M6S4 - Hypothesis Tests

Professor Jarad Niemi

STAT 226 - Iowa State University

November 1, 2018

Outline

- Hypothesis Tests
    - Review
    - Decision making
    - Practical vs Statistical Significance
    - Relationship between confidence intervals and pvalues
    - Plot your data and calculate summary statistics

Hypothesis test for a population mean μ

1. Specify the null and alternative hypothesis.
   - $H_0 : \mu = m_0$ is the default or current belief
   - $H_a : \mu > m_0$ or $\mu < m_0$ or $\mu \neq m_0$ is what you believe
2. Specify a significance level α.
3. Calculate the t-statistic.
4. Calculate the p-value.
5. Make a conclusion:
   - If p-value $< \alpha$, reject null hypothesis.
   - If p-value $\geq \alpha$, fail to reject null hypothesis.

Paired data

Definition

Two data sets and paired when each data point in one data set is related to one, and only one, data point in the other data set.

Examples:

- Record the moisturizing effect of hand lotion by using the hand lotion on only one of two hands for each study participant, but measure both hands.
- Record participant weight before and after a weight loss program.
- Assess environmental affects by studying identical twins who have grown up in different households.

Using paired data will increase your power where power is the probability of reject a null hypothesis that is not true, i.e. it is one minus the probability of a Type II error. Thus paired data will decrease the probability of a Type II error.

Water quality hypothesis test

The Ames Water Treatment Plant is considering two different processing methods for removing sediments from drinking water: active vs passive. They would like to know which method is better. They set up a pilot study where each method was implemented in parallel and observations were taken simultaneously from each method at random times. After 25 random times, they find the mean difference (active-passive) is 77 ppm with a standard deviation of 364 ppm.

1. Let μ be the true mean difference (active-passive) in sediment
2. $H_0 : \mu = 0$ versus $H_a : \mu \neq 0$
3. t-statistic is:
$$t = \frac{77 - 0}{364/\sqrt{25}} = 1.058$$
4. p-value is:
$$p\text{-value} = 2P(T_{24} > |1.058|) = 0.30$$
5. Fail to reject the null of no difference between active and passive methods based on a significance level $\alpha = 0.05$.

Water quality confidence interval

The plant manager thinks maybe a confidence interval will show a "significant" result by not including 0. So he asks a data scientist to construct a 95% confidence interval based on the sample size of 25, the sample mean of 77 ppm of the difference (active-passive), and the sample standard deviation of 364 ppm.

The data scientists finds the t-critical value:

$$t_{24,0.025} = 2.064$$

and constructs a confidence interval for the difference (active-passive)

$$77 \pm 2.064 \cdot 364/\sqrt{25} = (-73 \text{ ppm}, 227 \text{ ppm}).$$

This interval includes 0 which is consistent with no difference, but it is suggestive that the passive method is better because lower sediments is better and the interval covers more positive values than negative values.

Water quality sample size

The plant manager asks the data scientist how many samples they will need to reject the null hypothesis. The data scientists finds an online app, e.g. `https://www.stat.ubc.ca/~rollin/stats/ssize/n1.html`, and plugs in some numbers to find a sample size of $n = 176$.

Inference for a Mean: Comparing a Mean to a Known Value

(To use this page, your browser must recognize JavaScript.)

Choose which calculation you desire, enter the relevant values for mu0 (known value), mu1 (mean of the population to be sampled), and sigma (standard deviation of the sampled population) and, if calculating power, a sample size. You may also modify α (type I error rate) and the power, if relevant. After making your entries, hit the **calculate** button at the bottom.

- (●) Calculate Sample Size (for specified Power)
- ( ) Calculate Power (for specified Sample Size)

Enter a value for mu0: 0

Enter a value for mu1: 77

Enter a value for sigma: 364

- ( ) 1 Sided Test
- (●) 2 Sided Test

Enter a value for α (default is .05): .05

Enter a value for desired power (default is .80): .80

The sample size is: 176

Calculate

Water quality ample size (cont.)

The manager asks a statistician to verify this sample size. The statistician explains that with a sample size of 176 and significance level $\alpha = 0.05$ we reject if $|t| > 1.984$ since $2P(T_{100} > 1.984) = 0.05$. Assuming $X_i \stackrel{iid}{\sim} N(77, 364^2)$, we have

$$\frac{\overline{X} - 77}{364/\sqrt{176}} = \frac{\overline{X} - 0}{364/\sqrt{176}} - \frac{77}{364/\sqrt{176}} = T_{175} - 2.806 \sim t_{175}$$

and the power is

$$\begin{aligned}
&P\left(\frac{\overline{X}-0}{364/\sqrt{176}} < -1.984 \text{ or } \frac{\overline{X}-0}{364/\sqrt{176}} > 1.984\right) \\
&= \quad P\left(T_{175} < -1.984 + 2.806 \text{ or } T_{175} > 1.984 + 2.806\right) \\
&= \quad P(T_{175} < 0.822) + P(T_{175} > 4.79) \\
&\approx 1 - P(T_{175} > 0.822) + 0 \\
&\approx 1 - 0.2 \\
&= 0.8
\end{aligned}$$

Thus, the app is correct.

Water quality big data

Since samples are automated, the manager goes overboard and takes $17,600$ random samples. He doesn't even bother looking at the data or calculating summary statistics. Instead, he immediately calculates a pvalue of 0.04 and rejects the null hypothesis of no difference between active and passive and runs around the water treatment plant screaming in excitement.

Had he bothered to calculate summary statistics, he would have found a mean difference (active-passive) of 4.1 ppm with a standard deviation of 257 ppm. This results in a 95% confidence interval of

$$4.1 \pm 1.962 \cdot \frac{257}{\sqrt{17600}} = (0.3 \text{ ppm}, 7.9 \text{ ppm}).$$

Compared to the EPA limit of 500 ppm, it is likely that even an 8 ppm difference is not important.

Summary

This example demonstrated a

- Difference between practical and statistical significance
- Correspondence between confidence intervals and pvalues
- Informativeness of confidence intervals compared to pvalues

Practical versus statistical significance

Definition

A result is statistical significant if your p-value is less than your significance level. A result is practically significant if the size of the effect is meaningful.

In our example, we had two situations:

- pilot study:
  - statistically insignificant result with p-value$= 0.3 > 0.05$
  - practically significant result with estimated 77 ppm difference
- big data study:
  - statistically significant result with p-value$= 0.04 < 0.05$
  - practically insignificant result with estimated difference < 8 ppm

Correspondence between confidence intervals and pvalues

For a null hypothesis $H_0 : \mu = m_0$ and an alternative hypothesis $H_a : \mu \neq m_0$ with a p-value p:

- if $p < \alpha$ then a $100(1-\alpha)$% CI will not include m_0
- if $p \geq \alpha$ then a $100(1-\alpha)$% CI will include m_0

In our example, we had two situations:

- pilot study:
  - p-value$= 0.3 > 0.05$ and
  - 95% CI of (-73 ppm,227 ppm) included 0
- big data study:
  - p-value$= 0.04 < 0.05$ and
  - 95% CI of (0.3 ppm, 7.9 ppm) did not include 0

Reasons to ignore hypothesis tests and p-values

- Point null hypotheses, e.g $H_0 : \mu = m_0$, are never true
- A p-value and decision (reject/fail to reject) is never enough information
- When we reject, we don't know what assumption is to blame:
  - $\mu = m_0$?
  - independent and identically distributed with common variance? (random sample)
  - normal? (procedure is robust)
- A confidence interval provides an estimate with uncertainty and thus allows you to assess statistical and practical significance.