STATISTICS 305/605: Intro to Biostatistics

September 10, 2021

Lecture 5: Review of Statistical Inference: Hypothesis Tests

(Reading: PB Chapters 10, 11, 14.4, 14.6)

1 Goals of Lecture

- We have seen that confidence intervals allow us to quantify the uncertainty in an estimated value of a population parameter.
  - They try to answer questions like, "What values of the parameter might be consistent with the data I have observed?
  - You can use them to "rule out" certain values, at least at a particular level of probability or confidence.
- We often ask questions that are more specific: Is *this* particular value of the population parameter supported by the evidence we have in the data?
  - We could answer this indirectly with a CI
- Some questions involve concepts that can't be represented as a single parameter
  - Is this proposed model consistent with the data?
  - Are the population means for several groups the same, or are some different?
  - Is there a relationship between one variable and a group of other variables?
- We need another construct to help us address yes/no questions we might ask of the data
  - We introduce it in the simple problem where there is a single parameter, so you can understand how it works.

2 Hypothesis Test Concepts

2.1 Hypotheses

- To get started, let's again denote the parameter we are studying by θ
- We want to know whether θ might be equal to a particular value, which I will call θ_0 ("theta-naught")
  - In other words, we are asking the question, "Is $\theta = \theta_0$?"
  - In fact, it might be more interesting to find out that $\theta \neq \theta_0$ (or $\theta > \theta_0$ or $< \theta_0$).
    - If θ represents the effect of some treatment for disease, we really want to know whether the effect is positive, not particularly whether the effect is zero
  - Regardless, one is the complement of the other, so we can address the question from either direction
- For mathematical and logical reasons, we start from the question that specifies a restriction on the value of the parameter, $\theta = \theta_0$
  - We call this restriction the NULL HYPOTHESIS and denote it as "H_0"
  - We call other hypotheses ($\theta \neq \theta_0$, $\theta > \theta_0$, or $\theta < \theta_0$) the ALTERNATIVE HYPOTHESIS and denote it as "H_A"
- Alternative hypotheses of the form $\theta > \theta_0$ and $\theta < \theta_0$ are referred to as ONE-SIDED or ONE-TAILED because only one direction of difference from θ_0 matters.
  - Again, for treatment effects, we will only consider using the treatment if the effect is positive. If it is zero *or* negative, we won't bother with it.
  - In this case, we reformulate the null hypothesis to consider the other direction
    - In treatment-effect example, we would have $H_0 : \theta \leq 0$, $H_A : \theta > 0$
- Alternative hypotheses of the form $\theta \neq \theta_0$ are referred to as TWO-SIDED or TWO-TAILED because both directions of difference from θ_0 matters.
  - If θ is effect of certain foods on mean blood cholesterol measurement in a population of elderly, we might want to know in either case whether the cholesterol goes up or down, so that recommendations can be made about eating less or more of the food.

2.2 Overview of process

- The idea is to hypothesize temporarily that the restriction $\theta = \theta_0$ is true, then ask whether the data says anything to contradict this hypothesis.
  - "Innocent until proven guilty"
- We use the data to estimate the parameter, giving us $\hat{\theta}$.

- We then use the sampling distribution of the estimate to measure "how far" $\hat{\theta}$ is from θ_0 .
  - How likely is it that they would be this far apart?
    - More specifically how *unlikely* it is that a value as extreme as $\hat{\theta}$ could have been observed if the true parameter value is θ_0?
- So center the sampling distribution on θ_0 a
  - If this probability is too small, then we reject the null hypothesis
  - If the probability is not that small, then we don't reject the null hypothesis.
  - This probability is called the P-VALUE
- The p-value is computed in the direction that suggests evidence in favour of the alternative hypothesis
- When we reject the null, we conclude whatever the alternative hypothesis was.
- When we don't reject the null hypothesis, we can't conclude that it was right!
  - There are lots of other values we could try whose hypotheses also would not be rejected
  - We just say that there was not sufficient evidence to contradict the null.

