Non-Parametric Methods

References: Any introductory statistics text will have a reasonable chapter on non-parametric methods. If you need a suggestion:

> "Fundamentals of Biostatistics". B Rosner. 2000. Duxbury Press

Background

P-Values: Statistical tests are interpreted through p-values. A p-value is a probability which tells us, roughly, the chance that our experimental results are consistent with a pre-specified (null) hypothesis.

Usual Null Hypothesis: There is no effect

Usual Interpretation:

If $p < .05$, then our experimental results are inconsistent with the null hypothesis. Therefore, we reject the null hypothesis and conclude that there has truly been an effect.

→ Positive finding

If $p > .05$, then our experimental results are not inconsistent with the null hypothesis. Therefore, we do not reject the null hypothesis. De facto, we are left to believe that there was no effect.

→ Negative finding

Problems

In drawing conclusions from statistical tests (i.e., there is an effect; or, there is not an effect), we can make mistakes:

> False Positives: We conclude that there is an effect, when, in truth, there is not.
> i.e., For our study, $p<.05$, but the finding can not be reproduced.
>
> False Negatives: We conclude that there is no effect, when, in truth, there is.
> i.e., For our study, $p>.05$, but subsequent studies all show the effect to be significant.

Convention 1: False positives are more important than false negatives.

Convention 2: The false positive rate should be stated clearly in any manuscript (usually set to be 5%), and should be accurate.

Convention 3: While we want the false negative rate to be as small as possible, we acknowledge that it will vary from study to study, depending on sample size and the magnitude of the experimental effect.

Statistical Testing

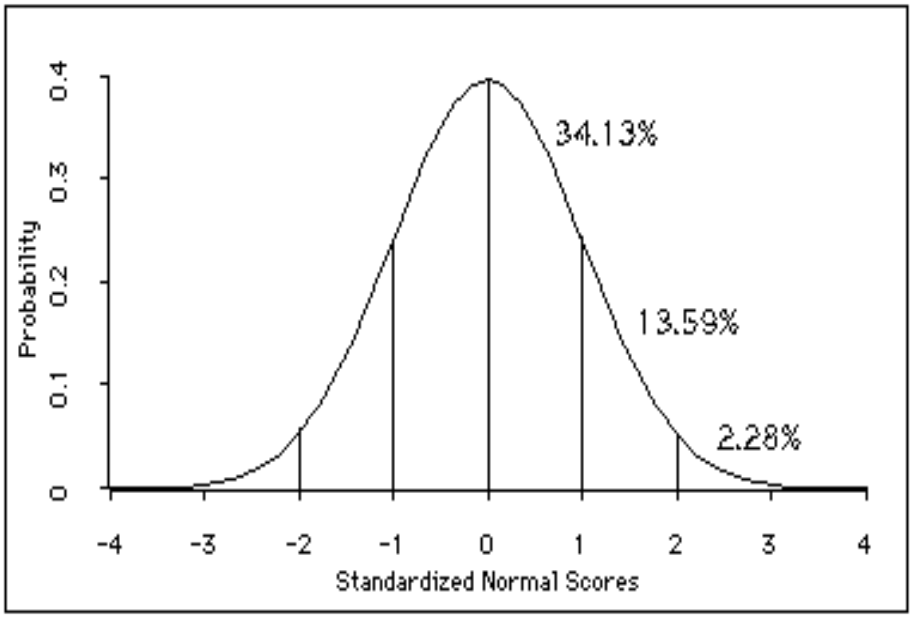
First Priority: Have an accurate assessment of the false positive rate.

Second Priority: Keep the false negative rate as low as possible. (The Power of a study equals one minus the False Negative Rate.)

How Do We Accomplish Both of These?

By understanding the behavior of our data.

For example, if we are interested in changes in Body Mass Index, and "know" that such changes follow a Normal distribution with mean 0 and standard deviation 1, then the following graph lets us interpret the data:



This graph can be used to interpret individual BMI changes, or group-to-group changes (since averages of Normal data are also Normal).

