2.1 Introduction to random variables

Random variables

A ***random variable*** is a rule that assigns a number to every outcome in the sample space of an experiment. Ex: In the experiment of a coin toss, a random variable may assign 1 to heads and 0 to tails. A random variable is typically defined using a capital letter, such as $X = 1$ (if the coin toss yielded heads) or 0 (if the coin toss yielded tails).

Random variables can be used to model quantities that can change if an experiment is repeated.

PARTICIPATION ACTIVITY 2.1.1: Defining a random variable.

Animation captions:

1. Flipping a coin is an experiment with the sample space {heads, tails}.
2. The outcome of a coin flip may be defined by X, where $X = 1$ is heads and $X = 0$ is tails.
3. The definition of a random variable can be written as a conditional equation.

PARTICIPATION ACTIVITY 2.1.2: Random variables.

1) A radar gun records a car's speed. A sign displays the text SPEEDING or NOT SPEEDING. The text displayed is a random variable.
   - ○ True
   - ○ False

2) A survey asks a participant to indicate eye color. Possible responses are blue, brown, or green. Y is assigned 1 for blue, 2 for brown, and 3 for green. Y is a random variable.
   - ○ True
   - ○ False

3) A survey asks a participant to indicate age. Y is assigned the value of the

response, which ranges from 1 to 120. Y is a random variable.

- True
- False

4) A person observes cars that drive past. Y is assigned the car's manufacturer, which may be Ford, Chevy, Toyota, or Honda. Y is a random variable.

- True
- False

5) Some airports, like John Wayne Airport in California, measure noise levels when a plane flies over houses near the airport and fine airlines that exceed thresholds. Y is assigned 1 if a plane's measured sound exceeds 100 dB, and 0 otherwise. Y is a random variable.

- True
- False

6) A person wins $\$100$ per dot on the roll of a die. Z is the amount won on a die roll. Z is a random variable.

- True
- False

Discrete and continuous random variables

A random variable may be discrete or continuous. A ***discrete random variable*** can take on a countable number of distinct values like the integers between 0 and 100. Ex: The possible outcomes of a die roll are 1, 2, 3, 4, 5, or 6. A ***continuous random variable*** can take on any value within a range of values like the real numbers between 0 and 1. Ex: The time used by a student to complete a timed 60-minute test is a continuous random variable with possible outcomes in the range $[0, 60]$, such as 30.2 minutes, 33.657 minutes, 59 minutes, or 49.0000001 minutes. Continuous random variables are typically measured, while discrete random variables are typically counted.

Understanding discrete and continuous random variables is important in determining the best model for a situation. Ex: An analyst interested in modeling the average global temperature over time would use a continuous random variable. On the other hand, the number of days a particular city experiences temperatures above 90 degrees is best modeled using a discrete random variable.

PARTICIPATION ACTIVITY 2.1.3: Discrete and continuous random variables.

Specify if the random variable is discrete or continuous.

1) $X =$ number of red cars in a parking lot

- Discrete
- Continuous

2) $D =$ Distance a ball is thrown

- Discrete
- Continuous

3) $W =$ Weight of a watermelon

- Discrete
- Continuous

4) $Z =$ Fraction of games a team wins in a 16-game football season

- Discrete
- Continuous

5) $X = 1$ if a person passes a driver's test and $X = 0$ if the person fails.

- Discrete
- Continuous

6) $B =$ The alcohol concentration in a person's blood

- Discrete
- Continuous

7) $Y =$ The water temperature at different ocean depths

- Discrete
- Continuous

2.2 Properties of discrete probability distributions

The probability mass function of a discrete probability distribution

Random variables have many powerful uses in modeling real world situations. A probability mass function captures useful information about a discrete random variable. A ***probability mass function (pmf)*** assigns the probability that a discrete random variable is exactly equal to some value (typically depicted as a table, plot, or equation). The notation $p(X = x)$ or $p(x)$ is typically used for the pmf of X. The probabilities assigned in a pmf are between 0 and 1, and the total probability must sum to 1.

Example 2.2.1: The probability mass function of a discrete random variable.

An adult can be classified as either underweight, normal weight, overweight, or obese. The discrete random variable W is used to model the weight classification of a randomly selected person from the U.S. population. W is assigned 0 if the person is underweight, 1 if normal weight, 2 if overweight, and 3 if obese.

In recent years, 2% of the adult population of the U.S. is classified as underweight, 27% as normal weight, 33% as overweight, and 38% as obese. Construct a table showing the probability mass function (pmf) of W.

Solution

The pmf of W is $p(W = 0) = 0.02$, $p(W = 1) = 0.27$, $p(W = 2) = 0.33$, and $p(W = 3) = 0.38$.

W	0	1	2	3
$p(W)$	0.02	0.27	0.33	0.38

PARTICIPATION ACTIVITY 2.2.1: Probability mass functions.

An experiment involves randomly selecting one review of the 100 online reviews for Joe's Restaurant. Often potential customers only read one review, so the rating for the one chosen is important. The rating corresponding to the randomly selected review ranges from 1

(worst) to 5 (best) stars. 70 raters gave 5 stars, 20 gave 4 stars, and 10 gave 3 stars for Joe's Restaurant.

1) Is the star rating, labeled S, a discrete or a continuous random variable?

- a discrete random variable
- a continuous random variable

2) What is the probability that the displayed rating is 5 stars?

- 70
- 0.70
- 0.07

3) Does the table below represent an experiment, an outcome, or a probability mass function?

Stars	Probability
5	0.70
4	0.20
3	0.10
2	0.00
1	0.00

- an experiment
- an outcome
- a probability mass function

4) Does the table below represent a probability mass function?

Stars	Observed proportion
5	0.75
4	0.15
3	0.10

2	0.05
1	0.05

- Yes
- No

The cumulative distribution function of a probability distribution

Another function used to describe the probabilities for all possible outcomes of a random variable is a cumulative distribution function. The ***cumulative distribution function (cdf)*** of a discrete random variable is the probability that for any number x, the observed value of the random variable will be at most x or $p(X \leq x)$. Ex: When a fair die is rolled and the value facing up recorded, the cdf describes the probability of getting less than or equal to any value x such that the probability X is less than or equal to 3, that is, $p(X \leq 3) = p(1) + p(2) + p(3)$. The notation $F(x)$ is typically used for the cdf of X. Ex: $F(3) = p(X \leq 3) = \frac{1}{2}$ is read "the probability X is less than or equal to 3 is one half". The cdf always starts at 0 and ends at 1 and never decreases as the value of X increases.

Example 2.2.2: The cumulative distribution function of a discrete random variable.

An adult can be classified as either underweight, normal weight, overweight, or obese. The discrete random variable W is used to model the weight classification of a randomly selected person from the U.S. population. W is assigned 0 if the person is underweight, 1 if normal weight, 2 if overweight, and 3 if obese.

In recent years, 2% of the adult population of the U.S. is classified as underweight, 27% as normal weight, 33% as overweight, and 38% as obese. Construct a table showing the cumulative distribution function (cdf) of W.

Solution

The pmf of W is used to find the cumulative distribution function. The probability that W is less than or equal to 0 is 0.02. Thus, $F(0) = 0.02$. $F(1)$ is the probability a randomly selected person is underweight or normal weight, or either $W = 0$ or $W = 1$. Thus, $F(1) = 0.02 + 0.27 = 0.29$. The rest of the cdf is computed similarly.

W	0	1	2	3
$p(W)$	0.02	0.27	0.33	0.38

$F(W)$	0.02	0.29	0.62	1

PARTICIPATION ACTIVITY 2.2.2: Cumulative distribution function.

Bea's Herbs and Teas offers five teas at a tea tasting. X is the number of teas a customer purchases after the tasting. Based on information from previous tastings, the pmf of X is shown below.

X	0	1	2	3	4	5
$p(X)$	0.05	0.1	0.2	0.15	0.2	0.3

1) What is the cdf value $F(0) = p(X \leq 0)$?

- 0
- 0.05
- Cannot determine from the information provided

2) What is the quantity that provides the probability a customer purchases no more than 3 teas?

- $F(2)$
- $F(3)$
- $1 - F(3)$

3) What is the correct value for $F(3)$?

- 0.15
- 0.35
- 0.5

4) What is $F(6)$?

- 0
- 0.3
- 1

5) What is $F(4) - F(2)$?

- 0

- 0.15
- 0.35

Mean or expected value of a discrete random variable

The ***mean*** or ***expected value*** μ of a discrete random variable X is the sum of the possible values of X multiplied by the probability of the value. The mean is calculated as follows:

$$\mu = E(X) = \sum (x \cdot p(x))$$

The mean is a weighted average of the possible values of X with the probabilities as the weights. The mean represents the average value for the population modeled by the random variable X. Computing the mean does not always make sense for a discrete random variable. Ex: X models the race of a randomly selected voter with 1 denoting white, 2 black and 3 other. The mean of X is 1.63, but reporting the mean race is not useful.

Example 2.2.3: The mean of a discrete random variable.

Grace earns money testing websites. Let X represent Grace's weekly earnings. Given the amount of time Grace has available to work each week, Grace estimates that the probability of earning $\$0$ a week is 20%, of $\$100$ is 70%, and of $\$150$ is 10%. What is the mean of X?

Solution

The possible outcomes of X are $X = 0$ with probability 0.2, $X = 100$ with probability 0.7, and $X = 150$ with probability 0.1. First, the pmf of X is constructed.

X	0	100	150
$p * (X)$	0.2	0.7	0.1

To find the mean, each value of X is multiplied by the corresponding probability, and the products are added.

$$\begin{aligned}\mu &= E(X) \\ &= \sum (X \cdot p(X)) \\ &= 0(0.2) + 100(0.7) + 150(0.1) \\ &= 0 + 70 + 15 \\ &= 85\end{aligned}$$

The expected value of $\$85$ means that the web tester earns $\$85$ per week on average.

PARTICIPATION ACTIVITY 2.2.3: Mean of discrete random variables.

Refer to the example above.

1) Grace decides she needs to earn more money and changes her probabilities. Without knowing the probabilities, which of the following could NOT be the resulting mean?

- ○ $\$100$
- ○ $\$149$
- ○ $\$155$

2) The new probabilities Grace assigns are 0.05 for $X = 0$, 0.75 for $X = 100$, and

0.2 for $X = 150$. What is Grace's new expected earnings?

- $\$100$
- $\$83.333$
- $\$105$

3) If the probabilities remain the same for each of the 3 possible X values, but Grace decides that on weeks she has less time she will still test at least 1 website instead of 0, will the expected value decrease, increase, or remain the same?

