Statistics Review

1. "Kindergartners and first-graders who watched as little as one hour of television a day were more likely to be overweight or obese compared to children who watched TV for less than 60 minutes each day."
   `http://www.sciencedaily.com/releases/2015/04/150426110453.htm`

2. "The first malaria vaccine candidate (RTS,S/AS01) to reach phase 3 clinical testing is partially effective against clinical disease in young African children up to 4 years after vaccination, according to final trial data."
   `http://www.sciencedaily.com/releases/2015/04/150423211320.htm`

3. "State laws banning texting while driving led to significant reductions in the number of teens using their cell phones while behind the wheel, but nearly one-third still admitted to engaging in this risky behavior, according to new research." `http://www.sciencedaily.com/releases/2015/04/150425215628.htm`

4. A company that produces bottled water claims that each bottle contains 12 ounces of water. If 100 bottles are randomly selected and found to contain an average of 11.5 ounces of water, should we reject the company's claim?

Data consists of individual pieces of information, usually collected through experimentation. We will be especially interested in numerical data.

Statistics is the branch of science and mathematics that deals with collecting and analyzing data. In particular, *inferential statistics* involves the analysis of data in order to make conclusions and predictions applying to large populations.

Population vs. Sample

A **population** is the collection of all data outcomes that are of interest, while a **sample** is a subset of a population.

Parameter vs. Statistic

A **parameter** is a numerical quantity describing a population. A **statistic** is a numerical quantity that describes a sample.
(Notation: parameters are usually denoted by Greek letters.)

Exercise: In the examples at the top of the page,

1. Identify the population of interest.

2. Provide an example of a sample from such a population

3. Provide an example of a parameter and an example of a statistic.

In inferential statistics, the goal is generally to make an inference about a usually very large population from a smaller, randomly selected sample. For basic statistical tests, the sample should not only be random but should be a simple random sample (SRS). An SRS is a sample that is selected in such a way that all individuals from the population have the same probability of being chosen. While an SRS ensures an unbiased selection from a given population, there is no guarantee that a particular SRS represents the population in an unbiased way.

Review of Normal Distributions and Random Variables

Definitions:

(i) A **random variable** is a function from subsets of a sample space to the real numbers. For this review, it is sufficient to assume that a random variable is a function that assigns a real number to each outcome in the sample space. Random variables are often denoted by capital letters such as X, Y, or Z.

(ii) A **discrete random variable** takes on finitely or countably many values.

(iii) A **continuous random variable** takes on values in an interval.

Examples of random variables:

(a) Recall the *Bernoulli random variable*, X, is equal to 1 if an experiment resuts in a success and 0 if the experiment does not result in a success. This random variable is discrete with $P(X = 1) = p$, the success probability, and $P(X = 0) = 1 - p = q$.

(b) Recall the *binomial random variable*, X = the number of successes in n independent trials. The binomial random variable is also discrete, with

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \qquad k = 0, 1, \ldots, n$$

(c) A *normal random variable*, $X \sim N(\mu, \sigma^2)$ with mean μ and variance σ^2 may take on a value of any real number. Probabilities associated with X can be computed by the formula, for $a \leq b$, with $a, b \in [-\infty, \infty]$,

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx$$

The integrand is called the **probability density function** (p.d.f.) of $X \sim N(\mu, \sigma^2)$.

There is no closed form for the antiderivative of the p.d.f. of a normal random variable. To compute probabilities associated with a normal random variable, X, one can

1. (optional) First *standardize* X by converting it to a standard normal random variable, $Z \sim N(0,1)$:
$$Z = \frac{X - \mu}{\sigma}$$

2. Then use a table, calculator, or software to find $P(a \leq Z \leq b)$. Usually, tables and some calculators will provide $P(Z \leq b)$ only. If that is the case,

   - $P(Z \geq b) = 1 - P(Z \leq b)$
   - $P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$

Exercises:

