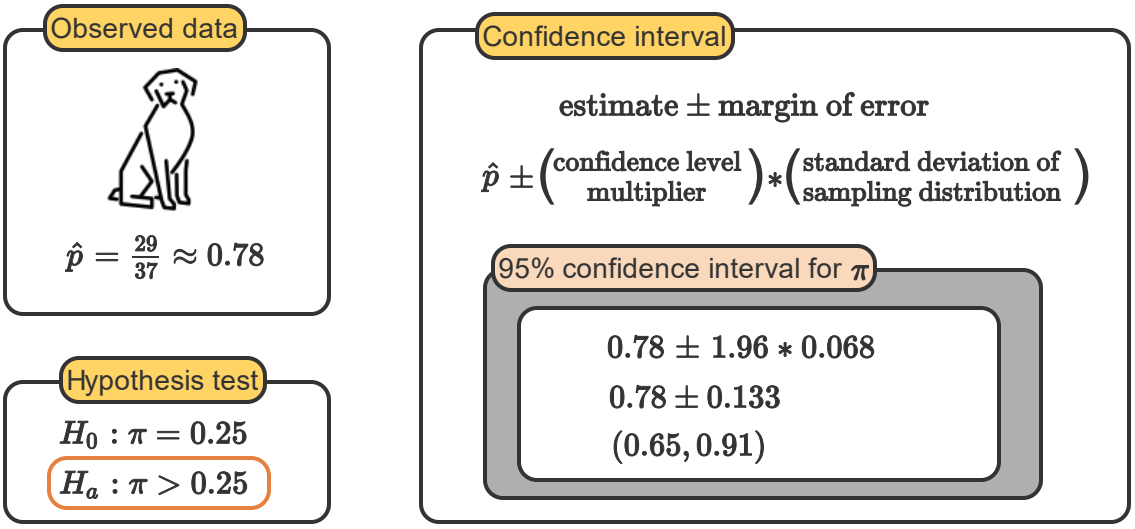
Estimation

Another inference method provides an estimate for the value of a population parameter. Ex: What is the mean moisture content for all compost produced in home compost bins? A ***confidence interval*** provides an interval of possible values for the parameter being estimated. A confidence interval is constructed using the general equation $estimate \pm margin\ of\ error$.

The estimate is a statistic calculated from the sample data and gives an initial best guess for the parameter's value. The ***margin of error*** measures the precision of the estimate and includes:

- the standard error, or measure of sampling variability, which comes from the statistic's sampling distribution, and
- the confidence level, or measure of interval reliability.

PARTICIPATION ACTIVITY 3.5.12: Constructing a confidence interval.



Animation content:

Animation illustrates constructing a confidence interval for the study about whether or not a dog can be trained to detect cancer by smell.

- Step 1: A box labeled Observed data appears with a dog icon and the sample proportion $\hat{p} = \frac{29}{37} \approx 0.78$ is given. A second box appears labeled Hypothesis test and includes the null hypothesis $H_0 : \pi = 0.25$ and the alternative hypothesis $H_a : pi > 0.25$. The alternative hypothesis is circled.

- Step 2: A box appears labeled Confidence interval with the general equation $\textbf{estimate} \pm \textbf{margin of error}$.
- Step 3: The estimate portion of the confidence interval equation is highlighted and then the sample proportion symbol, $\hat{p}$, flies over from the observed data box.

- Step 4: The margin of error portion of the confidence interval equation is highlighted and the two parts of the margin of error are added to the sample proportion to complete the equation:
  $\hat{p} \pm$ **confidence level multiplier** $*$ **standard deviation of sampling distribution**.

- Step 5: A box labeled 95% confidence interval for π appears with the following equations: $0.78 \pm 1.96 * 0.068 = 0.78 \pm 0.133$. The resulting confidence interval is (0.65, 0.91).

Animation captions:

1. The hypothesis test conclusion is that the data suggest the trained dog makes correct detections better than guessing.
2. A common follow-up analysis is to construct a confidence interval to estimate the dog's true probability of making a correct detection.
3. The sample proportion $\hat{p} \approx 0.78$ provides an initial estimate of π, the dog's true probability of making a correct detection.
4. The margin of error provides a measure of the estimate's precision by taking into account sampling variability and the level of confidence.
5. With 95% confidence, the dog's long-run proportion of correct detections is estimated to be between 0.65 and 0.91.

PARTICIPATION ACTIVITY 3.5.13: Confidence Interval for a population mean.

A survey[3] to understand train passengers' satisfaction with train usage in the five largest cities in Australia was used to collect information about the passengers' backgrounds and about the passengers' most recent train travel experiences.

An estimate is needed for the population mean total travel time for the population of all passengers in the five cities. A passenger's total travel time includes time spent waiting, on-boarding, and traveling on the train. The sample mean total travel time of the 2,927 survey responses is 62.60 minutes with a standard deviation of 25.83 minutes.

1) An initial estimate of the parameter μ, the population mean total travel time for all passengers in the five cities, is _____ minutes.

- 25.83
- 62.60
- 2,927

2) The margin of error for constructing a 95% confidence interval for μ from the observed data is 0.94. The 95% confidence interval is _____.

- (-0.94, 0.94)
- (61.66, 63.54)
- (60.76, 64.44)

3) The interpretation of the 95% confidence interval is that with 95% confidence, the _____ is at least 61.66 minutes but no more than 63.54 minutes.

- population mean total travel time for all passengers in the five cities
- the sample mean total travel time for the 2,927 passengers surveyed
- total travel time of a passenger from the population

(*1) Vázquez, M. A., and M. Soto. "The efficiency of home composting programmes and compost quality." *Waste Management* 64 (2017): 39-50. https://doi.org/10.1016/j.wasman.2017.03.022

(*2) Kitiyakara, Taya, Susan Redmond, Nattawut Unwanatham, Sasivimol Rattanasiri, Amarin Thakkinstian, Pongsatorn Tangtawee, Somkit Mingphruedhi, Abhasnee Sobhonslidsuk, Pongphob Intaraprasong, and Chomsri Kositchaiwat. "The detection of hepatocellular carcinoma (HCC) from patients' breath using canine scent detection: a proof-of-concept study." *Journal of breath research* 11, no. 4 (2017): 046002. https://iopscience.iop.org/article/10.1088/1752-7163/aa7b8e

(*3) Zheng, Zuduo, Albert Wijiweera, Keith Sloan, Simon Washington, Paul Hyland, Hong To, Michael Charles et al. "Understanding urban rail travel for improved patronage forecasting-Final report." (2013).