- Decrease
- Increase
- Remain the same

4) Grace's friend Mark also begins working. His schedule is less consistent and he believes his probabilities for $X = 0$, $X = 100$, and $X = 150$ will be $\frac{1}{8}$, $\frac{6}{10}$, and $\frac{2}{10}$, respectively. He also believes he will achieve $X = 200$ (twenty sites tested) the remaining 7.5% of the weeks. How does the expected value for Mark compare to the $\$105$ Grace expects?

- Mark's is lower at $\$90$.
- Mark has the same expected value, $\$105$.
- Mark's is higher at $\$120$.

Variance and standard deviation of a discrete random variable

The mean of a discrete random variable is a measure of the center of the distribution. The ***variance*** of a discrete random variable X is a measure of the spread of a distribution. The variance is calculated as follows.

$$\sigma^2 = V(X) = \sum((x - \mu)^2 \cdot p(x))$$

Like the mean, the variance is a weighted average with the probabilities as the weights. The variance measures the average of the squared distance of each possible value of X from the mean.

The standard deviation is another measure of spread. The ***standard deviation*** is the square root of the variance.

$$\sigma = \sqrt{\sigma^2}$$

Example 2.2.4: The mean of a discrete random variable.

Grace earns money testing websites at $\$10$ per site. Let X represent Grace's weekly earnings. Grace estimates the probability of testing 0 sites in a week is 20%, of 10 sites is 70%, and of 15 sites is 10%.

- What is the variance of X?
- What is the standard deviation of X?

Solution

The variance σ^2 is a measure of the spread of the distribution and is calculated as $\sigma^2 = V(X) = \sum((X-\mu)^2 \cdot p(X))$.

The mean was previously calculated as $\mu = 85$. To find σ^2, the squares of the difference of each value of X from μ are found. σ^2 is the sum of the squared differences weighted by the probability of each value of X.

X	μ	$X-\mu$	$(X-\mu)^2$	$p(X)$	$p(X)\cdot(X-\mu)^2$
0	85	-85	7225	0.2	1445
100	85	15	225	0.7	157.5
150	85	65	4225	0.1	422.5

The variance is

$$\begin{aligned}\sigma^2 &= V(X)\\ &= \sum((X-\mu)^2 \cdot p(X))\\ &= 7225(0.2) + 225(0.7) + 4225(0.1)\\ &= 2025\end{aligned}$$

The standard deviation σ is the square root of σ^2, or $\sigma = \sqrt{2025} = 45$.

PARTICIPATION ACTIVITY 2.2.4: Variance and standard deviation of discrete random variables.

1) Which gives the measure of the center of a distribution?
   - ○ the variance
   - ○ the mean
   - ○ the standard deviation

2) Consider the example above. Grace decides she will never have a week with no earnings, but instead increases the $0 weeks to $20. If the probabilities remain the same, will the variance increase, decrease, or remain the same?
   - ○ Increase
   - ○ Decrease
   - ○ Remain the same

3) If Grace earns $20, $100, and $150 weekly, the variance is $1409. Grace receives a guaranteed amount of $10 per week regardless of how much she works. In other words, her earnings are $30, $110, and $160 with the same probabilities. What will happen to the variance?
   - ○ Becomes less than $1409
   - ○ Remains the same at $1409
   - ○ Becomes higher than $1409

Python-Practice 2.2.1: Mean, variance, and standard deviation of a discrete random variable.

To find the mean, variance, and standard deviation of a discrete random variable, the `rv_discrete` class must be imported from the scipy.stats library. Next, a list containing the outcomes in a sample space and a list containing the probabilities of each outcome are defined. The outcome and probability lists are then linked.

```
from scipy.stats import rv_discrete

# Defines a list containing the outcomes in the sample space
x = [0,1,2,3,4,5,6]

# Defines a list containing the probabilities for each outcome
p = [0.1,0.2,0.3,0.1,0.1,0.0,0.2]

# Links the values in x to the probabilities in p
discvar = rv_discrete(values=(x,p))
```


To find the mean, the `.mean()` method is used.

```
# Returns the mean of the discrete random variable
print(discvar.mean())
```

```
2.7
```

To find the variance, the `.var()` method is used.

```
# Returns the variance of the discrete random variable
print(discvar.var())
```

```
3.81
```

To find the standard deviation, the `.std()` method is used.

```
# Returns the standard deviation of the discrete random variable
print(discvar.std())
```

```
1.95192212959
```

2.3 Properties of continuous probability distributions

The probability density function of a continuous random variable

A ***probability density function (pdf)*** describes the relative likelihood of all values for a continuous random variable. Ex: The amount of time for Casey to do his chores is a random variable X, where all values between 1 hour and 2 hours are equally likely. The notation $f(x)$ is typically used for the pdf. For Casey's chores, $f(x) = 1$ for all values of x between 1 and 2 and 0 everywhere else.

The pdf can be written as a function, but often a graphical representation is the most descriptive. The area under portions of the curve given by the pdf provide the probabilities. A pdf must be non-negative and the total area under the curve must be 1. Ex: The probability Casey spends between 1 and 1.5 hours doing chores is the area under the curve $f(x) = 1$ for values of x between 1 and 1.5. Calculus concepts such as integration are often required to find areas.

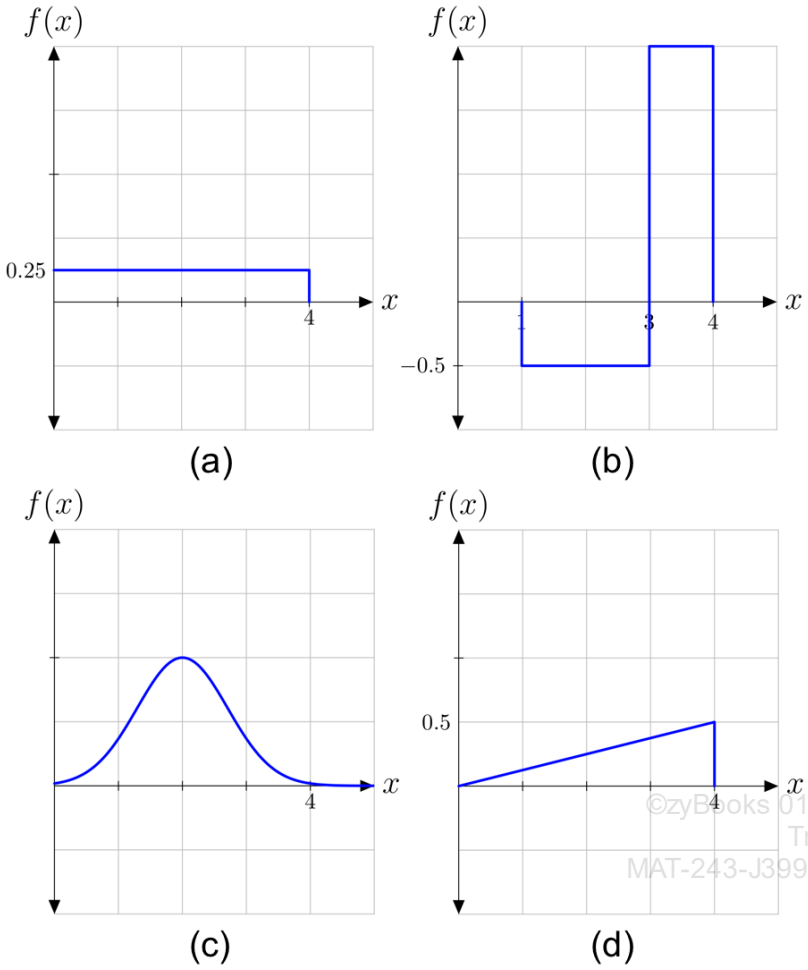
PARTICIPATION ACTIVITY 2.3.1: A probability density function.

Animation captions:

1. The amount of time Casey takes to do chores is a random variable X where all values between 1 hour and 2 hours are equally likely.
2. $f(x) = 1$ for $1 \leq x \leq 2$, and 0 for all other values of x.
3. The area under $f(x)$ between $x = 1$ and $x = 2$ is the probability Casey spends between 1 and 2 hours on chores. The area of the rectangle is $1 \cdot 1 = 1$.

PARTICIPATION ACTIVITY 2.3.2: Probability density functions.

Riley manages an auto repair shop and is modeling the time required for a car service. Four potential pdfs for this situation are shown below. Match each pdf with the best description of the model.



Select the definition that matches each term

1) pdf c

- Not a valid pdf
- Lowest chance of extreme (high or low) service times
- More longer service times than shorter service times
- Same probability for a time from 0 and 1 as 1 and 2

2) pdf b

- Not a valid pdf
- Lowest chance of extreme (high or low) service times
- More longer service times than shorter service times
- Same probability for a time from 0 and 1 as 1 and 2

3) pdf d

- Not a valid pdf
- Lowest chance of extreme (high or low) service times
- More longer service times than shorter service times
- Same probability for a time from 0 and 1 as 1 and 2

4) pdf a

- Not a valid pdf
- Lowest chance of extreme (high or low) service times
- More longer service times than shorter service times
- Same probability for a time from 0 and 1 as 1 and 2

Reset

The cumulative distribution function of a continuous random variable

A ***cumulative distribution function (cdf)*** of a continuous random variable is the probability that for any number x, the observed value of the random variable will be at most x or $p(X \leq x)$. Ex: When Casey does housework, the cdf describes the probability of Casey finishing in time less than or equal to any value x such that the probability X is less than or equal to 1.5. The notation $F(x)$ is typically used for the cdf of X, in contrast to lower-case $f(x)$ for the pdf. Ex: $F(1.5) = P(X \leq 1.5)$ is read "the probability X is less than or equal to 1.5". As with discrete random variables, the cdf always starts at 0 and ends at 1 and never decreases as the value of X increases. The cdf may approach the limits of 0 and 1 in cases where the possible values of x are infinite.