1. Suppose that $X \sim N(0,1)$, that is with $\mu = 0$ and $\sigma^2 = 1$. Compute:

   (a) $P(X \leq 0)$

   (b) $P(X \geq 0)$

   (c) $P(X \leq 1)$

   (d) $P(X \geq 1)$

   (e) $P(0 \leq X \leq 1)$

   (f) $P(-1 \leq X \leq 1)$

   (g) $P(-1.5 \leq X \leq 1)$

2. Suppose that $X \sim N(5,4)$, that is with $\mu = 5$ and $\sigma^2 = 4$. Compute:

   (a) $P(X \leq 5)$

   (b) $P(X \geq 5)$

   (c) $P(X \leq 9)$

   (d) $P(X \geq 9)$

   (e) $P(5 \leq X \leq 9)$

   (f) $P(1 \leq X \leq 9)$

   (g) $P(0 \leq X \leq 4)$

3. The average speed of vehicles traveling on a stretch of highway is 65 miles per hour with a standard deviation of 3.5 miles per hour. Assume that the speeds of vehicles are normally distributed. A vehicle is selected at random.

   (a) What is the probability that the vehicle is violating the 70 mph speed limit?

   (b) What is the probability that the vehicle is traveling between 60 and 70 mph?

   (c) What is the probability that the vehicle is traveling under 62 mph?

   (d) What is the probability that the vehicle is traveling exactly 65 mph?

   (e) Below what speed are 75% of the vehicles traveling?

(f) Below what speed are 55% of the vehicles traveling?

(g) Above what speed are 45% of the vehicles traveling?

(h) Above what speed are 95% of the vehicles traveling?

Inference About the Mean of a Population

Notation: the population mean is typically denoted by μ, and the sample mean is typically denoted by $\bar{x}$.

The population variance is denoted σ^2 and the population standard deviation is denoted σ.

The *sample* variance is denoted s^2 and the sample standard deviation is denoted s. However, the formulas for calculating sample variance and standard deviation involve a correction and are thus slightly different from the formulas that were provided in the data analysis section.

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

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

The normal random variable plays a central role in probability and statistics due to the *Central Limit Theorem* (CLT). Before stating this theorem, consider a data set with numerical values. One can view each entry of the data as a random variable. These random variables need not be normal but for the CLT, should be independent. The CLT states that if the data set is large enough, the *average* of the data entries will be approximately a normal random variable with. Thus, the sample average, $\bar{x}$, may be used to infer the population mean, μ.

Confidence Intervals for the Mean

Given a (simple random) sample from a population with average $\bar{x}$ and standard deviation s, a *confidence interval* for the population mean, μ is an interval centered at $\bar{x}$ in which one expects μ to lie with high confidence.

In particular, a 95% confidence interval for μ is an interval that is constructed in such a way that if 100 simple random samples were taken from the same population, about 95 out of the 100 confidence intervals constructed from these samples would contain the true population mean μ.

A 90% confidence interval for μ is an interval that is constructed in such a way that if 100 simple random samples were taken from the same population, about 90 out of the 100 confidence intervals constructed from these samples would contain the true population mean μ.

Note: the greater the confidence desired, the longer the confidence interval is for the same sample. In particular, if 100% confidence is desired, the interval would be $(-\infty, \infty)$.

Graphing calcularators generally have the capability to calculate confidence interval. The user should indicate a percentage of confidence (usually close to 100), the sample average $\bar{x}$, and the standard deviation:

- If σ, the population standard deviation is known and $n \geq 30$, where n is the sample size, one may use a z-interval (using z-scored from the standard normal distribution).
- If σ is unknown and $n \geq 30$, s, the sample standard deviation can always be found, and one should then use a t-interval, with $n-1$ degrees of freedom. Some people use z-intervals in this case and obtain similar intervals, although this is not technically correct.
- If $n < 30$ a confidence interval should only be constructed if the data can reasonably be assumed to be approximately normally distributed; a z-interval can be used for known σ and a t-interval with $n-1$ degrees of freedom should be used for unknown σ.

