6 Inference

6.1 Hypothesis Testing

Suppose that a real-world process is modeled by a binomial distribution for which we know n but do not know π. Examples abound.

Example 6.1.1

1. We have said that a fair coin is equally likely to be heads or tails when tossed. But now suppose we have a coin and toss it 100 times. How do we know it is fair? That is, how do we know $\pi = 0.5$?
2. A factory produces the ubiquitous widget. It claims that the probability that any widget is defective is less than 0.1%. We receive a shipment of widgets. We wonder whether the claim about the defective rate is really true. If we test 100 widgets, this is an example of a binomial experiment with $n = 100$ and π unknown.
3. A National Football League team is trying to decide whether to replace its field goal kicker with a new one. The current kicker makes about 30% of his kicks from 45 yards out. The team tests the new kicker by asking him to try 20 kicks from 45 yards out. This might be modeled by a binomial distribution with $n = 20$ and π unknown. The team is hoping that $\pi > .3$.
4. A standard test for ESP works as follows. A card with one of five printed symbols is selected without the person claiming to have ESP being able to see it. The purported psychic is asked to name what symbol is on the card while the experimenter looks at it and "thinks" about it. A typical experiment consists of 25 trials. This is an example of a binomial experiment with $n = 25$ and unknown π. The experimenter usually believes that $\pi = .2$.

In each of the instances of Example 6.1.1 we have a hypothesis about π that we could be considered to be testing. In the four cases we could be considered to be testing the hypotheses $\pi = .5$, $\pi \leq 0.001$, $\pi \leq 0.3$, and $\pi = .2$. A hypothesis proposes a possible state of affairs with respect to a probability distribution governing an experiment that we are about to perform. There are a variety of kinds of hypotheses that we might want to test.

1. A hypothesis stating a fixed value of a parameter: $\pi = .5$.

2. A hypothesis stating a range of values of a parameter: $\pi \leq .3$.

3. A hypothesis about the nature of the distribution itself: X has a binomial distribution.

To test the hypothesis that the coin is fair ($\pi = .5$) we must actually collect data. Suppose that we toss the coin $n = 100$ times and get $x = 40$ heads. What should we conclude about our hypotheses? The first thing to note is that we cannot conclude anything with certainty in this case. Any value of $x = 0, 1, \ldots, 100$ is consistent with both $\pi = 0.5$ and any other value of π. However, if the coin really is fair, some results for x are more surprising than others. In this case, for example, if our hypothesis is true, then $P(X \leq 40) = 0.02844$, so we would only get 40 or fewer heads about 2.8% of the times that did this test. In other words, getting only 40 heads is pretty unusual, but not extremely unusual. This gives us some evidence to suggest that the coin in biased. After all, one of two things must be true. Either

- the coin is fair ($\pi = 0.50$) and we were just "unlucky" in our particular 100 tosses, or
- the coin is not fair, in which case the probability calculation we just did doesn't apply to the coin.

That in a nutshell is the logic of a **statistical hypothesis test**. We will learn a number of hypothesis tests, but they all follow the same basic outline.

Step 1: State the null and alternative hypotheses

In a typical hypothesis test, we pit two hypotheses against each other.

1. **Null Hypothesis.** The null hypothesis, usually denoted $\mathbf{H_0}$, is generally a hypothesis that the data analysis is intended to investigate. It is usually thought of as the "default" or "status quo" hypothesis that we will accept unless the data gives us substantial evidence against it.

2. **Alternate Hypothesis.** The alternate hypothesis, usually denoted $\mathbf{H_1}$ or $\mathbf{H_a}$, is the hypothesis that we are wanting to put forward as true if we have sufficient evidence against the null hypothesis.

In the example of the supposedly fair coin, it is clear that the hypotheses should be

$$\begin{aligned} \mathbf{H_0}&: \quad \pi = 0.5 \\ \mathbf{H_a}&: \quad \pi \neq 0.5 \end{aligned}$$

The null hypothesis simply says that the coin is fair while the alternate hypothesis says that it is not. We want to choose between these two hypotheses. In this example, the alternate hypothesis is **two-sided.** There are also situations when we wish to consider a

one-sided alternate hypothesis. Consider the ESP example. Our null hypothesis is surely that the subject cannot do better than chance ($\pi = .2$) but our alternate hypothesis is that the subject can do better than chance ($\pi > .2$). In our particular test, we do not allow for the possibility that the the subject somehow typically does worse than chance (although this is logically possible).

Step 2: Calculate a test statistic

In our example, we compute the number of heads (40). This is the number that we will use to test our hypothesis. The number 40 in this instance is called a statistic. Since we use this statistic to test our hypothesis, we will sometimes call it a **test statistic**. In fact we will use the term statistic in two different ways. In this case, the number 40 is a specific value that is computed from the data. But also, 40 is the value of a certain random variable that is computed from the experiment of tossing a coin 100 times. We will refer to both the random variable and its value as statistics. In keeping with our notation for random variables and data, upper-case letters will denote random variables and lower-case letters their particular values.

A test statistic should be some number that measures in some way how true the null hypothesis looks. In this case, a number near 50 is in keeping with the null hypothesis. The farther x is from 50, the stronger the evidence against the null hypothesis.

Step 3: Compute the p-value

.

Now we need to evaluate the evidence that our test statistic provides. To do this requires that we think about our statistic as a random variable. In the case of the supposedly fair coin, our test statistic $X \sim \mathsf{Binom}(100, \pi)$. As a random variable, our test statistic has a distribution. The distribution of the test statistic is called its **sampling distribution**.

Now we can ask probability questions about our test statistic. The general form of the question is "How unusual would my test statistic be if the null hypothesis were true?" To do this, it is important that we know about the distribution of X when the null hypothesis is true. In this case, $X \sim \mathsf{Binom}(100, 0.5)$. So how unusual is it to get only 40 heads? Assuming that the null hypothesis is true (i.e., that the coin is fair),

$$\mathrm{P}(X \leq 40) = \texttt{pbinom(40,100,.5)} = 0.0284\,,$$

and

$$\mathrm{P}(X \geq 60) = \texttt{1 - pbinom(59,100,.5)} = 0.0284\,.$$

So the probability of getting a test statistic at least as extreme (unusual) as 40 is 0.0568. This probability is called a **p-value**.

There is some subtlety to the above computation and we shall return to it.

Step 4: Draw a conclusion

Drawing a conclusion from a p-value is a judgment call and it is a scientific rather than mathematical decision. Our p-value is 0.0568. This means that if we flipped 100 fair coins many times, between 5 and 6% of these times we would have fewer than 41 or more than 59 heads. So our result of 40 is a bit on the unusual side, but not extremely so. Our data provide some evidence to suggest that the coin may not be fair, but the evidence is far from conclusive. If we are really want to know, we probably need to gather more data.

Other hypothesis tests will proceed in a similar fashion. The details of how to compute a test statistic and how to convert it into a p-value will change from test to test, but the interpretation of the p-value is always the same. The p-value measures how surprising the value of the test statistic would be if the null hypothesis were true. The next example illustrates the steps of the hypothesis testing paradigm in a case where the alternate hypothesis is one-sided.