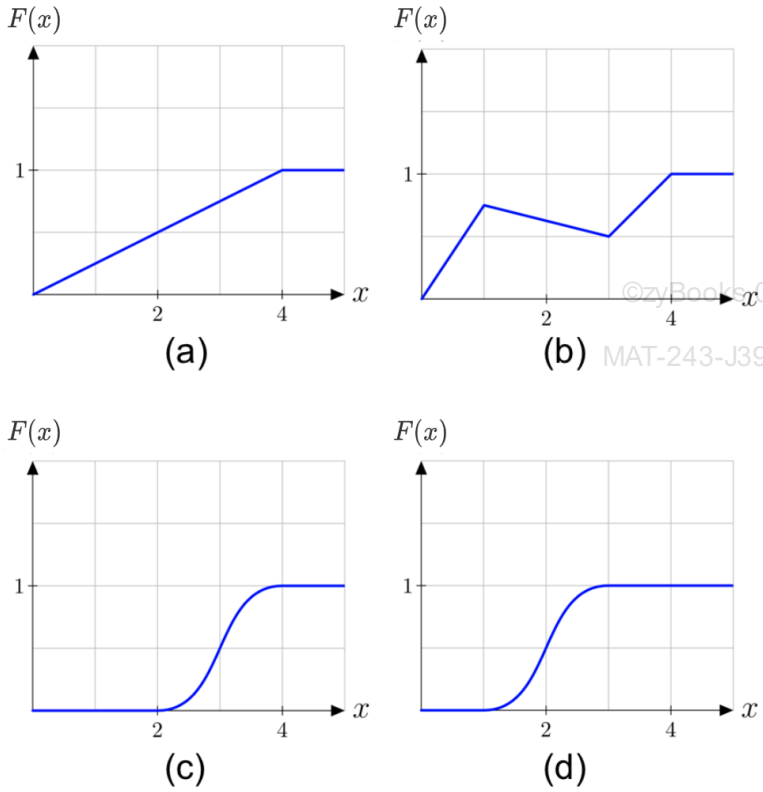
PARTICIPATION ACTIVITY 2.3.3: A cumulative distribution function.

Animation captions:

1. The amount of time Casey requires to do chores is a random variable X where all values between 1 hour and 2 hours are equally likely.
2. A continuous density function $F(x)$ can be constructed from the probability density function $f(x)$.
3. The probability Casey takes 1.5 hours or fewer to complete the chores is 0.5.
4. In the pdf, $f(x) \geq 0$ for all x, and the area under $f(x)$ totals 1. In the cdf, $F(x) = 1$ for all $x > 2$.

PARTICIPATION ACTIVITY 2.3.4: Continuous cumulative distribution function (cdf).

Riley manages an auto repair shop and is modeling the time required for a car service. Four potential cdfs for this situation are shown below.



1) The cdf which is not valid is:

- cdf a
- cdf b
- cdf d

2) If the correct model is cdf c, the probability of a service time less than 2 hours is:

- near 0
- 0.2
- 0.5

3) According to the model depicted in cdf a, the probability of a service time completion in the first hour is _____ the probability the service is completed in the second hour.

- lower than
- the same as
- higher than

4) Jeff takes his car in for service. He hopes to pick the car up no more than 3 hours later. Which cdf should Jeff hope is the correct model?

- cdf a
- cdf c
- cdf d

Mean, variance, and standard deviation of a continuous random variable

In a previous section, the mean, variance, and standard deviation for a discrete random variable were defined. The interpretation of the three measures is similar for continuous random variables.

- The ***mean*** μ or expected value $E(X)$ of a continuous random variable X is a measure of the center of the distribution. The mean is a weighted average of the possible values of the random variable, with the pdf providing the weights. Graphically, the mean is where a pivot is placed so that the pdf balances.
- The ***variance*** σ^2 of a continuous random variable X is a measure of the spread of a distribution. The variance, like the mean, is a weighted average. The variance averages the squared distance of each possible value of X from the mean, with weights provided by the pdf.
- The ***standard deviation*** σ is another measure of the spread of the distribution. The standard deviation is the square root of the variance, $\sigma = \sqrt{\sigma^2}$.

Integral calculus is required to compute the three quantities in the case of continuous random variables. The details are not discussed in this material.

PARTICIPATION ACTIVITY | 2.3.5: The mean, variance, and standard deviation of a continuous random variable.

Animation captions:

1. Let X be the percentage of correct responses for a test. A pdf of the test results can be constructed.
2. The mean is the center of the distribution which balances the density function. The density is greater below 0.5 percent than above.
3. The correct mean percentage is 0.33.
4. The variance and standard deviation measure the spread of the data. A smaller variance and standard deviation correspond to a distribution with values closer to the mean.

PARTICIPATION ACTIVITY | 2.3.6: Mean, variance, and standard deviation of a continuous random variable.

Select the definition that matches each term

1) Mean
   - ○ A measure of the center of a distribution
   - ○ Measure involving the squared differences of X from $E(X)$
   - ○ A measure of spread with units the same as the random variable

2) Variance
   - ○ A measure of the center of a distribution
   - ○ Measure involving the squared differences of X from $E(X)$
   - ○ A measure of spread with units the same as the random variable

3) Standard deviation
   - ○ A measure of the center of a distribution
   - ○ Measure involving the squared differences of X from $E(X)$
   - ○ A measure of spread with units the same as the random variable

Reset

2.4 Normal distribution

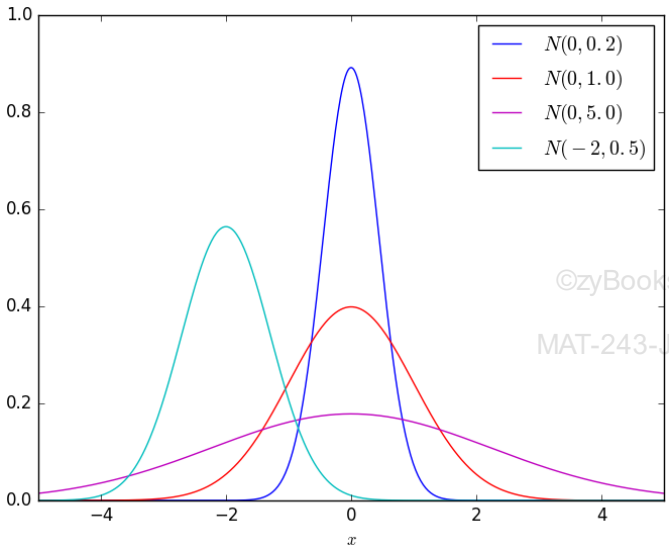
Normal distribution

The ***normal distribution*** is a continuous probability distribution characterized by a bell-shaped probability distribution function and is symmetric around the mean μ. The normal distribution is also referred to as the Gaussian distribution. The normal distribution is pervasive because the distribution is a model for quantities that are computed as sums (totals) or averages. Often data is summarized using either the total or average. The normal distribution also occurs in many settings, including exam scores and heights or other physical measurements.

The normal distribution

- Models: Averages, totals, many natural phenomenon such as measurement errors, test scores, body measurements.
- Notation: $X \sim N(\mu, \sigma^2)$, read "X has a normal distribution with parameters μ and σ^2." The distribution $N(0, 1)$ is known as the standard normal distribution $Z \sim N(0, 1)$.
- Parameters:
  - μ: Mean of the distribution
  - σ: Standard deviation of the distribution
  - σ^2: Variance of the distribution
- Possible values: All real numbers

Figure 2.4.1: Graphs of normal distributions with different values of μ and σ^2.



Empirical rule

Unimodal and symmetric distributions, such as the normal distribution, follow a definite pattern useful for obtaining probabilities and interpreting outcomes. A ***unimodal distribution*** is a distribution with exactly one mode. In such distributions, the mean, median, and mode are equal.

The ***empirical rule*** states that for any unimodal and symmetric distribution: (1) 68% of the data fall within one standard deviation of the mean, (2) 95% of the data fall within two standard deviations of the mean, and (3) 99.7% of the data fall within three standard deviations of the mean. Mathematically,

$$\begin{aligned} 68 \text{ percent of data are located on the interval } [\mu - \sigma, \mu + \sigma] \\ 95 \text{ percent of data are located on the interval } [\mu - 2\sigma, \mu + 2\sigma] \\ 99.7 \text{ percent of data are located on the interval } [\mu - 3\sigma, \mu + 3\sigma] \end{aligned}$$

A rule of thumb is that quantities or observations within two standard deviations of the mean are considered common or usual. Quantities outside of two standard deviations are considered uncommon or unusual.

The empirical rule is illustrated in the animation below.

PARTICIPATION ACTIVITY 2.4.1: Empirical rule.

Animation captions:

1. A distribution that is unimodal and symmetric is centered around the mean, μ.
2. The standard deviation σ determines the spread of the distribution.
3. Approximately 68% of the data is within 1 standard deviation from the mean.
4. Approximately 95% of the data is within 2 standard deviations from the mean.

5. Approximately 99.7% of the data is within 3 standard deviations from the mean.
6. The percentage of data within specific intervals can be obtained as a consequence of the empirical rule.

PARTICIPATION ACTIVITY 2.4.2: Empirical rule.

In 2015, the College Board SAT mathematics score of high school seniors, which follow an approximately normal distribution, had a mean of 511 and a standard deviation of 120[1].

1) What percentage of seniors who took the test in 2015 scored between 271 and 751?
   - 50%
   - 68%
   - 95%

2) What percentage of seniors who took the test in 2015 scored above 631?
   - 16%
   - 68%
   - 84%

3) What percentage of seniors who took the test in 2015 scored between 271 and 511?
   - 68%
   - 2.5%
   - 47.5%

4) Should a score of 770 be considered unusual?
   - Yes
   - No

z-scores

A ***z*-score** is a signed value that indicates the number of standard deviations a quantity is from the mean. A positive z-score indicates that the quantity is above the mean and a negative z-score

indicates that the quantity is below the mean. A z-score with high absolute value implies that the quantity is farther from the mean, and thus more unusual.

The z-score is given by

$$z = \frac{x - \mu}{\sigma}$$

where x is the raw score, μ is the mean, and σ is the standard deviation.

z-scores are particularly important for determining whether a data point is an outlier and for comparing quantities from different unimodal, symmetric distributions.

Example 2.4.1: Earnings surprise.

Earnings per share (eps) is a one of the major indicators of a company's profitability. Of interest, therefore, is the difference between expected earnings and reported earnings of a company. In the fiscal quarter ending April 2016, Walmart had an eps of 0.98, compared to a consensus eps forecast of 0.88[2]. The earnings surprise as a percentage is

$$\frac{0.98 - 0.88}{0.88} \cdot 100\% = 11.36\%$$

The table below lists the earnings surprise percentage of Walmart's competitors in the fiscal quarter ending in August 2016 or earlier.