3.6 Inference for proportions and means

Learning goals

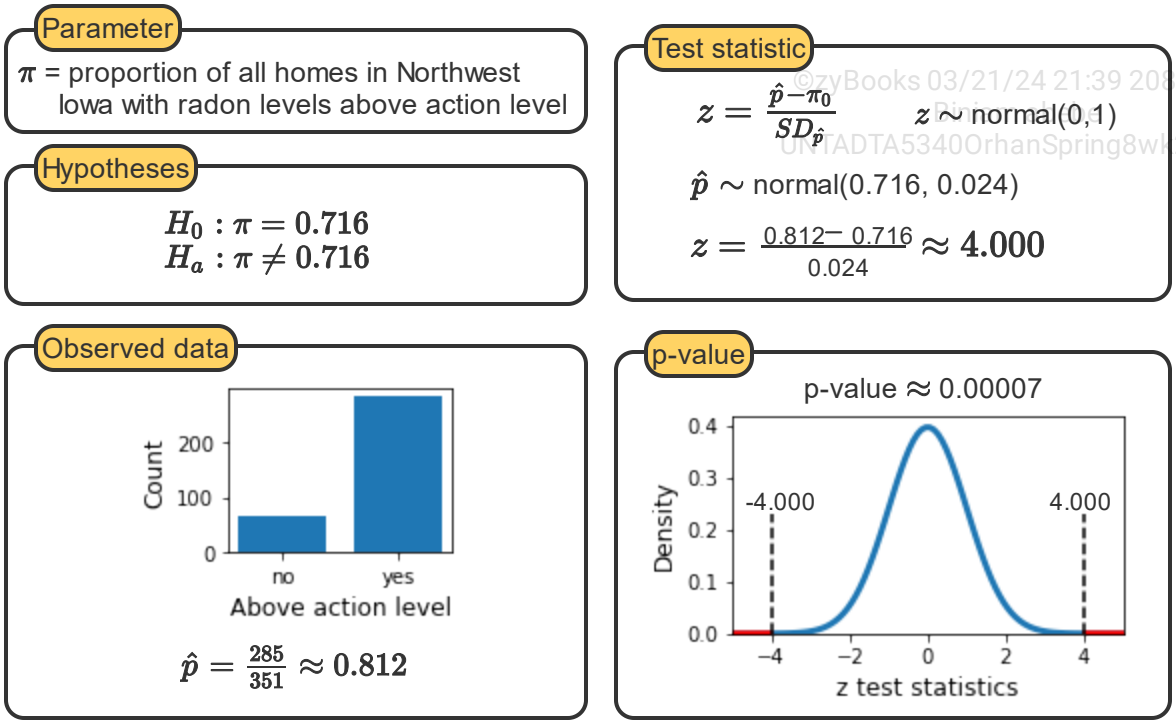
- Identify the appropriate inference method for a research question.
- Write the null and alternative hypotheses for a hypothesis test given a research question.
- Interpret the test statistic and p-value.
- Make a conclusion about the population parameter(s) based on a p-value.
- Use Python to obtain the test statistic and p-value for a hypothesis test about a proportion or mean.
- Use Python to construct a confidence interval for a proportion or mean.

Theory-based inference methods can help answer important questions about parameters of one or more populations. The general process of conducting a hypothesis test or constructing a confidence interval is mostly the same, but with slight variations for calculating the test statistic or for the test statistic's sampling distribution. Typically in data science, statistical software is used for the calculations. This section provides some calculation details when working with means and proportions and provides examples that further illustrate the use of statistical inference methods.

Inference for one proportion

Iowa Radon Survey reports that 71.6% of homes in Iowa have radon levels above the US Environmental Protection Agency (EPA) action level of 4pCi/L. A 2015 study [1] found that 285 of 351 homes tested for radon in Northwest Iowa had random levels about the EPA action level. A hypothesis test can answer the question: Based on the data collected, is the proportion of homes in Northwest Iowa with radon levels above action level the same as or different from the state proportion of 0.716?

PARTICIPATION ACTIVITY 3.6.1: Northwest Iowa radon one-proportion hypothesis test.



Animation content:

Animation shows an example of a one-proportion hypothesis test using the information from the study on radon levels of homes in Northwest Iowa.

- Step 1: A box appears labeled Parameter and contains the text π = proportion of all Northwest Iowa homes with radon levels above action level. A box appears labeled Hypotheses and contains the null hypothesis stated in symbols as $H_0 : \pi = 0.716$ and the alternative hypothesis stated in symbols as $H_a : \pi \neq 0.716$.
- Step 2: A box appears labeled Observed data and contains a bar graph of the data. The horizontal axis is labeled above action level with possible outcomes of no and yes. The vertical axis is labeled count. The no bar shows a count of 66 and the yes bar shows a count of 285. Below the graph the sample proportion is given as $\hat{p} = \frac{285}{351} \approx 0.812$.
- Step 3: A box appears labeled Test statistic. The z test statistic equation $z = \frac{\hat{p} - \pi_0}{SD_{\hat{p}}}$ and $z \sim$ normal(0,1) appear. The distribution of the statistic appears as $\hat{p} \sim$ normal(0.716, 0.024). An empty equation to calculate the z test statistic appears and is first filled in with 0.812 flying

over from the observed data box into the equation, next 0.716 and 0.024 fly from the distribution of $\hat{p}$ into the equation resulting in $z = \frac{0.812-0.716}{0.024} \approx 4.000$.

- Step 4: A box labeled p-value appears with a graph of the normal(0,1) density curve. The horizontal axis is labeled z test statistics and ranges from about -5 to 5. The vertical axis labeled Density and ranges from 0 to about 0.4. The density curve is unimodal, symmetric, and centered at 0. The values -4.000 and 4.000 appear and fly from the Test statistic box onto the density curve graph. The area under the curve for horizontal axis values less than -4.000 and greater than 4.000 are shaded red to indicate the area for computing the p-value. Above the graph, p-value ≈ 0.00007 appears.

Animation captions:

1. The population parameter of interest is the proportion, π, of all homes in Northwest Iowa with radon levels above action level. The null and alternative hypotheses identify the two possible explanations.
2. The observed sample proportion of homes in Northwest Iowa with radon levels above action level is $\hat{p} = 285/351 \approx 0.812$.
3. A standardized test statistic is typically used. Assuming $n = 351$ is large enough to apply the Central Limit Theorem, the sampling distribution of $\hat{p}$ will be approximately normal.
4. The normal(0,1) distribution is used to find the p-value, or probability, of obtaining the observed statistic, $z \approx 4.000$, or one more extreme in the direction of the alternative hypothesis.

PARTICIPATION ACTIVITY 3.6.2: Northwest Iowa radon one proportion inference example.

Consider the Iowa radon hypothesis test example in the animation above investigating whether the proportion of homes in Northwest Iowa with radon levels above action level are the same as or different from the state proportion of 0.716.

1) The alternative hypothesis is the proportion of all homes in Northwest Iowa with radon levels above action level is _____ the state proportion of 0.716.

- ○ equal to
- ○ greater than
- ○ not equal to

2) The standardized test statistic of $z \approx 4.000$, can be interpreted as the observed sample proportion is 4.000 standard deviations above _____.