Exercises:

1. The average speed of a simple random sample of 50 vehicles traveling on a stretch of highway is 67 miles per hour with a standard deviation of 3 miles per hour.
   (a) Construct a 98% confidence interval for μ.
   (b) Construct a 95% confidence interval for μ.
   (c) Construct a 92% confidence interval for μ.
   (d) Construct a 90% confidence interval for μ.
   (e) Which interval is longest? Which is shortest? Why?
   (f) Suppose it is known that $\sigma = 3$. Construct a 95% confidence interval for μ and compare with part (b).
2. The average speed of a simple random sample of 20 vehicles traveling on a stretch of highway is 67 miles per hour with a standard deviation of 3 miles per hour. Assume that vehicle speeds are normally distributed.
   (a) Construct a 98% confidence interval for μ.
   (b) Construct a 95% confidence interval for μ.
   (c) Construct a 92% confidence interval for μ.
   (d) Construct a 90% confidence interval for μ.
   (e) Which interval is longest? Which is shortest? Why?
   (f) Suppose it is known that $\sigma = 3$. Construct a 95% confidence interval for μ and compare with part (b).

Hypothesis Testing for the Mean

A hypothesis test may be conducted on a particular population parameter (one example is the population mean μ). The first key ingredient in a hypothesis test is the *null hypothesis.* We set up a null hypothesis about the parameter being tested. At the conclusion of the hypothesis test, we will either

(i) reject the null hypothesis, or,

(ii) fail to reject the null hypothesis

based on the statistical evidence we obtain about the population parameter from the sample. One never actually accepts the null hypothesis and should keep this in mind when selecting the null hypothesis. The test can only provide evidence that the null hypothesis is unlikely to be true. Thus, the null hypothesis is generally one that an experimenter does not believe to be true. For example, if one is testing the effect of a medication, having reason to believe that the medication is effective, the null hypothesis should be that the medication has no effect.

Below is the general setup:

(i) An SRS of size n is taken from a population with unknown mean and standard deviation.

(ii) A null hypothesis is made about the *true* population mean μ.

(iii) The null hypothesis is tested using the method of **hypothesis testing**, to be explained below.

The general steps in hypothesis testing are below and are demonstrated on the following example.

Example 1: Suppose that a manufacturer advertises that its new hybrid bus has a mean gas mileage of 5.5 miles per gallon (compared with about 4 mpg for a non-hybrid bus). Consider a simple random sample of $n = 30$ hybrid buses and test their gas mileage. Suppose that in this sample, the average is $\bar{x} = 5.3$ miles per gallon. Suppose it is known that the population (of hybrid buses produced by this manufacturer) has a standard deviation of $\sigma = 0.4$ miles per gallon. Suppose that you believe the manufacturer's claim is false and that the average fuel economy is actually lower. Conduct a hypothesis test to test whether the manufacturer's claim is reasonable.

Step 1:
Determine the **null hypothesis** and express it mathematically. The null hypothesis is generally denoted by H_0.

$$H_0 : \mu = 5.5$$

Assume, to start, that the null hypothesis is true. Throughout the test, continue to assume that H_0 is true, until, the end, when you may reject H_0 in favor of:

Step 2:
The alternative to the null hypothesis is typically the claim that you believe is true. It is called the **alternative hypothesis**, denoted H_a or H_1

$$H_a : \mu < 5.5$$

Step 3:
Set a (or use the given) **level of significance**, denoted α.

Definition: The P**-value** is the probability that the sample has the sample mean $\bar{x}$ or a sample mean more extreme (based on H_a) than $\bar{x}$, given that H_0 is true.

If $P \leq \alpha$, we shall reject H_0. α is often chosen to be 0.05 (or, for example 0.01 or 0.10).

$$\alpha = 0.05$$

To determine the P-value when testing for the population mean μ, we use the Central Limit Theorem to claim that when n is large enough (≥ 30), the sample average is distributed approximately normally.

Step 4:
Calculate the appropriate *test statistic* based on the sample. The test statistic is a function of the observed data from the sample, and the decision in the hypothesis test will be made based on this test statistic. For this example, the test statistic is the z-score. If s, the sample standard deviation is given or can be found and σ is not given, use $\sigma \approx s$:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{5.3 - 5.5}{0.4/\sqrt{30}} \approx -2.739$$

Now there are two courses of action, A and B

Step 5A:
Find the P-value based on the test statistic from the previous step.

$$P = P(\bar{x} \leq 5.3 | H_0) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \frac{5.3 - 5.5}{0.4/\sqrt{30}}\right) \approx P\left(Z \leq -2.739\right) \approx 0.00308,$$

where Z is a standard normal random variable.

