

QERM 598

First instructor: Mike Keim

A good educational goal

Develop the ability to *translate* a biological hypothesis into statistical terms for objective evaluation and subsequent *testing*.

Example: western white pine (*Pinus monticola*)

Western white pine is a 5-needle soft pine with needles of variable length that grows in the western U.S. and Canada. At northern latitudes, the amount of light and photosynthetically active radiation (PAR) striking the foliage varies by aspect, with southern facing foliage receiving more light than northern facing foliage. Scientists are curious about whether this variation in PAR levels may influence leaf morphology.

Ex: western white pine (*P. monticola*)

- In a perfect world:

Grow 10,000 white pines in greenhouses that perfectly re-create the range of natural conditions under which this tree grows in the wild, *except* for altering the PAR **randomly** by aspect. Measure resulting needle lengths; draw *causal* inference.

- In our world:

Sample from existing trees by aspect, a surrogate for PAR variation. Sample may or may not indicate a statistically significant difference in needle length by aspect. Inference is uncertain (because needle length is a random process) and *not causal* because aspect may be a surrogate for some other factor controlling needle length. More on this later...

Ex: western white pine (*P. monticola*)

Suppose 10,000 trees were randomly sampled across the U.S. and 100 needles were sampled randomly from the north and south aspects of these trees (random w/r/t location in tree and along branch). These samples show needles are longer on the south than the north side of the trees. What can we conclude from these results? What *can't* we conclude from these results?

Ex: western white pine (*P. monticola*)

- There is an association between aspect and needles size in western white pine.
- This may be due to varying levels of PAR by aspect.
- We cannot conclude that PAR is what's causing this difference in sizes, because there are other factors that vary by aspect, e.g. temperature, surrounding vegetation, etc. Hence our inference is *not causal*.
- We will soon discuss what implicit statistics were behind our conclusions and how we translated our biological ideas into statistical hypotheses.

Recall our objective

“Develop the ability to *translate* a biological hypothesis into statistical terms for objective evaluation and subsequent *testing*.”

Hypothesis testing proposes a specific value for a parameter or a numerical relationship between 2 or more parameters and uses sample data to support or refute the hypothesis.

We can **never** prove that a hypothesis is true or false by a statistical test, only that there is evidence for rejecting or not rejecting a hypothesis. Because we have a sample, rather than complete data on a population, there is always the possibility of a mistake in rejecting (or not) a hypothesis. If we do have complete data (e.g. a census), then there is no need for a hypothesis test; descriptive statistics tell us what we need to know.

Rejecting or not-rejecting a hypothesis

The rules of inductive logic only allow us to reject a hypothesis, never to accept it, because we never know if a counter-example is going to show up somewhere down the line.

Example: Suppose we hypothesize that all UW students wear glasses. We then observe 100 students all of whom wear glasses. Can we therefore accept the null hypothesis: all students at UW wear glasses? No: If the 101st student shows up without glasses, then we will need to reject our null hypothesis. When we've only seen 100, all wearing glasses, we say we "fail to reject" the null hypothesis, not that we accept it. (A subtle, but important point.)

Vocabulary for hypothesis testing I

- **null hypothesis** H_0 : The hypothesis we try to reject through our data.
- **alternative hypothesis** H_A or H_1 : The complement of H_0 , this usually describes what we actually think is true about the data.
- **test statistic**: A number we can calculate using our sample data, with a “known” probability distribution if the null hypothesis is true. (Sometimes we don’t know everything about this distribution, but enough to conduct a test.)
- **null distribution**: The probability distribution of the test statistic if the null hypothesis is true.

Vocabulary for hypothesis testing II

- **decision rule:** A pre-defined rule (Important! You must do this before you do your analysis!) that says how unlikely is unlikely enough to reject the null hypothesis, e.g. We will reject the null hypothesis if the observed value of the test statistic is greater than 5.
- **level of the test, a.k.a. α -level:** The probability (under the null) of observing a value more extreme than your decision rule threshold. In the example above, α is the probability of observing a test statistic greater than 5.

