6 Hypothesis Testing, Confidence Intervals and p-values

6.1 Small Sample Confidence Interval (Unknown σ_x)

For small samples, the sampling distribution of $\frac{\bar{x}-\mu_x}{s/\sqrt{n}}$ is a t-distribution with n-1 degrees of freedom. To compute that confidence interval, all we need to know is how to compute areas under the t-distribution between two numbers, which is similar to what we did with the normal distribution. To find t^*, we just replace qnorm(.05 / 2) with t(0.05 / 2,sample.size-1).

In terms of coverage, the results will be very similar if the sample size is large (e.g., 100+). For small samples (e.g., 10), the CI computed using a t-distribution will cover the population mean the correct number of times, while the CI computed using the normal distribution will not.

Note that computing the confidence interval for small samples using the t-distribution **assumes that the population is normal**. If the population is non-normal, the bootstrap method should be used instead.
Bootstrapping should be used when:

1. The population is not normal and the sample size is small.

2. No formulas for computing standard errors of the statistics of interest exist.

6.2 Confidence Intervals and Hypothesis Tests

In general, this is the way to decide how to set up the null and alternative hypothesis: Convert the statement of the problem to make it sound like "Does **data** provide evidence for blah?" Then that "blah" is what should go into H_1. The reason is that the hypothesis testing procedure starts by assuming whatever is under H_0. And so, if you are trying to see if the *data* provide evidence for X, then you should not start by assuming X is true. Similarly, if the problem asks "Does the data contradict blah?", then the blah should go into H_0.

Some problems don't readily lend themselves to that kind of translation. They ask something like "Test the prior belief that blah." In that case, the blah should go into H_0. The reason is similar to what I said above: Data provides evidence for H_1, against H_0. And "prior" means prior to data. So, any "prior belief" should go into H_0.

So far, we have learned 3 ways of constructing confidence intervals and doing hypothesis tests:

1. Using CI formulas.

2. Using bootstrapping.

3. Using the R function t.test().

The following example will focus on the last method.

Example 1: Exercise 8.28

Fusible interlinings are being used with increasing frequency to support outer fabrics and improve the shape and drape of various pieces of clothing. The article "Compatibility of Outer and Fusible Interlining Fabrics in Tailored Garments" (Textile Res. J., 1997: 137???142) gave the accompanying data on extensibility (%) at 100 gm/cm for both high-quality fabric (H) and poor-quality fabric (P) specimens:

```
H <- c(1.2, 0.9, 0.7, 1.0, 1.7, 1.7, 1.1, 0.9, 1.7, 1.9, 1.3, 2.1, 1.6, 1.8,
       1.4, 1.3, 1.9, 1.6, 0.8, 2.0, 1.7, 1.6, 2.3, 2.0)
P <- c(1.6, 1.5, 1.1, 2.1, 1.5, 1.3, 1.0, 2.6)
```

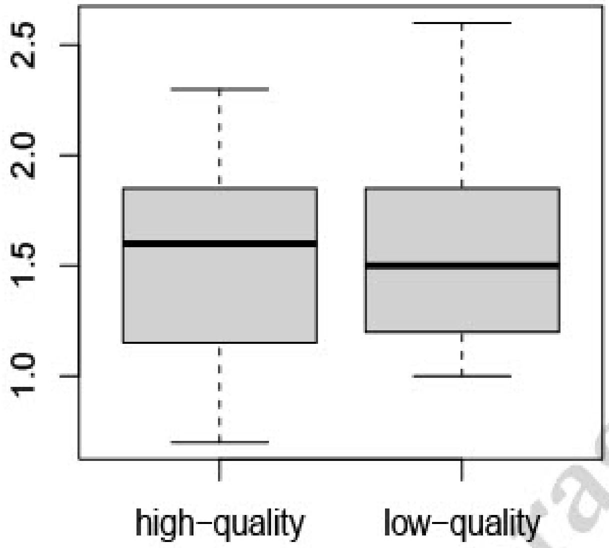
Suppose the problem asked us to estimate the true means for the the two populations separately. Then, we would compute 2-sided, 1-sample, CIs for each of the two population means.

```
t.test(H)


One Sample t-test

data:  H
t = 17, df = 23, p-value = 3e-14
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.321 1.696
sample estimates:
mean of x
    1.508
```

```
t.test(P)


One Sample t-test

data:  P
t = 8.5, df = 7, p-value = 0.00006
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.144 2.031
sample estimates:
mean of x
    1.588
```

