Data analysis and nonparametric statistical tests*

1 Introduction

Prior to running the experiment the researcher should consider what type of data will result from the experiment. The alternative design for the exit experiment (the one without the real-time progress of the market) would have yielded more data for analysis. However, this data would come at a cost, the cost being that the market may have been less similar to reality.

Once the data has been collected how should it be analyzed? The analysis will certainly depend on the questions that are asked – is the focus on individual behavior, group behavior, overall market outcomes, etc.? The key with the data is to let it tell its story – if the data does not tell the story you thought it would, explain the story that the data does tell, not the one that you wished it would tell.

1.1 Qualitative analysis

Given that the data set from the experiment is likely to be unique, a first pass at data analysis should be describing the data. The easiest thing to do is make tables and charts of relevant decisions and outcomes. Again, exactly which tables and charts are made will depend on the questions asked, but becoming familiar with the data is the first step, even before running a conditional logit or performing ARIMA analysis.[1] Personally, I typically begin by making tables, but then I remember that people do not like looking at rows of numbers, so I switch to charts and graphs that are more pleasing to the eye.

If experimental work came before the econometric revolution it is possible that qualitative analysis would be all that is necessary to support a point, with a formal statistical test used to make the point stronger. However, many economists are trained as applied econometricians and *expect* to see formal statistical testing. Also, there are still errors that occur in experiments. Thus, there is a need for quantitative analysis.

1.2 Types of errors

Error can occur in an experiment through measurement error, loss of control, or sampling error. Measurement error can occur when the experiment is hand run and the researcher incorrectly records information provided by the subjects. Loss of control might occur due to social interaction (recall that social interaction may be subject-experimenter social interaction in addition to subject-subject social interaction). Sampling error might occur if the group of subjects is non-representative of the population at large.

Error can be removed by running many independent trials (recall the completely randomized experimental design in the previous set of notes). But what is an independent trial? If subjects are randomly and anonymously rematched after each round, are their decisions independent of what occurred in the previous rounds, or are they dependent? This is not an easy question for experimentalists to answer. The most stringent view of independent trial is that each session is an independent trial (or, if groups never change throughout a session, that each group is an independent trial). However, experiments of this type will generally require a large amount of funding, or that the experiment is conducted as a one-shot experiment. Again, there are tradeoffs – if the experimental task is slightly complex perhaps subjects need more exposure

*This is based on part from Friedman and Cassar, but mostly on Sheskin's Handbook of Parametric and Nonparametric Statistics, 4^{th} edition, published by Chapman and Hall/CRC in 2007.

[1] After all, aren't there certain assumptions that need to be met for those estimation techniques to be unbiased or consistent estimators?

to the task than just one round. On the other hand, if learning does play a role, then treating each round as an independent trial is not correct.

1.3 Quantitative analysis

This is where your econometric training becomes handy. I am not going to delve into the specifics of parametric statistical tests as you all should have had plenty of training in that area, but I will discuss some nonparametric statistical tests. With parametric statistical tests an assumption is made about the error distribution, usually that errors are normally distributed. This facilitates statistical testing as it allows use of some of the standard statistical tables (Z, t, and F). But the normality of the error distribution is an assumption, and the benefit of nonparametric statistical tests is that they do not make this assumption. The drawback is that if the assumption about the error distribution is true (if errors are normally distributed) then the nonparametric tests are less powerful (in a statistical sense) than parametric tests.

2 Nonparametric Statistical Tests

In this section we will discuss a half-dozen or so nonparametric statistical tests in varying detail. Much of what is here is taken from the 4^{th} edition of David Sheskin's Handbook of Parametric and Nonparametric Statistical Procedures, published by Chapman and Hall/CRC in 2007. For more in-depth coverage of the material you should consult that text or another text on nonparametric statistics. We will begin with some single sample tests and then discuss some two sample tests.

2.1 Single sample tests

These single sample tests typically discern whether or not a sample is consistent with a distribution consisting of certain parameters. Many of these tests require that the data are ordinal and can be ranked.

2.1.1 Wilcoxon Signed Ranks Test

This is a test to determine whether or not the median value of a sample equals a specified value. Let θ represent the median value. The null and alternative hypotheses are listed below. I will list 3 alternative hypotheses – hopefully you all know enough about statistical testing to know that we only specify one of the three.