What "magic" allowed us to accomplish so much:

1. We assumed that the changes in BMI were from a distribution centered at 0.
   → Not really an assumption. This is really a statement of the null hypothesis.

2. We assumed that the changes in BMI were from a distribution with standard deviation 1.
   → It's easy to avoid this assumption. Simply use the collected data to estimate the standard deviation. The graph will change slightly but in an exactly computable way.

3. We assumed that the changes in BMI were from a Normal distribution. This means that the curve in the graph followed the formula:

$$f(x) = (1/\sqrt{2\pi\sigma}) \exp\{ - \tfrac{1}{2} [(x - \mu) / \sigma]^2 \}$$

which defines a Normal(μ,σ^2) distribution.

→ This is an example of a **parametric** distribution since the behavior of the data is dictated by two parameters: μ and σ^2

1. Parametric Testing

1. Look at the distribution of your outcome measure.

2. Assume that your data follows a particular distributional form.

3. Carry out the parametric test that is "optimal" for that distribution.

Optimal: The false positive rate is exactly accurate (i.e., if the test yields p=.02, then there is exactly a 2% chance of a false positive),
and the false positive rate is as low as it can possibly be (i.e., the power or the test is maximal).

Most Common Distributional Assumption: Normal

Other Possibilities:
a. Count data: Poisson
b. Right-skewed data: Exponential; Gamma; Lognormal; Weibull
c. Left-skewed data: Weibull
…

Normal Based Testing

Fact: Although there are many parametric models, many people mistakenly assume that Normal-based tests and parametric tests are synonymous.

Why?

1. Empirically, many distributions look at least roughly Normal
2. The Central Limit Theorem justifies Normal-based tests when sample sizes are large enough

And therefore, in practical terms

3. Tests and calculations were developed for the Normal situation
4. Teaching focuses on Normal-based tests
5. Software packages focus on Normal-based tests

More importantly, in scientific terms

6. The results of Normal-based tests are clinically interpretable
7. The Normal-based tests can be paired immediately with confidence intervals
8. The Normal-based tests can be paired with power calculations

Judging Normality

General: The Normal distribution is characterized by its symmetry and the scarcity (though not total lack) of outliers.

Graphic: Histogram - the histogram should look "bell-shaped" (as in my previous slide)

Symmetry: Skewness – a measure ranging from $-\infty$ to $+\infty$, where 0 means symmetric; positive numbers indicate a right skew (i.e., a long right tail to the distribution); and negative numbers indicate a left skew (i.e., a long left tail to the distribution). As an alternative, some people like to compare mean to median.

Outliers: Kurtosis – a measure ranging from –1 to $+\infty$, where 0 means Normal; positive values indicate too many outliers (relative to a Normal distribution); and negative values indicate too few outliers (relative to Normal).

Test: Shapiro-Wilk Test – tests whether the distribution is significantly different from Normal. P-values less than .05 indicate non-Normality.

Examples of Normal-Based Analyses

<u>One-Sample (Paired Sample):</u> Assume each patient provides a
baseline measurement of LDL, X_i
and a follow-up measurement of LDL, Y_i
Our interest is in whether there has been any significant change in LDL: $\Delta_i = Y_i - X_i$

Process: a. Calculate the average of the Δ's
b. Calculate the standard error of the Δ's
c. Use the ratio of the average over the standard error to test the null hypothesis that $\Delta=0$

<u>Two-Sample:</u> Assume each patient in the treatment group provides
a measurement of LDL, X_i
and each patient in the control group provides
a measurement of LDL, Y_i
Our interest is in whether there is any significant difference in LDL between the two groups

Process: a. Calculate the average and standard error of the X's
b. Calculate the average and standard error of the Y's
c. Calculate the difference between the two averages
d. Use the ratio of the difference over the standard errors to test the null hypothesis that the outcomes in the two groups are identical

Multiple Groups (Anova): Assume we want to compare outcomes between a control group, a low dose group, and a high dose group

Process:

a. Calculate the average outcome in the control group
b. Calculate the average outcome in the low dose group
c. Calculate the average outcome in the high dose group
d. Compare the three average to each other, relative to the standard errors

Continuous Predictor of a Continuous Outcome: Pearson Correlation Coefficient

Multiple Predictors of a Continuous Outcome: Linear Regression

Multiple Predictors of Correlated Continuous Outcomes: Repeated Measures Linear Regression

Etc.

