**Module 8: Statistical Hypothesis Testing**

# Setting up a Hypothesis Test[[1]](#footnote-1)

Consider a decision faced by a realty agency…

The agency manages rental properties and is considering expanding into the Denver metropolitan area. To justify the costs of opening a new office, the agency needs rents in the area to average at least $500 per month. Lower rental generate smaller fees that would make the office unprofitable. Are rents in Denver high enough to justify the cost of the move?

The general problem of statistical hypothesis testing concerns using data to make a decision. In this setting, the decision is whether a hypothesis of interest should be rejected. A statistical hypothesis is a claim about some property of a population.

Jargon:

The hypothesis to be tested is called the *null hypothesis* and is denoted *H*0.

An *alternative hypothesis*, denoted *H*a, is considered as an alternative to *H*0.[[2]](#footnote-2)

For the rental agency, these two hypotheses make different, conflicting claims about the mean rental cost *μ* in Denver:

*H*0: *μ* ≤ $500 versus *H*a: *μ* > 500

Notice that the null hypothesis H0 implies no expansion, to maintain the agency as it is. The alternative hypothesis implies change, expanding the agency into Denver.

Don’t take such hypotheses too literally. If *μ* =500.00001, the null hypothesis is false; expanding to Denver would be profitable, but not by very much. The value that separates the hypotheses ($500 in this example) is often set by some type of break-even analysis.

# The One-Sample *t* Test

Suppose that we have an iid sample *x*1, *x*2, …, *xn* from a population with unknown mean *μ*. A common set of hypotheses often considered for this setup is[[3]](#footnote-3)

*H*0: *μ* ≤ *μ*0 versus *H*a: *μ* > *μ*0 or *H*0: *μ* ≥ *μ*0 versus *H*a: *μ* < *μ*0

Note that the intent to purchase hypotheses is the special case with *μ*0 = $500.

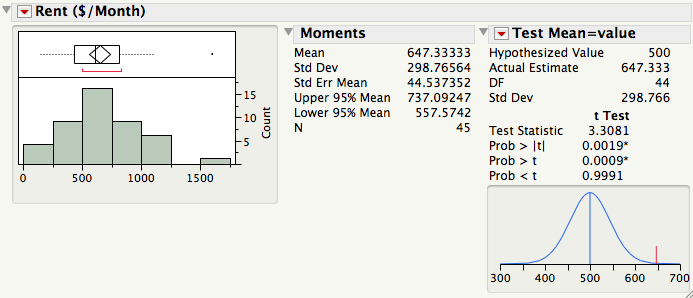
The key statistic for testing *H*0 here is



which is called a *t statistic* or a *t ratio*. The *t* statistic counts the number of standard errors between the observed statistic  and the hypothesized population parameter (*μ*0).

Intuition: a large *t* statistic in the direction of *Ha* implies that the data are implausible if the null hypothesis were true. We interpret such a large *t* statistic as evidence against *H*0.

Example: Testing[[4]](#footnote-4) the rental cost hypothesis *H*0: *μ* ≤ $500 vs. *H*a: *μ* > $500.



The estimate of the unknown average rental *μ* is the sample mean  = $647, differing from $500 in the direction specified by the alternative hypothesis, (it’s bigger than *μ*0).

In units of standard errors, how far is  = $647 from the hypothesized value *μ*0 = 500?

# The p-value: How Extreme is Enough to Reject H0?

Key issue: How large should *t* be in magnitude (positive or negative) in order to convince us to reject *H*0?

To answer this question in the previous example, JMP provides the quantity:

Prob > *t* = .0009

by which is meant that

If in fact *H*0: *μ* = $500 were true, the probability of observing a *t* more extreme than 3.31 (more positive) is .0009.[[5]](#footnote-5)

Thus, if *μ* were $500, observing *t* > 3.31 would occur only 0.09% of the time! We reject *H*0. If you don’t reject *H*0, however, it does not mean that it’s true.

The quantity Prob > t = .0009 is called a *p-value* and it measures the rarity of the data when *H*0 is true.

