

Sequential Analysis & Testing Multiple Hypotheses,

CS57300 - Data Mining
Spring 2016

Instructor: Bruno Ribeiro

This Class:

- Sequential Analysis
- Testing Multiple Hypotheses
- Nonparametric Tests
- Independence Tests

Sequential Analysis

Sequential Probability-ratio Test

- ▶ Inspector needs max of 5 broken bulbs out of 100 to reject lot
- ▶ We can clearly stop after $k < 100$ bulbs if we observe 5 broken ones
- ▶ But if $k = 10$ and we have observed 4 broken bulbs.
Aren't we confident we can stop?

Motivation: The New York Times Dilemma

- ▶ Select 50% users to see headline A
 - Titanic Sinks

- ▶ Select 50% users to see headline B
 - Ship Sinks Killing Thousands

- ▶ Hypothesis?
 - $H_0: \text{ \% page views A} = \text{ \% page views B}$
- ▶ Assign half the readers to headline A and half to headline B?

If A is **much** better than B then we could reject hypothesis H_0 quickly



Reject hypothesis before end of experiment
(see slide 11 for context)

- ▶ Early hypothesis rejection
 - ▶ We should not have to wait to declare hypothesis is rejected
 - ▶ Problem:



Consider experiment order

| Opponent | Tournament | Outcome | Correct |
|-----------|----------------|-----------|--|
| Australia | World Cup 2010 | Correct |  Australia |
| Serbia | World Cup 2010 | Correct |  Serbia |
| Ghana | World Cup 2010 | Correct |  Ghana |
| England | World Cup 2010 | Correct |  England |
| Argentina | World Cup 2010 | Correct |  Argentina |
| Spain | World Cup 2010 | Correct |  Spain |
| Uruguay | World Cup 2010 | Correct |  Uruguay |
| Spain | Euro 2008 | Incorrect |  Spain |
| Croatia | Euro 2008 | Incorrect |  Croatia |
| Poland | Euro 2008 | Correct |  Poland |
| Croatia | Euro 2008 | Incorrect |  Croatia |
| Spain | Euro 2008 | Incorrect |  Spain |

Complete experiment:

$$P[Y_{13} \geq 11 | H_0] = \sum_{k=11}^{13} \binom{13}{k} 0.5^k (1-0.5)^{n-k} = 0.0112 < 0.01$$

First 7 of Paul's predictions:

