Week 8 - Hypothesis Testing

Nicholas Kortessis

2025-03-05

Hypothesis Testing

So far we have done estimation (and graphing). It's now time to think about another technique and objective in the biostats toolbox: hypothesis testing. Estimation is really about trying to bounds on population parameters of interest. Those bounds are usually quantified using confidence intervals. In hypothesis testing, we are evaluating whether or not there is sufficient evidence in support of a particular statement about the world.

Hypothesis Testing: The Basic Procedure

1. To do hypothesis testing, we need a hypothesis to test. **A hypothesis is a statement about population parameters that reflects some understanding of the biology of your system under study**. For example, you might make be interested in the sex ratio of males to females in a particular species. A hypothesis would be a statement about the proportion of the population that is female (or male). And this could be motivated by biological theory (either the theory of independent assortment and segregation of chromosomes or the evolutionary stability of particular sex ratios under selection).
2. Once you have a hypothesis, you have to think about **a sampling design,** which details how many individuals you take from the population and how they are to be sampled.
3. Once you have a sampling design, you need to detail what you might measure about your sample that estimates the population parameter related to your hypothesis. The thing you measure from your data that links to the population parameter is called an **estimator.**
4. With a sampling design and an estimator, **you need a sampling distribution that tells you the probability of seeing a any given estimate from your estimator under the stated hypothesis.**
5. With the sampling distribution of the estimator, you then can decide on **rejection regions**, which are estimates that have sufficiently low probability that they are consider inconsistent with the hypothesis.
6. After you have your rejection region, you actually sample the population, calculate your estimate, and then see if it is in the rejection region. If it is, you can reject the hypothesis. If not, you *fail to reject your hypothesis* and keep it.
7. After the fact, you can calculate a ***p*-value** that gives the probability of seeing an outcome at least that extreme as your estimate under the hypothesis. If the estimate from your sample is consistent with the hypothesis, it will not be extreme and there is a high value (large p-value) of observing an outcome more extreme. If you observe something inconsistent with the stated hypothesis, the probability of seeing something more extreme will be low (a small p-value). Small p-values indicate that we should be surprised by our sample if the hypothesis is true.

A Note on Null Hypotheses

How do we pick a hypothesis to test? We default to what are called **null hypotheses**. These are simple statements about the world that set a baseline expectation to refute before moving on to more complicated and subtle hypotheses about the world. Typical null hypotheses include statements such as "there is no effect of treatment (or some other categorization)".

What does "**effect**" mean in this context? Effect means the difference between a population parameter and that expected under some basic biological theory. For example, sex ratios often differ from 50:50 because there are average fitness benefits of being male or female in a population that has an equal sex ratio. A null hypothesis of 50:50 sex ratio then relates to the fact that it assumes no effects of selection (and so all that is left is chromosome segregation).

Example 1: Binomial Test of Sex Ratio in Birds

Let's do the whole process with some data on sex ratios from a study on great reed warblers. ***Acrocephalus arundinaceus*** is the great reed warbler (Figure 1). In 1999, Bensch et al. wrote a paper called "Do females adjust the sex of their offspring in relation to the breeding sex ratio?" in the Journal of Evolutionary Biology (paper here). They noticed that the ratio of breeding birds varied predictably from year to year. Theory of natural selection suggests that if these birds could control the sex of their offspring, they should bias their offspring ratio to the less common sex. For example. if there are more breeding males than females, it would be advantageous for their offspring to have more females than males. Each female offspring would then have more potential mates than each female offspring. And that's a selective benefit to a biased sex ratio.

Figure 1: Figure 1. Great Reed Warbler. By Andreas Trepte - Own work, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=33325648

To test their hypothesis, they measured sex ratios of breeding birds and used PCR on blood samples from 9-day-old nestlings to determine the sex ratio of offspring. They looked for biased sex ratio of offspring in years with the most extreme breeding sex ratio. The year when the male sex ratio was lowest (between 30% and 40%) they had the following data.

Male Offspring	Female Offspring
117	110

Step 1: Set the null hypothesis

Let's begin with a simple null hypothesis (typically written as H_0). The basics of chromosome segregation and independent assortment suggest that each offspring has the same chance of being male or female. As such, let's set the null hypothesis as the probability of being male as $\Pr(\text{Male}) = p = 0.5$. Hence,

$$H_0 : p = 0.5$$

and

$$H_A : p \neq 0.5.$$

This is an implicit statement that there is no effect of natural selection (or sexual selection) going on in this population.

Step 2. Specify the sampling design

Let's assume they sampled offspring randomly. During their random sample, they had $n = 227$ individuals.