2. Normal-Based Testing, Justified by the Central Limit Theorem

Central Limit Theorem: The false positive rate of a test will be valid as long as the test is based on averages (interpreted loosely, including weighted averages) **and as long as the sample size is large enough.**

Notes:

1. Even if our outcome data are not Normally distributed, the Central Limit Theorem assures the validity of the false positive rate.
2. Normal-based tests all utilize averages and so the Central Limit Theorem is applicable.
3. How large the sample size has to be is vague. Generally, the more non-Normal the data are, the larger the sample size has to be.
4. The Central Limit Theorem says nothing about power:
   → Apply a Normal test to Normal data: valid false positive rate and optimal false negative rate
   → Apply a Normal test to non-Normal data: valid false positive rate, if the sample size is large enough, but ?? false negative rate

Bottom Line: If you feel that your sample size is large enough to apply the Central Limit Theorem, and you are willing to risk a loss of power, then you can use all of the familiar Normal-based tests, confidence intervals and power calculations.

3. Normal-Based Testing, After Transformation

Process: Look at your outcome measure, Y_i , and choose a mathematical transformation which makes the distribution of the Y_i appear more Normal (i.e., $Z_i = \log(Y_i)$). Now analyze the Z_i using standard Normal-based tests.

Note: It may or may not be possible to reverse the transformation and re-gain interpretability of effect estimates and confidence intervals. But, as long as the Z_i are (roughly) Normally distributed, you will have:

a (roughly) valid false positive rate

and

a (roughly) optimal false negative rate

4. Non-Parametric Methods

If none of the previous approaches can be justified, or if they don't appeal to you, then you can always use a non-parametric test.

Premise: We make no assumption about the data following any particular distribution. We work only from the collected data.

Implication: Without assumptions about the distribution, we lose the ability to interpret distances between observations.

i.e., If we assumed that the data were from a Normal distribution, the two observations that are 2 standard deviations apart are very different from each other. Similarly, two means that are 2 standard errors apart are very different from each other.

However, with no assumptions about the distribution, two observations that are 2 standard deviations apart could be unremarkable, if the distribution has many outliers.

Consequence: We need to recover some way to interpret differences between observations.

→ Dozens of suggestions have been made

Common Element: All of the non-parametric tests guarantee the validity of the false positive rate, regardless of the sample size and regardless of the underlying distribution of the data (i.e., non-parametric tests can be used for Normal data and the false positive rate will still be valid)

Differences: The power of a non-parametric test will vary according to the distribution of the underlying data (and it can be higher or lower than the power of a parametric test)

Most Common Approach: Non-parametric tests based on ranks

Take the original data: $Y_1, Y_2, Y_3, Y_4, \ldots, Y_n$

And substitute their ranks, from smallest (1) to largest (n)

Then, analyze the ranks.

What have we won? The data ($Y_1, Y_2, Y_3, Y_4, \ldots, Y_n$) whose distribution we could not interpret, have now been converted to ranks (1, 2, 3, 4, … , n). The distances between these new "data points" and their variability are immediately obvious.

Non-Parametric Methods Based on Ranks

One-Sample (Paired Sample): Wilcoxon Signed Rank Test

Assume each patient provides a
baseline measurement of LDL, X_i
and a follow-up measurement of LDL, Y_i

Our interest is in whether there has been any significant change in LDL: $\Delta_i = Y_i - X_i$

Process: a. Assign a rank to each of the Δ's according to its absolute distance from 0 (i.e., the closest point to 0 gets rank 1; the farthest point from 0 gets rank n).

b. Calculate the sum of the ranks of the Δ's which were positive.

c. Calculate the sum of the ranks of the Δ's which were negative.

d. Calculate the difference between these two sums (relative to an appropriate standard error). If the difference is small, then the data are consistent with the null hypothesis that the median $\Delta=0$ and there is no reason to reject the null hypothesis. If the difference is large, then reject the null hypothesis.