“Small” p-values indicate one of two things: either *H*0 is false or else something very unusual has happened. Faced with these two choices, statistical practice is to reject *H*0 when the p-value is “small” enough.

# The “Official” Rules for hypothesis Testing

Procedure to test a pair of hypotheses is:

1) Pick a value ** called the *significance level* (traditionally ** = .05 or .01).

2) If the p-value ≤ **, reject *H*0 and declare the result to be *statistically significant*(at the ** level of significance).

3) If the p-value > **, the result is said to be *not statistically significant.*

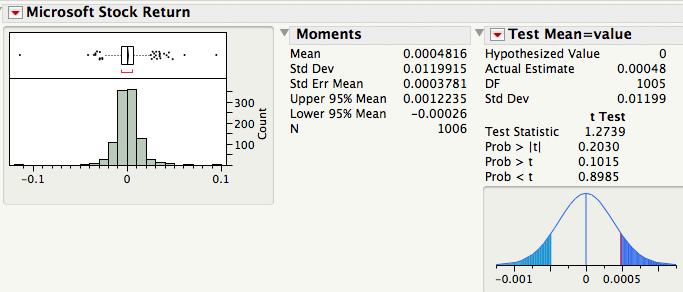
Clearly, the rental cost hypothesis *H*0: *μ* ≤ $500 should be rejected at the .05 level of significance or even at the .01 level of significance.

We rely on confidence intervals for testing two-sided hypotheses.

Example: 2004-2007 Microsoft returns. Based on the four years of data in *Microsoft\_Subset.JMP*, let’s test

*H*0: *μ* = 0 versus *H*a: *μ* ≠ 0

where *μ* is the unknown mean of the population of Microsoft daily returns. JMP yields[[6]](#footnote-6)



Can H0 be rejected at the .05 level of significance? For this, the 95% confidence interval is easier to use than a hypothesis test.

# Using Confidence Intervals to Test Hypotheses

The traditional choice of the significance level of a test is **= 5%. Also, the traditional choice of the level of confidence of a confidence interval is 95% (= 100% - 5%).

This correspondence[[7]](#footnote-7) is not accidental!

A convenient way to test *H*0: *μ* = *μ*0 versus *H*a: *μ* ≠ *μ*0 at the **=.05 level of significance uses a confidence interval:

Reject *H*0: *μ*0 = *μ*0 when *μ*0 lies outside the 95% CI for *μ*.

For example, for testing the stock returns, the 95% CI for the population proportion is   
(-0.0003, 0.0012). Since *μ*0 = 0 lies inside the confidence interval, we cannot reject *H*0.

This seems reasonable since a confidence interval is the set of plausible values for *μ* given the data.

# The "2 Standard Error" Rule of Thumb

For practical purposes, a good *approximate* test of

*H*0: *μ* = *μ*0 versus *H*a: *μ* ≠ *μ*0

is to

Reject *H*0: *μ* = *μ*0 at the =.05 level when the |t-statistic| ≥ 2

(i.e., reject *H*0 when  is more than 2 standard errors away from *μ*0).

# Summary of Testing *H*0: *μ* = *μ*0 versus *H*1: *μ* ≠ *μ*0

The following conditions are equivalent.[[8]](#footnote-8)

(a) The *p*-value is less than 0.05.

(b) The absolute value of the *t*-statistic is larger than 2 (i.e., | *t* | > 2).

(c) The 95% confidence interval does not contain *μ*0.

In any of these cases, we reject *H*0: *μ* = *μ*0.

# The Two-Sample *t* Test

Suppose we have *two* *independent* random samples[[9]](#footnote-9):

*x*1,…, *xm* from a population with unknown mean *μx*

*y*1,…, *yn* from a population with unknown mean *μ*y

and we want to test the null hypothesis that the means of the populations are the same [[10]](#footnote-10)

*H*0: *μx* = *μy* versus *H*a: *μx* ≠ *μy*

These hypotheses can also be written as

*H*0: (*μx* − *μy*) = 0 versus *H*a: (*μx* − *μy*) ≠ 0

To test this hypothesis, the test statistic of interest is the two-sample *t* statistic



counts the number of standard errors between the observed statistic  and 0. [[11]](#footnote-11)

The inference proceeds just as before, starting with a confidence interval:

Reject *H*0 if 0 lies outside the 95% confidence interval for the difference in the two means.