Step 3. Pick an estimator

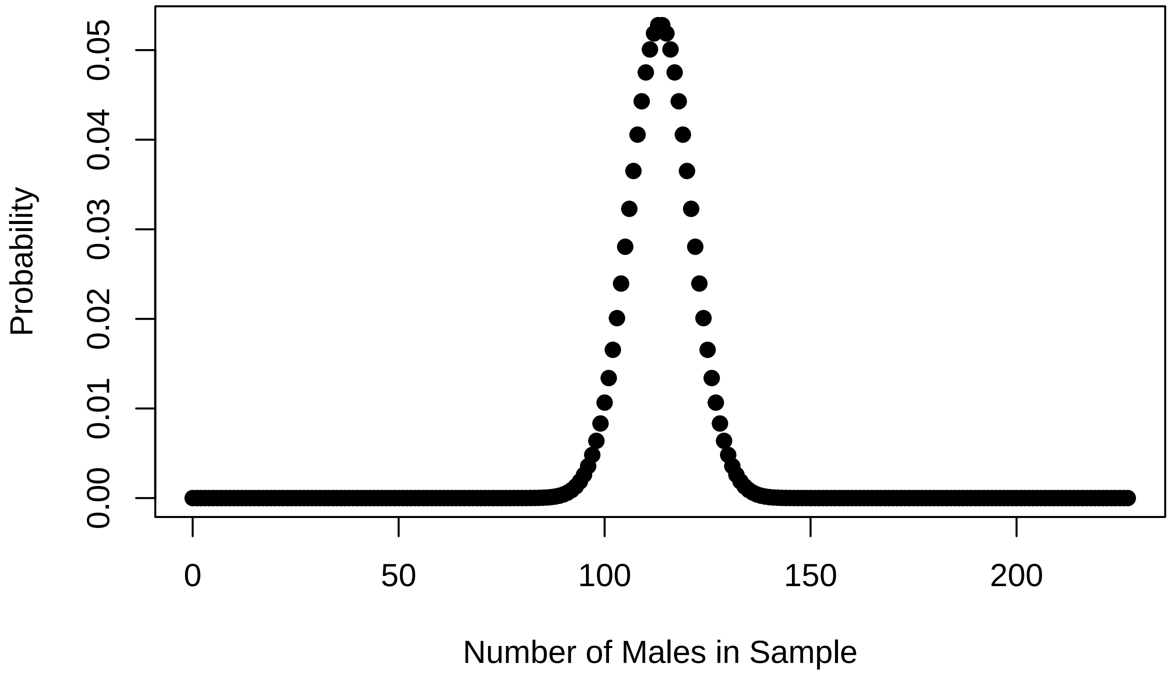
The estimator we will use is M, the number of males in the sample of offspring. This is easily related to the hypothesis p because M/n is the proportion of offspring in the sample that are male, just as p gives the proportion of offspring in the population that are male.

Step 4. Specify the sampling distribution under the null hypothesis.

Since we have random sampling of a categorical character with only two categories (sex: male/female), then a suitable sampling distribution of this character in terms of the number of males is the binomial distribution. Here is the sampling distribution for the number of males in 227 offspring **based on our null hypothesis**.

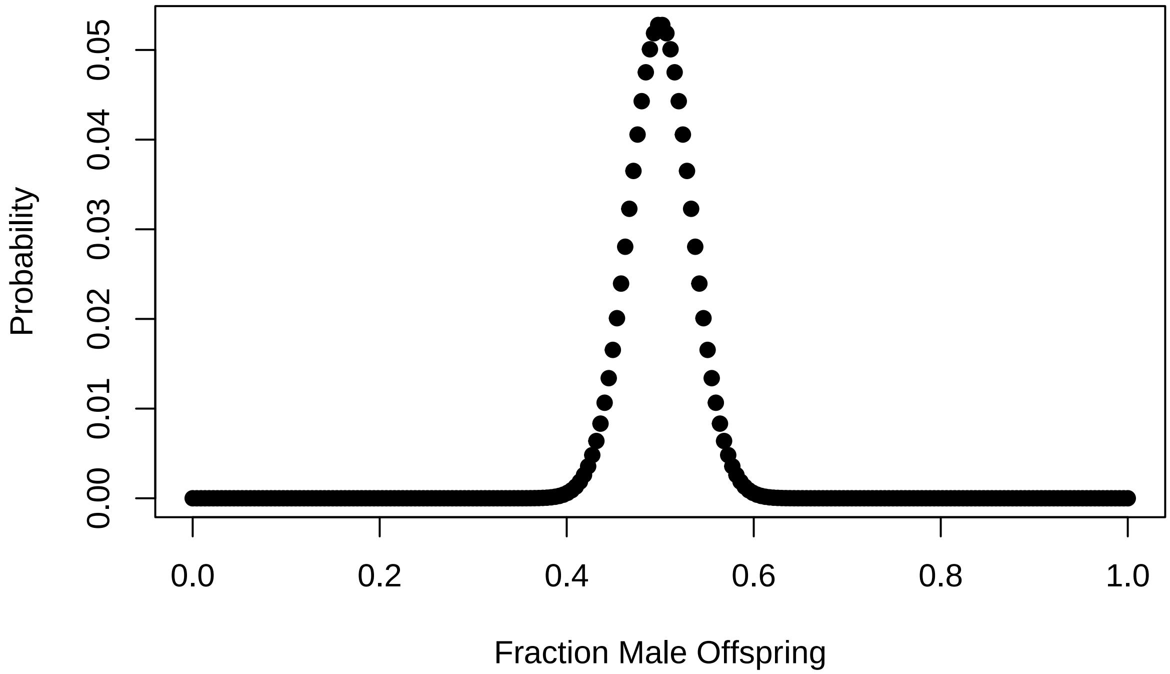
$$\text{Number of Males} \sim \text{Binomial}(n = 227, p = 0.5).$$

```r
n <- 227 # Sample size
p0 <- 0.5 # Null hypothesis

# Possible outcomes
num.males <- 0:n
# Sampling distribution probabilities
prob <- dbinom(num.males, size = n, prob = p0)

plot(num.males, prob, xlab = 'Number of Males in Sample',
     ylab = 'Probability', pch = 19)
```



Here is the same in terms of the fraction of males.

```r
frac.male <- num.males/n
plot(frac.male, prob, xlab = 'Fraction Male Offspring', ylab = 'Probability',
     pch = 19)
```



