Significance Tests

- Confidence intervals are one of the two most common types of statistical inference.
- Use a CI when your goal is to estimate a population parameter.
- The second type of inference, called tests of significance, has a different goal: to assess the evidence provided by data about some claim concerning a population.

Steps

1. **Step 1: The Claims**
   - Our aim is to decide between two opposing points of view, Claim 1 and Claim 2. In hypothesis testing, **Claim 1** is called the null hypothesis (denoted "\( H_0 \)"), and **Claim 2** plays the role of the alternative hypothesis (denoted "\( H_a \)").
   - The null hypothesis suggests nothing special is going on, no change from the status quo, no difference from the traditional state of affairs, no relationship.
   - In contrast, the alternative hypothesis disagrees with this, stating that something is going on, or there is a change from the status quo, or there is a difference from the traditional state of affairs. The alternative hypothesis, \( H_a \), usually represents what we want to check or what we suspect is really going on.

2. **Step 1: Stating the hypotheses \( H_0 \) and \( H_a \).**
   - The null hypothesis has the form:
     - \( H_0 : \mu = \mu_0 \) (where \( \mu_0 \) is a specific number).
   - The alternative hypothesis takes one of the following three forms (depending on the context):
     - \( H_a : \mu < \mu_0 \) (one sided)
     - \( H_a : \mu > \mu_0 \) (one sided)
     - \( H_a : \mu \neq \mu_0 \) (two sided)

How to choose?

What determines the choice of a one-sided versus two-sided test is what we know about the problem before we perform a test of statistical significance.

It is important to make that choice before performing the test or else you could make a choice of "convenience" or fall in circular logic.
### One-sided or two-sided?

**Ex.1:** A health advocacy group tests whether the mean nicotine content of a brand of cigarettes is greater than the advertised value of 1.4 mg. This is a one-sided test:

- $H_0: \mu = 1.4$ mg
- $H_a: \mu > 1.4$ mg

### One-sided or two-sided?

**Ex.2:** The FDA tests whether a generic drug has an absorption extent similar to the known absorption extent of the brand-name drug which is 8 hours. Higher or lower absorption would both be problematic, thus we test:

- $H_0: \mu = 8$ hours
- $H_a: \mu \neq 8$ hours

### Step 2: Collecting data and summarizing them

However, before we can do that we need to check the conditions (same conditions we had for CI):

- SRS
- $\sigma$ is known
- Either
  - The population is normally distributed OR
  - The sample size is 30 or greater

### Step 2: Collecting data and summarizing them

- If the conditions are met we can collect data and find the test statistic:
  - Since our parameter of interest is the population mean $\mu$, once we collect the data, we find the sample mean $\bar{x}$.
  - The test statistic is the z-score (standardized value) of the sample mean $\bar{x}$ assuming that $H_0$ is true (in other words, assuming that $\mu = \mu_0$).

### Step 2: test statistic

To test the hypothesis $H_0: \mu = \mu_0$ (where $\mu_0$ is a known or pre-determined constant) based on an SRS of size $n$ from a Normal population with unknown mean $\mu$ and known standard deviation $\sigma$, we rely on the properties of the sampling distribution $N(\mu, \sigma^2/n)$.

The test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

### Example 1

- The SAT is constructed so that scores have a national average of 500 and standard deviation of 100. The distribution is close to normal. The dean of students of Ross College suspects that in recent years the college attracts students who are more quantitatively inclined. A random sample of 4 students from a recent entering class at Ross College had an average math SAT (SAT-M) score of 550.

Does this provide enough evidence for the dean to conclude that the mean SAT-M of all Ross college students is higher than the national mean of 500? Assume that the standard deviation of 100 applies also to all of Ross College students.
Step 1:

- $H_0$: $\mu = 500$
- $H_a$: $\mu > 500$

**Step 2: Conditions**
- The sample is random
- $\sigma$ is known, $\sigma = 100$
- The variable of interest -- SAT-M -- is assumed to vary normally, so the fact that the sample size is small ($n=4$) does not present a problem.