Example 6.1.2

A company receives a shipment of printed circuit boards. The claim of the manufacturer is that the defective rate is at most 1%. If 100 boards are tested, should we dispute the claim of the manufacturer if we find 3 defective boards in this test? In this situation, the pair of hypotheses to test are

$$\mathbf{H_0}: \quad \pi = 0.01$$
$$\mathbf{H_a}: \quad \pi > 0.01$$

The following R session is relevant to this example.

```
> 1-pbinom(c(0:5),100,.01)
[1] 0.633968 0.264238 0.079373 0.018374 0.003432 0.000535
```

From this computation, we find that even if the null hypothesis is true, we could expect to find 3 or more defective boards 7.9% of the time if we test 100. This result doesn't seem surprising enough to reject the null hypothesis or the shipment. (But perhaps you disagree!) In this example, we have illustrated how we proceed when the alternate hypothesis is one-sided. Namely, we only consider results to favor the alternate hypothesis when they are in the correct direction of the null hypothesis. That is, we wouldn't consider having too few defectives as evidence against the null hypothesis in favor of the alternate hypothesis.

It is often the case that we must make a decision based on our hypothesis test. In Example 6.1.2, for example, we must finally decide whether to reject the shipment. There are of course two different kinds of errors that we could make.

Definition 6.1.1 (Type I and Type II errors)**.** A **Type I error** is the error of rejecting $\mathbf{H_0}$ even though it is true.

A **Type II error** is the error of not rejecting $\mathbf{H_0}$ even though it is false.

Of course, if we reject the null hypothesis, we cannot know whether we have made a Type I error. Similarly, if we do not reject the null hypothesis, we cannot know whether we have made a Type II error. Whether we have committed such an error depends on the true value of π which we cannot ever know simply from data. What we can do however is to compute the probability that we will make such an error given our decision rule and our true state of nature.

To illustrate the computation of these two kinds of errors, let's return to the computation of the p-value in the case of the (un)fair coin. Suppose that we decide that whenever we toss a coin 100 times, we will consider it unfair if we have 40 or fewer or 60 or more heads. Then the p-value computation (recall the p-value was 0.0568) tells us that

> If the null hypothesis is true, our decision rule will make a Type I error with probability 5.68%

Is this the right decision rule to use? If instead we decide to reject the null hypothesis only if $X \leq 39$ or $X \geq 61$ we find that we will make a type I error with probability only `pbinom(39,100,.5) + (1-pbinom(60,100,.5))=0.035`. Which decision rule should we use? A common convention is to make some canonical choice of a probability of Type I error that we are willing to tolerate. A probability of Type I error of 5% is often chosen. If 5% were the greatest type I error probability we were willing to tolerate then we would not reject a null hypothesis if our p-value was greater than 5%. In the coin example, 40 heads would be acceptable but 39 would not. The choice of 5% is conventional but somewhat arbitrary. It is usually better to report the result of a hypothesis test as a p-value rather than simply reporting that the null hypothesis is rejected. We usually denote by α the probability of a Type I error that we are willing to accept in our decision rule.

Notice that if we lower α it becomes more difficult to reject the null hypothesis. This means that if the null hypothesis is false, the probability of a Type II error increases with decreasing α. (Oddly enough, the probability of a Type II error is named β.) We cannot compute the probability of a Type II error however without knowing the true value of π. Consider the case of the un(fair) coin. Suppose we choose $\alpha = .05$ and so we choose to reject the null hypothesis only if $X \leq 39$ or $X \geq 61$. What is the probability that we make a type II error if the true value of $\pi = 0.55$? It is easy to see that this is computed by

```
> pbinom(39,100,.55) + (1-pbinom(60,100,.55))
[1] 0.1351923
```

Notice that we will reject the null hypothesis only 13.5% of the time so that the probability that we make a Type II error is 86.5%! Obviously, our test is very conservative and will not detect an unfair coin very often. That is the penalty we pay for wanting to be reasonably sure that we do not make a Type I error. The next example illustrates these considerations in the case of a one-sided alternate hypothesis.

Example 6.1.3

As described in Example 6.1.1, the conventional test for ESP is a card test. The subject is asked to guess what is on 25 consecutive cards each of which contains one of five symbols. The appropriate pair of hypotheses in this case are

$$\mathbf{H_0}: \quad \pi = 0.2$$
$$\mathbf{H_a}: \quad \pi > .2$$

The following computation from R will help us develop our test.

```
> 1-pbinom(c(5:10),25,.2)
[1] 0.38331 0.21996 0.10912 0.04677 0.01733 0.00555
```

Obviously, our decision rule should say to reject the null hypothesis if the number of successes is too large. Note that the probability that $P(X \geq 9) = 4.7\%$ if the null hypothesis is true. Therefore if we choose α to be 5% as is a custom, we should reject the null hypothesis in favor of the alternate hypothesis if the number of successes is at least 9. If we follow this rule, the probability that we will make a Type I error is 4.7% if the null hypothesis is true.

What if the null hypothesis is false? For example what if the true value of $\pi = .3$? (This is a rather modest case of ESP but such a person would be interesting!) In this case, our decision rule would reject the null hypothesis with probability `1-pbinom(8,25,.3)=.323`. Note that even if our subject has ESP, our test could very well not detect this.

What one should notice in our treatment of decision rules is the asymmetry between the two hypotheses. We are generally not willing to tolerate a large probability of a Type I error – we often set $\alpha = 5\%$. However this seems to lead to a rather large probability of a Type II error in the case that the null hypothesis is false. This asymmetry is intentional however as the null hypothesis usually has a preferred status as the "innocent until proven guilty" hypothesis.

6.2 Inferences about the Mean

One of the most important problems in inferential statistics is that of making inferences about the (unknown) mean of a population.

Example 6.2.1

1. What is the average height of a Calvin College student? It not being feasible to measure each student, we might take a random sample of Calvin students and compute the sample mean, $\bar{x}$ of these students. How close is $\bar{x}$ likely to be to the true mean?

2. We have a number of chickens that we feed a diet of sunflower seeds. The average weight of the chickens after 30 days is 330 grams. How close is this number to the average weight of the (theoretical) population of "all" similar chickens?
3. We take a number of measurements of the speed of the light. How close is the average of these measurements likely to be to the "true" value?

In this section, we will conceptualize the above examples as instances of this question.

Given i.i.d. random variables $X_1, \ldots, X_n$ with unknown mean μ_X, what can we infer about μ_X from a particular outcome $x_1, \ldots, x_n$?

Estimates and Estimators

We will call $\overline{x}$ an **estimate** of μ_X and $\overline{X}$ and **estimator** of μ_X. The difference is that $\overline{X}$ is a random variable – you can think of it as a procedure for producing an estimate — and $\overline{x}$ is a number. The estimator $\overline{X}$ has two very important properties that make it a desirable estimator. The first is that $E(\overline{X}) = \mu_X$. In other words, in the long run, the sample mean will average the population mean. Because of this, we say that the estimator $\overline{X}$ is **unbiased.** An unbiased estimator doesn't have a tendency to under- or over-estimate the quantity in question. The general definition is this.

Definition 6.2.1 (unbiased estimator)**.** Suppose that θ is a parameter of a distribution and that Y is a statistic computed from a random sample $X_1, \ldots, X_n$ from that distribution. Then Y is an unbiased estimator of θ if $E(Y) = \theta$.