Step 5. Find rejection regions

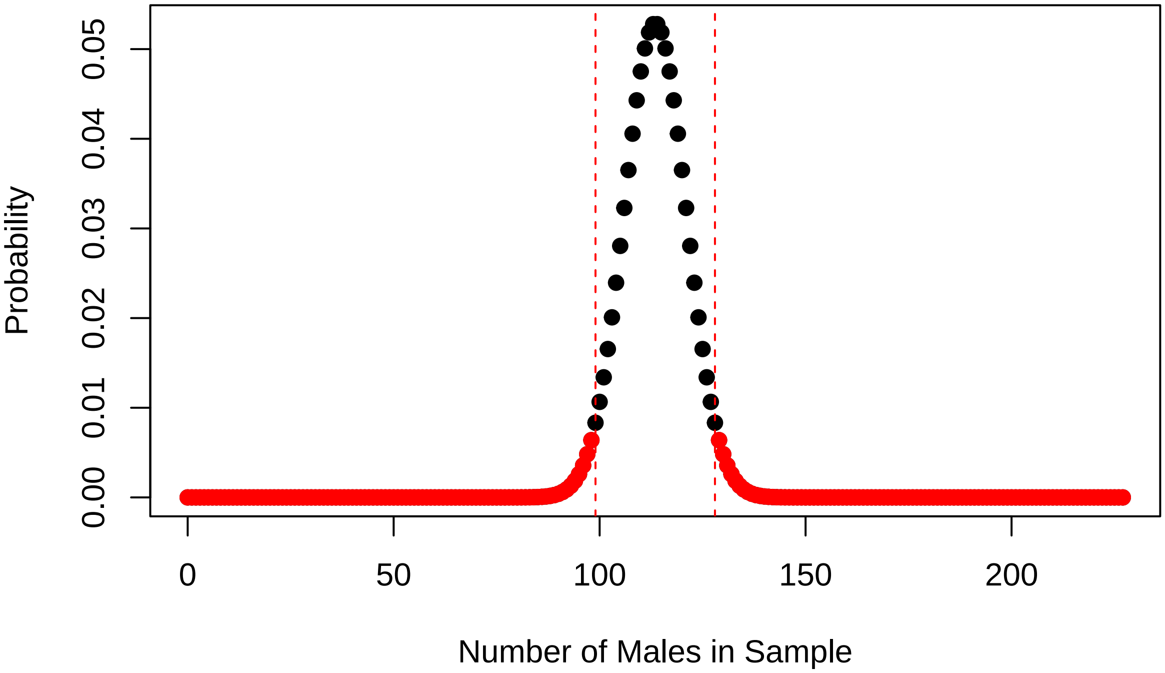
To find the rejection regions, we need to specify the Type I error (the false negative probability). A common choice is $\alpha = 0.05$ meaning we are comfortable falsely rejecting the null hypothesis (if it is true) 5% of the time we run such a test.

We then need to find the 5% least likely outcomes. Let's do this with the quantile function and color those outcomes in red.

```
alpha <- 0.05
(crit.values <- qbinom(c(alpha/2, 1-alpha/2), size = n, prob = p0))
```

```
## [1]  99 128
```

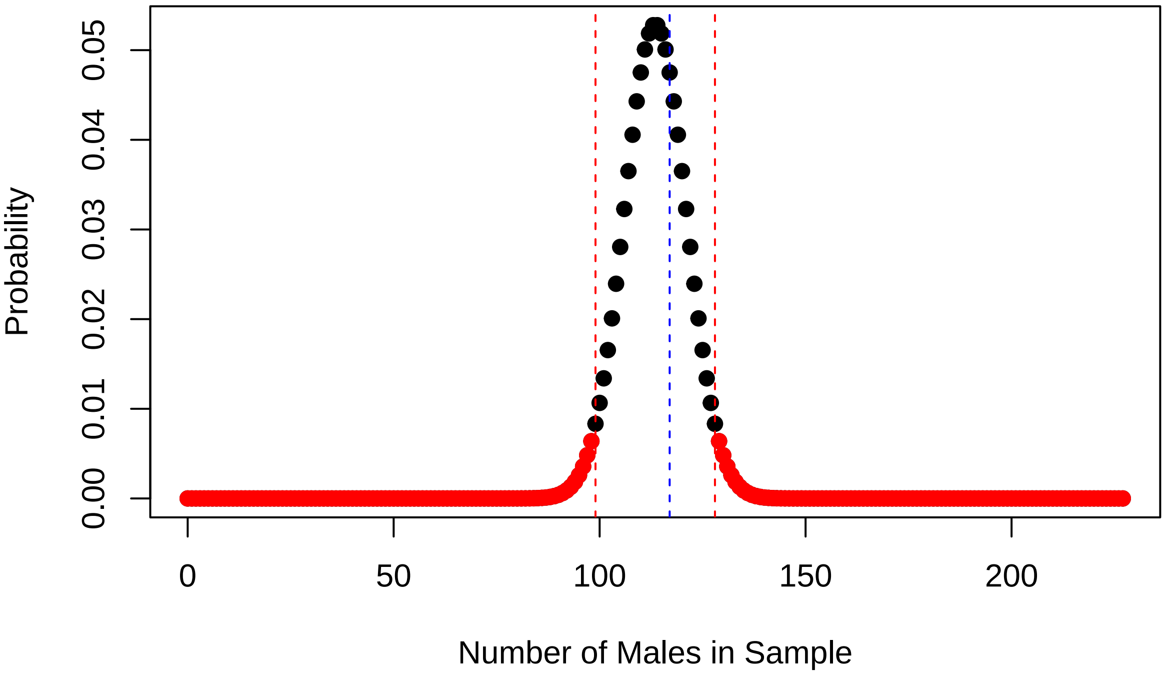
```
plot(num.males, prob, xlab = 'Number of Males in Sample',
     ylab = 'Probability', pch = 19)
abline(v = crit.values, lty = 2, col = 'red')
points(num.males[num.males<crit.values[1]], prob[num.males<crit.values[1]],
       pch = 19, col = 'red')
points(num.males[num.males>crit.values[2]], prob[num.males>crit.values[2]],
       pch = 19, col = 'red')
```



All the red points constitute the rejection region.

Step 6. Collect data and make a decision about the test

The data in this sample had 117 males. This is within the area of 95% of most likely outcomes and so is not in the rejection region. As such, we fail to reject the null hypothesis of an even sex ratio of offspring.