Note: We use information about which values of Δ_i are bigger and which are smaller, and which values of Δ_i are positive and which are negative. However, the values of Δ_i are not used in the calculations.

Warning: The test assumes that the Δ's are from a symmetric distribution.

Two-Sample: Wilcoxon Rank Sum Test

Assume each patient in the treatment group provides
a measurement of LDL, X_i

and each patient in the control group provides
a measurement of LDL, Y_i

Our interest is in whether there is any significant difference in LDL between the two groups

Process:

a. Combine the X's and the Y's and rank the combined data from smallest (rank 1) to largest (rank n+m).

b. Calculate the average rank of the X's.

b. Calculate the average rank of the Y's.

c. Calculate the difference between the two average ranks.

d. Use the ratio of the difference of the average ranks over an appropriate standard error to test the null hypothesis that the outcomes in the two groups are identical

Notes:

a. The Wilcoxon test does not compare medians
b. Showing the actual estimates that are being compared (i.e., the average ranks), does not help clinical readers
c. A confidence interval around the difference in average ranks is not useful either

Multiple Groups: Kruskal-Wallis Test

Assume we want to compare outcomes between a control group, a low dose group, and a high dose group

Process:

a. Combine the data from all 3 groups and rank the combined data from smallest (rank=1) to largest (rank=n+m+j).

b. Calculate the average rank in each group.

c. Compare the average ranks across the groups.

Continuous Predictor of a Continuous Outcome: Spearman Correlation Coefficient

Process:

Rank the predictor from smallest to largest. Rank the outcome from smallest to largest. Calculate a Pearson correlation between the two sets of ranks.

Regression: Rank the outcomes and perform a linear regression on the ranks. → Not commonly pursued because of lack of interpretability.

An Alternative Non-Parametric Test

One-Sample (Paired Sample): Sign Test

Assume each patient provides a
baseline measurement of LDL, X_i
and a follow-up measurement of LDL, Y_i

Our interest is in whether there has been any significant change in LDL: $\Delta_i = Y_i - X_i$

Process: a. Count the number of Δ's that are positive.

b. Count the number of Δ's that are negative.

c. If the null hypothesis that the median Δ is 0 is true, then about half of the Δ's should be positive and half should be negative. (i.e., 50% chance of being positive; 50% chance of being negative). Test whether the split is close to 50/50.

Trade-Offs

	Normal-Based		Non-Parametric			
			Wilcoxon Test		Sign Test	
	False+	False-	False+	False-	False+	False-
1. Normal Data	Valid	Optimal	Valid	Good	Valid	Okay
2. Large Sample Size, Non-Normal	Valid	??	Valid*	??	Valid	??
3. Not Large Sample Size, Non-Normal	Not Valid	??	Valid*	??	Valid	??
4. Large Sample Size, Slightly non-Normal	Valid	Good	Valid*	Okay	Valid	Not Good
5. Large Sample Size, Moderately non-Normal	Valid	Okay	Valid*	Good	Valid	Okay
6. Large Sample Size, Very non-Normal	Valid	Not Good	Valid*	Okay	Valid	Good
7. Not large Sample Size, Very non-Normal	Not Valid	Not Good	Valid*	Okay	Valid	Good

* Assuming a symmetric distribution; Otherwise, not valid.

