PROBABILITY AND STATISTICS PRIMER

DEFINITIONS

Quantitative: measurable/numeric (e.g., height, weight, age)

Qualitative: categorical (e.g., given a diagnosis)

Random: cannot be predicted in advance because it results from chance factors (e.g., predicting adult height from birth height)

Discrete: has gaps in the values it can assume

Continuous: no gaps in assumed values

Scales of Measurement

Nominal Scale: indicates *class* of the entity. No implied order, one category isn't greater or better than another (e.g., dog = 1, cat = 2).

Ordinal Scale: indicates *qualitative* amount of the entity. Has an implied order or range (e.g., Cancer Stages I, II, III, IV)

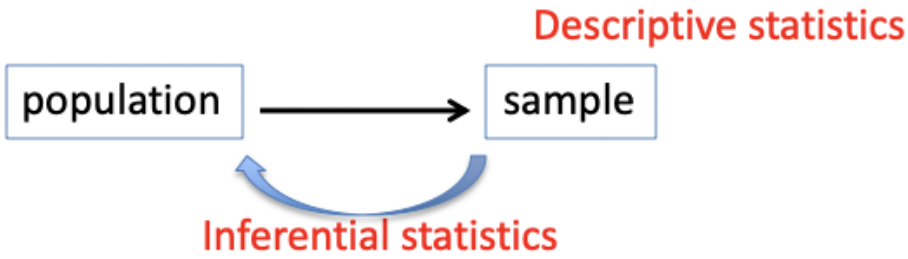
Types of Statistics

Descriptive: Collections, organization, summarization of data. "Here's what it is"

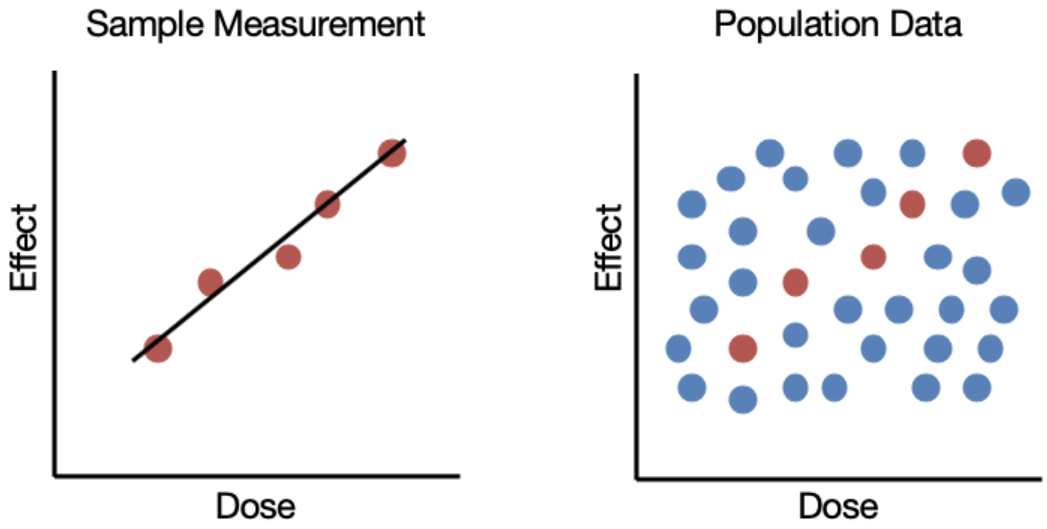
Inferential: Drawing inferences about a body of data when only part of the data is observed. Techniques to make generalizations about a population from a sample

Population vs. Sample Statistics

Population: all subjects of interest in a study whose properties will be analyzed
Sample: a subset of subjects selected from a population



Representative Samples are Critical!
Important to randomly sample with high enough numbers to avoid false positives or negatives



Before you analyze the stats, understand your data.

1. how data was collected
2. validity of collection procedure
3. accuracy of measurements (bias, such as over/underestimates; reliability/precision)
4. validity of measurements
5. all relevant information

DESCRIPTIVE STATISTICS

Range = max - min

Frequency: number of times a value occurs in the data set

Cumulative Frequency: sum of frequencies per class in a table

Relative Frequency: frequency of a given class divided by the total number of observations

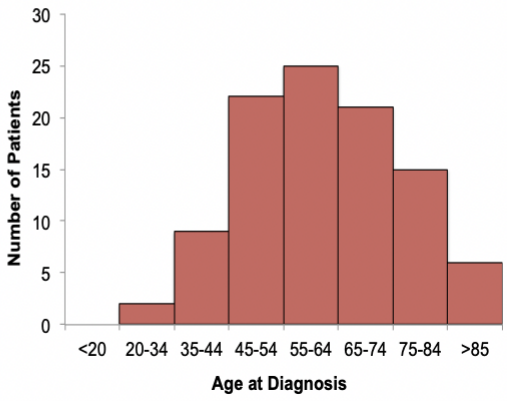
Frequency Distributions

	Range	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
Class	35-44	1	1	0.1	0.1
	45-54	2	3	0.2	0.3
	55-64	2	5	0.2	0.5
	65-74	2	7	0.2	0.7
	75-84	2	9	0.2	0.9
	85-94	1	10	0.1	1.0

Frequency tables summarize data,
while histograms give a nice visual representation

Age	Frequency
<20	0
20-34	2
35-44	9
45-54	22
55-64	25
65-74	21
75-84	15
>84	6

Data approximated from the percentages reported in the SEER Database, 2006-2010, All Races, Female



In R, one can quickly get the histogram of some array (e.g., X) by using the command, hist(X)

MEASURES OF CENTRAL TENDANCY

Population Mean

$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$; $\frac{5+2+6+3+6}{5} = 4.2$

Sample Mean

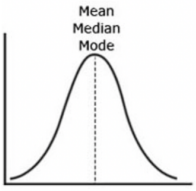
$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{n}$; $\frac{5+2+6+3+6}{5} = 4.2$; sample means are sensitive to outliers