Step 7. Calculate a p-value

To calculate a p-value, we need to find out how extreme our sample was compared to the null hypothesis. Under 227 individual, we expect 227/2 = 113.5 males. But the paper found 117. That's 3.5 more males than predicted exactly by the null hypothesis. An equivalently extreme outcomes is having 3.5 fewer males than expected exactly by the null hypothesis. That is, the p-value is the probability of 117 or more males OR 110 or fewer males.

$$p\text{-value} = \Pr(P \geq 117/227) + \Pr(P \leq 110/227).$$

This can be rewritten as

$$p\text{-value} = 1 - \Pr(P \leq 116/227) + \Pr(P \leq 110/227).$$

We can find these using cumulative densities.

```r
p.value <- 1 - pbinom(116, size = n, prob = p0) + pbinom(110, size = n, prob = p0)
p.value
```

```
## [1] 0.6905491
```

There we go. The p-value is quite large, saying we had a pretty good chance of seeing an observation at least this extreme under the null hypothesis. As such, we should not be surprised by this outcome if the null hypothesis is true.

Checkpoint 1: Write code to calculate the 95% confidence interval for the fraction of offspring in the population that are male. Does the 95% confidence interval contain the null hypothesis $p = 0.5$? This is the exact process for running hypothesis tests. But honestly, it can get a bit tiresome to do all this. R has a way to do it in a much more straightforward way. If you want to run a binomial test, simply use the function `binom.test()` and give it the number of successes and failures and then the hypothesis stated as a probability. Here is how it works here (you can also add in a `conf.level` argument to determine the level of the confidence interval to compute).

```
binom.test(c(117,110), p = 0.5, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  c(117, 110)
## number of successes = 117, number of trials = 227, p-value = 0.6905
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4483723 0.5820568
## sample estimates:
## probability of success
##              0.5154185
```

Does your 95% confidence interval match this? The p-value that we calculated above does!

What this means is that there isn't much evidence that the sex ratio is different from 50:50 in these birds among their offspring. Maybe it is and maybe it isn't. But given the data we have, we don't have enough evidence to say the sex ratio isn't 50:50.

Checkpoint 2: The breeding bird sex ratio in one of the years was 38% male. Run a binomial test for on the offspring data to evaluate the statement that offspring sex ratios are the same as breeding bird sex ratios. Do this test with a Type I error probability of 0.01.

Example 2: t-Test of soil pH

Binomial tests are pretty straightforward. What about non-categorical data? As before with estimation, there is a lot that can be done with normally distributed individual characters. If we want to make a hypothesis about the population mean, μ, then we use the estimator $\bar{X}$. As with the confidence interval example, the relationship between the population mean and the sample mean is given by a t-distribution. Hence, we can create a sampling distribution of the sample mean $\bar{X}$ relative to the hypothesized mean μ with a t-distribution.

This is what gives the name "t-test" it's name. Just like the binomial test, the t-test is referred to as such because the sampling distribution of the estimator of interest is a t-distribution.

To give an example, let's return to the worms data set.

```
worms.df <- read.table(file = "worms.txt", header = TRUE)
str(worms.df)
```

```
## 'data.frame':    20 obs. of  7 variables:
##  $ Field.Name  : chr  "Nashs.Field" "Silwood.Bottom" "Nursery.Field" "Rush.Meadow" ...
##  $ Area        : num  3.6 5.1 2.8 2.4 3.8 3.1 3.5 2.1 1.9 1.5 ...
##  $ Slope       : int  11 2 3 5 0 2 3 0 0 4 ...
##  $ Vegetation  : chr  "Grassland" "Arable" "Grassland" "Meadow" ...
##  $ Soil.pH     : num  4.1 5.2 4.3 4.9 4.2 3.9 4.2 4.8 5.7 5 ...
##  $ Damp        : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ Worm.density: int  4 7 2 5 6 2 3 4 9 7 ...
```

The worms data set include information about the pH of each of the soils. As a baseline, we could assume the pH is neutral on average in these different fields (there are 20). A neutral pH is about 7. So let's set our null hypothesis to

