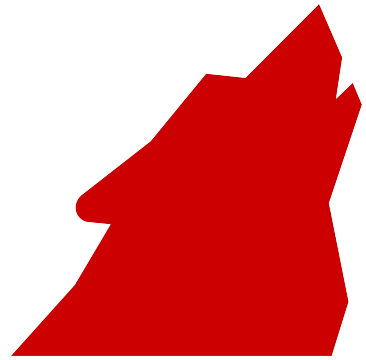
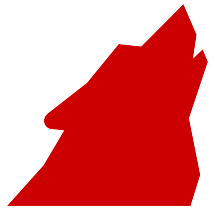


Introduction to p-values and Hypothesis Testing

Eric Yanchenko SEAS Module NC State Spring 2021

Outline



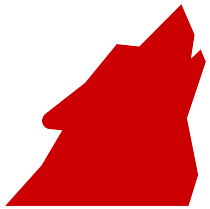
- Motivating Example
- Philosophy of Hypothesis Testing
- p-values Definition and Properties
- Common Misuses
- Some Solutions

Motivating Example

Motivating Example

- You think that most of the time you can get away with speeding on your way to work
- Set up an experiment to test this
- For the next 365 days, you will drive 10 mph over the speed limit and record the number of tickets you get
- At the end of 365 days, decide whether to keep speeding

Motivating Example



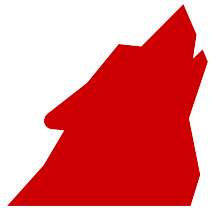
- *Hypothesis*: I can get away with speeding
- *Data*: Number of tickets in 365 days
- *Decision*: Keep speeding or not?
- Questions:
  - Why is this the hypothesis?
  - Is this the best data to collect?
  - How will you analyze the data?
  - How will you make your decision?
  - What are the consequences if you make the "wrong" decision?

Hypothesis Testing Philosophy

Fundamentals of Statistics

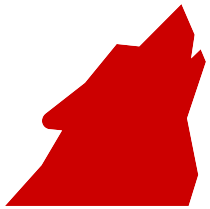
- "Statistics may be defined as a body of methods for making decisions in the face of uncertainty." - W. Allen Wallis
- Paradigm of hypothesis $\rightarrow$ data $\rightarrow$ decision
- Hypothesis testing is a fundamental method

Hypothesis Testing



1. Develop a scientific question that we want to test
    - I can speed without getting caught, smoking causes lung cancer, Drug A is more effective than Drug B, etc.
2. Formulate that hypothesis into something that can be tested statistically
    - Null (status quo) vs. Alternative (what you want to prove)
    - Innocent until proven guilty
    - This should be done carefully!
3. Derive test statistic and its distribution
    - Mean of data, largest observation, scaling of data, etc.
4. Determine decision criteria
    - Reject or fail to reject
5. Each step is deep!

Speeding Example Notation



- θ: *true* probability of being pulled over on any given day. *Unknown parameter.*
- T: $\#$ of tickets. *Test Statistic. Random variable.*
- t^*: $\#$ tickets received. A realization or draw from T. *Data. Observed. Not random.*

We assume that

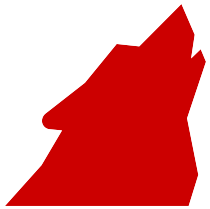
$$T \sim \text{Binomial}(n = 365, \theta)$$

Speeding Example

1. Question: Can I get away with speeding?
2. Hypothesis: H_0: $\theta \geq 0.01$ vs. H_1: $\theta < 0.01$
3. Test statistic: $T = \#$ of tickets
4. Decision rule: Reject H_0 if T is "small".
    - How small is "small"?