- 0.024
- 0.716
- 0.812

3) The p-value is the probability of observing a z test statistic of _____ assuming $\pi = 0.716$.

- 4.000 or less
- 4.000 or more
- -4.000 or less or 4.000 or more

4) Because the p-value is very small, the decision from the hypothesis test is to _____.

- reject the null hypothesis
- fail to reject the null hypothesis
- reject the alternative hypothesis

5) The 95% confidence interval for π is (0.771,0.853), and can be interpreted as the researchers are 95% confident that the proportion of _____ with radon levels above action level is between 0.771 and 0.853.

- all homes in Northwest Iowa
- sampled homes in Northwest Iowa
- all homes in Iowa

©zyBooks 03/21/24 21:39 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Inference for two independent proportions

A hypothesis test is also useful for comparing the parameters of two independent populations.

The 2015 study on radon in Northwest Iowa included a total of 621 homes: 351 homes had valid radon results and 270 were incomplete results due to either unreturned radon tests or invalid test results. Home demographics, like whether or not the home has a basement, were collected on all 621 homes. Whether or not a home has a basement is related to radon level in the home. Is the proportion

of homes with a basement the same for both the population of homes with valid results and the population of homes with incomplete results?

PARTICIPATION ACTIVITY 3.6.3: Two-proportion hypothesis test.

Parameters

π_V = proportion of homes with basement for the population of homes with valid results
π_I = proportion of homes with basement for the population of homes with incomplete results

Hypotheses

$H_0 : \pi_V = \pi_I$ or $\pi_V - \pi_I = 0$
$H_a : \pi_V \neq \pi_I$ or $\pi_V - \pi_I \neq 0$

Theory-based assumptions

- Random sample from each population
- Each sample is large enough to include at least 10 "successes" and 10 "failures"

Test statistic

$$z = \frac{\hat{p}_V - \hat{p}_I - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_V} + \frac{\hat{p}(1-\hat{p})}{n_I}}} \qquad z \sim \text{normal}(0,1$$

$$\hat{p} = \frac{count_V + count_I}{n_V + n_I}$$

Results

$\hat{p}_V = \frac{341}{351} \approx 0.972$ $\quad \hat{p}_V - \hat{p}_I \approx 0.060$

$\hat{p}_I = \frac{246}{270} \approx 0.911$ $\quad z \approx 3.280$

$\hat{p} = \frac{341+246}{351+270} \approx 0.945$ $\quad$ p-value ≈ 0.001

Animation content:

Animation shows an example of a two-proportion hypothesis test using the information from the study on radon levels of homes in Northwest Iowa.

- Step 1: A box labeled Parameters appears identifying the two population parameters of interest, π_V = proportion of homes with basement for the population of homes with valid results and π_I = proportion of homes with basement for the population of homes with incomplete results
- Step 2: A box labeled Hypotheses appears identifying the null hypothesis in symbols, $H_0 : \pi_V = \pi_I$ or $pi_V - \pi_I = 0$, and the alternative hypothesis in symbols, $H_a : \pi_V \neq \pi_I$ or $pi_V - \pi_I \neq 0$
- Step 3: A box labeled Theory-based assumptions appears and states the following two assumptions: a random sample from each population, and each sample is large enough to include at least 10 "successes" and 10 "failures".

- Step 4: Box labeled Test statistic appears. The $z = \frac{\hat{p}_V - \hat{p}_I - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_V} + \frac{\hat{p}(1-\hat{p})}{n_I}}}$ test statistic equation is shown and $z \sim$normal(0,1). Below the test statistic equation, $\hat{p} = \frac{count_V + count_I}{n_V + n_I}$ appears and is highlighted with an orange box. The four $\hat{p}$'s in the denominator of the test statistic equation are also highlighted in orange.

- Step 5: Box labeled Results appears with the sample proportions, $\hat{p}_V = \frac{341}{351} \approx 0.972$ and $\hat{p}_I = \frac{246}{270} \approx 0.911$, and difference in sample proportions, $\hat{p}_V - \hat{p}_I \approx 0.060$. The overall proportion then appears $\hat{p} = \frac{341+246}{351+270} \approx 0.945$ in the results box and then the test statistic $z \approx 3.280$ and p-value, p-value≈ 0.001 appears highlighted with a gray box to indicate the hypothesis test results.

Animation captions:

1. A hypothesis test can be used to determine whether the proportion of homes with basements is the same for the two populations: homes with valid test results and homes with incomplete test results.
2. The null hypothesis can be stated as either the two proportions are equal, or the difference between the two proportions is 0. The alternative hypothesis is the not-equal explanation.
3. Theory-based inference methods for two population proportions require a large enough random sample from each population for the Central Limit Theorem to apply and the test statistic to be normally distributed.
4. To calculate a z test statistic assuming a true null hypothesis, the sampling distribution of the difference in proportions has a mean equal to 0, and the standard deviation is estimated using the pooled proportion, $\hat{p}$.
5. The observed difference in proportions of homes with basements is $\hat{p}_V - \hat{p}_I \approx 0.060$. At an $\alpha = 0.05$ level, the null hypothesis is rejected because p-value $\approx 0.001 \leq 0.05$.

PARTICIPATION ACTIVITY 3.6.4: One-sided vs. two-sided hypothesis test for two proportions.

Homes without basements may be associated with lower radon levels, and an incomplete result may be caused by an undetectable level of radon. Suppose instead the researchers are specifically concerned about whether or not the population of homes with valid results have a higher proportion of homes with basements compared to the population of homes with incomplete results. Use this additional information to answer the following questions.

1) In symbols, the one-sided alternative hypothesis is _____.

- $H_a : \pi_V < \pi_I$
- $H_a : \pi_V > \pi_I$
- $H_a : \pi_V \neq \pi_I$

2) The value of the z test statistic for the one-sided hypothesis test is _____ 3.280, the value of the z test statistic for the two-sided hypothesis test.

- less than
- equal to
- greater than

3) The p-value for the one-sided hypothesis test is _____ 0.001, the p-value of the two-sided hypothesis test.

- less than
- equal to
- greater than

4) The p-value is less than α = 0.05 for the one-sided hypothesis test, so the data suggest the proportion of homes with basements is _____.

- lower for the population of homes with valid results compared to the population of homes with incomplete results
- the same for the population of homes with valid results and the population of homes with incomplete results.
- higher for the population of homes with valid results compared to the population of homes with incomplete results

©zyBooks 03/21/24 21:39 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Statistical vs. practical difference

When a hypothesis test results in a small p-value and the null hypothesis is rejected, the data provide statistical evidence of the parameter(s) being different from the null hypothesis claim. However, a statistical difference does not imply a practical difference. Whether or not the difference is enough to be meaningful in real life should also be considered. Ex: The researchers may consider the observed difference of 0.06 in the proportion of homes with basements between the valid and the incomplete results not large enough to meaningfully impact the project.