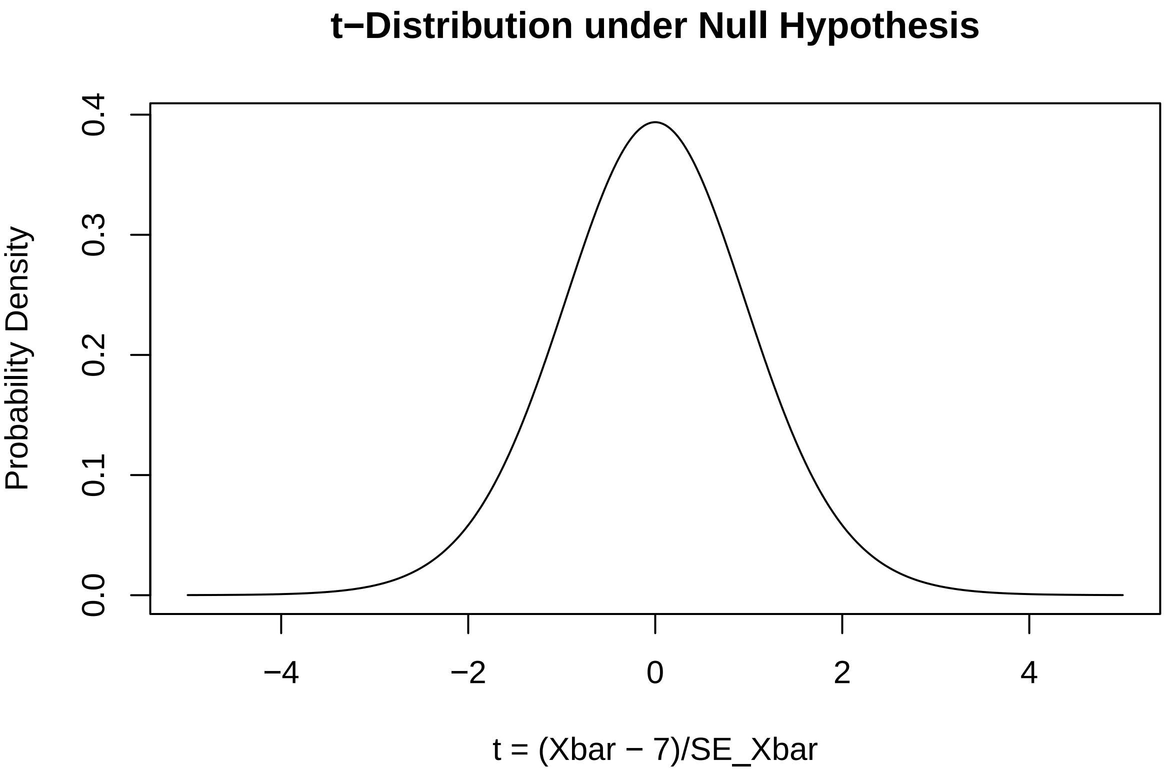
$$H_0 : \mu = 7$$

and

$$H_A : \mu \neq 7.$$

This means that *the average pH across all the fields* is 7. To estimate this from some data, we could measure the sample average, $\bar{X}$. If the null hypothesis is true, the difference between $\bar{X}$ from 7 (i.e., μ) when scaled by the standard error ($SE_{\bar{X}}$) is given by a t-distribution with df = n-1. Let's look at this null sampling distribution.

```
n <- length(worms.df$Soil.pH)
df <- n-1

poss.t.values <- seq(from = -5, to = 5, length = 1000)
t.prob <- dt(poss.t.values, df)
plot(poss.t.values, t.prob, typ = 'l',
     xlab = 't = (Xbar - 7)/SE_Xbar', ylab = 'Probability Density',
     main = 't-Distribution under Null Hypothesis')
```



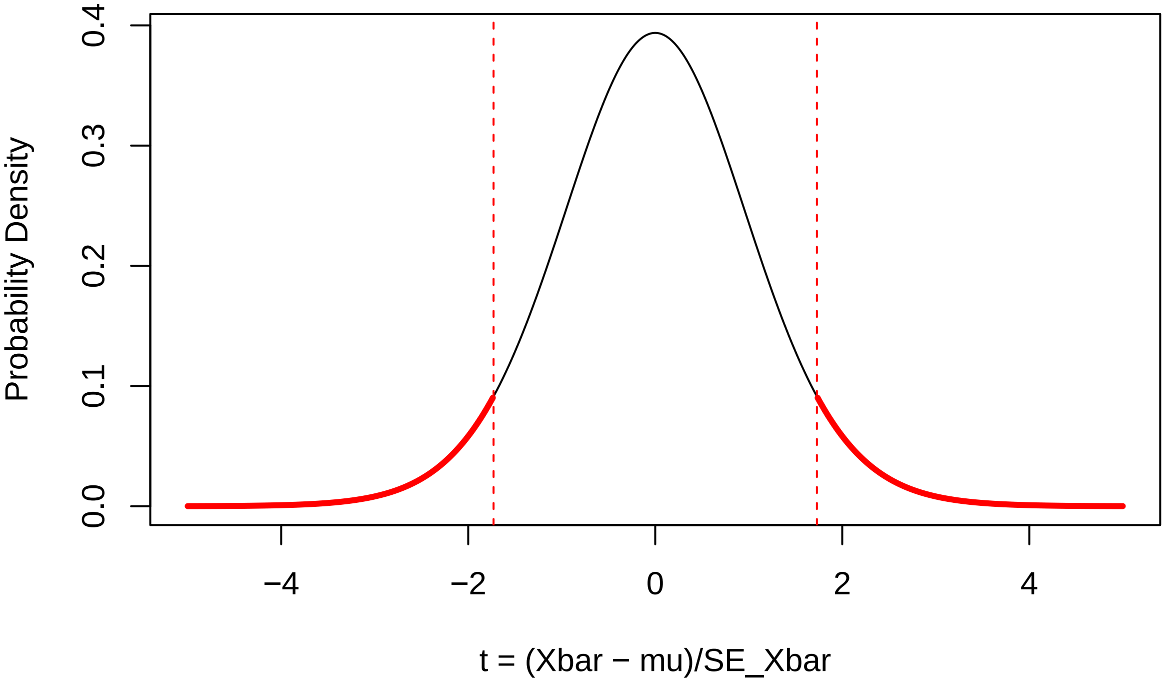
Okay. Now let's decide on a Type I error probability. I'm okay being a bit more wrong here. It's just fields and I want to know whether we are dealing with acidic or basic soils. Let's say Pr(Type I error) = α = 0.1.

```
alpha = 0.1
crit.values <- qt(c(alpha/2, 1- alpha/2), df)
crit.values
```

```
## [1] -1.729133  1.729133
```

```
plot(poss.t.values, t.prob, typ = 'l',
     xlab = 't = (Xbar - mu)/SE_Xbar', ylab = 'Probability Density')
abline(v = crit.values, col = 'red', lty = 2)
lines(poss.t.values[poss.t.values<crit.values[1]], t.prob[poss.t.values<crit.values[1]],
      col = 'red', lwd = 3)
```

```
lines(poss.t.values[poss.t.values>crit.values[2]], t.prob[poss.t.values>crit.values[2]],
      col = 'red', lwd = 3)
```



This says that if the t-value from our sample is more than 1.73 or less than -1.73, then we reject the hypothesis at the $\alpha = 0.1$ level. Stated differently, if our sample mean $\bar{X}$ is more than 1.73 standard errors from $\mu = 7$, then that is sufficiently unlikely to warrant rejecting the null hypothesis that $\mu = 7$.

```
Xbar <- mean(worms.df$Soil.pH)
s <- sd(worms.df$Soil.pH)
SE.Xbar <- s/sqrt(n)
(t.sample <- (Xbar - 7)/SE.Xbar)
```

```
## [1] -18.97462
```

That t-value is clearly many more than 1.73 standard errors away from 7. In fact, our sample is almost 19 standard errors below a neutral soil pH. That's very strong evidence against the null hypothesis and towards the idea that these soils are acidic.

