Two-Sample and Paired Tests

A national retailer of computer hardware and software is considering adopting a new type of advertising. Before it adopts this new format nationally, the retailer would like evidence that the new form of advertising is better than what it has been doing. The chain decided to test the effectiveness of a proposed type of advertising in the following way. It will use the new format of advertising for ads associated with a sample of 20 of its outlets while retaining the traditional form in 20 other stores. Is the new format better? Worse? The same?

The data for this example are in the JMP file **software_ads.jmp** which can be found on the class home page at www-stat.wharton.upenn.edu/~bob/stat102. The JMP file has two columns, labeled "Sales" and "Ad Type". The column "Sales" records the sales in the week following the introduction of the new type of advertisement; the "Ad Type" column has the value "New" or "Traditional". Before running any tests, what plots would you inspect and what would you look for?

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean
New	20	79.7488	22.5400	5.0401
Traditional	20	76.4067	20.7076	4.6304

To see if there is a significant difference in the sales in the two groups, we can use a test or a confidence interval. Before using these methods, we should check the underlying assumptions. The data in both groups appear normal and the SDs in the groups are similar (how would you check these?). Using the pooled variance comparison gives the following summary. The difference in the sample averages is 3.34, with the new method on average garnering about $3,340 more in weekly sales. The reported

t-Test

	Difference	t-Test	DF	Prob>|t|
Estimate	3.34	0.488	38	0.6281
Std Error	6.84			
Lower 95%	-10.51			
Upper 95%	17.20			
Assuming equal variances				

confidence interval, however, includes zero so that this difference is **not significant** with $\alpha = 0.05$: the population means could be the same.

Similarly, we can think of this as a hypothesis test. Taking the two-sided approach, we see that the t-ratio (number of SEs away from zero) is only 0.488. Such a difference occurs quite often if the two groups share a common mean. If the two groups are samples from normal populations that have common means, we'd see a difference this large about 63% of the time (the p-value). Again, there is little reason to assume that the two types of advertising produce different sales. What would be the conclusion of a one-sided comparison?

Although the SDs in the two groups are similar, we can still use the comparison that allows the standard deviations in the two populations to differ. Here's the output.

Welch Anova testing Means Equal, allowing Std's Not Equal

F Ratio	DF Num	DF Den	Prob>F
0.2384	1	37.73	0.6282

t-Test
0.4883

Again, the differences are not significant and are almost identical to what we found when we assumed the variances were the same in the populations.

While we're comparing the different analyses, we might as well also try one of the nonparametric comparisons. Below are the results from the rank-sum comparison discussed in class (a.k.a. the Wilcoxon or Kruskal-Wallis test). The p-value (0.5792) is a little bit smaller, but the conclusion is the same: no real difference is indicated between the sales associated with the two types of advertising.

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
New	20	431	21.5500	0.555
Traditional	20	389	19.4500	-0.555