Inference for one mean

For a numerical characteristic of a population, typically the population mean is of interest. Theory-based inference methods for a mean are similar to the methods for a proportion, but the statistics tend to follow the t-distribution rather than the standard normal distribution.

A primary source for fuel economy information in the United States, fueleconomy.gov provides the US Environmental Protection Agency (EPA) claimed miles per gallon (mpg) for vehicles and allows users to share gas mileage information on their vehicles. The EPA claims a 2019 Toyota RAV4 Hybrid AWD gets 40 mpg. 14 users who drive 2019 Toyota RAV4 Hybrid AWD vehicles shared their vehicle's miles per gallon. Do the user-reported miles per gallon data provide evidence in support of or against the EPA's 40 mpg claim?

PARTICIPATION ACTIVITY 3.6.5: One mean hypothesis test.

Parameter

μ = mean miles per gallon for all users who drive 2019 Toyota RAV4 Hybrid AWD vehicles

Hypotheses

$H_0 : \mu = 40$ $H_a : \mu \neq 40$

Observed data

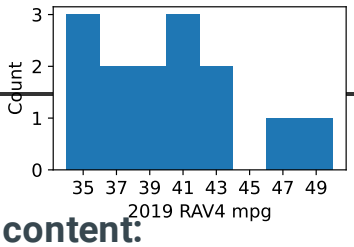
Sample mean: $\overline{x} \approx 40.04$

Sample standard deviation: $s \approx 4.33$

Test statistic

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} \qquad t \sim t(df = n - 1)$$

$$t = \frac{40.04 - 40}{\frac{4.33}{\sqrt{}}} \qquad t \sim t(df = 13)$$



$t \approx 0.03$ 14

Animation content:

Animation shows an example of a one mean hypothesis test using the information about the 2019 Toyota RAV4 Hybrid AWD.

- Step 1: A box labeled Parameter appears identifying the population parameter of interest, μ= mean miles per gallon for all users who drive 2019 toyota RAV4 AWD vehicles.
- Step 2: A box labeled Hypotheses appears identifying the null hypothesis in symbols, $H_0 : \mu = 40$, and the alternative hypothesis in symbols, $H_a : \mu \neq 40$.
- Step 3: Box labeled Observed data appears and a histogram of the user reported data appears. The horizontal axis is labeled 2019 RAV4 miles per gallon and ranges from 34 to 50. The vertical axis is labeled Count and ranges from 0 to 3.
  Histogram bins and counts are given in the table below. Above the graph the sample mean: $\bar{x} \approx 40.04$ and sample standard deviation: $s \approx 4.33$ are provided.

mpg bin	[34-36)	[36-38)	[38-40)	[40-42)	[42-44)	[44-46)	[46-48)	[48-50)
count	3	2	2	3	2	0	1	1

- Step 4: Box labeled Test statistic appears. The equation for the test statistic $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ is given and then a blank test statistic equation appears. The blank equation is filled in with values of 40.04 and 4.33 from the observed data and the value of 40 from the null hypothesis to complete the equation $t = \frac{40.04 - 40}{\frac{4.33}{\sqrt{14}}} \approx 0.03$.
- Step 5: The distribution of the t test statistic is given as $t \sim t(df = n - 1)$, and specifically for the miles per gallon example, $t \sim t(df = 13)$.

Animation captions:

1. Miles per gallon is a numerical measurement and the parameter of interest is the mean miles per gallon, μ, for all users who drive 2019 Toyota RAV4 Hybrid AWD vehicles.

2. The null hypothesis states that the mean miles per gallon for all users who drive this vehicle is equal to the EPA claimed miles per gallon of 40. The alternative hypothesis states the mean is different from 40.
3. The 14 user-reported miles per gallon for 2019 Toyota RAV4 Hybrid AWD vehicles have a mean of 40.04 and standard deviation of 4.33.
4. The standardized t test statistic's calculation assumes the null hypothesis is true. The standard deviation of the sample mean's sampling distribution is estimated by $\frac{s}{\sqrt{n}}$.
5. The t statistic will follow a t-distribution. The t-distribution accounts for additional sampling variability due to estimating the unknown population standard deviation with the sample standard deviation.

PARTICIPATION ACTIVITY 3.6.6: One mean hypothesis test.

Consider the example in the animation above investigating whether the user-reported miles per gallon data provide evidence in support of or against the EPA's 40 mpg claim.

1) The true value of the population parameter, μ, is _____.
   - ○ 40
   - ○ 40.04
   - ○ unknown

2) For the theory-based method to be valid, either the population distribution needs to be symmetric or the sample needs to be large enough, at least $n \approx 20$, and not strongly skewed. Is this condition for the theory-based method reasonably met for the RAV4 miles per gallon example?

- Yes, the sample size is small but the distribution of the sample does not indicate the sample came from a skewed population distribution.
- Yes, because the sample size is large enough and the distribution of the sample is not strongly skewed.
- No, the sample size is small and the distribution of the sample indicates the sample came from a skewed population distribution.

3) The p-value of 0.976 was found using the _____ distribution.

- normal(0,1)
- t(0,1)
- t(13)

4) The user-reported data _____ 40 as the mean miles per gallon for all users who drive 2019 Toyota RAV4 hybrid AWD vehicles.

- provide statistical evidence against
- do not provide statistical evidence against
- prove

©zyBooks 03/21/24 21:39 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Inference for two independent means

Eight users reported the miles per gallon for their 2020 Toyota RAV4 Hybrid AWD vehicles on fueleconomy.gov. Based on the two samples of user-reported data, is the mean miles per gallon for all users who drive 2019 Toyota RAV4 Hybrid AWD vehicles the same, or different from, the mean miles per gallon for all users who drive 2020 Toyota RAV4 Hybrid AWD vehicles?

PARTICIPATION ACTIVITY 3.6.7: Comparison of two population means.

Parameters

μ_{2019} = mean miles per gallon for all users who drive 2019 Toyota RAV4 Hybrid AWD vehicles
μ_{2020} = mean miles per gallon for all users who drive 2020 Toyota RAV4 Hybrid AWD vehicles

Hypotheses

$H_0 : \mu_{2019} = \mu_{2020}$ or $\mu_{2019} - \mu_{2020} = 0$
$H_a : \mu_{2019} \neq \mu_{2020}$ or $\mu_{2019} - \mu_{2020} \neq 0$

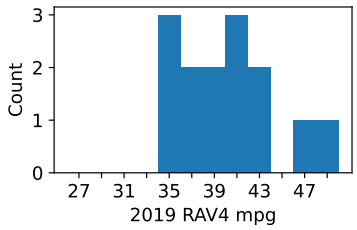
Assumptions

- Random sample from each population
- Symmetric population distributions or large enough sample sizes and not strongly skewed

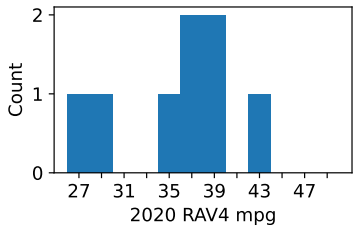
Results

$t \approx 2.11$ p-value ≈ 0.05

Observed data



$n_{2019} = 14$
$\bar{x}_{2019} \approx 40.04$
$s_{2019} \approx 4.33$



$n_{2020} = 8$
$\bar{x}_{2020} \approx 35.64$
$s_{2020} \approx 5.33$

$\bar{x}_{2019} - \bar{x}_{2020} \approx 4.40$

Animation content:

Animation shows an example of a hypothesis test for two means using the information about the 2019 and 2020 Toyota RAV4 Hybrid AWD vehicle miles per gallon.

