Lecture 29

Nancy Pfenning Stats 1000

Reviewing Confidence Intervals and Tests for Ordinary One-Sample, Matched-Pairs, and Two-Sample Studies About Means

Example

Blood pressure X was measured for a sample of 10 black men. It was found that $\bar{x} = 114.9$, $s = 10.84$. Give a 90% confidence interval for mean blood pressure μ of all black men. [Note: we can assume that blood pressure tends to differ for different races or genders, and that is why a separate study is made of black men—the confounding variables of race and gender are being controlled.] This is an ordinary one-sample t procedure.

A level .90 confidence interval for μ is $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$, where t^* has $10 - 1 = 9$ df. Consulting the *df* $= 9$ row and .90 confidence column of Table A.2, we find $t^* = 1.83$. Our confidence interval is $114.9 \pm 1.83 \frac{10.84}{\sqrt{10}} = (108.6, 121.2)$.

Here is what the MINITAB output looks like:

```
              N      MEAN    STDEV  SE MEAN   90.0 PERCENT C.I.
 calcbeg     10    114.90    10.84     3.43  (  108.62,  121.18)
```

Example

Blood pressure for a sample of 10 black men was measured at the beginning and end of a period of treatment with calcium supplements. To test at the 5% level if calcium was effective in lowering blood pressure, let the R.V. X denote *decrease* in blood pressure, beginning minus end, and μ_D would be the population mean decrease. This is a matched pairs procedure.

To test $H_0 : \mu_D = 0$ vs. $H_a : \mu_D > 0$, we find differences X to have sample mean $\bar{d} = 5.0$, sample standard deviation $s = 8.74$. The t statistic is $t = \frac{\bar{d}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{5-0}{\frac{8.74}{\sqrt{10}}} = 1.81$, and the P-value is $P(T \geq 1.81)$. We refer to Table A.2 for the $t(9)$ distribution, and see that 1.81 is just under 1.83, which puts our P-value just over .05. Our test has not quite succeeded in finding the difference to be significantly greater than zero, in a statistical sense. Populations of black men treated with calcium may experience no decrease in blood pressure.

MINITAB output appears below.

```
 TEST OF MU = 0.00 VS MU G.T. 0.00

              N      MEAN    STDEV   SE MEAN        T    P VALUE
 calcdiff    10      5.00     8.74      2.76     1.81      0.052
```

It is possible that our sample size was too small to generate statistically significant results. Another concern is the possibility of confounding variables influencing their blood pressure change. The placebo effect may tend to bias results towards a larger decrease. Or, time may play a role: if the beginning or end measurement date happened to be in the middle of a harsh winter or a politically stressful time, results could be affected.

Example

Data for a control group (taking placebos) of 11 black men at the beginning and end of the same time period produced control sample mean difference $\bar{d}_2 = -.64$, and $s_2 = 5.87$. Now we test $H_0 : \mu_1 - \mu_2 = 0$ [same as $H_0 : \mu_1 = \mu_2$, or mean difference for calcium-takers same as mean difference for placebo-takers] vs.

$H_a : \mu_1 - \mu_2 > 0$ [same as $H_0 : \mu_1 > \mu_2$, or mean difference for calcium-takers greater than mean difference for placebo-takers].

The t statistic is

$$t = \frac{(\bar{d}_1 - \bar{d}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5 - (-.64)}{\sqrt{\frac{8.74^2}{10} + \frac{5.901^2}{11}}} = \frac{5.64}{3.282} = 1.72.$$

Since $10 - 1 < 11 - 1$, use 9 df. In this row of Table A.2, we see that 1.72 is smaller than 1.83, so the P-value is larger than .05. Once again there is not quite enough evidence to reject the null hypothesis. [Note that in the MINITAB output below, degrees of freedom were calculated (in a very complicated way) to be 15, whereas we simply took the smaller sample size minus one, which was 9.]

```
TWOSAMPLE T FOR calcdiff VS contdiff
           N      MEAN     STDEV   SE MEAN
calcdiff  10      5.00      8.74      2.76
contdiff  11     -0.64      5.87      1.77

TTEST MU calcdiff = MU contdiff (VS GT): T= 1.72  P=0.053  DF=  15
```

Robustness: the sample sizes are quite small, and so MINITAB plots of the above distributions should be consulted to verify that they have no pronounced outliers or skewness.

Data values are shown in the following table:

Calcium			Control		
Beginning	End	Difference	Beginning	End	Difference
107	100	7	123	124	-1
110	114	-4	109	97	12
123	105	18	112	113	-1
129	112	17	102	105	-3
112	115	-4	98	95	3
111	116	-5	114	119	-5
112	102	10	112	114	-2
136	125	11	110	121	-11
102	104	-2	117	118	-1
107	106	1	119	114	5
			130	133	-3

```
                N      MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
calcbeg        10    114.90   111.50   113.88    10.84     3.43
calcend        10    109.90   109.00   109.25     7.80     2.47
calcdiff       10      5.00     4.00     4.62     8.74     2.76
contbeg        11    113.27   112.00   113.11     9.02     2.72
contend        11    113.91   114.00   113.89    11.33     3.42
contdiff       11     -0.64    -1.00    -0.89     5.87     1.77
```

Example

A biologist suspects that the antiseptic Benzamil actually impairs the healing process. To test her suspicions with a matched-pairs design, 9 salamanders are randomly selected for treatment. Each has one wounded hind leg treated with Benzamil, the other wounded leg treated with saline (as a control). The healing X (area in square millimeters covered with new skin) is measured after a certain period of time, with the following results:

Animal No.	Benzamil	Control	Difference B-C
1	.14	.32	-.18
2	.08	.15	-.07
3	.21	.42	-.21
4	.13	.13	.00
5	.10	.26	-.16
6	.08	.07	+.01
7	.11	.20	-.09
8	.04	.16	-.12
9	.19	.18	+.01
			$d = -.09$ $s_D = .084$

First find a 95% confidence interval for μ_d (the population mean difference benzamil minus control). Then test $H_0 : \mu_d = 0$ vs. $H_a : \mu_d < 0$. Note that because treatment and control are applied to both legs of the same salamander, the design is matched pairs and a one-sample t procedure should be used on the single sample of differences.

The critical value t^* for our 95% confidence interval is taken from the $9 - 1 = 8$ df row and the .95 confidence column of Table A.2: $t^* = 2.31$. Our interval is

$$-.09 \pm 2.31 \frac{.084}{\sqrt{9}} = -.090 \pm .065 = (-.155, -.025)$$

Note that the above interval contains only negative values, leading us to expect the difference to be statistically significant.

To test $H_0 : \mu_d = 0$ vs. $H_a : \mu_d < 0$, use $t = \frac{-.09}{\frac{.084}{\sqrt{9}}} = -3.21$. Because H_a has the $<$ sign, the P-value is $P(T \leq -3.21) = P(T \geq +3.21)$ by the symmetry of the T distribution. Our test statistic is between 2.90 and 3.36 in the 8 df row, which means the P-value is between .01 and .005. No level α has been specified, so we will simply judge the P-value to be "small" and reject H_0. Our conclusion is that Benzamil does indeed impair the healing process.

```
                  N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
benzamil          9   0.1200   0.1100   0.1200   0.0543   0.0181
control           9   0.2100   0.1800   0.2100   0.1071   0.0357
diff              9  -0.0900  -0.0900  -0.0900   0.0843   0.0281

             N      MEAN    STDEV  SE MEAN   95.0 PERCENT C.I.
diff         9   -0.0900   0.0843   0.0281  ( -0.1548, -0.0252)

TEST OF MU = 0.0000 VS MU L.T. 0.0000

             N      MEAN    STDEV   SE MEAN        T    P VALUE
diff         9   -0.0900   0.0843    0.0281    -3.20     0.0063
```

Example

Suppose a biologist uses a two-sample design to test if Benzamil impairs the healing process: one random sample of 9 salamanders has a wounded hind leg treated with Benzamil, a *different* sample of 9 salamanders has a wounded hind leg treated with saline. Now find a 95% confidence interval for the difference between the mean healings, and test $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 < 0$. [Note: we use t^* for df the smaller of $9 - 1$ and $9 - 1$, which is of course 8. Once again we use the .95 confidence column.]

