Introduction to Hypothesis Testing

January 20th, 2016

Melissa Fazzari, PhD
Department of Biostatistics
Winthrop University Hospital
Goals of this lecture

1. Understand the concepts of hypothesis testing, p-values, Type I and Type II error

2. Get a foundation for lectures 2&3, which will cover common types of statistical tests and study design
Example

Pac Man

Suppose your friend Dan told you that he is an amazing Pac Man expert with a long-run average (LRA) score of 400,000.

As much as you like Dan, you are not sure if you believe him.

Let’s set up a way to formally evaluate Dan’s claim through hypothesis testing. First we need some basic definitions.....
What is Hypothesis Testing?

A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

So first, we need to set up our hypotheses...
Null hypothesis (Ho): A statement that a population parameter equals some claimed value.

Ho: Dan’s LRA Pac Man score = 400,000

Alternative hypothesis (Ha): Our research hypothesis, what we are interested in concluding.

Possible alternative hypotheses:

Ha: Dan’s LRA score is > 400,000 (one-sided)
Ha: Dan’s LRA score is ≠ 400,000 (two-sided)
Ha: Dan’s LRA score is < 400,000 (one-sided)
Hypothesis

• Ho: Dan is an expert at Pac Man with a LRA score of 400,000 (claim)

• Ha: Dan is not an expert at Pac Man and his LRA score is below 400,000

Rejecting the null means that you can conclude that Dan’s score is lower than 400,000 (and call Dan a liar)

Failing to reject the null means that you do not have enough evidence to dispute Dan’s claim.
Decision Rule

• What evidence would you require to reject or not reject Dan’s claim?

• We cannot watch him play every game of Pac Man forever, so we need to watch a sample of games

• Suppose Dan offers to play 10 games in front of you. You will compute his average score over these 10 games
Decision Rule

• Suppose his average score for these 10 games is really poor 120,000 – would you believe his claim?

• Suppose his 10-game average score is 395,000?

• At what point do you make the decision to believe that Dan is an expert player? What is your cutoff?
Decision Rule

- You would probably think to yourself *ok….if Dan’s average score over 10 games is at least 350,000 I will believe him*. That’s your cutoff score for believing Dan.

- After all, you cannot expect Dan to always score an average of 400,000 or more in a set of 10 games. He might have a pulled muscle or be tired or stressed with work or the game’s joy stick may be faulty.

- This kind of decision rule is quite intuitive! But it is also arbitrary.
What is Hypothesis Testing?

What we are really trying to do here is reject the null when the sample statistic (Dan’s 10-game Pac Man average) is unusual enough under the null hypothesis.

If the observed score is unlikely under the null, we can reject the null hypothesis for the entire population (Dan’s true LRA Pac Man score).

How do we quantify “unusual”?
Probability Density Function

• A pdf quantifies the probability of a random variable to take a particular value, for all possible values

• Discrete Example: If it is sunny 70% of the time, and not sunny 30% of the time, the probability function is:

<table>
<thead>
<tr>
<th>Value of the random variable</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>0.70</td>
</tr>
<tr>
<td>Not Sunny</td>
<td>0.30</td>
</tr>
<tr>
<td>Sunny OR Not sunny</td>
<td>Total probability = 1.0</td>
</tr>
</tbody>
</table>
Probability density function under the \textbf{null hypothesis} (Dan is an expert)

Note: The area under the curve sums to 1.0

High probability average scores

Dan had a bad day and hurt his back

Dan had 12 hours sleep and was on top of his game

All possibilities: Dan’s average score in Pac Man over 10 games
PDF under the null hypothesis
(Dan is an expert)

Dan’s average score in Pac Man over 10 games

Rejection region = You call Dan a liar

Your Critical Value
PDF under the null hypothesis
(Dan is an expert)

Dan’s average score in Pac Man over 10 games

Rejection region =
You call Dan a liar

α

Your Critical Value

350,000          400,000
Definition: Type I error ($\alpha$)

- $\alpha$ is also called the **significance level** of the test.

- It is also the probability of making a **Type I error**, or *rejecting the null when the null is true* (Dan truly is an expert, but just had a really bad set of 10 games).

- What is the probability of making a type I error? It is the area under the curve (or the sum of all of the probabilities) associated with all values in the rejection region.
Type I error

• With a critical value of 350,000, we have about a 15% chance ($\alpha=0.15$) to call Dan a liar and reject his Pac Man expertise when he is telling the truth

• Which is fine with you, because he’s just an OK friend and you do not care if he gets mad. Your significance level can be a bit less conservative here...
Decision Rule

• **Now** suppose your mom is making the same claim.