Median

- **If n is odd**: the middle value of an ordered array of observations.
- example: [75, 79, **82**, 85, 90], median = 82
- **if n is even**: the mean of the 2 middle observations in an ordered array
- example: [50, 75, **79, 82**, 85, 90], median = 80.5
- sample median is more robust to outliers than the sample mean

Mode

- most frequently occurring value
- can have 0, 1, or multiple modes
- example: [24, 32, 27, 20, 32, 27, 32], mode = 32



Normal Distribution

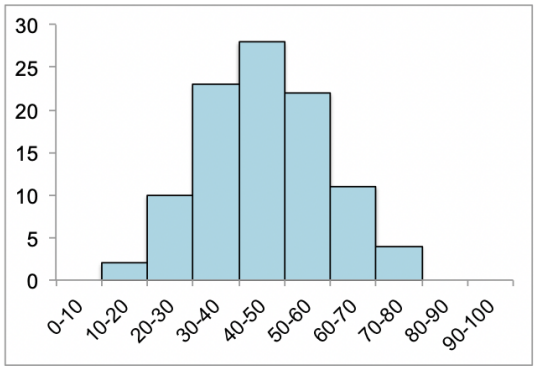
Represents perfectly symmetrical data distribution

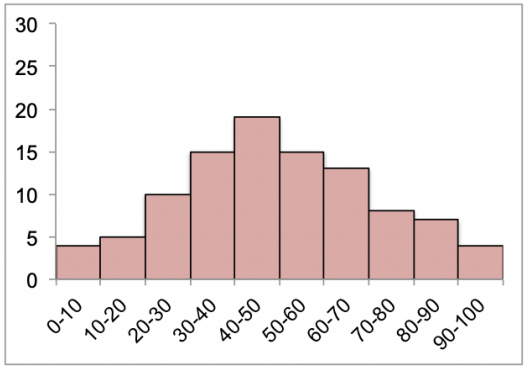
Mean = Median = Mode

Measure:	Mean	Median	Mode
Advantages:	Sensitive to all data	Insensitive to extremes	Is an actual characteristic of several individuals
Disadvantages:	Sensitive to extremes → can be misleading	May ignore useful data → loss of information	Indicates nothing about the rest of the data set (no group characteristics)
When to use?	1. Quantitative variables are not highly skewed 2. Want to report the typical score 3. Anticipate additional statistical analysis	1. Quantitative variables have highly skewed distributions 2. Want to report the central score	1. Variables are categorical 2. Need a quick & easy measure 3. Want to report the most common score

MEASURES OF DISPERSION

Datasets with the same measures of central tendency may have very different variability





Range = max - min

Population Variance (σ^2)

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

- Example for [50, 75, 79, 82, 85, 90] where $\mu = 76.83$ and $N = 6$
- $\frac{(50-76.83)^2+(75-76.83)^2+(79-76.83)^2+(82-76.83)^2+(85-76.83)^2+(90-76.83)^2}{6} = 165.8$

Sample Variance (s^2)

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

- Example for [50, 75, 79, 82, 85, 90] where $\bar{x} = 76.83$ and $n = 6$
- $\frac{(50-76.83)^2+(75-76.83)^2+(79-76.83)^2+(82-76.83)^2+(85-76.83)^2+(90-76.83)^2}{6-1} = 198.97$
- A sample will never have as much variance as a whole population, so dividing by n-1 addresses this problem. Proofs on the internet for those with further interest

Population Standard Deviation (σ)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

Sample Standard Deviation (s)

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{n-1}}$$

- Units for standard deviation the same units as the mean, which is why we report mean $\pm$ standard deviation

Coefficient of Variation (CV) of a Population

$CV = \frac{\sigma}{\mu} \cdot 100$

. CV allows us to compare relative variation rather than absolute variation

. Example: standard deviation in weight amongth older kids might appear higher than that of younger kids just because the numbers are larger, not because they vary more.

Coefficient of Variation (CV) of a Sample

$CV = \frac{s}{\bar{x}} \cdot 100$

Location Parameters: designate positions on the axis when data is graphed (e.g., mean, median, percentiles)

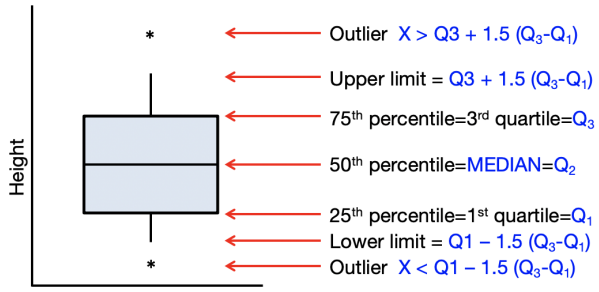
Percentiles

. Given a set of n observations $x_1, x_2, ..., x_n$, the pth percentile (P) is the value of x such that p percent or less of the observations are less than P, and (100-p) percent or less of the observations are greater than P.

. For example a baby with low weight, such that only 5% of all babies have a lower weight, has a 5th percentile weight.

. Quartiles: Q1 = 25th percentile, Q2 = 50th percentile (Median), Q3 = 75th percentile

Box-and-Whisker Plots with Outliers



Find the interquartile range of these 20 numbers:

14.6	31.5
24.3	31.6
24.9	32.3
27	32.8
27.2	33.3
27.4	33.6
28.2	34.3
28.8	36.9
29.9	38.3
30.7	44

Q_1=((n+1)/4)th observation
Q_3=(3(n+1)/4)th observation

Q_1=(21/4)=5.25th observation
Q_3=3(21/4)=15.75th observation

But there aren't a 5.25th or 15.75th observation! What do we do?