Benzamil	Control
.14	.32
.08	.15
.21	.42
.13	.13
.10	.26
.08	.07
.11	.20
.04	.16
.19	.18
$\bar{x}_1 = .12$	$\bar{x}_2 = .21$
$s_1 = .054$	$s_2 = .107$

Now our 95% confidence interval is

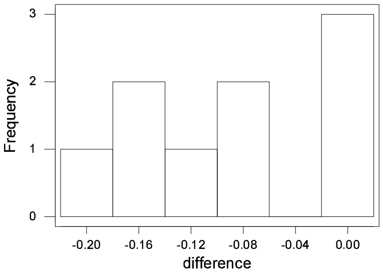
$$(.12 - .21) \pm 2.31\sqrt{\frac{.054^2}{9} + \frac{.106^2}{9}} = -.090 \pm .092 = (-.182, +.002)$$

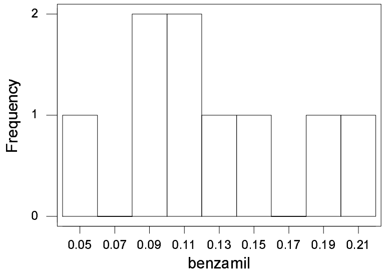
Note that this confidence interval *does* contain zero.

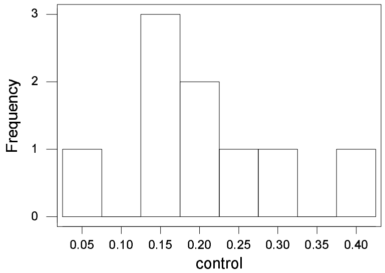
Our test statistic can be calculated to be -2.25, producing a P-value between .05 and .025. We could reject at the .05 level of significance, but not at the .01 level of significance. Thus, we see how the matched pairs study had a better chance of pinning down a difference. Subtracting values for each individual cancels out the variation among individuals, helping us to concentrate on the variation between treatment and control.

```
95 PCT CI FOR MU benzamil - MU control: ( -0.1781,  -0.001861)

TTEST MU benzamil = MU control (VS LT): T= -2.25  P=0.023  DF=  11
```

Robustness: What plots should be made first to validate use of the above procedures? Sample sizes are small, so the data should show no outliers or skewness. Looking at the shapes of the histograms below, results of the matched pairs procedure are dubious because of skewness; the two-sample results should be fine.







Example

Producers of gasoline want to test which is better, Gas A or Gas B. Miles per gallon are measured for 6 cars using Gas A and for *another* set of 6 cars using Gas B.

Gas A	Gas B
15	13
20	17
25	23
25	24
30	28
35	34
$\bar{x}_1 = 25.00$	$\bar{x}_2 = 23.17$
$s_1 = 7.07$	$s_2 = 7.52$
$n_1 = 6$	$n_2 = 6$

The two-sample t statistic is $\frac{25.00-23.17}{\sqrt{\frac{7.07^2}{6}+\frac{7.52^2}{6}}} = .43$. The P-value, according to MINITAB, is extremely large: .674. There is no evidence at all that $\mu_1 \neq \mu_2$, because the P-value is large, because t is small, because s_1 and s_2 are large. High variation *among* mileages for various cars prevented us from pinning down the effects of using a different gas.

A matched pairs design would be better:

Example

Producers of gasoline measure mpg for 6 cars using Gas A and for the *same* 6 cars using Gas B. Which gas is used first is determined by a coinflip.

Gas A	Gas B	Difference
15	13	2
20	17	3
25	23	2
25	24	1
30	28	2
35	34	1
		$d = 1.833$
		$s_d = .753$

Now the t statistic is $\frac{1.833-0}{.753/\sqrt{6}} = 5.97$ and, according to MINITAB, the P-value is extremely small: .002. We have strong evidence against $H_0 : \mu_d = 0$ because the P-value is very small, because

t is large, because s is small. Concentrating on the difference between mileages, Gas A minus Gas B, wipes out the differences among mileages for various cars, and helped control this outside variable.

Lecture 30

Chapter 14: More About Regression

In Chapter 5, we displayed the relationship between two quantitative variables with a scatterplot, and summarized it by reporting direction, form, and strength. If the form appeared linear, then we made a much more specific summary by describing the relationship with the equation of a straight line, called the least squares regression line. We also used the correlation r to specify the direction and strength of the relationship. All of this was done for the *sample* of explanatory and response pairs only. By construction our line was the one that best fitted the sample data points, but we did not attempt to draw conclusions about how the explanatory and response variables were related in the entire population from which the sample was taken.

Now that we are familiar with the principles of statistical inference, our goal in this chapter is to use sample explanatory and response values to draw conclusions about how the variables are related for the population. It should go without saying that such conclusions will only be meaningful if the sample is truly representative of the larger population. As usual, all results are based on probability distributions, which tell us what we can expect from *random* behavior.

The first step in this inference process is an important one: examine the scatterplot to decide if the form of the relationship really does appear linear. The methods of inference that we will develop cannot help us in producing evidence that a straight-line relationship holds for the larger population—this is something that we must decide for ourselves, based on the appearance of the scatterplot. If the points seem to cluster around a curve rather than a straight line, then other, more advanced options must be explored. In more advanced treatments of relationships between two quantitative variables, methods are presented for transforming variables so that the resulting relationship is linear. In this book, we will proceed no further if the relationship is non-linear. If linearity seems to be a reasonable assumption, then we can use inference to draw conclusions about what the line should be like for the entire population, and also about how much spread there is around the line. Whereas in Chapter 5 we only went so far as to predict a single value for the response to a given explanatory value, we will now have the tools to make interval estimates.

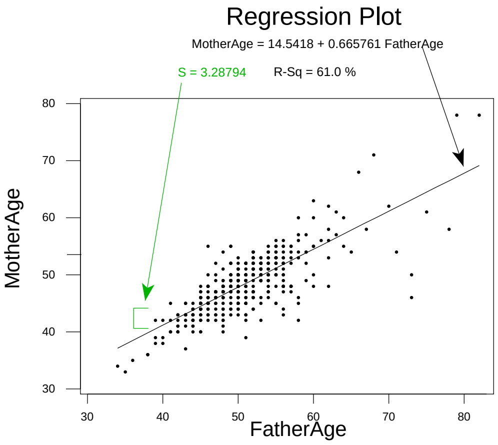
Our first example concerns two variables which common sense suggests should have a positive linear relationship: ages of students' mothers and fathers.

Example

Are ages of all students' mothers and fathers related? If so, what can we say about this relationship for a population of age pairs?

Because couples tend to have ages that are reasonably close to one another, we would expect the mother to be on the young side if the father is young, and on the old side if the father is old. There is reason to expect a rather steady increase in the variable MotherAge as values of the variable FatherAge increase. Therefore, we do expect the relationship to be positive and linear, not just for a sample of age pairs but also for the larger population.

Using the methods of Chapter 5, we can look at a scatterplot of father and mother ages and decide that it appears linear. The least squares method can be used to fit a line that comes "closest" to the data points, in the sense that it minimizes the sum of squared residuals, which are the sample prediction errors. The typical size of sample prediction error is S in the output; it calculates the square root of the "average" squared distance of observed response values minus predicted response values. (This average is calculated dividing by $n-2$, which will be our regression degrees of freedom.)

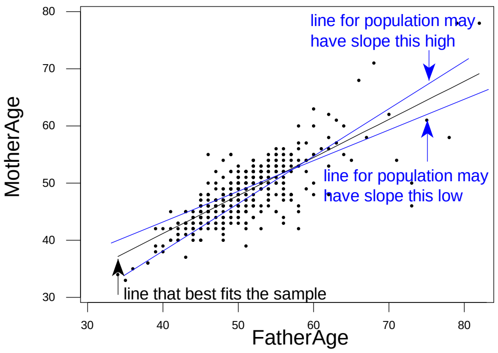


```
The regression equation is
MotherAge = 14.5418 + 0.665761 FatherAge
S = 3.28794      R-Sq = 61.0 %      R-Sq(adj) = 60.9 %
431 cases used 15 cases contain missing values
Predictor        Coef     SE Coef          T        P
Constant       14.542       1.317      11.05    0.000
FatherAge     0.66576     0.02571      25.89    0.000

Pearson correlation of FatherAge and MotherAge = 0.781
P-Value = 0.000
```

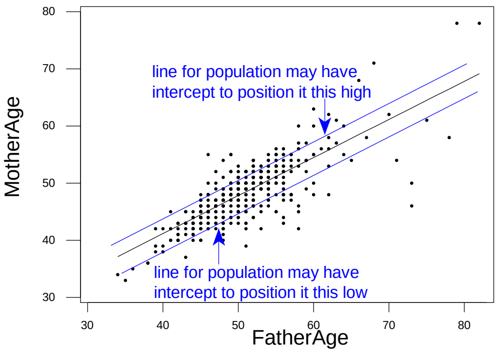
You may have noticed that a p-value is always reported along with the correlation r, and that t statistics and p-values are included with the regression output. We will pay more attention to these once we have established a meaningful hypothesis test procedure for the regression context. First, we consider relationships for *populations* as opposed to for *samples* of quantitative explanatory and response values.

