The Basics of Hypothesis Testing

36-600

Hypothesis Testing: Basic Idea

- A *hypothesis test* is an evaluation of a pre-conceived notion (or *null hypothesis*) about the value of distribution parameter θ
  - e.g., the true mean human height is $H_o : \mu_o = 65$ inches
- Before looking at any data, we specify...
  - the null and alternative hypotheses (where the alternative could be $\theta < \theta_o$, $\theta > \theta_o$, or $\theta \neq \theta_o$)
  - a Type I error, conventionally denoted α, that represents a "tolerance" for rejecting the null hypothesis when it is indeed correct; α is typically set to 0.05
- To perform the hypothesis, we...
  - specify a statistic (e.g., the sample mean $\bar{X}$)
  - specify the sampling distribution for that statistic, *assuming the null is true*
  - determine a *p-value*, which is the probability that we would observe the value of test statistic *or something more extreme*, *assuming the null is true*
  - reject the null hypothesis if $p < \alpha$ (or fail to reject otherwise)
- Note that in a hypothesis test, **we are not proving anything!** A hypothesis test is simply a mechanism by which to make a (perhaps incorrect) decision about the viability of the null hypothesis

Hypothesis Test: Extended Illustrative Example

- In an experiment, I will collect n measurements, where n is my *sample size*
  - H_o (the null hypothesis): the mean of the distribution from which the data are to be sampled is $\mu_o = 10$
  - H_a (the alternative hypothesis): the mean of the distribution from which the data are to be sampled is $\mu > \mu_o$ (I will thus conduct an *upper-tail test*, as opposed to a *lower-tail test* ($\mu < \mu_o$) or a *two-tail test* ($\mu \neq \mu_o$)
  - $\alpha = 0.05$
  - my statistic is the sample mean, $\bar{X}$
  - I assume the individual data are normally distributed with true mean μ and true standard deviation σ, so under the null, $(\bar{X} - \mu_o)/(S/\sqrt{n})$ is sampled from a t distribution for $n - 1$ degrees of freedom (note that this is a statement, and not something I would expect you to just know)

Hypothesis Test: Extended Illustrative Example

- Let's assume that the value of the statistic increases, on average, as the parameter value increases (this is usually true: e.g., the average value of $\bar{X}$ would increase if we increase the normal mean μ)
  - the p-value for an lower-tail test is

    ```
    p <- [cumulative distribution function]([arguments])
    ```

  - the p-value for an upper-tail test is

    ```
    p <- 1 - [cumulative distribution function]([arguments])
    ```

  - the p-value for a two-tail test is

    ```
    p <- 2*min(c([p for lower tail],[p for upper tail]))
    ```

  - if the value of the statistic *decreases* on average as the parameter value increases, we'd simply flip the equations for the lower- and upper-tail tests
- **¡MUY IMPORTANTE! A p-value is not the probability that the null hypothesis is correct!**
  - a hypothesis is an idea, not a random variable, and thus there is no probability associated with it!)

Hypothesis Test: Extended Illustrative Example

- Now I collect the data

```
##  [1]  6.331929  7.457654  7.385734 12.776347  7.768998  3.800809 12.986752
##  [8]  6.902247 17.230201  7.661783 10.010315  9.822167 15.526207 10.563105
## [15] 18.822190 12.649759 15.520070 15.646156 14.798670 12.748517 12.216513
## [22] 13.874132 13.049744  8.117794 11.988662 16.317378  7.056131  5.376409
## [29]  7.451565  7.340971 12.762597  6.768779 21.296951 13.978057 10.495503
## [36]  7.833064 12.610566 21.407899  6.371005  4.883713  7.195884 10.553674
## [43] 23.949660 13.511982 14.573277 11.426570 11.657881 13.290343 10.502460
## [50] 12.506495
```

```
## The sample mean is  11.53551
```

- The p-value is the probability that I would observe the value 11.54, or something greater (i.e., more extreme, in the context of an upper-tail test), if $\mu = \mu_o = 10$
    - in code:

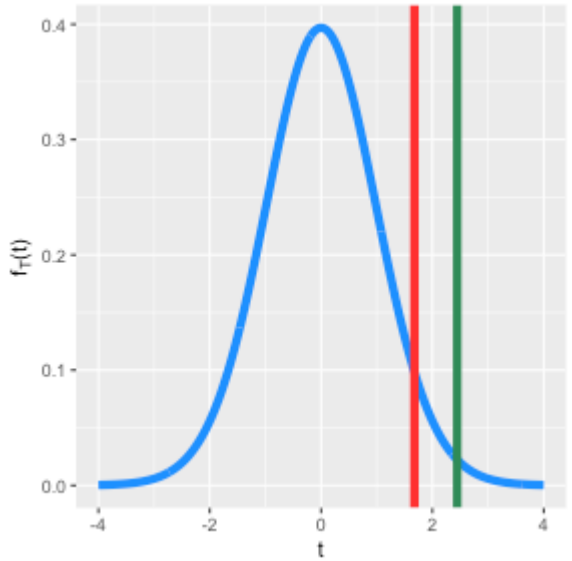
```
n <- length(x)
1 - pt((mean(x)-10)/(sd(x)/sqrt(n)),n-1)
```

```
## [1] 0.009056937
```

- $p < \alpha = 0.05$, so I make the decision to reject the null hypothesis

Hypothesis Test: Extended Illustrative Example

- Let's draw a t distribution for $n - 1 = 49$ degrees of freedom:



- the blue curve is the sampling distribution for $(\bar{X} - \mu_o)/(S/\sqrt{n})$
- the red line is the rejection-region boundary, i.e., the statistic value for which $p = \alpha$ exactly
- the green line is the observed statistic (for which $p = 0.009$)
- For more details about hypothesis testing, see, e.g., Chapters 1 and 2 of this book

Hypothesis Testing: Beware Multiple Comparisons!

- Suppose I perform 100 independent hypothesis tests with Type I error α, and that for all 100 tests, the underlying truth is that H_o is correct
  - p-values are distributed uniformly between 0 and 1 when the null hypothesis is correct, so we expect 5% of the p-values to be less than 0.05
  - thus we will, on average, falsely reject the null hypothesis five times
- Upshot: if you test enough (true!) null hypotheses, you *will* find an apparently significant result
- Testing multiple hypotheses, without correcting for *multiple comparisons*, is *p-hacking*
- Corrections for multiple comparisons include the following:
  - the *Bonferroni correction*: if you run k tests, change α to α/k (this tends to be too conservative)
  - the *False Discovery Rate*: this implements a correction that changes error rate to something between α and α/k (this tends to work well in practice)

Hypothesis Testing for Distributions

- In the extended example above, we assumed that our data were distributed normally
  - what if we want to test that assumption?
- Examples of hypothesis tests for distributions include
  - the *one- and two-sample Kolmogorov-Smirnov (KS) tests*: was a sample of data drawn from a hypothesized continuous distribution?
  - the *Shapiro-Wilk test*: was a sample of data drawn from a normal distribution?