$$P[Y_7 = 7 | H_0] = 0.5^7 = 0.0078 < 0.01$$

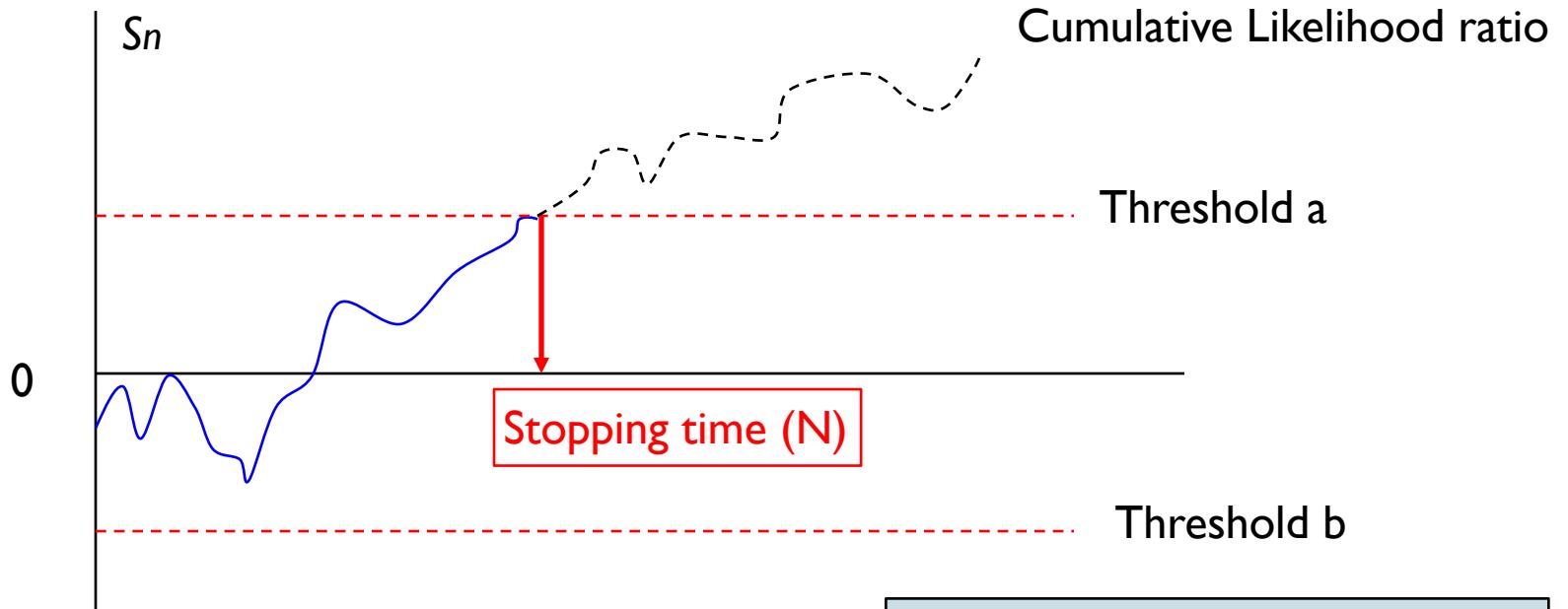
Example Application: Jung et al. 2004

- ▶ Detect whether a remote host is a port scanner or a benign host
- ▶ Ground truth: based on percentage of local hosts which a remote host has a failed connection
- ▶ Example of parameters:
 - for a scanner, the probability of hitting inactive local host is 0.8
 - for a benign host, that probability is 0.1

Wald's Sequential Probability Ratio Test

- ▶ Starts with two hypotheses
- ▶ Example:
 - A remote host attempts to connect a local host at time i
let $X_i = 0$ if the connection attempt is a success,
 1 if failed connection
 - As outcomes X_1, X_2, \dots are observed we wish to determine whether host is a scanner or not
 - Two competing hypotheses:
 - H_0 : remote host is benign: $P[X_i = 1|H_0] = 0.1$
 - H_1 : remote host is a scanner: $P[X_i = 1|H_1] = 0.8$
- ▶ Calculate the cumulative sum of the log-likelihood ratio
 - $S_0 = 0$
 - $S_i = S_{i-1} + \log P[X_i|H_0] - \log P[X_i|H_1]$

Thresholds vs. Errors



- False alarm rate $\alpha = P[R = 1|H_0]$
- Misdetection rate $\beta = P[R = 0|H_1]$

Wald's approximation :

$$a \geq \log \frac{\beta}{1-\alpha} \Rightarrow a \approx \log \frac{\beta}{1-\alpha}$$

$$b \leq \log \frac{1-\beta}{\alpha} \Rightarrow b \approx \log \frac{1-\beta}{\alpha}$$

$$\text{So, } \alpha \approx \frac{1-e^a}{e^b - e^a} \text{ and } \beta \approx \frac{e^{-b} - 1}{e^{-b} - e^{-a}}$$

Hypothesis Testing Example

Paul the Octopus (2008-2010)

- ▶ Paul was an animal oracle
- ▶ Paul's keepers would present him with two boxes containing food
- ▶ Whichever teams is in the box Paul chooses first is the predicted winner



Results involving Germany

| Opponent | Tournament | Outcome |
|-----------|----------------|-----------|
| Poland | Euro 2008 | Correct |
| Croatia | Euro 2008 | Incorrect |
| Austria | Euro 2008 | Correct |
| Portugal | Euro 2008 | Correct |
| Turkey | Euro 2008 | Correct |
| Spain | Euro 2008 | Incorrect |
| Australia | World Cup 2010 | Correct |
| Serbia | World Cup 2010 | Correct |
| Ghana | World Cup 2010 | Correct |
| England | World Cup 2010 | Correct |
| Argentina | World Cup 2010 | Correct |
| Spain | World Cup 2010 | Correct |
| Uruguay | World Cup 2010 | Correct |

Hypothesis Testing Paul the Octopus as an Oracle

- ▶ Random variable (i.i.d.)

$$X_i = \begin{cases} 1 & , \text{ if Paul predicts correct outcome} \\ 0 & , \text{ otherwise} \end{cases}$$

- ▶ Variable of interest: $Y_{13} = \sum_{i=1}^{13} X_i$

- ▶ What is the Null Hypothesis?

- Paul is not an animal oracle
 - Mathematical definition?

- $H_0 := P[X_i = 1] = p = 0.5$

$$\Rightarrow P[Y_{13} = k | H_0] = \binom{13}{k} 0.5^k (1 - 0.5)^{n-k}$$

- ▶ Should we reject H_0 with significance level 0.05? **(one-sided test)**

$$P[Y_{13} \geq 11 | H_0] = \sum_{k=11}^{13} \binom{13}{k} 0.5^k (1 - 0.5)^{n-k} = 0.0112 < 0.05$$



Anything Wrong in our Hypothesis Test?