Notice that -2.739 is precisely the test statistic found in the previous step.

Step 6A: Make the final conclusion based on the P-value and α.

Since $0.00308 = P < \alpha = 0.05$, we conclude that there is enough statistical evidence to reject the null hypothesis that $\mu = 5.5$ in favor of the alternative hypothesis, $\mu < 5.5$.

OR

Step 5B: Determine the *rejection region.* This is the interval of possible values for the test statistic that will make us reject H_0.

This is, in this example the interval of z-values for which $P \leq \alpha$. We shall find the *critical value(s)* that define this rejection region, z_0 based on α and H_0, and H_a.

Since $\alpha = 0.05$ and $H_0 : \mu = 5.5$, $H_a : \mu < 5.5$, we are looking for z_0 so that $P(Z \leq z_0) = 0.05$. Using a table or computer software, one finds that $z_0 \approx -1.645$.

Thus, the rejection region consists of z values in the interval $(-\infty, -1.645]$.

Step 6B: If the test statistic is in the rejection region, make the decision to *reject* H_0. Otherwise, *fail to reject* H_0.

In this example, the test statistic $z = -2.739$ and is in the rejection region. Thus, we reject H_0.

(Note: Failing to reject H_0 is not the same as accepting it – H_0 was assumed to be true from the start. Failing to reject H_0 simply means that there is not enough statistical evidence to reject it.)

It is not possible to determine with absolute certainty whether or not H_0 is true, unless the entire population is tested, which is rarely feasible.

This table shows all of the possible scenarios that can occur in hypothesis testing:

	H_0 **is true**	H_0 **is false**
Fail to reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

In summary, the inputs in a hypothesis test for the mean are:

H_0 : null hypothesis[1]
H_a : alternative hypothesis
α : significance level
$\bar{x}$: sample mean
s : sample standard deviation
P : the P-value, or
z_0 : the critical value that determines the rejection region

The output is a decision to reject H_0 or not, based on the relationship between z, which you calculate, and the P or z_0, the P-value or the threshold between the rejection region and the non-rejection region.

[1] H_0 should *always* contain the equals part of the inequality.

Exercises:

1. Determine the conclusion of the hypothesis test in example 1 if $\alpha = 0.1$ and if $\alpha = 0.01$. In each case, do this both ways: using the P-value and using the rejection region.
2. Repeat example 1, but assume that you have no prior expectation on whether the true fuel efficiency is higher or lower than the manufacturer's claim. In this case, $H_a : \mu \neq 5.5$. How are the P-value and rejection region affected? Use $\alpha = 0.05$.
3. Repeat example 1, but assume that the manufacturer claims that the fuel efficiency is 5.0 mpg and that you believe that the true fuel efficiency is greater. Use $\alpha = 0.05$.

When σ is unknown, and s is used to approximate σ, a t-value is calculated, rather than the z-value,

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

Inference About the Variance of a Population

In some situations, it is important to infer the variance of a population, σ^2, from a simple random sample of size n.

Hypothesis Testing for the Variance

Key Assumption: the population must be approximately normally distributed for this test to be valid.

Consider the following example:

Example 3: A yogurt company produces yogurt containing an average of 32 ounces of yogurt per container. The company claims that the standard deviation is 0.02 ounces. A simple random sample of 25 items has been selected and weighed. The standard deviation was 0.03 ounces. Does this support the company's claim? Assume that the population is normally distributed.

If s^2 is the variance of a sample of size n from a normally distributed population with variance σ^2, then $\frac{(n-1)s^2}{\sigma^2}$ is a random variable following a χ^2 distribution (chi-squared distribution) with $n - 1$ degrees of freedom (we will not prove this).