• Suppose she makes the SAME claim as Dan and scores an average of 250,000 in 10 games?

• You can’t call her a liar *quite* so easily, and so you want to be *extra* sure if you reject the null that it is indeed false
Probability function under the null hypothesis (Mom is an expert)

Mom’s average score in Pac Man over 10 games

Rejection region – make it smaller!

\[ \alpha = 0.005 \]
Decision Rule

• We have made the rejection region smaller

• Meaning mom can have an average score above only a pitiful 200,000 and you will believe her and fail to reject the null hypothesis

• The consequence of this is that your alpha level (the area under the curve corresponding to the rejection region) is now much lower, about .5%
Decision Rule

• By making the threshold for statistical significance stricter (i.e. alpha level is lower) we are less likely to reject the null hypothesis.

• The probability of Type I error has been reduced – you are not likely to call mom a liar if she is, in fact, telling the truth about her Pac Man abilities.
Decision Rule

• Remember – we determine the type I error prior to the study

• In most studies we typically limit Type I error to 5%.

• Meaning that you want to set your critical value such that there is at most a 5% chance that you are calling Dan a liar when he is telling the truth.
Probability function under the null hypothesis (Dan is an expert)

Dan’s average score in Pac Man over 10 games

α=0.05

Critical value

Rejection region

Probability

300,000

400,000
Calculate test statistics and p-values

• Suppose now that Dan plays 10 games and his observed average score is 375,000

• 375,000 is not in the rejection region (our critical value assuming a 5% significance level is 300,000) we set up at the start of our experiment so we fail to reject the null at the 0.05 level of significance
Probability distribution function under the null hypothesis (Dan is an expert)

Dan’s average score in Pac Man over 10 games

Rejection region

α = 0.05

300,000  400,000
p-values

What is the actual probability of observing a score of 375,000 or less given that the null is true and Dan is an expert player?

This quantity is called the p-value

Note: This statistical test and p-value is one-sided, since we only care about showing Dan stinks at Pac Man
Probability function under the null hypothesis (Dan is an expert)

Total area in here is the p-value

375,000 = observed average score after 10 games
Dan’s average score in Pac Man over 10 games
The p-value associated with this experiment is obviously greater than 0.05 (375,000 is not in the rejection region)

The probability of Dan scoring a 375,000 or lower when his true average is really 400,000 is about 35%. So it is quite reasonable to not reject Dan’s claim
Probability function under the null hypothesis (Dan is an expert)

- Dan's average score in Pac Man over 10 games: 400,000
- Observed average score after 10 games: 375,000
- Total area in the graph is the p-value, P = 0.35
- Dan’s average score in Pac Man over 10 games: 375,000 = observed average score after 10 games
P-values

• The p-value is computed based on theoretical knowledge of the total area (total probability) under the curve associated with your observed value

• A p-value of $\leq 0.05$ means that the observed value is in the critical region of a test with an alpha level of 5%
Duality: p-values and hypothesis testing

Assuming a test with a 5% significance level:

• P-value ≤ 0.05 -> reject the null
• P-value > 0.05 -> fail to reject the null
Type II error

There is another error we can make – what happens if we fail to reject the null (we decide to believe Dan) when he is actually lying?

How does this happen?

It’s easy to imagine that someone with mediocre skills could have a great set of 10 games just by chance
Probabilities under a specific $Ha$
(Dan’s LRA score = 300,000)

We do not know what the true LRA score is. All we have is an observed average score of 375,000 and it seems to be reasonable to assume it is coming from the distribution under the null.

Dan’s average score in Pac Man over 10 games
Type II error

How can we control Type II error?

1. Larger sample sizes – reduces noise, the variation in Dan’s average score becomes less noisy if you are basing it on 100 games

2. Setting an appropriate significance level for the test - by being generous to mom (believing her unless her score is really low), you have dramatically increased your Type II error probability
How do we figure out these probabilities?

• For continuous outcomes we assume a specific shape (Normal) and certain width (SD) and a certain mean value

• Once we know the distribution, mean and SD the probability distribution is completely defined and thus we can compute a probability for every region we are interested in

• For binary outcomes (example: treatment response) we assume a specific probability for Pr(response)
Review of terms

- Null hypothesis
- Alternative Hypothesis
- One-sided vs. two-sided alternatives
- Probability distribution function
- Critical value
- Significance level of the test
- Type I error
- Type II error
Department of Biostatistics

Melissa Fazzari  mfazzari@winthrop.org
Shah Islam  sislam@winthrop.org
Rose Calixte  rcalixte@winthrop.org