Simulation Results

1. Uniform Distribution: symmetric but non-Normal because of lack of outliers
   Example: Skewness = .28; Kurtosis = -1

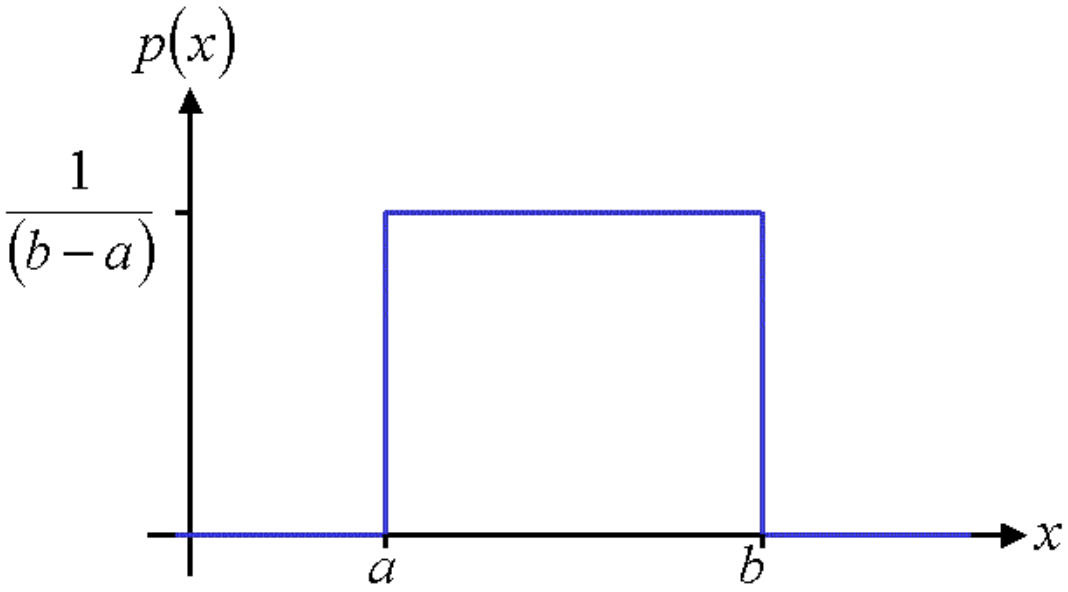
2. Normal Distribution:
   Example: Skewness = -.04; Kurtosis = -.17

3. Exponential Distribution: moderately right-skewed with too many outliers in the right tail
   Example: Skewness = 1.9; Kurtosis = 5.5

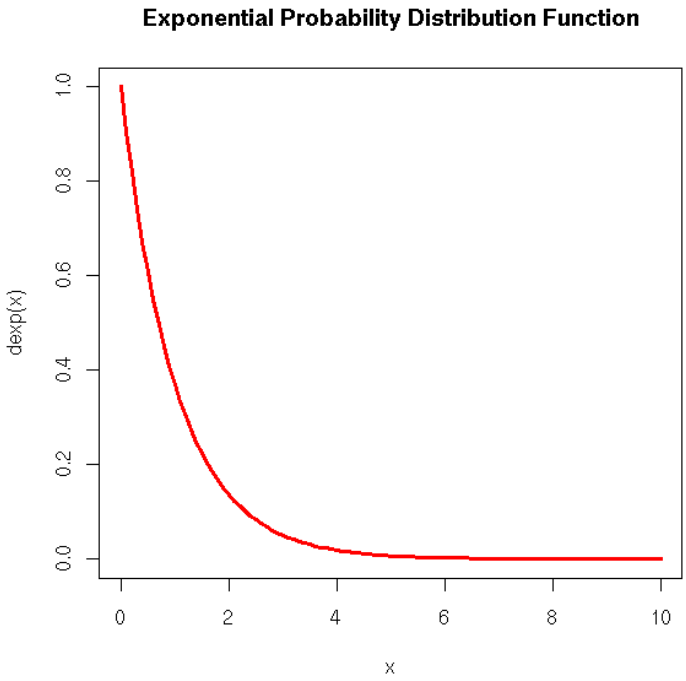
4. Cauchy Distribution: pathological distribution with too many outliers in both tails
   Example: Skewness = 1.7; Kurtosis = 35.4

Sample size for each simulation: n=100 (1-sample test)