$H_0 : \theta = \mu$

$H_A : \theta \neq \mu$ or $\theta > \mu$ or $\theta < \mu$

Assumptions:

1. The sample has been randomly selected from the population it represents.
2. The original scores are in interval/ratio format.
3. The underlying population distribution is symmetric.

The Wilcoxon signed ranks test ranks difference scores, which are the differences between the actual score and the hypothesized median.

Process of performing the test:

1. Take all the observations and subtract the proposed median
2. Order all NONZERO differences in ascending order by *absolute value of the difference* (thus, a difference of (-7) will receive a higher rank than a difference of 4). It is important in this test that the differences be ranked from lowest to highest. In other rank tests whether the order is low to high or high to low does not matter, but usually when the order does matter it must be from low to high.

3. Assign ranks, giving the lowest absolute difference score a 1 and the highest a value of N, where N represents the number of ranked scores (recall that any score that equals the hypothesized median will have a difference score of zero and will be unranked, so N may be less than the number of observations). For ties, give an average rank to all tied values. For example, if the 2^{nd}, 3^{rd}, and 4^{th} lowest differences are all equal, than give each score a 3 (this comes from $\frac{2+3+4}{3}$). If the 7^{th} and 8^{th} lowest differences are equal, then each gets a rank of 7.5 (from $\frac{7+8}{2}$).

4. Looking at the original differences, usually some will be positive and some will be negative. For those differences which are negative, multiply the rank of that difference by (-1). Essentially we are creating two groups of ranks, those that are above the proposed median and those that are below the proposed median.

5. Sum the ranks of the two groups independently, so that ΣR_+ is the sum of the ranks of the positive differences and ΣR_- is the sum of the ranks of the negative differences (although this sum will be negative, it is assumed that you drop the negative signs when summing the ranks – again, the importance of step 4 is to separate the difference scores into two groups). As a check, the sum of the signed ranks should equal $\frac{N(N+1)}{2}$, where N is the number of ranked observations.

6. The test statistic is the smaller of ΣR_+ and ΣR_-. To reject the null hypothesis, the test statistic must be less than or equal to the critical value. Many of these nonparametric tests have their own tables of critical values – again, the benefit of using parametric tests is that most tests use either the Z, t, or F tables.

Example:

Suppose there are 10 observations: 9, 10, 8, 4, 8, 3, 0, 10, 15, and 9. The hypothesized median is 5. This leads to difference scores of 4, 5, 3, -1, 3, -2, -5, 5, 10, and 4 respectively. The table below provides steps 3 and 4. The first row is the observation #, the second row is the actual observation, the third row is the difference, the fourth row is the absolute value of the difference, the fifth row is the rank, and the sixth row is the signed rank. Note that the observations are already in ascending order.

Obs #	4	6	3	5	1	10	2	7	8	9
Observation	4	3	8	8	9	9	10	0	10	15
Difference	-1	-2	3	3	4	4	5	-5	5	10
$\|difference\|$	1	2	3	3	4	4	5	5	5	10
Rank	1	2	3.5	3.5	5.5	5.5	8	8	8	10
Signed Rank	-1	-2	3.5	3.5	5.5	5.5	8	-8	8	10

Now, the two sums are $\Sigma R_+ = 44$ and $\Sigma R_- = 11$. Note that $N = 10$, so $\frac{N(N+1)}{2} = 55$. The test statistic is 11, since $11 < 44$. Some critical values for this test for standard significance levels are provided below:

	$\alpha = .05$	$\alpha = .01$
Two-tailed	8	3
One-tailed	10	5

Since 11 is greater than all the values in the table, we *cannot* reject the null hypothesis for either the one-tailed test or the two-tailed test at the 5% level.

2.1.2 Kolmogorov-Smirnov goodness of fit test for a single sample

This is a test to determine if a distribution of sample observations conforms to a specified probability distribution. The distribution may be a theoretical one (like the uniform or normal) or may be one derived from previous empirical observation.

$H_0 : F(X) = F_0(X)$

$H_A : F(X) \neq F_0(X)$

We will not go through all of the procedures for this test. The term $F_0(X)$ is the specified distribution. With this test, the researcher typically wants to fail to reject the null hypothesis, as the researcher would like to conclude that the data is distributed in some particular manner.

Assumptions:

1. KS test used for continuous variables
2. Data must be ordinal as a cumulative frequency distribution must be constructed

Process of performing the test:
In practice it is best to use a statistical package for this test, but here is an idea of how the test proceeds. Construct both the specified cumulative frequency distribution and the sample cumulative frequency distribution. Find the point at which the two distributions are the greatest distance apart. This distance is your test-statistic, and if it is too large then the null hypothesis is rejected.