It turns out that the sample variance S^2 is an unbiased estimator of σ_X^2 which is the real reason we use $n - 1$ rather than n in the definition of S^2.

The second important property is that $\overline{X}$ is likely to be close to μ_X if n is large. Formally, we can say that for every $\epsilon > 0$ we have that $\lim_{n\to\infty} \mathrm{P}(|\overline{X_n} - \mu_X| > \epsilon) = 0$. While we will not prove this, it follows from the fact that the variance of $\overline{X_n}$ is σ^2/n.

These two properties together suggest that $\overline{X}$ is a good choice for an estimator of μ.

The Idea of a Confidence Interval

While the estimator $\overline{X}$ may be a good procedure to use, we recognize that in any particular instance, the estimate $\overline{x}$ will not be equal to to μ_X. We next will use the Central Limit Theorem to say something about how close to μ_X the estimate is likely to be. The Central

Limit Theorem allows us to say that $\overline{X}$ is approximately normally distributed with mean μ_X and variance σ^2/n. Thus the following random variable has a distribution that is approximately standard normal:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Therefore we can write

$$\mathrm{P}\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \approx .95 \, .$$

(The number 2 in the 68%-95%-99.7% law is actually 1.96.) Using algebra, we find that

$$\mathrm{P}\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx .95 \, .$$

What this probability statement says is that the interval

$$\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is likely to contain the true mean of the distribution. This interval is a random interval.

Definition 6.2.2 (confidence interval)**.** Suppose that $X_1, \ldots, X_n$ is a random sample from a distribution that is normal with mean μ and variance σ^2. Suppose that $x_1, \ldots, x_n$ is the observed sample. The interval

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is called an approximate 95% confidence interval for μ.

How does this notion of a confidence interval help us? Actually not much since this interval is defined in terms of σ, the standard deviation of the original distribution. But σ is not likely to be known (after all, we don't even know the mean μ of the original distribution). Let's set that issue aside and consider an example.

Example 6.2.2

A machine creates rods that are to have a diameter of 23 millimeters. It is known that the standard deviation of the actual diameters of parts created over time is 0.1 mm. A random sample of 40 parts are measured precisely to determine if the machine is still producing rods of diameter 23 mm. The data and 95% confidence interval are given by

```
> x
 [1] 22.958 23.179 23.049 22.863 23.098 23.011 22.958 23.186 23.015 23.089
[11] 23.166 22.883 22.926 23.051 23.146 23.080 22.957 23.054 22.995 22.894
[21] 23.040 23.057 22.985 22.827 23.172 23.039 23.029 22.889 23.019 23.073
[31] 22.837 23.045 22.957 23.212 23.092 22.886 23.018 23.031 23.059 23.117
> mean(x)
[1] 23.024
> c(mean(x)-(1.96)*.1/sqrt(40),mean(x)+(1.96)*.1/sqrt(40))
[1] 22.993 23.055
```

It appears that the process could still be producing rods of average diameter 23 mm. ■

We use the term confidence interval for this interval since we are are reasonably confident that the true mean of the rods is in the interval (22.933, 23.055). We even have a number that quantifies that confidence, 95%. But we need to be very careful in what we are saying. We are not saying that

> (**BAD - DO NOT SAY)** the probability that the true mean is in the interval (22.993, 23.055) is 95%.

There is no probability after the data are collected. Either the mean is in the interval or it isn't. Rather we are making a statement before the data are collected:

> If we are to generate a 95% confidence interval for the mean from a random sample of size 40 from a normal distribution with standard deviation 0.1, then the probability is 95% that the resulting confidence interval will contain the mean.

On the frequentist conception of probability we could say

> If we generate many 95% confidence intervals by this procedure, approximately 95% of them will contain the mean of the population.

After the data are collected, a good way of describing the confidence interval that results is

> Either the population mean is in (22.993, 23.055) or something surprising happened.

Notice that the confidence interval says something about the precision of our estimate. A wide confidence interval means that our estimate is not very precise.

But σ Isn't Known!

Using the Central Limit Theorem, we have seen that

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx .95\,. \tag{6.1}$$

The next step is to make another approximation. We need to get rid of σ. Since S^2, the sample variance is an unbiased estimate of σ^2, the trick is to use $S = \sqrt{S^2}$, the sample standard deviation, to estimate σ. Thus we have

$$P\left(\overline{X} - 1.96\frac{S}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{S}{\sqrt{n}}\right) \approx .95\,.$$

Now, after the experiment we have values for both $\overline{X}$ and S. We illustrate the procedure for getting our new confidence interval using Example 6.2.2. Note that the following R code computes a 95% confidence interval for μ_X.

```
> sd(x)
[1] 0.098755
> c( mean(x) - 1.96* sd(x)/sqrt(40), mean(x) + 1.96 * sd(x)/sqrt(40))
[1] 22.993 23.054
```

Removing the Approximations

Our new 95% confidence interval for the mean

$$\left(\overline{x} - 1.96\frac{s}{\sqrt{n}}, \overline{x} + 1.96\frac{s}{\sqrt{n}}\right)\,.$$

makes two approximations:

- We use the CLT to say that we can use the normal distribution (that's where the 1.96 comes from)
- We use S instead of σ simply because we do not know σ

The CLT Approximation

There are two ways of getting around the fact that we use the CLT in our approximation. First, we could assume that the underlying distribution is normal. Then there is no need to approximate since the distribution of $\overline{X}$ is exactly normal. Or we could use facts about the particular distribution in question. For example if X is binomial, we could use similar facts about the binomial distribution to develop a different kind of confidence interval. In general however we are just going to have to be content with the fact that our confidence intervals are approximate and hope that our sample size n is large enough.

The Approximation of using S for σ

The bottom line here is that we will change the 1.96 used in our current approximation to a slightly larger number to compensate for the approximation that results from not knowing σ. It seems right to do this: if we are less sure that we are using the right endpoints for the interval, we should make the interval a little wider to ensure that we have a 95% chance of capturing the mean. How much wider we should make the interval is a somewhat tricky (and long) story that we will tell in the next section.

Before we modify our intervals to take into account the approximation of σ by S, we note that we could modify our confidence intervals in a number of ways. For example, the number 95% is not sacred. It should be clear how to generate a 68% confidence interval or even a 80% confidence interval. We merely need to look up the appropriate fact about the standard normal distribution. A second way in which we might modify our intervals is to make them one-sided. For example, if we wanted a lower-bound for our rod diameters, since `qnorm(.05,0,1)=-1.644854` we could use

$$\mathrm{P}\left(\overline{X} - 1.64\frac{S}{\sqrt{n}} < \mu < \infty\right) \approx .95 \, .$$

6.3 The t-Distribution

In the last section, we left the problem of finding a confidence interval for μ at the point where we had a perfectly reasonable, but approximate, confidence interval. There were two approximations: the use of the CLT and the approximation of σ by S. We focus on the later problem here. We will begin by assuming that the random sample $X_1, \ldots, X_n$ are normal random variables so that we need not concern ourselves with the CLT approximation. Then the question is, what is the effect of replacing $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ by $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$? The t-distribution holds the key.