Vocabulary for hypothesis testing III

- **p-value:** The probability of observing a test statistic as extreme or more extreme as the value calculated from the sample data. Loosely, “the probability of observing your data under the null hypothesis” (not a true definition since the p-value is only concerned with the test-statistic, not the full distribution of all the data.)
- **the phrase: “under the null”:** “Under the null” is shorthand for “if the null hypothesis were true”. Although we usually doubt that the null hypothesis is, in fact, true, we act *as if it were* and try to show that the resulting test statistic is highly unlikely.

Let’s see these in action through an example.

Step 1: Formulate hypotheses

- H_0 There is no difference in length between needles sampled from the north and south sides of western white pines on the UW campus.
- H_1 There is a difference in length between needles sampled from the north and south sides of western white pines on the UW campus.

Step 2: Translate hypotheses to mathematical form and state decision rule

- $H_0 \mu_N = \mu_S$ where μ_A is the mean leaf length for leaves drawn from the tree at points facing in direction A.
- $H_1 \mu_N \neq \mu_S$

We will reject the null hypothesis if the p-value of our observed test statistic is less than 0.05. (This is a common, BUT ARBITRARY! α -level to use for hypothesis testing.)

Step 3: Choose a test statistic T_0

- Difference in the means

Our test statistic the difference in the sample means, i.e. $T_0 = \mu_N - \mu_S$ which we will estimate by $t_0 = \bar{x}_S - \bar{x}_N$.

If the null hypothesis is true, i.e. there is no difference between leaves drawn from the north and south sides of the tree, what would we expect t_0 to be?

Do we expect it to be *exactly* this value? Why or why not?

Step 4: Determine the null distribution of the test statistic

- Is the observed value of the test statistic unexpected under the null?
- Randomization test

Randomization is one way to generate a null distribution. We will see others, but randomization illustrates what a null distribution, p-value and α -level are.

Randomization I

Here are the ordered lengths (in cm) of 30 needles from 3 trees around campus, sorted by aspect.

N: 3.6, 4.5, 6.9, 7.2, 8.2, 8.3, 8.4, 8.5, 8.8, 8.8, 9.4, 9.7, 10.0, 11.3, 11.4

S: 7.1, 8.0, 8.4, 8.6, 8.7, 9.6, 10.6, 10.8, 10.9, 11.1, 11.5, 11.7, 12.3, 12.5, 12.6

If the null hypothesis is true, then the difference in the average for N and the average for S is just due to sampling variation, not to any real difference in needle lengths. We will use this fact to create a null distribution.

Randomization II

- Randomly draw 15 values from all 30 values, ignoring the distinction between N and S, because if the null hypothesis is true, ignoring that distinction shouldn't matter.
- Calculate the sample mean for these 15, and subtract the sample mean of the other 15, and save this value.
- Repeat this whole procedure 999 times.

The resulting 999 values, plus the actual value for the observed data, give us 1000 samples from the null distribution.

- The R code for doing this is contained in the homework file.

Step 5: Calculate the observed value of the test statistic and find the associated p-value

- Observed test statistic $t_0 = 1.96$

$$\bar{x}_S = 10.29 \text{ cm } \bar{x}_N = 8.33\text{cm, so } t_0 = 1.96.$$

- p-value for t_0

Among the 1000 values generated under the randomization, t_0 is the 996th most extreme, meaning $\Pr(T_0 > t_0) = 1 - 996/1000 = 0.004$

Because there is sampling variability in how R does the re-sampling randomization, the p-value you find when you run the code may be different from .004, but it should be close to this value.

Step 6: Based on p-value, α -level and decision rule, either reject or do not reject H_0 .

Since $p = 0.004$ is very small (well below our α -level, we conclude that it would be very unlikely to observe this test statistic if the null hypothesis were true, so we reject H_0 , the hypothesis that there is no difference in leaf length by aspect.

Some more topics

- What can go wrong with hypothesis testing?
- Common mistakes about p-values
- hypothesis testing \neq not thinking
- One and two-tailed tests

Do we compare $p < \alpha$ or $p < \alpha/2$?

Kinds of error in hypothesis testing

We've discussed the limitations of inference based on a sample. The possibility of some errors can be reduced but not eliminated. What are these errors?

- Rejecting H_0 when H_0 is true.
(Type I error)
- Not rejecting H_0 when H_0 is false.
(Type II error)

The names Type I and Type II are not helpful, but they are in common use in the statistical and scientific community, so you need to know them.