Hypothesis Test Possible Outcomes

Actual Situation “Truth”

| | | H_0 True | H_0 False |
|------------------------|--|---|-------------|
| Do Not Reject H_0 | $P[H_0 H_0]$ | Type II error (false negative) $P[H_0 \neg H_0]$ | |
| | $1 - \alpha$ | β | |
| Reject H_0 | Type I error (false positive) $P[\neg H_0 H_0]$ | $P[\neg H_0 \neg H_0]$ $1 - \beta$ | |

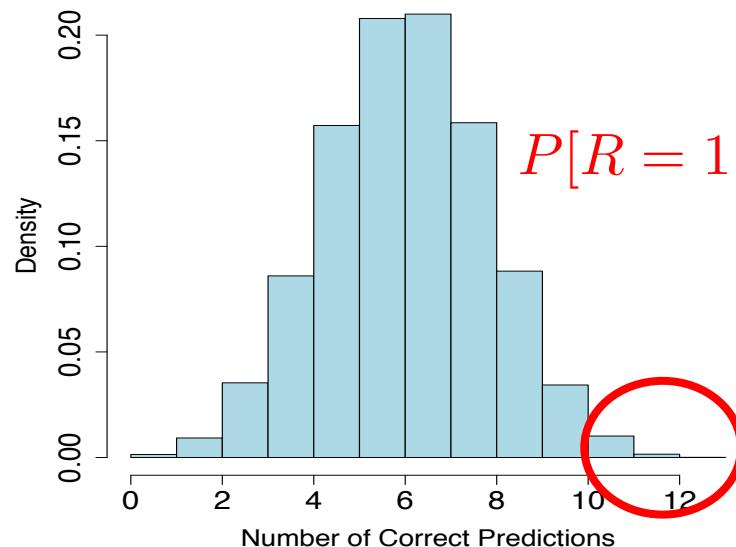
Hypothesis Test as Random Variable

$$X_i = \begin{cases} 1 & , \text{ if Paul predicts correct outcome} \\ 0 & , \text{ otherwise} \end{cases}$$



- ▶ R is random variable that defines if hypothesis is rejected
if $P[Y_{13} \geq k|H_0] < 0.05$ then $R = 1$; otherwise $R = 0$
k correct predictions by animal

Binomial Distribution ($p=0.5$)

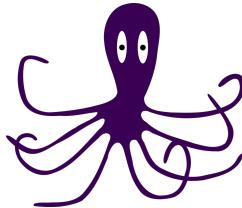


$$P[R = 1|H_0] = P[Y_{13} \geq 10|H_0] = 0.046$$

Testing Multiple Hypotheses

Familywise Error

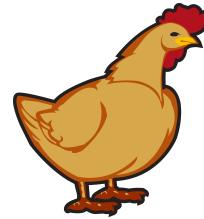
(probability of rejecting a true hypothesis in multiple hypotheses tests)



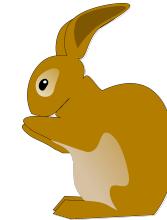
Paul



Peter



Paloma



Philis

- ▶ Probability we reject "not an oracle" hypothesis of Paul based on chance alone?

$$P[R = 1|H_0] = 0.046$$

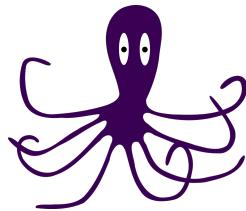
- ▶ Probability we reject "not an oracle" hypothesis of one or more animals (Paul, Peter, Paloma, Philis)

$$1 - (1 - P[R = 1|H_0])^4 = 0.17$$

$P[R=0|H_0]^4$ = Probability we correctly reject all 4 hypotheses

Bonferoni's correction

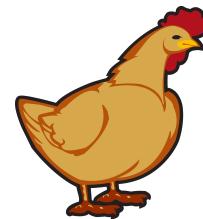
- ▶ Used when there aren't too many hypotheses
- ▶ Tends to be too conservative for large number of hypotheses



Paul



Peter



Paloma



Philis

- ▶ Per-hypothesis significance level of m hypotheses: α/m
- ▶ In our animal oracle example:
 - Old significance level $\alpha=0.05$
 - Bonferoni's corrected significance level $\alpha'=0.05/4 = 0.0125$
 - Hypothesis test: "Paul is not an animal oracle"

$$P[Y_{13} \geq 11 | H_0] = \sum_{k=11}^{13} \binom{13}{k} 0.5^k (1 - 0.5)^{n-k} = 0.0112 < 0.0125$$

- ▶ Effect on $P[H_0 | \neg H_0]$?

