Sets 25 to 27

Example: The maximum safe level of bacteria per mL sample of water is 50. Suppose that 60 samples of water are taken randomly from our water source. For these 60 observations, the average bacteria content is 51.7 bacteria per mL, with a standard deviation of 7 bacteria per mL.

Are our observations consistent with the water being safe to drink?

- Our sample mean, $\overline{x} = 51.7$ is larger than the maximum safe level of 50.
- This by itself does not mean that *all* of the water is unsafe to drink.
- It is possible the water is safe to drink, but by chance, our 60 observations were higher in bacteria content than others from the same water source.

Are our observations extreme enough to suggest that the statement "the water is safe to drink" might not be true?

- If $\overline{x}$ were even larger than 51.7, this would be a stronger indication of the water no longer being safe to drink.
- If, for example the sample mean were 70, we might feel there is a good chance the water is unsafe. If the sample mean were 200, we would feel almost certainly that the water was unsafe.

In our scenario we are trying to decide between two hypotheses: "the water is safe to drink" and "the water is not safe to drink". A **hypothesis test** is a method in inferential statistics that can help us make this decision.

In many cases, including our example, the hypotheses really are about the true values of one or more parameter.

In a hypothesis test, we have one hypothesis, called the **null hypothesis**, which we assume to be true, unless we have strong enough evidence to the contrary. We denote the null hypothesis by H_0.

The second hypothesis is called the **alternative hypothesis** or **research hypothesis**, and is denoted H_a or H_1. If the evidence against H_0 is strong enough, then we may have reason to believe that H_1 might be true.

For our example, what are we *really* doing with the hypothesis test we want to carry out?

- We are testing proposed values of μ, using $\overline{x}$ as evidence.
- Is $\overline{x} = 51.7$ large enough to make us disbelieve the null hypothesis, that $\mu \leq 50$?
- Does it give us reason to believe that the alternative, $\mu > 50$ might be true?

Is the sample mean alone enough to make a decision?

- The sample mean, like any other measurement, has some variability. (If we were to repeat the experiment again, we would likely find a different value for $\overline{x}$.)
- In our scenario, we found the sample mean from the observed values of 60 iid random variables. We know that the estimated standard error of $\overline{x}$ would be $s/\sqrt{n}$.
- Suppose that H_0 were true, and that $\mu \leq 50$:
  - If $s/\sqrt{n}$ were large, then a value of $\overline{x} = 51.7$ might not be an unusual value.
  - However, if $s/\sqrt{n}$ is small, then a value of $\overline{x} = 51.7$ might be an unusual (or extreme) value.

There are many approaches to hypothesis testing. For now, we will focus on the p-value approach to hypothesis testing.

The **p-value** is a probability. Supposing that H_0 is true, and we were to repeat the experiment, the p-value is the probability of seeing results that are **as extreme or more extreme** than the observations we've already made.

If the p-value is large, our observations wouldn't be unusual if H_0 were true. We might say our observations are consistent with the null hypothesis.

If the p-value is small, then our observations would be unusual if H_0 were true. They are **not** consistent with the null hypothesis.

The **smaller** the p-value is, the **stronger** the evidence we have **against** H_0.

The **larger** the p-value is, the **weaker** the evidence we have **against** H_0.

Why do I say "evidence *against* H_0"?

- H_0 is our default choice; we assume H_0 is true, and will only disbelieve it if we have enough evidence to the contrary.
- The p-value gives us a way of measuring how consistent our observations are with our choice of H_0.
- It is *never* "evidence in favour of H_0" because the same observations might be consistent with several different choices of null hypothesis that you might wish to test.

The p-value is calculated using some appropriate **test statistic**. It be a function of our observations, and of the hypothesized parameters. Since we use it to calculate a probability, then a test statistic should have a known probability distribution.

The p-value approach

1. Define the parameters to be tested.
2. Define H_0 and H_1.
3. Specify the test statistic and the distribution under H_0.
4. Find the observed value of the test statistic.
5. Find the p-value.
6. Report the strength of evidence against H_0:
   - *Very strong* if $p \leq 0.01$
   - *Strong* if $0.01 < p \leq 0.05$
   - *Moderate* if $0.05 < p \leq 0.1$
   - *Little or none* if $0.1 < p$
7. Answer any other questions given (i.e. report the value of the estimate, report the value of the estimated standard error, etc.)

Example: The maximum safe level of bacteria per mL sample of water is 50. Suppose that 60 samples of water are taken randomly from our water source. For these 60 observations, the average bacteria content is 51.7 bacteria per mL, with a standard deviation of 7 bacteria per mL. Is there reason to believe that the water is unsafe to drink?

The test we just performed is called a **one-tailed test**, as H_1 consists of μ being in one region. If our H_1 consists of the parameter being in one of two regions, then we call the test a **two-tailed test**.

Example: A certain medication is supposed to contain $350\ mg$ of the active ingredient per pill. It is known from previous work that this content is normally distributed with a standard deviation of $3.5\ mg$. Suppose a random sample of 5 pills are taken, and the average content is $346.4\ mg$. Is the mean pill content not $350\ mg$?

Note: When our test statistic has a symmetric distribution (Z or t), then the p-value for a two-tailed test is exactly twice that of the equivalent one-tailed test.