- Step 1: A box labeled Parameters appears with the following two parameters given: μ_{2019} = mean miles per gallon for all users who drive 2009 Toyota RAV4 AWD hybrid vehicles and μ_{2020} = mean miles per gallon for all users who drive 2020 Toyota RAV4 Hybrid AWD vehicles. A box labeled Hypotheses appears identifying the null hypothesis in symbols, $H_0 : \mu_{2019} = \mu_{2020} or \mu_{2019} - \mu_{2020} = 0$, and the alternative hypothesis in symbols, $H_0 : \mu_{2019} \neq \mu_{2020} or \mu_{2019} - \mu_{2020} \neq 0$.

- Step 2: A box labeled Observed data appears and a histogram of the user reported data for each year appears. The horizontal axes are labeled either 2019 RAV4 miles per gallon or 2020

RAV4 miles per gallon and range from 26 to 50. The vertical axis is labeled Count and ranges from 0 to 3 for the 2019 graph and 0 to 2 for the 2020 graph. The sample statistics for the 2019 data are given where $n_{2019} = 14$, $\bar{x}_{2019} \approx 40.04$, and $s \approx 4.33$. The sample statistics for the 2020 data are given where $n_{2020} = 8$, $\bar{x}_{2020} \approx 35.64$, and $s \approx 5.33$. The difference in sample means, $\bar{x}_{2019} - \bar{x}_{2020} \approx 4.40$ is also given.
Histogram bins and counts are given in the table below.

mpg bin	[26-28)	[28-30)	[30-32)	[32-34)	[34-36)	[36-38)	[38-40)	[40-42)	[42-44)	[44-46)	[46-48)	[48-50)
2019 count	0	0	0	0	3	2	2	3	2	0	1	1
2020 count	1	1	0	0	1	2	2	0	2	0	1	1

- Step 3: A box labeled Assumptions appears and states the following two assumptions: a random sample from each population, and symmetric population distributions or large enough sample sizes and not strongly skewed.

- Step 4: A box labeled Results appears and gives a test statistic of $t \approx 2.11$ and p-value of approximately 0.05.

Animation captions:

1. A hypothesis test for two independent population means can be used to determine whether or not the data provide statistical evidence of a difference in the mean miles per gallon between 2019 and 2020 RAV4s.
2. Graphs and descriptive statistics summarize the samples of miles per gallon for each year. If the mean miles per gallon is the same for both years, how likely would it be to observe a difference of 4.40 or more extreme?
3. With small sample sizes, each sample's miles per gallon distribution should be roughly symmetric with no indication the sample came from a skewed population distribution for the theory-based inference methods to be valid.
4. The t test statistic and the p-value can be obtained using statistical software. The theory-based hypothesis test assumes the t test statistic approximately follows a t-distribution.

PARTICIPATION ACTIVITY 3.6.8: Inference for two means.

Consider the example in the animation above investigating whether the mean miles per gallon for all users who drive 2019 Toyota RAV4 Hybrid AWD vehicles is the same, or different from, the mean miles per gallon for all users who drive 2020 Toyota RAV4 Hybrid AWD vehicles.

1) Based on a p-value of 0.05 and a significance level of α=0.05, the decision is to _____.
   - reject the null hypothesis
   - accept the alternative hypothesis
   - fail to reject the null hypothesis

2) A 95% confidence interval for difference in means, $\mu_{2019} - \mu_{2020}$, is (0.05, 8.75). The confidence interval agrees with the conclusion from the hypothesis test that there is a difference in population means because the _____ the interval.
   - observed sample difference in means of 4.40 is within
   - p-value of 0.05 is within
   - null hypothesis difference of 0 is not within

Inference for proportions in Python

The `statsmodels` library contains a `stats` module that has functions for conducting hypothesis tests and constructing confidence intervals for proportions. The two main functions for inference for proportions are described in the table below. Additional functions and further information about parameters can be found in the statsmodels stats documentation.

Table 3.6.1: Functions for inference about proportions.

Function	Parameters	Description
`proportions_ztest()`	`count`: number/array of successes `nobs`: number/array of	Returns the test statistic and p-value for a hypothesis test based on a normal (z) test.

	observations `value`: value in the null hypothesis `alternative`: type of the alternative hypothesis `prop_var=False`: estimate variance based on sample proportions	`count` and `nobs` take a single value for a one proportion test and an array of values for a two proportion test.
`proportion_confint()`	`count`: number of successes `nobs`: number of observations `alpha`: significance level `method='normal'`: use normal approximation to calculate interval	Returns a (1-`alpha`)∗100% confidence interval for a population proportion.

Inference for proportions in Python.

Full screen

The National Health and Nutrition Examination Survey (NHANES) is conducted every year to survey Americans about their health and nutrition. The dataset includes physical characteristics and behaviors, such as exercise and eating habits.

Another study reported 7% of the US population has been diagnosed with diabetes in 2012. Determine whether the NHANES data provides statistical evidence in support or against the proportion of the US population diagnosed with diabetes in 2012 is 0.07.

The code below uses the NHANES dataset to conduct a single-proportion hypothesis test using the additional study information given above, construct a confidence interval for a single proportion, and conduct a two-proportion hypothesis test comparing the proportion of the US population with diabetes for the 2009-10 and 2011-12 survey years.

©zyBooks 03/21/24 21:39 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

- Modify the code to construct a 90% confidence interval for the proportion of the US population diagnosed with diabetes in 2012.
- Modify the code to test whether the population proportion who own homes is the same or different for the 2009-10 and 2011-12 survey years. In the 2009-10 survey year 3,349 of 4,965 participants owned their home compared to 3,076 of 4,972 participants in the 2011-12 survey year.