Let's calculate a p-value. We want to know the probability of getting an outcome at least this extreme. In our case, we want to know the probability of t-values smaller than -18.97 or larger than 18.97. We can calculate this with cumulative densities.

```
p.value <- pt(t.sample, df) + 1 - pt(-t.sample, df)
p.value
```

```
## [1] 8.260059e-14
```

That's a very small p-value, indicating that we should be VERY surprised to see this result if the null hypothesis is actually true.

Like the binomial test, R has a way to do this faster. We use the function `t.test`. This function always sets the null hypothesis to 0. To make that work here, we just have it model the difference of the data from our null hypothesis of 7.

```
t.test(worms.df$Soil.pH - 7)
```

```
##
##  One Sample t-test
##
## data:  worms.df$Soil.pH - 7
## t = -18.975, df = 19, p-value = 8.272e-14
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.714699 -2.175301
## sample estimates:
## mean of x
##    -2.445
```

You can see that this matches our numbers almost exactly (any differences are due to approximations made by the function `t.test` that don't have any real effect). You can also see that this is called a "One Sample t-test", which literally means we do this with a single sample and are comparing the mean of that sample with a specific number, here 7.

There is another way to write this using function notation.

`t.test(Soil.ph - 7 ~ 1, data = worms.df)` gives the exact same result.

Checkpoint 3: I looked up typical soil pH for grasslands and found the average is near 5.5. Run a t-test to evaluate whether there is evidence that the grasslands in this data set have atypical soil pH. If they are, are they more acidic or more basic than typical grasslands? What is the 95% confidence interval for soil pH?

Checkpoint 4: Make a figure that shows the distribution of the data as well as the hypothesis that $\mu = 5.5$. Does this figure reinforce your understanding of the outcome of the test?

Example 3: A two-sample t-test

A two-sample t-test evaluates a hypothesis about the differences in means of two samples. In nearly all cases, the null hypothesis is a statement that the means of the two samples are the same, i.e., their difference is zero. (This can be generalized to be a test of differences of specific amounts, but that is rarely done).

When your hypothesis is about the relationship between means, then the estimator has to be about the difference in means. A useful one to choose is $\bar{X}_1 - \bar{X}_2$. which estimates how different the means are. If the hypothesis is that the means differ by an amount $\mu_1 - \mu_2$, then the sampling distribution under this hypothesis is again given by a t-distribution. By now you should be comfortable with this idea of sampling distributions and t-distributions. The important point to take home is that the sampling distributions **are constructed under the assumption that null hypothesis is true.**

If the null hypothesis is that the means are the same, then the null distribution is one that describes how different the sample means are under random sampling of the two groups.

Let's give this a try by evaluating the evidence against the null hypothesis that the soil pH is different damp and dry fields. Let's pick an α value of 0.01, meaning that the probability of a Type I error is 1%. We can reject this hypothesis when the p-value is 0.01 or smaller.

```
damp.fields <- subset(worms.df, subset = (Damp == TRUE))
non.damp.fields <- subset(worms.df, subset = (Damp == FALSE))
test <- t.test(damp.fields$Soil.pH, non.damp.fields$Soil.pH)
test
```

```
##
##  Welch Two Sample t-test
##
## data:  damp.fields$Soil.pH and non.damp.fields$Soil.pH
## t = 3.0935, df = 15.596, p-value = 0.007144
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1693188 0.9116336
## sample estimates:
## mean of x mean of y
##  4.933333  4.392857
```

OK. This says that we can reject the null hypothesis that the two kinds of fields have the same pH because the *p*-value is below our α of 0.01. We can extract the p-value specifically using the code

```
test$p.value
```

```
## [1] 0.00714401
```

But in rejecting that, it only means that we go with the alternative hypothesis, i.e., the fields are NOT the same. But which field has the lower (more acidic pH)?

To find that, we have two options. One, we could make a figure to see. Two, we can look at the output of the test. The test does a few things other things besides report a p-value, and these other things are quite valuable to know. We can ask about each of them from the `test` object we saved using `test$<something>` just like we asked for the p-value using the code `test$p.value`.

1. `test$estimate`- The test estimates the mean in each sample. It says x (the first sample we put in, damp fields) has a mean soil pH of 4.93 and y (the second sample we put in, non-damp fields) has a mean soil pH of 4.39. Hence, the non-damp fields are more acidic (with lower pH).

   ```
   test$estimate
   ```

   ```
   ## mean of x mean of y
   ##  4.933333  4.392857
   ```

2. `test.conf.int`-The test also estimates how different the two means are. It is given by the "95 percent confidence interval". It says this interval is [0.17,0.91]. We can interpret this as saying the population difference in means of damp and non-damp sites is in this range with probability 0.95. That's really useful!

   ```
   test$conf.int
   ```

   ```
   ## [1] 0.1693188 0.9116336
   ## attr(,"conf.level")
   ## [1] 0.95
   ```

3. `test$statistic`- The test also reports the calculated t-value. The t-value for our data is positive, saysing the difference between damp and non-damp means is about 3.1 standard errors above null expectation of no difference in the mean of the two groups. That outcome is pretty unlikely.

   ```
   test$statistic
   ```

   ```
   ##        t
   ## 3.093502
   ```

We can do the same thing using the function notation. This is helpful because it includes the names of factors for the sites (Damp can be FALSE or TRUE).

```
test <- t.test(Soil.pH ~ Damp, data = worms.df)
test
```

```
##
##  Welch Two Sample t-test
##
## data:  Soil.pH by Damp
```

```
## t = -3.0935, df = 15.596, p-value = 0.007144
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to
## 95 percent confidence interval:
##  -0.9116336 -0.1693188
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            4.392857            4.933333
```

Note in this case that the t-value has opposite sign and the confidence intervals also have opposite sign.

```
test$statistic
```

```
##         t
## -3.093502
```

```
test$conf.int
```

```
## [1] -0.9116336 -0.1693188
## attr(,"conf.level")
## [1] 0.95
```

This is because R picked `Damp == FALSE` as the first group and `Damp == TRUE` as the second group. This ordering is opposite of the ordering we had before, so the signs have all flipped. R chooses the ordering based on alphabetical order; `FALSE` comes before `TRUE` in that ordering. You can see that if we ask for the estimates.

```
test$estimate
```

```
## mean in group FALSE  mean in group TRUE
##            4.392857            4.933333
```

When we gave R the groups, it orders them in the order we give them in the function.

