Mathematical Biostatistics Boot Camp 2: Lecture 1, Hypothesis Testing

Brian Caffo

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

August 4, 2016

Table of contents

Hypothesis Testing

- Hypothesis testing is concerned with making decisions using data
- A null hypothesis is specified that represents the status quo, usually labeled H_0
- The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis

Example

- A respiratory disturbance index of more than 30 events / hour, say, is considered evidence of severe sleep disordered breathing (SDB).
- Suppose that in a sample of 100 overweight subjects with other risk factors for sleep disordered breathing at a sleep clinic, the mean RDI was 32 events / hour with a standard deviation of 10 events / hour.
- We might want to test the hypothesis that
  - $H_0 : \mu = 30$
  - $H_a : \mu > 30$

  where μ is the population mean RDI.

Hypothesis testing

- The alternative hypotheses are typically of the form $<$, $>$ or $\neq$
- Note that there are four possible outcomes of our statistical decision process

	Decision	
Truth	H_0	H_a
H_0	Correctly accept null	Type I error
H_a	Type II error	Correctly reject null

Discussion

- Consider a court of law; the null hypothesis is that the defendant is innocent
- We require evidence to reject the null hypothesis (convict)
- If we require little evidence, then we would increase the percentage of innocent people convicted (type I errors); however we would also increase the percentage of guilty people convicted (correctly rejecting the null)
- If we require a lot of evidence, then we increase the the percentage of innocent people let free (correctly accepting the null) while we would also increase the percentage of guilty people let free (type II errors)

Example

- Consider our example again
- A reasonable strategy would reject the null hypothesis if $\bar{X}$ was larger than some constant, say C
- Typically, C is chosen so that the probability of a Type I error, α, is .05 (or some other relevant constant)
- α = Type I error rate = Probability of rejecting the null hypothesis when, in fact, the null hypothesis is correct

Continued

- Note that

$$\begin{aligned} .05 &= P\left(\bar{X} \geq C \mid \mu = 30\right) \\ &= P\left(\frac{\bar{X} - 30}{10/\sqrt{100}} \geq \frac{C - 30}{10/\sqrt{100}} \mid \mu = 30\right) \\ &= P\left(Z \geq \frac{C - 30}{1}\right) \end{aligned}$$

- Hence $(C - 30)/1 = 1.645$ implying $C = 31.645$
- Since our mean is 32 we reject the null hypothesis

Discussion

- In general we don't convert C back to the original scale
- We would just reject because the Z-score; which is how many standard errors the sample mean is above the hypothesized mean

$$\frac{32 - 30}{10/\sqrt{100}} = 2$$

is greater than 1.645

General rule

- The Z test for $H_0 : \mu = \mu_0$ versus
  - $H_1 : \mu < \mu_0$
  - $H_2 : \mu \neq \mu_0$
  - $H_3 : \mu > \mu_0$
- Test statistic $TS = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- Reject the null hypothesis when

$H_1 : \; TS \leq -Z_{1-\alpha}$
$H_2 : \; |TS| \geq Z_{1-\alpha/2}$
$H_3 : \; TS \geq Z_{1-\alpha}$

Notes

- We have fixed α to be low, so if we reject H_0 (either our model is wrong) or there is a low probability that we have made an error
- We have not fixed the probability of a type II error, β; therefore we tend to say "Fail to reject H_0" rather than accepting H_0
- Statistical significance is no the same as scientific significance
- The region of TS values for which you reject H_0 is called the rejection region

More notes

- The Z test requires the assumptions of the CLT and for n to be large enough for it to apply
- If n is small, then a Gossett's T test is performed exactly in the same way, with the normal quantiles replaced by the appropriate Student's T quantiles and $n - 1$ df
- The probability of rejecting the null hypothesis when it is false is called **power**
- Power is a used a lot to calculate sample sizes for experiments

Example reconsidered

Consider our example again. Suppose that $n = 16$ (rather than 100). Then consider that

$$.05 = P\left(\frac{\bar{X} - 30}{s/\sqrt{16}} \geq t_{1-\alpha,15} \mid \mu = 30\right)$$

So that our test statistic is now $\sqrt{16}(32 - 30)/10 = 0.8$, while the critical value is $t_{1-\alpha,15} = 1.75$. We now fail to reject.

Two sided tests

- Suppose that we would reject the null hypothesis if in fact the mean was too large or too small
- That is, we want to test the alternative $H_a : \mu \neq 30$ (doesn't make a lot of sense in our setting)
- Then note

$$\alpha = P\left(\left|\frac{\bar{X} - 30}{s/\sqrt{16}}\right| > t_{1-\alpha/2,15} \;\middle|\; \mu = 30\right)$$

- That is we will reject if the test statistic, 0.8, is either too large or too small, but the critical value is calculated using $\alpha/2$
- In our example the critical value is 2.13, so we fail to reject.

Connections with confidence intervals

- Consider testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$
- Take the set of all possible values for which you fail to reject H_0, this set is a $(1 - \alpha)100\%$ confidence interval for μ
- The same works in reverse; if a $(1 - \alpha)100\%$ interval contains μ_0, then we **fail to** reject H_0

Proof

- Consider that we do not reject H_0 if

$$\left| \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \right| \leq t_{1-\alpha/2,n-1}$$

implying

$$\left| \bar{X} - \mu_0 \right| \leq t_{1-\alpha/2,n-1} s/\sqrt{n}$$

implying

$$\bar{X} - t_{1-\alpha/2,n-1} s/\sqrt{n} < \mu_0 < \bar{X} + t_{1-\alpha/2,n-1} s/\sqrt{n}$$

P-values

- Notice that we rejected the one sided test when $\alpha = 0.05$, would we reject if $\alpha = 0.01$, how about 0.001?
- The smallest value for alpha that you still reject the null hypothesis is called the **attained significance level**
- This is equivalent, but philosophically a little different from, the **P-value**
- The P-value is the probability under the null hypothesis of obtaining evidence as extreme or more extreme than would be observed by chance alone
- If the P-value is small, then either H_0 is true and we have observed a rare event or H_0 is false

Example

In our example the T statistic was 0.8. What's the probability of getting a T statistic as large as 0.8?

```
pt(0.8, 15, lower.tail = FALSE) ##works out to be 0.22
```

Therefore, the probability of seeing evidence as extreme or more extreme than that actually obtained is 0.22

Notes

- By reporting a P-value the reader can perform the hypothesis test at whatever α level he or she choses
- If the P-value is less than α you reject the null hypothesis
- For two sided hypothesis test, double the smaller of the two one sided hypothesis test Pvalues
- Don't just report P-values, give CIs too!

Criticisms of the P-value

- P-values only consider significance, unlike CIs
- It is difficult with a P-value or result of a hypothesis test to distinguish practical significance from statistical significance
- Absolute measures of the rareness of an event are not good measures of the evidence for or against a hypothesis
- P-values have become abusively used