**Step 2: The Test Statistic**

- The sample mean is $\bar{x} = 550$, and so the test statistic is:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{550 - 500}{100 / \sqrt{4}} = 1$$

**Example 2**

- A certain prescription medicine is supposed to contain an average of 250 parts per million (ppm) of a certain chemical. If the concentration is higher than this, the drug may cause harmful side effects; if it is lower, the drug may be ineffective. The manufacturer runs a check to see if the mean concentration in a large shipment conforms to the target level of 250 ppm or not. A simple random sample of 100 portions is tested, and the sample mean concentration is found to be 247 ppm. It is assumed that the concentration standard deviation in the entire shipment is $\sigma = 12$ ppm.
Step 2: Conditions

- The sample is random.
- $\sigma$ is known, $\sigma = 12$
- The sample size ($n=100$) is large enough for the Central Limit Theorem to apply (note that in this case the large sample is essential since the concentration level is not known to vary normally)

Step 2: The Test Statistic

- The sample mean is $\bar{x} = 247$, and so the test statistic is:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{247 - 250}{12/\sqrt{100}} = -2.5$$

Step 3. Finding the p-value of the test

- The p-value -- the probability of getting data (test statistic) as extreme as that observed or even more extreme when $H_0$ is true -- for the $z$-test for the population mean is found exactly like the probability when we learned about probabilities for a normal distribution.
- You can use the $z$-table, or your calculator: `normalcdf(…)`

First case: upper tail

Second case: lower tail
Third case: lower AND upper tail

Since the distribution is symmetric, find one of the tails and multiply by 2:

\[ 2 \cdot \text{normalcdf}(\mid z \mid, 99999) \]

OR

\[ 2 \cdot \text{normalcdf}(-99999, -\mid z \mid) \]

**Example 1**

**Step 4. Conclusion**

- If the \( p\text{-value} < \alpha \), then the data we got is considered to be rare (or surprising) enough when \( H_0 \) is true, and we say that the data provide significant evidence against \( H_0 \), so we reject \( H_0 \) and accept \( H_a \).
- If the \( p\text{-value} > \alpha \), then our data are not considered to be surprising enough when \( H_0 \) is true, and we say that our data do not provide enough evidence to reject \( H_0 \) (or equivalently, that the data does not provide enough evidence to accept \( H_a \)).

**Comment about wording:** Another common wording is:

- "The results are statistically significant" - when \( p\text{-value} < \alpha \)
- "The results are not statistically significant" - when \( p\text{-value} > \alpha \)

**Example 1: Conclusion**

The \( p\text{-value} \) is quite large (.16), \( p\text{-value} > 0.05 \). Therefore, the results are not significant, and so we do not have enough evidence to reject \( H_0 \) and conclude that the mean SAT-M of all Ross College students is higher than the national mean (500). Note that even though the average SAT-M in our sample was 550 (which is substantially larger than 500), since this result was based on a sample of only 4 students it does not provide enough evidence to conclude that the mean SAT-M is higher than 500.
Step 4: Conclusion

- In this example the p-value is quite small (.012). In particular, for a significance level of 5%, the p-value indicate that the results are significant. (0.012 < 0.05)
- The data provide enough evidence for us to reject $H_0$ and conclude that the mean concentration level in the shipment is not the required 250.

Example 2: Conclusion

Confidence intervals to test hypotheses

Because a two-sided test is symmetrical, you can also use a confidence interval to test a two-sided hypothesis. ONLY for two-sided tests!

For a significance level $\alpha$, you can use a $1 - \alpha$ confidence interval to test the hypothesis.

If $\mu_0$ is inside the CI $\rightarrow$ do not reject the null hypothesis.
If $\mu_0$ is outside of the CI $\rightarrow$ reject the null hypothesis.

Logic of confidence interval test

Ex: Your sample gives a 99% confidence interval of

$$\bar{x} \pm m = 0.84 \pm 0.011$$

With 99% confidence, could samples be from populations with $\mu = 0.867, \mu = 0.85\mu$?

Type I and Type II errors

- The only way to be absolutely certain of whether $H_0$ is true or false is to test the entire population. Because your decision is based on a sample, you must accept the fact that your decision might be incorrect.
- You might reject a null hypothesis when it is actually true, or you might fail to reject a null hypothesis when it is actually false.
Type I and Type II errors

- The goal is to avoid these errors.
- Every cook knows how to avoid a Type I error: just remove the batteries! But this increases the incidence of Type II errors.
- Similarly, reducing the chances of Type II error, for example by making the alarm hypersensitive, can increase the number of false alarms.

<table>
<thead>
<tr>
<th>Type I and Type II errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Fire</td>
</tr>
<tr>
<td>No alarm</td>
</tr>
<tr>
<td>Alarm</td>
</tr>
</tbody>
</table>

Now think of the null hypothesis as the condition of NO FIRE, while the alternative hypothesis is that a FIRE is burning.

