Recap: Hypothesis testing framework

1. Set the hypotheses.
2. Check conditions.
3. Calculate a *test statistic* and a p-value.
4. Make a decision, and interpret it in context of the research question.
Recap: Hypothesis testing for a population mean

1. Set the hypotheses
   - $H_0: \mu = \text{null value}$
   - $H_A: \mu < \text{or} > \text{or} \neq \text{null value}$

2. Check conditions
   - **Ind**: random sampling/assignment
     - $n < 10\%$ (when sampling without replacement)
   - **NN**: if $n \geq 30$ and no extreme skew use $Z$ distribution
     - if $n < 30$ and no extreme skew use $t$ distribution
     - otherwise use simulation based approaches

3. Calculate a test statistic and a $p$-value (draw a picture!)
   \[ Z = \frac{\bar{x} - \mu}{SE} \text{ or } T_{df} = \frac{\bar{x} - \mu}{SE}, \text{ where } SE = \frac{s}{\sqrt{n}} \text{ and } df = n - 1 \]

4. Make a decision, and interpret it in context of the research question
   - $p$-value $< \alpha$, reject $H_0$, data provide evidence for $H_A$
   - $p$-value $> \alpha$, fail to reject $H_0$, data do not provide sufficient evidence for $H_A$
Agreement of CI and HT

- Confidence intervals and hypothesis tests (almost) always agree, as long as the two methods use equivalent levels of significance / confidence.
  - A two sided hypothesis with threshold of $\alpha$ is equivalent to a confidence interval with $CL = 1 - \alpha$.
  - A one sided hypothesis with threshold of $\alpha$ is equivalent to a confidence interval with $CL = 1 - (2 \times \alpha)$.

- If $H_0$ is rejected, a confidence interval that agrees with the result of the hypothesis test should not include the null value.

- If $H_0$ is failed to be rejected, a confidence interval that agrees with the result of the hypothesis test should include the null value.
Significance level vs. confidence level

Two sided

One sided
Example - Waiting Times

A 95% confidence interval for the average waiting time at an emergency room is (128 minutes, 147 minutes).

Determine if the following statements are true or false,

(a) a hypothesis test of $H_A : \mu \neq 120$ min at $\alpha = 0.05$ is equivalent to this CI.
(b) a hypothesis test of $H_A : \mu > 120$ min at $\alpha = 0.025$ is equivalent to this CI.
(c) This interval does not support the claim that the average wait time is 120 minutes.
(d) The claim that the average wait time is 120 minutes would not be rejected using a 90% confidence interval.
Example - Sample Size

Suppose $\bar{x} = 50$, $s = 2$, $H_0 : \mu = 49.5$, and $H_A : \mu \geq 49.5$.

Will the p-value be lower if $n = 100$ or $n = 10,000$?
Real differences between the point estimate and null value are easier to detect with larger samples.

However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (effect size), even when the difference is not practically significant.

This is especially important to research: if we conduct a study, we want to focus on finding meaningful results (we want observed differences to be real but also large enough to matter).

The role of a statistician is not just in the analysis of data but also in planning and design of a study.

“To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.” — R.A. Fisher
Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.
There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

<table>
<thead>
<tr>
<th>Truth</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>fail to reject $H_0$</td>
</tr>
<tr>
<td>$H_A$ true</td>
<td>reject $H_0$</td>
</tr>
</tbody>
</table>

- A **Type 1 Error** is rejecting the null hypothesis when $H_0$ is true.
- A **Type 2 Error** is failing to reject the null hypothesis when $H_A$ is true.
- We (almost) never know if $H_0$ or $H_A$ is true, but we need to consider all possibilities.
Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

\[ H_0 : \text{Defendant is innocent} \]
\[ H_A : \text{Defendant is guilty} \]

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty
- Declaring the defendant guilty when they are actually innocent

Which error do you think is the worse error to make?

“better that ten guilty persons escape than that one innocent suffer”

– William Blackstone
Type 1 error rate

- As a general rule we reject $H_0$ when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where $H_0$ is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error.

\[
P(\text{Type 1 error}) = \alpha
\]

- This is why we prefer to small values of $\alpha$ – increasing $\alpha$ increases the Type 1 error rate.
## Decision errors

### Error rates & power

Filling in the table...

<table>
<thead>
<tr>
<th>Truth</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>fail to reject $H_0$</td>
</tr>
<tr>
<td>$1 - \alpha$</td>
<td>$\alpha$ Type 1 Error, $\alpha$</td>
</tr>
<tr>
<td>$H_A$ true</td>
<td>Type 2 Error, $\beta$</td>
</tr>
</tbody>
</table>

- **Type 1 error** is rejecting $H_0$ when you shouldn’t have, and the probability of doing so is $\alpha$ (significance level)
- **Type 2 error** is failing to reject $H_0$ when you should have, and the probability of doing so is $\beta$ (a little more complicated to calculate)
- **Power** of a test is the probability of correctly rejecting $H_0$, and the probability of doing so is $1 - \beta$
- In hypothesis testing, we want to keep $\alpha$ and $\beta$ low, but there are inherent trade-offs.
Type 2 error rate

If the alternative hypothesis is actually true, what is the chance that we make a Type 2 Error, i.e. we fail to reject the null hypothesis even when we should reject it?

- The answer is not obvious.
- If the true population average is very close to the null hypothesis value, it will be difficult to detect a difference (and reject $H_0$).
- If the true population average is very different from the null hypothesis value, it will be easier to detect a difference.
- Clearly, $\beta$ depends on the effect size ($\delta$)
Example - Blood Pressure

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg.