Definition 6.3.1 (t-distribution)**.** A random variable T has a t distribution (with parameter $\nu \geq 1$, called the degrees of freedom of the distribution) if it has pdf

$$f(t) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{(1 + t^2/\nu)^{(\nu+1)/2}} \qquad -\infty < t < \infty \, .$$

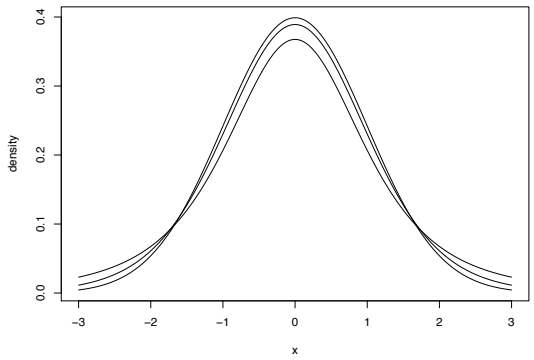
(The Γ function in the definition of the pdf above is an important function from analysis that is a continuous extension of the factorial function. But in this instance, it doesn't really matter what it is since its purpose is simply as a constant to ensure that the integral of the density is 1.)

Some properties of the t-distribution include

1. f is symmetric about $t = 0$ and unimodal. In fact f looks bell-shaped.
2. The mean of T is 0 if $\nu > 1$ (and does not exist if $\nu = 1$).
3. The variance of T is $\nu/(\nu - 2)$ if $\nu > 2$.
4. For large ν, T is approximately standard normal.

R knows the t-distribution of course and the appropriate functions are `dt(x,df)`, `pt()`, `qt()`, and `rt()`. The graphs of the normal distribution and two t-distributions are shown below.

```
> x=seq(-3,3,.01)
> y=dt(x,3)
> z=dt(x,10)
> w=dnorm(x,0,1)
> plot(w~x,type="l",ylab="density")
> lines(y~x)
> lines(z~x)
```



The importance of the t-distribution is contained in the following Theorem.

Theorem 6.3.2. If $X_1, \ldots, X_n$ are i.i.d. normal random variables with mean μ and variance σ^2, then the random variable

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t distribution with $n - 1$ degrees of freedom.

It is now clear how to generate an exact confidence interval for μ in the case that the data come from a normal distribution. For any number β, let $t_{\beta,\nu}$ be the unique number such that

$$P(T > t_{\beta,\nu}) = \beta$$

where T is a random variable that has a t distribution with ν degrees of freedom.

Theorem 6.3.3. If $x_1, \ldots, x_n$ are the observed values of a random sample from a normal distribution with unknown mean μ and $t^* = t_{\alpha/2,n-1}$, the interval

$$\left(\bar{x} - t^*\frac{s}{\sqrt{n}}, \bar{x} + t^*\frac{s}{\sqrt{n}}\right)$$

is an $100(1 - \alpha)\%$ confidence interval for μ.

In Example 6.2.2 where we considered the diameter of manufactured rods, we had $n = 40$. If we assume that the measurements come from a normal distribution, we would use the t-distribution with $\nu = 39$. To find a 95% confidence interval we need $t_{.025,39}$. R of course computes this as `qt(.975,39)= 2.022691` . So the effect of not knowing σ in this case is to use 2.02 in determining the width of the confidence interval rather than 1.96.

Notice that in this confidence interval there are three components. The first is an estimate $\overline{x}$ of the quantity it is a confidence interval for. Second there is a number t^* determined from the t-distribution by the level of confidence and the degrees of freedom. This number is usually referred to as a **critical value**. Finally, there is an estimate $s/\sqrt{n}$ of the standard deviation of the estimator. The number $\sigma/\sqrt{n}$ is often called the **standard error (of the estimator or of the mean)** and is often denoted σ_e. The estimate $s/\sqrt{n}$ of this standard error is often denoted s_e. Therefore we have that the confidence interval is of the form

$$\text{(estimate)} \pm \text{(critical value)} \cdot \text{(estimate of standard error)} .$$

Many other confidence intervals in statistics have the same form. The critical values and estimates change based on the situation but the general form of the interval is the same.

Because of the importance of confidence intervals for μ that are generated by the t-distribution, there is a function in R that does the table lookup and the arithmetic for us. We illustrate in the next example.

Example 6.3.1

Returning to the iris data, we might want to know the average sepal width of virginica irises. There is a lot to ignore in the following output but note that two confidence intervals are generated (95% and 90%) and that the t-distribution is used with 49 degrees of freedom (as $n = 50$).

```
> data(iris)
> sw=iris$Sepal.Width[iris$Species=="virginica"]
> hist(sw)
> t.test(sw)

        One Sample t-test

data:  sw
t = 65.208, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.882347 3.065653
sample estimates:
mean of x
    2.974
> t.test(sw,conf.level=.9)

        One Sample t-test
```

```
data:  sw
t = 65.208, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 2.897536 3.050464
sample estimates:
mean of x
    2.974
```

Now that we have exact confidence intervals in the case that data come from a normal distribution even if σ is unknown, we turn to the case that the underlying distribution is unknown. In this case we advocate using the t-distribution just as above, recognizing that the result is just an approximation. We illustrate in the following example.

Example 6.3.2

Thirty seniors are chosen at random from the collection of 1,333 seniors at a certain midwest college. The average GPA of the thirty seniors chosen is 3.2981. What inferences can we make about the mean GPA of the 1,333 seniors? We first simplify and assume that the 30 seniors represent the result of thirty i.i.d. random variables. Though sampling was without replacement, this seems like a relatively harmless assumption. We next realize that the underlying distribution of GPAs is not likely to be normal but rather to be negatively skewed. (This does not mean that we expect to find negative GPAs!) The technology of the last section suggests to use the normal distribution with s in place of σ. Using the t-distribution instead produces

```
> sr=read.csv('http://www.calvin.edu/~stob/data/actgpa.csv')
> sr$GPA
 [1] 3.992 2.533 3.377 3.009 3.509 3.969 3.917 3.547 3.416 3.287 4.000 3.446
[13] 3.905 2.926 3.100 3.446 2.785 3.663 3.368 3.352 3.929 2.750 3.620 3.765
[25] 2.763 1.986 2.836 2.696 3.119 2.662
> t.test(sr$GPA)

        One Sample t-test

data:  sr$GPA
t = 35.1095, df = 29, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.097500 3.480700
sample estimates:
mean of x
   3.2891
```

In this case, the effect of using the t-distribution is to replace 1.96 for 2.045 in the computation of the width of the confidence interval. It seems prudent to use a

wider confidence interval since we are only approximating the "true" 95% confidence interval in light of the fact that we are using the CLT.

Most statisticians recommend the approach of the last example. Namely, when constructing an approximate confidence interval in the case when our data are not from a normal distribution, we use the t-distribution with its slightly wider intervals than would be constructed by using the normal distribution. Statisticians have found that when this is done, the confidence intervals constructed work well for a wide variety of underlying non-normal distributions. That is, 95% confidence intervals produced from the t-distribution tend to be approximately 95% confidence intervals even though the distributional hypothesis is not satistfied. We say that this method of producing confidence intervals is **robust** meaning it is not particularly sensitive to departures from the hypothesis (normality) on which it is based. (Older books suggest that one could use the normal distribution instead of the t-distribution if $n \geq 30$ but this was a computational simplification. R knows all the t-distributions and can use one as easily as another.)