Accept $H_0$: No error | Type II error
Reject $H_0$: Type I error | No error

The null hypothesis is the condition of NO FIRE, while the alternative hypothesis is that a FIRE is burning.

A Type I error occurs if the null hypothesis is rejected when it is true.
A Type II error occurs if the null hypothesis is not rejected when it is false.

The probability of Type I error is the significance level, $\alpha$.

Statistical significance

A consumer advocate is interested in evaluating the claim that a new granola cereal contains “4 ounces of cashews in every bag.” The advocate recognizes that the amount of cashews will vary slightly from bag to bag, but she suspects that the mean amount of cashews per bag is less than 4 ounces. To check the claim, the advocate purchases a random sample of 40 bags of cereal and calculates a sample mean of 3.68 ounces of cashews.

The consumer advocate should declare statistical significance only if there is a small probability of

- Observing a sample mean of 3.68 oz. or less when $\mu = 4$ oz.
- Observing a sample mean of exactly 3.68 oz. when $\mu = 4$ oz.
- Observing a sample mean of exactly 3.68 oz. when $\mu = 4$ oz.
- Observing a sample mean of less than 4 oz. when $\mu = 4$ oz.

Suppose the consumer advocate computes the probability described in the previous question to be 0.0048. Her result is

- Statistically significant.
- Not statistically significant.
A consumer advocate is interested in evaluating the claim that a new granola cereal contains “4 ounces of cashews in every bag.” The advocate recognizes that the amount of cashews will vary slightly from bag to bag, but she suspects that the mean amount of cashews per bag is less than 4 ounces. To check the claim, the advocate purchases a random sample of 40 bags of cereal and calculates a sample mean of 3.68 ounces of cashews.

If the probability described in the previous question is 0.0048, then the consumer advocate should conclude that the granola is

a) Correctly labeled.
b) Incorrectly labeled.

We reject the null hypothesis whenever

a) $P$-value > $\alpha$
b) $P$-value $\leq$ $\alpha$
c) $P$-value $\neq$ $\alpha$
d) $P$-value $\neq \mu$

The significance level is denoted by

a) $\mu$
b) $\sigma$
c) $\alpha$
d) $P$-value

Which of the following is the most common choice for significance level?

a) 0 
b) 0.01 
c) 0.25 
d) 0.50 
e) 0.05

To calculate the $P$-value for a significance test, we need to use information about the

a) Sample distribution 
b) Population distribution 
c) Sampling distribution of $\bar{x}$
Conclusions

Suppose the $P$-value for a hypothesis test is 0.304.
Using $\alpha = 0.05$, what is the appropriate conclusion?

a) Reject the null hypothesis.
b) Reject the alternative hypothesis.
c) Do not reject the null hypothesis.
d) Do not reject the alternative hypothesis.

Conclusions

Suppose the $P$-value for a hypothesis test is 0.0304.
Using $\alpha = 0.05$, what is the appropriate conclusion?

a) Reject the null hypothesis.
b) Reject the alternative hypothesis.
c) Do not reject the null hypothesis.
d) Do not reject the alternative hypothesis.

P-value

True or False: The $P$-value should be calculated BEFORE choosing the significance level for the test.

a) True
b) False

Conclusions

Suppose a significance test is being conducted using a significance level of 0.10. If a student calculates a $P$-value of 1.9, the student

a) Should reject the null hypothesis.
b) Should fail to reject the null hypothesis.
c) Made a mistake in calculating the $P$-value.

Stating hypotheses

If we test $H_0$: $\mu = 40$ vs. $H_a$: $\mu < 40$, this test is

a) One-sided (left tail).
b) One-sided (right tail).
c) Two-sided.

Stating hypotheses

If we test $H_0$: $\mu = 40$ vs. $H_a$: $\mu \neq 40$, this test is

a) One-sided (left tail).
b) One-sided (right tail).
c) Two-sided.
A researcher is interested in estimating the mean yield (in bushels per acre) of a variety of corn. From her sample, she calculates the following 95% confidence interval: \((118.74, 128.86)\). Her colleague wants to test (at \(\alpha = 0.05\)) whether or not the mean yield for the population is different from 120 bushels per acre. Based on the given confidence interval, what can the colleague conclude?

a) The mean yield is different from 120 and it is statistically significant.
b) The mean yield is not different from 120 and it is statistically significant.
c) The mean yield is different from 120 and it is not statistically significant.
d) The mean yield is not different from 120 and it is not statistically significant.