Statistics and Probability

Hypothesis testing; Chi-square tests

Truth \ Decision	not rejected	rejected
true H	☺	X
false H	X	☺

http://www.isrec.isb-sib.ch/~darlene/geneve/

Hypothesis testing

- 2 'competing theories' regarding a population parameter:
  - *NULL* hypothesis *H* ('straw man')
  - *ALTERNATIVE* hypothesis *A* ('claim', or theory you wish to test)
- *H:* NO DIFFERENCE
  - any observed deviation from what we expect to see is due to *chance variability*
- *A:* THE DIFFERENCE IS *REAL*

Hypotheses

- Examples of NULL hypothesis:
  - The coin is fair (no preference for heads or tails)
  - This new drug is no better (or worse) than a placebo
  - There is no difference in weight between two given strains of mice
- Examples of Alternative hypothesis:
  - The coin is biased (either towards tail or head)
  - The drug is better than a placebo

Test statistic

- Measure how far the observed data are from what is expected *assuming the NULL H* by computing the value of a *test statistic* (TS) from the data
- The particular TS computed depends on the parameter
- For example, to test the population mean, the TS is the *sample mean* (or standardized sample mean)

Steps in hypothesis testing

1. Identify the population parameter being tested
2. Formulate the NULL and ALT hypotheses
3. Compute the TS
4. Compute the p-value
5. (Optional) *Decision Rule:* REJECT H if the p-value $\le \alpha$
   (This is a type of argument by contradiction)

Example

- An experiment is conducted to study the effect of exercise on the reduction of the cholesterol level in slightly obese patients considered to be at risk for heart attack. 80 patients are put on a specified exercise plan while maintaining a normal diet. At the end of 4 weeks the change in cholesterol level will be noted. It is thought that the program will reduce the average cholesterol reading by more than 25 points.
- Data:
  - sample mean = 27
  - sample SD = 18

Steps in hypothesis testing (I)

1. Identify the population parameter being tested

   Here, the parameter being tested is the population mean cholesterol reading μ

2. Formulate the NULL and ALT hypotheses

   $H: \mu = 25$ (or $\mu \leq 25$)

   $A: \mu > 25$

3. Compute the TS

   $t = (27 - 25)/(18/\sqrt{80}) = .99$

Errors in hypothesis testing

Truth \ Decision	not rejected	rejected
true H	☺ specificity (True -)	X Type I error (False +) α
false H	X Type II error (False -) β	☺ Power $1 - \beta$; sensitivity (True +)

p-value

- Decide on whether or not to *reject* the NULL hypothesis *H* based on the chance of obtaining a TS *as or more extreme* (as far away from what we expected or even farther, in the direction of the ALT) than the one we got, *ASSUMING THE NULL IS TRUE*
- This chance is called the *observed significance level*, or *p-value*
- A TS with a p-value less than some pre-specified false positive *level* (or *size*) α is said to be 'statistically significant' at that level

More on p-values

- Calculate p based on distribution of the TS
- Historically, compare observed to values in a table for pre-defined p-values (e.g. .05, .01)
- Tables specific for the TS you are using
- Computers can now calculate exact p-values, which are reported as output

p-value interpretation

- The interpretation of a p-value is a little tricky
- The p-value represents the chance that we would see a difference as big as we saw (or bigger) *if* there were really nothing happening other than chance variability
- 'a single convenient number giving a measure of the degree of surprise which the experiment should cause a believer of the null hypothesis' (Hodges and Lehmann)

A p-value is ***NOT***

- Confusion about p-values:
  - It is ***not*** the chance that the NULL hypothesis is true
  - It is ***not*** the chance of making an error
  - It is ***not*** the chance that a followup experiment would give a different result
- *Handout:* What's a p-value?

Steps in hypothesis testing (II)

4. Compute the p-value

   Here, p = P(T_{79} > .99) = .16

5. (Optional) *Decision Rule:* REJECT H if the p-value ≤ α.
   (This is a type of argument by contradiction)

   A typical value of α is .05, but there's no law that it needs to be. If we use .05, the decision here will be DO NOT REJECT H

Example: Trees

- A study is conducted to investigate the growth of a certain type of tree at an elevation of 675 meters
- The variable of interest is the core measurement (in cm) for a 10 year period
- The theory is that the mean should be at least 1.75
- In a random sample of 10 measurements, the mean was 2 with an SD of .5