There are two very important cautions to be made here. Although the t-distribution works well over a wide range of distributions and sample sizes, it still is an approximation and in particular can give poor results if the sample size is small and the underlying distribution is quite skew. And the t-distribution will often fail disastrously if the independence assumption is violated.

6.4 Inferences for the Difference of Two Means

In this section we consider the problem of making inferences about the difference of two unknown means. We first give some examples.

Example 6.4.1

1. One might hypothesize that females get better grades at Calvin than males on average. One way of stating this claim precisely is to claim that the average GPA of females is greater than the average GPA of males. Since Calvin does not publish the average GPA by gender, we might test this claim by choosing a random sample of males and a separate random sample of females and comparing the two sample means.
2. One might claim that Tylenol is better than ibuprofren for treating pain from fractures in young children. To test this, one might assign children with leg fractures at random to treatment by Tylenol or ibuprofren. One would then compare the averages of some measure of pain relief in the two groups.
3. Kaplan claims to be able to raise SAT scores by 100 points with its tutoring program. To test the claim, they take a number of individuals who have already

taken the SAT test and subject them to their program. The students then take the SAT test after the program and their before and after scores are compared.

In the first case of the example, it is easy to see that we are choosing a random sample from each of two different populations. The second case is somewhat different. The "populations" of ibuprofren and Tylenol takers are really theoretical and not actual populations. But we can still think of the results as random sample from these theoretical populations (e.g., the population of all children with similar injuries who might be given ibuprofren), in part because we randomized the assignment of individuals to the two groups. The third case of the example is clearly different. The before and after scores do not represent two independent populations since we measured these scores on the the same individuals. In this section we address the issue of determining whether there is a difference in means between the two populations. In this section, we consider the situation that arises in the the first two cases of the example. We will call this the "two independent samples" case.

Assumptions for two independent samples:

1. $X_1, \ldots, X_m$ is a random sample from a population with mean μ_X and variance σ_X^2.
2. $Y_1, \ldots, Y_n$ is a random sample from a population with mean μ_Y and variance σ_Y^2.
3. The two samples are independent one from another.
4. The samples come from normal distributions.

We first write a confidence interval for the difference in the two means $\mu_X - \mu_Y$. Just as did our confidence intervals for one mean μ, our confidence interval will have the form

$$\text{(estimate)} \pm \text{(critical value)} \cdot \text{(estimate of standard error)} .$$

The natural choice for an estimator of $\mu_X - \mu_Y$ is $\overline{X} - \overline{Y}$. To write the other two pieces of the confidence interval, we need to know the distribution of $\overline{X} - \overline{Y}$. The necessary fact is this:

$$\frac{\left(\overline{X} - \overline{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathsf{Norm}(0, 1) .$$

Analogously to confidence intervals for a single mean, it seems like the right way to proceed is to estimate σ_X by s_X, σ_Y by s_Y and to investigate the random variable

$$\frac{\left(\overline{X}-\overline{Y}\right)-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}}. \tag{6.2}$$

The problem with this approach is that the distribution of this quantity is not known in general (unlike the case of the single mean where the analogous quantity has a t-distribution). We need to be content with an approximation.

Lemma 6.4.1. (Welch) The quantity in Equation (6.2) has a distribution that is approximately a t-distribution with degrees of freedom ν where ν is given by

$$\nu = \frac{\left(\frac{S_X^2}{m}+\frac{S_Y^2}{n}\right)^2}{\frac{(S_X^2/m)^2}{m-1}+\frac{(S_Y^2/n)^2}{n-1}} \tag{6.3}$$

(It isn't at all obvious from the formula but it is good to know that $\min(m-1, n-1) \leq \nu \leq n+m-2$.)

We are now in a position to write a confidence interval for $\mu_X - \mu_Y$.

An approximate $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is

$$\overline{x}-\overline{y} \pm t^* \left(\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}\right) \tag{6.4}$$

where t^* is the appropriate critical value $t_{\alpha/2,\nu}$ from the t-distribution with ν degrees of freedom given by (6.3).

We note that ν is not necessarily an integer and we leave it R to compute both the value of ν and the critical value t^*.

Example 6.4.2

The t-test is due to "Student" (a pseudonym of William Sealy Gossett whose employer, Guinness Brewery, did not allow him to publish under his own name). In a famous paper in 1908 addressing the issue of the inference about means, Student considered data from a sleep experiment. Two different soporifics were tried on a number of subjects and the amount of extra sleep that each subject attained was recorded. The question is whether one soporific worked better than another.

```
> sleep
   extra group
1    0.7     1
2   -1.6     1
3   -0.2     1
4   -1.2     1
5   -0.1     1
...................

> t.test(extra~group,data=sleep)

        Welch Two Sample t-test

data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.0794
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

We see that each group averaged more sleep and the excess was more for those subjects in group 2. However it does not appear that we could say one drug was clearly better than the other (after all, 0 is in the confidence interval so that the mean difference could be 0). A 95% confidence interval for the difference in mean effect of the two drugs is $(-3.37, 0.21)$. We can see that the degrees of freedom is 17.776 and we can be grateful that we didn't have to compute it or the critical value. Note too that R refers to this as a Welch test.

We should remark at this point that older books (and the Fundamentals of Engineering Exam) suggest an alternate approach to the problem of writing confidence intervals for $\mu_X - \mu_Y$. These books suggest that we assume that the two standard deviations σ_X and σ_Y are equal. In this case the exact distribution of our quantity is known. The problem with this approach is that there is usually no reason to suppose that σ_X and σ_Y are equal and if they are not the proposed confidence interval procedure is not as robust as the one we are using. In these notes we take the approach of not even mentioning what this alternate procedure is since it has fallen into disfavor.

Hypotheses and Cautions

Confidence intervals generated by Equation (6.4) are probably the most common confidence intervals in the statistical literature. But those who generate such intervals are not always sensitive to the hypotheses that are necessary to be confident about the confidence

intervals generated. It should first be noted that the confidence intervals constructed are based on the hypothesis that the two populations are normally distributed. It is often apparent from even a cursory examination of the data that this hypothesis is unlikely to be true. However, if the sample sizes are large enough, we can rely on the Central Limit Theorem to tell us our results are approximately true. There are a number of different rules of thumb as to what large enough means, but $n, m > 15$ for distributions that are relatively symmetric and $n, m > 40$ for most distributions are common rules of thumb. A second approximation concerns the approximation made in computing the Welch interval. The rule of thumb here is that we are surer of confidence intervals where the quotients s_X^2/m and s_Y^2/n are not too different in size than those in which they are quite different.

Turning Confidence Intervals into Hypothesis Tests

It is often the case that we are interested in testing a hypothesis about $\mu_X - \mu_Y$ rather than computing a confidence interval for that quantity. For example, the null hypothesis $\mu_X - \mu_Y = 0$ in the context of an experiment is a claim that there is no difference in the two treatments represented by X and Y. Hypothesis testing of this sort has fallen into disfavor in many circles since the knowledge that $\mu_X - \mu_Y \neq 0$ is of rather limited interest unless the size of this quantity is known. A confidence interval answers that question more directly. Nevertheless, since the literature is still littered with such hypothesis tests, we give an example here.

Example 6.4.3

Returning to our favorite chicks, we might want to know if we should believe that the effect of a diet of horsebean seed is really different that a diet of linseed. Suppose that $x_1, \ldots, x_m$ are the weights of the m chickens fed horsebean seed and $y_1, \ldots, y_n$ are the weights of the n chickens fed linseed. The hypothesis that we really want to test is $\mathbf{H_0}$: $\mu_X - \mu_Y = 0$. We note that if the null hypothesis is true, then $T = (\overline{X} - \overline{Y})/\sqrt{S_x^2/m + S_y^2/m}$ has a distribution that is approximately a t-distribution with the Welch formula giving the degrees of freedom. Thus the obvious strategy is to reject the null hypothesis if the value of T is too large. Fortunately, R does all the appropriate computations. Notice that the mean weight of the two groups of chickens differs by 58.5 but that a 95% confidence interval for the true difference in means is $(-99.1, -18.0)$. On this basis we expect to conclude that the linseed diet is superior, i.e., that there is a difference in the mean weights of the two populations. This is verified by the hypothesis test of $\mathbf{H_0} : \mu_X - \mu_Y = 0$ which results in a p-value of 0.007. That is, this great a difference in mean weight would have been quite unlikely to occur if there was no real difference in the mean weights of the populations.