When we introduced the process of performing statistical inference about a single parameter (such as population mean) based on a statistic (such as sample mean), we acknowledged that although sample mean may be our best estimate for population mean, it is almost surely "off" by some amount. Similarly, although the least squares regression line is our best guess for the line that describes the relationship for the entire population, it is also probably "off" to some extent. Unlike inference about a single parameter like unknown population mean, when we perform inference about a relationship between two quantitative variables, there are actually three unknown parameters.

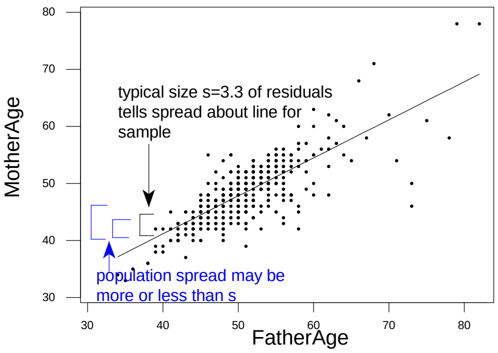
- First of all, we only know the slope $b_1 = .666$ of the line that best fits our *sample*. The line that best fits the entire *population* may have more or less slope. We use β_1 to denote the unknown slope of the population regression line. The graph below shows that other slopes are plausible candidates for β_1.



- Next, we only know the intercept $b_0 = 14.542$ of the least squares line fitted from the *sample*. The line that best fits the entire *population* has an unknown intercept β_0. There is a whole range of plausible values for β_0, resulting in lines that may be lower or higher than the line constructed from our sample.



- Thirdly, the typical size of the $n = 431$ residuals for our sample is reported in the regression output to be $S = 3.288$. The spread σ about the population regression line for *all* age pairs is unknown. The population may exhibit less spread than what is seen in our sample, or more.

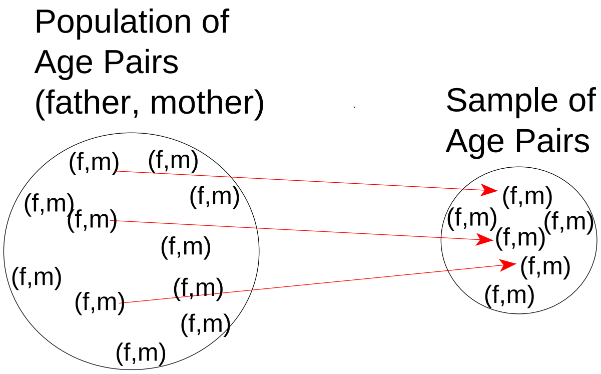


When inference methods were first introduced in Chapter 10, we learned to perform various forms of inference: first a point estimate such as $\bar{x}$

around the point estimate, and then a hypothesis test. For practical purposes, it is usually enough to use b_0 as a point estimate for unknown intercept β_0 of the population regression line, and s as a point estimate for unknown spread σ about the population regression line. Because of the special role played by slope in the relationship between two quantitative variables, we will not merely use b_1 as a point estimate for unknown slope β_1 for the population. In addition, we will be interested in an interval estimate for β_1, and perhaps most importantly we will carry out a hypothesis test about β_1. Later on in this chapter, we will use inference to make response predictions in the form of intervals, given a particular explanatory value.

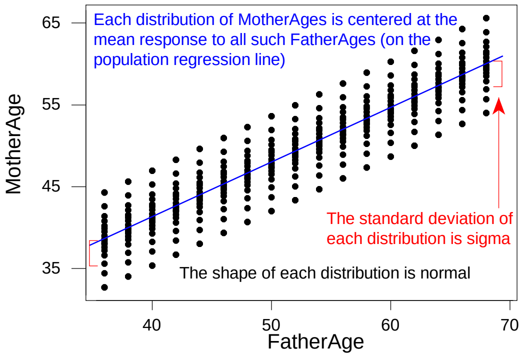
Behavior of slope for a sample

Before performing inference about population proportion in Chapters 10 and 11 and population mean in Chapters 12 and 13, we took a great deal of care in Chapter 9 to think about the behavior of sample proportion relative to population proportion, and sample mean relative to population mean. Similar considerations will be helpful now so that we can grasp the workings of the process of inference for regression. We can imagine a large population of explanatory and response values (ages of all students' fathers and mothers), from which a sample is taken.



Intuitively, it makes sense that if the ages are related linearly in the population, they should also be related more or less linearly in the sample. If a certain slope β_1 holds for the population, then the slope b_1 in the sample should be in the same ballpark. Similarly, if a certain intercept β_0 holds for the population, then the intercept b_0 for the sample should be somewhere in that vicinity. Also, if responses for the entire population are spread about the line with some standard deviation σ, then the sample standard deviation s should be similar.

The behavior of statistics like sample slope b_1 in random samples taken from the larger population of explanatory/response pairs is perfectly predictable as long as the population relationship meets certain requirements. As we stated at the beginning of this chapter, the relationship must be linear. In addition, the distribution of standardized sample slope is exactly **t** if the residuals are exactly normally distributed. For any explanatory value, responses then should vary normally about the population regression line, and their standard deviation is the parameter σ that is estimated by s. The graph below is an oversimplification of the situation, in that it shows only 20 normally distributed MotherAges for each FatherAge, instead of an infinite number. Likewise, the idealized model assumes FatherAges to be a (continuous) normal distribution, instead of just taking whole even-numbered age values as shown. Otherwise, it is a fair representation of how we imagine the population relationship between ages: it is positive and linear, with constant spread σ about the regression line following a normal pattern.



Population relationship expressed as $\mu_{\mathbf{y}} = \beta_{\mathbf{0}} + \beta_{\mathbf{1}}\mathbf{x}$

Notice that a new symbol "μ_y" has been used to model the relationship in the larger population. Because statistics concerns itself with drawing conclusions about populations, based on samples, we must always be sure to distinguish between parameters (describing populations) and statistics (describing samples). In the case of a regression line, we have already used the notation $\hat{y}$ to refer to the response predicted for the sample: $\hat{y} = b_0 + b_1 x$, where b_0 and b_1 are calculated from the sample data. The corresponding parameter is $\mu_y = \beta_0 + \beta_1 x$, the unknown population mean response to a given explanatory value x, which responds linearly with an unknown intercept β_0 and slope β_1.

[Note: the mean response μ_y is the same thing as expected response $E(y)$.]

Distribution of sample slope b_1

As always, we report the long-run behavior of a sample statistic by describing its distribution, specifically by telling its center, spread, and shape.

- **(Center:)** If the previously mentioned requirements are met—linear scatterplot, normally distributed residuals, and apparently constant spread about the line—then slope b_1 of the least squares line for a random sample of explanatory/response pairs has mean equal to the unknown slope β_1 of the least squares line for the population.

- **(Spread:)** The standard deviation of sample slope b_1 is

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}}$$

which we estimate with the standard error of b_1,

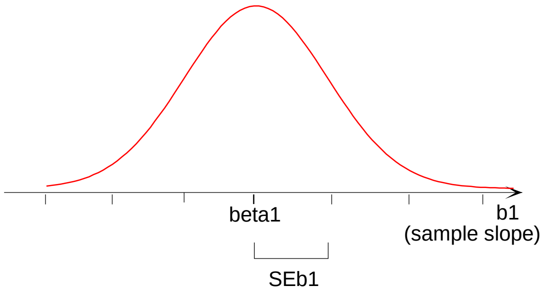
$$SE_{b_1} = \frac{s}{\sqrt{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}}$$

where s, the estimate for spread σ about the population regression line, measures typical residual size.

Although the above formula need not be used for calculations as long as software is available, it is worth examining SE_{b_1} to see how the residuals contribute to the spread of the distribution of sample slope. The appearance of s in the numerator of SE_{b_1} should make perfect intuitive sense: if the residuals as a group are small, then there is very little spread about the line and we should be able to pinpoint its slope fairly precisely. Conversely, if the residuals are large, then there is much spread about the line and there is a much wider range of plausible slopes. Note that the quantity $\sqrt{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}$, which appears in the denominator, measures *combined* distances of explanatory values from their mean. This will be larger for larger sample sizes, and so b_1 has less spread for larger samples. Again, our intuition tells us that we should be better able to pinpoint the unknown population slope β_1 if we obtain a sample slope from a larger sample.

- **(Shape:)** Finally, b_1 itself has a normal shape if the residuals are normal, or if the sample size is large enough to offset non-normality of the residuals.

The graph below depicts what we have established about the distribution of sample slope b_1 for large enough sample sizes: it is centered at population slope β_1, has approximate standard deviation SE_{b_1}, and follows a normal distribution.