Company	Earnings surprise (%)
Walmart	11.36
Target	7.89
Kroger	1.96
Whole Foods	0
Costco	−3.12
Supervalu	−9.52

Analysis

To determine whether the eps of Walmart compared to the other companies is unusual, the sample mean and sample standard deviation should be computed. Using Python, these

quantities can be calculated as follows.

```
import pandas as pd

# Create a DataFrame containing the eps data
earnings_surprise = pd.DataFrame([11.36, 7.89, 1.96, 0, -3.12, -9.52])
print(earnings_surprise.mean())
print(earnings_surprise.std())
```

```
0    1.428333
dtype: float64
0    7.526849
dtype: float64
```


Run example

Computing the z-score for Walmart's earnings surprise percentage yields

$$z = \frac{11.36 - 1.43}{7.53} = 1.32$$

Thus, Walmart's earnings surprise percentage is not considered too unusual in the grocery business because the z-score is less than 2.

PARTICIPATION ACTIVITY 2.4.3: Using z-scores to compare data from two different distributions.

According to the United States Department of Health and Human Services, the mean height for Americans is 1.757 m for men and 1.618 m for women[3]. The standard deviation is 0.074 m for men and 0.069 m for women.

1) What z-score corresponds to a man who is 1.853 m tall? Type as: #.###

Check Show answer

2) What z-score corresponds to a woman who is 1.758 m tall? Type as: #.###

Check Show answer

3) Is the man or the woman taller with respect to their gender? Type as: man or woman

Check Show answer

Python-Function 2.4.1: norm.cdf() and norm.sf().

The `norm.cdf()` and `norm.sf()` functions are used to find probabilities related to the normal distribution. The `scipy.stats` library must be imported to use these functions.

`norm.cdf(z, mean, sd)` returns the probability of z being less or equal to than the critical value z for a normal distribution with the specified mean and standard deviation.

```
import scipy.stats as st

# For a normal distribution, if the mean is 0 and
# the standard deviation is 1, what is P(z <= -0.25)?
print(st.norm.cdf(-0.25, 0, 1))

# For a normal distribution, if the mean is 0 and
# the standard deviation is 1, what is P(z <= 1.5)?
print(st.norm.cdf(1.5, 0, 1))
```

```
0.401293674317
0.933192798731
```

`norm.sf(z, mean, sd)` returns the probability of z being greater than or equal to the critical value z for a normal distribution with the specified mean and standard deviation.

```
# For a normal distribution, if the mean is 0 and
# the standard deviation is 1, what is P(z >= -0.25)?
print(st.norm.sf(-0.25, 0, 1))

# For a normal distribution, if the mean is 0 and
# the standard deviation is 1, what is P(z >= 1.5)?
print(st.norm.sf(1.5, 0, 1))
```

```
0.598706325683
0.0668072012689
```

To find the probability between two critical values, the difference between the two probabilities is calculated.

```
# For a normal distribution, if the mean is 0 and
# the standard deviation is 1, what is P(-0.25 <= z <= 1.5)?
print(st.norm.cdf(1.5, 0, 1) - st.norm.cdf(-0.25, 0, 1))

# For a normal distribution, if the mean is 0 and
# the standard deviation is 1, what is P(1.5 <= z <= 2.85)?
print(st.norm.cdf(2.85, 0, 1) - st.norm.cdf(1.5, 0, 1))
```

```
0.531899124414
0.0646212398139
```


Both `norm.cdf()` and `norm.sf()` can also be used for non-standard normal distributions, that is, when the mean is not 0 or the standard deviation is not 1.

```
# For a normal distribution, if the mean is 55 and
# the standard deviation is 7.5, what is P(x <= 62)?
print(st.norm.cdf(62, 55, 7.5))

# For a normal distribution, if the mean is 55 and
# the standard deviation is 7.5, what is P(x >= 51)?
print(st.norm.sf(51, 55, 7.5))

# For a normal distribution, if the mean is 55 and
# the standard deviation is 7.5, what is P(49 <= x <= 60)?
print(st.norm.cdf(60, 55, 7.5) - st.norm.cdf(49, 55, 7.5))
```

```
0.824676055148
0.703098571396
0.53565206387
```


Run example

PARTICIPATION ACTIVITY 2.4.4: The normal distribution.

The GRE (Graduate Record Exam) scores for both verbal and quantitative reasoning are approximately normally distributed and scaled to have a mean of 150 and 8.75 as the standard deviation[4].

1) What is the probability a randomly selected student scored higher than 150 on verbal reasoning?
   - ○ 0.5
   - ○ 0.68
   - ○ 0.95

2) What is the probability the quantitative reasoning score for a randomly selected student is between 132.5 and 167.5?
   - ○ 0.5
   - ○ 0.68
   - ○ 0.95

3) What is the approximate probability of scoring at least 159 or higher on verbal reasoning?
   - ○ 0.16
   - ○ 0.32
   - ○ 0.5

4) What is the probability of scoring between 165 and 170?

- Less than the probability of a score between 150 and 155.
- The same as the probability of a score between 150 and 155.
- Greater than the probability of a score between 150 and 155.

Python-Function 2.4.2: norm.ppf() and norm.isf().

The `norm.ppf()` and `norm.isf()` functions are used to convert percentiles to z-scores. The scipy.stats library must be imported to use these functions.

`norm.ppf(p, mean, sd)` returns the critical z-score for which the probability of z being below that z-score is p, for a normal distribution with the specified mean and standard deviation.

```
import scipy.stats as st

# For a normal distribution, if the mean is 0 and
# the standard deviation is 1, what is z* if P(z < z*) = 0.135?
print(st.norm.ppf(0.135, 0, 1))
```

```
-1.1030625562
```


`norm.isf(p, mean, sd)` returns the critical z-score for which the probability of z being above that z-score is p, for a normal distribution with the specified mean and standard deviation.

```
# For a normal distribution, if the mean is 0 and
# the standard deviation is 1, what is z* if P(z > z*) = 0.405?
print(st.norm.isf(0.405, 0, 1))
```

```
0.240426031142
```

Both `norm.ppf()` and `norm.isf()` can also be used with non-standard normal distributions.

```
# For a normal distribution, if the mean is 55 and
# the standard deviation is 7.5, what is x* if P(x < x*) = 0.8247?
print(st.norm.ppf(0.8247, 55, 7.5))

# For a normal distribution, if the mean is 55 and
# the standard deviation is 7.5, what is x* if P(x > x*) = 0.95?
print(st.norm.isf(0.95, 55, 7.5))
```

```
62.0006958915
42.6635977979
```

Run example

PARTICIPATION ACTIVITY 2.4.5: Finding a z-score given a probability.

The GRE (Graduate Record Exam) scores for both verbal and quantitative reasoning are approximately normally distributed and scaled to have mean 150 with standard deviation of 8.75.

1) Below what score do 40% of all scores fall? Type as: #.###

Check Show answer

2) Above what score do 20% of all scores fall? Type as: #.###

Check Show answer

Sampling distribution for sample means

Often, statisticians look at the distribution of a test statistic, such as the sample mean. Suppose a sample of size n is taken from a population and the sample mean is computed. Repeating this process for multiple samples and creating a relative frequency plot of the obtained test statistic yields a sampling distribution.

The ***sampling distribution*** of the mean, denoted by $\bar{X}$, is the distribution of sample means when taking random samples of the same size. The ***mean of the sample means***, denoted by $\mu_{\bar{X}}$, is the population mean. That is, $\mu_{\bar{X}} = \mu$. The ***standard error*** (SE) is the standard deviation of the sampling distribution, denoted by $\sigma_{\bar{X}}$, when sampling with replacement. That is, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. The standard deviation requires a correction factor when sampling without replacement. This correction factor is $\sqrt{\frac{N-n}{N-1}}$, where N is the population size.

Example 2.4.2: Guessing squash weights.

At a carnival booth, the population is the weights of 5 squash with unknown distribution. To play the game, a child weighs 2 randomly selected squash and uses the sample mean weight of the selected squash to guess the population mean weight.

Squash	A	B	C	D	E
Weight	5	6	7	9	13

a. Find the population mean μ.
b. Find the distribution of the sample means for a sample size of $n = 2$.
c. Find the mean of the sample means for a sample size of $n = 2$.
d. Find the standard deviation of the sample means given $\sigma = 2.83$.

Solution

a. The population mean is $\mu = \frac{5 + 6 + 7 + 9 + 13}{5} = 8$.

b. 10 samples of size $n = 2$ exist.

Sample	Weights	Sample mean $\bar{x}$	Probability
A, B	5, 6	5.5	1/10
A, C	5, 7	6	1/10
A, D	5, 9	7	1/10
A, E	5, 13	9	1/10
B, C	6, 7	6.5	1/10
B, D	6, 9	7.5	1/10
B, E	6, 13	9.5	1/10
C, D	7, 9	8	1/10
C, E	7, 13	10	1/10
D, E	9, 13	11	1/10

The sampling distribution of the mean $\bar{X}$ is

$\bar{x}$	5.5	6	6.5	7	7.5	8	9	9.5	10	11
$P(\bar{x})$	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10

Only when squash C and D are picked would the child guess the population's mean weight correctly. Thus, the child has a 1 in 10 chance of correctly guessing the population mean μ.

c. The mean of the sample means is

$$\frac{5.5 + 6 + 6.5 + 7 + 7.5 + 8 + 9 + 9.5 + 10 + 11}{10} = 8$$

as expected. The mean of the sampling distribution is always equal to the population mean.

d. Because the same squash cannot be chosen twice, the game involves sampling without replacement. Thus, a correction factor is needed when using the formula for

standard deviation.

$$\sigma_{\bar{X}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{5-2}{5-1}} \cdot \frac{2.83}{\sqrt{2}} = 1.733$$

PARTICIPATION ACTIVITY 2.4.6: Guessing squash weights.

Suppose that 4 squash instead of 2 are used to guess the the population mean weight of 5 squash. The possibilities and the corresponding sampling distribution are given below.

Sample	Weights	Sample mean $\bar{x}$	Probability
A, B, C, D	5, 6, 7, 9	6.75	1/5
B, C, D, E	6, 7, 9, 13	8.75	1/5
A, B, D, E	5, 6, 9, 13	8.25	1/5
A, B, C, E	5, 6, 7, 13	7.75	1/5
A, C, D, E	5, 7, 9, 13	8.5	1/5

$\bar{x}$	6.75	8.75	8.25	7.75	8.5
$P(\bar{x})$	1/5	1/5	1/5	1/5	1/5

1) Can the child correctly guess the mean population weight if 4 squash are chosen? Type as: yes or no

Check Show answer

2) What is the mean of the sample means?

