Lecture 7: Binomial Test, Chi-square Test, and ANOVA

May 22, 2012
GENOME 560, Spring 2012

Su-In Lee, CSE & GS
suinlee@uw.edu

1

Goals

- ANOVA
- Binomial test
- Chi-square test
- Fisher's exact test

2

Whirlwind Tour of One/Two-Sample Tests

	Type of Data		
Goal	Gaussian	Non-Gaussian	Binomial
Compare one group to a hypothetical value	One sample t-test	Wilcoxon test	Binomial test
Compare two paired groups	Paired t-test	Wilcoxon test	McNemar's test
Compare two unpaired groups	Two sample t-test	Wilcoxon-Mann-Whitney test	Chi-square or Fisher's exact test

How about 3 or more groups?

3

General Form of a *t*-Test

	One Sample	Two Sample
Statistic	$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$	$t = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$
df	$t_{\alpha, n-1}$	$t_{\alpha, n+m-2}$

4

Non-Parametric Alternatives

- **Wilcoxon Test:** non-parametric analog of one sample t-test
- **Wilcoxon-Mann-Whitney test:** non-parametric analog of two sample t-test

5

Hypothesis Tests of 3 or More Groups

- Suppose we measure a quantitative trait in a group of *N* individuals and also genotype a SNP in our favorite candidate gene. We then divide these *N* individuals into the 3 genotype categories to test whether the average trait value differs among genotypes.
- What statistical framework is appropriate here?
- Why not perform all pair-wise t-test?

6

Do Three Pair-wise *t*-Test?

- This will increase our type I error
- So, instead, we want to look at the pairwise differences "all at once."
- To do this, we can recognize that variance is a statistic that let us look at more than one difference at a time

7

The F-Test

- Is the difference in the means of the groups more than background noise (=variability within groups) ?

Summarizes the mean differences between all groups at once.

$$F = \frac{Variability\ between\ groups}{Variability\ within\ groups}$$

Analogous to pooled variance from a ttest.

8

Basic Framework of ANOVA

- Want to study the effect of one or more ***qualitative*** variables on a ***quantitative*** outcome variable
- Qualitative variables are referred to as ***factors***
  - e.g., SNP
- Characteristics that differentiates factors are referred to as ***levels***
  - e.g., three genotypes of a SNP

9

One-Way ANOVA

- Simplest case, also called single factor ANOVA
  - The *outcome* variable is the variable you're comparing
  - The *factor* variable is the categorical variable being used to define the groups
    - We will assume *k* samples (groups)
  - The *one-way* is because each value is classified in exactly one way
- ANOVA easily generalizes to more factors

10

Assumptions of ANOVA

- **Independence**
- **Normality**
- **Homogeneity of variances**

11

One-Way ANOVA: Null Hypothesis

- The null hypothesis is that the means are all equal

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

- The alternative hypothesis is that at least one of the means is different

12

Motivating ANOVA

- A random sample of some quantitative trait was measured in individuals randomly sampled from population

- Genotype of a single SNP
  - AA: 82, 83, 97
  - AG: 83, 78, 68
  - GG: 38, 59, 55

  There are N (= 9) individuals and k (= 3) groups ...

13

Rational of ANOVA

- Basic idea is to partition total variation of the data into two sources
  - 1. Variation within levels (groups)
  - 2. Variation between levels (groups)

- If H_0 is true the standardized variances are equal to one another

14

The Details

- Our Data:
  - AA: 82, 83, 97 $\quad \bar{x}_{1.} = (82+83+97)/3 = 87.3$
  - AG: 83, 78, 68 $\quad \bar{x}_{2.} = (83+78+68)/3 = 76.3$
  - GG: 38, 59, 55 $\quad \bar{x}_{3.} = (38+59+55)/3 = 50.6$

- Let x_{ij} denote the data from the i[th] level (group) and j[th] observation

- Overall, or ***grand mean***, is:

$$\bar{x}_{..} = \sum_{i=1}^{K}\sum_{j=1}^{J}\frac{x_{ij}}{N}$$

$$\bar{x}_{3.} = \frac{82+83+97+83+78+68+38+59+55}{9} = 71.4$$

15

Partitioning Total Variation

- Recall that variation is simply average squared deviations from the mean

$$\text{SST} = \text{SST}_G + \text{SST}_E$$

$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{i.})^2$$

Sum of squared deviations about the grand mean across all N observations

Sum of squared deviations for each group mean about the grand mean

Sum of squared deviations for all observations within each group from that group mean, summed across all groups

