Intro. to Hypothesis Tests

Two of the most common types of statistical inference:

1. Confidence intervals
   Goal is to estimate (and communicate uncertainty in our estimate of) a population parameter.

2. Tests of Significance
   Goal is to assess the evidence provided by the data about some claim concerning the population.

Basic Idea of Tests of Significance

Example: Each day Tom and Heather decide who pays for lunch based on a toss of Tom's favorite quarter.

Heads - Tom pays

Tails - Heather pays

- Tom claims that heads and tails are equally likely outcomes for this quarter.
- Heather thinks she pays more often.

Heather steals the quarter, tosses it 10 times, and gets 7 tails (70% tails). She is furious and claims that the coin is not fair.

There are two possibilities:

1. Tom is telling truth – the chance of tails is 50% and the observation of 7 tails out of 10 tosses was only due to sampling variability.
2. Tom is lying – the chance of tails is greater than 50%.

Suppose they call you to decide between the two possibilities.

To be fair to both of them, you toss the quarter 25 times. Suppose you get 21 tails.

What would you conclude? Why?

=> The coin is probably not fair. Even with sampling variability it is unlikely that a fair coin would result give such a high percentage of tails. (The actual probability of getting 21 or more tails in 25 tosses is 0.000455 *if the coin is fair*.)

Moral of the story: an outcome that would rarely happen *if a claim were true* is good evidence that the claim is in fact not true.

This is the idea behind **Hypothesis Testing**.

- A **hypothesis** is a statement about the parameters in a population; we will be making statements about **μ** in Section 6.2.

- A **hypothesis test** (or **significance test**) is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess.

- The results of a test are expressed in terms of a probability that measures how well the data and hypothesis agree.

Performing a Hypothesis Test

1. <u>State Hypotheses</u>

   State your research question as two hypotheses - the **null** and the **alternative** hypotheses. These hypotheses are written in terms of the population parameters.

   The **null hypothesis (*H_0*)** is the statement being tested. This is assumed "true" and compared to the data to see if there is evidence *against* it. A null hypothesis that we will see often is that the mean μ is equal to some standard value. Usually, null hypotheses give a statement of "no difference" or "no effect."

Suppose we want to test the null hypothesis that μ is some specified value, say μ_0. Then

$$H_0: \mu = \mu_0$$

Note: We will always express H_0 using an equality sign.

The **alternative hypothesis (*H_a*)** is the statement about the population parameter that we hope or suspect is true. We are interested in seeing if the data support this hypothesis.

- H_a can be **one-sided**: H_a: $\mu > \mu_0$ or H_a: $\mu < \mu_0$
- H_a can be **two-sided**: H_a: $\mu \neq \mu_0$

Example: Strawberry Bars

Kellogg's says that its strawberry bars weigh, on average, 16 oz. 10TV's consumer reporter is suspicious that the bars weigh less than what is claimed. In order to check his suspicion, he weighs the contents of 20 randomly chosen bars. These 20 bars have an average weight of 15.6 oz. Assume that the weights follow a normal distribution with a standard deviation of 0.7 oz. Is there evidence that the reporter's suspicion is correct?

- The hypotheses are: H_0: $\mu = 16$
  H_a: $\mu < 16$

- Is this a one-sided test or a two-sided test?

  This is a one sided test. The reporter thought the bars were smaller than 16 oz.

2) <u>Calculate P-value</u>

We ask: "Does the sample give evidence against the null hypothesis?"

P*-value:** The probability that the sample mean would take a value as extreme or more extreme than the one we actually observed ***assuming H_0 is true.

In the strawberry bar example, this means:

"Is the sample average much less than we would expect to see if μ really is 16?"

- We need to find the probability that we get a sample of 20 bars whose mean is 15.6 or less **given that μ = 16**.

  Question: Why 15.6 or less? This is more extreme evidence against our null hypothesis and in support of the alterative hypothesis.

- What is the distribution of $\overline{x}$, the average weight of the bars sampled if the null hypothesis is true?

  N (16, 0.7/√20)

- What area under the Normal (16, 0.7/√20) curve corresponds to the p-value?

  The area to the left of 15.6.

- Calculate the test statistic (z-score)

$$z = \frac{15.6 - 16}{0.7 / \sqrt{20}} = -2.56$$

- Calculate the P-value:

  Using Table A, the area to the left of -2.56 is 0.0052.

=> The probability of getting more extreme evidence against H_0: μ = 16, given that μ = 16, is 0.0052.

- More on p-values:

  P-values correspond to tail areas under the density curve. For the one-sided test in the strawberry bar example, the p-value was the area to the left of the test statistic.

  What area corresponds to the p-value of a two-sided test? In the case of two-sided tests, we could observe something "as extreme or more extreme" in either direction. The p-value includes two tail areas.

P-values in terms of the test statistic:

H_a	*P*-value	Area under curve
$\mu < \mu_0$	$P(Z \leq z)$	Left of z
$\mu > \mu_0$	$P(Z \geq z)$	Right of z
$\mu \neq \mu_0$	$2P(Z \geq \lvert z \rvert)$	Tails

where z is the observed value of the test statistic and the probabilities are found using the standard normal distribution given in Table A.

- A *P*-value is exact if the population distribution is normal.
- If the population is not normal, the P-value approximates the true probability *for large n* because of the Central Limit Theorem.

3) State Your Conclusions

- The final step is to decide if there is a strong amount of evidence to reject H_0 in favor of H_a. This is accomplished using the *P*-value.
- In our example, we got a *P*-value =0.0052. What does this tell us?