Check Show answer

3) What is the correction factor for the standard deviation of the sampling distribution? Type as: #.#

Check Show answer

4) What is the standard deviation of the sampling distribution? Type as: #.###

Check Show answer

Central Limit Theorem

Previously, finding the probabilities for normally distributed random variables was discussed. However, most situations do not involve parameters that are normally distributed. Luckily, the distribution of sample means can be assumed to be approximately normal because of a powerful result of the Central Limit Theorem (CLT). The CLT is the basis for assuming averages and totals follow the normal distribution and underlies many of the tests and results used in data analysis. The CLT states that as the sample size drawn from the population with distribution X becomes larger, the sampling distribution of the means $\bar{X}$ approaches that of a normal distribution $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Thus, the z-score is

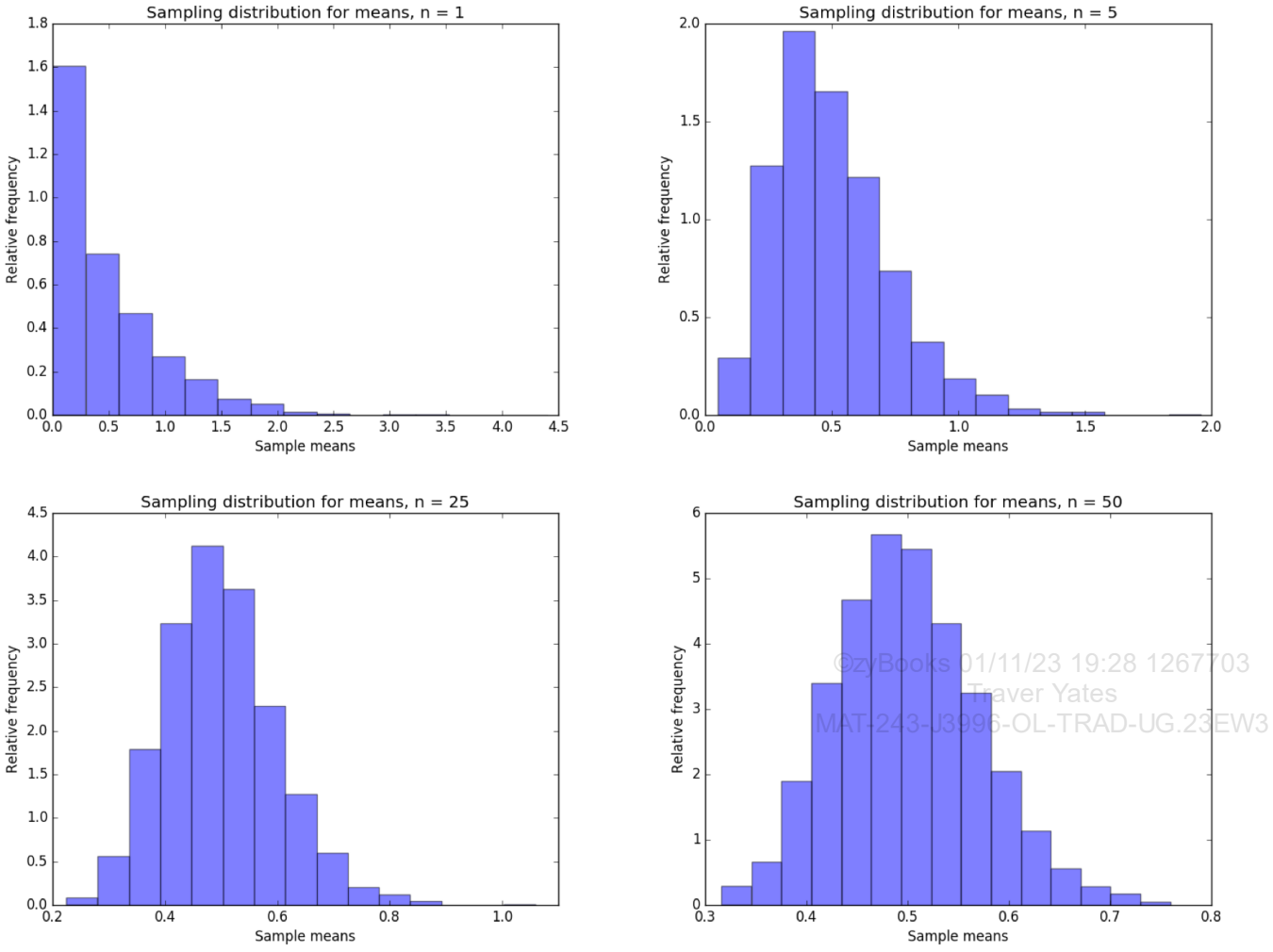
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Central Limit Theorem: Assumptions and conditions

- *Randomness assumption - samples must be randomly selected.*
- *Independence condition - sample values must be independent from each other.*
- *Sample size assumption - sample size must be large enough. A rule of thumb is that sample sizes should be at least 30.*
- *10% condition - sample size must be at most 10% of the population size.*

Example 2.4.3: Sampling distribution for sample means of an exponential distributed random variable.

Consider an exponentially distributed random variable $X \sim \text{Exp}(\lambda = 2)$. 2000 samples of size n are collected. In the figure below, the sampling distribution $\bar{X}$ are shown for $n = 1, 5, 25$, and 50. Although $\bar{X}$ is not normally distributed, the shape of $\bar{X}$ approaches the shape of the normal distribution as n gets larger, which is consistent with the CLT.



Example 2.4.4: Applying the Central Limit Theorem.

An airline studies the mean arrival time of flights from city A to city B. Arrival times (in hours) follow an exponential distribution T with $\mu_T = 3.57$ hours and $\sigma_T = 0.59$ hours. Suppose the study involves 36 randomly selected flights out of the $123,501$ overall number flights in the time period. What is the probability that the mean flight time is greater than 3.8 hours?

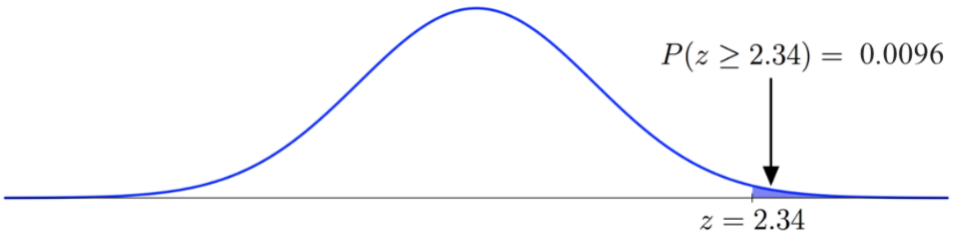
Solution

Although T does not follow a normal distribution, the normal distribution can be used to find the mean flight time because the situation satisfies all conditions for the CLT. Thus, the sampling distribution for $\bar{T}$ approaches the normal distribution $N\left(\mu = \mu_T, \sigma = \frac{\sigma_T}{\sqrt{n}}\right)$. The mean and standard deviation of the sampling distribution are

$$\mu = \mu_T = 3.57 \sigma = \frac{\sigma_T}{\sqrt{n}} = \frac{0.59}{\sqrt{36}} = 0.098$$

The z-score that corresponds to $\bar{T} = 3.8$ is

$$z = \frac{\bar{t} - \mu_T}{\frac{\sigma_T}{\sqrt{n}}} = \frac{\bar{t} - \mu}{\sigma} = \frac{3.8 - 3.57}{0.098} = 2.34$$



Thus, the probability that the mean arrival time is greater than 3.8 hours is $P(\bar{T} \geq 3.8) = P(z \geq 2.34) = 0.0096$. On average, approximately 1 in 104 flights from city A to city B have arrival times of at least 3.8 hours.

PARTICIPATION ACTIVITY 2.4.7: Weights of adult males.

The weights of adult individuals in a certain country are normally distributed with a population mean of $\mu = 172$ pounds and a population standard deviation of $\sigma = 29$ pounds. Suppose $n = 36$ individuals are sampled.

1) What is the mean of the sampling distribution of the means?

Check Show answer

2) What is the standard deviation of the sampling distribution of the means? Type as: #.###

Check Show answer

3) What is the probability that $n = 36$ randomly selected individuals will have a mean weight of at least 180 pounds? Type as: #.###

Check Show answer

Sampling distribution for sample proportion

Sometimes, a study deals with binary categorical variables instead of continuous variables. A ***binary categorical variable*** is a random variable that can only take on two possible names or labels. Ex: A free throw shot results in success or failure. A widget produced by a company is either functional or defective. For samples drawn from a population with a binomial distribution, the Central Limit Theorem for proportions apply.

The Central Limit Theorem for proportions states that if $X \sim B(n, p)$ where n is the number of trials and p is the probability of success, then the sampling distribution for proportions $\hat{p}$ follows a normal distribution $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. Thus, the z-score is

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Central Limit Theorem for proportions: Conditions

- $np \geq 5$
- $n(1 - p) \geq 5$

Example 2.4.5: Applying the Central Limit Theorem for proportions.

A power tools manufacturer reviews the production history for all drill bits produced and found that 8% of the drill bits made are defective. A change in the manufacturer's quality assurance process reduced the percentage of defective drill bits to 5% of the 300 drill bits sampled. If the quality assurance process was not changed, what is the probability that at most 5% of drill bits are defective? The population proportion is $p = 0.08$ and the sample proportion is $\hat{p} = 0.05$.

Solution

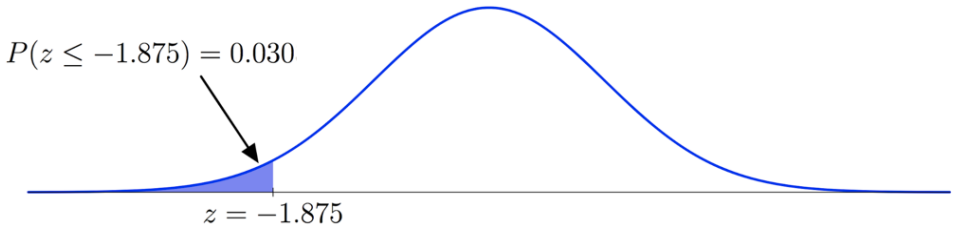
The population proportion is $p = 0.08$. Both $np = 300(0.08) = 24$ and $n(1 - p) = 300(0.92) = 276$ are greater than or equal to 5. Thus, the conditions of the CLT are satisfied and the shape of the binomial distribution approaches that of the normal distribution $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

The standard deviation is

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08(1-0.08)}{300}} = 0.016$$

The corresponding z-score is

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.05 - 0.08}{0.016} = -1.875$$



Thus, the probability that at most 5% of the drill bits are defective is $P(\hat{p} \leq 0.05) = P(z \leq -1.875) = 0.0303$ or 3.03%.

