

MMA 863

Introduction to Analytic Modelling

Master of Management Analytics

Topic 5 – Hypothesis Testing, Conviction and Power

Keith Rogers



Smith
SCHOOL OF BUSINESS

Queen's
University

Agenda

Motivation

Background Concepts

Constructing Hypothesis

Establishing Conviction

Hypothesis Testing

Why do you believe some things and not others?

Some key concepts:

- Parameters *versus* statistics
- Null *versus* alternative hypotheses
- Types of errors
- PDFs, CDFs and the probability of events
- Distributions of random variables
- Sampling distributions of sample statistics
- An event that is at least as extreme as...
- p-value, α and power

As analytically-minded people, we should use data to help us make decisions. In other words, when considering an opportunity, we should consider the evidence and only undertake the opportunity if the evidence supports it at a level that justifies any risks involved.

Hypothesis testing does this in a formal way by establishing a 'null' and 'alternative' hypotheses. The null is generally consistent with inaction; the alternative with some kind of action.

For technical reasons, the null will always contain the equality bound (for continuous data) and the null and alternative will always be complementary sets.

Here is how to create a null and alternative hypothesis.

1. Confirm the business decision under consideration.
2. Write out the ***ALTERNATIVE*** hypothesis in English in the form:
H1: <We will do X> if <the data shows Y>
3. Nudge that towards mathematics in as many steps as it takes.
4. Construct the null as the complement of the alternative.

Creating Hypothesis – Example

Set a formal hypothesis testing process for the following:

The average duration of a cold is 5 days; my mom recommends chicken soup. But I don't like it, should I try the soup?

Creating Hypothesis – Example

1. Confirm the business decision under consideration.

Presumably I don't want to eat the soup but would if doing so would reduce the duration of my time spent sick.

2. Write out the ***ALTERNATIVE*** hypothesis in English in the form :

H1: <We will do X> if <the data shows Y>

H1: <I will eat the soup> if <the data shows a reduction in average duration of a cold>

3. Nudge that towards mathematics in as many steps as it takes.

H1: Average duration with chicken soup < 5 day

H1: $\mu_{CS} < 5$

4. Construct the null as the complement of the alternative.

H0: $\mu_{CS} \geq 5$

Identifying Hypotheses

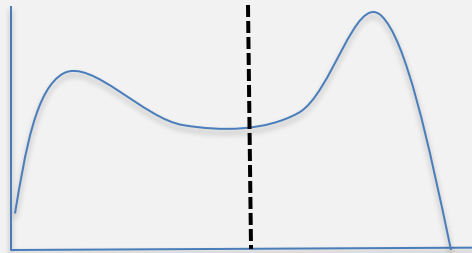
Consider the following scenarios and attempt to generate an appropriate test, null and alternative hypothesis for each.

1. Housing prices in Toronto have increased 15% since last year.
2. A housing developer is considering offering higher priced homes, but can only do so if the market is sufficiently strong.
3. A key machine in your manufacturing process has a defect rate of 2%. This leads to higher repair and warranty costs. A salesman from Acme products claims his new machine, which costs 10% more, would reduce the defect rate to 0.5%.
4. Could I even do this one: I would like to show that the market is stable – that the program's starting salaries this year are the same as last year – to report this in my news paper.

Hypothesis Testing in Three Parts

Assumption $H_0 : \mu_x = 0; H_1 : \mu_x \neq 0$

Sample Distribution



$$\mu_X = \mu_{H0} = 0$$

Sample

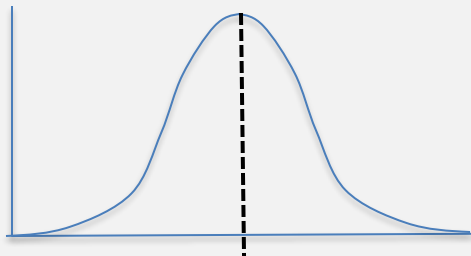
$$x_1, x_2, \dots, x_n$$

Mean

$$\bar{x} = \frac{1}{n} \sum_{i=1..n} x_i$$

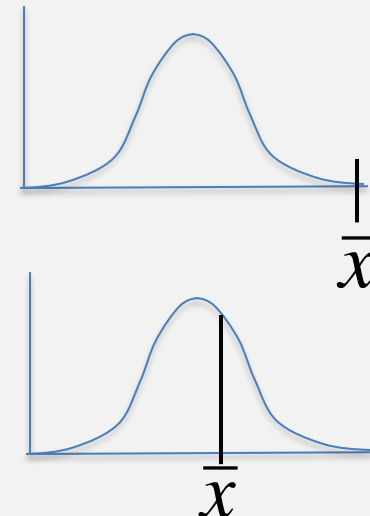
By CLT: $\bar{x}_{n \geq 30} \sim_a N\left(\mu_{\bar{x}} = \mu_x, \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\right)$

Sampling Distribution
of Sample Mean



$$\mu_{\bar{X}} = \mu_X = \mu_{H0} = 0$$

Can Assess
Probabilities



Variations on the Theme with Single Populations

For random variables that look normal, you do not need $n \geq 30$ to use the t-distribution.

For proportions, treat the samples as approximately normal if pn and $(1-p)n > 5$. Note, that we use the normal, not the t-distribution because H_0 gives us p and one can show that given p , $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, so we essentially know the standard deviation 'under the null.'

Types of Errors



If we do everything right, there are two possible errors:

- **Type I error**: reject a true null hypothesis.
- **Type II error**: do not reject a false null hypothesis.

For any statistical test, there are exactly two types of errors:

In reality, the null is:

Our Decision

H_0	True	False
Reject	Type I	
Do Not Reject		Type II

Types of Errors

In a jury context, we can think of these errors as representing the following:

Type I error: Convicting and innocent person (I thought the evidence pointed to guilt, but in reality the person was innocent)

Type II error: Acquitting a guilty defendant (I didn't think the evidence pointed to guilt, but in reality the person was guilty).