In [1]:
```
# Import pandas package and functions from statsmodels
import pandas as pd
from statsmodels.stats.proportion import proportions_ztest
from statsmodels.stats.proportion import proportion_confint
```

In [2]:
```
# Load the dataset
nhanes = pd.read_csv('nhanes.csv')

# View dataset (first/last 5 rows and the first/last 10 columns)
nhanes
```

Out[2]:

	ID	SurveyYr	Gender	Age	AgeDecade	AgeMonths	Race1	Race3	Educa
0	51624	2009_10	male	34	30-39	409.0	White	NaN	Sc
1	51624	2009_10	male	34	30-39	409.0	White	NaN	Sc
2	51624	2009_10	male	34	30-39	409.0	White	NaN	Sc
3	51625	2009_10	male	4	0-9	49.0	Other	NaN	
4	51630	2009_10	female	49	40-49	596.0	White	NaN	S Col
...	...	...	...	...	...	...	...	...	
9995	71909	2011_12	male	28	20-29	NaN	Mexican	Mexican	9 - G
9996	71910	2011_12	female	0	0-9	5.0	White	White	

PARTICIPATION ACTIVITY 3.6.9: Inference for proportions in Python.

Consider the example above and the documentation for the inference Python functions.

1) From the example above, identify the p-value for the two-proportion hypothesis test comparing the proportion of the US population with diabetes for the 2009-10 and 2011-12 survey years.

- -0.5694
- 0.0771
- 0.5691

2) State the null hypothesis corresponding to the code `proportions_ztest(count=722, nobs=3637, value=0.16, alternative='two-sided', prop_var=0.16)`

- $H_0 : \pi = 0.16$
- $H_0 : \pi \neq 0.16$
- $H_0 : \pi = 0.21$

3) In code block 3 of the example, if `prop_var=0.07` is removed, or set to the default `prop_var=False`, the resulting test statistic is _____ the resulting test statistic when `prop_var=0.07` is specified.

- equal to
- not equal to

4) For a hypothesis test with an alternative hypothesis $H_a : \pi < 0.07$, in the `proportions_ztest()` function, specify `alternative=` _____.

- `'smaller'`
- `'left-sided'`
- `'less than'`

Inference for means in Python

The `SciPy` library contains a `stats` module which has functions for conducting hypothesis tests for means. The two functions for testing population means are described in the table below. Additional functions and further information about parameters can be found in the SciPy stats documentation.

Table 3.6.2: Functions for inference about means.

Function	Parameters	Description
`ttest_1samp()`	`a`: array of values `popmean`: value in null hypothesis `alternative`: type of alternative hypothesis	Returns the t-statistic and p-value from a one-sample t-test for the null hypothesis that the population mean of a sample, `a`, is equal to a specified value.
`ttest_ind()`	`a`: array of values from sample 1 `b`: array of values from sample 2 `equal_var=False`: assumes non-equal variances `alternative`: type of alternative hypothesis	Returns the t-statistic and p-value from a two-sample t-test for the null hypothesis that two independent samples, `a` and `b`, have equal population means.

Inference for means in Python.

Full screen

Doctors recommend that adults sleep at least 7 hours per night. The `SleepHrsNight` feature in the NHANES dataset is the self-reported number of hours participants usually get at night for participants aged 16 and older. Determine whether the NHANES data provides statistical evidence that the population mean self-reported number of hours of sleep per night is 7 or whether the population mean is less than 7.

The code below uses the NHANES dataset to conduct a single mean hypothesis test using the additional information given above, construct a confidence interval for a single mean, and conduct a hypothesis test for two independent means comparing the population

mean self-reported number of hours of sleep per night for the 2009-10 and 2011-12 survey years.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Modify the code to find the test statistic and p-value for a hypothesis test comparing the population means of another numerical feature in the NHANES dataset for the 2009-10 and 2011-12 survey years. Possible numerical features to use include:
  - `PhysActiveDays`—the number of days in a typical week participants do moderate or vigorous-intensity activity
  - `BMI`—Body mass index for participants aged 2 or older
  - `DaysPhysHlthBad`—self-reported number of days out of the past 20 days a participant's health was not good
  - `Poverty`—a ratio of family income to poverty guidelines for which smaller numbers indicate more poverty

©zyBooks 03/21/24 21:39 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

In [1]:
```
# Import pandas and numpy packages and functions from scipy.stats
import pandas as pd
import numpy as np
from scipy.stats import ttest_1samp
from scipy.stats import ttest_ind
from scipy.stats import t
```


In [2]:
```
# Load the dataset
nhanes = pd.read_csv('nhanes.csv')

# View dataset (first/last 5 rows and the first/last 10 columns)
nhanes
```

Out[2]:

	ID	SurveyYr	Gender	Age	AgeDecade	AgeMonths	Race1	Race3	Educa
0	51624	2009_10	male	34	30-39	409.0	White	NaN	Sc
1	51624	2009_10	male	34	30-39	409.0	White	NaN	Sc
2	51624	2009_10	male	34	30-39	409.0	White	NaN	Sc
3	51625	2009_10	male	4	0-9	49.0	Other	NaN	
4	51630	2009_10	female	49	40-49	596.0	White	NaN	S Col
...	...	...	...	...	...	...	...	...	
9995	71909	2011_12	male	28	20-29	NaN	Mexican	Mexican	9 -

PARTICIPATION ACTIVITY 3.6.10: Inference for means in Python.

Consider the example above and the documentation for the inference Python functions. Determine whether each of the following statements is True or False.

1) The functions `ttest_1samp()` and `ttest_ind()` both return two values. The first value is the test statistic and the second is the p-value.
   - ○ True
   - ○ False

2) Using `alternative='smaller'` is how to specify a less-than alternative hypothesis in the functions `ttest_1samp()` and `ttest_ind()`.

- True
- False

3) In code cell 7 of the example, specifying `nan_policy='propagate'`, the default, produces different results than `nan_policy='omit'`.

- True
- False

CHALLENGE ACTIVITY 3.6.1: Inference for proportions and means.

537150.4174434.qx3zqy7

Start

A local farm sells eggs classified as "large", requiring the mean weight of the eggs to be 2 oz Agriculture samples 500 eggs to determine if the mean weight is less than 2 oz.

What is the appropriate hypothesis test for this scenario?

Pick

The null hypothesis is that the average egg weight is ____ 2 oz.

Pick

The alternative hypothesis is that the average egg weight is ____ 2 oz.

Pick

1 2

Check Next

CHALLENGE ACTIVITY 3.6.2: Hypothesis tests using SciPy and statsmodels.

537150.4174434.qx3zqy7

Start

A local parks and recreation board is considering building a new set of bike trails. Before building ne at how many riders are using an existing bike trail over a random sample of 90 days. Each day is rat volume" (1 means yes, 0 means no).

- The parks and recreation board believes that less than 45% of days are high volume. Complete hypothesis using `proportions_ztest()`.