2.1.3 Chi-square goodness of fit test for a single sample

The chi-square goodness of fit test is a test to determine if observed cell frequencies differ from expected frequencies. It is used with categorical data, meaning this test can be used to answer questions such as whether or not a die is fair.

H_0 : Observed cell frequencies are equal to expected cell frequencies for ALL cells
H_A : Observed frequency is not equal to expected frequency for at least one cell
Thus, if only one cell frequency is not equal to its expected frequency then the null hypothesis is rejected.
Assumptions:

1. Data is categorical/nominal
2. Random sample of n independent observations
3. Expected frequency of each cell is at least 5

Process of performing the test:

1. Specify the number of cells (k) and categorize the sample data into cells
2. Find the difference between the actual frequency (O_i) and the expected frequency (E_i) for each cell $(O_i - E_i)$
3. For each cell, square the difference found in step 2 $(O_i - E_i)^2$
4. For each cell, divide the squared difference by the expected cell frequency (steps 2-4 are essentially squaring a variable that has been converted into a standard normal) $\frac{(O_i - E_i)^2}{E_i}$
5. Sum the results $\sum_{i=1}^{k} \left(\frac{(O_i - E_i)^2}{E_i} \right)$
6. The test statistic is the sum. It is distributed as chi-square with $k - 1$ degrees of freedom If the test statistic is greater than the appropriate critical value then we reject the null hypothesis.

Example:
Suppose a die has been rolled 120 times. We observe a sample of 20, 14, 18, 17, 22, and 29 for a roll of 1, 2, 3, 4, 5, and 6 respectively. The table below contains the operations in steps 1-4. Note that $E_i = 20$ for each cell because we want to see if the die is fair.

Die roll	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	20	20	0	0	0
2	14	20	−6	36	1.8
3	18	20	−2	4	.2
4	17	20	−3	9	.45
5	22	20	2	4	.2
6	29	20	9	81	4.05
	$\sum O_i = 120$	$\sum E_i = 120$			$\sum \frac{(O_i - E_i)^2}{E_i} = 6.7$

Thus, the test statistic is 6.7. Since there are 6 categories there are 5 degrees of freedom. The critical value for a chi-square with 5 degrees of freedom for $\alpha = .05$ is 11.07 and for $\alpha = .01$ is 15.09. Thus, we fail to reject the null hypothesis since $6.7 < 11.07$. Even when $\alpha = .10$ we fail to reject, as the critical value when $\alpha = .10$ is 9.24.

2.1.4 Binomial sign test for a single sample

This test is used when the data can be categorized into 2 groups. The test determines if the proportion of observations in one group equals a specific value.

$H_0 : \pi_1 = \mu$

$H_A : \pi_1 \neq \mu$

In the hypotheses, π_1 refers to the proportion of observations in group 1 and μ refers to the specified value. If the exercise were to determine whether or not a coin was fair, then $\mu = .5$.

Assumptions:

1. Each of n independent observations is randomly selected from a population
2. Each observation can be classified into 1 of 2 mutually exclusive groups

Process of performing the test:

1. Compute the probability that exactly x out of n observations will fall into category 1, where x is the number of observations in category 1. This is done as

$$P(x) = \binom{n}{x} (\pi_1)^x (1 - \pi_1)^{n-x}$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

2. We actually need to know what the probability of the exact outcome or more extreme outcomes are. Thus, we need to determine $P(z)$ for all $z \in [x, n]$.
3. Sum the probabilities found in steps 1 and 2. This sum is the test statistic. If the test statistic is less than or equal to $\frac{\alpha}{2}$, where α is the level of significance, then we reject the null hypothesis.

Example:

Suppose a coin is flipped 10 times and the result is 8 heads and 2 tails. Is this coin fair? We have $P(8) = .0439$, $P(9) = .0098$, and $P(10) = .001$, so $P(8, 9, 10) = .0547$. We fail to reject the null hypothesis at $\alpha = .05$ since $.0547 > .025$. We can then conclude that the coin is fair. Note that if we had more observations where 80% of the observations were heads we would not be able to reach this hypothesis. For example, if we had 20 coin flips and 16 heads, then $P(16, 17, 18, 19, 20) = .005909$, which is significant at the 5% level.