In a business context, we can think of these errors as representing the following costs/risks:

Type I error: Cost of failure – I launch a business that turns out not to be successful (I thought the conditions were right, but they were not).

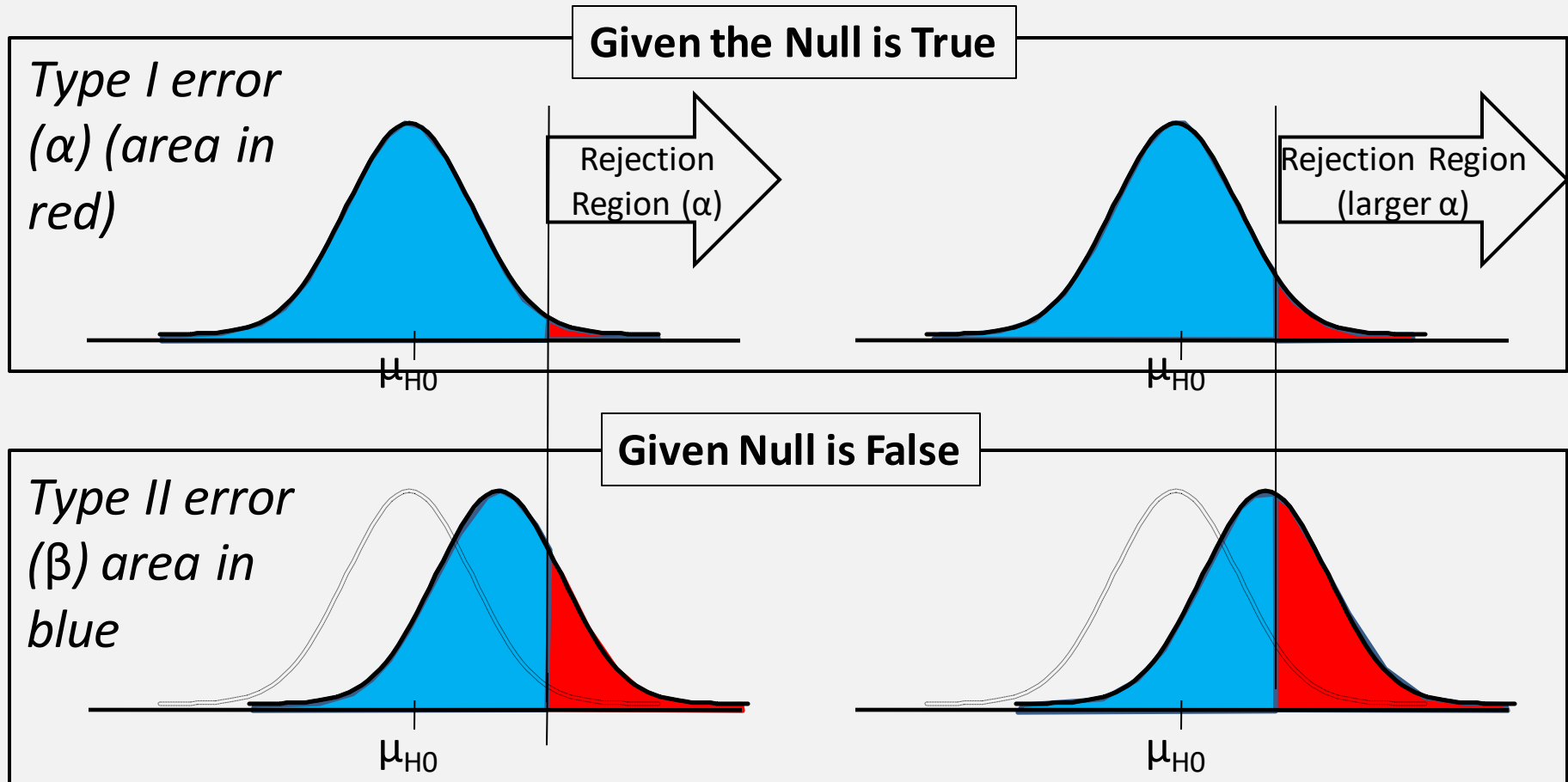
Type II error: Opportunity cost – I don't launch a business that would have been successful (I didn't think the conditions were right, but they were).

Cost of Failure / Opportunity Cost

H0	True	False
Reject	<p>Type I Error</p> <ul style="list-style-type: none">- Conclude H1 is true and it is not.- We do the thing we were considering when we should not so we fail.- We experience the cost of failure.- Small alpha protects against this.	<p>Good Decision</p> <ul style="list-style-type: none">- Conclude H1 is true and it is.- We do the thing we were considering- This was the right choice.- We enjoy the benefit of success.
Not Reject	<p>Good Decision</p> <ul style="list-style-type: none">- Stick with H0 when it is true.- We do not do the thing we were considering- We avoid the cost of failure.	<p>Type II Error</p> <ul style="list-style-type: none">- Stick with H0 when we shouldn't.- We do not do the thing we were considering though we should have.- We experience the opportunity cost of forgoing a good option.- High power protects against this.

