Statistics for Human Genetics and Molecular Biology
Lecture 4: Some Statistical Tools

Dr. Yen-Yi Ho (yho@umn.edu)

Sep 18, 2015

Objectives of Lecture 4

- Categorical Data
  - Tabulating and Plotting Categorical Data
  - Conditional Probability
  - Odds ratio
  - Test for Contingency Tables
  - Cochran-Armitage Trend Test

Summarizing and Presenting Categorical Data

FAMuSS Example

	Genotype			
BMI $>$ 25	AA	GA	GG	Total
0	30	246	380	656
1	30	130	184	344
Total	60	376	564	1000

Plotting Categorical Data

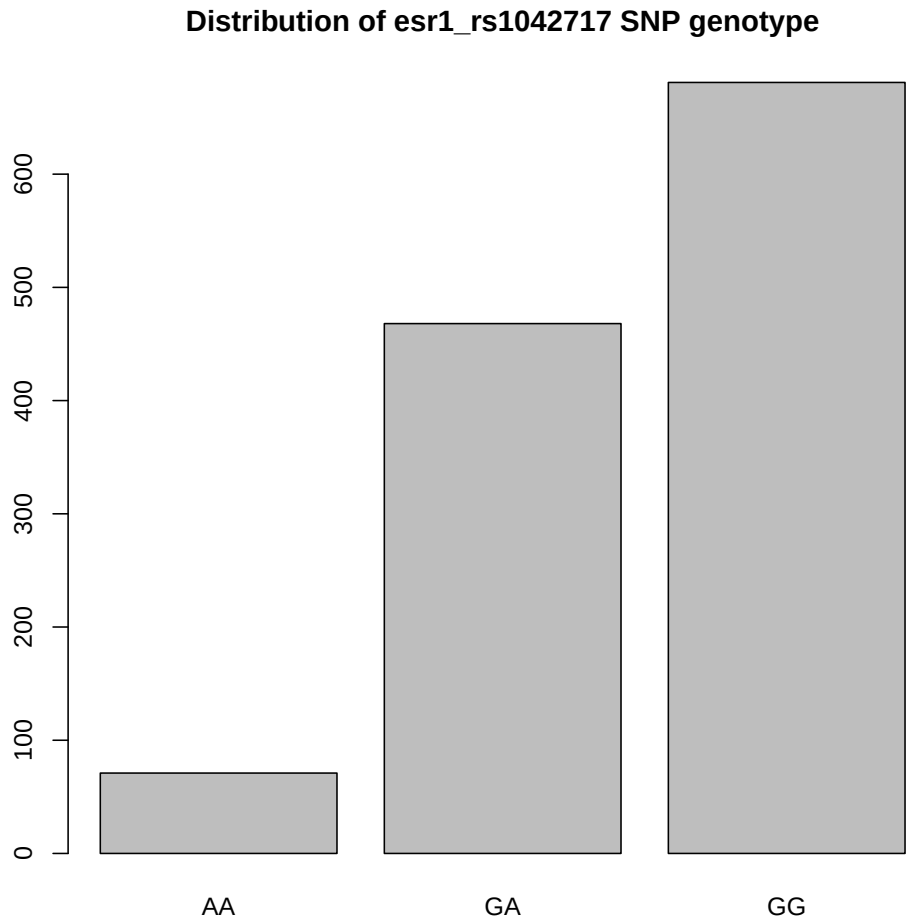
```
fmsURL<-‘‘http://people.umass.edu/foulkes/asg/data/FMS_data
ams <-read.delim(file=fmsURL, header=TRUE, sep="\t")
attach(fms)
Geno <-esr1_rs1042717
trait <- as.numeric(pre.BMI > 25)
table(Geno)
```

Plotting Categorical Data

>str(Geno)

Factor w/ 3 levels "AA","GA","GG": 3 2 3 3 3 3 3 3 3 3 ...

>plot(Geno)



Odds

The odds in favor of an event are the ratio of the probability that the event will happen to the probability that it will not happen.

$$Odds = \frac{p}{1-p}$$

Odds ratio: Measuring Association

	Genotype	
BMI > 25	AA	(GA or GG)
1	a	c
0	b	d
	a+b	c+d

$$\text{Odds of disease among AA} = \frac{\Pr(D^+|E^+)}{[1-\Pr(D^+|E^+)]} = \frac{\frac{a}{(a+b)}}{\frac{b}{(a+b)}} = \frac{a}{b},$$

$$\text{Odds of disease among GA and GG} = \frac{\Pr(D^+|E^+)}{[1-\Pr(D^+|E^+)]} = \frac{\frac{c}{(c+d)}}{\frac{d}{(c+d)}} = \frac{c}{d}.$$

