Two-Sample Hypothesis Tests

Note that in some of the examples previously discussed we have assumed (for simplicity) that the standard deviations of observations we made are known. In real life, this is of course rarely the case. What happens, if the true population standard deviation σ gets replaced by our best guess estimate, the sample standard deviation s? If the sample is large enough ($n \geq 40$), then nothing happens. But if the sample is small, then this substitution will change the distribution of a sample mean from a Normal distribution to a t-distribution. t-distributions have similar shapes but slightly thicker tails than Normal distributions. Their exact shape is determined by a "degree of freedom" which depends on the sample size.

We have now reviewed confidence intervals and hypothesis tests for a single population mean μ or a single population proportion p. More interesting are cases in which two means or two proportions are compared to each other.

Example:

- Can Calcium supplements lower blood pressure more than placebo treatments?
- Do women have lower science GPA's than men?

To compare two population means μ_1 and μ_2 we need data from two independent populations: Let $x_1, \ldots, x_n$ be observations from the first population which has mean μ_1 and standard deviation σ_1. Further, let $y_1, \ldots, y_m$ be observations from the second population which has mean μ_2 and standard deviation σ_2. Assume that the observations were taken independently.

Case I: If both population distributions are Normal and both standard deviations are known, then the test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \text{ Normal}(0, 1)$$

has a Normal distribution.

Case II: If both sample sizes are large ($m > 40$ and $n > 40$), then the test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \sim \text{ Normal}(0, 1)$$

has approximately a Normal distribution, even if the unknown population standard deviations are replaced by their sample estimates.

Case III: If the population distributions are Normal but the samples are not very large and the population standard deviations are not known then the test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \sim t_\nu$$

has approximately a t-distribution with df ν estimated by

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)}{\frac{(s_1^2/m)}{m-1} + \frac{(s_2^2/n)}{n-1}}$$

(round ν down to the nearest integer). In a pinch, one can also use the approximation $\nu = \min(n, m) - 1$.

When performing a two sample t-test (in the case of small samples) you can decide whether you want to assume that the population standard deviations of the two populations are the same ($\sigma_1 = \sigma_2$, homoscedastic) or different ($\sigma_1 \neq \sigma_2$, heteroscedastic). The homoscedastic assumption is only justified if the sample standard deviations are of at least comparable magnitude (the larger one should be no more than twice the smaller one).

The test statistic for the difference in population means can be used to conduct the usual 5-step hypothesis test to decide whether

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

against one of the alternatives

$$H_a : \mu_1 - \mu_2 > \Delta_0 \quad \text{or} \quad H_a : \mu_1 - \mu_2 < \Delta_0 \quad \text{or} \quad H_a : \mu_1 - \mu_2 \neq \Delta_0$$

The same information can also be used to derive a $(1 - \alpha)\%$ confidence interval for the difference in means.

Example: The firmness of a piece of fruit is an important indicator of fruit ripeness. The Magness-Taylor firmness (N) was determined for one sample of 20 golden apples after 0 days of shelf life resulting in a sample mean of 8.74 and a sample standard deviation of 0.66. Another sample of twenty of the same kind of apples was tested after 20 days of the shelf with a sample mean of 4.96 and standard deviation 0.39. Compute a 95% confidence interval for the loss in firmness after 20 days.

Note: There is a one-to-one correspondence between a confidence interval for the difference in population means and the two-sided hypothesis test for $H_0 : \mu_1 - \mu_2 = 0$.

Example: A $(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ will contain zero if and only if the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$ will be accepted at significance level α.

Analysis of Paired Data

In many cases studies yield data which cannot be assumed to be independent. For example, *before* and *after* measurements on the same patient are clearly related (since they concern the same person). Whenever an object or experimental subject is measured repeatedly, we cannot make the independence assumption necessary for the two-sample z- or t-test. In this case, we say that the (X_i, Y_i) are PAIRED data. The same question - whether the means of the *before* and *after* populations are the same or not - can still be answered, but the test statistic changes slightly.

PROPOSITION: Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be paired observations on n randomly selected individuals with $E(X_i) = \mu_1$ and $E(Y_i) = \mu_2$. Let $D_i = X_i - Y_i, i = 1, \ldots, n$ be the pairwise differences of the observations. We assume these differences to be normally distributed with mean value $\mu_D = \mu_1 - \mu_2$ and standard deviation σ_D. Then the null hypothesis

$$H_0 : \mu_D = \Delta$$

can be tested using the test statistic

$$t = \frac{\bar{d} - \Delta}{s_D / \sqrt{n}} \sim t_{n-1}$$

which has a t-distribution with $n - 1$ degrees of freedom.