Tests and Errors

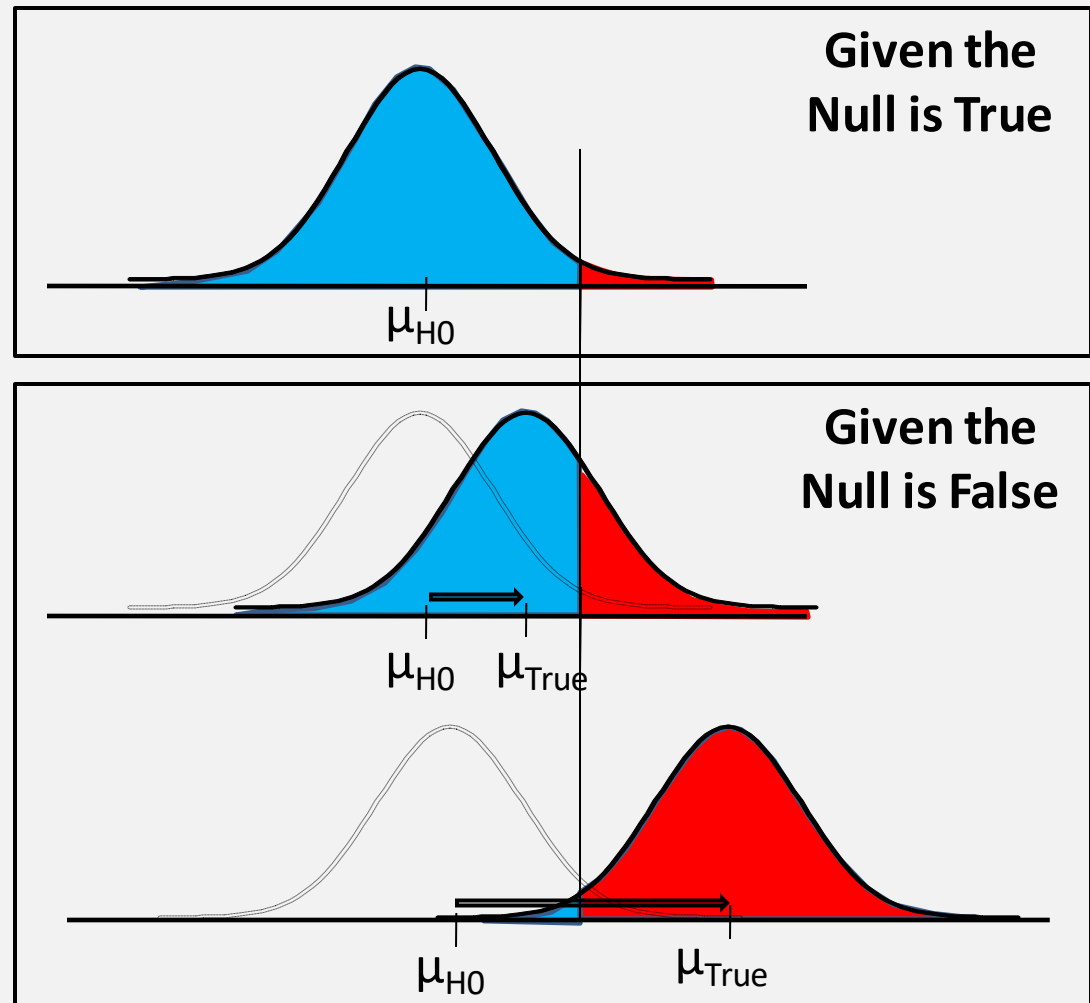
The more willing we are to risk a type I error, the less risk of a type II error.



Tests and Errors

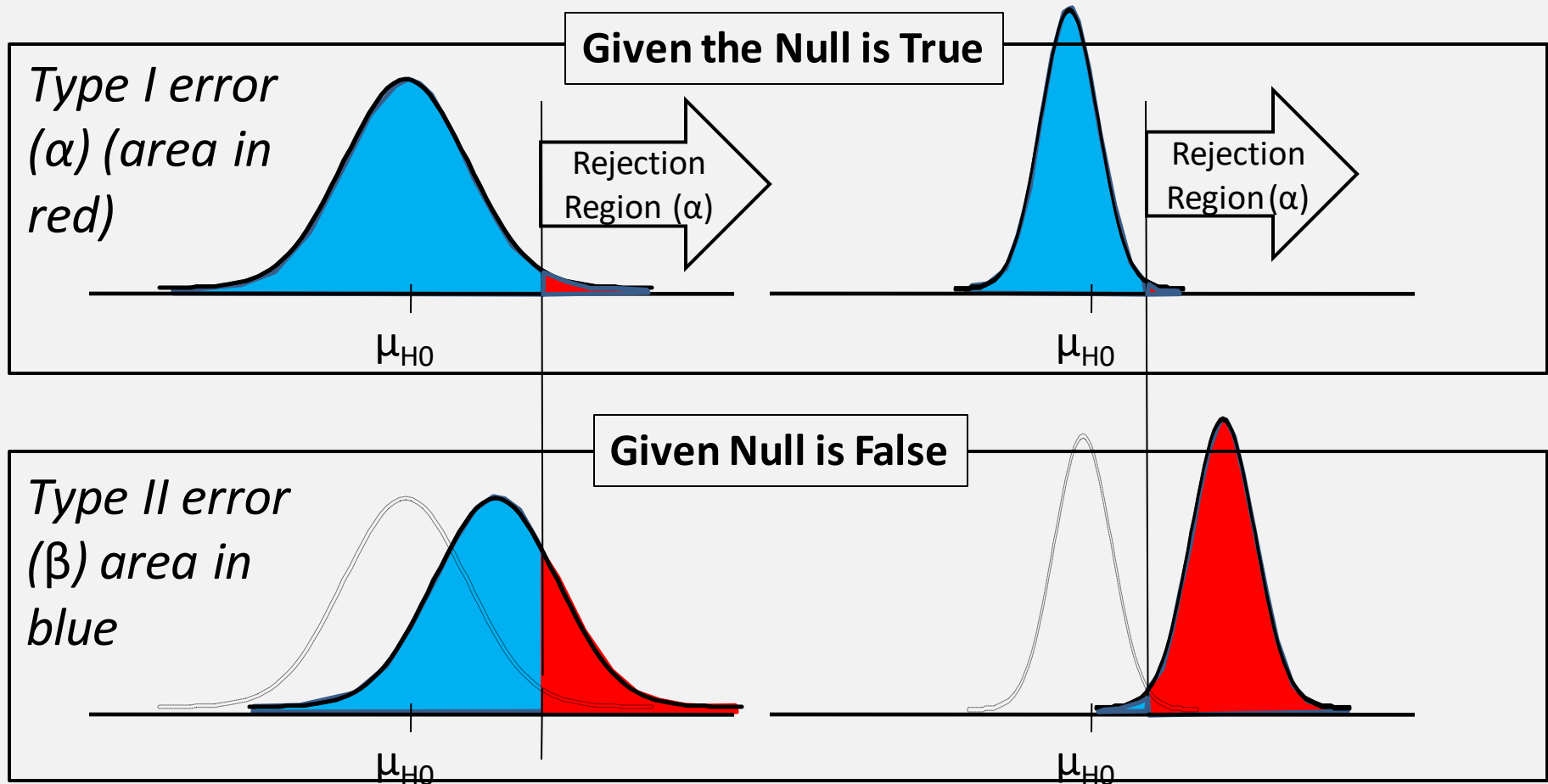
The further the true parameter value is from that assumed by the null, the smaller the probability of a type II error.

In other words, the more 'untrue' the null, the greater the chance of rejecting it.



Tests and Errors

The larger n , the smaller $\sigma_{\bar{x}}$, the lower the probability of a type II error.



This Brings us to Power...

Power represents the probability of rejecting the null hypothesis when it is false. The power of a test is defined as $1 - \beta$.

H_0	True	False
Reject	Type I	Power
Do Not Reject		Type II

The power of a test depends on how far the true parameter value is from that assumed by the null hypothesis. The distance between the null hypothesis value and the truth is called the **effect size**.

Power increases as $n \uparrow$; effect size \uparrow ; $\alpha \uparrow$; $\sigma \downarrow$.

Cost of Failure / Opportunity Cost

H0	True	False
Reject	<p>Type I Error</p> <ul style="list-style-type: none"> - Conclude H1 is true and it is not. - We do the thing we were considering when we should not so we fail. - We experience the cost of failure. - Small alpha protects against this. 	<p>Good Decision</p> <ul style="list-style-type: none"> - Conclude H1 is true and it is. - We do the thing we were considering - This was the right choice. - We enjoy the benefit of success.
Not Reject	<p>Good Decision</p> <ul style="list-style-type: none"> - Stick with H0 when it is true. - We do not do the thing we were considering - We avoid the cost of failure. 	<p>Type II Error</p> <ul style="list-style-type: none"> - Stick with H0 when we shouldn't. - We do not do the thing we were considering though we should have. - We experience the opportunity cost of forgoing a good option. - High power protects against this.

Hypothesis Testing in a Business Context

Discuss the following business scenarios. Think about: what should the null be; what would Type I / Type II errors represent; what value of alpha makes sense; and how much Power do you need? The scenarios are deliberately brief think about what omitted details are important.

1. Should you run a 10% or 20% price discount this weekend to increase traffic to your store?
2. Targeted ads on social media are five times as expensive but appear to be twice as effective.
3. If sales growth projections are favourable, you are planning to build a series of 15 distribution locations to service the growing market in Brazil.
4. Chris has been performing below average in terms of sales for the past 5 days in a row and you are contemplating a dismissal.

Hypothesis Testing

A local grocery store is considering a service where customers order online and receive delivery by truck the next day. To be profitable the average order must exceed \$85.

Steps:

1. Identify the two hypotheses and the size of the test, α
2. Identify / code data
3. Draw a picture of the distribution
4. Calculate the test statistic
5. Use Z or t to determine p-value and compare to α
6. Reject or fail to reject the null
7. Write a managerially relevant conclusion

Hypothesis Testing Exercise 2

We will now extend hypothesis testing to two populations. We develop this through an example using tire data. Later you can repeat the tire data analysis and do similar analysis using gasoline data.

Two Population Hypothesis Testing Exercise

You have just been hired as the fleet quality control manager for a New York based taxi company. The company has been using Brand X tires for years, but you suspect that Brand Y would have a longer life. You order 40 of each brand of tire and run an experiment to determine how quickly each tire wears down. Assume the data will be the amount of wear, in mm, after 1 month.

Should you switch tires brands?

1. Formally state the null and alternative hypotheses.
2. Use the data to perform the appropriate test assuming:
 - a. You distribute 40 new tires of each brand randomly across the fleet.
 - b. You put one of each tire type on the back of 40 cars.

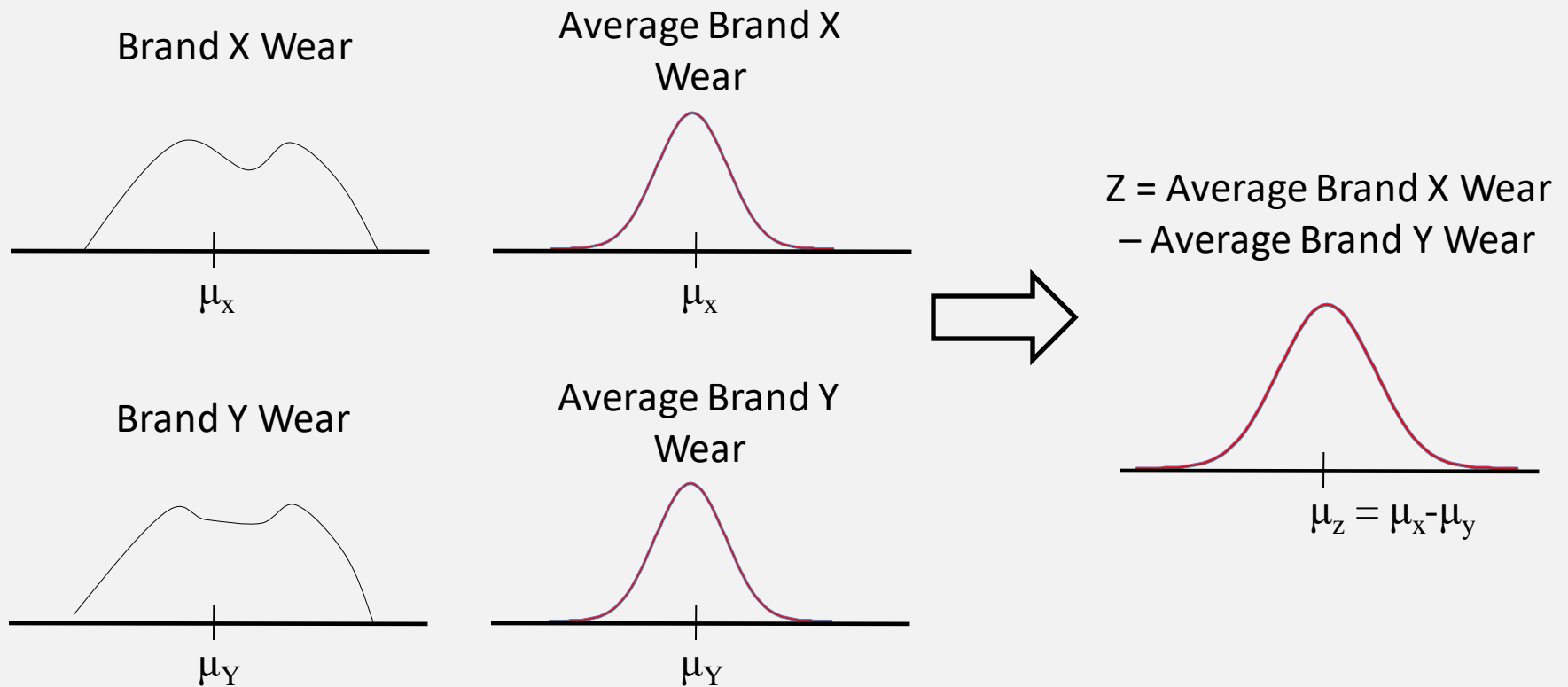
There are a Couple of Ways to Think About This

Unlike previous hypothesis testing questions, we now have data from two different populations:

X_i = wear on one Brand X tire; Y_i = wear on one Brand Y tire.

There are a number of ways we can put these together as well as a variety of assumptions we could make about them.

The difference in averages is approximately normal.



One Way to Look at the Data

Since we have samples of $n > 30$, we know under the CLT that:

$$\bar{X} \sim N(\mu_x, \sigma_{\bar{x}}) \text{ and } \bar{Y} \sim N(\mu_y, \sigma_{\bar{y}}).$$

Based on this, we could calculate

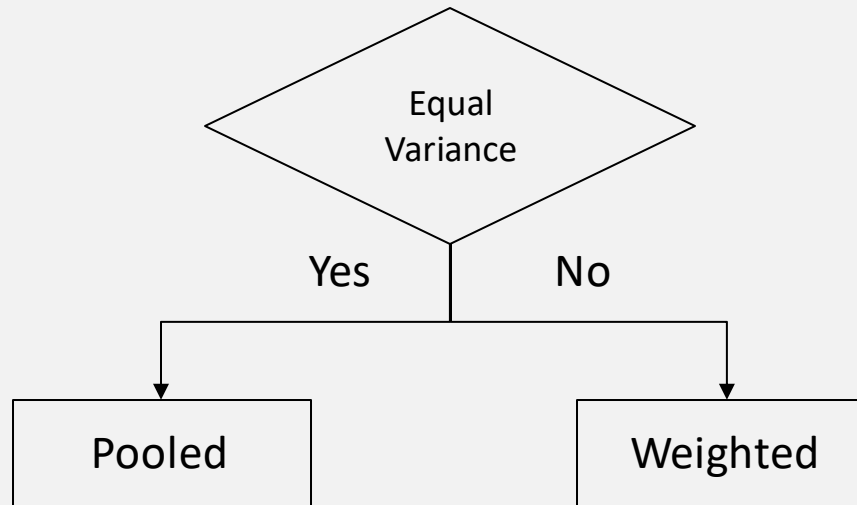
$$\bar{Z} = \bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \sigma_{\bar{Z}})$$

From here, the null and alternative hypotheses are easy:

- H0: Brand X has the same or less wear than Brand Y: $Z \leq 0$
- H1: Brand X has more wear than Brand Y: $Z > 0$

The only problem is to identify (or estimate) $\sigma_{\bar{Z}}$, which will require a short explanation and a bit of work.

Two-Population Testing Flow Chart



REFERENCE ONLY

With two random variables you generally convert into a single one.

For continuous random variables:

- If independent and $\sigma_X = \sigma_Y$, pool to estimate $s_x = s_y$.
- If independent and $\sigma_X \neq \sigma_Y$, use a weighted estimate of stdev.
- If dependent then calculate a difference $d_i = x_i - y_i$.

For proportions:

- If H_0 implies $P_1 = P_2$ then pool to estimate stdev.
- If H_0 implies $P_1 \neq P_2$ then weight to estimate stdev.

REFERENCE ONLY

If we knew σ_x and σ_y we could calculate σ_z using what we already know about normal distributions and standard deviations. **Assuming independence** could show that.

$$\sigma_{\bar{z}} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

In practice, we don't know σ_x and σ_y . So we estimate σ_z using s_x and s_y . The method of estimation depends on whether or not we are willing to assume that $\sigma_x = \sigma_y$.

Estimating σ_z when $\sigma_x = \sigma_y$ and $\sigma_x \neq \sigma_y$

REFERENCE ONLY

$$\sigma_x = \sigma_y$$

Estimate one pooled value

$$S_{pool} = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}$$

$$S_{\bar{z}} = \sqrt{S_{pool}^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}$$

$$df = n_x + n_y - 2$$

$$\sigma_x \neq \sigma_y$$

Estimate weighted values

$$S_{\bar{z}} = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

$$df = \frac{(s_x^2 / n_x + s_y^2 / n_y)^2}{\frac{(s_x^2 / n_x)^2}{n_x - 1} + \frac{(s_y^2 / n_y)^2}{n_y - 1}}$$

$$df^* \cong \text{Min}(n_x - 1, n_y - 1)$$

*Since the df formula is very complex, for this course, we will use this approximation.

REFERENCE ONLY

By now, we have $S_{\bar{z}}$ so we can complete the test using the rejection region or P-value methods just like we did before.

$$\begin{array}{l} \begin{array}{l} \sigma_x = \sigma_y \longrightarrow \\ \left| \right. \\ t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S_{\bar{z}}} \end{array} \quad t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{S_{pool}^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}} \quad df = n_x + n_y - 2 \\ \begin{array}{l} \left| \right. \\ \sigma_x \neq \sigma_y \longrightarrow \end{array} \quad t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \quad df \cong \text{Min}(n_x - 1, n_y - 1) \end{array}$$

Example

You have just been hired as the fleet quality control manager for a New York based taxi company. The company has been using Brand X tires for years, but you suspect that Brand Y would have a longer life. You order 40 of each brand of tire and run an experiment to determine how quickly each tire wears. Test with $\alpha = 0.05$.

After your experiment, you collect the following data:

$$\bar{x} = 1.3; \bar{y} = 1.19; S_x = 0.3; S_y = 0.25;$$

Hypothesis Testing 'Cook Book'



Steps:

1. Identify the two hypotheses and the size of the test, α .
2. Identify / code data
3. Draw a picture of the distribution
4. Calculate the test statistic
5. Use Z or t to determine p-value and compare to α
6. Reject or fail to reject the null
7. Write a managerially relevant conclusion

Only we need to pay extra attention to step 4.

How to Tell if $\sigma_x = \sigma_y$

Whether $\sigma_x = \sigma_y$ can be formally tested. In the absence of a formal test, there are more practical issues.

1. If the estimates for s_x and s_y or the sample sizes, n_x and n_y are similar the assumption that $\sigma_x = \sigma_y$ has minor risk.
2. Tests tend to be more powerful if you assume $\sigma_x = \sigma_y$.
3. If the samples are large, you have plenty of power so assuming $\sigma_x \neq \sigma_y$ is more appropriate.
4. If you assume that $\sigma_x = \sigma_y$, and they are not, that wrong assumption can be the source of a rejection. This may be a good thing or a bad thing.

If the Data are Not Independent

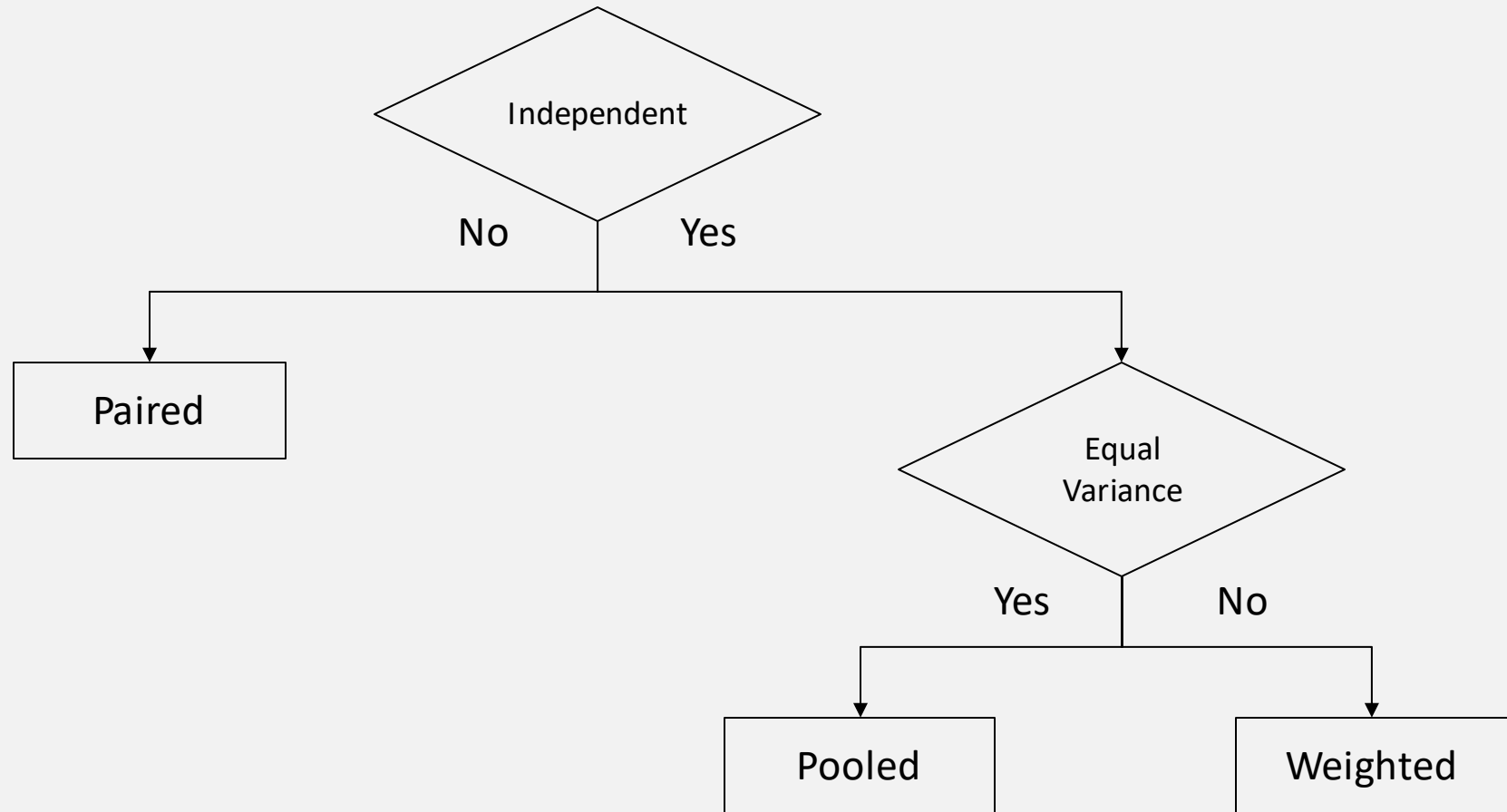
This might be better motivated with a different example. Suppose you were running the HR department of a large company. You are considering an expensive new training process for your sales team. Before you implement it company-wide you would like to know if it is effective.

So you send 50 people for training as an experiment. You could consider the average sales before and after training for this sample. However, you would be missing an important feature of this experiment, namely that the before and after data comes in pairs associated with individual sellers.

Since the resulting data comes in pairs – a before-training score and an after-training score for each employee – the individual seller's characteristics will impact both results. An excellent sales person will tend to do well both before and after training; a bad sales person would do poorly.

These characteristics lead to extra variability in the resulting data, making it harder to detect differences. Some of the variability can be eliminated by taking the difference between each individual's data elements. Constructing a test to eliminate this extra variability leads to greater power.

Two-Population Testing Flow Chart



You have just been hired as the fleet quality control manager for a New York based taxi company. The company has been using Brand X tires for years, but you suspect that Brand Y would have a longer life. You order 40 of each brand of tire and run an experiment to determine how quickly each tire wears. Test with $\alpha = 0.05$.

Now suppose you ran your experiment by putting one of each tire on 40 cars...

Two of Each Brand per Car

Part of the variability in tire wear is due to of the characteristics of the tire, and in part due to how it was used. By having one of each tire on each of 40 cars, you could 'net out' the effect of variability.

It actually does this to some degree without your even being able to identify the source or quantity of that common variability.

Repeat This with All 40 Tires...

To do this, we pair observations by car and then take the difference between the two observations in each group:

Let $d_i = x_i - y_i$

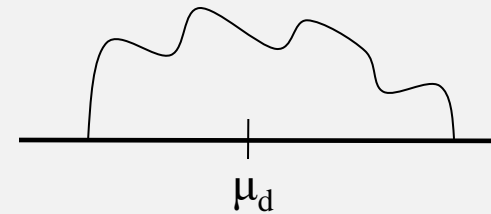
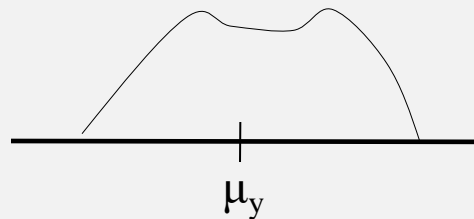
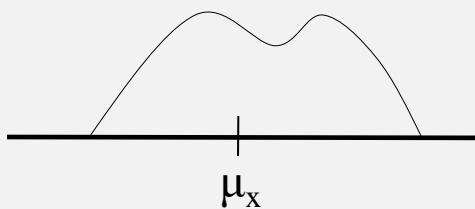
Brand X Wear (x_i)

–

Brand Y Wear (y_i)

=

Difference in Wear (d_i)

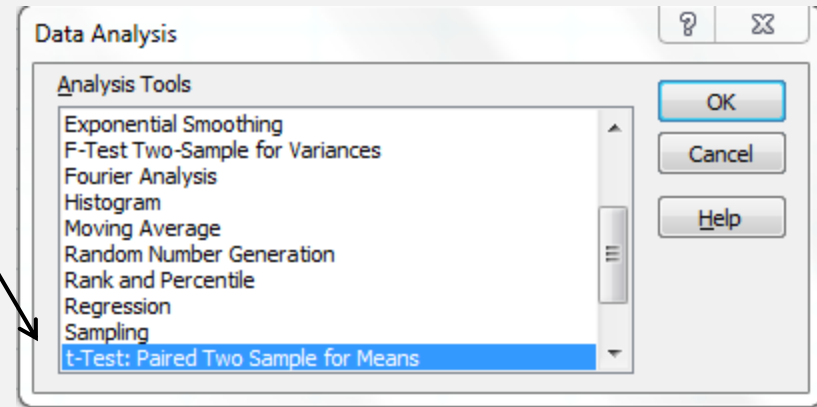


Now We Essentially Have One Sample Again

This time it is different because we have netted out some individual differences.

Hypothesis Testing Exercise in Excel

1. Open the Class Data File to the tire wear tab. You will have to manipulate the data a bit for the matched t-test.
3. Choose Data Analysis ToolPak / t-Test: Paired for Two Sample Means.
4. Fill in the appropriate boxes and click OK.



	Brand X	Brand Y
1	0.82	0.43
2	1.07	0.87
3	1.09	1.00
4	0.72	0.77
5	0.91	0.45
6	0.74	0.07
7	0.89	0.72
8	0.33	0.31
9	0.25	0.17
10	0.92	0.61
11	0.85	0.09
12	0.95	0.91
13	0.69	0.76
14	0.46	0.41
15	0.63	0.58
16	0.85	0.10

The 't-Test: Paired Two Sample for Means' dialog box is shown. In the 'Input' section, 'Variable 1 Range' is \$B\$4:\$B\$44 and 'Variable 2 Range' is \$C\$4:\$C\$44. 'Hypothesized Mean Difference' is empty. The 'Labels' checkbox is checked, and 'Alpha' is 0.05. In the 'Output options' section, 'Output Range' is \$E\$4 (selected), 'New Worksheet Ply' is unselected, and 'New Workbook' is unselected. Buttons for 'OK', 'Cancel', and 'Help' are on the right.

Great! You Rejected the Null!

(So what?)

The Role of Power and Alpha in Conviction

Alpha (α) is the probability that you will erroneously reject a true null given that it is true; **power** is the probability of rejecting a false null given that it is false with a particular value.

In my experience the importance of these issues is often lost in the math. Power and alpha relate to how much a test should alter your conviction of a prior belief. You should only adjust your belief if a test has a sufficiently high power and low alpha; consequently low power tests are not worth doing even if you are lucky enough to get a rejection.

To understand why requires a Bayesian-inspired argument which I will present using stylized facts.

Suppose you are launching a new product. Experience tells you the product will either be a hit and sell to 20% of the market or be a dud and sell to only 5% of the market.* You also know that 99% of products are duds, so you want to run a trial in a test market to demonstrate that it will be a success before you launch.

H_0 : Product is a dud

H_1 : Product is a hit

Suppose you run a test with $\alpha = 0.05$ and reject the null. Should you conclude that the product will be a hit? If so, how much conviction should you have in your new belief?

*Note: Constructing the alternate hypotheses as a single value is not required but it simplifies things considerably.

To answer this question, we need to define some events and link them with Bayes' formula. Our events will be that the product is a **Dud** or a **Hit**; and either that the test **Conclude Hit** it (i.e. reject H_0) or **Conclude Dud** it (maintain H_0). From Bayes' theorem we get the following:

$$\begin{aligned} P(\text{Hit} \mid \text{Conclude Hit}) &= \frac{P(\text{Conclude Hit} \mid \text{Hit})P(\text{Hit})}{P(\text{Conclude Hit})} \\ &= \frac{P(\text{Conclude Hit} \mid \text{Hit})P(\text{Hit})}{P(\text{Conclude Hit} \mid \text{Hit})P(\text{Hit}) + P(\text{Conclude Hit} \mid \text{Dud})P(\text{Dud})} \end{aligned}$$

Establishing Conviction (continued...)

From the previous page:

$$P(\textit{Hit} \mid \textit{Conclude Hit}) =$$

$$\frac{P(\textit{Conclude Hit} \mid \textit{Hit})P(\textit{Hit})}{P(\textit{Conclude Hit} \mid \textit{Hit})P(\textit{Hit}) + P(\textit{Conclude Hit} \mid \textit{Dud})P(\textit{Dud})}$$

$$P(\textit{Conclude Hit} \mid \textit{Hit})P(\textit{Hit}) + P(\textit{Conclude Hit} \mid \textit{Dud})P(\textit{Dud})$$

Since power was the probability of rejecting the null when it was false and alpha is the probability of rejecting the null when it is true, we can make the following substitutions:

$$= \frac{(\textit{Power})P(\textit{Hit})}{(\textit{Power})P(\textit{Hit}) + (\textit{Alpha})P(\textit{Dud})}$$

Establishing Conviction (continued...)

The resulting probability can be very low for a low power test, for tests with low likelihood of the alternate being true or with high alpha. If the test was run on a test market of one individual (who loved it). The probability that it is actually a hit is given by:

$$P(Hit | Conclude Hit) = \frac{(Power)P(Hit)}{(Power)P(Hit) + (Alpha)P(Dud)}$$
$$= \frac{0.2(0.01)}{0.2(0.01) + 0.05(0.99)} = 0.0388$$

If test were conducted with Power = 0.999 we would get:

$$= \frac{0.999(0.01)}{0.999(0.01) + 0.05(0.99)} \cong 0.168$$

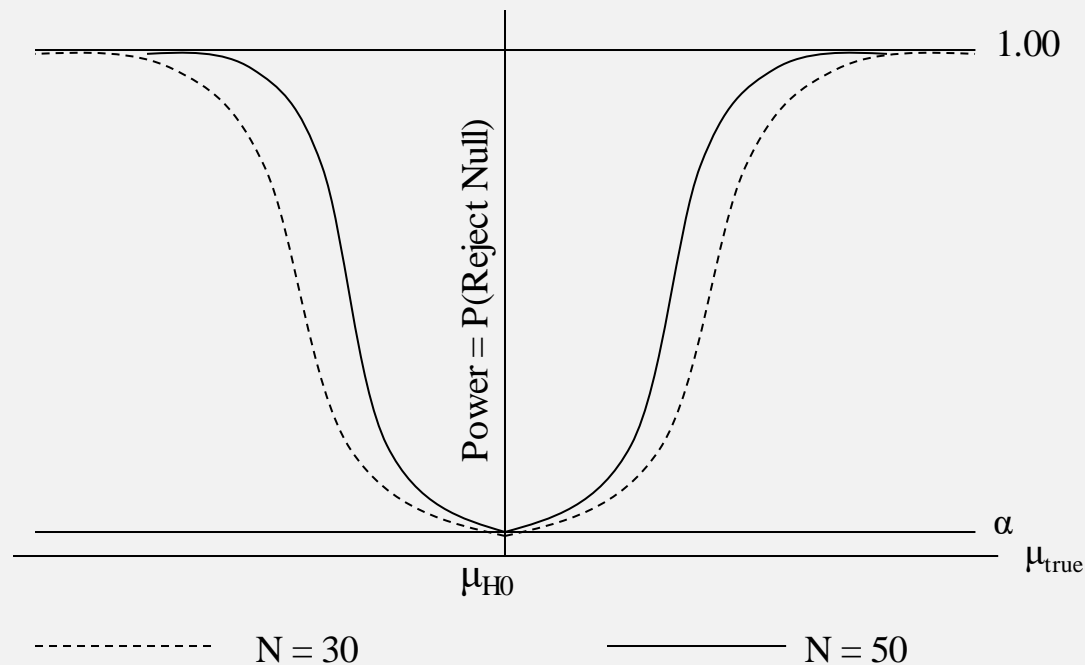
For an interesting discussion of related issues in medical research see:

<http://www.economist.com/node/4342386>

Determining Power

Power increases with effect size and sample size. Power decreases with population standard deviation and α .

Typical Power Curve



Grocery Delivery Problem

A local grocery store is considering a service where customers order online and receive delivery by truck the next day. To be profitable the average order must exceed \$85. They would like to test the appropriate null hypothesis with an $\alpha = 0.05$.

1. We can test hypothesis by building on the logic of the CLT.
2. Hypothesis testing starts with a formal statement that we assume to be true in the absence of evidence – the null hypothesis.
3. The alternate hypothesis covers every possible 'state of the world' that is not captured by the null hypothesis so exactly one of them must be true.
4. There are two types of errors you can make in hypothesis testing and there are trade-offs between them.
5. Alpha is the probability of a type I error; Beta is the probability of a type II error. Beta is a function of the true parameter value.

6. Rejecting a null does not mean that the alternative is true; failing to reject a null does not mean that it is. Testing does, however, provide evidence that should impact your degree of belief.
7. In a Bayesian context, power and prior probabilities should also impact your conviction. It does this in a formal way.
8. Power can be calculated using the rejection region method when sigma is assumed to be known. There are ways to do it when sigma is not known but it is more complex / requires simulation.
9. The greater the effect size (i.e. the more untrue the null) the greater the power.
10. You should be able to design appropriate hypothesis tests and use Excel (and later R) to implement them.
11. You should be able to calculate the power of a simple test.

Questions