Alternatively, we can perform the test directly from the *p*-value, which is the probability (if *H*0 were true) of observing a larger *t* statistic (positive or negative) than the observed *t*.

If p-value ≤ **, reject H0 and declare the difference to be statistically significant (at the ** level of significance).

# Example

A car manufacturer uses the price of used cars to determine the initial cost charged to customers who lease its automobiles.

The file *UsedCars.JMP* contains the prices of 155 used BMW automobiles divisions. Some are the *xi* model (4-wheel drive) and the others are the standard *i* model.

Comparison boxplots of the prices of the two types of cars show considerable overlap.[[12]](#footnote-12)



To judge if the population means differ, it is useful to compare the 95% confidence intervals for *μi* and *μxi*.[[13]](#footnote-13)



Because these intervals *do not* overlap, we are assured that the difference between the means is significantly different from zero.[[14]](#footnote-14)

We can confirm these results (and get a more detailed comparison) with the two-sample test. A two-sample *t* test of

*H*0: *μi* = *μxi* versus *H*a: *μi* ≠ *μxi*

yields[[15]](#footnote-15)



This result is statistically significant because the 95% CI for (*μi* − *μxi*), here (619, 2780), *does not* contain 0. Thus, we can reject *H*0 at the .05 level.

Note that we can also reach this conclusion by noting that | *t* | = 3.11is larger than 2 or that the p-value = 0.0023 is less than 0.05.

# The Paired *t* Test

When comparing the population means of two samples, say *x*1,…, *xn* and *y*1,…, *yn*, it is sometimes useful to consider the paired differences *di* = (*xi* − *yi*), and treat*d*1,…, *dn* as a random sample from a population with unknown mean *μd* = (*μx* − *μy*).

When might such pairings be natural?

In this case, testing

*H*0: *μx* = *μy* versus *H*a: *μx* ≠ *μy*

is equivalent to testing

*H*0: *μd* = 0versus *H*a: *μd* ≠ 0

which can then be tested using the previous one-sample *t* test based on *d*1,…, *dn*.

# Example

Management of a newly merged pharmaceutical company needs to reduce its sales force. Does the sales force from the “GL” division differ from the sales force from the “BW” division? The sales data in *Pharmasal-split.JMP* suggest a natural pairing within each of the 20 sales districts which leads to the variable: Differences = BW – GL. The hypothesis *H*0: *μ*BW = *μ*GL is then equivalent to *H*0: *μ*d = 0. What should we conclude?[[16]](#footnote-16)



# Testing Other Statistical Hypotheses

Hypothesis testing is not restricted to statements about mean values alone.

Many other hypotheses are of interest in statistical analysis. For example, useful null hypotheses make claims about other features of populations.

*H*0: population is normal

*H*0: population correlation is 0

*H*0: two populations are identical

*H*0: time series is iid

In all of these settings, conventional statistical practice proceeds as follows:

1) Calculate a p-value for the disparity between the null hypothesis H0 and data.

2) If p-value ≤ **, reject *H*0 and declare the result to be statistically significant (at the ** level of significance).

3) If p-value > **, the result is not statistically significant.

We will describe some of these tests in more detail in Stat 621.

# What You Will Need to Know For 621

*Graphical tools* - Histogram, boxplot, comparison boxplots, and scatterplot.

*Mean, variance and correlation* - Mean is the average value. Variance is the average squared deviation from mean. Correlation measures the strength of linear association.

*Normal distribution* - 95% of the distribution lies within*μ* ± 2*σ*.Normal quantile plot as a diagnostic.