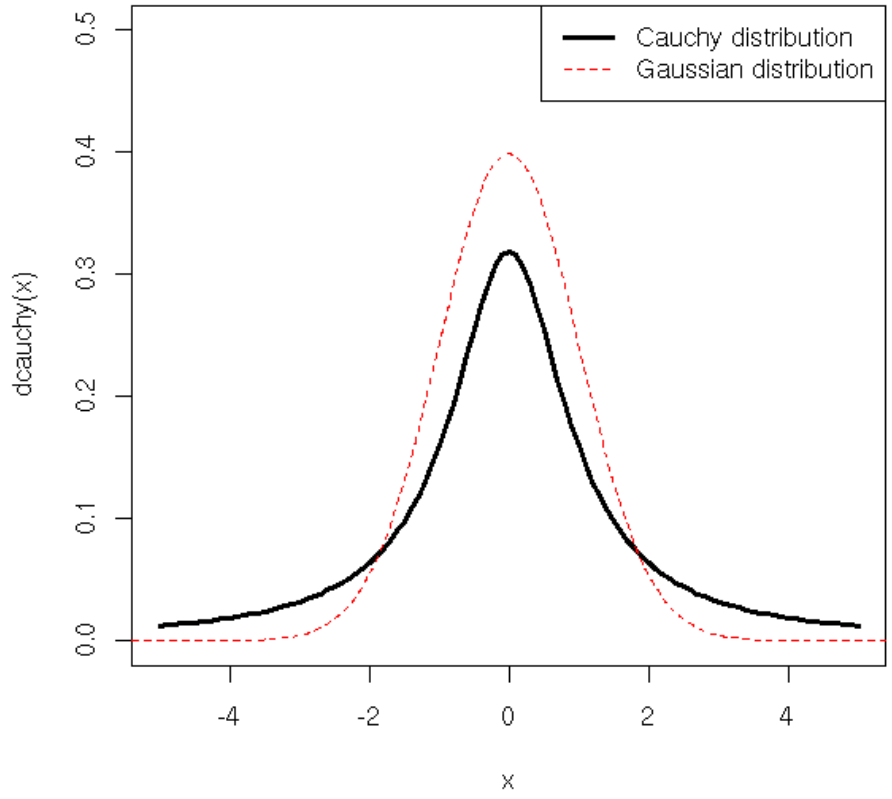
Uniform Distribution



Exponential Distribution:



Cauchy Distribution:



Simulated False Positive Rates and Power

	T-Test		Wilcoxon		Sign Test	
	False +	Power	False +	Power	False +	Power
Uniform Data						
Mean 0	5%		5%		4%	
Mean .05		43%		39%		14%
Mean .10		95%		90%		48%
Normal Data						
Mean 0	6%		6%		5%	
Mean .20		49%		48%		29%
Mean .30		86%		84%		60%
Mean .40		97%		96%		85%
Exponential Data						
Mean 0	6%		40%		4%	
Mean .20		51%		NA		56%
Mean .30		90%		NA		94%
Cauchy Data						
Mean 0	2%		5%		3%	
Mean .20		3%		20%		20%
Mean .30		4%		36%		39%
Mean .40		6%		59%		65%
Mean .50		8%		74%		82%

Data Results (n about 90)

	Skewness	Kurtosis	S-W Test	P-Value From The: T-test	Wilcoxon	Sign Test
HDL: 6 to Baseline						
Group a	- .93	6.4	Fail	.017	.002	.031
Group ace	+ .90	3.5	Fail	.40	.46	.58
Group p	+ .23	1.7	Fail	.71	.56	.33
LDL: 12 to 6 Months						
Group a	- 1.8	7.1	Fail	.058	.066	.015
Group ace	+ .77	2.9	Fail	.74	.40	.66
Group p	+ .28	.25	Fail	.51	.48	1.0
Triglycerides: 12 to 6 Months						
Group a	+ .31	2.1	Fail	.27	.17	.21
Group ace	+ 4.3	41.6	Fail	.98	.23	.22
Group p	- 2.0	13.0	Fail	.42	.31	.13
Total Cholesterol: 12 to 6 Months						
Group a	- 1.5	6.0	Fail	.048	.049	.040
Group ace	+ .17	1.6	Pass	.59	.47	.58
Group p	+ .04	-.01	Pass	.55	.57	.52