However, if we were only comparing the means, we would not be able to tell much about the difference in the true means because there is a lot of overlap between the two CIs.

```
boxplot(H, P, names = c("high-quality", "low-quality"))
```



If we are interested in the difference between the two means, it's better to compute a 2-sample CI, instead of comparing two 1-sample CIs.

Suppose the problem asks "Does the data provide evidence to support the claim that the two populations have different means?" Then, we need to construct a 2-sided, 2-sample, CI. The two hypotheses are:

$$H_0\colon \mu_H - \mu_P = 0$$
$$H_1\colon \mu_H - \mu_P \neq 0$$

```
t.test(H, P, alternative = "two.sided")


Welch Two Sample t-test

data:  H and P
t = -0.38, df = 10, p-value = 0.7
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5404  0.3820
sample estimates:
mean of x mean of y
    1.508     1.588
```

Note that the 95% CI includes zero. Also note that p-value > 0.05. Both of these observations imply that we cannot reject the null hypothesis that the two means are equal. One often says "there is no statistically significant difference between the means of H and P." Keep reminding yourself that this does NOT mean that there is no difference; it just means that if there is a difference, your data is not seeing it.

Also note that the sample mean of H is smaller than the sample mean of P. Suppose the problem had asked us if the data provide evidence that the population mean of H is less than the population mean of P. Then the appropriate "interval" would be a (1-sided) **upper** confidence bound for $\mu_H - \mu_P$. The two hypotheses would be:

$$H_0\colon\ \mu_H - \mu_P > 0$$
$$H_1\colon\ \mu_H - \mu_P < 0$$

```
t.test(H, P, alternative = "less")


Welch Two Sample t-test

data:  H and P
t = -0.38, df = 10, p-value = 0.4
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
   -Inf 0.2966
sample estimates:
mean of x mean of y
    1.508     1.588
```

The upper confidence bound is positive, and so the difference $(\mu_H - \mu_P)$ may be positive. So, the data do NOT provide evidence that $\mu_H - \mu_P < 0$. Although the p-value is lower than the 2-sided p-value above, it's still not less than $\alpha = 0.05$. So, the conclusion is that the data do NOT provide evidence to reject $\mu_H - \mu_P > 0$ in favor of H_1. Either way, the conclusion is the same.

Note that this (1-sided) upper confidence bound is smaller than the upper limit of the 2-sided CI. This is consistent with what confidence intervals are supposed to do, i.e., cover the true parameter some percentage of the time.

Had the problem asked us if there is evidence for $\mu_H - \mu_P > 0$, then the hypotheses would be:

$$H_0\colon\ \mu_H - \mu_P < 0$$
$$H_1\colon\ \mu_H - \mu_P > 0$$

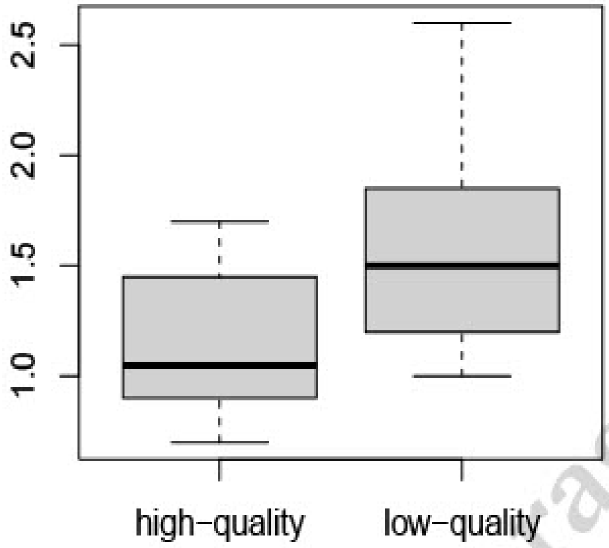
```
t.test(H, P, alternative = "greater")


Welch Two Sample t-test

data:  H and P
t = -0.38, df = 10, p-value = 0.6
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.4549     Inf
sample estimates:
mean of x mean of y
    1.508     1.588

# Note that the two 1-sided p-values, above, do sum to 1, as they should.

# If you want to extract only the p-value from t.test(), do the following:
t.test(H, P)$p.value
```

```
[1] 0.7115

# If the above command was placed inside a loop, R will not print the values on
# the screen, unless you put the whole thing in a print(), i.e.,
print(t.test(H, P)$p.value)

[1] 0.7115
```