Checkpoint 5: Make a graph showing the data distributions for soil pH for both damp and non-damp locations.

Checkpoint 6: t-tests rely on the fact that the data can be modeled with a normal. Run a check to see if soil pH in this data set can reasonably be modeled with a normal distribution. Provide reasoning on whether these t-tests are justified or not based on these checks.

Example 4: χ^2 tests for independence

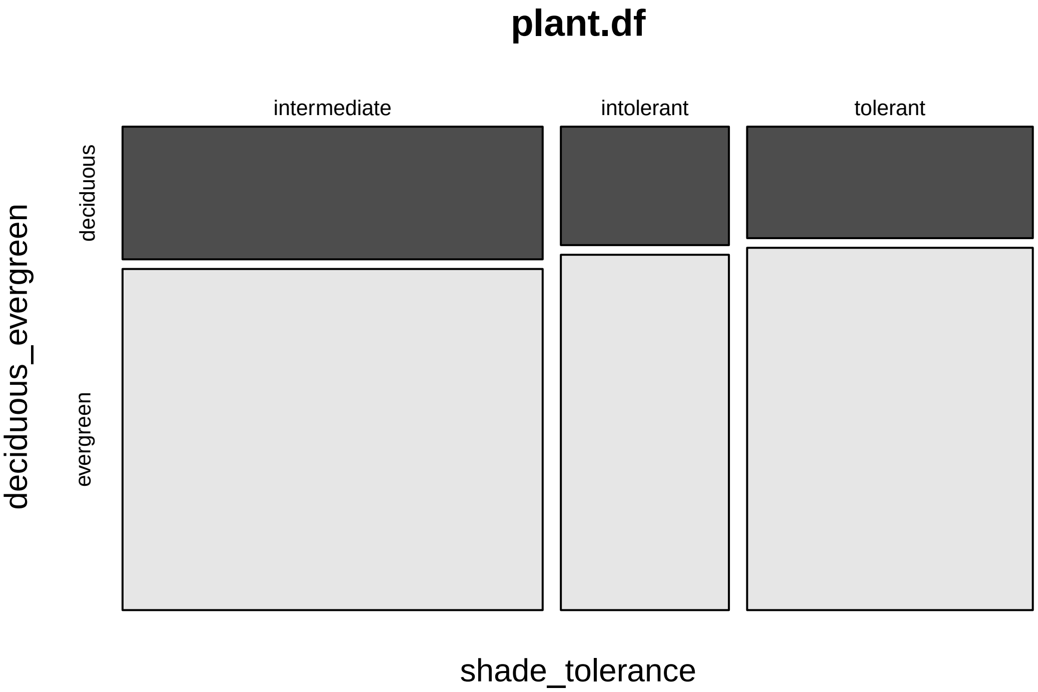
In the week when we talked about probability, we were looking at plant traits. Let's load that data package again.

```
plant.df <- read.csv(file = "species_attributes.csv")
str(plant.df)
```

```
## 'data.frame':    104 obs. of  17 variables:
##  $ species_name          : chr  "Abies_amabilis" "Abies_concolor" "Abies_grandis" "Abies_lasiocarpa"
##  $ family                : chr  "Pinaceae" "Pinaceae" "Pinaceae" "Pinaceae" ...
##  $ genus                 : chr  "Abies" "Abies" "Abies" "Abies" ...
##  $ epithet               : chr  "amabilis" "concolor" "grandis" "lasiocarpa" ...
##  $ seed_development_years: int  2 2 2 2 2 2 2 2 2 2 ...
##  $ pollinator_code       : chr  "wind" "wind" "wind" "wind" ...
##  $ mycorrhiza_type       : chr  "EM" "EM" "EM" "EM" ...
##  $ needleleaf_broadleaf  : chr  "needleleaf" "needleleaf" "needleleaf" "needleleaf" ...
##  $ deciduous_evergreen   : chr  "evergreen" "evergreen" "evergreen" "evergreen" ...
##  $ seed_maturation_timing: chr  "late summer" "fall" "late summer" "late summer" ...
##  $ seed_mass_mg          : num  46.2 34.3 21.1 13.7 78.4 ...
##  $ sexual_system        : chr  "monoecious" "monoecious" "monoecious" "monoecious" ...
##  $ shade_tolerance      : chr  "tolerant" "tolerant" "tolerant" "tolerant" ...
##  $ growth_form          : chr  "tree" "tree" "tree" "tree" ...
##  $ seed_bank            : chr  "no" "no" "no" "no" ...
##  $ fleshy_fruit         : chr  "no" "no" "no" "no" ...
##  $ dispersal_syndrome   : chr  "abiotic" "abiotic" "abiotic" "abiotic" ...
```

We used mosaic plots to ask about the commonness of different categories of one character based on another character. For example, we might want to know whether shade tolerance is more common in deciduous or evergreen trees.

```
mosaicplot(shade_tolerance ~ deciduous_evergreen, data = plant.df,
           color = TRUE)
```



Hmm. It looks like there might be some differences. It looks a bit like intermediate shade tolerant trees are a bit more likely to be deciduous than the trees with other shade tolerance. It appears there is some signal for non-independence between these two characters. How do we test this?

We test it using a χ^2 goodness of fit test. You can run goodness of fit tests whenever you have a number of possible outcomes and you can assign frequencies to those possible outcomes based on some hypothesis.