Definition: a random variable X is distributed according to a χ^2 *distribution* with $r \in \mathbb{N}$ degrees of freedom if it has probability density function

$$f_X(x) = \frac{x^{\frac{r}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{r}{2}} \Gamma\left(\frac{r}{2}\right)}, \ x > 0$$

Now, let us answer the question in example 3 using a hypothesis test for variance. Note that the sample variance multiplied by $(n-1)/\sigma^2$ has a χ^2 distribution and *not* the sample standard deviation.

Step 1:
Determine the **null hypothesis** and express it mathematically.

$$H_0 : \sigma^2 = 0.02^2 = 0.0004$$

Assume, to start, that the null hypothesis is true. Throughout the test, continue to assume that H_0 is true, until, the end, when you may reject H_0 in favor of:

Step 2:
The alternative hypothesis

$$H_a : \sigma^2 > 0.0004$$

Step 3:
Set a (or use the given) **level of significance**, denoted α.

$$\alpha = 0.05$$

To determine the P-value when testing for the population variance σ^2, we use the fact that since the population is normally distributed, then $(n-1)s^2/\sigma^2$ follows a χ^2 distribution.

Step 4:
Calculate the appropriate *test statistic* based on the sample. For a test of variance, calculate a χ^2 value.

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma^2} = \frac{(25-1)\,(0.03)^2}{(0.02)^2} \approx 54$$

Now there are two courses of action, A and B

Step 5A:
Find the P-value based on the test statistic from the previous step.

$$P = P(s^2 > (0.03)^2 | H_0) \approx P\left(X > \frac{(25-1)\,(0.03)^2}{(0.02)^2}\right) = P\,(X > 54) \approx 0.0004,$$

where X is a χ^2 random variable with 24 degrees of freedom.

Step 6A: Make the final conclusion based on the P-value and α.

Since $0.0004 = P < \alpha = 0.05$, we conclude that there is enough statistical evidence to reject the null hypothesis that $\sigma^2 = 0.0004$ (equivalent to $\sigma = 0.02$).

OR

Step 5B: Determine the *rejection region.* This is the interval of possible values for the test statistic that will make us reject H_0.

This is, in this example the interval of x-values for which $P \leq \alpha$. We shall find the *critical value(s)* that define this rejection region, x_0 based on α and H_0, and H_a.

Since $\alpha = 0.05$ and $H_0 : \sigma^2 = 0.0004$, $H_a : \sigma^2 > 0.0004$, we are looking for x_0 so that $P(X > x_0) = 0.05$. Using a table or computer software, one finds that $x_0 \approx 36.4$.

Thus, the rejection region consists of x values in the interval $[36.4, \infty)$.

Step 6B: If the test statistic is in the rejection region, make the decision to *reject* H_0. Otherwise, *fail to reject* H_0.

In this example, the test statistic $\chi^2 = 54$ is in the rejection region. Thus, we reject H_0.

Exercises:

1. Determine the conclusion of the hypothesis test in example 3 if $\alpha = 0.1$ and if $\alpha = 0.01$. In each case, do this both ways: using the P-value and using the rejection region.
2. Repeat example 3 using $H_0 : \sigma^2 = 0.02$, $H_a : \sigma^2 \neq 0.02$, $s^2 = 0.03$. How are the P-value and rejection region affected? Use $\alpha = 0.05$.
3. Repeat example 3 using $H_0 : \sigma^2 = 0.02$, $H_a : \sigma^2 < 0.02$, $s^2 = 0.01$. How are the P-value and rejection region affected? Use $\alpha = 0.05$.
4. Repeat example 3, but assume that the sample size is $n = 20$.

Correlation

Is there a relationship between two variables? For example, is there a relationship between the number of employee training hours and the number of on-the-job accidents? Is there a relationship between the number of hours a person sleeps and their reaction time? Is there a relationship between the number of hours a student spends studying for a calculus test and the student's score on that calculus test?

Definition: a **correlation** is a relationship between two variables.