Example 2: Paired and Unpaired Two-sample t-test

Suppose the data (from example 1) on H and P were of the same size. Assume the data on H was just the first *8* cases. Suppose the question had asked "is there a difference?"

```
H <- H[1:8]  # Keep only the first 8 cases in above H.
boxplot(H, P, names = c("high-quality", "low-quality"))
t.test(H, P, alternative = "two.sided")


Welch Two Sample t-test

data:  H and P
t = -1.9, df = 13, p-value = 0.08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.93351  0.05851
sample estimates:
mean of x mean of y
    1.150     1.588

# Note that although the p-value is pretty small, it's still greater than
# alpha = 0.05.
```

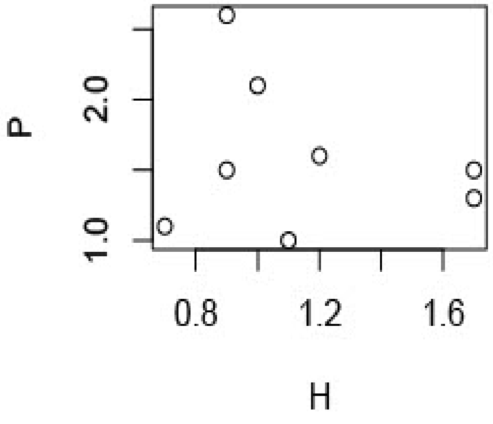


Now suppose the problem had said that the two sets of measurements, H, P, are taken on the same unit of study. For example, the two measurements are made on a given fabric, but in two different conditions, say wet and dry. Then we are dealing with paired data. Then the appropriate test would be:

```
t.test(H, P, paired = T, alternative = "two.sided")


Paired t-test

data:  H and P
t = -1.8, df = 7, p-value = 0.1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.0252  0.1502
sample estimates:
mean of the differences
                -0.4375
```

The CI is now much wider and the p-value is much larger. So, the pairing of the data means that it provides even less evidence than otherwise. This makes sense in this example, only because the data on H and P are not paired anyway. You can see that they are not paired by looking at the their scatterplot, and noting that there is no correlation:

```
plot(H, P)
```



But, if the data truly were paired, the CI for the paired data would be narrower than that of the unpaired data. Similarly, the p-value for the paired test would be smaller than the p-value from an unpaired test. That makes sense too, because by taking the difference between two columns of data, all the variability **within** each column is "subtracted out," and so the test can focus only on the variability in the **difference between** the two columns, which is all we really care about anyway.

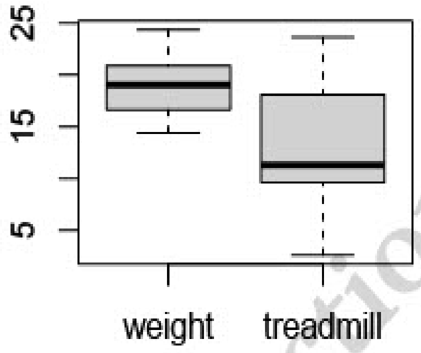
Note that comparative boxplots of paired data are misleading. For example, it's possible that the boxplots will show a huge overlap, but each case in H is higher than the corresponding/paired case in P. If H is greater than P, case-by-case, then the conclusion that H has a higher mean than P is warranted, and yet the boxplots will simply not show that.