2.3 Errors in hypothesis tests

- Recall that $\hat{\theta}$ is a random variable
- Basing a decision on the value of $\hat{\theta}$ means you are injecting some randomness into the decision
- You can be unlucky with the sample you draw! No matter what the truth is, you could make the wrong decision
  - It is possible to decide to reject H_0 when it is actually true
    - This is called a TYPE 1 ERROR or FALSE POSITIVE
  - It is possible to decide not to reject H_0 when it is actually false
    - This is called a TYPE 2 ERROR or FALSE NEGATIVE
- Given the randomness in the problem, there is a probability associated with each type of error
  - We use α to represent the probability of a type 1 error and β to represent the probability of a type 2 error
  - POWER of a test is the probability of rejecting H_0 when it is, indeed, false (power= $1 - \beta$)

- We would like these probabilities to be small, but they work in opposite directions
  - In order to make α small, we need to make the null hypothesis harder to reject. But if we do this, then we increase the chance that we fail to reject it when we should
  - In order to make β small, we need to make the null hypothesis easier to reject, but this raises α!
- Approach we take is to set one error rate at a small fixed value and try to design data collection and analysis to make other one as small as possible
  - Fix α to something like 0.01, 0.05, or 0.10.
  - Use estimates that have low standard errors to make β as small as it can be
    - Increasing sample size will increase power

3 Hypothesis Test Process

There is a systematic process to a hypothesis test that follows a series of steps:

1. **State question and define parameters**
   - What is the research trying to find out?
     - *Is the mean log-PSA for older men different from the mean log-PSA for younger men among those who are candidates for radical prostatectomy?*
   - If symbols will be used to represent parameters, what do they mean?
     - Don't use symbols that haven't been explained!
     - *Let μ_O be the mean log-PSA for a man 65 or older awaiting radical prostatectomy, and let μ_Y be the corresponding mean log-PSA for men under 65.*
2. **State Hypotheses**
   - Typically in symbols for class problems, but in technical reports the alternative hypothesis of interest is often stated in words.
     - $H_0 : \mu_O = \mu_Y$, *or equivalently,* $H_0 : \mu_O - \mu_Y = 0$
     - $H_A : \mu_O \neq \mu_Y$, *or equivalently,* $H_A : \mu_O - \mu_Y \neq 0$
3. **State type 1 error rate,** α
   - Default is 0.05, like in confidence intervals
   - Larger values are used when sample size is small and power could be very low if α is too small
   - Smaller values can be used when there is a particular reason to want to avoid type 1 errors

4. **State test statistic**

   - Name the statistic that will be used to measure the distance between the observed value $\hat{\theta}$ and the hypothesized value θ_0
   - Choice usually depends on sampling distribution of $\hat{\theta}$
     - When it can be approximated by a normal distribution, then typically use a standardized test statistic that has the form

       $$\frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$$

       where $SE(\hat{\theta})$ is the estimated standard error of the estimate $\hat{\theta}$
     - *We estimate the population means μ_O and μ_Y with the respective sample means $\bar{x}_O$ and $\bar{x}_Y$. Then our estimate of $\mu_O - \mu_Y$ is $\bar{x}_O - \bar{x}_Y$ which we believe has (approximately) a normal distribution, because we have seen that the parent distribution of log-PSA values is mound-shaped. The test statistic is*

       $$t = \frac{(\bar{x}_O - \bar{x}_Y) - 0}{SE(\bar{x}_O - \bar{x}_Y)},$$

       where we will compute the standard error assuming equal population variances in the two groups. This statistic has approximately a t-distribution.

5. **Compute test statistic and p-value**

   - Plug in the numbers and do the calculations.
   - Report both the test statistic value and the p-value

6. **State Decision**

   - If p-value$< \alpha$ then reject H_0
   - Otherwise, if p-value$\geq \alpha$ then do not reject H_0
     - Don't say, "Accept H_0!"

7. **State conclusion**

   - Write a statement that addresses what the test told you about the question.
     - Use words in the context of the question, not symbols.

