There are two problems with using data to prove a hypothesis:

1. It is always possible, no matter what the data results, that any apparent differences between our groups are just a result of data sampling
   - The tumor sizes in the mouse example are approximately normally distributed
   - What if the three treated with the drug just happened to be to the left of the mean, and the three controls just happened to lie to the right?
   - We can only really talk about the chances of this happening

2. Even the question “Given my data, how likely is my hypothesis?” cannot be answered
   - The best we can do is: “If we assume the hypothesis is false, how likely is it I would see data like this?”
   - This is not a proof by contradiction, but with an estimate of the degree of uncertainty!

A proof by contradiction

- Assume \( \sqrt{2} \) can be written as a fraction:
  - Can assume \( n \) and \( m \) have no common factors (i.e., fraction is not in lowest terms)
  - Then \( m^2/2m^2 \), so \( n^2=2m^2 \)
  - So \( n^2 \) is even. This means \( n \) is even, so we can write \( n=2k \) for some integer \( k \).
  - This means \( m^2 \) is even, so \( m \) is even.
  - Now \( m \) and \( n \) are both even, so they have a common factor (fraction is not in its lowest terms)
  - Since our assumption led to a contradiction, the assumption has to be false
  - So \( \sqrt{2} \) cannot be written as a fraction

The problem with proving hypotheses

- There are two problems with using data to prove hypotheses in this manner:
  - About 2,600 years ago, the Greek mathematician/philosopher Pythagoras discovered that \( \sqrt{2} \) cannot be written as a fraction
  - How do we know this for sure?
    - Certainly cannot try all possible fractions
    - Have to construct a mathematical proof
    - Classic example of a proof by contradiction
    - Assume \( \sqrt{2} \) can be written as a fraction
      - Show this results in a contradiction

How statisticians think

- Inferential statistics is often concerned with using data to “prove” a hypothesis
  - Called hypothesis testing
- Example:
  - Take 6 mice and initiate tumor development in them
  - Treat 3 with an experimental drug
  - Measure the size of the tumors after some fixed time
  - Use the data to “prove” the hypothesis that the drug inhibits tumor growth

Simple example

- Suppose you toss a coin 20 times, and 16 times it comes up heads.
- What are the chances of that?
  - Normally that question is asked rhetorically
  - But in a scientific context we ask that question about our observations all the time, and expect a quantitated answer
    - Implicit in the question “What are the chances of that” the assumption that the coin is fair (i.e. there is a 50% chance of a head on any given throw):
    - “What are the chances of that assuming that the coin is fair?”

How Mathematicians Think

- About 2,600 years ago, the Greek mathematician/philosopher Pythagoras discovered that \( \sqrt{2} \) cannot be written as a fraction
- How do we know this for sure?
  - Certainly cannot try all possible fractions
  - Have to construct a mathematical proof
  - Classic example of a proof by contradiction
  - Assume \( \sqrt{2} \) can be written as a fraction
    - Show this results in a contradiction

BMS 617

Lecture 4 – p-values, Statistical Significance, and Hypothesis Testing
Hypothesis Testing

- We can reword the coin toss example in terms of hypothesis testing
- The hypothesis we wish to test is “This is not a fair coin”
- We want to use the data (16 heads from 20 tosses) to support our hypothesis
- To do this, we assume the coin is fair (the opposite of what we want to prove) and compute the probability of getting data got” under this assumption
- This probability is called a p-value

The null hypothesis

- The assumption we make in computing the p-value is called the null hypothesis
- It is typically what we would expect to be true without any data to the contrary
- In the coin-tossing example, the null hypothesis is “The coin is fair”
- In the example with the mouse tumors, the null hypothesis is “There is no difference in tumor size between the treated and untreated mice”
- In most scientific applications, you are trying to prove the null hypothesis is false

The p-value

- Given a null hypothesis and a set of data, the p-value is the probability of seeing data at least as extreme as the observed data, assuming the null hypothesis is true