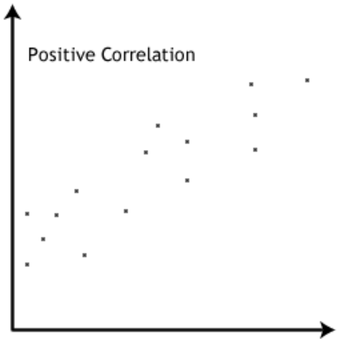
Typically, **x** is taken to be the **independent variable**, **y** is taken to be the **dependent variable**. Data is represented by a collection of ordered pairs (x, y).

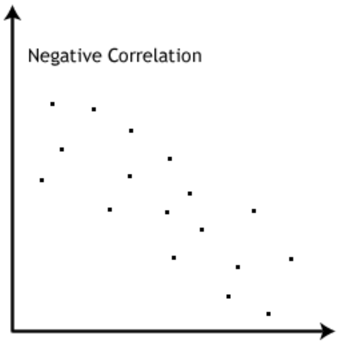
Mathematically, the strength and direction of a linear relationship between two variables is represented by the **correlation coefficient**. Suppose that there are n ordered pairs (x, y) that make up a sample from a population. The correlation coefficient r is given by:

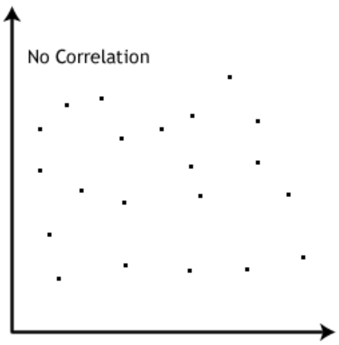
$$r = \frac{n\sum(xy) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

This will always be a number between -1 and 1 (inclusive).

- If r is close to 1, the variables are said to be *positively correlated.* This means there is likely a strong linear relationship between the two variables, with a positive slope.
- If r is close to -1, the variables are said to be *negatively correlated.* This means there is likely a strong linear relationship between the two variables, with a negative slope.
- If r is close to 0, the variables are *not correlated.* This means that there is likely no *linear* relationship between the two variables, however, the variables may still be related in some other way.







The correlation coefficient of the population is denoted by ρ – and is usually unknown.

Exercises:

1. The time x in years that an employee spent at a company and the employee's hourly pay, y, for 10 employees are listed in the table below. Calculate and interpret the correlation coefficient r. Include a plot of the data in your discussion.

Employee	1	2	3	4	5	6	7	8	9	10
Years at the company	5	3	4	10	15	6	7	7	12	10
Hourly pay	25	20	31	35	38	22	25	28	30	30

2. The table below shows the number of absences, x, in a Calculus course and the final exam grade, y, for 9 students. Find the correlation coefficient and interpret your result.

x	1	0	2	6	4	3	3	2	5
y	95	90	90	55	70	80	85	95	60

3. The table below shows the height, x, in inches and the pulse rate, y, per minute, for 9 people. Find the correlation coefficient and interpret your result.

x	68	72	65	70	62	75	78	64	68
y	90	85	88	100	105	98	70	65	72

Interpreting the Correlation Between Two Variables:

Suppose that you find a strong positive or negative correlation between two variables. Is there a cause-and-effect relationship between these variables?

- There could be a direct cause-and-effect relationship: that is, x causes y.
- There could be a reverse cause-and-effect relationship: that is, y causes x.
- There could be a third (or fourth? or more?) variable that leads to the relationship between x and y.
- The "relationship" between x and y may just be a coincidence.

Linear Regression

If there is a "significant" linear correlation between two variables, the next step is to find the equation of a line that "best" fits the data. Such an equation can be used for prediction: given a new x-value, this equation can predict the y-value that is consistent with the information known about the data. This predicted y-value will be denoted by $\hat{y}$. The line represented by such an equation is called the **linear regression** line.

The equation for a line is

$$\hat{y} = mx + b,$$

where m is the slope of the line and b is the y-intercept (the y-value for which x is 0).