Q_1=27.2 + 0.25 * (27.4-27.2) = 27.2 + 0.05 = 27.25
Q_3=33.3 + 0.75 * (33.6-33.3) = 33.3 + 0.225 = 33.525

note: there are other ways to classify outliers.

Interquartile Range (IQR)

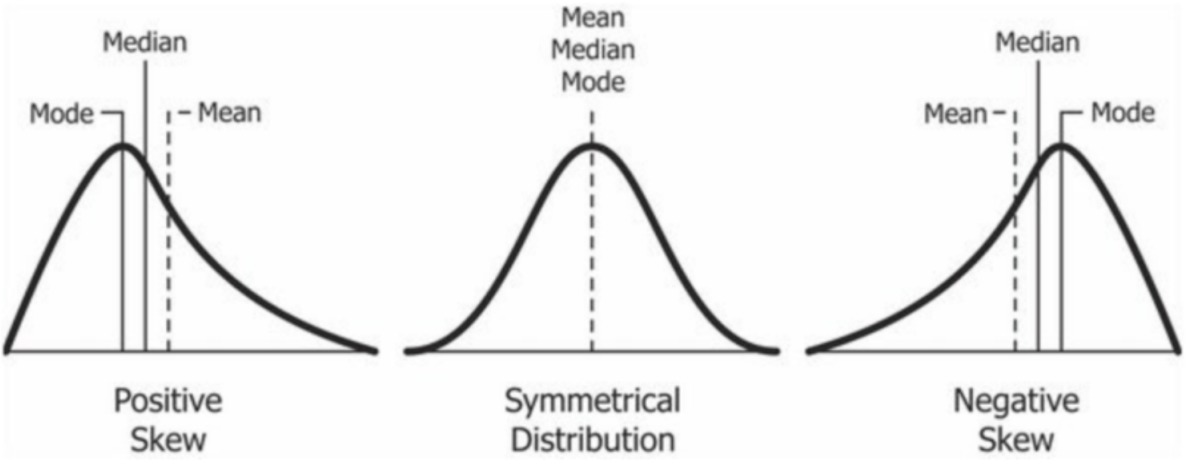
. Q3 - Q1

. Robust Measure of Dispersion

MEASURES OF SHAPE

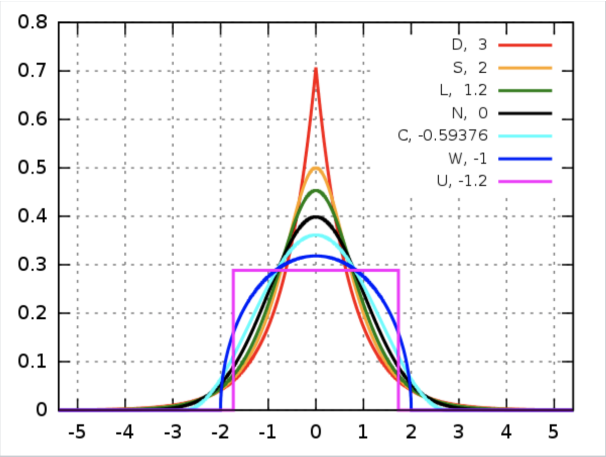
Skew ($\tilde{\mu_3}$)

$$\tilde{\mu_3} = \frac{\sum_{i=1}^{N} (\frac{x_i - \mu^3}{\sigma})}{N}$$



Kurtosis ($\tilde{\mu_4}$)

$$\tilde{\mu_4} = \frac{\sum_{i=1}^{N} (\frac{x_i - \mu^4}{\sigma})}{N}$$



Probability density functions for selected distributions with mean 0, variance 1 and different excess kurtosis

Mesokurtic: distributions with zero excess kurtosis (e.g., Normal distribution)

Leptokurtic: Distributions with positive excess kurtosis. These distributions are 'slender' (lepto) at the peak and have fatter tails (e.g., Laplace Distribution)

Platykurtic: Distributions with negative excess kurtosis. These distributions are 'broad' (platy) at the peak and have thinner tails (e.g., Uniforrm Distribution)