Example: coin tosses

- For the coin toss example:
  - the null hypothesis is “For each toss of the coin, there is a 50% chance of heads and a 50% chance of tails”
  - the observed data is that we saw 16 heads in 20 tosses of the coin
  - to compute the p-value, we must calculate the probability of at least 16 heads, or at least 16 tails, in 20 coin tosses
    - This is a two-tailed p-value: the null hypothesis is “the coin is fair”
    - If the null hypothesis were “the coin does not favor heads”, we would compute the one-tailed p-value, which would be the probability of 16 or more heads.

p-value for coin tosses

- The p-value for the coin toss example is computed from the binomial distribution
  - p-value in this case is 0.0119, or 1.19%
  - The interpretation is If the coin were fair, there is a 1.19% chance of seeing at least this many heads, or this many tails, in 20 coin tosses

Is the null hypothesis true?

- Given these data, there are two possibilities:
  - by chance
    - The coin is not fair
  - Which is true?
    - Impossible to know!
  - Which is most likely?
    - It depends on the context
Example: Body Temperature

- Motulsky uses an example of measured body temperature in two samples
  - One sample of 130 people
  - One sample of 12 people
- For the larger sample, the mean was 36.82°C, with SD 0.41°C
- For the smaller sample, the mean was 36.77°C, with SD 0.40°C
- Average body temperature is often stated as 37.0°C. Note that these samples have mean which is different to that

p-values for body temperature data

- We can calculate p-values for these data, under the null hypothesis that the average body temperature is 37.0°C
- The p-value for the larger sample is the probability that the average body temperature of 130 randomly selected people would be at least 0.18°C different to 37.0°C, assuming the population average is 37.0°C

Computing the p-value for body temperature data

- To compute the p-value for body temperature data, we use a one sample t-test:
  - Compute the difference between the sample mean and the hypothetical mean: 36.82-37.0=-0.18C
  - Divide this by the standard error of the mean: -0.18C/0.0365=-5.0
  - Under the assumption, this quantity is distributed according to the t-distribution with 129 degrees of freedom
  - The probability of a value less than -5.0 or greater than 5.0 in this distribution is 1.8x10^-6 (0.0000018)

Interpretation of the p-value

- If average body temperature is 37.0°C, and these 130 subjects are representative of the population (selected at random), then the chances of seeing a discrepancy of at least 0.18°C from the average are 0.0000018 (0.00018%)
- To further interpret this, we need to ask 3 questions:
  -
  -
  -

p-value for the smaller sample

- If we repeat this calculation using the smaller sample, we find p=0.0687
- What is the interpretation of this p-value?
- Why is the p-value so different to the previous one?

One- and two-tailed p-values

- In computing the p-values for body temperature, we computed two-tailed p-values:
  - The probability of the t-value being less than -5.0 or more than 5.0
  - i.e. the probability that the mean in our sample was at least 0.18C different to 37.0C
- If we’d computed a one-tailed p-value, it would be the probability that the mean in our sample was at least 0.18C less than 37.0C
Statistical significance in practice

In practice, many scientists use phrases like “Significant”, “Highly Significant”, “Extremely Significant”, or “Borderline Significant.”

- Remember that these all refer to the statistical meaning of the word significant
- The chances of observing these experimental results, under the assumption that the null hypothesis holds.
- Still doesn’t say anything about the importance of the result
- Often, numerical ranges are defined for these terms and a system of stars is used to denote these levels on a graph
- These are non-standard: make sure you know what the threshold values are for the different levels

One-tailed p-values

- Compute a one-tailed p-value only when it is impossible for the value to be in a specific direction
- Ask yourself the question “if the values I got were in the ‘wrong’ direction, would I consider it just due to sampling/experimental error?”
- If not, then use a two-tailed p-value
- If in doubt, use a two-tailed p-value