Example: For each of the following experiments, decide whether the data should be analyzed as two independent samples or as paired data. Explain your reasoning. Phrase the null hypothesis and alternative hypothesis you would use for each test:

(a) Is math harder than economics? To study this question, 30 students took both a math and an economics test and their scores in both tests were compared.

(b) Are girls better in math than boys? To study this question 30 boys and 30 girls took the same math test and their scores were compared.

Example: Two identical footballs, one air-filled and the other helium filled were kicked by 39 athletes at The Ohio State University on a windless day. The distance that each football traveled was recorded:

Athlete	Helium	Air	Difference
1	25	25	
2	16	23	
3	25	18	
⋮	⋮	⋮	

Summary statistics for this experiment are $\bar{d} = 0.462$ and $s_D = 6.867$.

(a) Why is it necessary to conduct the analysis of this experiment in a paired-sample way rather than as two independent samples?

(b) Test the hypothesis that Helium filled balls travel further on average than air filled balls at level $\alpha = 0.05$

Inference For Two Sample Proportions

Recall: Previously we have discussed how to compute confidence intervals and conduct hypothesis tests for a single population proportion p. In applications it is often of interest to compare two population proportions to each other.

Example:

- Is the passing rate in Calculus classes with workshops higher than in classes without workshops?
- Is the proportion of cars with hybrid technology in California higher than in Indiana?

Suppose that we have two independent populations in which every individual can be classified either as a "success" or a "failure" (e.g., student passing or failing a Calculus class). Let

$$p_1 = \text{ proportion of successes in population 1}$$

$$p_2 = \text{ proportion of successes in population 2}$$

Suppose further that now a sample of size m is selected from the first population. Let X denote the number of successes in that sample. Than the distribution of X is:

Independently, a sample of size n is selected from population 2. Let Y denote the number of successes in the second sample.

$$Y \sim$$

Example: Consider the random variables X and Y described above. Use the distributions of X and Y to find a point estimate for $p_1 - p_2$. What is the variance of this estimate?

Test Statistic: Suppose that two samples of size m and n, respectively, are taken from two independent populations which contain proportions p_1 and p_2 of successes. Let X and Y denote the number of successes in those samples and consider

$$\hat{p}_1 = \frac{X}{m}, \qquad \hat{p}_2 = \frac{Y}{n}$$

Then the quantity

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim \text{Normal}(0,1)$$

can be used as a test statistic for the null hypothesis

$$H_0 : p_1 - p_2 = 0$$

against one of the alternatives

$$H_a : p_1 - p_2 > 0 \text{ or } H_a : p_1 - p_2 < 0 \text{ or } H_a : p_1 \neq p_2$$

Here $\hat{p}$ is the overall proportion of successes in both samples combined and $\hat{q} = 1 - \hat{p}$.

Example: A university financial aid office polled a random sample of undergraduate students about their summer employment. Not all students were employed the previous summer. Here are the results separated by gender

	Men	Women
Employed	728	603
Not Employed	89	149
Total	817	752

(a) Is there (statistical) evidence that the proportion of employed students differs for male and female student? Conduct a 5-step hypothesis tests.

(b) Derive a 95% confidence interval for the difference between the proportions of employed male and female students.

(c) Does the difference in percentages seem important to you?

Note: Not every difference (in either means or proportions) that is statistically significant is actually important. Even a tiny difference will become statistically significant if the sample sizes are very large.

Example: Suppose the average duration of a cold is 102 hours if you do not take medication but just sleep a lot and drink plenty fluids (i.e., follow your mom's advice). On the other hand, if you take medication, than the average duration of a cold may be 101.5 hours. This difference, while probably not having any practical relevance, would be considered statistically significant in an experiment where the sizes of both samples would be very large compared to the variation in each sample.

Do not confuse statistical significance with practical relevance!

Confidence Intervals and Hypothesis Tests through Software

SPSS can be used to compute confidence intervals for means and conduct hypothesis tests for a single mean or to compare two means. Confidence intervals and hypothesis tests for proportions are very easily computed by hand.

