A/B Testing
(Hypothesis Testing)

CS57300 - Data Mining
Spring 2016

Instructor: Bruno Ribeiro

How big is infinity?

∞ is not a number

∞ a relationship with other quantities in your equation

How much does it cost?

- population A
- population B

How much does it cost?

- population A

- When writing use the format:

  $100 (no decimal points)

How much does it cost?

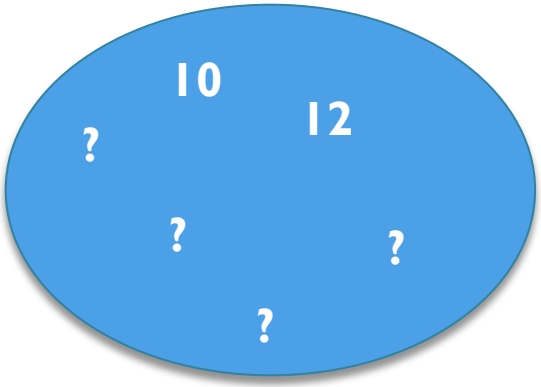
- population B
- When writing use the format:

  $10 (no decimal points)

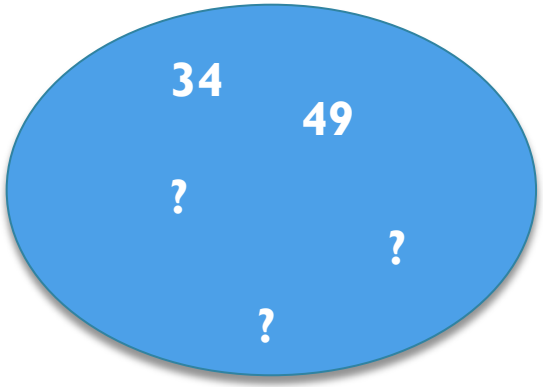
The two-sample t-test

Is difference in averages between two groups more than we would expect based on chance alone?

Testing Hypotheses over Two Populations



Average μ_1



Average μ_2

Are the averages different?
Which one has the largest average?

t-Test (Independent Samples)

The goal is to evaluate if the average difference between two populations is zero

$$x^{(1)} = \text{population A prices}$$

$$x^{(2)} = \text{population B prices}$$

Two hypotheses:

population A average price

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

In the t-test we make the following assumptions

- The values in $x^{(1)}$ and $x^{(2)}$ follow a normal distribution (we will see why)
- Observations are independent

t-Test Calculation

General t formula

$$t = \frac{\text{sample statistic - hypothesized population difference}}{\text{estimated standard error}}$$

Independent samples t

Empirical averages

$$t = \frac{(\bar{x}^{(1)} - \bar{x}^{(2)}) - (\mu_1 - \mu_2)}{\mathrm{SE}}$$

Estimated standard deviation??

t-Statistics p-value

$$H_0: \ \mu_1 - \mu_2 = 0$$

$$H_1: \ \mu_1 - \mu_2 \neq 0$$

- What is the p-value?

$$P[\bar{x}^{(1)} - \bar{x}^{(2)} > \text{observed value} \,|H_0] = p$$

- Can we ever accept hypothesis H_1 ?

R code

```
x1 <- c(1,3)
x2 <- c(2,4)

p <- t.test(x1,x2, alternative = "two.sided")$p.value

print(p)
```

A/B Testing

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

"All the News That's Fit to Print."

The New York Times.

THE WEATHER.

NEW YORK, TUESDAY, APRIL 16, 1912.—TWENTY-FOUR PAGES. ONE CENT

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG; 866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH; ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

Col. Astor and Bride, Isidor Straus and Wife, and Maj. Butt Aboard.

"RULE OF SEA" FOLLOWED

Women and Children Put Over in Lifeboats and Are Supposed to be Safe on Carpathia.

PICKED UP AFTER 8 HOURS

Vincent Astor Calls at White Star Office for News of His Father and Leaves Weeping.

FRANKLIN HOPEFUL ALL DAY

Manager of the Line Insisted Titanic Was Unsinkable Even After She Had Gone Down.

HEAD OF THE LINE ABOARD

J. Bruce Ismay Making First Trip in Gigantic Ship That Was to Surpass All Others.

Biggest Liner Plunges to the Bottom at 2:20 A. M.