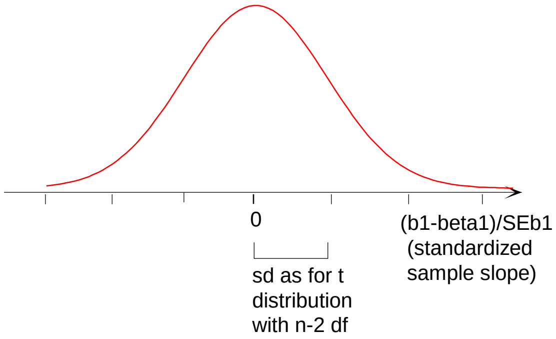


Distribution of standardized sample slope

Recall that in Chapters 12 and 13, when we standardized sample mean using sample standard deviation s instead of unknown population standard deviation σ, the resulting random variable $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ followed a **t** distribution instead of **z**. This could only be asserted if the sample size was large enough to offset any non-normality in the population distribution, so that the Central Limit Theorem could guarantee sample mean to be approximately normally distributed. In this chapter, we standardize sample slope b_1 using SE_{b_1}, calculated from s because σ is unknown. The resulting standardized slope

$$t = \frac{b_1 - \beta_1}{SE_{b_1}}$$

follows a **t** distribution, and its degrees of freedom are $n-2$, the same as for s. Again, this can only be asserted if b_1 follows a normal distribution, which is the case if the sample is large enough to offset non-normality in the residuals. Remember also that for large samples, the **t** distribution is virtually identical to that of **z**. The distribution of standardized sample slope is displayed below: centered at zero as is any **t** distribution, standard deviation subject to degrees of freedom which are determined by sample size (in particular, standard deviation close to 1 if the sample size is large enough to make **t** roughly the same as **z**), and bell-shaped like any **t** distribution.



Now that we know more about the behavior of b_1 relative to β_1, we will make use of the critical role played by β_1 in the relationship between explanatory and response variables, so as to set up a test for evidence of a relationship in the larger population. Because the construction of confidence intervals tends to be more intuitive than carrying out hypothesis tests, we start by setting up a confidence interval for the unknown slope of the linear relationship in the population. After that we will establish a procedure for testing the null hypothesis that slope for the population relationship is zero. In practice, it may make sense to carry out the test first, and then report the confidence interval for slope if there is statistical evidence of a relationship.

Inference about β_1

If the relationship between sampled values of two quantitative variables appears linear, then methods of Chapter 5 can be used to produce the line that best fits those sample values. For example, ages of students' fathers and mothers produced the following regression output.

```
Pearson correlation of FatherAge and MotherAge = 0.781
P-Value = 0.000

The regression equation is
MotherAge = 14.5 + 0.666 FatherAge
431 cases used 15 cases contain missing values
Predictor        Coef     SE Coef          T        P
Constant       14.542       1.317      11.05    0.000
FatherAge     0.66576     0.02571      25.89    0.000
S = 3.288       R-Sq = 61.0%     R-Sq(adj) = 60.9%
```

The fact that r is $+.781$ tells us there is a fairly strong positive relationship between x and y data values. Based on the fact that $b_1 = .666$, our best guess for how MotherAge responds to FatherAge is to predict that if one student's father is 1 year older than a second student's father, his mother would be .666 years older than the second student's mother. By now we know enough about behavior of samples to realize that there must be some margin of error attached to this slope. For every additional year of FatherAge in the population, does MotherAge tend to be an additional .666 years, give or take about .1 years? Or .666 years, give or take about 1 year? As usual, the size of the margin of error will supply important information. In the former case, having evidence that population slope is in the interval $(+.566, +.766)$ would convince us of a positive relationship, whereas in the latter case, where the range of plausible values $(-.334, +1.666)$ for unknown population slope straddles zero, we could not claim to have statistical evidence of a relationship. Knowing enough about the distribution of sample slope relative to population slope will help us find the answer to our earlier question about the relationship between ages. Thus, we are ready to begin the process of statistical inference to draw conclusions about the relationship between two quantitative variables in a larger population, based on sample data about those variables.

Confidence interval for β_1

Example

For a population of students' parents, what does the age of the father tell us about the age of a mother? Specifically, if one father is a year older than another, how much older (if at all) do we expect the mother to be?

The estimate b_1 for the unknown slope β_1 of the line that relates the variables MotherAge and FatherAge in the larger population is shown not only in the regression equation, but also as the coefficient of FatherAge in the second row of the output table.

```
Predictor        Coef     SE Coef          T        P
Constant       14.542       1.317      11.05    0.000
FatherAge     0.66576     0.02571      25.89    0.000
```

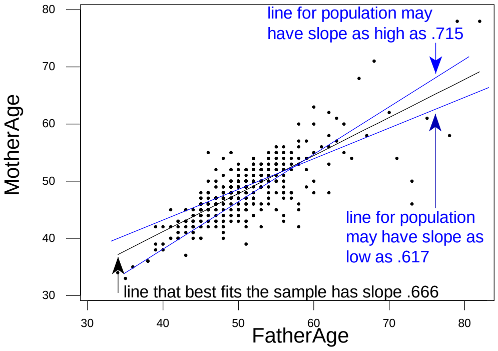
It is reported to five decimal places as .66576, and its standard error, .02571, appears in the next column. A 95% confidence interval for β_1 is constructed in the usual way, as

$$\text{estimate} \pm \text{margin of error}.$$

The estimate is of course b_1, and the margin of error is a multiple of the standard error SE_{b_1}, where the multiplier is the value of the relevant **t** distribution that corresponds to a symmetric area of 95%. As established in the previous section, sample slope b_1 follows the **t** distribution with $n-2$ degrees of freedom. The output shows our sample size to be 431, and so there are 429 degrees of freedom. With such a large sample, the t multiplier is virtually identical to the z multiplier for 95% confidence, which is approximately 2. Our 95% confidence interval for β_1 is

$$.666 \pm 2(.02471) = .666 \pm .049 = (.617, .715)$$

The fact that this interval contains only positive numbers supplies us with statistical evidence of a positive relationship between fathers' and mothers' ages in the population. More specifically, for every additional year of FatherAge, we are 95% confident that the corresponding value of MotherAge is an additional .617 to .715 years.



You may perhaps wonder why MotherAge doesn't increase by a full year for every increase of one year in FatherAge: as a student's father gets older, doesn't his or her mother have to age at exactly the same rate? It is important to recognize that ages are not being recorded as a time series, year by year for only one mother and father. Rather, we are thinking about an entire population of age pairs from which—at one point in time—we extract a sample of 431 *independent* age pairs. If one of these fathers is older than another by one year, then that mother may be older than the first mother, too, but not necessarily. On average, we expect her to be older than the first mother by about .666 years.

Independence of the observations from one another is an additional condition for our inference procedure methods to yield accurate results, and the sampling process should always be considered in case there may be a violation of this condition.

Example

Next lecture, we will look at the relationship between male students' heights and weights. The data must consist of height/weight pairs obtained randomly and independently from a population of male students. Methods developed in this chapter would not apply if our data consisted of height and weight measurements for the same student recorded each month over several years' time.

Motivated by the earlier example on the relationship between parents' ages, we now state our general confidence interval result.

95% Confidence Interval for β_1

An approximate 95% confidence interval for slope β_1 of the line that best fits the population of explanatory and response values, based on a random sample with *large* size n, is

$$\text{estimate} \pm \text{margin of error} = b_1 \pm 2(SE_{b_1}).$$

For a *small* sample size n, the approximate 95% interval is

$$b_1 \pm \text{multiplier}(SE_{b_1}).$$

where the multiplier is the value of the **t** distribution for $n-2$ degrees of freedom associated with 95% confidence (right-tail area under the curve is .025). This multiplier is greater than 2, but as long as there are at least six explanatory/response pairs in our sample, it will be no more than 3.

This interval is only appropriate if

- the scatterplot appears linear
- the sample size is large enough to offset any non-normality in the response values
- spread of responses is fairly constant over the range of explanatory values
- explanatory/response pairs are independent of one another

Hypothesis test about β_1

Our first step in learning to perform inference about proportions in Chapter 10 was to set up a confidence interval. By checking if the interval contained a proposed value of population proportion, we were able to make a rather informal decision as to whether that value was plausible, based on whether or not the value was contained in the interval. In Chapter 11 we learned to carry out a formal test of hypotheses about unknown population proportion, following five basic steps.

Similarly, we used the confidence interval in our example above to informally conclude that the value of β_1 is not zero. A more formal way to reach this conclusion is by carrying out a test of hypotheses.