Hypothesis testing and statistical jargon

- Notice that statistics give special meaning to the following terms:
  - A specific process involving choosing a level of statistical significance, performing an experiment and a statistical test, and comparing the p-value to the significance level
  - Different to the general idea of testing a scientific hypothesis
  - Merely means that if the null hypothesis were true, the results observed would be less likely than some predetermined value
  - Gives evidence to reject the null hypothesis
  - Says nothing about the importance of the result

Strict interpretations of statistical significance

- In a strict interpretation of statistical significance, there are only two possible outcomes to testing a hypothesis:
  - The result is statistically significant
  - The result is not statistically significant
- The p-value is at least as big as α
- In this strict interpretation, if ‘α’ is set at 0.05, there is no difference between a p-value of 0.04 and a p-value of 0.00000001
- They are both statistically significant
- However, there is a big difference between a p-value of 0.049 and 0.051
  - The first is statistically significant, while the second is not

Hypothesis Testing

- Based objective mechanism for making a decision.
  - E.g. Make a decision as to whether a particular line of study is worthy of pursuing
  - Or whether or not a new drug is more effective than current treatment
- Idea is to define a threshold p value
  - Called the significance level of the test
  - Denoted α
  - Commonly set to 0.05
- Then compute the p-value
  - If the p value of the test is less than α, declare the result to be statistically significant and reject the null hypothesis

Type I and Type II Errors

- Remember, hypothesis testing is used to make a decision:
  - Should we reject the null hypothesis?
- There are two types of mistake that can occur when we make this decision:
  - The null hypothesis was true (i.e. there is no real difference (or association, or correlation, etc) between the data from the populations), but random sampling resulted in data which showed a statistically significant difference (or etc).
  - This is a Type I Error (or false positive)
  - The null hypothesis was false (i.e. there really is a difference, etc) but random sampling (and not enough samples) caused the data not to show a statistically significant difference
  - This is a Type II Error (or false negative)
Type I and II Errors Summarized

<table>
<thead>
<tr>
<th>Decision: Reject null hypothesis</th>
<th>Decision: Do not reject null hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reality: Null hypothesis is true</td>
<td>Type I error (false positive) No error (true negative)</td>
</tr>
<tr>
<td>Reality: Null hypothesis is false</td>
<td>No error (true positive) Type II error (false negative)</td>
</tr>
</tbody>
</table>

• Note that for each experiment, you choose the column in this table based on the p-value (and significance level).
• However, you never know which row you are in!
• The smaller the value of α, the less likely it is that a Type I error will occur
• So why not choose a really small value for α?
• The problem is that the smaller α, the harder it becomes to detect true effects
• So Type II errors become more likely
• Choosing α becomes a trade-off between the two types of error
• The only way to decrease the chances of both types of error is to increase the sample size.

What happens in practice

• In practice, scientists rarely weigh up the balance of type I and type II errors.
  – Almost always, α is chosen to be 0.05.
  – The value used by Fisher in his pioneering statistical work
  – Never intended to be a universal standard
• When designing experiments, the goal for statistical power is usually chosen to be 0.8
  – This is the probability of a type II error, assuming the null hypothesis is false
  – We will examine this idea later

Guidelines for statistical significance

• Never use the word “significant” on its own to mean “statistically significant”
  – The statistical meaning of the word is far from its meaning in everyday language
• In fact, ask yourself if you really need to discuss statistical significance at all
  – Perhaps it is better just to report the statistics
  – p-value, confidence interval, etc
  – Leave interpretation to the reader
• If possible, avoid using the word “significant” if it is meant in a non-statistical sense
  – Prone to misinterpretation as “statistically significant”!
  – Use words like “important”, “consequential”, “meaningful”, etc instead