The code provided loads the dataset and packages, calculates the number of high volume days, and p-value.

main.py **trails.csv**

```
# Import packages and functions
import pandas as pd
from statsmodels.stats.proportion import proportions_ztest

# Load the dataset
trails = pd.read_csv('trails.csv')

x = trails['highvolume'].value_counts()
n = trails.shape[0]

# Find the test statistic and p-value
test = # Your code goes here

print('Test statistic:', test[0])
print('p-value:', test[1])
```


1 2

Check Next level

(*1) Levy, Barcey T., Cynthia K. Wolff, Paul Niles, Heather Morehead, Yinghui Xu, and Jeanette M. [illegible] "Radon testing: community engagement by a rural family medicine office." *The Journal of the American Board of Family Medicine* 28, no. 5 (2015): 617-623. https://doi.org/10.3122/jabfm.2015.05.140346

3.7 Case study: Flight delays

Learning goals

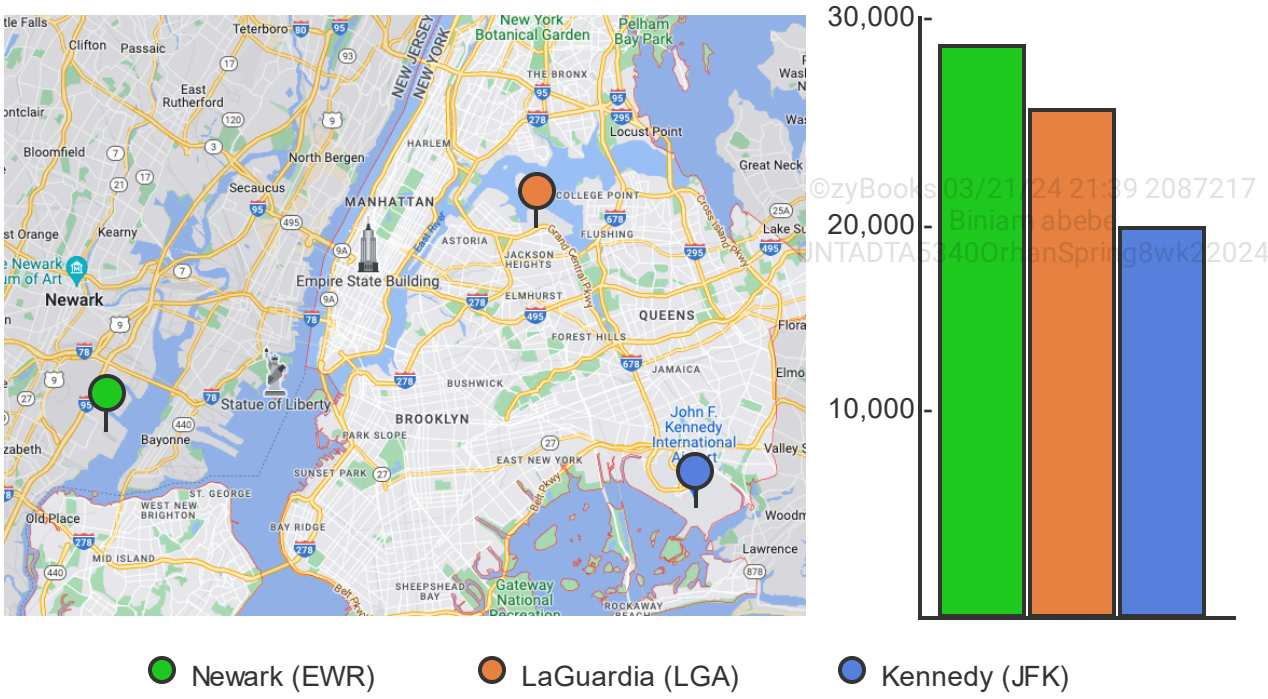
- Identify appropriate hypothesis tests to compare two groups.
- Use Python to compare two groups with an appropriate hypothesis test.
- Interpret the test statistic and p-value.
- Conclude whether a finding is statistically significant or practically significant.

Avoiding flight delays

Flight delays cause stress and inconvenience for thousands of travelers every year. Travelers comparing different airport and flight options may consider factors like price, time, and possibility of delay. If an airport develops a reputation for delayed flights, then that airport may lose millions of dollars in annual revenue and airlines may switch to other nearby airports.

The Port Authority of New York and New Jersey monitors travel from New York City area airports and uses flight data to drive decisions about local airports. The flights dataset contains data on all flights from three major New York City airports (Newark, John F. Kennedy, and LaGuardia). Data includes flight destination, scheduled departure time, flight duration, and whether the flight was delayed.

PARTICIPATION ACTIVITY 3.7.1: New York City airports.



Animation content:

Map of New York City with locations of each airport highlighted. Steps 2-4: A bar chart appears with the number of flights from each airport.

Animation captions:

1. The New York City area has three major international airports: Newark, LaGuardia, and John F. Kennedy.
2. Newark Liberty International Airport (EWR) is between the cities of Newark and Elizabeth in New Jersey. In 2013, a total of 28,344 flights departed from EWR.
3. LaGuardia Airport (LGA) is in northern Queens, near Manhattan. In 2013, a total of 25,178 flights departed from LGA.
4. John F. Kennedy International Airport (JFK) is currently the largest of the New York City area airports. In 2013, a total of 19,212 flights departed from JFK.

Maps Data: Google, © 2022

Data source: Hadley Wickham. "nycflights13: Flights that Departed NYC in 2013", 2021. R package version 1.0.2. https://CRAN.R-project.org/package=nycflights13

PARTICIPATION ACTIVITY 3.7.2: Flight traffic in 2013.

1) Which airport had the most departing flights in 2013?

- EWR
- JFK
- LGA

2) Which airport had the least departing flights in 2013?

- EWR
- JFK
- LGA

Are flight delays more likely at JFK or LaGuardia?

The Port Authority is concerned that a difference exists between the proportion of flights delayed at JFK Airport compared to LaGuardia Airport. Whether or not a flight is delayed is categorical with two possible outcomes: "delay" (`'delay'=1`) or "no delay" (`'delay'=0`). Since the proportion of delays is being compared for two airports, a hypothesis test for two independent proportions is most appropriate.

Comparing two proportions with a hypothesis test.

Full screen

The Python code below imports the flights dataset, and compares the mean duration of flight delays from JFK and LGA with a hypothesis test.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

In [1]:
```
# Import packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.stats.proportion import proportions_ztest
from scipy.stats import ttest_ind

# Load the flights dataset
flights = pd.read_csv('flights.csv').dropna()
flights
```


Out[1]:

	year	month	day	dep_time	sched_dep_time	delay	dep_delay	arr_time	sc
1	2013	1	1	533.0	529	1	4.0	850.0	
4	2013	1	1	554.0	600	0	-6.0	812.0	
7	2013	1	1	557.0	600	0	-3.0	709.0	
9	2013	1	1	558.0	600	0	-2.0	753.0	
14	2013	1	1	559.0	600	0	-1.0	941.0	
...	...	...	...	...	...	...	...	...	
332572	2013	9	26	1141.0	1145	0	-4.0	1457.0	

In [2]:
```
# Define dataframes for each origin airport
```

PARTICIPATION ACTIVITY 3.7.3: Are flight delays more likely at JFK or LaGuardia?

Assume the flights dataset is a representative sample of all flights leaving New York City area airports.