PARTICIPATION ACTIVITY 2.4.8: Nearsightedness.

Nearsightedness affects 8% of children in a certain country. The eyesight of 256 randomly selected children are checked for nearsightedness.

1) What is the mean of the sampling distribution of proportions? Type as #.##

Check Show answer

2) What is the standard deviation of the sampling distribution of proportions? Type as: #.###

Check Show answer

3) What is the probability that $n = 256$ randomly selected children will have a proportion of nearsightedness of at least 9%? Type as #.###

Check Show answer

Calculators

PARTICIPATION ACTIVITY 2.4.9: Probability to Z-score calculator.

PARTICIPATION ACTIVITY 2.4.10: Z-score to probability calculator.

References

(*1) "Total Group Profile Report SAT." *College Board*, 2015, https://research.collegeboard.org/programs/sat/data/archived/cb-seniors-2015.

(*2) "Walmart Inc. Earnings Surprise." *NASDAQ*, www.nasdaq.com/symbol/wmt/earnings-surprise

(*3) US Department of Health and Human Services. "Anthropometric Reference Data for Children and Adults: United States, 2011-2014 ." *Center for Disease Control Vital and Health Statistics*, Series 3 Number 39, August 2016, www.cdc.gov/nchs/data/series/sr_03/sr03_039.pdf

(*4) "Guide to the Use of Scores." *Educational Testing Service Graduate Record Exam*, 2018, www.ets.org/s/gre/pdf/gre_guide.pdf

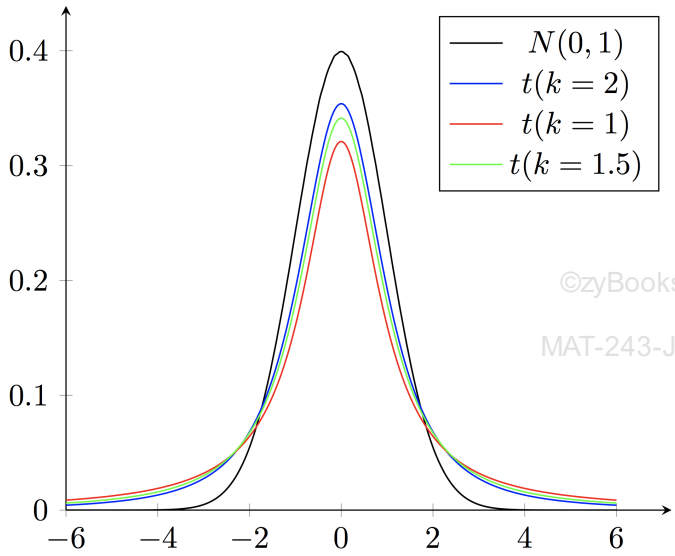
2.5 Student's t-Distribution

t-distribution

The central limit theorem provides a useful tool to calculate probabilities associated with non-normal distributions assuming the sample sizes are sufficiently large. In practice, obtaining a large enough sample may not be possible or the population standard deviation is unknown. In both cases, the sample standard deviation divided by the square of the sample size can be used in place of the population standard deviation.

The ***Student's t-distribution*** or ***t-distribution*** is used in place of the normal distribution in situations where the sample size is too small or the population standard deviation is unknown. The t-distribution has one parameter, degrees of freedom or df, which is equal to $n - 1$. As the sample size n (and consequently df) increases, the t-distribution approaches the normal distribution with a mean of 0 and standard deviation of 1.

Figure 2.5.1: The t-distribution has the same shape as the normal distribution, but has a wider spread because of a slightly larger standard deviation.



The t-statistic is obtained from a sample assumed to have a t-distribution and involves the population mean and a larger variability from estimating the population standard deviation. The same process applies to the computation of probabilities involving the t-distribution as shown in an earlier section with normal distribution. The formula for the t-statistic is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and n is the sample size.

Python-Function 2.5.1: t.cdf() and t.sf().

The `t.cdf()` and `t.sf()` functions are used to find probabilities related to the t-distribution. The scipy.stats library must be imported to use these functions.

`t.cdf(t, df, mean, sd)` returns the probability of t being less than the critical value t for a t-distribution with the specified degrees of freedom, mean, and standard deviation.

```
import scipy.stats as st

# For a t-distribution, if the degrees of freedom is 30, the mean is
0,
# and the standard deviation is 1, what is P(t < -0.25)?
print(st.t.cdf(-0.25, 30, 0, 1))

# For a t-distribution, if the degrees of freedom is 30, the mean is
0,
# and the standard deviation is 1, what is P(t < 1.5)?
print(st.t.cdf(1.5, 30, 0, 1))
```



```
0.40214570454
0.927967035436
```

`t.sf(t, df, mean, sd)` returns the probability of t being greater than the critical value t for a t-distribution with the specified degrees of freedom, mean, and standard deviation.

```
import scipy.stats as st

# For a t-distribution, if the degrees of freedom is 30, the mean is
0,
# and the standard deviation is 1, what is P(t > -0.25)?
print(st.t.sf(-0.25, 30, 0, 1))

# For a t-distribution, if the degrees of freedom is 30, the mean is
0,
# and the standard deviation is 1, what is P(t > 1.5)?
print(st.t.sf(1.5, 30, 0, 1))
```

```
0.59785429546
0.0720329645643
```


To find the probability between two critical values, the difference between the two probabilities is calculated.

```
import scipy.stats as st

# For a t-distribution, if the degrees of freedom is 30, the mean is
0,
# and the standard deviation is 1, what is P(-0.25 < t < 1.5)?
print(st.t.cdf(1.5, 30, 0, 1) - st.t.cdf(-0.25, 30, 0, 1))

# For a t-distribution, if the degrees of freedom is 30, the mean is
0,
# and the standard deviation is 1, what is P(1.5 < t < 2.85)?
print(st.t.cdf(2.85, 30, 0, 1) - st.t.cdf(1.5, 30, 0, 1))
```

```
0.525821330895
0.0681176765133
```

Both `t.cdf()` and `t.sf()` can also be used for t-distributions with different degrees of freedom and when the mean is not 0 or the standard deviation is not 1.

```
import scipy.stats as st

# For a t-distribution, if the degrees of freedom is 59, the mean is
55,
# and the standard deviation is 7.5, what is P(t < 62)?
print(st.t.cdf(62, 59, 55, 7.5))

# For a t-distribution, if the degrees of freedom is 34, the mean is
55,
# and the standard deviation is 7.5, what is P(t > 51)?
print(st.t.sf(51, 34, 55, 7.5))

# For a t-distribution, if the degrees of freedom is 59, the mean is
55,
# and the standard deviation is 7.5, what is P(49 < t < 60)?
print(st.t.cdf(60, 59, 55, 7.5) - st.t.cdf(49, 59, 55, 7.5))
```

```
0.82277404211
0.701363849613
0.532747974268
```


Run example

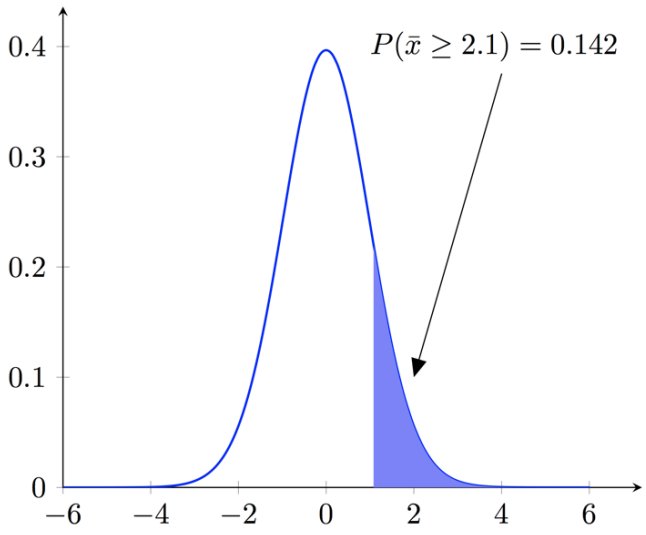
Example 2.5.1: Finding probability using the t-statistic.

The United States Census Bureau determined that the mean number of children in an American household is 1.86. Suppose 50 households are polled and the sample mean is found to be 2.1 and the standard deviation is found to be 1.57. What is the probability of another 50 household sample with a sample mean of at least 2.1?

Solution

The t-statistic is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{50}}} = \frac{2.1 - 1.86}{\frac{1.57}{\sqrt{50}}} \approx 1.081$$



Thus, the probability of another 50 household sample with a sample mean of at least 2.1 is

$$P(\bar{x} \geq 2.1) = P(t \geq 1.081) \approx 0.143$$

PARTICIPATION ACTIVITY 2.5.1: Finding probabilities associated with a t-distribution.

The United States Census Bureau determined that the mean number of children in an American household is 1.86. Polls of 15 households were conducted in a certain city.

1) What is the number of degrees of freedom?

Check Show answer

2) A poll of 15 households had a sample mean of 3.26 children and a sample standard deviation of 2.12 children. What is the t-statistic? Type as: #.###

Check Show answer

3) What is the probability of finding 15 households with a sample mean of 3.26 children or less? Type as: #.###

Check Show answer

4) A different poll of 15 households had a sample mean of 2.72 children and a sample standard deviation of 2.12 children. What is the t-statistic? Type as: #.###

Check Show answer

5) What is the probability of finding 15 households with a sample mean between 2.72 and 3.26 children and a standard deviation of 2.12? Type as: #.###

Check Show answer

Python-Function 2.5.2: t.ppf() and t.isf().