```
> hb=chickwts$weight[chickwts$feed=="horsebean"]
> ls=chickwts$weight[chickwts$feed=="linseed"]
> t.test(hb,ls)
```

```
        Welch Two Sample t-test

data:  hb and ls
t = -3.0172, df = 19.769, p-value = 0.006869
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -99.05970 -18.04030
sample estimates:
mean of x mean of y
   160.20    218.75
```

Variations

One-sided confidence intervals and one-sided tests are possible as are intervals of different confidence levels. All that is needed is an adjustment of the critical numbers (for confidence intervals) or p-values for tests.

Example 6.4.4

A random dot stereogram is shown to two groups of subjects and the time it takes for the subject to see the image is recorded. Subjects in one group (VV) are told what they are looking for but subjects in the other group (NV) are not. The quantity of interest is the difference in average times. If μ_X is the theoretical average of the population of the NV group and μ_Y is the average of the VV group, then we might want to test the hypothesis

$$\begin{aligned} \mathbf{H_0}&: \quad \mu_X - \mu_Y = 0 \\ \mathbf{H_a}&: \quad \mu_X > \mu_Y \end{aligned}$$

```
> rds=read.csv('http://www.calvin.edu/~stob/data/randomdot.csv')
> rds
       Time Treatment
1  47.20001        NV
2  21.99998        NV
3  20.39999        NV
.....................
77  1.10000        VV
78  1.00000        VV
> t.test(Time~Treatment,data=rds,conf.level=.9,alternative="greater")

        Welch Two Sample t-test

data:  Time by Treatment
t = 2.0384, df = 70.039, p-value = 0.02264
alternative hypothesis: true difference in means is greater than 0
```

```
90 percent confidence interval:
 1.099229      Inf
sample estimates:
mean in group NV mean in group VV
        8.560465         5.551429

>
```

From this we see that a lower bound on the difference $\mu_X - \mu_Y$ is 1.10 at the 90% level of confidence. And we see that the p-value for the result of this hypothesis test is 0.023. We would probably conclude that those getting no information take longer than those who do on average.

.

6.5 Regression Inference

In Section 4.4, we tried to describe the relationship between two quantitative variables by fitting a line to the data that came to us in pairs $(x_1, y_1), \ldots, (x_n, y_n)$. In this section, we describe a statistical model that attempts to account for both the linear relationship in the data and also the fact that the data are not exactly collinear. What results is known as the **standard linear model.**

The standard linear model is given by the following equation that relates the values of x and y.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

where

1. β_0, β_1 are (unknown) parameters,
2. ϵ_i is a random variable with mean 0 and (unknown) variance σ^2,
3. thus Y_i is a random variable with mean $\beta_0 + \beta_1 x_i$ and variance σ^2,
4. the random variables ϵ_i (and hence the variables Y_i) are independent,
5. the random variables ϵ_i are normally distributed.

We can write this model more succinctly in terms of linear algebra. Let $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Then the model says that $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is a random vector. There are three unknown parameters to estimate in this model: β_0, β_1, and σ^2.

Estimating β_0 and β_1

One obvious choice for the estimates of β_0 and β_1 is given by the coefficients b_0, b_1 of the least squares regression line. It turns out that there are good statistical reasons for using b_0, b_1 to estimate β_0, β_1.

Lemma 6.5.1. The estimates b_0 and b_1 are unbiased estimates of β_0 and β_1 respectively. Therefore, $\hat{y}_i = b_0 + b_1 x_i$ is an unbiased estimate of $\beta_0 + \beta_1 x_i$.

Since b_0 and b_1 are the estimates, we will use B_0, B_1 for the estimators (just as we used $\overline{X}$ and $\overline{x}$ for the estimator and estimate of the mean). Unbiased estimators are not much good to us if they have large variance. It is fairly easy to show (using equation 4.2, say) that

$$\mathrm{Var}(B_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$\mathrm{Var}(B_0) = \frac{\sigma^2}{n} \frac{\sum x_i^2}{\sum (x_i - \bar{x})^2} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)$$

An inspection of these formulas for the variances of the coefficients shows that the variances of the estimators decrease as the number of observations increase (provided that the values x_i are not all identical). The variance depends not only the the error variance but also on the spread of the independent variables x_i. Qualitatively, at least, the variances of the estimators behaves as we would want them to. But could we find estimators with even smaller variance? The following famous theorem says that the least-squares estimates of β_0 and β_1 are the best estimators in a certain precise sense.

Theorem 6.5.2 (Gauss-Markov Theorem)**.** Assume that $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, and the random variables ϵ_i are independent. Then the estimators B_0 and B_1 are the unbiased estimators of minimum variance among all unbiased estimators that are linear in the random variables Y_i. (We say that these estimators are BLUE which stands for Best Linear Unbiased Estimator.)

Estimating σ^2

The random variables ϵ_i have mean 0, variance σ^2 and are independent. Thus $\mathrm{E}(\epsilon_i^2) = \sigma^2$. So we could estimate σ^2 by

$$\frac{\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2}{n} \,. \tag{6.5}$$

This fraction would give an unbiased estimate of σ^2. This is not much good however as we do not know β_0 and β_1. Substituting estimates for β_0 and β_1 and changing the denominator of the fraction gives us the estimate we need

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2}{n-2} = \frac{\text{SSResid}}{n-2}\,.$$

This estimate, denoted MSE, is called the **mean squared error**. The justification for substituting $n-2$ in the denominator rather than n which would be more natural is the same as that for using $n-1$ in the definition of s^2 in Section 6.2. Namely, the use of $n-2$ ensures that MSE is an unbiased estimate of σ^2. Notice that the denominator in each case accounts for the number of parameters estimated (one in the case of s^2 and two in the case of MSE.

Example 6.5.1

A class taught at a college in the midwest took three tests and a final exam. There were 32 students in the class. The final exam scores are related to the scores on Test 1. The result of a regression analysis appears below.