As with our other hypothesis test procedures about the relationship between two variables, there are two formulations of the null and alternative hypotheses: one about a key parameter; the other about the variables and their relationship. When there are two quantitative variables of interest, the null hypothesis states that the slope β_1 of the least squares line for the population is zero. Equivalently, it claims that the variables are *not* related, because the equation $\mu_y = \beta_0 + \beta_1 x$ reduces to $\mu_y = \beta_0$ when β_1 is zero, and the mean population response does not depend on the so-called explanatory variable x. The alternative may be one-sided or two-sided. The two-sided alternative, $\beta_1 \neq 0$, is equivalent to the statement that the variables *are* related in the population. The one-sided alternatives $\beta_1 > 0$ or $\beta_1 < 0$ are more specific in that they express a claim not only that the variables are related, but also with regards to the direction of the purported relationship. In order to determine which formulation is appropriate, the wording and background of a problem must be carefully considered.

Example

Is there statistical evidence of a relationship between FatherAge and MotherAge?

We could equivalently pose the question as $H_0 : \beta_1 = 0$ vs. $H_a : \beta 1 \neq 0$ where $beta_1$ is the slope of the line that relates ages of fathers and mothers for the entire population of students. Because common sense would tell us to expect a positive relationship, we may go so far as to formulate the alternative as one-sided: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 > 0$.

Example

Based on information from a sample of 4 states, can we conclude that for all states there is a negative relationship between percentage voting democratic and percentage voting republican?

In this case we would write $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 < 0$.

Example

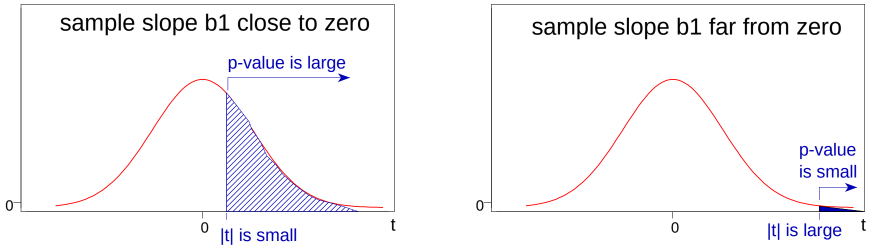
A website called "ratemyprofessor.com" reports students' ratings of their professors at universities around the country. These are unofficial in that they are not monitored by the universities themselves. Besides listing average rating of the professors' teaching on a scale of 1 to 5, where 1 is the worst and 5 is the best, there is also a rating of how easy their courses are, where 1 is the hardest and 5 is the easiest. Is there a relationship between the rating of teaching and the rating of ease?

Offhand, we may suspect that students would favor easy teachers, in which case the relationship would be positive. On the other hand, teachers who are more conscientious may maintain higher standards, not just for their students but also for themselves. In this example, because the direction could really go either way, we should keep a more general two-sided alternative, and write $H_0 : \beta_1 = 0$ vs. $H_a : \beta 1 \neq 0$.

We have already established that the distribution of sample slope b_1, if certain conditions are met, is normal with mean β_1 and approximate standard deviation SE_{b_1}. Under the null hypothesis that $\beta_1 = 0$, the standardized test statistic

$$t = \frac{b_1 - 0}{SE_{b_1}}$$

follows a **t** distribution with $n-2$ degrees of freedom. If the sample slope b_1 is relatively close to zero (taking sample size and spread into account), then the standardized test statistic t is not especially large, and so the p-value is not small and there is no compelling evidence of a non-zero population slope β_1. Thus, if b_1 isn't large enough, we cannot produce evidence that the two quantitative variables are related in the larger population. Conversely, if sample slope b_1 is relatively far from zero, then t is large, the p-value is small, and we have statistical evidence that the population slope β_1 is *not* zero. In other words, a large t results in a small p-value and a conclusion that the variables *are* related. If a one-sided alternative has been formulated and the sample slope b_1 tends in the direction claimed by that alternative, then the p-value is the one-tailed probability of t being as extreme as the one observed.



Example

Let's revisit the output for the regression of MotherAge on FatherAge, carrying out a five-step test of hypotheses. This will require us to focus our attention on the size of the t statistic and p-value.

```
Predictor        Coef     SE Coef         T       P
Constant       14.542       1.317     11.05   0.000
FatherAge     0.66576     0.02571     25.89   0.000
```

1. The null hypothesis states that the slope β_1 of the line that relates MotherAge to FatherAge in the population is zero; alternatively, it may state that MotherAge and FatherAge are not related. Since common sense would suggest a positive relationship between the two variables, our alternative hypothesis would be that $\beta_1 > 0$; alternatively, it could make a more general

claim that MotherAge and FatherAge are related for the general population of parents' ages, that is, that $\beta_1 \neq 0$.

2. When we set up a confidence interval for unknown population slope β_1, we noted that b_1 and its standard error SE_{b_1} are reported in the *second* row of the regression table. (The first row concerns intercept b_0, which is not of particular interest to us, since we tend not to perform inference about intercept β_0. Remember that it is the slope that provides key information about if and how the explanatory and response values are related.) Thus, the standardized sample slope t is reported as 25.89. Notice that the t statistic is easily calculated from b_1 and SE_{b_1}:
$$t = \frac{b_1 - 0}{SE_{b_1}} = \frac{.66576 - 0}{.02571} = 25.89.$$
(Remember that 0 is subtracted from sample slope b_1 because for random samples b_1 is centered at population slope β_1, which is proposed to be 0 in the null hypothesis.) For a large sample like this, our cut-off point for "large" values of t is like that for z, namely 2. Obviously 25.89 is extremely large compared to 2.

3. The p-value corresponding to our t statistic is shown to be 0.000 (in the second row, not the first).

4. Just as t was extremely large, the p-value is extremely small. The fact that the p-value is so small tells us that obtaining a sample slope as far from zero as $+.66576$ would be extremely unlikely if population slope were zero, and so we conclude population slope is *not* zero. This p-value actually corresponds to a two-sided alternative; technically, if we suspected all along that the slope would be positive and formulated the alternative as $\beta_1 > 0$, the p-value should be half of the one shown in the output. This only serves to strengthen our conclusion that ages are related.

5. To summarize, we have strong statistical evidence that MotherAge and FatherAge have a positive relationship, not just in our sample, but also in the larger population of students.

Motivated by the example above, we summarize the process of testing for a relationship between two quantitative variables, by testing the null hypothesis that slope of the regression line for the population of explanatory and response values is zero.

Hypothesis Test about β_1

Just as for any hypothesis procedure, there are five basic steps to test for a relationship between two quantitative variables in the population of interest, based on a random sample of size n.

1. Assuming the relationship (if it exists) between two quantitative variables to be linear rather than curved, we test the null hypothesis that the variables are *not* related, which is equivalent to the claim
$$H_0 : \beta_1 = 0$$
where β_1 is the slope of the population least squares regression line. The alternative may state more generally that the two variables *are* related, which is equivalent to the claim
$$H_a : \beta_1 \neq 0$$
or a more specific one-sided alternative may be formulated as $H_a : \beta_1 < 0$ if we suspect in advance that the relationship is negative, or $H_a : \beta_1 > 0$ if we suspect the relationship is positive.

2. Software should be used to produce the standardized sample slope $t = \frac{b_1 - 0}{SE_{b_1}}$, which, when conditions below are met, follows a **t** distribution with $n - 2$ degrees of freedom. This t statistic is a standardized measure for how far sample slope b_1 is from zero.

3. The p-value to accompany the t test statistic is the probability of a **t** random variable being as extreme as the one observed. It is reported alongside t as part of the regression output. A small p-value suggests that t is unusually extreme if the null hypothesis were true; that is, that the sample slope could be considered unusually steep if it were coming from a population where the explanatory and response variables were not related.

4. If the p-value is small, we reject the null hypothesis of no relationship (equivalent to rejecting the claim that slope β_1 for the population is zero). If the p-value is not small, we conclude that the null hypothesis may be true.

5. Conclusions should be stated in context: if the null hypothesis has been rejected, we conclude that there is statistical evidence of a relationship between the explanatory and response variables. If it has been rejected against a one-sided alternative, we conclude there is evidence of a negative or of a positive relationship, depending on how the alternative has been expressed. If the null hypothesis has not been rejected, we conclude there is not enough statistical evidence to convince us of a relationship between the two quantitative variables.