2.1.5 Single sample runs test

This test can be used to determine if the distribution of a series of *binary* events in a population is random. Most of the other tests have considered whether or not the sample data are consistent with a particular distribution. The single sample runs test aims to determine if there is a bias in the distribution of the events.

H_0 : The events in the underlying population represented by the sample are distributed randomly.

H_A : The events are distributed nonrandomly.

In order to conduct this test we need to know the number of runs in the series of observations. A run is a sequence within a series in which one of the 2 alternatives occurs on consecutive trials.

Process of performing the test:

Small sample – Less than 20 observations in both groups:

1. Calculate the number of runs in the series. This is the test statistic.
2. Find the range of critical values for n_1 and n_2, where n_1 is the number of observations in group 1 and n_2 is the number of observations in group 2. If the number of runs is within this range, then we fail to reject the null hypothesis.

Large sample:

1. Calculate the z-statistic for the sample using the formula below, where r is the number of runs

$$z = \frac{r - \left[\frac{2n_1n_2}{n_1+n_2} + 1\right]}{\sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}}$$

2. Then use the z-table. If $\alpha = .05$, then the critical value for z will be 1.96, and we will reject the null hypothesis if $z > 1.96$. If $\alpha = .01$, then we will reject the null hypothesis if $z > 2.57$.

Examples:
Consider 3 series of coin tosses from different coins:
Series 1: $H, T, H, T, H, T, H, T, H, T, H, T, H, T, H, T, H, T, H, T$
Series 2: $H, H, H, H, H, H, H, H, H, H, T, T, T, T, T, T, T, T, T, T$
Series 3: $H, T, T, T, H, H, T, H, T, H, H, T, T, T, H, T, H, T, H, H$

Note that in all 3 series there are 10 heads and 10 tails, so we can use the same range of critical values. In Series 1 there are 20 runs. In Series 2 there are 2 runs. In Series 3 there are 13 runs. The range of runs for 10 of each group is $[6, 16]$. Thus, for Series 1 and 2 we reject the null hypothesis and for Series 3 we fail to reject the null hypothesis.

Although 10 of each group is not a large sample, we will use Series 3 to calculate the z-statistic. For series 3, $z = .91$. Since this $.91 < .96$ we fail to reject the null hypothesis. For Series 1 and Series 2 the |z-statistic| is 4.12, and we still reject the null hypothesis.

2.2 Two sample tests

We will now discuss some two sample tests. These are fairly similar to one sample tests, and many of the names look the same.

2.2.1 Mann-Whitney U test

This test can be used to determine if two independent samples represent two populations with different median values. The test is used with two independent samples, and the data must be ordinal in data as the data will need to be ranked.

$H_0 : \theta_1 = \theta_2$ (the two medians are equal)
$H_A : \theta_1 \neq \theta_2$
Process of performing the test:

1. Arrange scores in order from lowest to highest

2. Assign ranks, giving the lowest score a 1 and the highest score N, where N is the number of scores. Adjust for ties by taking the average rank of all tied scores as in the Wilcoxon signed ranks test

3. Sum the ranks of both groups

4. Compute the U-value for both groups as follows:

$$\begin{aligned} U_1 &= n_1n_2 + \frac{n_1(n_1+1)}{2} - \Sigma R_1 \\ U_2 &= n_1n_2 + \frac{n_2(n_2+1)}{2} - \Sigma R_2 \end{aligned}$$

   where n_1 is the number of observations in group 1, n_2 is the number of observations in group 2, ΣR_1 is the sum of the ranks of group 1, and ΣR_2 is the sum of the ranks of group 2

5. The smaller of U_1 and U_2 is the test statistic. Compare the test statistic with the critical values for this test and if the test statistic is less than or equal to the critical value we reject the null hypothesis.

As two methods of checking your method, note that $U_1 \geq 0$ and $U_2 \geq 0$ and that $n_1 n_2 = U_1 + U_2$.

Example:

There are 10 participants in one treatment and 8 in another treatment. The earnings of the participants in the treatments are provided below:

Treatment 1: $1, 2, 10, 12, 14, 15, 15, 25, 32, 72$

Treatment 2: $23, 96, 109, 110, 116, 117, 117, 119$

We have $n_1 = 10$, $n_2 = 8$, $\Sigma R_1 = 58$ (the sum of the numbers from 1 to 7 is 28, and 7 observations in group 1 are the lowest among both groups, followed by 1 observation from group 2 in 8^{th} place, and then 3 more observations for group 1 in 9^{th}, 10^{th}, and 11^{th} places), and $\Sigma R_2 = 113$. This gives $U_1 = 77$ and $U_2 = 3$. The critical values for a two-tailed test for $\alpha = .05$ and $\alpha = .01$ are 17 and 11 respectively. For a one-tailed test with $\alpha = .01$, the critical value is 13.