Multimodal Distributions

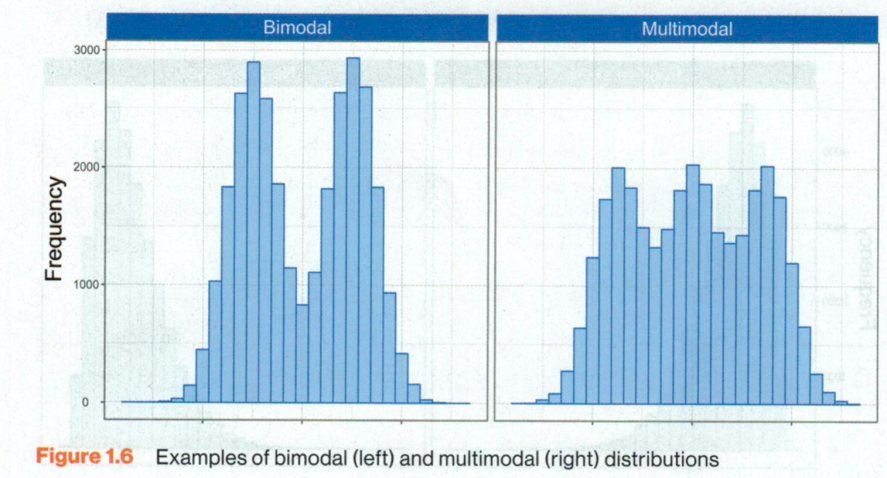


Figure 1.6 Examples of bimodal (left) and multimodal (right) distributions

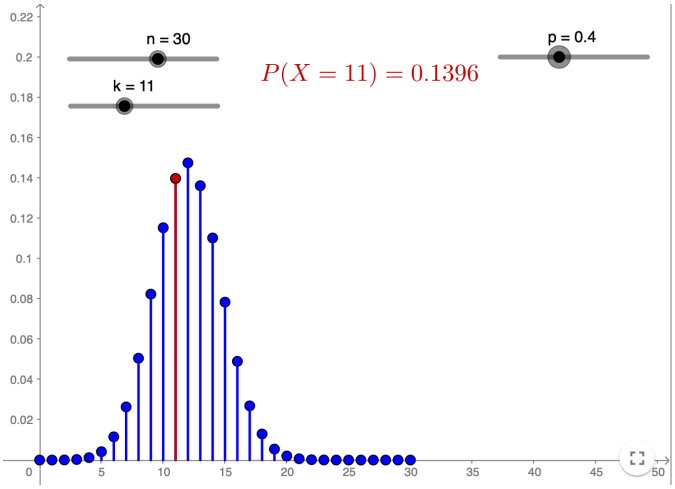
PROBABILITY DISTRIBUTIONS

Random Variable (X): a variable whose values depend on outcomes of a random phenomenon

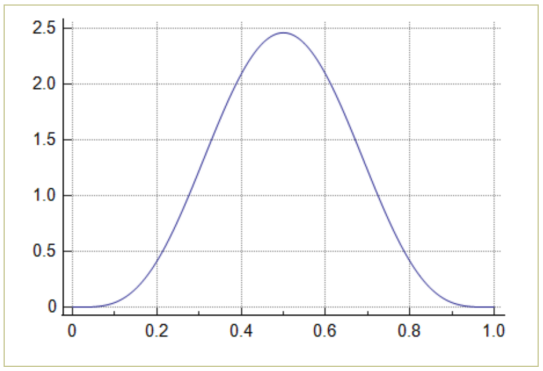
Independent and Identically Distributed (i.i.d.): a collection of random variables that have the same probability distribution as the others and all are mutually independent. For example, the sequence of: a) fair or loaded dice rolls, b) fair or unfair roulette wheels, or c) fair or unfair coin flips are all i.i.d.

Probability Distribution: mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. The sum of a probability distribution should add to 1 (or 100%).

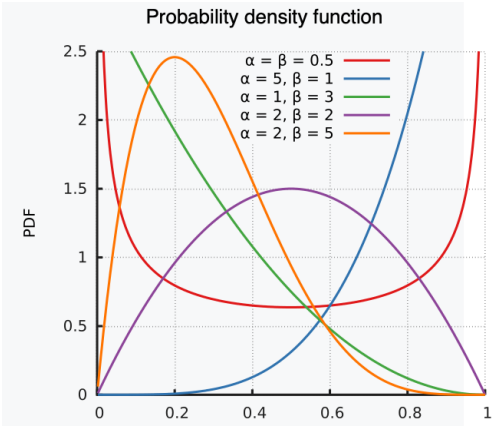
Discrete Probability Distribution (**Probability Mass Function, PMF**): Table or Formula that lists the probabilities for each outcome of the random variable, X, and can take on a countable number of values (e.g., Poisson Distribution, Binomial Distribution)



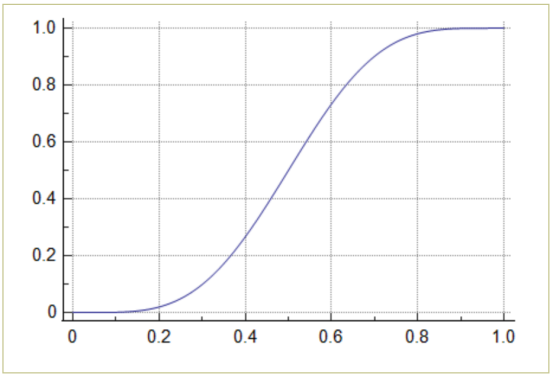
Continuous Probability Distribution (**Probability Density Function, PDF**): The random variable, X, can take on any continuous value (e.g., Normal Distribution, Exponential Distribution, Beta Distribution). The left graph below shows the PDF of a single Beta Distribution. The right graph below shows the PDFs of several Beta Distributions with differing parameters.



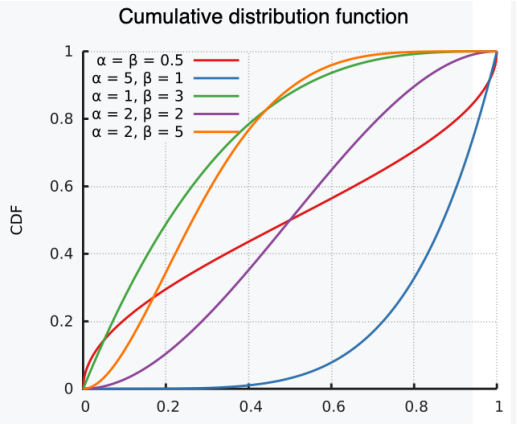
Function: PDFBeta(x ,5 ,5)



Continuous Distribution Function (CDF): The probability that X will take a value less than or equal to x. Applies to any probability distribution. The left graph below shows the CDF of a single Beta Distribution. The right graph below shows the CDFs of several Beta Distributions with differing parameters. Note that if you integrate the entire PDF the CDF is equal to 1.0.