EXAMPLE: Hypothesis tests for differences of means and proportions (Lecture 4 scripts.R, Prostate.csv)

We now perform "2-sample tests" comparing different population parameters across age groups for the prostate data. We first compare mean values of log-PSA for older and younger men (where 65 is the boundary). Later we compare proportions with PSA>3 between these same groups. In all cases, we use $\alpha = 0.05$. We could also do 1-sample tests if we had particular values we wanted to examine for any of these variables.

Test for differences in population means (Intro R modules 10, 11, and 12) This is the problem that we were using as an example above. We use the function `t.test()` for comparisons of means using a 2-sample t-test. This is the same function that we used for confidence intervals. In fact, the output from the function includes both the test and the confidence interval. The only new thing to notice here is that we can control what value of θ_0 we want to test using `mu=` , which has a default value of 0. We specify it in our code below anyway.

1. **State question and define parameters**

   - *Is the mean log-PSA for older men different from the mean log-PSA for younger men among those who are candidates for radical prostatectomy?*
   - *Let μ_O be the mean log-PSA for a man 65 or older awaiting radical prostatectomy, and let μ_Y be the corresponding mean log-PSA for men under 65.*

2. **State Hypotheses**

   - $H_0 : \mu_O = \mu_Y$, *or equivalently,* $H_0 : \mu_O - \mu_Y = 0$
   - $H_A : \mu_O \neq \mu_Y$, *or equivalently,* $H_A : \mu_O - \mu_Y \neq 0$

3. **State type 1 error rate, α**

   - Set $\alpha = 0.05$

4. **State test statistic**

   - *We estimate the population means μ_O and μ_Y with the respective sample means $\bar{x}_O$ and $\bar{x}_Y$. Then our estimate of $\mu_O - \mu_Y$ is $\bar{x}_O - \bar{x}_Y$ which we believe has (approximately) a normal distribution, because we have seen that the parent distribution of log-PSA values is mound-shaped.* ***The test statistic is***

$$t = \frac{(\bar{x}_O - \bar{x}_Y) - 0}{SE(\bar{x}_O - \bar{x}_Y)},$$

   where we will compute the standard error assuming equal population variances in the two groups. This statistic has approximately a t-distribution with 95 degrees of freedom (total number in two groups, minus 2).

5. **Compute test statistic and p-value**

```
> t.test(x=lpsa.old, y=lpsa.young, mu=0,  var.equal=TRUE)

        Two Sample t-test

data:  lpsa.old and lpsa.young
t = 1.8975, df = 95, p-value = 0.0608
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02030854  0.89849597
sample estimates:
mean of x mean of y
 2.691144  2.252050
```

The test statistic is $t = 1.90$*, and the p-value is* 0.0608

6. **State Decision**

   - *Since* $0.0608 > 0.05 = \alpha$*, we do not reject* H_0*.*

7. **State conclusion**

   - *There is insufficient evidence of a difference in mean log-PSA values between men 65+ or under 65 who are candidates for radical prostatectomy.*

Tests for differences in proportions (Intro R modules 13 and 14) We again analyze the same parameter we made a confidence interval for in the last lecture, the difference in probabilities of PSA>3 between older and younger men. The test is performed using the `prop.test()` function. The function requires the two groups' counts of successes in `x=` and the numbers of trials in `n=` . The test statistic that this function uses is actually the square of the Z test that is described in PB Section 14.4, and it used a CHI-SQUARED sampling distribution with 1 degree of freedom. The p-value for tests—using the argument `alternative=` with values `"two.sided"`, `"greater"`, or `"less"`—is identical to what the Z test would give.

1. **State question and define parameters**

   - *Is the probability that a man's PSA> 3 different for men awaiting radical prostatectomy who are 65 or older compared to those under 65?*
   - *Let* p_O *be the true probability that a man 65 or older awaiting radical prostatectomy has PSA>3, and let* p_Y *be the corresponding probability of PSA>3 for men under 65.*

2. **State Hypotheses**

   - $H_0 : p_O = p_Y$*, or equivalently,* $H_0 : p_O - p_Y = 0$
   - $H_A : p_O \neq p_Y$*, or equivalently,* $H_A : p_O - p_Y \neq 0$