Results of the above test are only valid if the following conditions are met:

- the scatterplot appears linear
- the sample size is large enough to offset any non-normality in the response values
- spread of responses is fairly constant over the range of explanatory values
- explanatory/response pairs are independent of one another

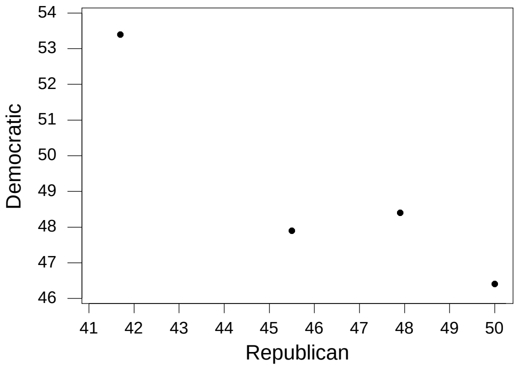
In the previous example, not only did we have strong evidence of a relationship (by virtue of the p-value being close to zero), but we also could assert that the relationship was strong (by virtue of the correlation r being .78, which is pretty close to one). It is nevertheless possible to produce weak evidence of a strong relationship, or strong evidence of a weak relationship. These possibilities will be explored in the following examples. We will also consider an example where there is no statistical evidence of a relationship.

Example

While most voters in a presidential election vote for the democratic or republican candidate, other parties do account for a small percentage of the popular vote in each state. The table below looks at the relationship between percentages voting democratic and republican in the year 2000 for just a few states.

State	Democratic	Republican
Alabama	48.4	47.9
California	53.4	41.7
Ohio	46.4	50.0
Minnesota	47.9	45.5

The points in the scatterplot below do appear to cluster around some straight line, rather than a curve. The line has a negative slope because when the percentage voting republican is low, then the percentage voting democratic is high, and vice versa.



When a regression is carried out, the correlation r is found to be quite close to -1 ($r = -.922$), suggesting a strong negative relationship. On the other hand, the p-value (.078) may not necessarily be considered small enough to provide statistical evidence of a relationship.

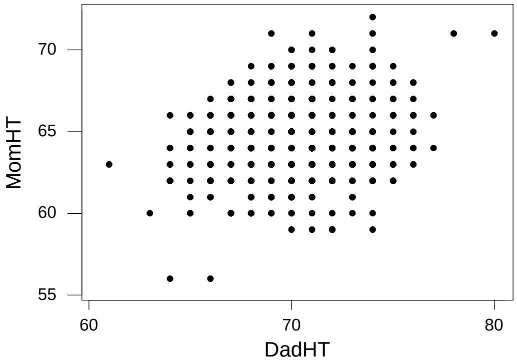
```
Pearson correlation of dem and rep = -0.922
P-Value = 0.078
%Pearson correlation of Democratic and Republican = -0.910
%P-Value = 0.090
```

Due to the small sample size of only 4, we do not have especially strong evidence of a relationship in the larger population of states, even though for the sample the relationship is apparently quite strong. In other words, we have *weak evidence of a strong relationship* between percentage voting republican and percentage voting democratic. A larger sample of states would have certainly supplied very strong evidence of such a relationship.

In the preceding example, we saw that although a linear relationship between two quantitative variables may be quite strong, with too small a sample we may only produce weak evidence of that relationship. In the next example, we see that with a large sample we may produce very strong evidence of a rather weak relationship in the population.

Example

As a contrast to the rather strong relationship between MotherAge and FatherAge, we now look at the relatively weak relationship between MotherHt and FatherHt. A scatterplot for the latter is shown below.



There is apparently a slight tendency for relatively short fathers to be paired with relatively short mothers, and for relatively tall fathers to be paired with relatively tall mothers. Since height is such a minor factor when it comes to couples' compatibility, the relationship is naturally quite weak. According to the output below, the correlation is only $r = .225$. On the other hand, a test of whether the slope of the regression line could be zero for the general population of parents' heights produces a very large t statistic (4.79) and a very small p-value (0.000).

```
Pearson correlation of MomHT and DadHT = 0.225

The regression equation is
MomHT = 50.4 + 0.200 DadHT
431 cases used 15 cases contain missing values
Predictor        Coef     SE Coef          T        P
Constant       50.431       2.936      17.18    0.000
DadHT         0.20019     0.04178       4.79    0.000
S = 2.551       R-Sq = 5.1%      R-Sq(adj) = 4.9%
```

In this case, by virtue of a large sample (431 height pairs), we are able to produce very strong evidence of a relationship in the general population of parents' heights, but the relationship itself is rather weak.

Lecture 31

Other Interval Estimates in Regression

In the previous lecture, we learned two important regression inference procedures: testing for statistical evidence of a relationship between the two quantitative variables of interest, and estimating the slope of the line that relates those variables in the larger population. For practical purposes, two other types of estimation are quite common.

Example

Reassessment of property values in Allegheny County, Western Pennsylvania, in 2002 were extremely controversial, and some property owners believed the assessment was too high, resulting in higher taxes. Suppose a homeowner was told that his land (not including house) was reassessed at $40,000, and he wants to contest it as being unreasonably high.

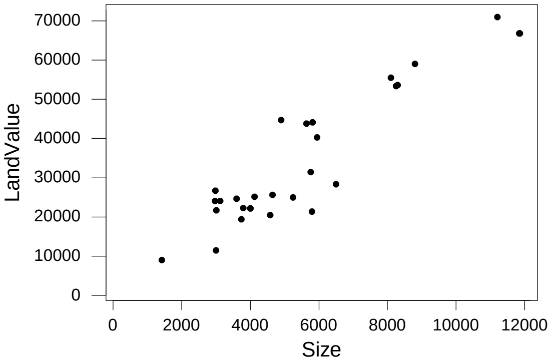
As a first step, he could look at a sample of assessment values of other properties in his neighborhood. A sample of 29 land values in the neighborhood have mean $34,624 and standard deviation $17,494. A value of $40,000 at this point doesn't seem unusually high, at $\frac{40000-34624}{17494} = .31$ standard deviations above the mean.

```
Variable            N        Mean      Median      TrMean       StDev    SE Mean
LandValu           29       34624       25600       34226       17494       3249
Variable      Minimum     Maximum          Q1          Q3
LandValu         9000       71000       22200       49050
```

But the homeowner suspects his property, at 4,000 square feet, is smaller than average, and so he researches size of those neighborhood properties.

```
Variable            N        Mean      Median      TrMean       StDev    SE Mean
Size               29        5619        4900        5544        2755        512
Variable      Minimum     Maximum          Q1          Q3
Size             1425       11853        3671        7299
```

By now we have established that his property's assessed value ($40,000) is higher than average ($34,624), although its size (4,000 square feet) is smaller than average (5,619 square feet). This in itself is not enough evidence to argue that the assessment is unfair; the homeowner needs to show that in general the relationship between size and value is such that $40,000 would be an unreasonably high value for a lot of size 4,000 square feet. First let's look at a scatterplot of the 29 size and value pairs:



Certainly the relationship appears to be positive, linear, and quite strong, suggesting that a smaller-than-average lot should be given a smaller-than-average assessment. This is confirmed by the output below, which shows the correlation to be quite close to one. Furthermore, the p-value is close to zero, providing evidence that this relationship should hold in the larger population from which the sample of land sizes and values was obtained.

```
Pearson correlation of Size and LandValue = 0.927
P-Value = 0.000
```

An option when using software to perform a regression is to request a "prediction interval for new observation", with results shown below for an observed size of 4,000 square feet:

```
Predicted Values for New Observations
New Obs     Fit     SE Fit         95.0% CI             95.0% PI
1         25094       1446   (   22127,   28060)  (   11066,   39121)
Values of Predictors for New Observations
New Obs      Size
1            4000
```

The output includes two very different intervals: one labeled "95.0% CI" that ranges roughly from $22,000 to $28,000, and one labeled "95.0% PI" that ranges roughly from $11,000 to $39,000. Both intervals are centered at $25,094, the predicted value for a lot of size 4,000 square feet. The first of the intervals is *not* especially relevant to the homeowner, because it presents a set of plausible values for *mean* value of *all* 4,000-square-foot lots in the neighborhood. The second interval reports a 95% prediction interval for the value of one individual lot whose size is 4,000 square feet. Since the assessed value of $40,000 falls above the interval ($11,066, $39,121), the homeowner does have statistical evidence that the assessment is unusually high, given the size of his lot.

In order to put new inference skills in perspective, the following example includes a variety of estimates: estimating an individual or mean value of a quantitative variable; estimating an individual or mean response for a given explanatory value; and estimating an individual or mean response for a different explanatory value. We present a series of questions, all alike in that they seek estimates concerning male weight, but all different in terms of whether an estimate is sought for an individual or a mean, and also in terms of what height information, if any, is provided.