Statistical significance and confidence intervals

• It is usually possible to state a null hypothesis in a manner that specifies a single value:
  – Instead of “The mean expression of the treated sample is the same as the mean expression of the untreated sample”, we could say
  – “The difference between the mean expression of the treated sample and the mean expression of the untreated sample is zero”
  – Notice we now have a single value of interest: the difference between the mean expression from the two groups
• The confidence interval for this value is closely related to the concept of statistical significance for the corresponding hypothesis test

Choosing α

• Ideally, the choice of significance level should be made by balancing the consequences of type I and type II errors.
• Suppose you are testing a new cancer drug. The hypothesis is that it is more effective (has a higher remission rate) than the currently used drug. However, it has also been shown to have serious side effects in a small number of patients.
• What about the same scenario if there is currently no effective treatment?

Statistical Significance and Confidence Intervals

• Compare the two ideas:
  – When computing a 95% confidence interval, we compute a region in which we are 95% certain the true value lies
  – When performing a hypothesis test with a significance level of 0.05, we compute a range which is 95% likely to contain any experimental results if the null hypothesis were true
• These regions must have exactly the same size.
  – The first is centered on the sample mean.
  – The second is centered on the value from the null hypothesis
Example: body temperature data (n=12)

95% CI contains the null hypothesis value

- In the previous graph, the top bar shows the region not considered statistically significant with α=0.05
  - It is the region in which 95% of experimental data would fall if the null hypothesis were true
- The bottom bar shows the 95% confidence interval
  - It is the region in which we are 95% certain the true value lies
- That the 95% confidence interval contains the null hypothesis value is equivalent to the result not being statistically significant, or not rejecting the null hypothesis
  - Note, however, that you can compute a 95% confidence interval without specifying a null hypothesis! You cannot compute a p-value without a null hypothesis.

Example: body temperature data (n=130)

95% CI does not contain the null hypothesis value

- In this example
  - hypothesis value
    - Equivalently, the observed mean lies outside of the region of no statistical significance
- In general:
  - if the 95% confidence interval contains the null hypothesis value, the result is not statistically significant (p>0.05)
  - if the 95% confidence interval does not contain the null hypothesis value, the result is statistically significant (p<0.05)
  - There is nothing special about 95% CIs and α=0.05; the same applies, e.g. to a 99% CI and α=0.01.

Interpreting Statistically Significant Results

- Remember that a result is said to be statistically significant if the p-value is less than some predefined significance level α (usually 0.05)
- This means that, if the null hypothesis were true, these results would be unlikely
  - And that's all it means!
- How important or "significant" the results are is an entirely different question.

Checkpoint

- An investigator performs an experiment, collects the data, and performs a statistical test with an appropriately stated null hypothesis and a significance level of 0.05. The result is statistically significant.
- Is the following statement true? The result is statistically significant, so there is less than a 5% chance that the null hypothesis is true.
Prior Probability and False Discovery Rate

- Suppose we do lots of experiments (we do...)
- In some of these, the null hypothesis will be true, and in some it will not
  - We don’t know which, of course
- So there is a probability (which we can’t measure) of whether any given experiment we attempt is a “true positive” (null hypothesis is false)
  - This is the prior probability
    - We don’t know this, either...
- In some of our experiments, we reject the null hypothesis
  - Some of these will be false positives
  - The proportion of false positives among all positives is called the false discovery rate (FDR)

FDR is not the p-value

<table>
<thead>
<tr>
<th>Decision: Reject null hypothesis</th>
<th>Decision: Do not reject null hypothesis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reality: Null hypothesis is true</td>
<td>A</td>
<td>B + A</td>
</tr>
<tr>
<td>Reality: Null hypothesis is false</td>
<td>C</td>
<td>D + C</td>
</tr>
<tr>
<td>Total</td>
<td>A + C</td>
<td>B + D</td>
</tr>
</tbody>
</table>

- The FDR is the proportion of false positives among all positive results:
  \[ \text{FDR} = \frac{A}{A + C} \]
- The significance level is the probability of a false positive assuming the null hypothesis holds:
  \[ \alpha = \frac{A}{A + B} \]