There is also a normal approximation for large samples for this test. Compute the z-statistic as:

$$z = \frac{U - n_1 n_2}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

where U is the smaller of U_1 and U_2 (computed as above). Take the absolute value of z and compare it with the critical value from the z-table. If $|z|$ >critical value, reject the null.

2.2.2 Kolmogorov-Smirnov test for 2 independent samples

This test is similar to the 1 sample KS-test except that now we are testing to determine if two independent samples represent 2 different populations. This test compares the 2 observed cumulative frequency distributions to one another, rather than an observed cumulative frequency distribution to a pre-specified one.

$H_0 : F_1(X) = F_2(X)$

$H_A : F_1(X) \neq F_2(X)$

Again, we will not discuss the method in detail. Compare the cumulative frequency distributions from the samples and find the point of greatest difference between the two. That is your test statistic.

2.2.3 Wilcoxon matched-pairs signed-ranks test

Recall the definition of a matched-pairs experimental design. In this design type, one individual makes a decision for two treatments, holding all else constant. Thus, each observation is dependent – in the exit experiment, if a participant saw one cost for both the fixed cost and opportunity cost treatments and had to make an exit decision for both, that would be a matched-pairs design. The two samples would be dependent because the decisions are made by the same individual. The Wilcoxon matched-pairs signed-ranks test can be used to determine if two dependent samples represent two different populations.

$H_0 : \theta_D = 0$

$H_A : \theta_D \neq 0$

The null hypothesis is that the median of the difference scores equals zero. Thus, the Wilcoxon matched-pairs signed-ranks test is a rank test based on *difference scores*. This is similar to the one sample Wilcoxon test we have already seen.

Assumptions:

1. Sample of n subjects has been randomly selected from the population it represents
2. Data is ordinal
3. Distribution of the difference scores in the populations represented by the two samples is symmetric about the median of the population of difference scores

Process of performing the test:

1. Find the difference for each matched-pair, D_i

2. Take the absolute value of each difference, $|D_i|$

3. Order the absolute value of the differences from lowest to highest (the order is important in this test)

4. Assign ranks to each NONZERO $|D_i|$, giving the smallest nonzero $|D_i|$ a rank of 1 and the highest nonzero $|D_i|$ a rank of k, where k is the number of RANKED $|D_i|$. As in the single sample test, any difference of zero is not ranked. For any ties, give each observation the average of the tied ranks.

5. Apply the sign of the difference (from step 1) to each of the ranks in step 4. Again, this process is simply to separate the differences into two groups.

6. Find ΣR_+ and ΣR_-, where ΣR_+ is the sum of the ranks of the observations with positive ranks and ΣR_- is the absolute value of the sum of the ranks with negative ranks

7. The smaller of ΣR_+ and ΣR_- is the test statistic. Compare the the test statistic with the critical value from the table and if the test-statistic is less than the critical value reject the null hypothesis.

Example:

Consider 10 matched-pair observations in the table below. The steps of the test are performed in the table as well.

Obs. #	Treat. 1	Treat. 2	D	Rank of $\|D\|$	Signed rank of $\|D\|$
1	9	8	1	2	2
2	2	2	0	–	–
3	1	3	−2	4.5	−4.5
4	4	2	2	4.5	4.5
5	6	3	3	7	7
6	4	0	4	9	9
7	7	4	3	7	7
8	8	5	3	7	7
9	5	4	1	2	2
10	1	0	1	2	2

From this table we see that $\Sigma R_+ = 40.5$ and $\Sigma R_- = 4.5$. Thus, the test-statistic is 4.5. The table of critical values for $k = 9$ is:

	$\alpha = .05$	$\alpha = .01$
2-tailed	5	1
1-tailed	8	3

Note that we reject the null hypothesis at the 5% level for both the 2-tailed and 1-tailed tests, but not at the 1% level (although only one difference is negative, there are only 9 ranked observations). For the one-tailed test, one would conclude that $\theta_D > 0$ because $\Sigma R_+ > \Sigma R_-$. The one-tailed test where $\theta_D < 0$ is not supported because $\Sigma R_+ > \Sigma R_-$.