Example

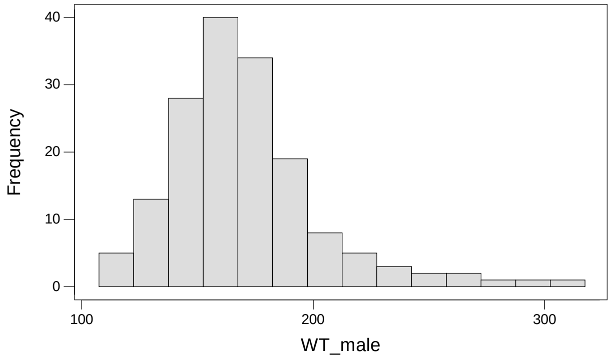
1. Based on a sample of male weights, how do we estimate weight of an individual male?
2. Based on a sample of male weights, how do we estimate mean weight of all males?
3. Based on a sample of male heights and weights, how do we estimate weight of an individual 71-inch-tall male?
4. Based on a sample of male heights and weights, how do we estimate mean weight of all 71-inch-tall males?
5. Based on a sample of male heights and weights, how do we estimate weight of an individual 76-inch-tall male?
6. Based on a sample of male heights and weights, how do we estimate mean weight of all 76-inch-tall males?

Estimating an individual weight with no height information

In Chapter 2, we learned that if a distribution is roughly normal, and we know its mean and standard deviation, we can report a range for most of its values using the 68-95-99.7 Rule, which is based on a normal distribution. For example, the output below shows male weights to have mean 170.83 and standard deviation 33.06.

```
Variable            N          N*       Mean     Median     TrMean      StDev
WT_male           162           2     170.83     165.00     168.24      33.06

Variable      SE Mean     Minimum    Maximum         Q1         Q3
WT_male          2.60      115.00     315.00     150.00     185.00
```

If the shape of the distribution of weights is approximately normal, then about 95% of the time any one individual weight should fall within 2 standard deviations of the mean, from $170.83-2(33.06)$ to $170.83+2(33.06)$; that is, in the interval $(104.71, 236.95)$. The accuracy of this interval is not necessarily to be trusted, because the shape of the distribution of weights shown in the histogram below is *not* entirely normal, but is rather right-skewed.



Although weights are often purported to be normal for specific age and gender groups, the reality is that most populations include individuals with weights that are unusually high to the point where they cannot be balanced out by unusually low weights. In our sample, for instance, the highest weight (315) is $(315-170.83)/33.06 = 4.4$ standard deviations above the mean; a man would have to weigh just 25 pounds to be this many standard deviations *below* the mean! In fact, the lowest weight (115) has a z-score of $(115-170.83)/33.06 = -1.7$, so it is only 1.7 standard deviations below the mean.

Estimating mean weight with no height information

In Chapter 12, we learned to perform inference about the mean of a single quantitative variable. These methods can be used to set up a confidence interval for the mean weight of all male college students, based on a sample of weights.

```
One-Sample T: WT_male
Variable          N      Mean     StDev   SE Mean         95.0% CI
WT_male         162    170.83     33.06      2.60  (  165.70,  175.96)
```

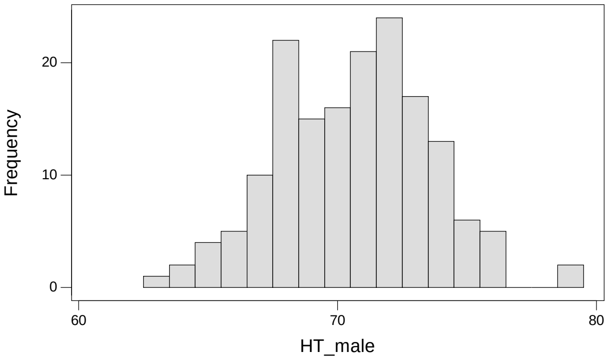
Thus, a 95% confidence interval for the *mean* weight of all male college students is (165.70, 175.96). Notice how much narrower this interval is than the interval that should contain an individual weight. The interval for individuals ranged all the way from about 105 to 237 pounds, with a width of 132 pounds. In contrast, the interval for mean weight ranged only from about 166 to 176 pounds, with a width of only 10 pounds. It is much harder to pinpoint an individual as opposed to a mean value. Remember that the spread of all values is estimated with s, while the spread of sample mean is estimated with $\frac{s}{\sqrt{n}}$. Whereas the non-normality of the distribution of weights presented a problem in setting up a range for 95% of individual weights, by virtue of the Central Limit Theorem, the large sample size guarantees sample mean weight to be approximately normal, and so this interval should be quite accurate.

Including Height Information

In fact, the confidence interval above is of limited usefulness, because instead of asking, "What is a typical weight for any male college student?" we would be more inclined to wonder, "What is a typical weight for a male college student who is x inches tall?" A range of plausible values for the mean weight of *all* male college students is no longer appropriate if we are specifically interested in what is plausible for the mean of, say, all *71-inch-tall* male college students. In order to really do justice to the variable weight, the variable height should be taken into account. Inference for regression can be used to produce a range of plausible values for the *mean* weight of all male college students of a *given* height. Along the way, we will also take a look at the range of plausible values for the weight of an *individual* male college student of a given height, in order to contrast such intervals.

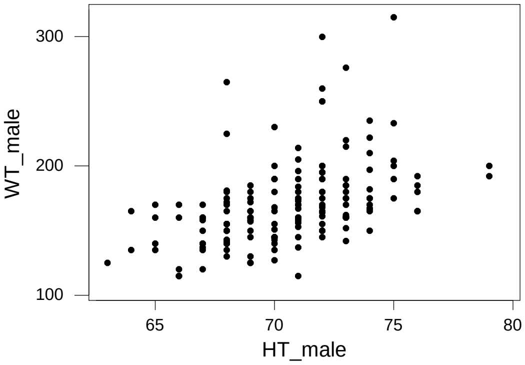
We have already examined the distribution of weights alone. Now let's examine the heights of our sample of 162 male college students, and look at the relationship between height and weight. Then, we'll produce a 95% prediction interval for the weight of an individual 71-inch-tall male, along with a 95% confidence interval for the mean weight of all 71-inch-tall males. These in turn will be compared to 95% prediction and confidence intervals for a given height of 76 inches. In the end, we will contrast these to the intervals already discussed, which do not take height into account.

```
Variable             N          N*        Mean      Median      TrMean       StDev
HT_male            163           1      70.626      71.000      70.626       2.940

Variable       SE Mean     Minimum     Maximum          Q1          Q3
HT_male          0.230      63.000      79.000      68.000      73.000
```



Based on the descriptive statistics and histograms, we can say that sampled heights appear normally distributed (symmetric, bulging in the middle and tapering at the ends), with mean 70.626 and standard deviation 2.940.

The separate summaries and histograms produced so far do not supply any information about the relationship between height and weight; this requires a regression procedure, starting off with a scatterplot for display.



The scatterplot shows a moderately strong positive relationship between male heights and weights. The right skewness that we saw in the histogram of weights is seen in the scatterplot as a looser scattering of points in the higher weight ranges. Fortunately, the sample size of 163 is large enough to offset this non-normality. Next we look at regression output.

```
The regression equation is
WT_male = - 188 + 5.08 HT_male
162 cases used 2 cases contain missing values
Predictor        Coef     SE Coef          T        P
Constant      -187.55       56.12      -3.34    0.001
HT_male        5.0759      0.7942       6.39    0.000
S = 29.60       R-Sq = 20.3%     R-Sq(adj) = 19.8%
```

The fact that the correlation r is the positive square root of .203, or +.45, tells us that the relationship is of moderate strength for the sample of height/weight values. Height does tell us something about weight, but its prediction power is far from perfect. The fact that $p = 0.000$ tells us we have very strong evidence that a relationship holds in the larger population from which the sample was taken. And the fact that the slope is +5.08 tells us that if one male college student is 1 inch taller than another, his weight should be about 5 pounds more.

Estimating individual weight for a given height of 71 inches

Output for a request of confidence and prediction intervals for a height of 71 inches will help us estimate weight of a particular male student who is 71 inches tall.

```
Predicted Values for New Observations
New Obs     Fit     SE Fit         95.0% CI             95.0% PI
1       172.83       2.35   (  168.20,  177.47)  (  114.20,  231.47)
Values of Predictors for New Observations
New Obs   HT_male
1            71.0
```

When we only considered mean and standard deviation for weights, with no additional information provided by heights, we could say that if the distribution were normal, then 95% of the time *any* individual male weight should fall within 2 standard deviations of the mean, in the interval (104.71,236.95). With height taken into account, the prediction interval "95.0% PI" reported in the regression prediction output tells us that 95% of the time the weight for a 71-inch-tall individual male should fall in the interval $(114.20, 231.47)$. This interval is about 15 pounds narrower (and therefore more precise) than the above interval that did not utilize information about height.

Estimating mean weight for a given height of 71 inches

If our goal is to produce a range of plausible values for *mean* height of all 71-inch-tall male college students, it is the confidence interval, labeled "95.0% CI", that is relevant. We are 95% confident that the mean weight of all 71-inch-tall male college students is somewhere between 168.20 and 177.47 pounds. Once again, we see a dramatic difference between the extent to which we can pinpoint an individual (interval width $231.47 - 114.20 = 117.27$) and a mean for all individuals (interval width $177.47 - 168.20 = 9.27$).