In general, the regression line, will not pass through each data point. For each data point, there is an error: the difference between the y-value from the data and the y-value on the line, $\hat{y}$. By definition, this linear regression line is such that the sum of the squares of the errors is the least possible. It turns out, given a set of data, there is only one such line. The slope m and y-intercept b are given by

$$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum(x^2) - \left(\sum x\right)^2} \qquad\qquad b = \frac{\sum y}{n} - m\frac{\sum x}{n}$$

Exercises:

1. The time x in years that an employee spent at a company and the employee's hourly pay, y, for 10 employees are listed in the table below.

Employee	1	2	3	4	5	6	7	8	9	10
Years at the company	5	3	4	10	15	6	7	7	12	10
Hourly pay	25	20	31	35	38	22	25	28	30	30

(a) Find the equation of the regression line.

(b) Plot the data and the line.

(c) Does the regression line seem to "fit" the data well? What does your answer indicate about the relationship between x and y?

(d) Use the equation in (a) to predict the hourly pay rate of an employee who has worked for 20 years. Is this a good prediction?

(e) Use the equation in (a) to predict the hourly pay rate of an employee who has worked for 10 years. Is this a good prediction?

2. The table below shows the number of absences, x, in a Calculus course and the final exam grade, y, for 9 students.

x	1	0	2	6	4	3	3	2	5
y	95	90	90	55	70	80	85	95	60

(a) Find the equation of the regression line.

(b) Plot the data and the line.

(c) Does the regression line seem to "fit" the data well? What does your answer indicate about the relationship between x and y?

(d) Use the equation in (a) to predict the test grade of a student with 7 absences. Is this a good prediction?

(e) Use the equation in (a) to predict the test grade of a student with 4 absences. Is this a good prediction?

3. The table below shows the height, x, in inches and the pulse rate, y, per minute, for 9 people. Find the correlation coefficient and interpret your result.

x	68	72	65	70	62	75	78	64	68
y	90	85	88	100	105	98	70	65	72

(a) Find the equation of the regression line.

(b) Plot the data and the line.

(c) Does the regression line seem to "fit" the data well? What does your answer indicate about the relationship between x and y?

(d) Use the equation in (a) to predict the pulse rate of a person who is 70 inches tall. Is this a good prediction?

(e) Use the equation in (a) to predict the pulse rate of a person who is 65 inches tall. Is this a good prediction?

More on Regression

Some data sets of ordered pairs do not appear linear when plotted. Even if r is close to 1 or -1, a linear regression line is not appropriate for making predictions based on such a data set.

If the plot of the ordered pairs appears to have a polynomial shape, a **polynomial regression** curve may be calculated.

Another technique that may be used is to test by graphing whether one of the following modifications of the original data (x, y) yields a plot that looks more linear:

- $(\log x, y)$
- $(x, \log y)$
- $(\log x, \log y)$

If one of such plot appears linear, a linear regression line may be calculated on that plot. Here are the scenarios and terminology for each situation:

Transformed Data	**Linear Regression Equation**	**Equation for $\hat{y}$**	**Type of Equation**
(x, y)	$\hat{y} = mx + b$	$\hat{y} = mx + b$	Linear
$(\log x, y)$	$\hat{y} = m \log x + b$	$\hat{y} = m \log x + b$	Logarithmic
$(x, \log y)$	$\log \hat{y} = mx + b$	$\hat{y} = e^{mx+b}$	Exponential
$(\log x, \log y)$	$\log \hat{y} = m \log x + b$	$\hat{y} = e^b x^m$	Power

Exercises: Fit the appropriate type of regression line to each data set and name the type of relationship between the data (type of equation that best represents the data).

1. The distance in feet travelled by a ball (y) thrown at various initial angles (x):

x	15	20	25	30	35	40	45	50	55
y	161	185	221	257	271	289	288	275	268

2.

x	8	70	4	14	6	52	45	20	30
y	32	48	19	27	20	50	35	30	42

3. The temperature of the water in a pot (y), x minutes after the water has been boiled and the heat source turned off:

x	2	3	4	5	6	7	8
y	185	176	171	165	158	155	151

4. The age of a tree in years (x) and its height in feet (y):

x	2	3	4	5	6	7	8	9	10
y	10	14	16	17	18	18.5	18.9	19.2	19.3