Prior probability affects FDR

- Suppose we do 1000 experiments
- We set a significance level of 0.05, and we’ll assume we have a statistical power of 80%
  - This is the probability of a false negative, assuming the null hypothesis is false
  - We will discuss statistical power later
- We’ll examine the FDR under assumptions of prior probabilities of 10%, 80%, and 1%

Prior probability of 10%

<table>
<thead>
<tr>
<th>Decision: Reject null hypothesis</th>
<th>Decision: Do not reject null hypothesis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reality: Null hypothesis is true</td>
<td>5% \times 1000 = 45</td>
<td>90% \times 1000 = 950</td>
</tr>
<tr>
<td>Reality: Null hypothesis is false</td>
<td>30% \times 100 = 30</td>
<td>20% \times 100 = 20</td>
</tr>
<tr>
<td>Total</td>
<td>125</td>
<td>875</td>
</tr>
</tbody>
</table>

FDR = 45 / (45 + 875) = 45 / 125 = 36%. Given a positive result, chances the null hypothesis is false are 64%.
### Prior probability of 80%

<table>
<thead>
<tr>
<th>Decision: Reject null hypothesis</th>
<th>Decision: Do not reject null hypothesis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reality: Null hypothesis is true</td>
<td>5% * 200 = 10</td>
<td>95% * 200 = 190</td>
</tr>
<tr>
<td>Reality: Null hypothesis is false</td>
<td>80% * 800 = 640</td>
<td>20% * 800 = 160</td>
</tr>
<tr>
<td>Total</td>
<td>650</td>
<td>350</td>
</tr>
</tbody>
</table>

FDR = 10 / (10+640) = 10/650 = 1.53%. Given a positive result, chances the null hypothesis is false are 98.5%.

### Prior probability of 1%

<table>
<thead>
<tr>
<th>Decision: Reject null hypothesis</th>
<th>Decision: Do not reject null hypothesis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reality: Null hypothesis is true</td>
<td>1% * 990 = 50</td>
<td>99% * 990 = 940</td>
</tr>
<tr>
<td>Reality: Null hypothesis is false</td>
<td>80% * 10 = 8</td>
<td>20% * 10 = 2</td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>942</td>
</tr>
</tbody>
</table>

FDR = 50 / (50+48) = 50/98 = 8%: Given a positive result, chances the null hypothesis is false are 14%.

### Statistical significance is not a measure of “importance”

- In the Mascaro study, the correlation coefficient that was calculated was $r = -0.29$.
  - The interpretation of this is that $(-0.29)^2 = 8.41\%$ of the one variable (care-giving score) can be “explained by” the other variable (testes size), $91.59\%$ of the variability is explained by other factors.
  - Is this a scientifically important result?
    - The authors try to make that case

### Statistical significance is important!

- This does not mean that statistical significance is not important to calculate!
  - Without knowing the p-value, there is no way to know if these results would be likely if the null hypothesis were true
- Merely means that it is not the only measure
- The “effect size” (e.g. correlation coefficient, actual difference between groups, odds ratio, etc.) is also important
- Whether or not there is strong prior evidence for the null hypothesis is also important

### Summary

- The p-value represents the chances of seeing data as extreme as that observed under the assumption that the null hypothesis is true
  - i.e. (usually) under the assumption that what the investigator is trying to prove is false
- A large p-value does not mean the null hypothesis is true
  - but that there is little evidence to reject it
- The p-value does not represent the probability the null hypothesis is true, given the data
  - That is really the FDR
- It is ok – advisable, even – to interpret a p-value in the context of prior evidence to support the null hypothesis
- If the p-value is small enough to reject the null hypothesis, that does not prove the result is important
  - The size of the effect (how much difference, how strong is the correlation, etc.) can help guide you in that

---

Marshall University School of Medicine