Since heights and weights have a positive relationship, we expect that weight estimates for taller men should be higher.

Estimating individual weight for a given height of 76 inches

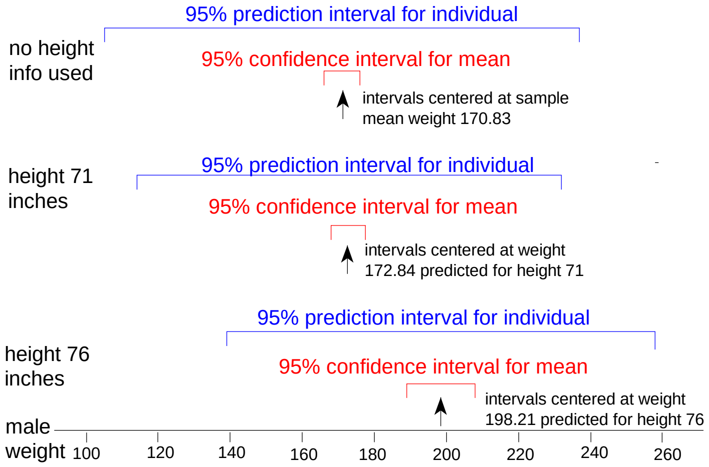
According to the output below, if an individual male is 76 inches tall, we predict his weight to be somewhere between 138.97 and 257.45 pounds. Since 76 is 5 inches taller than 71, and since the slope $b_1 = 5.08$ tells us that each additional inch in height is accompanied by about 5 more pounds in weight, this entire interval is about 25 pounds higher than the interval of predicted weight for a height of 71 inches. The width of this interval is $257.45 - 138.97 = 118.48$, just slightly wider than the interval for an individual 71-inch-tall male (interval width 117.27 pounds).

```
Predicted Values for New Observations
New Obs     Fit     SE Fit         95.0% CI             95.0% PI
1       198.21       4.88   (  188.58,  207.84)  (  138.97,  257.45)
Values of Predictors for New Observations
New Obs   HT_male
1            76.0
```

Estimating mean weight for a given height of 76 inches

The output also includes a 95% confidence interval for the mean weight of all 76-inch-tall men. The width of this interval ($207.84 - 188.58 = 19.26$) is more than twice the width of the 95%

confidence interval for the mean weight of all 71-inch-tall men (9.27). In the next section, we will see that this difference is due to the fact that 76 is much further from the mean height than 71 is. For now, we summarize our interval estimates with the display below.


95% prediction interval for individual
no height
info used
95% confidence interval for mean
intervals centered at sample
mean weight 170.83
95% prediction interval for individual
height 71
inches
95% confidence interval for mean
intervals centered at weight
172.84 predicted for height 71
95% prediction interval for individual
height 76
inches
95% confidence interval for mean
intervals centered at weight
198.21 predicted for height 76
male
weight
100
120
140
160
180
200
220
240
260

Especially in the case of prediction intervals, if there is a substantial relationship between two quantitative variables, we can produce a narrower interval if we include information in the form of a given explanatory value. Confidence intervals for means will be considerably narrower than prediction intervals for individuals. In the next section we will see that sample size plays an important role in the width of the confidence interval. Naturally enough, if the relationship is positive, then for higher values of the explanatory variable, both confidence and prediction intervals are centered at a higher response.

Role of s in Confidence and Prediction Intervals

As is the case for any interval estimates, our confidence and prediction intervals are of the form

$$\text{estimate} \pm \text{margin of error} = \text{estimate} \pm \text{multiplier} * \text{standard error}.$$

No matter if we are constructing a prediction interval for an individual response, or a confidence interval for the mean response to a given explanatory value, the estimate at the center of our interval is the regression line's predicted response for that explanatory value.

Example

The regression line for estimating male weight from height is

```
WT_male = - 188 + 5.08 HT_male.
```

The confidence and prediction intervals for male weight when height is 71 are both centered at the estimate $-188 + 5.08(71) = 172.68$. The confidence and prediction intervals for male weight when height is 76 are both centered at the estimate $-188 + 5.08(76) = 198.08$.

Both confidence and prediction intervals are centered at the predicted response $\hat{y} = b_0 + b_1 x*$, but the confidence interval for mean response is narrower. When sample size is large, the prediction interval extends roughly $2s$ on either side of the predicted response. If there were no relationship between explanatory and response values, this interval would be no different from the interval that extends two ordinary standard deviations in y (s_y) on either side of the mean response. If there is a strong relationship between explanatory and response values, this interval is noticeably more precise than the interval obtained without taking explanatory value into account.

Example

For the regression of male weight on height, based on a large sample of 162 height/weight pairs, the regression output showed $s = 29.6$. The prediction interval should have a margin of error equal to roughly twice this, and so its entire width should be about four times 30, or 120.

```
New Obs     Fit     SE Fit         95.0% CI             95.0% PI
1        172.83       2.35   (  168.20,  177.47)  (  114.20,  231.47)
Values of Predictors for New Observations
New Obs   HT_male
1            71.0

New Obs     Fit     SE Fit         95.0% CI             95.0% PI
1        198.21       4.88   (  188.58,  207.84)  (  138.97,  257.45)
Values of Predictors for New Observations
1            76.0
```

In fact, the width of the prediction interval for weight when height equals 71 is $231.47 - 114.20 = 117.27$ and the width of the prediction interval for weight when height equals 76 is $257.45 - 138.97 = 118.48$. Both of these are quite close to our *ad hoc* calculation of 120.

When sample size is large and a confidence interval for mean response is desired for an explanatory value that is close to the mean $\bar{x}$, this interval extends roughly $2\frac{s}{\sqrt{n}}$ on either side of the predicted response.

Example

Since the mean male height is 70.626, as shown in our summary output, a height of 71 is close to the mean, and so for our sample of size 162, the standard error should be roughly $\frac{s}{\sqrt{n}} = \frac{29.6}{\sqrt{162}} = 2.3$. If our confidence interval for mean weight extends roughly 2 standard errors on either side of the predicted weight 172.83, its width should be about $4(2.3) = 9.2$. In fact, the 95% confidence interval has width $177.47 - 168.20 = 9.27$.

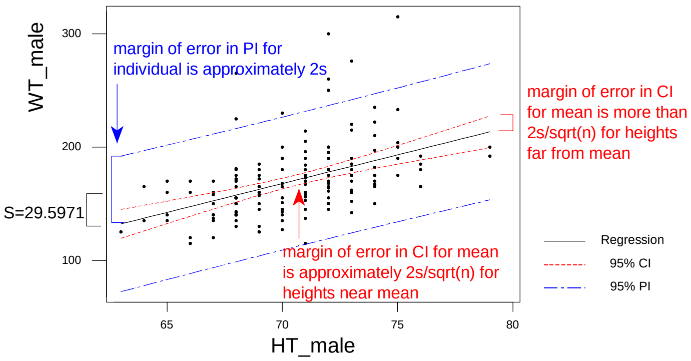
```
Variable             N         N*       Mean     Median     TrMean      StDev
HT_male            163          1     70.626     71.000     70.626      2.940
WT_male            162          2     170.83     165.00     168.24      33.06

New Obs     Fit     SE Fit         95.0% CI             95.0% PI
1        172.83       2.35   (  168.20,  177.47)  (  114.20,  231.47)
Values of Predictors for New Observations
New Obs   HT_male
1            71.0
```

On the other hand, *when predicting the mean response to an explanatory value far from the mean of all explanatory values*, the standard error is considerably larger than $\frac{s}{\sqrt{n}}$.

Example

A height of 76 inches is rather far from the mean height of 70.626. The confidence interval for mean weight of all 76-inch-tall men has a width of $207.84 - 188.58 = 19.26$, which is more than eight times the standard error, 2.3, instead of just four times, as was the case for estimating mean weight when height was 71, close to the mean of all heights. The illustration shows that whereas prediction interval width remains fairly uniform throughout the range of explanatory values, the confidence interval band widens considerably for explanatory values far below or above average.

```
New Obs      Fit      SE Fit          95.0% CI              95.0% PI
1        198.21       4.88    (  188.58,  207.84)  (  138.97,  257.45)
Values of Predictors for New Observations
1             76.0
```



These rough estimates are presented here merely as a reference point; in practice, the precise prediction interval and confidence interval should be found using software. Sample size plays its usual role, in that smaller samples result in wider intervals.

Exercise: Find two quantitative variables from our survey, summarize their relationship as in Chapter 5, and then test $H_0 : \beta_1 = 0$. State your conclusions in terms of the variables of interest.