P-VALUES

Definition



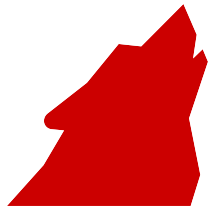
Assume that we are testing $H_o : \theta = 0.01$ vs. $H_1 : \theta < 0.01$ and let T be the test statistic where T comes from some distribution when H_o is true. We collect data and observe t^*. Then

$$p - \text{value} = P_{H_o}(T \leq t^*)$$

- probability of observing a test statistic as extreme or more extreme than we did, assuming null is true
- evidence against the null
- measure of "surprise"

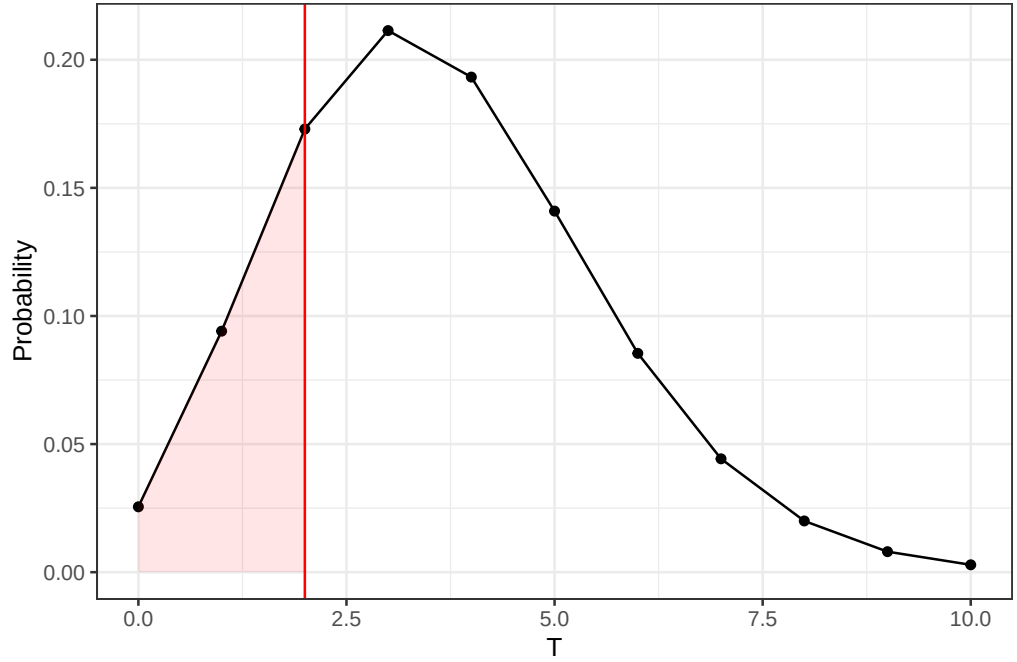
Question: Is the p-value a probability or a statistic?

Speeding Example



We have that $T \sim \text{Binomial}(n = 365, \theta = 0.01)$. We drive for a year and get $t^* = 2$ tickets. Then

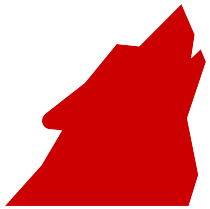
$$p\text{−value} = P_{H_0}(T \leq 2) = P(T = 0) + P(T = 1) + P(T = 2) = 0.29.$$



Is this enough evidence to reject H_0?

Errors

Type I and II Errors

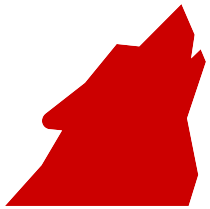


We can make two mistakes in hypothesis testing

- Reject H_0 when H_0 was true (Type I Error, α)
- Fail to reject H_0 when was false (Type II Error, $1 - \beta$)

α is the *level* of the test and is often set to 0.05, though there is nothing special about this value.
β is the *power* of the test, or the probability of correctly rejecting
Want this to be high power and both error rates low

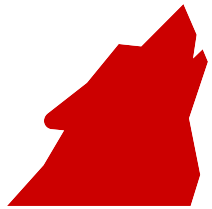
Type I and II Errors



- Reject H_0 when p-value $< \alpha$ (but always report p-value)
- Want to balance Type I and II errors
- Usually we set beforehand α because β is complicated.
- Remember! $\alpha = 0.05$ means there's a 5% chance that we reject H_0 even if H_0 is true

Misuses

Properties



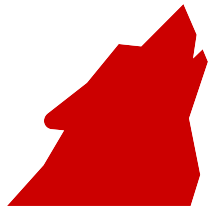
A *p*-value...