Function: CDFBeta(x ,5 ,5)

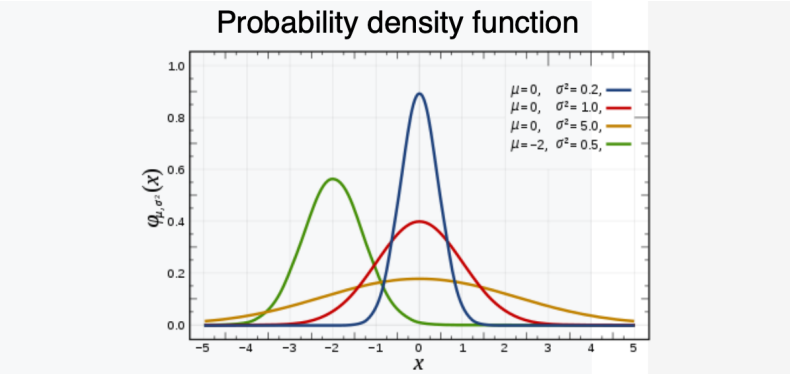


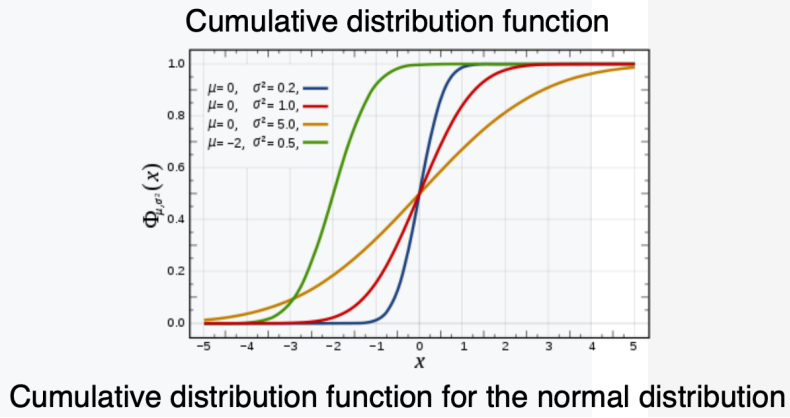
The Normal Distribution

PDF:
$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

CDF:
$\Phi(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
There is no closed form solution for this integral, thus erf (error function) is an approximated portion of this integral.
In R, you can call the error function by: erf(). Type the following in your console:
install.packages("NORMT3")
library(NORMT3)
You now have access to the function, erf().

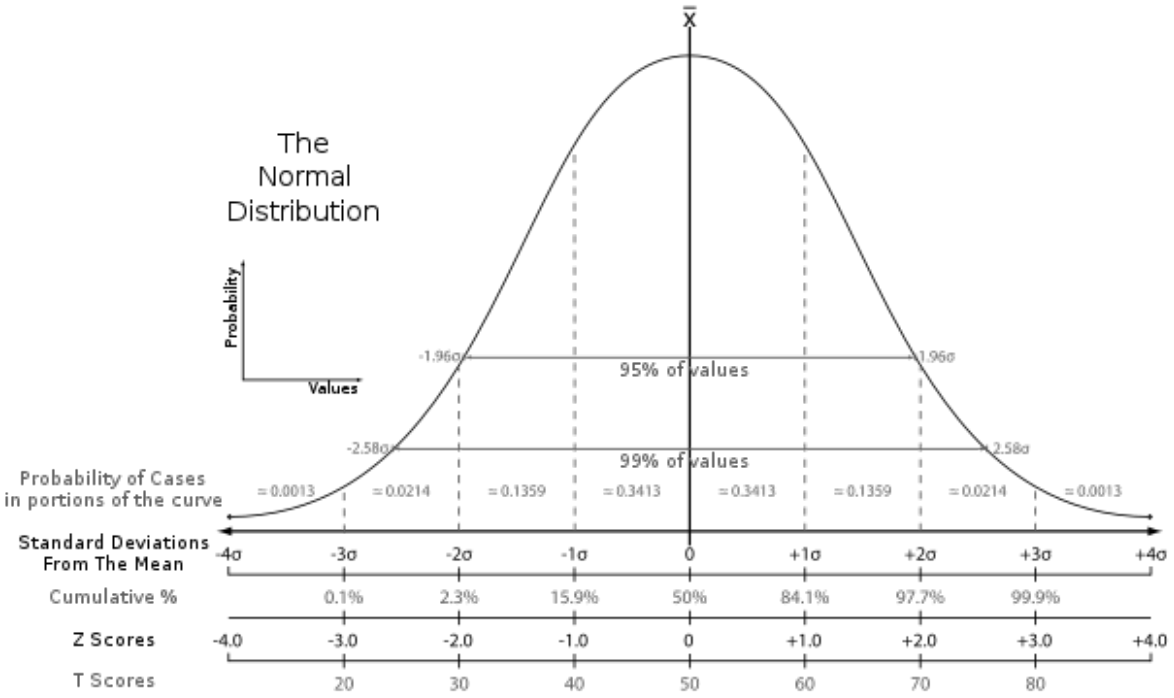
Below is the PDF and CDF of several normal distributions with different parameters, μ, σ.





z-score (or standard score): the number of standard deviations that a particular value is above or below the mean value of observed measurements.
$z = \frac{x-\mu}{\sigma}$ (population z-score)
$z = \frac{x-\bar{x}}{s}$ (sample z-score)

Standard Normal Distribution [i.e., $f(x|\mu = 0, \sigma = 1)$] .



For the figure above, use the CDF equation $\Phi(x|\mu, \sigma) = \left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$ to confirm [see instructions on the page above for calling erf()]:

1. the total area under the curve
2. the probability of being less than 2σ
3. the probability of being greater than 2σ
4. the probability of being between -1σ and 2σ
5. (tips: $x = 2 = 2\sigma$, $\mu = 0$, $\sigma = 1$, equation above has range [$-\infty$,x])

Note: In educational assessment, T-score is a standard score Z shifted and scaled to have a mean of 50 and a standard deviation of 10.

You can also find the integral of the Standard Normal curve by calling pnorm() in R

SAMPLING DISTRIBUTIONS

Sample Point: an event which contains only a single outcome in the sample space

Sample: A set of collected sample points

Statistic: Descriptive measure computed from the data of a sample (e.g., mean, median, etc.)