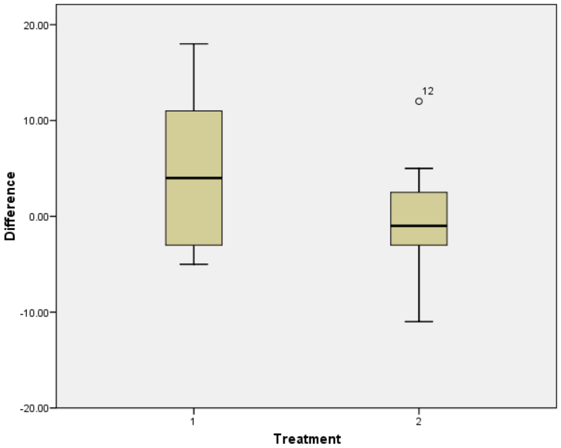
Note: Refer to the SPSS help pages on the "SPSS Help" course website for details on SPSS commands.

Example: Does calcium intake reduce blood pressure? Observational studies suggest that there is a link and that it is strongest in African-American men. Twenty-one African American men participated in an experiment to test this hypothesis. Ten of the men took a calcium supplement for 12 weeks while the remaining 11 men received a placebo. Researchers measured the blood pressure of each subject before and after the 12-week period. The experiment was double-blind.

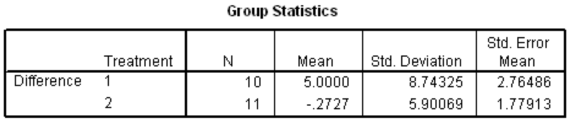
The data is available in the file "Calcium.txt" on the course website.

(a) Suggest some meaningful hypotheses that can be tested in this context. What type of hypothesis test is appropriate for each?

(b) Let's compare the drop in blood pressure (difference) between the two groups. Use SPSS to draw two (side-by-side) box plots for the differences. Your result should look like this:

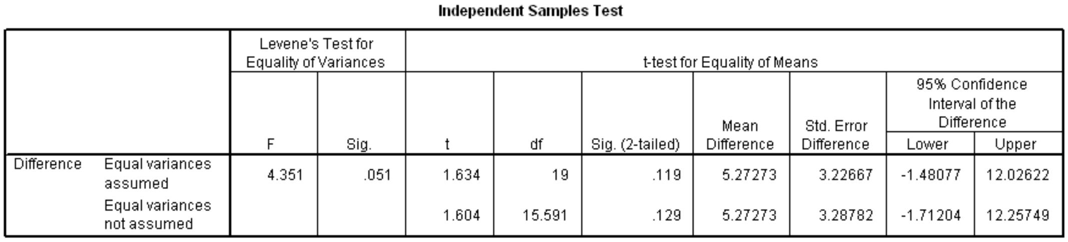


(c) Use SPSS to perform an independent sample t-test on the difference in blood pressure (before - after) for the calcium (treatment 1) and placebo (treatment 2) groups. Interpret the results:



Group Statistics

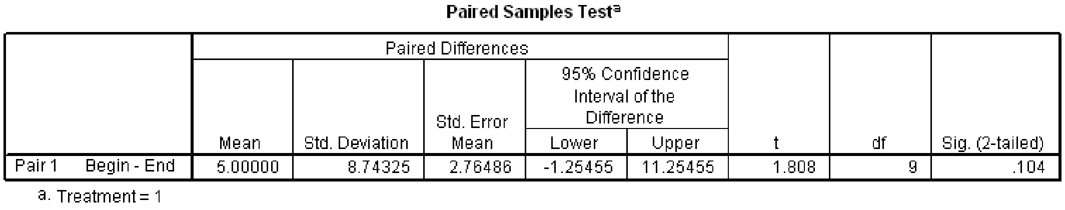
	Treatment	N	Mean	Std. Deviation	Std. Error Mean
Difference	1	10	5.0000	8.74325	2.76486
	2	11	-.2727	5.90069	1.77913



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Difference	Equal variances assumed	4.351	.051	1.634	19	.119	5.27273	3.22667	-1.48077	12.02622
	Equal variances not assumed			1.604	15.591	.129	5.27273	3.28782	-1.71204	12.25749

(d) Use SPSS to test whether Blood pressure has significantly decreased in the treatment group (men who received calcium treatment). Since the "Begin" and "End" measurements are taken on the *same* men, we cannot use an independent sample test. Instead, use a paired-samples t-test, after selecting only the calcium treatment group. The results of this analysis appear below:



Paired Samples Test[a]

	Paired Differences					t	df	Sig. (2-tailed)
				95% Confidence Interval of the Difference				
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1 Begin - End	5.00000	8.74325	2.76486	-1.25455	11.25455	1.808	9	.104

a. Treatment = 1