Level and power of a test

We want to avoid both type I and type II error. Given that we can't eliminate their possibility, it is important to know as much as possible about how often we're making these errors. A good hypothesis test procedure allows you to calculate your error rates.

- Level of a test

The level of a test, called the α -level, is the probability of making a type I error, i.e.

$$\alpha = Pr(\text{rejecting } H_0 \mid H_0 \text{ is true}).$$

(We read the RHS of that equation as “the probability of rejecting the null hypothesis, given that the null hypothesis is true.”)

Level and power of a test

- Power of a test

The power of a test, often denoted $1-\beta$, is the probability of not making a type II error.

$$\beta = Pr(\text{failing to reject } H_0 \mid H_0 \text{ is false})$$

so $1-\beta = Pr(\text{rejecting } H_0 \mid H_0 \text{ is false})$.

You want a test with high power and a small α -level, but these are at odds. For a fixed sample size, you cannot increase power without raising your α -level, and you cannot decrease α without losing power. Increasing sample size is almost always good for reducing the probability both types of error.

More on β and test power

H_0 usually specifies one value for a parameter, or a parameter relationship, e.g. $\mu_S - \mu_N = 0$, whereas the H_A usually specifies multiple values, e.g. $\mu_S - \mu_N \neq 0$. We can calculate a null distribution because our H_0 specifies one value to use, so making probability statements “under the null” is possible. Making probability statements “under H_A ” is much more difficult. As a result, the power of a test is often indeterminate.

In practice, researchers often want a certain power level to detect a particular difference with a set probability, and they want to know what sample size will give them that power. For our white pine example, we might want to know how many needles to sample to detect a difference in the means of 0.5 cm with 80% probability, i.e.

$$\beta = \Pr(\text{don't reject } H_0 \mid \mu_S - \mu_N = 1.5) = 0.2$$

A common mistake about p-values

- What a p-value is:

The probability of observing a test statistic as extreme as that calculated from the data, given that the null hypothesis is true. In other words, the probability under the null of drawing a sample with a test-statistic as far in the tails as the observed test statistic.

- What a p-value isn't:

The probability that the null hypothesis is true.

Whether probability can be assigned to a hypothesis is a matter of some debate, but *both* sides of that debate agree that a p-value does not give a probability about the truth or falsehood of a hypothesis.

One-tailed vs. two-tailed tests

- One-tailed test example

$$H_0 : \mu_S - \mu_N \leq 0$$

$$H_A : \mu_S - \mu_N > 0$$

- Two-tailed test example

$$H_0 : \mu_S - \mu_N = 0$$

$$H_A : \mu_S - \mu_N \neq 0$$

- What is the difference?

One-tailed vs. two-tailed tests

In both cases the null and alternative hypotheses are complementary: they cover all possible cases with no overlap. Some statisticians argue that you should not see $H_0 : \mu_S - \mu_N = 0$ and $H_A : \mu_S - \mu_N \geq 0$, but this does appear in some texts. My opinion is that this isn't a big deal; if H_A states $\mu_S - \mu_N \geq 0$, then for testing purposes, we're only interested in the case $H_0 : \mu_S - \mu_N = 0$, which is a subcase of the complementary $H_0 : \mu_S - \mu_N \leq 0$.

One-tailed vs. two-tailed tests

When should we use a one-tailed test?

- Within H_A , the treatment effect can only produce an effect in one direction, e.g. a toxicant (increased dose only increases a mortality rate.)
- Prior study demonstrate a one-directional effect. (Need to know your subject very well.)
- A decision will be made and only one alternative plan of action exists if H_A is true, e.g. herbicide where

$$H_0 : \mu_c \leq \mu_t \text{ vs. } H_A : \mu_c > \mu_t, \text{ where}$$

μ is mean weed abundance at 4 weeks for a control (c) and a treatment (t).

Two-tailed tests are more common.

What's clear at this point:

- Conceptual understanding of rejecting a null hypothesis
- Relation of a test statistic to a hypothesis
- Calculation of a test statistic
- Idea of a null distribution
- interpretation of α -level and p-value

What's not clear at this point:

- How do we *get* a null distribution?
- How do we find the p-value for the observed value of our test statistic?

We'll talk more about this in the context of t-tests next week.