16

In Our Example

$$SST = SST_G + SST_E$$

$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{i.})^2$$

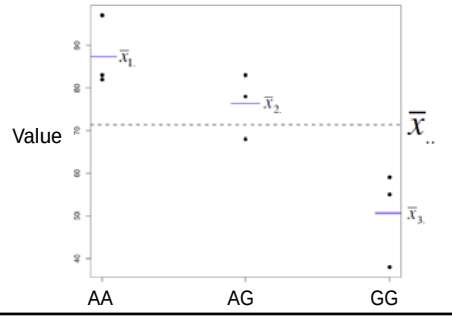
$$(82-71.4)^2+(83-71.4)^2+(97-71.4)^2+ \quad 3\bullet(87.3-71.4)^2+ \quad (82-87.3)^2+(83-87.3)^2+(97-87.3)^2+$$
$$(83-71.4)^2+(78-71.4)^2+(68-71.4)^2+ \quad 3\bullet(76.3-71.4)^2+ \quad (83-76.3)^2+(78-76.3)^2+(68-76.3)^2+$$
$$(38-71.4)^2+(59-71.4)^2+(55-71.4)^2= \quad 3\bullet(50.6-71.4)^2= \quad (38-50.6)^2+(59-50.6)^2+(55-50.6)^2=$$

2630.2 2124.2 506

In Our Example

$$SST = SST_G + SST_E$$

$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.}-\bar{x}_{..})^2 \qquad \sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{i.})^2$$



Calculating Mean Squares

- To make the sum of squares comparable, we divide each one by their associated degrees of freedom
  - SST_G : k – 1 (3 – 1 = 2)
  - SST_E : N - k (9 – 3 = 6)
  - SST_T : N – 1 (9 – 1 = 8)
- MST_G = 2142.2 / 2 = 1062.1
- MST_E = 506 / 6 = 84.3

Almost There… Calculating F Statistics

- The test statistic is the ratio of group and error mean squares

$$F = \frac{MST_G}{MST_E} = \frac{1062.2}{84.3} = 12.59$$

- If H_0 is true MST_G and MST_E are equal
- Critical value for rejection region is $F_{\alpha, k-1, N-k}$
- If we define $\alpha = 0.05$, then $F_{0.05, 2, 6} = 5.14$

ANOVA Table

Source of Variation	df	Sum of Squares	MS	F
Group	k-1	SST_G	$\frac{SST_G}{k-1}$	$\frac{\frac{SST_G}{k-1}}{\frac{SST_E}{N-k}}$
Error	N-k	SST_E	$\frac{SST_E}{N-k}$	
Total	N-1	SST		

21

Non-Parametric Alternative

- Kruskal-Wallis Rank Sum Test: non-parametric analog to ANOVA
- In R, kruskal.test()

22

Whirlwind Tour of One/Two-Sample Tests

	Type of Data		
Goal	Gaussian	Non-Gaussian	Binomial
Compare one group to a hypothetical value	One sample t-test	Wilcoxon test	Binomial test
Compare two paired groups	Paired t-test	Wilcoxon test	McNemar's test
Compare two unpaired groups	Two sample t-test	Wilcoxon-Mann-Whitney test	Chi-square or Fisher's exact test

23

Binomial Data

- Previously, given the following data, assumed to have a normal distribution:

$$x_1,\ x_2,\ \ldots,\ x_n$$

- We were wondering if the mean of the distribution is equal to a specified value μ_0.
- Now, let's consider a different situation…
- Say that we have a binary outcome in each of *n* trials and we know how many of them succeeded
- We are wondering whether the true success rate is likely to be *p*.

24

Example

- Say that you're interested in studying a SNP on a gene associated with Thrombosis. Its expected allele frequency is $p = 0.2$
- In a population with 50 subjects, you know that there are 5 having the mutation
- Then, is p=0.2 the "right" frequency?
- What range of p is *not* going to surprise you?

25

Confidence Limits on a Proportion

- Our question is whether 0.2 is a too frequency to observe 5 mutants (out of 50)
- In R, try:

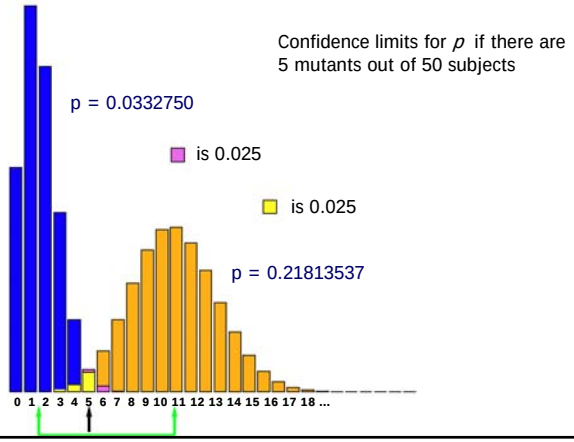
```
> binom.test (5, 50, 0.2)

        Exact binomial test

data:  5 and 50
number of successes = 5, number of trials = 50, p-value = 0.07883
alternative hypothesis: true probability of success is not equal to 0.2
95 percent confidence interval:
 0.03327509 0.21813537
sample estimates:
probability of success
                   0.1
```

26

Confidential Intervals and Tails of Binomials



27

Testing Equality of Binomial Proportions

- How do we test whether ***two populations*** have the same allele frequency?
- This is hard, but there is a good approximation, the **chi-square (χ^2) test**. You set up a 2 x 2 table of numbers of outcomes:

	Mutant allele	WT allele
Population #1	5	45
Population #2	10	35

- In fact, the chi-square test can test bigger tables: R rows by C columns.
- There is an R function chisq.test that takes a matrix as an argument.

