The Z-test (one-sample categorical data)

Ryan Miller

Introduction

Last week, we introduced **hypothesis testing**, the logic was as follows:

1) Begin with a *null hypothesis* that would be useful to disprove (this has nothing to do with the sample data)
2) Find a suitable *null model* corresponding to that hypothesis (we've focused on Normal models)
3) Use the *p*-value to measure how compatible the observed data are with what would be expected under the null model
4) Make a decision based upon the *p*-value

Introduction

Last week, we introduced **hypothesis testing**, the logic was as follows:

1) Begin with a *null hypothesis* that would be useful to disprove (this has nothing to do with the sample data)
2) Find a suitable *null model* corresponding to that hypothesis (we've focused on Normal models)
3) Use the *p*-value to measure how compatible the observed data are with what would be expected under the null model
4) Make a decision based upon the *p*-value

One challenge in streamlining this approach is that the null model is different for every null hypothesis. . .

The Z-test

- The premise of the Z-test is to *standardize* the hypothesis testing procedure
  - That is, we can standardize the estimate observed in our data relative to what would be expected under the null hypothesis, and then compare this standardized value to the Standard Normal distribution
  - This means the Standard Normal curve will *always* be the null distribution for the Z-test

The Z-test - A Quick Example

Let's see how the Z-test compares to our prior analysis of the *Nature* study where 14 of 16 infants chose the "helper" toy:

General hypothesis test:

- H_0: p $= 0.5$

Z-test:

- H_0: p $= 0.5$

The Z-test - A Quick Example

Let's see how the Z-test compares to our prior analysis of the *Nature* study where 14 of 16 infants chose the "helper" toy:

General hypothesis test:

- H_0: p $= 0.5$
- $\hat{p} \sim N(0.5, \sqrt{\frac{.5(1-.5)}{16}})$

Z-test:

- H_0: p $= 0.5$
- $Z = \frac{14/16-0.5}{\sqrt{.5(1-.5)/16}} = 3$

The Z-test - A Quick Example

Let's see how the Z-test compares to our prior analysis of the *Nature* study where 14 of 16 infants chose the "helper" toy:

General hypothesis test:

- H_0: p $= 0.5$
- $\hat{p} \sim N(0.5, \sqrt{\frac{.5(1-.5)}{16}})$
- Using this model, $Pr(\hat{p} \geq 14/16) = 0.001$
- two-sided p-value of 0.002

Z-test:

- H_0: p $= 0.5$
- $Z = \frac{14/16-0.5}{\sqrt{.5(1-.5)/16}} = 3$
- Using the Standard Normal, $Pr(Z \geq 3) = 0.001$
- two-sided p-value of 0.002

A More Detailed Example

- In genetics, a common question is whether a gene is related an observed trait (phenotype)
  - This is a tough question, as humans have more than 30,000 genes and environmental factors have a large influence of observed traits

A More Detailed Example

- In genetics, a common question is whether a gene is related an observed trait (phenotype)
  - This is a tough question, as humans have more than 30,000 genes and environmental factors have a large influence of observed traits
- A *transmission disequilibrium* study approaches this question using child-parent pairs where the parent was *heterozygous* (one copy of each gene variant)
  - Every child has the trait of interest
  - Because the parent is heterozygous, you'd expect half of the children to have either version of the gene if it's not related to the trait

FP1 and Type 1 Diabetes

- A paper published in *Genetic Epidemiology* studied 124 children with Type 1 diabetes whose parent was heterozygous for the gene FP1
  - 78 of 124 children had the "class 1" version of the FP1 gene, but is this sufficient evidence to link the FP1 gene to Type 1 diabetes?

FP1 and Type 1 Diabetes

- If the gene were unrelated, we'd expect half of the sample to have the FP1 version, or $H_0 : p = 0.5$
  - The null model describing what we'd expect to see in a sample is $\hat{p} \sim N(0.5, \sqrt{.5(1 - .5)/124})$

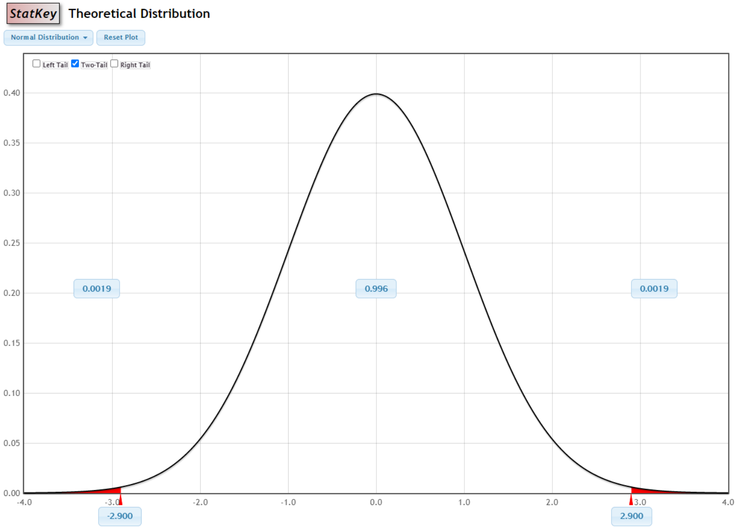
FP1 and Type 1 Diabetes

- If the gene were unrelated, we'd expect half of the sample to have the FP1 version, or $H_0 : p = 0.5$
  - The null model describing what we'd expect to see in a sample is $\hat{p} \sim N(0.5, \sqrt{.5(1-.5)/124})$
- We observed a sample proportion of $\hat{p} = 78/124 = 0.63$
  - This corresponds to a Z-value of $Z = \frac{.63-.5}{\sqrt{.5(1-.5)/124}} = 2.9$

FP1 and Type 1 Diabetes

- If the gene were unrelated, we'd expect half of the sample to have the FP1 version, or $H_0 : p = 0.5$
  - The null model describing what we'd expect to see in a sample is $\hat{p} \sim N(0.5, \sqrt{.5(1-.5)/124})$
- We observed a sample proportion of $\hat{p} = 78/124 = 0.63$
  - This corresponds to a Z-value of $Z = \frac{.63 - .5}{\sqrt{.5(1-.5)/124}} = 2.9$
- We can find the *p*-value by locating this Z-value on the Standard Normal curve

FP1 and Type 1 Diabetes



We can conclude there is strong evidence the FP1 gene and Type 1 diabetes are associated

Summary of the Z-test

1) State the null hypothesis
2) Based upon the null hypothesis, calculate a Z-value describing the sample estimate
3) Locate this Z-value in the Standard Normal curve to find the p-value
4) Use the p-value to make a decision

Next Steps

- We've now thoroughly covered statistical inference for a single proportion (one-sample categorical data)
  - For the remainder of this week, we will practice these concepts using lots of examples
- Next week we will see how things differ (slightly) in situations involving quantitative data