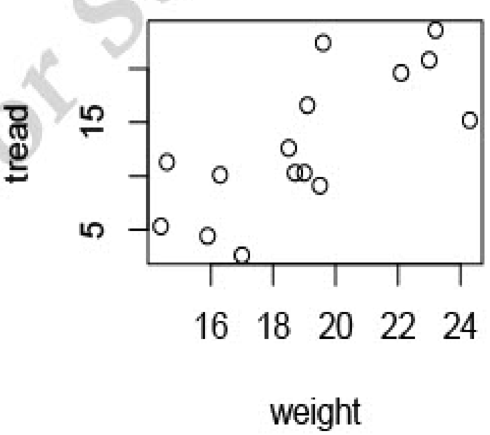
Example 3: Exercise 8.38

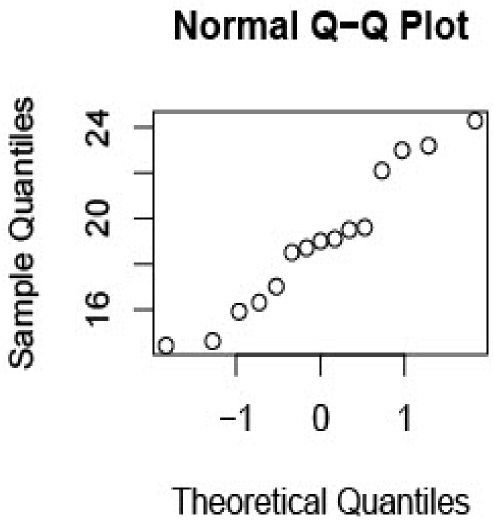
Elevated energy consumption during exercise continues after the workout ends. Because calories burned after exercise contribute to weight loss and have other consequences, it is important to understand this process. The paper "Effect of Weight Training Exercise and Treadmill Exercise on Post-Exercise Oxygen Consumption" (Medicine and Science in Sports and Exercise, 1998: 518??522) reported the accompanying data from a study in which oxygen consumption (liters) was measured continuously for 30 minutes for each of 15 subjects both after a weight training exercise and after a treadmill exercise. Carry out a formal test to decide whether there is compelling evidence for concluding that true average consumption after weight training exceeds that for the treadmill exercise by more than 5. Does the validity of your test procedure rest on any assumptions, and if so, how would you check the plausibility of what you have assumed?

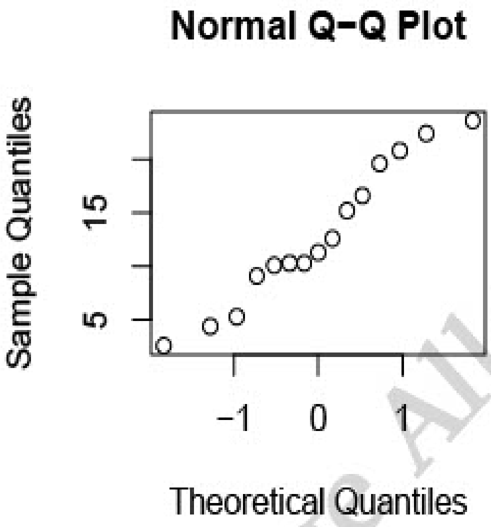
First, ask yourself if the data are paired. In this problem the answer is Yes, based simply on the statement of the problem regarding how the data were collected.