28

The Chi-Square Test

- We draw individuals and classify them in one way, and also another way.

	Mutant allele	WT allele	Total
Population #1	5	45	50
Population #2	10	35	45
Total	15	80	95

```
> tb <- matrix( c(5,45,10,35), c(2,2) )
> chisq.test(tb)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tb
X-squared = 1.8211, df = 1, p-value = 0.1772
```

29

How To Do a Chi-Square Test?

- 1. Figure out the expected numbers in each class (a cell in a contingency table). For an m x n contingency table this is (row sum) x (column sum) / (total)

	Mutant allele	WT allele	Total
Population #1	5	45	50
Population #2	10	35	45
Total	15	80	95

$$\frac{50}{95} = 0.5263$$

$$\frac{15}{95} = 0.1578$$

"Expected" number of subjects in pop #1 AND having mutant allele

$$\frac{15}{95} \times \frac{50}{95} \times 95 = 7.8947$$

30

How To Do a Chi-Square Test?

- 1. Figure out the expected numbers in each class (a cell in a contingency table). For an m x n contingency table this is (row sum) x (column sum) / (total)

	Mutant allele		WT allele		Total
Population #1	5	7.895	45	42.105	50
Population #2	10	7.105	35	37.894	45
Total	15		80		95

"observed"

"Expected" number of subjects in each cell

31

How To Do a Chi-Square Test?

- 1. Figure out the expected numbers in each class (a cell in a contingency table). For an m x n contingency table this is (row sum) x (column sum) / (total)
- 2. Sum over all classes:

$$\chi^2 = \sum_{\text{classes}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

  - The number of degrees of freedom is the total number of classes, less one because the expected frequencies add up to 1, less the number of parameters you had to estimate. For a contingency table you in effect estimated (n − 1) column frequencies and (m − 1) row frequencies so the degrees of freedom are [nm−(n−1)−(m−1)−1] which is (n−1)(m−1).
  - Look the value up on a chi-square table, which is the distribution of sums of (various numbers of) squares of normally-distributed quantities.

32

Chi-Square Test

	Mutant allele		WT allele		Total
Population #1	5	7.895	45	42.105	50
Population #2	10	7.105	35	37.894	45
Total	15		80		95

"observed" → Population #1 / Population #2 ← "Expected"

$$\chi^2 = \sum_{\text{classes}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\chi^2_1 = \frac{(5\text{-}7.895)^2}{7.895} + \frac{(45\text{-}42.11)^2}{42.11} + \frac{(10\text{-}7.11)^2}{7.11} + \frac{(35\text{-}37.89)^2}{37.89}$$

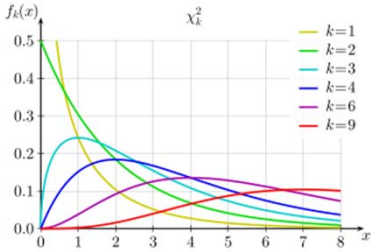
Degrees of freedom = (# rows-1) x (# columns-1) = 1

33

The Chi-Square Distribution

- The Chi-square distribution is the distribution of the sum of squared standard normal deviates.

$$\chi^2_{df} = \sum_{i=1}^{df} Z^2; \text{ where } Z \sim N\ 0,1)$$



- The expected value and variance of the chi-square
  - E(x) = df
  - Var(x) = 2(df)

34

Critical Values

- Here are some critical values for the χ^2 distribution for different numbers of degrees of freedom:

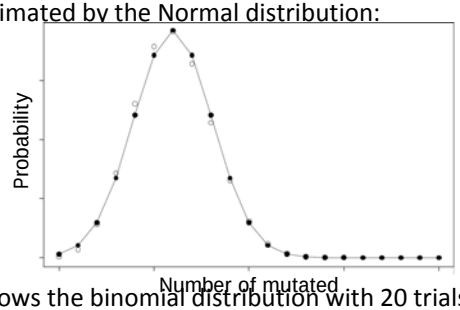
df	Upper 95% point	df	Upper 95% point
1	3.841	15	24.996
2	5.991	20	31.410
3	7.815	25	37.652
4	9.488	30	43.773
5	11.070	35	49.802
6	12.592	40	55.758
7	14.067	45	61.656
8	15.507	50	67.505
9	16.919	60	79.082
10	18.307	70	90.531

- Of course, you can get the correct p-values computed when you use R.

35

The Normal Approximation

- Actually, the binomial distribution is fairly well-approximated by the Normal distribution:



- This shows the binomial distribution with 20 trials and allele frequency 0.3, the class probabilities are the open circles. For each number of heads k, we approximate this by the area under a normal distribution with mean

36