We are interested in finding out if the average blood pressure of employees at a certain company is greater than the national average, so we collect a random sample of 100 employees and measure their systolic blood pressure. What are the hypotheses? We’ll start with a very specific question – “What is the power of this hypothesis test to correctly detect an increase of 2 mmHg in average blood pressure?”
Calculating power

The preceding question can be rephrased as – How likely is it that this test will reject $H_0$ when the true average systolic blood pressure for employees at this company is 132 mmHg?

Let’s break this down into two simpler problems:

1. **Problem 1:** Which values of $\bar{x}$ represent sufficient evidence to reject $H_0$?

2. **Problem 2:** What is the probability that we would reject $H_0$ if $\bar{x}$ had come from $N\left(mean = 132, SE = \frac{25}{\sqrt{100}} = 2.5\right)$, i.e. what is the probability that we can obtain such an $\bar{x}$ from this distribution?
Problem 1

Which values of \( \bar{x} \) represent sufficient evidence to reject \( H_0 \)?

(Remember \( H_0 : \mu = 130, \ H_A : \mu > 130 \))
Problem 2

What is the probability that we would reject $H_0$ if $\bar{x}$ did come from $N(\text{mean} = 132, SE = 2.5)$. 

Putting it all together

Systolic blood pressure

Null distribution
Putting it all together

Decision errors

Power

Null distribution

Power distribution

Systolic blood pressure

120 125 130 135 140
Putting it all together

Null distribution

Power distribution

Systolic blood pressure

120 125 130 135 140

0.05
Putting it all together
Putting it all together

Decision errors

Systolic blood pressure

Null distribution

Power distribution

Power

Statistics 102 (Colin Rundel)
Lec 9
February 13, 2013
Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

1. Increase the sample size.
2. Decrease the standard deviation of the sample, which essentially has the same effect as increasing the sample size (it will decrease the standard error). With a smaller $s$ we have a better chance of distinguishing the null value from the observed point estimate. This is difficult to ensure but cautious measurement process and limiting the population so that it is more homogenous may help.
3. Increase $\alpha$, which will make it more likely to reject $H_0$ (but note that this has the side effect of increasing the Type 1 error rate).
4. Consider a larger effect size. If the true mean of the population is in the alternative hypothesis but close to the null value, it will be harder to detect a difference.
Recap - Calculating Power

- Begin by picking a meaningful effect size $\delta$ and a significance level $\alpha$
- Calculate the range of values for the point estimate beyond which you would reject $H_0$ at the chosen $\alpha$ level.
- Calculate the probability of observing a value from preceding step if the sample was derived from a population where $\bar{x} \sim N(\mu_{H_0} + \delta, SE)$
Example - Calculating power for a two sided hypothesis test

Going back to the blood pressure example, what would the power be to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is different from 130 mmHg at a 95% significance level for a sample of 625 patients?

Step 1:

\[ H_0 : \mu = 130, \quad H_A : \mu \neq 130, \quad \alpha = 0.05, \quad n = 625, \quad \sigma = 25, \quad \delta = 4 \]

Step 2:

\[ P(Z > z \text{ or } Z < -z) < 0.05 \quad \Rightarrow \quad z > 1.96 \]

\[ \bar{x} > 130 + 1.96 \frac{25}{\sqrt{625}} \quad \text{or} \quad \bar{x} < 130 - 1.96 \frac{25}{\sqrt{625}} \]

\[ \bar{x} > 131.96 \quad \text{or} \quad \bar{x} < 128.04 \]

Step 3:

\[ \bar{x} \sim N(\mu + \delta, SE) = N(134, 1) \]

\[ P(\bar{x} > 131.96 \text{ or } \bar{x} < 128.04) = P(Z > [131.96 - 134]/1) + P(Z < [128.04 - 134]/1) \]

\[ = P(Z > -2.04) + P(Z < -5.96) = 0.979 + 0 = 0.979 \]
Example - Using power to determine sample size

Going back to the blood pressure example, how large a sample would you need if you wanted 90% power to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is different from 130 mmHg at a 95% significance level?

Step 1:
\[ H_0 : \mu = 130, \quad H_A : \mu \neq 130, \quad \alpha = 0.05, \quad \beta = 0.10, \quad \sigma = 25, \quad \delta = 4 \]

Step 2:
\[ P(\bar{X} > \mu + \delta \sqrt{n} / \sigma \text{ or } \bar{X} < \mu - 4 \sqrt{n} / \sigma) < 0.05 \quad \Rightarrow \quad \sigma > 1.96 \]
\[ \bar{X} > 130 + 1.96 \frac{25}{\sqrt{n}} \text{ or } \bar{X} < 130 - 1.96 \frac{25}{\sqrt{n}} \]

Step 3:
\[ \bar{X} \sim N(\mu + \delta, SE) = N(134, 25/\sqrt{n}) \]
\[ P \left( \bar{X} > 130 + 1.96 \frac{25}{\sqrt{n}} \text{ or } \bar{X} < 130 - 1.96 \frac{25}{\sqrt{n}} \right) = 0.9 \]
\[ P \left( Z > 1.96 - 4 \frac{\sqrt{n}}{25} \text{ or } Z < -1.96 - 4 \frac{\sqrt{n}}{25} \right) = 0.9 \]
Example - Using power to determine sample size (cont.)

So we are left with an equation we cannot solve directly, how do we evaluate it?

For \( n = 410 \) the power \( = 0.8996 \), therefore we need 411 subjects in our sample to achieve the desired level of power for the given circumstance.