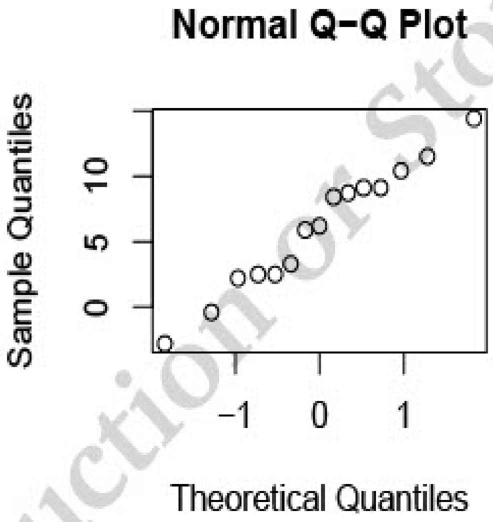
```
weight <- c(14.6, 14.4, 19.5, 24.3, 16.3, 22.1, 23, 18.7, 19, 17, 19.1, 19.6,
            23.2, 18.5, 15.9)
tread <- c(11.3, 5.3, 9.1, 15.2, 10.1, 19.6, 20.8, 10.3, 10.3, 2.6, 16.6, 22.4,
           23.6, 12.6, 4.4)

# Before doing any tests, always "look" at the data:
boxplot(weight, tread, names = c("weight", "treadmill"))
# Note that these data are paired; these boxplots do NOT reflect that fact.
# As such, the comparison of these boxplots is very misleading, and the conclusions
# can be completely wrong for paired data. But it is useful to look at for unpaired
# data.

# The scatterplot of the two variables shows a correlation
plot(weight, tread)
cor(weight, tread)

[1] 0.7419

# Now, the t.test assumes that the population is normal. So, let's see
# if our data are at least consistent with that assumption:
qqnorm(weight)
qqnorm(tread)

# These could look better! But with the small sample size we're dealing with,
# they are normal enough. Also, technically, since we need to do a paired test,
# it is the differences that should have a normal distribution.
qqnorm(weight - tread)
```











Now, suppose the problem had asked if the data suggest that the mean consumption associated with weight training is higher than that associated with treadmill exercise. The hypotheses would be:

$$H_0\colon \mu_{weight} - \mu_{tread} \leq 0$$
$$H_1\colon \mu_{weight} - \mu_{tread} > 0$$

The appropriate CI or test would be the two-sample, 1-sided, t-test. But, which side - the lower or the upper confidence bound?

```
  t.test(weight, tread, paired = T, alternative = "greater")


Paired t-test

data:  weight and tread
t = 4.9, df = 14, p-value = 0.0001
```

```
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.902   Inf
sample estimates:
mean of the differences
                  6.067
```

This particular arrangement of arguments in t.test(), and "alternative = greater" produce the lower confidence bound for $\mu_{weight} - \mu_{tread}$. Here, it's about 3.9, and it's greater than 0. So, we would say that we are 95% confident that the true difference between the means is greater than 3.9. So, there is evidence (from the data) that μ_{weight} is greater than μ_{tread}.

Also, note that the p-value is small (below any of the common α values). So, the conclusion would be "Yes, the data do provide sufficient evidence to reject H_0 in favor of the alternative (that $\mu_{weight} > \mu_{tread}$).

Now, returning to the exact statement of the problem, it asks if there is sufficient evidence that the difference exceeds 5 (not zero). The appropriate hypotheses are:

$$H_0\colon \mu_{weight} - \mu_{tread} \leq 5$$
$$H_1\colon \mu_{weight} - \mu_{tread} > 5$$

This is how you do the t.test():

```
t.test(weight, tread, mu = 5, paired = T, alternative = "greater")


Paired t-test

data:  weight and tread
t = 0.87, df = 14, p-value = 0.2
alternative hypothesis: true difference in means is greater than 5
95 percent confidence interval:
 3.902   Inf
sample estimates:
mean of the differences
                  6.067
```

The lower confidence bound is still about 3.9 . But the conclusion is now different. Because 3.9 is lower than 5 (i.e., 5 is inside the "interval"), we CANNOT say anything! The most we can say is that there is **insufficient** evidence to conclude that the true average consumption after weight training exceeds that for the treadmill exercise by more than 5.
The p-value approach leads to the same conclusion:

The p-value for this test is different from that of the previous part. Now, the p-value is large (larger than any common value of α). As a result, the data do NOT provide sufficient evidence to reject H_0 in favor of H_1 (that $\mu_{weight} - \mu_{tread} > 5$).