  If H_0 is true (i.e., true mean weight is 16 oz), then the chance of getting a sample whose mean weight is 15.6 oz or less is 0.52%

Does it give evidence against H_0?

Yes, it is very unlikely that we would observe a sample mean as low as we did if H_0 is true.

Conclusion: We reject the null hypothesis.

- A *small* *P*-value is strong evidence *against* H_0. Such a *P*-value says that if H_0 is true, then the observed data are unlikely to occur just by chance.
- The smaller the P-value, the stronger the evidence against the null.

- **Question:** How small does the P-value need to be?

  Prior to testing, it is determined how small the *P*-value must be to be considered decisive evidence against H_0.
  This value is called the **significance level** and is usually represented as α. Typical α levels used are 0.1, 0.05 and 0.01.

  If the *P*-value $\leq \alpha$, reject the null hypothesis.
  If the *P*-value $> \alpha$, do not reject the null hypothesis.

- If we use an α level of 0.01 and the p-value is smaller than α, we can say that there is *a less than one in one-hundred chance* that we would observe data as extreme or more extreme than what we saw if the null hypothesis is true.
- If the *P*-value $\leq \alpha$ we say the data are **statistically significant at level α**.

Note: When we do not reject H_0, we are not claiming H_0 is true. We are just concluding there is not sufficient evidence to reject it.

Example: Ameritech

Suppose last year Ameritech's repair service took an average of 3.1 days to fix customer complaints. One of the managers is assigned to check if this year's data show a different average time to fix problems. He collects a random sample of 30 customer complaints and finds that the average time taken to fix them is 2.1 days. Assume that the standard deviation of the time taken to fix the complaints is 2.5 days. Is this good evidence that the average time taken to fix the complaints is not 3.1 days?

- State the hypotheses: $H_0: \mu = 3.1$
  $H_a: \mu \neq 3.1$
- Calculate the test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{2.1 - 3.1}{2.5 / \sqrt{30}} = -2.19$$

- Calculate *P*-value: Since this is a two-sided test, we find the area to the left of -2.19 and then double it to get the p-value.

P-value = 2 x 0.0143 = 0.0286

- What is your conclusion at the 0.05 level? Since the p-value is less than 0.05, the test is significant at the 0.05 level. We reject the null.
- What is your conclusion at the 0.01 level? The p-value is larger than 0.01, so the test is not significant at this level. We would not reject the null at this level.

Example: EPA

The EPA limit on concentration of PCB in drinking water is 5 ppm. Wells are regularly tested to make sure they are not over the limit. A random sample of 100 water specimens from a well was collected and has an average PCB 5.1 ppm. Is there evidence at 5% level that the well is over the limit? Assume that the PCB concentration varies with standard deviation 0.8.

- State the hypotheses: H_0: $\mu = 5$
  H_a: $\mu > 5$

- Calculate the test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{5.1 - 5}{0.8 / \sqrt{100}} = 1.25$$

- Calculate the P-value: The P-value is the area under the normal curve to the right of 1.25

P-value = 0.1056

- Since the p-value is larger than 0.05, we do not have evidence at the 5% level that the PCB level exceeds the limit. We do not reject the null hypothesis.

Tests from Confidence Intervals

- We have covered two types of statistical inference procedures for the population mean μ:

Confidence Interval (CI) and Tests of Significance

- **Question:** Is there any relationship between hypothesis tests and confidence intervals?

  Answer: Yes, a level α two-sided test rejects a hypothesis $H_0{:}\mu = \mu_0$ exactly when the value μ_0 falls *outside* a level $(1-\alpha)$ CI for μ.

Example: Concrete Block

Bud's Home Center sells concrete blocks. Bud wants to estimate the average weight of all blocks in stock. A sample of 64 blocks has a mean weight of 65.5 lbs. Assume that the weights of blocks vary with standard deviation 4.6 lbs.

- Construct a 95% CI for the mean weight of all blocks.

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 65.5 \pm (1.96) \frac{4.6}{\sqrt{64}}$$

=> (64.373, 66.627) is a 95% CI for μ.

- Bud is interested in knowing if the mean weight of all blocks is 68 lbs or not. State the appropriate hypotheses. Using the above CI, what do you conclude?

H_0:μ=68
H_a:μ≠68

Our confidence interval was 64.373 to 66.627. Since 68 does not fall in this interval, we reject H_0.

The significance level of the above test is 5% (0.05), because we used a 95% confidence interval.

Example: Jupiter

The diameter of Jupiter is measured 100 times independently using a new unbiased process. Using these 100 measurements, a 99% CI for the true diameter is computed to be (88,707 miles to 88,733 miles). Is there evidence at 1% level that the true diameter of Jupiter is not 88,720 miles?

- What would the hypotheses be? $H_0: \mu = 88{,}720$
  $H_a: \mu \neq 88{,}720$

- What do you conclude?

  Since 88,720 falls in the 99% confidence interval (88,707 miles, 88,733 miles), we do not reject the null hypothesis when testing at the 1% level.

More on stating your conclusions:

- When working with hypothesis tests there are many ways to state your conclusion.

- The following four statements convey the same conclusion
  1. The test is significant.
  2. Reject the null hypothesis.
  3. The data show strong evidence against H_0.
  4. The p-value is smaller than α.

- These four statements also convey the same conclusion:
  1. The test is not significant.
  2. Do not reject the null hypotheses.
  3. The data do not show evidence against H_0.
  4. The p-value is larger than α.

- Usually 1,2, or 3 are given as the conclusion and 4 is given as the reason for or explanation of the conclusion.