1) Let π_J denote the population proportion of flights delayed at JFK and π_L denote the population proportion of flights delayed at LGA. The null hypothesis for comparing the proportion of flights delayed at the two airports is $H_0 : \pi_J = \pi_L$. What is the alternative hypothesis?

- $H_a : \pi_J \neq \pi_L$
- $H_a : \pi_J > \pi_L$
- $H_a : \pi_J < \pi_L$

2) Calculate the sample proportion of flights delayed at JFK.

- 0.389
- 0.440
- 0.456

3) The p-value for the hypothesis test is _____.

- 0.00000000000000000000000000178
- 1.78
- 10.86

4) Based on a significance level of $\alpha = 0.05$, the decision is to _____ the null hypothesis.

- reject
- fail to reject

Is there a significant difference in the duration of a delay?

The Port Authority found evidence of a difference in the proportion of delayed flights between JFK and LGA. But, whether or not a flight's departure is delayed is not the only consideration. A traveler might be willing to deal with a 5-minute departure delay, but a 30-minute departure delay is much more inconvenient. How does the average length of departure delay compare at JFK vs. LGA?

Comparing two means with a hypothesis test.

Full screen

The Python code below imports the flights dataset, and compares the average delay length for delayed flights from JFK and LGA with

descriptive statistics and a hypothesis test.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

In [1]:
```
# Import packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.stats.proportion import proportions_ztest
from scipy.stats import ttest_ind

# Load the flights dataset
flights = pd.read_csv('flights.csv').dropna()
flights
```


Out[1]:

	year	month	day	dep_time	sched_dep_time	delay	dep_delay	arr_time	sc
1	2013	1	1	533.0	529	1	4.0	850.0	
4	2013	1	1	554.0	600	0	-6.0	812.0	
7	2013	1	1	557.0	600	0	-3.0	709.0	
9	2013	1	1	558.0	600	0	-2.0	753.0	
14	2013	1	1	559.0	600	0	-1.0	941.0	
...	...	...	...	...	...	...	...	...	
332572	2013	9	26	1141.0	1145	0	-4.0	1457.0	
332575	2013	9	26	1151.0	1150	1	1.0	1429.0	

PARTICIPATION ACTIVITY 3.7.4: Are delays longer at JFK or LaGuardia?

Assume the flights dataset is a representative sample of all flights leaving New York City airports.

1) The sample average departure delay for all flights departing JFK is _____ minutes.

- 9.88
- 11.04
- 30.94

2) The sample median departure delay for all flights departing JFK is _____ minutes.

- -1
- 0
- 13

3) Let μ_J denote the population mean departure delay for all flights departing JFK and μ_L denote the population mean delay for all flights departing LGA. The alternative hypothesis for comparing the mean departure delay at the two airports is $H_a : \mu_J \neq \mu_L$. What is the null hypothesis?

- $H_0 : \mu_J < \mu_L$
- $H_0 : \mu_J > \mu_L$
- $H_0 : \mu_J = \mu_L$

4) The p-value for the hypothesis test is 0.001. Based on a significance level of $\alpha = 0.5$, the conclusion of this hypothesis test is that the mean departure delay for flights from JFK is _____ the mean departure delay for flights from LGA.

- different from
- greater than
- equal to

Practical significance

A difference between two groups is ***practically significant*** if the difference is large enough to have a real-life consequence. Based on the previous analysis, the average departure delay at JFK was about 11 minutes, and the average departure delay at LGA was about 10 minutes. The hypothesis test found strong statistical evidence of a difference in the average length of departure delay. However, statistical evidence does not necessarily equate to practical significance.

- Suppose the average flight delay from JFK was 1 minute, and the average flight delay from LGA was 2 minutes. An additional minute on a flight delay is not likely to cause someone to miss an important meeting or a connecting flight, so the difference is not practically significant.
- Suppose the average flight delay from JFK was 1 minute, and the average flight delay from LGA was 31 minutes. An additional 30 minutes on a flight delay is more likely to cause someone to miss an important meeting or a connecting flight, so the difference is practically significant.

Comparing two distributions with a histogram.

Full screen

The Python code below imports the flights dataset, and creates a histogram of flight delays from JFK and LGA.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

In [1]:
```
# Import packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.stats.proportion import proportions_ztest
from scipy.stats import ttest_ind

# Load the flights dataset
flights = pd.read_csv('flights.csv').dropna()
flights
```


Out[1]:

	year	month	day	dep_time	sched_dep_time	delay	dep_delay	arr_time	sc
1	2013	1	1	533.0	529	1	4.0	850.0	
4	2013	1	1	554.0	600	0	-6.0	812.0	
7	2013	1	1	557.0	600	0	-3.0	709.0	
9	2013	1	1	558.0	600	0	-2.0	753.0	
14	2013	1	1	559.0	600	0	-1.0	941.0	
...	...	...	...	...	...	...	...	...	
332572	2013	9	26	1141.0	1145	0	-4.0	1457.0	
332575	2013	9	26	1151.0	1150	1	1.0	1429.0	

PARTICIPATION ACTIVITY 3.7.5: Is the difference in delays practically significant?

The average length of a flight delay at JFK is 11.04 minutes, and the average length of a flight delay at LGA is 9.88 minutes.

1) The difference in the average flight delay at JFK compared to LGA is statistically significant.
   - ○ True
   - ○ False

2) The difference in the average flight delay at JFK compared to LGA is practically significant.

- ○ True
- ○ False

3.8 LAB: Measures of center

The `mtcars` dataset contains data from the 1974 Motor Trends magazine, and includes 10 features of performance and design from a sample of 32 cars.

- Import the csv file `mtcars.csv` as a data frame using a `pandas` module function.
- Find the mean, median, and mode of the column `wt`.
- Print the mean and median.

Ex: for the column `qsec`, the output would be:

```
mean = 17.84875, median = 17.710
```

537150.4174434.qx3zqy7

LAB ACTIVITY 3.8.1: LAB: Measures of center 0 / 1

main.py Load default template...

```
import pandas as pd

# Read in the file mtcars.csv
cars = # Your code here

# Find the mean of the column wt
mean = # Your code here

# Find the median of the column wt
median = # Your code here

print("mean = {:.5f}, median = {:.3f}".format(mean, median))
```

Develop mode | Submit mode

Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

If your code requires input values, provide them here.

Run program

Input (from above) → main.py (Your program) → Outp

Program output displayed here

Coding trail of your work What is this?

```
History of your effort will appear here once you begin
working on this zyLab.
```

3.9 LAB: Calculating probabilities using a normal distribution

The intelligence quotient (IQ) of a randomly selected person follows a normal distribution with a mean of 100 and a standard deviation of 15. Use the `scipy` function `norm` and user input values for `IQ1` and `IQ2` to perform the following tasks:

- Calculate the probability that a randomly selected person will have an IQ less than or equal to `IQ1`.
- Calculate the probability that a randomly selected person will have an IQ between `IQ1` and `IQ2`.

For example, if the input is: