Objectives

Quantifying the quality of hypothesis tests

- Type I and II errors
- Power of a test
- Cautions about significance tests
- Designing Experiments based on power
Evaluating a testing procedure

The testing procedure depends on two decisions made by the practitioner, the significance level $\alpha$ and the sample size (which the practitioner collects). Before or after doing a statistical test we need to understand how these two factors may have influenced the result (the decision made in the test).

When designing your experiment: we need to know whether the test can reject the null for `important’ alternatives. *Is the sample size large enough?*

If we conduct the test and we do not reject the null we need to investigate whether the test could have `easily’ rejected the mean if the important mean was used.

This comes under the canopy of evaluating the testing procedure.

In this chapter we review how these values may have influence the outcome of the test.

We will make many calculations, but most of this can be done using JMP or any other statistical software.
Next we consider this little guy.
A real story

- A few years ago while I was expecting that little creature, the doctor wanted to determine whether I had gestational diabetes. There exists accurate medical tests to determine if a patient has gestational diabetes, but this takes time and is also costly. Instead, a screening process is done to identify likely candidates. The screening process is based on taking a few blood samples and testing whether there is `evidence' of gestational diabetes.

- Gestational diabetes is diagnosed if the mean glucose level, $\mu$, is greater than 140.

- Here are the results of my blood tests: 139.3, 137.1, 146.9, 149.9. The sample mean was 143.3.

- What do you think, is there evidence of gestational diabetes? If there is, then the patient is given more invasive tests.

- The standard deviation can be estimated from the data, but this will mean the test is not very sensitive to detecting gestational diabetes. Instead a very accurate estimate of the standard deviation is obtained from a blood bank and it is assumed $\sigma = 4$. 
What not to do

- A bad procedure is to use the sample mean being greater than 140 as the decision rule on whether a patient should go for more tests.

- This method would make insurance companies crazy. The reason being is if a women has a healthy mean of 140, there is a 50% chance her sample mean will be greater than 140. Using this as the decision rule will lead to too many so called `false positives', ie. healthy patients referred for tests. Remember

\[ \bar{x} \sim N(\mu, 2) \]

we now illustrate the above with some examples.

On the left we have the distribution of the sample mean for two healthy patients. We see that the chance a sample mean being over 140 is about 16% and 50% depending on their true mean level.

One the left we have the distribution of the sample mean for two unhealthy patients. We see the chance of the sample mean being over 140 depends on their true mean level.
Gestational diabetes detection via testing

- Instead, a one-sided statistical test is done, $H_0 : \mu \leq 140$ (healthy) against $H_A : \mu > 140$ (has gestational diabetes). Given the data we decide how plausible it is to obtain that data under the assumption the null is true (the p-value). If this is `too' small (implausible), we reject the null and determine that it is plausible she has gestational diabetes.

- Typically the 5% significant level is used. This means if the p-value is less than 5% we reject the null. We recall from Chapter 7, that 5% is the boundary and we can construct from this a point such that if $\bar{x}$ is larger than this value we can reject the null. For the doctor’s surgery this is a very simple way to execute the test.

- Using the normal tables (since we are not estimating the standard deviation from the data) we know that the area to the right of a standard normal which gives 5% is 1.64. Therefore the area to the right tail of 140 which gives 5% is $\bar{x} > 143.28$.

This gives an automatic way to do the test. **Remember** using this rule there is a 5% chance of falsely diagnosing a healthy patient (ie, the proportion of healthy people diagnosed is 5%).
The rejection region for this test

- Implementing the rule: This means that when a doctor takes 4 blood samples she takes the average and compares this number to 143.28.
  - If this average is less than 143.28 the doctor cannot reject the null. In other words, the patient is not diagnosed with gestational diabetes.
  - On the other hand, if the average is greater than 143.28 than it is believed the patient has gestational diabetes and more tests are done.

- Examples:
  - A patient has measurements: 36.2, 36.8, 39.11, 39.27, the sample mean is 137.8. Since $137.8 < 140 < 143.28$ we definitely cannot reject the null.
  - A patient has measurements: 143, 143.2, 145, 144 average is 143.8. Since $143.8 > 143.28$ we reject the null, patient goes on to have further tests.
Type I errors: Example my case

- As we can only prove the alternative we need to state what are looking for in the alternative. If we want to discount the possibility of the patient being healthy without gestational diabetes (prove gestational diabetes), we set the hypotheses as:
  - \( H_0 : \mu \leq 140 \) against \( H_A : \mu > 140 \). The hypotheses have to be the opposite of each other. We do the test at the 5% level.
  - My sample means was 143.3. Since 143.3 is greater than 143.28 the null was rejected (the p-value is less than 5%)
  - To calculate the precise p-value.
    - Make a z-transform = \((143.3 - 140)/2 = 1.65\).
    - Since the alternative is pointing right, we calculate the area to the right of 1.65, which from the tables is 4.9%.
    - 4.9% < 5%, therefore I was referred for more tests.

- Fortunately for me, the more invasive medical showed that I did not have gestational diabetes.

- I was an example of a Type I error. The decision where we reject the null hypothesis when the null is true.
Roughly speaking there is a 4.9% chance of a healthy patient yielding the sample 139.3, 137.1, 146.9, 149.9.

The 5% significance level means that by using this level, 5% of healthy patients will be referred for extra tests. My insurance company is happy - it is better than all patients being given costly invasive tests.

But my insurance company would be even happier if they could reduce this level to 1%. At the 1% level I would not have been given the extra tests (the insurance company would have liked this). But there is a cost for this, by reducing the significance level, the doctor is less likely to identify patients who actually have gestational diabetes, because we need a lot more evidence against the null (we can demonstrate this using the Applet described later in the slides).

We need to make a balance. A large significance level means more healthy patients will be given unwanted tests, but a large number of unhealthy patients will be caught. A small significance level means fewer healthy patients will be given tests, but less unhealthy patients will be identified.
The decision rule and the errors made

<table>
<thead>
<tr>
<th>Truth</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$ true, no Gestational diabetes</td>
</tr>
<tr>
<td>Decide not to reject null</td>
<td>Correct decision</td>
</tr>
<tr>
<td>Decide to reject null, $H_A$ true</td>
<td>Incorrect decision</td>
</tr>
</tbody>
</table>

We can quantify the chance of our making the correct or wrong decision:

<table>
<thead>
<tr>
<th>Truth</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$ true, no Gestational diabetes</td>
</tr>
<tr>
<td>Decide not to reject null</td>
<td>Correct decision, chance = $(1 - \alpha)$</td>
</tr>
<tr>
<td>Decide to reject null, $H_A$ true</td>
<td>Type I error, chance = $\alpha$</td>
</tr>
</tbody>
</table>

- We have already covered the type I error (usually denoted $\alpha$), this is the chance of rejecting the null when it is true (chance of diagnosing gestational diabetes when the patient is healthy). The chance of this happening is the pre-set significance level. In this example it is 5%.
- The Type II error (usually denoted $\beta$) is the chance of not detecting gestational diabetes when the patient has gestational diabetes (not rejecting the null when the alternative is true).
- **Definitions** Power is the chance of correctly rejecting the null (diagnosing gestational diabetes when person has gestational diabetes), $\text{Power} = 1 - \beta$
- **Common mistake:** Remember $\alpha$ and $\beta$ are not related linearly: ie. $\beta \neq (1 - \alpha)$. 

Power: Detecting gestational diabetes

- Recall the hypothesis $H_0 : \mu \leq 140$ against $H_A : \mu > 140$.
- In the previous classes I discussed what it means to falsely reject the null (this is known as a type I error).
- Next we discuss the issue of rejecting the null, when the alternative is true (this is known as power).
  - Power = $1 - \text{(the chance of NOT detecting the alternative when a patient has diabetes)} = 1 - \text{Type II error}$.
  - To do the calculation we need to use alternatives which have particular interest. In this example it’s women who have severe gestational diabetes, in other words $\mu$ is far larger than 140.
- We want the statistical test to detect seriously ill patients. To see whether it can, we need to make what is known as a power calculation.
- This calculation has nothing to do with data. It is a method of checking how good the actual statistical test is. In this example, it assesses, how good the test is in diagnosing ill patients.
Motivation: Power = 1 – Type II error

- The sugar level of a pregnant woman after taking a sugary drink has mean level $\mu$. If the mean $\mu$ is above 140, she has gestational diabetes. The problem is we don’t know her mean level. So we make a decision using by doing a hypothesis test $H_0: \mu \leq 140$ against the alternative $H_A: \mu > 140$. We do the test at the 5% level.

- Based on the average of four blood samples, if her sample mean (based on 4 blood samples) is greater than 143.28 (see previous calculation) there is evidence to suggest she has gestational diabetes and a proper medical test is done. 143.28 is our decision threshold.

- People who do have diabetes can also have a sample mean which is less than 143.28, so this are people who are let through the net. Our statistical test has not caught this person, so she will not be given treatment – did we make a mistake? Are we not treating a person who may be ill?

- To answer this question we make use of the following calculators.
We will make use of the following

- Free Applet (made by Dr. Kolodziej): [http://shiny.stat.tamu.edu:3838/eykolo/power/](http://shiny.stat.tamu.edu:3838/eykolo/power/) If you use this applet make sure the range is specified correctly (by changing lower bound of plot to 135 and upper bound of plot to 150) – else it will look strange.

- The power calculator in JMP:

  Instructions: Go to DOE (short for Design of Experiments) -> Sample size and Power -> One sample (this will do the power calculation for you). We will go through it in class (for the column Difference you need to put the difference between the null and alternative of interest). Note the calculations that JMP do are done using the t-distribution (assuming the standard deviation is estimated from the data). This is fine if you simply want to calculate the power. On the other hand, it seems it uses the t-distribution when find a suitable sample size. This I find a little strange and I am unsure how it does its calculation.

I will be using Statcrunch to do most of the calculations.
This patient has mean level 142 (mild diabetes). The chance of this being detected is the red area to the left of 143.38. This probability is the area to the RIGHT of

\[ z = \frac{143.38 - 142}{2} = 0.69 \]

This is about 25% (look up z-tables). That is there is a 25% chance of detecting gestational diabetes when the mean is 142 (the power in this case). It’s not huge, but when the diabetes is mild it probably does not matter. Remember there is a 75% chance of not detecting diabetes in this patient (this is the Type II error in this case).
Example: Power for a patient with moderate diabetes $\mu=145$

This patient has mean level 145 (moderate diabetes). The chance of this being detected is the red area to the left of 143.38. This probability is the area to the RIGHT of

$$z = \frac{143.38 - 145}{2} = -0.81$$

This is about 80% (look up z-tables). That is there is a 80% chance of detecting gestational diabetes when the mean is 145.
This means there is a 21% chance of not detecting diabetes for this patient. Whether this matters or not depends on the practitioner. Will the patient be at risk if her diabetes is not detected?
Example: Power for severe diabetes $\mu \geq 148$

This patient has mean level 148 (severe diabetes). The chance of this being detected is the red area to the left of 143.38. This probability is the area to the RIGHT of

$$z = \frac{143.38 - 148}{2} = -2.31$$

This is about 99.0% (look up z-tables). That is there is a 99% chance of detecting gestational diabetes when the mean is 148. This means there is a 1% chance of not detecting diabetes for this patient. Thus there is a very good chance of our test detecting severe diabetes.
Using this test there is 99% we have calculated means that 99% of women who are severe (mean level of 148 and over).

To increase the detecting rate:

- Increase the significance level from 5-10%. This will move the rejection point from $143.28$ to $140 + 1.28 \times 2 = 142.56$ (1.28 correspond to the 90% on the normal tables). This means that a person with a sample mean greater than 142.56 is referred for further tests. The `cost' of this method is that 10% of women without gestational diabetes will be referred for more tests.

The chance of detection is the area to the RIGHT of 142.56. This is the probability to the RIGHT

\[ z = \frac{142.56 - 148}{2} = -2.72 \]

There is a 99.7% of detecting the severe diabetes, but a 10% of falsely diagnosing a person who is healthy.
An alternative method to increase power without changing the significance level is to **increase** the sample size. By increasing the sample size we increase the reliability of the sample mean (by decreasing the standard error). This is best seen with a calculation, increase the sample size from 4 to 9, construct the point at which rejection occurs and again the power.

By increasing the sample size:
- We push the rejection point closer to 140: \[ 140 + 1.64 \times \frac{4}{\sqrt{9}} = 142.2 \]
- We also make the sample means more centered about the patients mean level.
- This combination means the chance of detecting severe diabetes is the area to the right of \[ z = \frac{142.2 - 148}{\left(\frac{4}{3}\right)} = -4.35 \], which is 99.999%!
- Even the chance of detecting moderate diabetes ($\mu=145$) increases from 79% to 98.21% (check this).
Increasing power

- To summarize, there are two feasible ways to increase power:
  - Increase the significance level, but this will result in more false positives.
  - Increase the sample size, this also has a cost, since it may be costly to take several samples.
- A less feasible method:
  - Another way to increase power (but does not have any real meaning especially for this example), is to move the alternative further from the null. In other words, increase 148 to say 150. However, this really has no practical meaning, since severely ill people are those who level is over 148.

Choosing the Type I error, $\alpha$, is a balancing act between the risk of a Type I error and the risk of a Type II error, $\beta$.

Reducing $\alpha$ (type I error) always increases $\beta$ (type II error).
Why power calculations are important

The **power** of a test of hypothesis with fixed significance level \( \alpha \) is the probability that the test will reject the null hypothesis when the alternative hypothesis is true. That is, **the power is** \( 1 - \beta \).

In other words, power is the probability that the data gathered in an experiment will be sufficient to reject a null hypothesis that is false.

Knowing the power of your test is important:

- When designing your experiment: select a sample size large enough to detect an effect of a magnitude you think is meaningful. For example in the gestational diabetes example it was important the test should detect \( \mu \geq 148 \).

- When your test ends up being non-significant: check that your test would have had enough power to detect an effect of a magnitude you think is meaningful. If the power is large, then it is unlikely the alternative of interest is true.
Example: Gestational diabetes

6 blood samples are taken to test for gestational diabetes. $H_0 : \mu \leq 140$ against $H_A : \mu > 140$ with $\sigma=4$. The sample mean of these 6 blood samples is 142. There is no evidence to reject the null hypothesis. The patient is not given further treatment and sent home. Could we have missed a patient who actually had severe gestational diabetes ($\mu \geq 148$)?

- We need to do a power calculation. Using the Statcrunch app we see that

![Power calculation graph]

From the power calculator we see that the chance of missing a person with such severe diabetes is

$$1 - 0.99943129 = 0.00056871 = 0.056\%$$

This is a very small, chance. Thus we can be reassured we have not missed a seriously ill patient.
Review: Type I and Type II errors

- A **Type I error** is made when *the null hypothesis is rejected* but *the null hypothesis actually is true*. (But we do not know that it is true.)
  The probability of making a Type I error is the significance level $\alpha$.
  *The conclusion for the cola study may have been a Type I error.*

- A **Type II error** is made when *the null hypothesis is not rejected* but *the null hypothesis actually is false*. (We do not know that it is false.)
  The probability of making a Type II error is labeled $\beta$, but it is not a fixed number because it depends on the actual truth (true value of $\mu$).
  *The conclusion for the tomato study may have been a Type II error.*

Here, “error” does not mean we made an avoidable mistake. It simply means we got misleading evidence (by chance alone).
Examples of Type I and Type II errors

- Recall the gestational diabetes example:
  - $H_0 : \mu \leq 140$ against $H_A : \mu > 140$

- The truth is never known, if the patients sample mean has a p-value less than 5% (or equivalent the sample mean is greater than 143.28), then the patient is diagnosed with gestational diabetes and further tests are done. If the p-value is greater than 5% nothing is done.