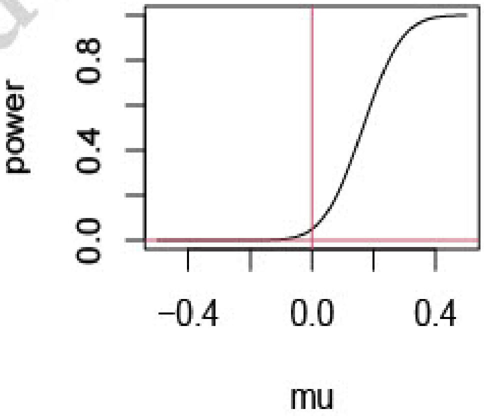
```
# Computing the last p-value "by hand." First, compute the observed statistic
# (z, t, ...)
t_obs <- (mean(weight - tread) - 5) / (sd(weight - tread) / sqrt(15))
t_obs   # Note that this agrees with t_observed given in t.test().

[1] 0.8681
```

```
# According to the formulas for the paired t-test, we should find the area
# under the t-distribution to the right (see H1) of this t_observed:
pt(t_obs, lower.tail = F, df = 15-1)  # p-value =  upper tail.

[1] 0.2

# Note that gives the same p-value as t.test().
```

6.3 Power of A Test

Recall that the power of a test is defined as *power* $= 1 - \beta$ where β is the probability of making a type II error. The main reason why we care about power is that just concentrating on α (which is what most people do) has some serious and adverse consequences in decision making.

```
n <- 100  # Sample size.
pop.sd <- 1  # Population standard deviation.
mu0 <- 0  # the null parameter.

alpha <- 0.05
a <- qnorm(1 - alpha, mu0, pop.sd / sqrt(n))  # Value of x_bar with right-area = 0.05.
a  # Note this is not 1.64, but 1.64/(sigma/root(n)).

[1] 0.1645

mu <- seq(-0.5, 0.5, 0.01)  # Different values of mu.
power <- pnorm(a, mu, pop.sd / sqrt(n), lower.tail = F)
# Note that we are doing a one-sided test because H1: mu > mu0.
plot(mu, power, type = "l")
abline(v = 0, col = 2)
abline(h = 0, col = 2)
```



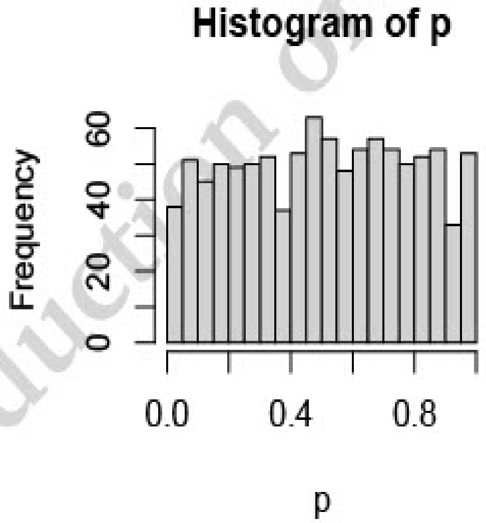
Recall what $\alpha = 0.05$ means: If we use the above procedure for testing H_0 vs. H_1 many many times, about 5% of the time we will commit a type I error. In other words, we will reject H_0 when it is in fact true. In this problem, we will say μ is not zero, when it really is zero. But what fraction of the time will we reject H_0 when it is in fact False? That's power. So, in this case if μ is 0.4, then nearly 99% of the time we will correctly reject $\mu = 0$.

6.4 Distribution of p-values

Have you wondered what the distribution of p-values (say, from a 2-sided, 2-sample t-test) is under the null hypothesis (of equal means)? The following simulation shows that the p-values have a uniform distribution.

```
mu.1 <- 0  # mu of population 1.
mu.2 <- 0  # mu of population 2.
n.trials <- 1000  # Number of samples to take.
p <- numeric(n.trials)  # Space for storing p-values.
for(i in 1:n.trials) {
  x1 <- rnorm(100, mu.1, 1)  # Sample of size 100 from population 1.
  x2 <- rnorm(100, mu.2, 1)  # and from population 2.
  p[i] <- t.test(x1, x2)$p.value
}
hist(p, breaks = 20, xlim = c(0, 1))  # The distribution is uniform.
range(p)  # Note that some p-values are really really small.
```

```
[1] 0.001452 0.998470
```



This result will seem either obvious or completely mysterious. It's not easy to make it intuitive, but think of it this way: If the null hypothesis is true, e.g., if there really is no difference between two population means, then what else can the distribution of p-values be? Any distribution other than uniform would have some nontrivial location (e.g., mean), or scale (e.g., std dev), which means that it cannot be a general/universal answer to the question.