For our example, there are 6 possible outcomes. The plants in the data set can be

1. Deciduous, shade intolerant
2. Deciduous, shade tolerant
3. Deciduous, shade intermediate
4. Evergreen, shade intolerant
5. Evergreen, shade tolerant
6. Evergreen, shade intermediate.

And that is it. For this question, there are only 6 outcomes for every tree. If we assume that these two characters are independent, that means that probability of being evergreen versus deciduous is the same regardless of the shade tolerance.

We can estimate the shade tolerance by the fraction that are in each shade tolerance category. We can also estimate the fraction evergreen in the population with the fraction evergreen in the sample. The `table` function gives this to us right away.

```
table(plant.df$deciduous_evergreen, plant.df$shade_tolerance)
```

```
##
##              intermediate intolerant tolerant
##   deciduous            14          5        8
##   evergreen            36         15       26
```

This counted them all up for us. 14 + 5 +8 = 27 trees are deciduous, and 36+15+26 = 77 are evergreen. Hence, 27/(27+77) = 0.26 fraction are deciduous. Moreover, 50 are intermediate shade tolerance, 20 are shade intolerant, and 34 are shade tolerant. That means 48% are intermediate shade tolerance, 19% are shade intolerant, and 33% are shade tolerant. If we cross multiply the shade tolerance fractions by the evergreen/deciduous fractions, we have expected fractions of trees in each category.

Checkpoint 7: Find the expected number of trees in each of the 6 categories under the null hypothesis that deciduous/evergreen classification and shade tolerance are independent. These expected counts are compared against the actual counts in the data. The total squared deviations are approximately distributed according to a χ^2 distribution. We can use run this test right away by putting the table from above into the function `chisq.test()`. Let's see

```
chi.test <- chisq.test(table(plant.df$deciduous_evergreen,
                        plant.df$shade_tolerance) )
chi.test
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(plant.df$deciduous_evergreen, plant.df$shade_tolerance)
## X-squared = 0.22234, df = 2, p-value = 0.8948
```

This test gives us the χ^2 statistic of 0.222, which is the total squared deviations of the observed frequencies from the expected under the hypothesis of independence between the two plant characters. Given such small differences between expected and observed, the p-value is quite large, indicating that this result is not surprising given the hypothesis that the two characters are independent.

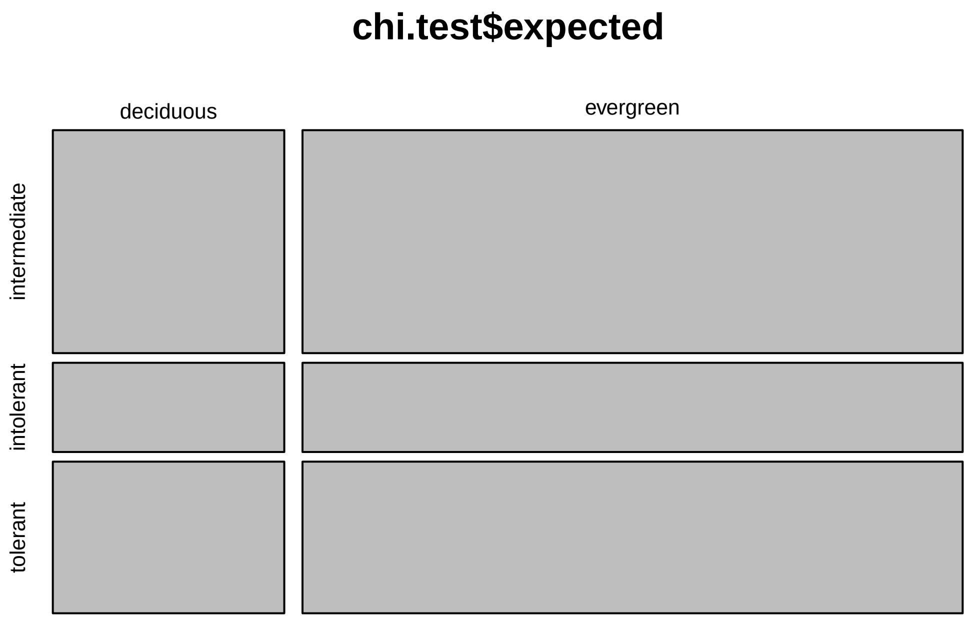
This test also gives us the expected number of individuals in each category based on an assumption of independence.

```
chi.test$expected
```

```
##
##              intermediate intolerant  tolerant
##   deciduous      12.98077   5.192308  8.826923
##   evergreen      37.01923  14.807692 25.173077
```

Here is what independence looks like in a mosaic plot.

```
mosaicplot(chi.test$expected)
```



Checkpoint 8: Make a mosaic plot showing the data of how common the kinds of mycorrhizae are across deciduous and evergreen species (compare only AM and EM mycorrhizae; ignore the others). Mycorrhiza are fungal root symbionts of plants. The association between trees and these mycorrhizal fungi are thought to be mutually beneficial: the trees provide carbon to the mycorrhizae and the mycorrhizae provide nitrogen and water to the plant.

Checkpoint 9: Evaluate the hypothesis that the each plant type (deciduous vs. evergreen) is equally likely to have AM or EM mycorrhizae.

Challenge

The last couple weeks, we were looking at Norwegian Cod. Let's return to that dataset and answer one question.

Checkpoint 10: Are fish different sizes based on their infection status? To answer this question, do the following

1. Choose weight or length to use for fish size and justify your decision using diagnostics we have talked about in the class.
2. Estimate fish size for infected and non-infected fish, including uncertainty.
3. Make a figure showing fish size as a function of infection status. Make sure the figure shows estimates and uncertainty from step 2.
4. Evaluate evidence for this hypothesis using a hypothesis test. Provide a clear statement about the chosen α level, the test outcome with respect to the null hypothesis (reject/fail to reject), and provide a p-value.
5. Estimate, with uncertainty, the differences between the sizes of infected and non-infected fish.
6. Make a clear statement about the "effect" of parasites on fish size. Pose one potential explanation for why this might be.