- My case was an example of a Type I error, where, as a healthy patient, I was diagnosed with gestational diabetes since my p-value was 4.9%.

- A patient with gestational diabetes who has not been diagnosed with gestational diabetes, because her p-value is larger than 5% is an example of a patient who `falls through the net’ and is an example of a Type II error.
Choosing a larger $\alpha$ makes it more likely that $H_0$ will be rejected and thus more likely that a Type I error will occur (if $H_0$ is true) but less likely that a Type II error will occur (if $H_0$ is false).

On the other hand, choosing a smaller $\alpha$ makes it less likely that $H_0$ will be rejected and less likely that a Type I error will occur (if $H_0$ is true) but more likely that a Type II error will occur (if $H_0$ is false).
Example 2: Low potassium

- Hypokalemia is diagnosed when the blood potassium level is below 3.5mEq/dl. The potassium in a blood sample varies from sample to sample and follows a normal distribution with unknown mean but standard deviation is known to be 0.2. The hypothesis of interest is \( H_0 : \mu \geq 3.5 \) against \( H_A : \mu < 3.5 \). Typically, the average of 9 blood samples are taken and the test is done at the 5% level.

- **Question:** Will the test be able reliable in detecting a critically ill patient with potassium level below 3.3?

- We translate the above into statistical speak. It asks whether the test will have very high power (over 99.5%) for detecting people whose mean level \( \mu \leq 3.3 \)?

- **Answer:** The point of rejection is when the sample mean is less than

\[
3.5 - 1.64 \times \frac{0.2}{\sqrt{9}} = 3.39
\]

The probability of the sample mean being less than 3.39 when the true mean \( \mu=3.3 \) is the area to the LEFT of the z-transform:

\[
z = \frac{3.39 - 3.3}{(0.2/3)} = 1.35
\]

Which is 91.1%
In this case, the difference between null and alternative of interest is $3.5 - 3.3 = 0.2$. The standard deviation is 0.2 and the sample size is 9 blood samples.

The calculator gives 91.23% chance of catching (detecting people) who have mean level of 3.3 or below.

**Application** A person is tested for low potassium using the above procedure. The test is done at the 5% level (using 9 blood samples). There is no evidence to reject the null, should we be worried that we have missed someone who is severely ill (mean level below 3.3)? The power calculation show that the chance of missing a person who is severely ill is 8.77%. The medical professional needs to decide whether this is too high and the consequences of it.

If the doctor does decide it is too high then the doctor will either need to collect more bloods samples or do the test at higher significance level.
Prospective sales people are now being offered a sales training program. Previous data indicate that without training the average sales person sold 33 items. The company wants to access the impact the sales training program has on sales. One month after training started, 40 people had training and it was found that on average 34.6 books were being sold.

- Using that the population standard deviation is 8.4. Has the training enhance sales (do the test at the 5% level).
- The company decides that the training is only cost effective if on average 38 or more books are sold. Evaluate the power for the test above and what this means about the results of the statistical test we have just done.
Solution 3

- We want to see whether mean sales have improved after some training, therefore the hypotheses are $H_0 : \mu \leq 33$ against $H_A : \mu > 33$. Assuming normality of the sample mean and the p-value is the area to the RIGHT of 34.6 we have

$$P(\bar{X} > 35) = P \left( Z < \frac{34.6 - 33}{8.4/\sqrt{40}} \right) = P(Z > 1.2) = 0.11$$

The p-value is 11%, therefore at the 5% level there is NO evidence to reject the null. Note the point of rejection is when the sample mean is greater than

$$33 + 1.64 \times \frac{8.4}{\sqrt{40}} = 35.17$$
Next we want to calculate the power for the specific alternative $H_A : \mu = 38$ (this is an average of 38 book sales). We see that the chance the test can detect this alternative (reject the null when $\mu = 38$) is 98%, which is extremely large.