*Sampling Distributions* - Random Sampling; iid samples. Sample-to-sample variation of a statistic. Standard Error (*se*), 

*Confidence interval* - 95% CI - estimate ± 2 *se*(estimate). Interpretation.

*Hypothesis test* - t-statistic/t-ratio counts the SE’s from conjectured value.  
p-value measures “plausibility” of *H*0. Reject *H*0 at the .05 significance level ↔ p-value < 0.05 ↔ | t-statistic| > 2 ↔ hypothesized value lies outside 95% CI.

*JMP Software*

See you in September!

1. This example appears in SF, beginning on page 388. [↑](#footnote-ref-1)
2. Some textbooks denote the alternative hypothesis as *H*1 rather than *H*a. Conventions for the notation of hypothesis tests vary. [↑](#footnote-ref-2)
3. Hypotheses of this form are sometimes called “one-sided” hypotheses to distinguish them from hypotheses of the form *H*0: *μ* = *μ*0 vs. *H*a: *μ* ≠ *μ*0 which are called “two-sided.” We’ll stick to confidence intervals if we need a two-sided comparison. [↑](#footnote-ref-3)
4. For the data in *CompPur.JMP*, apply Analyze > Distribution to the column *Intend*, select Test Mean after clicking on the title bar, and enter .25 for the Hypothesized Mean to obtain this output. [↑](#footnote-ref-4)
5. For testing the hypotheses *H*0: *μ* ≥ 500 versus *H*a: *μ* < 500, JMP reports the p-value as Prob < t = .9991. For testing the hypotheses *H*0: *μ* =500 versus *H*a: *μ* ≠ 500, JMP reports the p-value as Prob < t = .0019. [↑](#footnote-ref-5)
6. Again, use the Test Mean command with Analyze > Distribution. [↑](#footnote-ref-6)
7. The correspondence described here works for “two-sided” hypotheses. See SF, section 16.4, p. 393. [↑](#footnote-ref-7)
8. This equivalence assumes that the sample size is large enough so that we can use the form ±2 SE for a 95% interval. For small samples with *t* statistics close to ±2, use the p*-*value rather than the rule of thumb. [↑](#footnote-ref-8)
9. For inferential purposes, the ideal method for getting two independent samples is from a *randomized experiment* as done in the pharmaceutical industry to show the value of a new drug. Otherwise, we must sort out the possibility of *confounding* (SF, page 426). [↑](#footnote-ref-9)
10. It would be pretty rare to find two populations with *exactly* the same mean, measured to infinite precision. As before, interpret *H*0 as saying that the means are close and offer a description of the populations that is consistent with the data. [↑](#footnote-ref-10)
11.  stands for the standard error of  which has a formula that you can safely ignore. If you are interested, it is found in precisely the same way that we found the variance of portfolios, see BBS, p 212. It is based on the fact the variance of the difference of the two means is the sum of their variances when the two means are independent, a consequence of Fact 2 in Module 5. [↑](#footnote-ref-11)
12. To horizontally perturb the points so you can see them all, as in this plot, click on the red triangle and select Display Options > Points Jittered. Recall that in JMP’s comparison boxplots the software uses quantile boxplots rather than outlier boxplots. [↑](#footnote-ref-12)
13. To get this output, use Fit Y by X. But rather than select Box Plots as you would for comparison boxplots, select Mean Diamonds. To get the output underneath, select Means and Standard Deviation from the red triangle menu. [↑](#footnote-ref-13)
14. If the intervals don’t overlap, the difference is significant. If they overlap, you need to be more careful. See SF, section 18.3. [↑](#footnote-ref-14)
15. Use the JMP Fit Y by X command, then right-click on the title bar and select t Test. If you have good reason to justify assuming the two populations have equal variances, you can get a bit more precision by selecting the Means/Anova/Pooled t. [↑](#footnote-ref-15)
16. A two-sample t test without pairing fails to find a significant difference. Use Analyze >Distribution and select Test Mean to obtain these results. JMP also offers a specific method for paired comparisons. That procedure yields the same result. [↑](#footnote-ref-16)