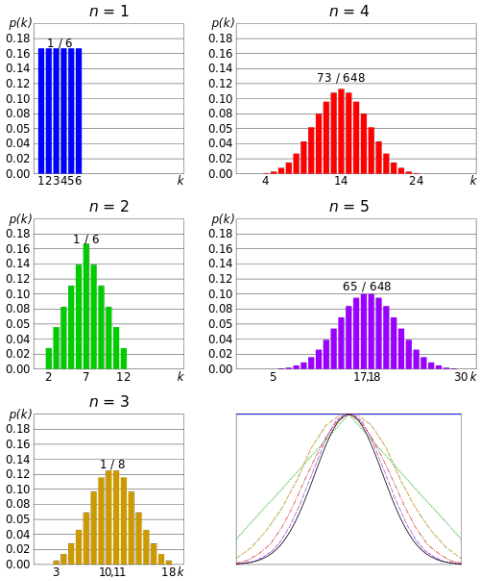
Sampling Distribution: The probability distribution of a given statistic taken from a random sample.

Constructing a Sampling Distribution

1. Randomly draw n sample points from a finite population with size N
2. Compute statistic of interest
3. List different observed values of the statistic with their corresponding frequencies

Central Limit Theorem: when i.i.d. random variables are added the sample distribution often approaches a normal distribution—even if they were drawn from a non-normally distributed population

Central Limit Theorem | Dice Rolling Example

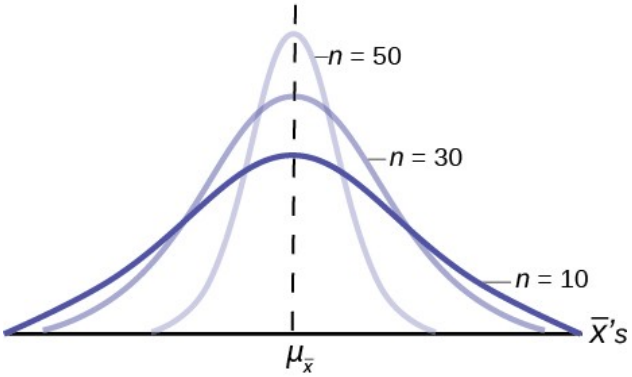


Let's roll n number of dice. The distribution of the sum (or average) of the rolled numbers will approximate a normal distribution, even though each single dice's value is sampled from a discrete uniform distribution. In the bottom-right graph, smoothed profiles of the previous graphs are rescaled, superimposed and compared with a normal distribution (black curve).

Distribution of the Sample Mean: (Known population variance):
Mean: $\mu_{\bar{X}} = \mu$,
where μ is the population mean.

Standard Error of the Mean: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$,
where σ is the standard deviation of the population distribution and n is the sample size. We divide by $\sqrt{n}$ because the more we sample the more confidence (smaller spread) we have in our sample mean estimate. For example, we become more and more confident on what the average population height is when we average together the height of more and more individuals. *This formula assumes σ is known, which is rarely the case.* This calculation is used to construct a confidence interval or forms the basis of a **z-test** when evaluating a hypothesis.

Notice in the figure below how as n increases we become more confident (less spread) of our estimate of the population mean.



- - - - -
Z-score of a sample mean: $z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$,
Here we standardize (z-transform) some sample mean ($\bar{X}$) relative to the sample mean distribution:

Distribution of the Difference Between Two Means (Known population variance):
Mean Difference: $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$,
where μ_1 and μ_2 are the population means.

Standard Error of the Mean Difference: $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$,
where σ_1 and σ_2 are the standard deviation of the population distributions and n_1 and n_2 are the sample sizes. *This formula assumes σ is known, which is rarely the case.* This calculation is used to construct a confidence interval or forms the basis of a **z-test** when evaluating a hypothesis.
- - - - -
Z-score of the difference between two sample means: $z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$

Distribution of the Sample Mean: (*Unknown* population variance):
Mean: $\mu_{\bar{X}} = \mu$,
where μ is the population mean.

Standard Error of the Mean: $s_{\bar{x}} = \frac{s}{\sqrt{n}}$,
where s is the sample standard deviation and n is the sample size. This calculation is used to construct a confidence interval, or forms the basis of a **z-test** (when $n \geq 30$) or a **t-test** (when $n < 30$) when evaluating a hypothesis.
- - - - -
T-score of a sample mean: $t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$,
Here we standardize (t-transform) some sample mean ($\bar{X}$) relative to the sample mean distribution. Note the equation for a t or z score is the same. If $n < 30$ one uses the Student's t-distribution (see below) to calculate percentiles and critical scores of significance, but if $n \geq 30$ one can use the normal distribution. Practically, the Student's t-distribution approximates the normal distribution so it is often standard practice to simply default to a t-score.

Distribution of the Difference Between Two Means (*Unknown* population variance):
Mean Difference: $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$,
where μ_1 and μ_2 are the sample means.

Standard Error of the Mean Difference: $s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$,
where s_1 and s_2 are the sample standard deviation and n_1 and n_2 are the sample sizes. This calculation is used to construct a confidence interval, or forms the basis of a **z-test** (when $n \geq 30$) or a **t-test** (when $n < 30$) when evaluating a hypothesis.
- - - - -
t-score of the difference between two sample means: $t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

Standard error can also be used for proportions (p).

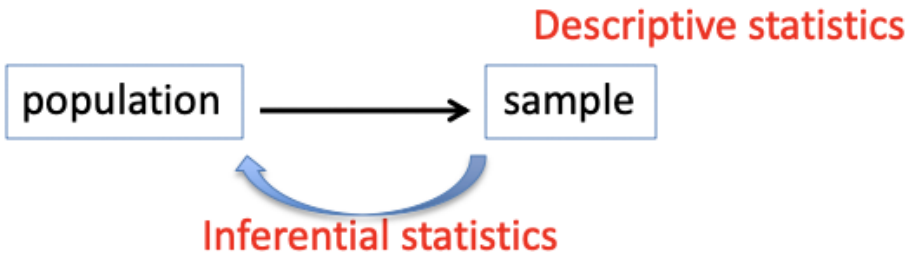
Sample proportion	$\mu_{\hat{p}}=p$	$\sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$
Difference between two sample proportions	$\mu_{\hat{p}_1-\hat{p}_2} = p_1-p_2$	$\sigma^2_{\hat{p}_1-\hat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

Notes on Standard Error

1. not all distributions are normally distributed (e.g., skewed), despite central limit theorem
2. standard error works well using the mean as a statistic, but not others (e.g., median, mode, interquartile range, etc).
3. alternatives include bootstrapping and MCMC (Markov Chain Monte Carlo) for non-normal, non-linear and other complex models

ESTIMATION | CONFIDENCE INTERVALS

Statistical Inference: is the procedure by which we reach conclusions about a population based on a sample.