We see that the power for this alternative is 98%, which is extremely large.

This means if the mean (population mean) increase in sales after training is actually 5 books or more, then the statistical test we used would reject the null (with a very large probability).

We we did not reject the null, it seems very unlikely that the mean increase in sales was over 5.
What is statistical significance?

We say the results of a test is statistically significant if the p-values is less than a pre-decided significance level.

Statistical significance only says that the evidence for $H_a$ is substantial enough to reject $H_0$, because it is unlikely that the observed difference from $H_0$ was due just to random sampling.

Statistical significance may not be practically important. That’s because statistical significance does not tell you about the magnitude of the effect, it only tells you that there is an effect.

An effect could be too small to be relevant or important. And with a large enough sample size, significance can be reached even for the tiniest effect. This is an example of impractical significance. This is when making confidence intervals for the mean can be more useful. The confidence interval tells how big the effect actually is.
Example: Impractical significance

- Let us return to the gestational diabetes example $H_0 : \mu \leq 140$ against $H_A : \mu > 140$. Suppose a patient’s mean level is 140.01. By definition this means she has gestational diabetes. But we can see that it is only very slightly greater than 140: though by definition she has the condition, for all practical purposes she does not.

- For sample sizes of order 5, 10, 15, it would not be possible to diagnose her condition (ie. difficult to reject the null). But this is not an issue, since a mean level so close to 140 her condition is so mild it does not require treatment.

Suppose we could take 50,000 blood samples. In this case the power calculation shows that there is a very large chance of rejecting the null, even when her diabetes is extremely mild. However, with such mild diabetes it is unnecessary to treat the patient. For very large sample sizes we are likely to reject the null even with the alternative is very close to the null and not very useful.
Interpreting effect size: It depends on the context

There is no rule on how big an effect has to be in order to be considered meaningful. In some cases, effects that may appear to be trivial can be very important.

- Example: Improving the format of a computerized test reduces the average response time by about 2 seconds. Although this effect is small, it is important since this is done millions of times a year. The cumulative time savings of using the better format is gigantic.

Always think about the context of the study. Try to plot your results, and compare them with a baseline or results from similar studies.
Do not ignore lack of significance

- Consider this provocative title from the British Medical Journal: “Absence of evidence is not evidence of absence.”
- Having no proof of who committed a murder does not imply that the murder was not committed.

**Failing to find statistical significance only results in not rejecting the null hypothesis.** This is very different from actually accepting the null hypothesis. Lack of significance does not imply that there is no effect. The sample size, for instance, could be too small to overcome large variability in the population.

Lack of significance also does not imply that there is nothing of interest in your study. A test for a population mean or for comparing means does only that; it does not consider other characteristics of the population(s) such as standard deviation(s).
Using power to determine sample size

- Just as we use Margin of Error to guide the sample size we choose, we can use power to guide the sample size we choose.

- **Example** Let us return to the book samples example where we wanted to test if training of a sales person increased the average number of book sales. Our hypotheses is $H_0 : \mu \leq 33$ against $H_A : \mu > 33$.

- Suppose we want to ensure (with at least 90%) to reject the null when sales have increased to 34 (one or over), how large a sample size should we choose?

- **Solution**

  Here we can see that we need to interview 604 salesmen (after training) in order to reject the null for a population mean change of one (from 33 to 34).

  This example shows that in order to detect a very small difference, we need a very large sample size.