The `t.ppf()` and `t.isf()` functions are used to convert percentiles to t-statistics. The scipy.stats library must be imported to use these functions.

t.ppf(p, df, mean, sd) returns the critical t-statistic for which the probability of t being below that t-score is p, for a t-distribution with the specified degrees of freedom, mean, and standard deviation.


```
import scipy.stats as st

# For a t-distribution, if the degrees of freedom is 49, the mean is 0
and
# the standard deviation is 1, what is t* if P(t < t*) = 0.135?
print(st.t.ppf(0.135, 49, 0, 1))
```

```
-1.11568202664
```

`t.isf(p, df, mean, sd)` returns the critical t-statistic for which the probability of t being above that t-score is p, for a t-distribution with the specified degrees of freedom, mean, and standard deviation.

```
import scipy.stats as st

# For a t-distribution, if the degrees of freedom is 49, the mean is 0
and
# the standard deviation is 1, what is t* if P(t > t*) = 0.405?
print(st.t.isf(0.405, 49, 0, 1))
```

```
0.241727638106
```

Both `t.ppf()` and `t.isf()` can also be used with non-standard t-distributions.

```
import scipy.stats as st

# For a t-distribution, if the degrees of freedom is 24, the mean is 55
and
# the standard deviation is 7.5, what is t* if P(t < t*) = 0.8247?
print(st.t.ppf(0.8247, 24, 55, 7.5))

# For a t-distribution, if the degrees of freedom is 24, the mean is 55
and
# the standard deviation is 7.5, what is t* if P(t > t*) = 0.95?
print(st.t.isf(0.95, 24, 55, 7.5))
```

```
62.1398037924
42.1683844007
```

Run example

PARTICIPATION ACTIVITY 2.5.2: Finding a percentile of a t-distribution.

1) Find the 25th percentile of the t-distribution with 30 degrees of freedom. Type as #.###

Check Show answer

2) Find the 60th percentile of the t-distribution with 4 degrees of freedom. Type as #.###

Check Show answer

2.6 F-distribution

This section has been set as optional by your instructor.

F-distribution

In a sample with multiple groups taken from independent populations, variability can be partitioned into between-group variability and within-group variability. The F-statistic is the ratio of between-group variance to within-group variance.

A number of quantities are involved when calculating the F-statistic.

The ***sum of squares between groups (SSB)*** measures the between-group variability.

$SSB = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2$ where n_i is the number of observations in the ith group, k is the number of groups, $\bar{X}_i$ is the mean of the ith group, and $\bar{\bar{X}}$ is the overall mean.

The ***mean squares between groups (MSB)*** measures the between-group variance.

$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})}{k-1}$ where $k-1$ is the between-group degrees of freedom df_B.

The ***sum of squares within groups (SSW)*** measures the within-group variability. $SSW = \sum_{ij} (X_{ij} - \bar{X})^2$ where X_{ij} is the ith observation of the jth group.

The ***mean squares within groups (MSW)*** measures the within-group variance. $MSW = \frac{\sum_{ij} (X_{ij} - \bar{X})^2}{n - k}$ where n is the overall sample size and $n - k$ is the within-group degrees of freedom df_W.

Since MSB and MSW represent the between-group variance and within-group variance respectively, $F = \frac{MSB}{MSW}$.

The following example shows the calculation of the F-statistic for a simple dataset. In practice, however, the F-statistic is usually calculated using statistical software.

Example 2.6.1: Calculating the F-statistic.

The F-statistic is generally obtained using a statistical package. However, knowing how to compute the F-statistic is useful in illustrating how the associated formulae work. Consider a sample with $k = 3$ groups $X_1 = \{7, 8, 9\}$, $X_2 = \{6, 8, 10, 12\}$, and $X_3 = \{2, 5, 5\}$. Find the F-statistic.

Solution

The mean of each group and the overall mean are obtained.

$$\bar{X}_1 = \frac{7 + 8 + 9}{3} = 8$$
$$\bar{X}_2 = \frac{6 + 8 + 10 + 12}{4} = 9$$
$$\bar{X}_3 = \frac{2 + 5 + 5}{3} = 4$$
$$\bar{X} = \frac{7 + 8 + 9 + 6 + 8 + 10 + 12 + 2 + 5 + 5}{3 + 4 + 3} = 7.2$$

The sum of squares between groups is obtained with $n_1 = 3$, $n_2 = 4$, and $n_3 = 3$.

$$SSB = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2 = 3(8 - 7.2)^2 + 4(9 - 7.2)^2 + 3(4 - 7.2)^2 = 45.6$$

Since the between-group degrees of freedom is $df_B = k - 1 = 3 - 1 = 2$, the mean squares between groups is

$$MSB = \frac{SSB}{df_B} = \frac{45.6}{2} = 22.8$$

To find the sum of squares within groups, a table is helpful.

$(X_{i1} - \bar{X}_1)^2$	$(X_{i2} - \bar{X}_2)^2$	$(X_{i3} - \bar{X}_3)^2$
$(7-8)^2 = 1$	$(6-9)^2 = 9$	$(2-4)^2 = 4$
$(8-8)^2 = 0$	$(8-9)^2 = 1$	$(5-4)^2 = 1$
$(9-8)^2 = 1$	$(10-9)^2 = 1$	$(5-4)^2 = 1$
	$(12-9)^2 = 9$	

Then,

$$SSW = \sum_{ij} (X_{ij} - \bar{X}_j)^2 = (1+0+1) + (9+1+1+9) + (4+1+1) = 28$$

Since the within-group degrees of freedom is $df_W = n - k = (3+4+3) - 3 = 7$, the mean squares within groups is

$$MSW = \frac{SSW}{df_W} = \frac{28}{7} = 4$$

Thus, the F-statistic is $F = \frac{MSB}{MSW} = \frac{22.8}{4} = 5.7$.

PARTICIPATION ACTIVITY 2.6.1: Calculating an F-statistic.

Use the dataset $X_1 = \{0, 2, 4\}, X_2 = \{1, 1, 4\}, X_3 = \{0, 6, 6\}$ to answer the following.

1) What is $\bar{X}_1$?

Check Show answer

2) What is $\bar{X}_2$?

Check Show answer

3) What is $\bar{X}_3$?

Check Show answer

4) What is $\bar{X}$? Type as: #.###

Check Show answer

5) What is df_B?

Check Show answer

6) What is df_W?

Check Show answer

7) Given that $SSB = 8$, what is MSB?

Check Show answer

8) Given that $SSW = 38$, what is MSW? Type as: #.###

Check Show answer

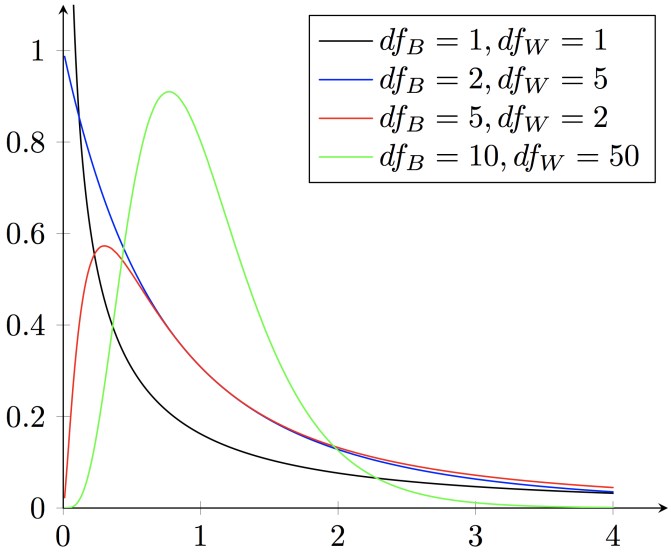
9) What is the F-statistic? Type as: #.###

Check Show answer

Finding probabilities associated with the F-distribution

The F-distribution is a right-skewed distribution, as shown in the figure below. The F distribution has two parameters: the degrees of freedom between samples df_B and the degrees of freedom within samples df_W.

Figure 2.6.1: The F-distribution is a right-skewed distribution.



Python-Function 2.6.1: f.cdf() and f.sf().

The `f.cdf()` and `f.sf()` functions are used to find probabilities related to the F-distribution. The scipy.stats library must be imported to use these functions.

`f.cdf(f, dfb, dfw)` returns the probability of F being less than a critical value f for an F-distribution with df_B equal to dfb and df_W equal to dfw.

```
import scipy.stats as st

# For an F-distribution, if the degrees of freedom between samples is
2
# and the degrees of freedom within samples is 5, what is P(F < 2)?
print(st.f.cdf(2, 2, 5))
```

```
0.769951854167
```


`f.sf(f, dfb, dfw)` returns the probability of F being greater than a critical value f for an F-distribution with df_B equal to dfb and df_W equal to dfw.

```
# For an F-distribution, if the degrees of freedom between samples is
2
# and the degrees of freedom within samples is 5, what is P(F > 3.5)?
print(st.f.sf(3.5, 2, 5))
```

```
0.112065490342
```

To find the probability between two critical values, the difference between the two probabilities is calculated.


```
# For an F-distribution, if the degrees of freedom between samples is
2
# and the degrees of freedom within samples is 5, what is P(2 < F <
3)?
print(st.f.cdf(3, 2, 5) - st.f.cdf(2, 2, 5))
```

```
0.0907506535888
```

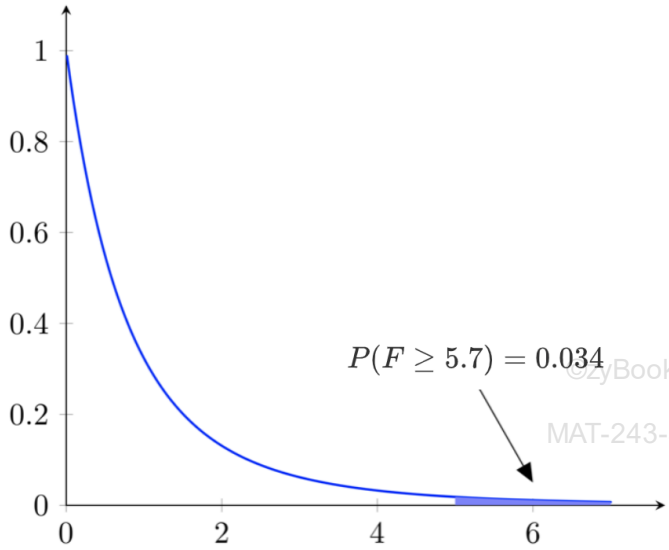
Run example

Example 2.6.2: Finding a probability using an F-score.

For an F-distribution with $df_B = 2$ and $df_W = 7$, what is $P(F \geq 5.7)$?

Solution

For an F-distribution with $df_B = 2$ and $df_W = 7$, the command `f.sf(5.7, 2, 7)` yields $P(F \geq 5.7) = 0.034$.



PARTICIPATION ACTIVITY

2.6.2: Finding probabilities associated with an F-distribution.