Two Types of Statistical Inference

1. Estimation (see below)
2. Hypothesis Testing (Lecture 2)

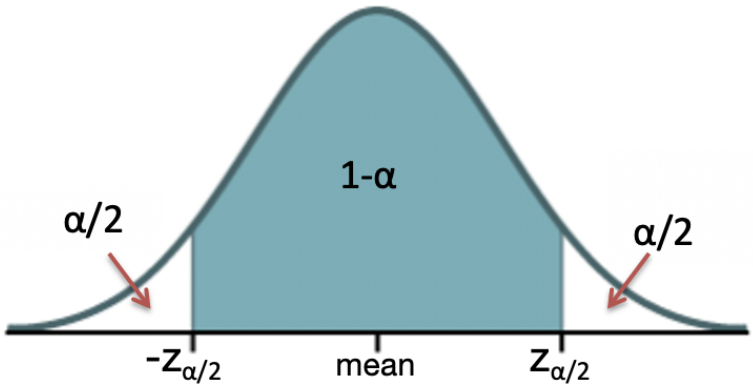
Normal Distribution

Point Estimation: A single numerical value used to estimate the unknown population parameter (e.g., just considering the mean)

Interval Estimation: Provides a range of values for an unknown population parameter

Unbiased Estimator: The estimator of a given parameter is unbiased if its expected value (average) is equal to the true value of the parameters

Confidence Intervals: A range of possible values for an unknown parameter (e.g., mean). A valid confidence interval has some probability, given some confidence level (e.g., where $1 - \alpha = 90\%, 95\%, 99\%$), of containing the true underlying parameter.



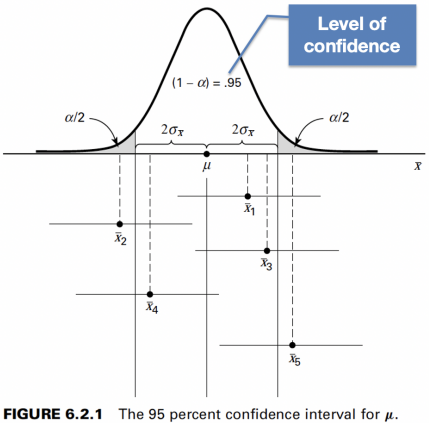


FIGURE 6.2.1 The 95 percent confidence interval for μ.

Deriving a Confidence Interval

Define $z_{\alpha/2}$:

- $p(z \geq z_{\alpha/2}) = \alpha/2$
- $p(z \leq -z_{\alpha/2}) = \alpha/2$
- $p(-z_{\alpha/2} \leq z \leq z_{\alpha/2}) = 1 - \alpha$

since, $z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

therefore, $p(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}) = 1 - \alpha$

now using some simple rearrangement of terms we show:

$p(-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}})$

$p(-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} - \bar{X} \leq -\mu \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} - \bar{X})$

$p(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}})$

Lower Bound: $\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Upper Bound: $\bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Thus,

Confidence Interval: $\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

confidence interval = estimator $\pm$ (reliability coefficient)(standard error)

Confidence Level	α	α/2	$z_{\alpha/2}$	Interval
90%	0.1	0.050	1.6449	$\bar{x} \pm 1.6449\ \sigma/\sqrt{n}$
95%	0.05	0.025	1.9600	$\bar{x} \pm 1.9600\ \sigma/\sqrt{n}$

In the table above, $z_{alpha/2}$ represents our reliability coefficient (note: these are also called critical values when used in hypothesis testing). Traditionally, researchers used to look these up in a very large z-table. However, now we can just call upon those values in a programming language. In R, we can find $z_{alpha/2}$ using the command, qnorm(), For example, $z_{alpha/2}$ for a 90% confidence interval is found by typing, qnorm(1 - 0.1/2.) = 1.644854. Note, we are dividing the 10% by 2 since we are considering the interval around both sides of the mean. In other words, for this example there is 5% on the two tails on either side of the confidence interval.

Example: Calculating the Confidence Interval using the Normal Distribution

The following are activity values of a certain enzyme measured in normal gastric tissue of 35 patients with gastric carcinoma.

.360	1.189	.788	.614	.273	2.464	.571
1.827	.537	.449	.374	.262	.448	.971
.372	.898	.348	.411	1.925	.550	.622
.610	.319	.413	.406	.767	.385	.674
.521	.603	.662	.533	1.177	.307	1.499

Suppose the population variance (σ^2) is 0.36. Construct a 95% confidence interval for the population mean.

Confidence Interval:
$\bar{X} = 0.718$
$z_{\alpha/2} = 1.96$ (found using the qnorm function in R)
$\sigma = \sqrt{0.36} = 0.6$
$n = 35$

$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ $0.718 \pm 1.96 \cdot \frac{0.6}{\sqrt{35}}$
0.718 ± 0.199

Thus, the confidence interval ranges from 0.519 (lower-bound = 0.718 - 0.199) to 0.917 (upper-bound = 0.718 + 0.199).

Interpretation of Confidence Interval:

1. practical interpretation: When sampling is from a normally distributed population with known standard deviation, we are $100(1-\alpha)$ percent confident that the single computed interval $X \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, contains the population mean μ.