- IS: evidence in favor or against the null
- IS NOT: probability that the null is true
- IS: a statistic, i.e. random so changes every time you collect data
- IS NOT: beholden to $\alpha = 0.05$. This is simply a convention (and is probably too lenient)
- IS NOT: as helpful for huge sample sizes
- IS NOT: helpful if your assumptions are wrong
- IS: designed for a *single* test
- IS NOT: designed for multiple testing
- Statistical significance IS NOT equivalent to practical significance

Multiple Testing Problem

- Set up: Looking for a new material with some special property
  - 100 potential materials
- H_0: material is not better than current standard vs. H_1: material is better than current standard
- Test statistic is maybe amount of strain before breaking
- Decide to reject if p-value $< \alpha = 0.05$.

Multiple Testing Problem



- The 87th material has a p-value of 0.03!
- You declare it significant!
- You call up *Nature*!
- You're going to be famous!

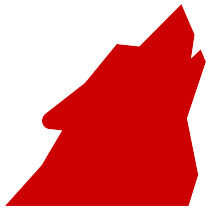
Question: What's the problem with this claim?

Multiple Testing Problem

Let p_i be the p-value for the ith material, $i = 1, 2, \ldots, 100$. For a single material/test,

$$P_{H_0}(\text{reject } i) = P_{H_0}(p_i \leq 0.05) = 0.05.$$

Multiple Testing Problem



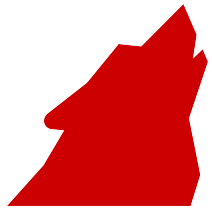
Let p_i be the p-value for the ith material, $i = 1, 2, \ldots, 100$.

$$\begin{aligned} P_{H_0}(\text{reject } i) &= P_{H_0}(p_i \leq 0.05) = 0.05. \\ P_{H_0}(\text{reject at least one } i) &= 1 - P_{H_0}(\text{reject none}) \\ &= 1 - (P_{H_0}(p_i > 0.05))^{100} \\ &= 1 - (0.95)^{100} \\ &= 0.994. \end{aligned}$$

So even if H_0 is true, there's a 99% chance that at least one of the materials will appear significant! Not great! Sounds like we have a *Nature* retraction :(

Adjustments

Corrections



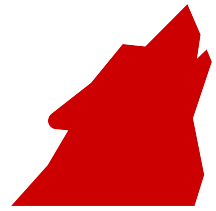
- The problem is that we are testing multiple hypotheses *simultaneously*
- α is designed to control the error for a *single* experiment

Bonferroni Correction

- Most popular method
- Simple, intuitive, quick
- Based on Bonferroni's inequality

Carlo Bonferroni (1892-1960)

Bonferroni Correction



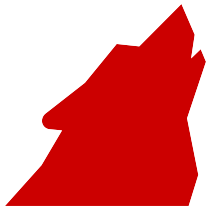
Assume that you are conducting m tests ($m = 100$ in our example) and you want your *family-wise error rate* to be α. Then for each hypothesis i, you reject H_0 if

$$p_i \leq \frac{\alpha}{m}$$

where p_i is the p-value for hypothesis i.

- Divide your significance level α by the number of tests m
- Our example should have been reject H_0 if $p_i \leq 0.05/100 = 0.0005$.
- Note that Bonferroni can be too conservative, i.e. making it too difficult to reject

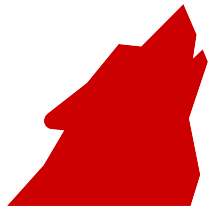
Take Home Message



- Be careful with multiple testing
- Must adjust *p*-values
- Can use Bonferroni correction
- Always report the number of tests and all statistical analyses you conducted
- Don’t draw conclusions based on *p*-values alone

Conclusions

Final Thoughts



- Statistics is about making decisions with uncertainty, something we do every day!
- Statistics is philosophical
- Take time to carefully construct your hypothesis
- Use *p*-values with caution
- Adjust for multiple tests

Thank You!

QUESTIONS?