1) For an F-distribution with $df_B = 10$ and $df_W = 5$, what is $P(F \leq 3)$?
Type as: #.###

Check Show answer

2) For an F-distribution with $df_B = 2$ and $df_W = 4$, what is $P(F \geq 2)$?
Type as: #.##

Check Show answer

2.7 Chi-square distribution

This section has been set as optional by your instructor.

Shape of the chi-square distribution

The chi-square statistic, denoted χ^2, is related to the standard normal distribution. For a random variable Z with a standard normal distribution, Z^2 has a chi-square distribution with 1 degree of freedom. For independent random variables $Z_1, Z_2, \cdots, Z_k$ where $Z_i \sim N(0, 1)$, $Z_1^2 + Z_2^2 + \cdots + Z_k^2$ has a χ^2 distribution with k degrees of freedom.

If a random sample of size n and standard deviation s is chosen from a standard normal distribution with standard deviation σ, then

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

has a chi-square distribution with $n - 1$ degrees of freedom.The probability density function of the chi-square distribution is given by the formula

$$Y = Y_0(\chi^2)^{\frac{n-1}{2}-1}e^{-\chi^2/2}$$

where Y_0 is a constant that makes the area under the chi-square curve equal to one, n is the size of the sample, and $n-1$ is the degrees of freedom. The shape of the distribution is solely dependent on the degrees of freedom, $n-1$. The mean of the distribution is $\mu = n-1$ and the variance is $\sigma^2 = 2(n-1)$.

PARTICIPATION ACTIVITY 2.7.1: Shape of the chi-square distribution.

Animation content:

undefined

Animation captions:

1. The chi-square distribution is never negative, and therefore is skewed to the right.
2. The shape of the χ^2 distribution is dependent on the degrees of freedom $n-1$. As n gets bigger, the shape of the distribution approaches a normal distribution.

PARTICIPATION ACTIVITY 2.7.2: Characteristics of the chi-square distribution.

1) A sample with $k = 5$ has 5 degrees of freedom.
   - True
   - False

2) A chi-square distribution with 8 degrees of freedom is more symmetric than a distribution with 4 degrees of freedom.
   - True
   - False

3) A chi-square distribution with $k = 9$ has a standard deviation of 4.
   - True
   - False

Example 2.7.1: Finding a χ^2 value.

A company makes LED light bulbs with a stated life expectancy of 50000 hours. The standard deviation is 6000 hours. A customer buys 8 bulbs and finds a standard deviation of 4000 hours. What is the chi-square statistic for this situation?

Solution

The chi-square statistic can be calculated using the formula

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

where $\sigma = 6000$, $s = 4000$, and $n = 8$.

$$\begin{aligned}\chi^2 &= \frac{(n-1)s^2}{\sigma^2} \\ &= \frac{(7)(4000)^2}{(6000)^2} \\ &= 3.111\end{aligned}$$

PARTICIPATION ACTIVITY 2.7.3: Finding χ^2.

The average height for a population of 10 year olds is 54.5 inches with a standard deviation of 1.8 inches. A sample of 10 children is randomly chosen and found to have a standard deviation in height of 2.1. What is the chi-square statistic?

1) What is σ?

Check Show answer

2) What is s?

Check Show answer

3) What is n?

Check Show answer

4) What is χ^2?

Check Show answer

Pearson chi-square test statistic

The chi-square distribution can be used to compare an observed distribution to an expected distribution or two distributions to each other. This comparison is most commonly done using an approximation to the chi-square distribution called the Pearson's chi-square test statistic. The Pearson's chi-square test statistic is defined as

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

where the O_i are the observed counts, E_i are the expected counts, and n is the number of categories. This expression approximates a chi-square distribution with $n - 1$ degrees of freedom.

Example 2.7.2: Using Pearson's chi-square test statistic.

A 10-sided die is thrown 100 times. The frequency with which each face comes up is shown in the table below.

Face	Frequency
1	17
2	10
3	11
4	14
5	6
6	10

7	7
8	11
9	9
10	5

What is the Pearson's chi-square test statistic for determining whether the die is fair?

Solution

If the die is fair, each face should come up with equal frequency. Since the die has 10 faces and is rolled 100 times, the expected counts for each face is $\frac{100}{10} = 10$. Using $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$, the Pearson's chi-square test statistic is

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(17-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(14-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(10-10)^2}{10} \\ &\quad + \frac{(7-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(5-10)^2}{10} \\ &= 11.8\end{aligned}$$

PARTICIPATION ACTIVITY 2.7.4: Throwing dice.

A six-sided die is thrown 90 times. The faces 1, 2, 3, 4, 5, and 6 come up 13, 14, 16, 16, 19, and 12 times, respectively. What is the Pearson's chi-square test statistic for determining if the die is fair?

1) What is the correct expression for the Pearson's chi-square test?

○ $\chi^2 = \frac{13-15}{15} + \frac{14-15}{15} + \frac{16-15}{15} + \frac{16-15}{15} + \frac{19-15}{15} + \frac{12-15}{15}$

- $\chi^2 = \frac{(13-15)^2}{15} + \frac{(14-15)^2}{15} + \frac{(16-15)^2}{15} + \frac{(16-15)^2}{15} + \frac{(19-15)^2}{15} + \frac{(12-15)^2}{15}$

- $\chi^2 = \frac{(13-15)^2}{13} + \frac{(14-15)^2}{14} + \frac{(16-15)^2}{16} + \frac{(16-15)^2}{16} + \frac{(19-15)^2}{19} + \frac{(12-15)^2}{12}$

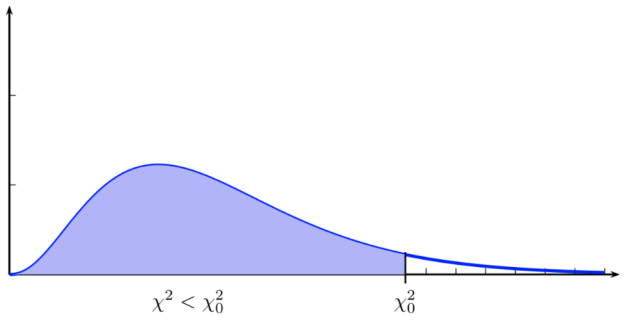
2) What is χ^2?

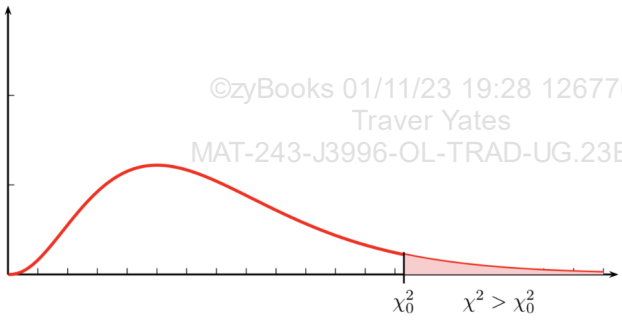
- 0.267
- 31.995
- 2.133

Finding areas and percentiles

The cumulative probability that a chi-square statistic for a distribution with $n-1$ degrees of freedom will have a value less than a particular value χ_0^2 is the area under the chi-square curve between 0 and χ_0^2. Similarly, the cumulative probability that a chi-square statistic for a distribution with $n-1$ degrees of freedom will have a value greater than a particular value χ_0^2 is the area under the chi-square curve between χ_0^2 and ∞, or $(1 - \text{area between } 0 \text{ and } \chi_0^2)$.

Figure 2.7.1: The probability that a chi-square statistic will have a value more or less than χ_0^2 is the shaded area.





Python-Practice 2.7.1: Finding areas and percentiles.

To find the area under a chi-square distribution or the χ_0^2 at which the distribution has a given area, the **chi2** object must be imported from the **scipy.stats** library. Next, the degrees of freedom and χ_0^2 are defined.


```
from scipy.stats import chi2

# Defines the degrees of freedom and chi-square_0
df = 7
x2 =  1.689
```

chi2.cdf can be used to calculate the area under the curve between 0 and χ_0^2.

```
# Calculates the area under the curve between 0 and chi-square_0
area = chi2.cdf(x2, df)

print(area)
```

```
0.02496315095256541
```

Conversely, if the area is defined, **chi2.ppf** gives the χ_0^2 value, or percentile, necessary to obtain that area.

```
# Defines an area under the curve

a = 0.025

# Calculates the percentile
perc = chi2.ppf(a, df)

print(perc)
```

```
1.6898691806773554
```

Run example

Example 2.7.3: Finding probability.

A company makes LED light bulbs with a stated life expectancy of 50000 hours. The standard deviation is 6000 hours. A customer buys 8 bulbs and finds a standard deviation of 4000 hours. If the customer bought 8 more bulbs, what is the probability that the standard deviation of the new sample would be less than 4000 hours?

Solution

As shown in the example above, the chi-square statistic for this situation is $\chi^2 = 3.111$. To find the probability that the standard deviation of a new sample of 8 bulbs would be less than 4000 hours, $\chi_0^2 = 3.111$ and $df = 7$ can be entered into statistical software.

```
from scipy.stats import chi2

df = 7
x2 = 3.111

area = chi2.cdf(x2, df)

print(area)
```

```
0.12545259857860666
```


Thus the probability is 0.125 that the standard deviation of another sample of 8 LED bulbs will be less than 4000 hours.

Run example

Example 2.7.4: Finding percentiles.

What is the 95th percentile of a chi-square distribution with 5 degrees of freedom?

Solution

The 95th percentile is the χ^2 value at which the volume under the chi-square curve is 0.95. This value can be found using statistical software with $a = 0.95$ and $df = 5$.

```
from scipy.stats import chi2

df = 5
a = 0.95

perc = chi2.ppf(a, df)

print(perc)
```

```
11.070497693516351
```

Thus, the χ^2 value that marks the 95th percentile for a chi-square distribution with 5 degrees of freedom is 11.070.

Run example

PARTICIPATION ACTIVITY 2.7.5: Finding probability and percentiles.

The average height for a population of 10 year olds is 54.5 inches with a standard deviation of 1.8 inches. A sample of 10 children is randomly chosen and found to have a standard deviation in height of 2.1 inches. $\chi^2 = 12.25$ for this scenario.

1) What is the probability that another randomly chosen sample of 10 children would have a standard deviation in height of less than 2.1 inches? Type as #.###

Check Show answer

2) What is the probability that another randomly chosen sample of 10 children would have a standard deviation in height of more than 2.1 inches? Type as #.###

Check Show answer

3) What is the 98th percentile for the given χ^2 distribution? Type as #.###

Check Show answer

©zyBooks 01/11/23 19:28 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3