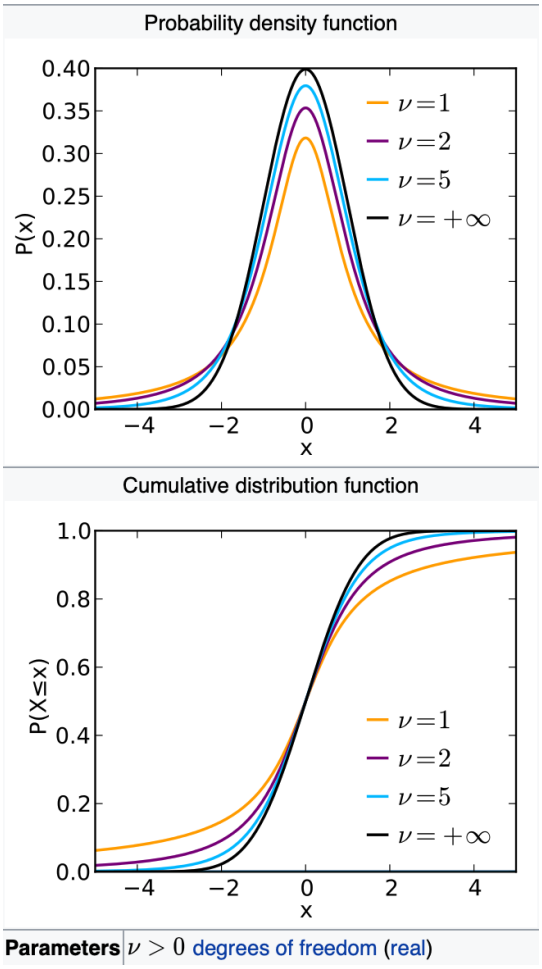
2. probabilistic interpretation: In repeated random sampling of a normally distributed population with known standard deviation, $100(1-\alpha)$ percent of all intervals of size $X \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ will contain μ.

Important Consideration for estimation:

1. Your sampled population must accurately reflect your target population. Thus, your data should come from *Random Samples*

2. When n is small (<30) or the population variance is unknown, we have to calculate confidence intervals using the Student's t-distribution. Otherwise you will underestimate the confidence interval.

Student's t-distribution



Parameters $\nu > 0$ degrees of freedom (real)

1. The single parameter ν represents the degrees of freedom (n - 1).
2. Fun Fact: This distribution was discovered by William Sealy Gosset, under the pen name Student, to monitor the quality of stout beer while working at Guinness.
3. The t distribution is heavy tailed (more spread out) than the Normal distribution for a small ν but approaches the Normal distribution as ν approaches infinity
4. Nearly the same procedure to obtain a confidence interval based on Student's t-distribution compared to using a Normal distribution.

Finding $t_{\alpha/2}$

1. Traditionally, researchers used a t-table since Student's t-distribution is a complex function. But we are in the 21st century, so let's just use a programming language.

2. In R, we can use the function pt(t, ν) to find the area below some t-value (shown as an x-value on the figure above) in the t-distribution curve.

3. In R, we can use the function qt() to find the $t_{\alpha/2}$ value used to construct a confidence interval

4. To find $t_{\alpha/2}$ for a 95% confidence interval given $n = 35$, in R we can type the following command: qt(1-0.05/2,(35-1)) = 2.032245.

Confidence Interval using the t-distribution: Example

Let's say we have the same 35 patients with gastric cancer as before, but we don't know the population variance (note: experimentally we rarely know the population variance). Find the 95% confidence interval.

.360	1.189	.788	.614	.273	2.464	.571
1.827	.537	.449	.374	.262	.448	.971
.372	.898	.348	.411	1.925	.550	.622
.610	.319	.413	.406	.767	.385	.674
.521	.603	.662	.533	1.177	.307	1.499

Confidence Interval: $(\bar{X} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}})$
$\bar{X} = 0.718$
$t_{\alpha/2} = 2.032245$ (found using the qt function in R)
$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n-1}}$ = 0.51
$n = 35$

$\bar{X} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$
$0.718 \pm 2.032245 \cdot \frac{0.51}{\sqrt{35}}$
0.718 ± 0.1751912

Thus, the confidence interval ranges from 0.54 (lower-bound = 0.718 - 0.175) to 0.89 (upper-bound = 0.718 + 0.175).

Confidence (Welsh's t) Interval using the t-distribution: Mean Difference Example

Treatment time for patients with different disorders shown in table. What is the 95% confidence interval?

Disorder	Patients (n)	Treatment time ($\bar{x}$)	Std. Deviation (s)
Schizophrenia	18	4.7	9.3
Bipolar	10	8.8	11.5

Confidence Interval: $\bar{X_1} - \bar{X_2} \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$\bar{X_1} - \bar{X_2} = -4.1$

$s_1 = 9.3, s_2 = 11.5$

$n_1 = 18, n_2 = 10$

$t_{\alpha/2}$ unknown. Lets define ν first, which gets substantially more complex when we have unknown and unequal variance of the two groups

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

$$\nu \approx \frac{\left(\frac{9.3^2}{18} + \frac{11.5^2}{10}\right)^2}{\frac{9.3^4}{18^2(18-1)} + \frac{11.5^4}{10^2(10-1)}}$$

$\nu \approx 15.63$

To find $t_{\alpha/2}$ for a 95% confidence interval given $\nu \approx 15.63$, in R we can type the following command: qt(1-0.05/2,15.63)) = 2.123991.

$$\bar{X_1} - \bar{X_2} \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$4.7 - 8.8 \pm 2.123991 \cdot \sqrt{\frac{9.3^2}{18} + \frac{11.5^2}{10}}$$

-4.1 ± 9.02

Thus, the confidence interval ranges from -13.1 (lower-bound) to 4.42 (upper-bound).