Odds ratio (OR)

	Genotype	
BMI > 25	AA	(GA and GG)
1	30	314
0	30	626
	60	940

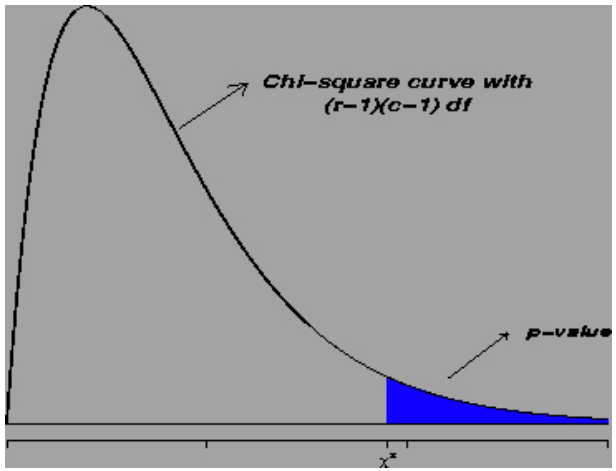
$$\begin{aligned} \text{OR}_{\frac{\text{AA}}{\text{GA and GG}}} &= \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc} \\ &= \frac{30 \times 626}{30 \times 314} \approx 1.99 \end{aligned}$$

Test of Association

Hypothesis: no association between genotype and disease
= Hypothesis : OR=1

$$\chi^2 = \Sigma_{allcells} \frac{(oberseved - expected)^2}{expected}$$

$$\text{p value} = \Pr(\chi^2_{df} > \chi^2_{obs})$$



$\rightarrow$ If p value is small, reject H_0 Hypothesis.

Expected Cell Count

Observed				
	Genotype			
	AA	GA	GG	Total
0	30	246	380	656
1	30	130	184	344
Total	60	376	564	1000

Expected				
	Genotype			
	AA	GA	GG	Total
0	$1000 \times 0.656 \times 0.06$			656
1				344
Total	60	376	564	1000

Degree of freedom

Pearson's χ^2 test for association

Observed			
	Genotype		
	AA	GA	GG
0	30	246	380
1	30	130	184

Expected			
	Genotype		
	AA	GA	GG
0	39.36	246.66	369.98
1	20.64	129.34	194.02

$$\chi^2_{obs} = \frac{(30-39.36)^2}{39.36} + \frac{(246-246.66)^2}{246.66} + \frac{(380-369.998)^2}{369.98}$$
$$+ \frac{(30-20.64)^2}{20.64} + \frac{(130-129.34)^2}{129.34} + \frac{(184-194.02)^2}{194.02} \approx 7.26$$

>chisq.test(tab)

Pearson's Chi-squared test

data: tab, X-squared = 7.2638, df = 2, p-value = 0.02647

Fisher's exact test for association

- Prefer Fisher's exact test when expected cell count < 5.

	Genotype		
	AA	GA	GG
0	30	246	380
1	30	130	184

> fisher.test(tab)

Fisher's Exact Test for Count Data

data: tab, p-value = 0.02941 alternative hypothesis: two.sided

Pearson's Chi-squared test

data: tab, X-squared = 7.2638, df = 2, p-value = 0.02647

Hardy-Weinberg equilibrium

Genotype	Frenquency*
AA	P_A^2
GA	$2 \times P_A P_G$
GG	P_G^2

*Frequencies under Hardy-Weinberg equilibrium assumption

Hardy-Weinberg equilibrium

Genotype	Observed	Expected
AA	60	61.5
GA	376	373.0
GG	564	565.5

$$\chi^2_{obs,df=1} = \frac{(60-61.5)^2}{61.5} + \frac{(376-373)^2}{373} + \frac{(564-565.5)^2}{565.5}$$
$$\approx 0.065$$

```
> library(genetics)
>gt <- genotype(Geno, sep="")
>HWE.chisq(gt)
```

Pearson's Chi-squared test

data: tab, X-squared = 0.065, df = NA, p-value = 0.8052

Cochran-Armitage (C-A) trend test

Detect a linear trend in proportions over levels of exposure variable.

	Genotype		
	AA	GA	GG
0	30	246	380
1	30	130	184

```
> install.packages("coin")
> library(coin)
> GenoOrd <- ordered(Geno)
> independence_test(Trait ~GenoOrd,teststat="quad",
    scores=list(GenoOrd=c(0,1,2)))
```

Asymptotic General Independence Test

data: Trait by GenoOrd (AA< GA < GG)

chi-squared = 4.4921, df = 1, p-value = 0.03405