2-Sample Test, Normal Approximation

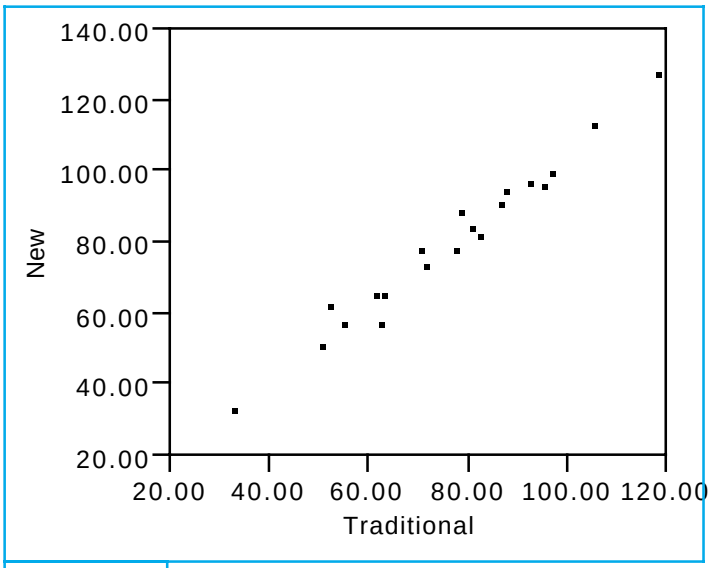
S	Z	Prob>|Z|
389	-0.55453	0.5792

1-way Test, Chi-Square Approximation

ChiSquare	DF	Prob>ChiSq
0.3227	1	0.5700

We have now used procedures that allow the populations to have unequal variance or come from non-normal populations. We have ignored, however, the first key assumption: **independence**. Rather than being two independent samples, these samples are very dependent. The retailer did a sensible thing. Instead of comparing 20 stores with new advertising to 20 other stores with the traditional advertising, the retailer **matched** the stores. Each type of advertising was used in stores with comparable baseline sales. This type of matching is likely to provide two test groups that start with common levels of sales. It avoids one group, for example, having higher sales just because that group happened to end up with a few more busy outlets.

How does one recognize a lack of independence? First and foremost, consider the context. The stores were paired, so that the first item in one sample is related to the first in the other sample. Second, we can look at a very useful plot – after we rearrange the data. The data come as 40 rows. We need to arrange them as two columns of 20 each. The JMP command "Split columns" (in the Tables menu) does just what we need. Once we have the two columns, we can plot the sales of the stores with the new advertising versus those of the matched store using traditional advertising. What would this plot look like if the two samples were independent of each other? (For fun, try out the "Nonpar density" option offered with this plot. It offers some wild plotting options.)



So, the samples are dependent. How should we make the comparison? The answer is easy. Compute the differences between the matched stores. With the column "Diff" formed as "New – Traditional", the histogram tool summary tool gives this summary.

Moments

Mean	3.342
Std Dev	4.146
Std Error Mean	0.927
Upper 95% Mean	5.282
Lower 95% Mean	1.402
N	20.000
Sum Weights	20.000

Since the differences appear close to normal, we should not hesitate to make a comparison based on averages (i.e., we don't need the nonparametric methods). The mean of "Diff" is the same mean that we observed when we treated the two samples as independent ($3,340), but there is a very big difference. The confidence interval for the population mean based on the differences is much shorter and **does not include zero**. The standard error is much smaller than the one observed when treating the two samples as independent, and the difference is significant. Why does this happen?

If we test for a mean of zero using the t-test, we get these results (which go along with our previous confidence interval). Use the "Test" button shown near the histogram of the differences. The average of the differences is over 3.6 standard errors from zero; the two-sided p-value of 0.0019 shows that it is quite rare to get such a large difference in

Test Mean=value

Hypothesized Value	0
Actual Estimate	3.34206
	t Test
Test Statistic	3.6051
Prob > |t|	0.0019
Prob > t	0.0009
Prob < t	0.9991

mean values (relative to standard error) if the true mean in the population is zero. The differences in retail sales are significant even at $\alpha = .005$.

To illustrate the corresponding nonparametric comparison, consider the signed-rank test (as described in Section 9.4 of the text). (Check the rank option in the testing dialog.) This nonparametric test (i.e., a test not assuming normality of the population of differences) also finds a very significant result, with an even smaller two-sided p-value.

Test Mean=value

Hypothesized Value	0	
Actual Estimate	3.34206	
	t Test	**Signed-Rank**
Test Statistic	3.6051	86.000
Prob > |t|	0.0019	0.001
Prob > t	0.0009	0.000
Prob < t	0.9991	1.000

Matching works magic in this problem, allowing us to detect a relatively small difference with but a few stores. Make sure you understand why the pairing helps. Here are some other things to think about.

First, might there be other differences between these stores? Has matching solved all of our problems?

Second, the difference in mean values is statistically significant, but is it meaningful? You might want to follow up to see if the two types of advertisements cost the same.

Third, pairing makes the test procedure harder to manage. You've got to find the matching items to pair and get both to participate. Lose one, and you have lost both.

Finally, pairing only "works" when the matching is effective (otherwise you drop from 40 measurements to only 20 differences with no gain in comparison.