One sample t-test

- Is the mean of a population equal to a given value ?
- *Example*
  - Given a gene and several replicate microarray measurements (log ratios) $g_1, g_2, \ldots, g_n$. Is the gene differentially expressed?
- Hypotheses:
  - NULL: mean equals 0
  - ALT: could be for example
    - mean different from 0
    - mean larger than 0

TS for the one sample t-test

- Test-statistic (Student's t-statistic):

$$T = \frac{\bar{x} - \mu_0}{\sqrt{S^2 / n}}$$

- Here
  - $\bar{x}$ is the average of the observations
  - S is the (estimated) standard deviation
  - n is the number of observations
  - μ_0 is the NULL value (= 0 for our example)

t-distribution

- If the NULL is true, the value T follows a *known distribution* (*t*-distribution)
- The shape of the *t*-distribution depends on the number of observations
- If the average is made of n observations, we use the *t*-distribution with *n-1 degrees of freedom (t_{n-1})*

Steps in hypothesis testing (I)

1. Identify the population parameter being tested

   Here, the parameter being tested is the population mean core measurement μ

2. Formulate the NULL and ALT hypotheses

   H: μ = 1.75 (or $\mu \leq 1.75$)

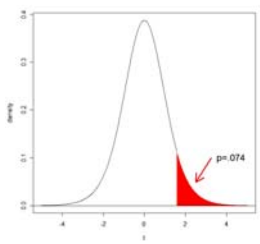
   A: μ > 1.75

3. Compute the TS

   t = (2– 1.75)/(.5/√10) = 1.58

One sample t-test: example

- Since we have 10 observations, compare observed t to *t*-distribution with 9 df



Steps in hypothesis testing (II)

4. Compute the p-value

   Here, p = P(T_9 > 1.58) = .074

5. (Optional) *Decision Rule:* REJECT H if the p-value ≤α

   If we use α = .05, the decision here will be DO NOT REJECT H (but just barely!)

Another example

- To monitor serum total protein level, a series of nine blood tests was run on a particular patient over several days. The normal level for a healthy adult is 7.25.
  - sample mean = 7.28
  - sample SD = 0.03
- Is there sufficient evidence that the total protein level differs from the average? ...

Comments

- The *t*-test assumes that the different observations are *independent* and that they follow a *normal distribution*
- The population SD does not have to be assumed to be 'known'
- If there are many observations, the *t*-distribution is essentially the same as the normal distribution ('*z*-test')

(BREAK)

Variables (review)

- Statisticians call characteristics which can differ across individuals *variables*
- Types of variables:
  - *Numerical*
    - *Discrete* – possible values can differ only by fixed amounts (most commonly counting values)
    - *Continuous* – can take on any value within a range (e.g. any positive value)
  - *Categorical*
    - *Nominal* – the categories have names, but no ordering (e.g. eye color)
    - *Ordinal* – categories have an ordering (e.g. `Always', `Sometimes', 'Never')

Categorical data analysis

- A categorical variable can be considered as a *classification* of observations
- Single classification
  - goodness of fit
- Multiple classifications
  - *contingency table*
  - homogeneity of proportions
  - independence

Mendel and peas

- Mendel's experiments with peas suggested to him that seed color (as well as other traits he examined) was caused by two different 'gene alleles' (he didn't use this terminology back then!)
- Each (non-sex) cell had two alleles, and these determined seed color:

y/y, y/g, g/y →

g/g →

Peas, cont

- Here, yellow is dominant over green
- Sex cells each carry one allele
- Also postulated that the gene pair of a new seed determined by combination of pollen and ovule, which are passed on *independently*

pollen parent			seed parent
y g			y g
yy	yg	gy	gg
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Did Mendel's data prove the theory?

- We know today that he was right, but how good was his experimental proof?
- The statistician R. A. Fisher claimed the data fit the theory *too well*:

  'the general level of agreement beween Mendel's expectations and his reported results shows that *it is closer than would be expected* in the best of several thousand repetitions.... I have no doubt that Mendel was deceived by a gardening assistant, who know only too well what his principal expected from each trial made'
- *How can we measure how well data fit a prediction??*

Testing for goodness of fit

- The NULL is that the data were generated according to a particular chance model
- The model should be *fully specified* (including parameter values); if parameter values are not specified, they may be estimated from the data
- The TS is the *chi-square statistic*:

  χ^2 = sum of [(observed – expected)^2 / expected]
- The χ^2 distribution depends on a number of *degrees of freedom*

Example

- A manager takes a random sample of 100 sick days and finds that 26 of the sick days were taken by the 20-29 age group, 37 by 30-39, 24 by 40-49, and 13 by 50 and over
- These groups make up 30%, 40%, 20%, and 10% of the labor force at the company. Test the hypothesis that age is not a factor in taking sick days ...