False Discovery Rate

- ▶ Often used for large number of tests
- ▶ Bonferoni's correction seeks to ensure that no true hypotheses are rejected
 - Low statistical power for large number of hypotheses (rejects no hypotheses $m \gg 1$)
- ▶ False Discovery Rate:
 - Controls: $mP[\neg H_0 | H_0]$
 - Greater statistical power at expense of more false positives
 - Order p-values of all m tests: $p_1 \leq p_2 \leq \dots \leq p_m$
 - Holm's Method:
 - $\tilde{p}_i = \min((m - i + 1)p_i, 1)$
 - Reject if adjusted p-value $< \alpha$
 - Benjamini-Hochberg method:
 - Reject j null hypothesis if $p_j \leq \alpha \frac{j}{m}$

Back to Single (Two-sample) Hypothesis Tests

Nonparametric Test Procedures

- ▶ Not Related to Population Parameters
Example: Probability Distributions, Independence
- ▶ Data Values not Directly Used
Uses Ordering of Data

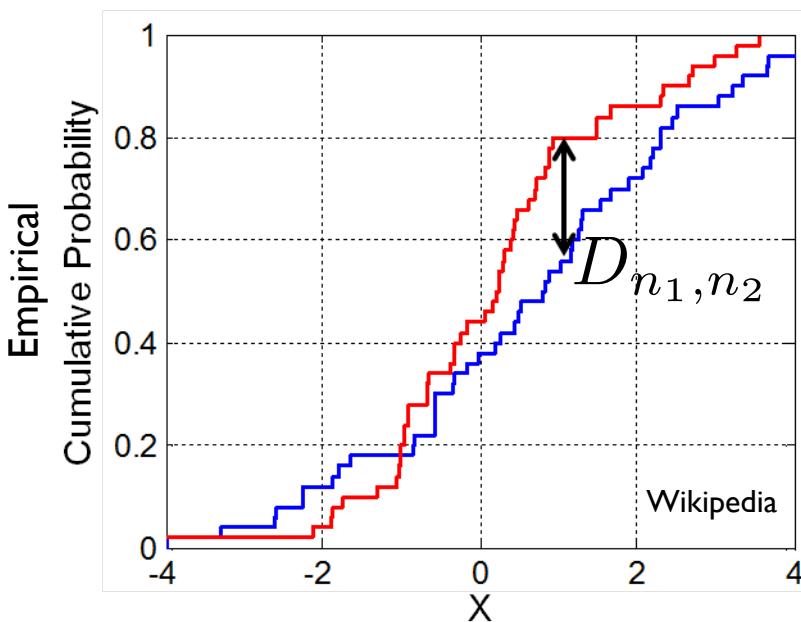
Examples:

Wilcoxon Rank Sum Test , Komogorov-Smirnov Test

Example of Nonparametric Test

Nonparametric Testing of Distributions

- ▶ Two-sample Kolmogorov-Smirnov Test
 - Do $X^{(0)}$ and $X^{(1)}$ come from same underlying distribution?
 - Hypothesis (same distribution) rejected at level p if



$$D_{n_1, n_2} > c(p) \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

Sample size correction

Confidence interval factor

The K-S test is less sensitive when the differences between curves is greatest at the beginning or the end of the distributions. Works best when distributions differ at center.

Good reading:

M.Tygart, Statistical tests for whether a given set of independent, identically distributed draws comes from a specified probability density. PNAS 2010

Are Two User Features Independent?

Chi-Squared Test

- ▶ Twitter users have features gender and number of tweets.
- ▶ We want to determine whether gender is related to number of tweets.
- ▶ Use chi-square test for independence

When to use Chi-Squared test

- ▶ When to use chi-square test for independence:
 - Uniform sampling design
 - Categorical features
 - Population is significantly larger than sample
- ▶ The hypotheses:
 - H_0 : variables are independent
 - H_1 variables are not independent
- ▶ O_i = the number of observations of type i
e.g. (gender = Male, $N > 100$ tweets)
- ▶ E_i = the expected frequency of type i if H_0 is true
e.g. (population_size \times P[gender = Male]P[N > 100 tweets])

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

If H_0 is true χ^2 is distributed as a sum square of random variables $\sim N(0,1)$

Example Chi-Squared Test

```
men = c(300, 100, 40)
```

```
women = c(350, 200, 90)
```

```
data = as.data.frame(rbind(men, women))
```

```
names(data) = c('large', 'med', 'small')
```

```
data
```

```
chisq.test(data)
```

Reject H_0 ($p < 0.05$) means ...

Next Class

- ▶ Select 50% users to see headline A
 - [Titanic Sinks](#)
- ▶ Select 50% users to see headline B
 - [Ship Sinks Killing Thousands](#)
- ▶ If we just care about page visitors we never need to decide which one is best

