Hypothesis testing

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

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

TRUTH	DECIDE	RESULT
H_0	H_0	Correctly accept null
H_0	H_a	Type I error
H_a	H_a	Correctly reject null
H_a	H_0	Type II error

Discussion

- Consider a court of law; the null hypothesis is that the defendant is innocent
- We require a standard on the available evidence to reject the null hypothesis (convict)
- If we set a low standard, 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 set a high standard, 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 sleep 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

Example continued

- Standard error of the mean $10/\sqrt{100} = 1$
- Under H_0 $\bar{X} \sim N(30, 1)$
- We want to chose C so that the $P(\bar{X} > C; H_0)$ is 5%
- The 95th percentile of a normal distribution is 1.645 standard deviations from the mean
- If $C = 30 + 1 \times 1.645 = 31.645$
  - Then the probability that a $N(30, 1)$ is larger than it is 5%
  - So the rule "Reject H_0 when $\bar{X} \geq 31.645$" has the property that the probability of rejection is 5% when H_0 is true (for the μ_0, σ and n given)

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
- Or, whenever $\sqrt{n}(\bar{X} - \mu_0)/s > Z_{1-\alpha}$

General rules

- 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
  - $TS \leq Z_\alpha = -Z_{1-\alpha}$
  - $|TS| \geq Z_{1-\alpha/2}$
  - $TS \geq Z_{1-\alpha}$

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)
- The statistic

$$\frac{\bar{X} - 30}{s/\sqrt{16}}$$

  follows a T distribution with 15 df under H_0
- Under H_0, the probability that it is larger that the 95th percentile of the T distribution is 5%
- The 95th percentile of the T distribution with 15 df is 1.7531 (obtained via `qt(.95, 15)`)
- So that our test statistic is now $\sqrt{16}(32 - 30)/10 = 0.8$
- 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 $H_a : \mu \neq 30$
- We will reject if the test statistic, 0.8, is either too large or too small
- Then we want the probability of rejecting under the null to be 5%, split equally as 2.5% in the upper tail and 2.5% in the lower tail
- Thus we reject if our test statistic is larger than `qt(.975, 15)` or smaller than `qt(.025, 15)`
  - This is the same as saying: reject if the absolute value of our statistic is larger than `qt(0.975, 15)` = 2.1314
  - So we fail to reject the two sided test as well
  - (If you fail to reject the one sided test, you know that you will fail to reject the two sided)

T test in R

```
library(UsingR); data(father.son)
t.test(father.son$sheight - father.son$fheight)
```

```
>
>   One Sample t-test
>
> data:  father.son$sheight - father.son$fheight
> t = 11.79, df = 1077, p-value < 2.2e-16
> alternative hypothesis: true mean is not equal to 0
> 95 percent confidence interval:
>  0.831 1.163
> sample estimates:
> mean of x
>     0.997
```

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

Two group intervals

- First, now you know how to do two group T tests since we already covered indepedent group T intervals
- Rejection rules are the same
- Test $H_0 : \mu_1 = \mu_2$
- Let's just go through an example

`chickWeight` data

Recall that we reformatted this data

```
library(datasets); data(ChickWeight); library(reshape2)
##define weight gain or loss
wideCW <- dcast(ChickWeight, Diet + Chick ~ Time, value.var = "weight")
names(wideCW)[-(1 : 2)] <- paste("time", names(wideCW)[-(1 : 2)], sep = "")
library(dplyr)
wideCW <- mutate(wideCW,
  gain = time21 - time0
)
```

Unequal variance T test comparing diets 1 and 4

```
wideCW14 <- subset(wideCW, Diet %in% c(1, 4))
t.test(gain ~ Diet, paired = FALSE,
       var.equal = TRUE, data = wideCW14)
```

```
>
>   Two Sample t-test
>
>  data:  gain by Diet
>  t = -2.725, df = 23, p-value = 0.01207
>  alternative hypothesis: true difference in means is not equal to 0
>  95 percent confidence interval:
>   -108.15  -14.81
>  sample estimates:
>  mean in group 1 mean in group 4
>            136.2           197.7
```

Exact binomial test

- Recall this problem, *Suppose a friend has* 8 *children,* 7 *of which are girls and none are twins*
- Perform the relevant hypothesis test. $H_0 : p = 0.5$ $H_a : p > 0.5$
  - What is the relevant rejection region so that the probability of rejecting is (less than) 5%?

REJECTION REGION	TYPE I ERROR RATE
[0 : 8]	1
[1 : 8]	0.9961
[2 : 8]	0.9648
[3 : 8]	0.8555
[4 : 8]	0.6367
[5 : 8]	0.3633
[6 : 8]	0.1445
[7 : 8]	0.0352
[8 : 8]	0.0039

Notes

- It's impossible to get an exact 5% level test for this case due to the discreteness of the binomial.
  - The closest is the rejection region [7 : 8]
  - Any alpha level lower than 0.0039 is not attainable.
- For larger sample sizes, we could do a normal approximation, but you already knew this.
- Two sided test isn't obvious.
  - Given a way to do two sided tests, we could take the set of values of p_0 for which we fail to reject to get an exact binomial confidence interval (called the Clopper/Pearson interval, BTW)
- For these problems, people always create a P-value (next lecture) rather than computing the rejection region.