Example, contd

Age	Observed	Expected	Difference	χ^2
20-29	26	.3*100=30	26-30=-4	$(-4)^2/30$ =.533
30-39	37			
40-49	24			
≥ 50	13 (total=100)			

- χ^2 = .533 + _____ + _____ + _____ ≈ 2.46
- To get the p-value in R:

```
> pchisq(2.46,3,lower.tail=FALSE)
```


Multiple variables: rxc contingency tables

- A *contingency table* represents all combinations of variable levels for the different classifications
- r = number of rows, c = number of columns
- *Example:*
  - Hair color = Blond, Red, Brown, Black
  - Eye color = Blue, Green, Brown
- Numbers in table represent *counts* of the number of cases in each combination ('*cell*')
- Row and column *totals* are called *marginal counts*

Hair/eye table

Hair \ Eye	Blue	Green	Brown	
Blond	n_{11}	n_{12}	n_{13}	$n_{1.}$
Red	n_{21}	n_{22}	n_{23}	$n_{2.}$
Brown	n_{31}	n_{32}	n_{33}	$n_{3.}$
Black	n_{41}	n_{42}	n_{43}	$n_{4.}$
	$n_{.1}$	$n_{.2}$	$n_{.3}$	*Grand Total* $n_{..}$

cells

row margins

column margins

Hair/eye table for our class

Hair \ Eye	Blue	Green	Brown	
Blond				
Red				
Brown				
Black				

Special Case: 2x2 tables

- Each variable has *2 levels*
- Measures of *association*
  - Odds ratio (cross-product) *ad/bc*
  - Relative risk *[a/(a+b) / (c/(c+d))]*

	+	-	Total
group 1	a (n_{11})	b (n_{12})	$n_{1.}$
group 2	c (n_{21})	d (n_{22})	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Chi-square Test of Independence

- Tests association between two categorical variables
  - NULL: The 2 variables (classifications) are independent
- Compare observed and expected frequencies among the cells in a contingency table
- The TS is the *chi-square statistic* :
  χ^2 = sum of [(observed – expected)^2 / expected]
- df = (r-1) (c-1)
  - So for a 2x2 table, there is *1 df*

Chi-square independence test: intuition

- Construct bivariate table as it would look under the NULL, ie if there were *no association*
- Compare the *real table* to this hypothetical one
- *Measure how different* these are
- If there are *sufficiently large differences*, we conclude that there is a *significant relationship*
- Otherwise, we conclude that our numbers vary just due to chance

Expected frequencies

- How do we find the *expected* frequencies?
- Under the NULL hypothesis of independence, the chance of landing in any cell should be the *product of the relevant marginal probabilities*
- ie, expected number n_{ij}
  $= N*[(n_{i.}/N) * (n_{.j}/N)]$
  $= n_{i.}*n_{.j}/N$

Chi-Square test assumptions

- Data are a *simple random sample* from some population
- Data must be *raw frequencies* (*not* percentages)
- Categories for each variable must be mutually exclusive (and exhaustive)
- The chi-square test is based on a large sample approximation, so the *expected numbers should not be too small* (at least 5 in most cells)

Example

- According to Mendelian theory, genes for color and shape for a variety of pea segregate independently. Round shape is dominant, angular shape is recessive; yellow color is dominant, green color is recessive. A random sample of 200 peas showed the frequency distribution in the table below
  - Rewrite this table as a 2x2 table
  - What is the expected proportion for each group under Mendelian theory?
  - Test whether the data contradict the theory

Pheno	R/Y	R/G	A/Y	A/G
Freq	110	40	42	8

Pitfalls in hypotheses testing

- Even if a result is 'statistically significant', *it can still be due to chance*
- Conversely, if a result is *not* statistically significant, it may be only because you do not have enough data
- *Statistical significance* is not the same as *practical importance*
- A test of significance does not say how *important* the difference is, or *what caused it*
- A test does not check the *study design*
- *Data-snooping* makes p-values hard to interpret: the test should be fully defined **BEFORE** looking at the data