3. **State type 1 error rate,** α

   - Set $\alpha = 0.05$

4. **State test statistic**

- *We estimate the population probabilities p_O and p_Y with the respective sample proportions $\hat{p}_O$ and $\hat{p}_Y$. Then our estimate of $p_O - p_Y$ is $\hat{p}_O - \hat{p}_Y$ which we believe has (approximately) a normal distribution, as proportions often do.* ***The test statistic is***

$$Z^2 = \left[\frac{(\hat{p}_O - \hat{p}_Y) - 0}{SE(\hat{p}_O - \hat{p}_Y)}\right]^2,$$

where we will compute the standard error assuming that the two probabilities are equal. This statistic has approximately a chi-squared with 1 degree of freedom.

5. **Compute test statistic and p-value**

- The `prop.test()` function puts out more information than we need, so I have saved the test output as an object I call `test.2` and just print the test statistic and p-value.

```
> test.2 = prop.test(x=c(sum(hpsa.old), sum(hpsa.young)),
+            n=c(length(hpsa.old), length(hpsa.young)),
+            alternative="two.sided", conf.level=0.95, correct=FALS
Warning message:
In prop.test(x = c(sum(hpsa.old), sum(hpsa.young)), n = c(length(h
:
  Chi-squared approximation may be incorrect
> c(test.2$statistic, test.2$p.value)
  X-squared
7.705391538 0.005505613
> test.2$estimate
   prop 1    prop 2
0.9800000 0.8085106
```

The test statistic is $Z^2 = 7.71$, and the p-value is 0.0055.

6. **State Decision**

- *Since $0.0055 < 0.05 = \alpha$, we reject H_0.*

7. **State conclusion**

- *There is evidence of a difference in probability that PSA>3 between men 65+ or under 65 who are candidates for radical prostatectomy. Looking at the two estimates, it is apparent that older men have a higher change of passing a PSA screening at 3 than younger men do.*

- Notice that both of these tests reach the same conclusions that the respective confidence intervals did.
  - This is no mistake, because the sampling distributions and estimates that the tests are based on are the same as those used for the confidence intervals (other than that `prop.test()` uses Z^2 instead of Z, but they are equivalent procedures)
- This is typically the case when we run tests where the null hypothesis can be written in the form,

$$H_0 : \text{Parameter} = \text{value}$$

  as these are.
- For this reason, most research disciplines no longer rely on hypothesis tests, which were once the workhorse of inferential statistics
  - Confidence intervals can tell you everything a test can (is θ_0 inside or outside the interval)
  - It also gives more information about conceivable values for the parameter, which a test can't do.

4 What to learn from this

1. Hypothesis tests are cousins of confidence intervals
   (a) They focus on asking a question about a specific hypothesized value for the parameter
   (b) They tell you less about the parameter than the CI does.
2. Null and alternative hypotheses are complements of one another.
   (a) The special value θ_0—which may be thought of as the boundary between H_0 and H_A—always goes in H_0
3. We assume that H_0 is true until sufficient evidence demonstrates otherwise.
4. You need to know what statistic R is using for its tests in order to properly report and interpret tests.
5. There is always a chance that you have made the wrong decision
   (a) Larger samples make false negatives less likely
   (b) We fix the probability of false positive (α) by design, so nothing in the data affects this.

5 Exercises

Use R for all calculations, unless otherwise specified.

1. Refer to the Prostate data. Recall that in Lecture 4, we computed a confidence interval for the difference of mean PSA values between older and younger men, again performing calculations directly on PSA values. Now perform a hypothesis test to see whether there is any difference between these two means (HINT: What does this mean about the alternative hypothesis?). Set $\alpha = 0.05$.

   (a) **Report all parts of the test as in the examples.**

   (b) **Compare the results to what we got from the confidence interval. Do the conclusions agree?**

2. Continuing with the Prostate data, recall that we cited a paper that defined "clinically insignificant" prostate cancer. We found a confidence interval for the difference in probability of clinically insignificant prostate cancer between older and younger men in this population. Now perform a hypothesis test for any difference between these probabilities, using $\alpha = 0.05$. **Report all parts of the test as in the examples.**