```
> class=read.csv('http://www.calvin.edu/~stob/data/m222.csv')
> class[1:3,]
  Test1 Test2 Test3 Exam
1    98   100    98  181
2    93    91    89  168
3   100    99    99  193

> l.class=lm(Exam~Test1,data=class)
> summary(l.class)
Call:
lm(formula = Exam ~ Test1, data = class)

Residuals:
     Min       1Q   Median       3Q      Max
-33.6930 -10.1574  -0.9462   8.5918  44.0759

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.9652    22.7916   1.051    0.301
Test1         1.6044     0.2729   5.880 1.95e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.86 on 30 degrees of freedom
Multiple R-Squared: 0.5354,        Adjusted R-squared: 0.5199
F-statistic: 34.57 on 1 and 30 DF,  p-value: 1.952e-06

> p.class=predict(l.class)
```

```
> mse=sum( (class$Exam-p.class)^2/30 )
> rse=sqrt(mse)
> rse
[1] 18.86495
```

Notice that R computes $\sqrt{\text{MSE}}$ which is called the **residual standard error**. The residual standard error is used as an estimate for σ (although it is not an unbiased estimate of σ). In keeping with our previous use of s to denote the estimate of the standard deviation of an unknown distribution, we will generally use s_e to denote the residual standard error (and S_e to denote the corresponding estimator).

What we have done until now does not depend on the normality assumptions on the random variables ϵ_i but only on the fact that they are independent with mean 0 and common variance σ^2. In order to make inferences about the parameters β_0 and β_1, we need to assume something about the distribution of the ϵ_i and so we now assume also that the random variables ϵ_i are normally distributed. This in turn implies that the random variables Y_i are normally distributed with $\text{E}(Y_i) = \beta_0 + \beta_1 x_i$ and $\text{Var}(Y_i) = \sigma^2$.

Under this assumption, it turns out the estimators B_0 and B_1 are normally distributed as well. So we have that

$$B_0 \sim \text{Norm}\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}}\right)$$

$$B_1 \sim \text{Norm}\left(\beta_1, \frac{\sigma}{\sqrt{\sum(x_i - \overline{x})^2}}\right)$$

We will primarily be concerned with constructing confidence intervals and hypothesis tests for β_1, the slope in the regression line. The reason for this is that the slope tells us the direction and size of the supposed linear relationship between x and y. The same reasoning can be used to write confidence intervals and tests for β_0.

Our procedure for writing a confidence interval for β_1 is very similar to that of constructing a confidence interval for the mean μ of an unknown distribution. Just as in that case, the unknown standard deviation σ is a nuisance parameter and we must substitute an estimate of σ for it. In this case we use s_e. This in turn means that the sampling distribution of our statistic becomes a t-distribution rather than a normal distribution. The resulting fact is the statistic T has a t-distribution with $n - 2$ degrees of freedom. (Here $n - 2$ matches the denominator in the definition of s_e.)

$$T = \frac{B_1 - \beta_1}{S_e/\sqrt{\sum(x_i - \overline{x})^2}}$$

We define $s_{b_1} = s_e/\sqrt{\sum(x_i - \overline{x})^2}$ This number s_{b_1} is called the estimate of the standard error of b_1. We now have the following result

Confidence Intervals for β_1
A $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\left(b_1 - t_{\alpha,n-2}s_{b_1}, b_1 + t_{\alpha,n-2}s_{b_1}\right)$$

Example 6.5.2

In Example 4.4.1 we used linear regression to write a relationship between iron content and material loss in certain Cu/Ni alloy bars. The dataset was the `corrosion` dataset in R. In what follows, we write a 95% confidence interval for the slope of the regression line.

