6.S085 Statistics for Research Projects	**IAP 2014**

Lecture 6: January 28

Lecturer: Ramesh Sridharan and George Chen — *Notes by: William Li*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

6.1 Categorical Data

```
                Outcome
                 1      2
Treatment   1    A      B
            2    C      C (table of counts)
```

Start with inputs and outputs that are categorical

Usually look at two-way table (contingency table)

Risk: of an outcome for a treatment is proportional of data with tha outcome

"risk fo outcome 1 (cancer) for treatment 1 (smoking) is $A/(A+B)$"

Relative risk: $\frac{A/(A+B)}{C/(C+D)}$

Odds ratio: $\frac{A/B}{C/D}$

6.2 Simpson's Paradox

Data from two hospitals on a risky procedure

```
                Live   Die   Survival Rate
Hospital   A     80    120      40%
           B     20     80      20%
```

Hospital B sees more patients than A

```
GOOD PATIENTS

                Live   Die   Survival Rate
Hospital   A     80    100      44%
           B     10     10      50%
```

```
BAD PATIENTS
              Live   Die   Survival Rate
Hospital   A    0     20        0%
           B   10     70       13%
```

Simpson's paradox

Caused by confounding variable at play

6.3 Testing Significance of Categorical Data

We gathered data and avoided confounds.

Question:

Is there a relationship between input and output?

Are they not independent?

Null hypothesis: treatment and outcome are independent

Test:

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

Note: in above example, "hospital" is the treatment and live/die is the outcome

Expected counts for hospitals:

```
A: 2/3
B: 1/3
L: 1/3
D: 2/3

Expected:
     L             D
-------------------------
A   300*A*L     300*A*D

B   300*B*L     300*B*D

     L             D
-------------------------
A    67            133

B    33             67
```

This is how you compute expected counts for independence

Now, we can go back to the χ^2 formula

For hospital example, $\chi^2 = 12$

This test statistic (χ^2) has a chi^2 distribution with $(r-1)(c-1)$ degrees of freedom

We can look it up and see that the p-value is 0.0053; we can reject the null hypothesis that the treatments (hospitals) are independent from the outcome

6.3.1 Why is it χ^2

Each entry is binomial

If the entries are large enough and the samples are indepnedent, each entry is approximately normal

AND the sum of squared standard Gaussians is χ^2

Key assumptions: "entries are large enough" and "samples are independent"

6.3.2 What if the entries are too small?

Fisher's exact test: Works for 2x2 tables

Permutation test

Fisher p-value: $p = \frac{\binom{A+B}{A}\binom{C+D}{C}}{\binom{N}{A+C}}$

Easy if entries are small

Monto Carlo approximation

Yates correction: substract .5 (makes approximation more accurate)

Recall:

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

6.4 Categorical Inputs, Numerical Outputs

6.4.1 Vocabulary

Factor: categorical varaible (e.g. color)

Level: value that the factor takes (e.g. "red")

6.4.2 ANOVA

Start with one factor, k levels

e.g. input data is color: red/blue/yellow

```
Data

MM color    taste score
```

```
R       8.1
R       8.0
B       2.1
Y       0.0
B       1.9
```

Recall, in linear regression:

$y = X\beta + \epsilon$

Now, let's have one predictor per level

$R = (1, 0, 0)$

$B = (0, 1, 0)$

$Y = (0, 0, 1)$

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$y = \begin{pmatrix} 8.1 \\ 8.0 \\ 2.1 \\ 0 \\ 1.9 \end{pmatrix}$$

fit linear regression with X, y to get β

$$\beta = \begin{pmatrix} 8.05 \\ 2.0 \\ 0 \end{pmatrix}$$

Idea: use F-test ("how good is the model?")

$SS_{total} = SS_{model} + SS_{error}$

F-test: how well does the model explain → ANOVA!

Is there a relationship between the input and the output?

The question you are asking: Do the categories predict the outcomes? (Not the difference in the categories)

Null hypothesis: $\beta_1 = \beta_2 = ... = 0$

6.4.3 What assumptions are we making?

- errors independent → data must be independent
- errors must be normally distributed → data in each group must be normally distributed

- All errors have the same variance *rightarrow* all groups (categories) have the same variance (homoskedasticity)

$\hat{\epsilon} = y - \hat{y}$ (residual)

ϵ (error)

6.4.4 Two-Way ANOVA

Two input factors

Additive (no interaction)

$$X = \begin{pmatrix} R & B & Y & sq & tr & st \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$