Errors and Hypothesis Tests: Two types of errors are possible in a hypothesis test. A **Type I** error (Rejection Error) is made when we reject the null hypothesis when it is true. A **Type II** error (Acceptance Error) is made when we do not reject the null hypothesis when it is false.

For the pill example, a Type I error would be made if we said that the pills didn't contain $350\ mg$ when they actually did.

A Type II error would be made if we said the pills contained $350\ mg$ when they actually didn't

We use α and β to represent the probability of committing a Type I or Type II error, respectively.

Relationship between the errors: If we require extremely strong evidence against H_0 before we'll believe H_1, then we are attempting to make α small; as a consequence, β becomes larger.

If we are willing to accept not particularly strong evidence against H_0 before we'll believe H_1, then we are letting α be larger; as a consequence, β becomes smaller.

The significance-level (rejection region) approach

With this approach, we are asked to test H_0 at some significance level α. We carry out the hypothesis test in much the same way: defining parameters, calculating the value of the test statistic, finding the $p-$value.

Rather than giving the level of strength against H_0, as in the $p-$value approach, we instead either reject or don't reject H_0 by the following rule:

- If $p \leq \alpha$, then reject the null hypothesis.
- If $p > \alpha$, then do not reject the null hypothesis. (Some will phrase this as "maintain the null hypothesis" or "fail to reject the null hypothesis")

Example: For the pill example, if we were asked to test our hypotheses at the level $\alpha = 0.01$, what would our conclusion be?

Example: What if we were testing at the level $\alpha = 0.05$?

Important: It is statistically dishonest to set your value of α after the data has been collected and examined; the value of α should be made by taking into account the consequences of Type I and II errors *before* the study is carried out.

Relationship between hypothesis testing and confidence intervals: Suppose we construct a $(1-\alpha)100\%$ confidence interval for μ.

It is true that for any number k in this interval, that if we were to test $H_0 : \mu = k, H_1 : \mu \neq k$, we'd have a p-value greater than α.

This means that if we were testing $H_0 : \mu = k, H_1 : \mu \neq k$ at the level of α, we would reject the null hypothesis if and only if k were not inside the $(1-\alpha)100\%$ confidence interval for μ.

Example: Using our pill data, we can find that a 95% confidence interval for μ is $(343.77, 349.03)$.

What would our conclusion be if we test $H_0 : \mu = 344, H_1 : \mu \neq 344$ at the level $\alpha = 0.05$?

Example: What would our conclusion be if we test $H_0 : \mu = 342$, $H_1 : \mu \neq 342$ at the level $\alpha = 0.05$?

Example: The lengths of mourning doves (from beak to tail) are known to be normally distributed. Suppose that 5 mourning doves are selected at random, and it is found that the average length of the mourning doves is $32.4\ cm$, with a standard deviation of $2.9\ cm$.

Let μ denote the true mean length of mourning doves. Test the hypotheses $H_0 : \mu = 30$, $H_a : \mu > 30$ at the level $\alpha = 0.1$.

Example: What would the conclusion be if we were testing at the level $\alpha = 0.05$?

Example: In a sample of 46 people, we find the average blood glucose level upon waking up is $5.3\ mmol/L$ with a standard deviation of $1.2\ mmol/L$. Is there reason to believe that the true mean blood glucose level upon waking for people is not $5\ mmol/L$?

Example: Brand X cola decides that if the proportion of customers who prefer their brand ever falls below 40%, they will launch a new ad campaign. They take a survey of 150 randomly selected customers, and find that 57 prefer Brand X over the competitors.

Is there evidence that the true proportion of customers preferring Brand X is less than 0.4? Test at the level $\alpha = 0.2$

Thoughts to consider:

- When we say that our results are **significant**, we mean that our observations were such that we calculated a p-value that was small enough for us to reject H_0.
- Significant does not necessarily mean a *large* difference:
  - A type of corn has a mean height of $100\ cm$. We want to see if a new fertilizer has an effect on the mean height, at the significance level $\alpha = 0.01$.
  - We plant $n = 5000$ plants, and find $\overline{x} = 100.01$ and $s = 0.1$.
  - We test $H_0 : \mu = 100$, $H_1 : \mu \neq 100$, and find the p-value for this test is about 1.55×10^{-12}.
  - We reject H_0, and would conclude that the fertilizer has an effect.
  - We could say there is a **significant difference** between the sample mean $\overline{x} = 100.01$, and the hypothesized mean $\mu = 100$. However, there isn't a really *large* difference numerically between 100 and 100.01.
- In this example, the small p-value (and the large test statistic that generated it) are not because $\overline{x}$ was particularly large, but because s was rather small and n was fairly large.

The **effect size** is a measurement that assesses the size/strength of an effect. The sample correlation coefficient, r is an example of an effect size.

In the case of testing a sample mean, a common effect size used is Cohen's d, where

$$d = \frac{\overline{x} - \mu}{s}$$

- Cohen's d measures how many standard deviations $\overline{x}$ is away from μ; it takes into account both the difference between the means, but also takes into account the variability of the measurements.
- Cohen proposed that:
  if $0.2 \leq |d| < 0.5$, the effect is small;
  if $0.5 \leq |d| < 0.8$ the effect is medium;
  if $|d| > 0.8$, then the effect is large.
- In our example, $d = (100.01 - 100)/0.1 = 0.1$. While the effect of the fertilizer is statistically significant, the effect itself is *very small*.