```
> summary(l.corrosion)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  129.787      1.403   92.52  < 2e-16 ***
Fe           -24.020      1.280  -18.77 1.06e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.058 on 11 degrees of freedom
Multiple R-Squared: 0.9697,     Adjusted R-squared: 0.967
F-statistic: 352.3 on 1 and 11 DF,  p-value: 1.055e-09
> qt(.975,11)
[1] 2.200985
> c(-24.020 - qt(.975,11)*1.280, -24.020+qt(.975,11)*1.280)
[1] -26.83726 -21.20274
```

The confidence interval constructed, $(-26.84, -21.20)$, is a 95% confidence interval for the slope of the "true" linear relationship between x and the mean of y. To interpret this, we might say something like "We are 95% confident that the an increase in iron content of 1% results in an average loss of between 21.2 and 26.8 milligrams per square decimeter of material." Notice the high R^2 value in this model. A very high percentage of the loss due to corrosion is explained by the percentage iron content of the bar.

6.6 Exercises

6.1 For each scenario, write appropriate null and alternative hypotheses. (Write mathematical statements using Greek letters appropriate to the situation, and include a sentence stating precisely what that Greek letter represents for the given situation.)

a) A Tufts University study finds that 40% of 12^{th} grade females feel they are overweight. Is this percentage higher for college-age females?

b) George finally got a new bowling ball, after years of averaging 211 (game score) with his old one. He wonders if there will be any (initial(?)) effect on his game.

c) Last year Jessica had an average time of 1:08.51 (68.51 seconds) in the 400-meters. She wonders if her rigorous off-season training regimen will result in faster times this year.

d) Is there a difference in the mean amount of hours that female and male students study per week?

6.2 A basketball player claims to be a 90% free-throw shooter. Namely, she claims to be able to make 90% of her free-throws. Should we doubt her claim if she makes 14 out of 20 in a session at practice? Set this problem up as a hypothesis testing problem and answer the following questions.

a) What are the null and alternate hypotheses?

b) What is the p-value of the result 14?

c) If the decision rule is to reject her claim if she makes 15 or fewer free-throws, what is the probability of a Type I error?

6.3 In Example 6.1.1(c), we are trying to decide whether to fire the old kicker and hire a new one on the basis of a trial of 20 kicks. Suppose that we decide to hire the new kicker if he makes 8 or more kicks.

a) Suppose that he makes exactly 8 kicks. What is the p-value of this result?

b) What is α, the probability of a Type I error, for this decision rule?

c) If the kicker truly has a 35% chance of making each kick, what is the probability of a Type II error (i.e., that we don't believe that he is better than the old kicker)?

6.4 Nationally, 79% of students report that they have cheated on an exam at some point in their college career. You can't believe that the number is this high at your own institution.

Suppose that you take a random sample of size 50 from your student body. Since 50 is so small compared to the size of the student body, you can treat this sampling situation as sampling with replacement for the purposes of doing a statistical analysis.

a) Write an appropriate set of hypotheses to test the claim that 79% of students cheat.

b) Construct a decision rule so that the probability of a Type I error is less than 5%.

6.5 In this problem, you will develop a hypothesis test for a random variable other than a binomial one. Suppose that you believe the waiting time until you are served at the MacDonald's on 28th street is a random variable with an exponential distribution but with unknown λ. The sign at the drive-up window says that the average wait time is 1 minute. You actually wait 2 minutes. Your friend in the car says that this is outrageous and that the claim on the sign must be wrong.

a) Write a pair of hypothesis about λ that captures the discussion between you and your friend.

b) What is the p-value of the single data point of a 2 minute wait?

c) Write a sentence that explains clearly to your friend the meaning of that p-value. Remember that your friend has not yet been fortunate enough to take a statistics course.

d) How long would you have to have waited to be suspicious of MacDonald's claim? There are many right answers to this question but any answer needs statistical justification.

6.6 In Example 6.2.2 we generated an approximate 95% confidence interval for μ assuming that σ is known.

a) Construct instead a 90% confidence interval for μ.

b) Construct both 90% one-sided confidence intervals for μ.

c) Describe clearly a situation in which you would want a one-sided confidence interval rather than a two-sided one.

6.7 Suppose that the standard deviation σ of a normal population is known. How large a random sample must be chosen so that a 95% confidence will be of the form $\overline{x} \pm 0.1\sigma$?

6.8 Which is wider, a 90% confidence interval or a 95% confidence interval generated from the same random sample from a normal population?

6.9 Suppose that $X_1, \ldots, X_n$ are i.i.d. from an exponential distribution with parameter λ unknown. In this problem we write a confidence interval for λ using $\overline{X}$.

a) Rewrite Equation (6.1) in this case by substituting for μ and σ the appropriate expressions involving λ.

b) Solve the inequality that results in part (a) for an inequality of form $a < \lambda < b$ where a and b do not involve λ.

c) Suppose that $n = 30$ and $\overline{X} = 4.23$. Using (b), write an approximate 95% confidence interval for λ. Note that this confidence interval relies on the CLT but makes no other approximation.

6.10 The `chickwts` dataset presents the results of an experiment in which chickens are fed six different feeds. Suppose that we assume that the chickens were assigned to the feed groups at random so that we can assume that the chickens can be thought of as coming from one population. For each feed, we can assume that the chickens fed that feed are a random sample of the (theoretical) population that would result from feeding all chickens that feed.

a) Write 95% confidence intervals for the mean weight of chickens fed each of the six seeds.

b) From an examination on the six resulting confidence intervals, is there convincing evidence that some diets are better than others?

c) Since you no doubt used the t-distribution to generate the confidence intervals in (a), you might wonder whether that is appropriate. Are there any features in the data that suggest that this might not be appropriate?

6.11 The dataframe in `http://www.calvin.edu/~stob/data/miaa05.csv` contains statistics on each of the 134 players in the MIAA 2005 Men's Basketball season. Choose 10 different random samples of size 15 from this dataset.

a) From each, compute a 90% confidence interval for the mean PTSG (points per game) of all players.

b) Of the 10 confidence intervals you computed in part (a), how many actually did contain the true mean? (Note: you can compute the true mean since you have the population in this instance.)

c) How many of the 10 confidence intervals in part (a) would you have expected (before you actually generated them) to contain the true mean?

d) In light of your answer in (c), are you surprised by your answer in (b)?

6.12 The dataset found at http://www.calvin.edu/~stob/data/normaltemp.csv contains the body temperature and heart rate of 130 adults.[1]

a) Assuming that the body temperatures of adults in the population is approximately normal, and that 130 adults sampled behave like a simple random sample, write a 95% confidence interval for the mean body temperature of an adult.

b) Comment on the result in (a).

c) Is there anything in the data that would lead you to believe that the normality assumption is incorrect?

6.13 The `R` dataset `morley` contains the speed of light measurements for 100 different experimental runs. The vector `Speed` contains the measurements (in some obscure units).

a) If we think of these 100 measurements as repeated independent trials of a random variable X, what is a good description of the population of which the measurements are a sample?

b) Write a 95% confidence interval for the mean of this population.

c) what is the value $t_{\beta,n-1}$ for the confidence interval generated in the previous part?

d) Is there anything in the histogram of the data values that suggests that the procedure might not be a good one for generating a confidence interval in this case?

6.14 Write 95% confidence intervals for the mean of sepal length of each of the three species of irises in the `R` dataset `iris`. Would you say that these confidence intervals give strong evidence that the means of the sepal lengths of these species are different?

6.15 The dataset http://www.calvin.edu/~stob/data/uselessdata.csv contains data collected the first day of the semester about each student in one of Professor Stob's classes. The class is *not* a random sample of Calvin students, but suppose that we consider it so.

a) Write a 90% confidence interval for the mean number of hours of sleep that a Calvin student got the night before the first day of classes. (The variable named `Sleep` records that for the sample.)

b) Is there anything in the data itself that concerns you in using the t-distribution to generate the confidence interval in (a)?

c) Write a 90% confidence interval for the average amount of cash that students carried on that first day of class?

[1] This data appeared in "What's normal?—Temperature, Gender, and Heart Rate," **Journal of Statistics Education**, Shoemaker, 1996.

d) Is there anything in the data that concerns you about using the t-distribution to generate the interval in (c)?

6.16 The dataset http://www.calvin.edu/~stob/data/reading.csv contains the results of an experiment done to test the effectiveness of three different methods of reading instruction. We are interested here in comparing the two methods DRTA and Strat. Let's suppose, for the moment, that students were assigned randomly to these two different treatments.

a) Use the scores on the third posttest (POST3) to investigate the difference between these two teaching methods by constructing a 95% confidence interval for the difference in the means of posttest scores.

b) Your confidence interval in part (a) relies on certain assumptions. Do you have any concerns about these assumptions being satisfied in this case.

c) Using your result in (a), can you make a conclusion about which method of reading instruction is better?

6.17 Surveying a choir, you might expect that there would not be a significant height difference between sopranos and altos but that there would be between sopranos and basses. The dataset `singer` from the `lattice` package contains the heights of the members of the New York Choral Society together with their singing parts.

a) Decide whether these differences do or do not exist by computing relevant confidence intervals.

b) These singers aren't random samples from any particular population. Explain what your conclusion in (a) might be about.

6.18 Returning to the sport of baseball one last time, let's reexamine the results of the 1994–1998 baseball seasons in http://www.calvin.edu/~stob/data/bball9498all.csv. Earlier, we tried to predict R (runs) by HR (homeruns). Let's refine that analysis here.

a) Instead of predicting R from HR, use regression to write a linear relationship to predict RG (runs per game) from HRG (homeruns per game).

b) Interpret the slope and intercept of the line in part (a) informally.

c) Write a 95% confidence interval for the slope of the line in (a).

6.19 The dataset http://www.calvin.edu/~stob/data/lakemary.csv contains the age and length (in mm) of 78 bluegills captured from Lake Mary, Minnesota. (Richard Frie, J. Amer. Stat. Assoc., (81), 922-929).

a) Write a linear function to predict the length from the age.

b) Interpret the slope and intercept of the line in (a).

c) Write a 95% confidence interval for the slope of the regression line.

d) Do you have any comments about the data or the model?

6.20 The dataset http://www.calvin.edu/~stob/data/home.csv contains the prices of homes in a certain community at two different points in time.

a) Write a linear function to predict the old price from the new.

b) Write a 90% confidence interval for the slope of the line in (a).

c) Write a sentence explaining what the confidence interval in (b) means.