RESCUERS THERE TOO LATE

Except to Pick Up the Few Hundreds Who Took to the Lifeboats.

WOMEN AND CHILDREN FIRST

Cunarder Carpathia Rushing to New York with the Survivors.

SEA SEARCH FOR OTHERS

The California Stands By on Chance of Picking Up Other Boats or Rafts.

OLYMPIC SENDS THE NEWS

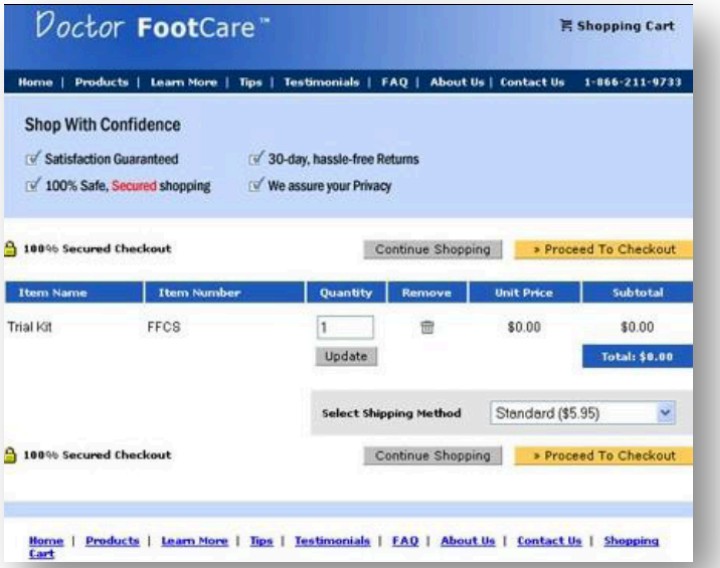
Only Ship to Flash Wireless Messages to Shore After the Disaster.

LATER REPORT SAVES 866.

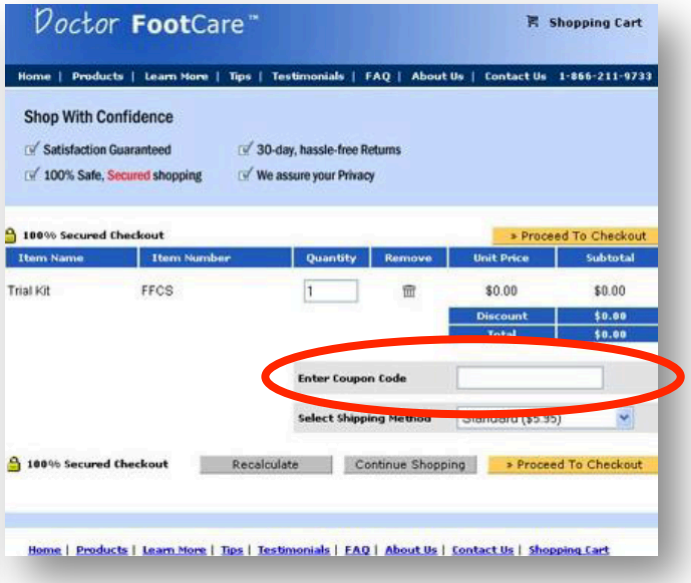
- Do people click more on headline A or B?

A/B Testing on Websites

- Can you guess which page has a higher conversion rate and whether the difference is significant?



A



B

Kumar et al. 2009

- When "upgraded" from the A to B the site lost 90% of their revenue
- Why? "*There maybe discount coupons out there that I do not have. The price may be too high. I should try to find these coupons.*" [Kumar et al. 2009]

Less Obvious Applications

- E.g. software updates
  - Perform incremental A/B testing before rolling a big system change on a website that should have no effect on users
    - What is the hypothesis we want to test?
    - H_0 = no difference in [engagement, purchases, delay, transaction time,…]
  - How?
    - Start with 0.1% of visitors (machines) and grow until 100% of visitors (machines)
    - If at any time H_0 is rejected, stop the roll out
    - Must account for testing multiple hypotheses (next class)

Types of Hypothesis Tests

- Fisher's test
  - Test can only reject H_0 (we **never** accept a hypothesis)
  - H_0 is likely wrong in real-life, so rejection depends on the amount of data
    - More data, more likely we will reject H_0

- Neyman-Pearson's test
  - Compare H_0 to alternative H_1
  - Reject H_0 in favor of H_1

- Bayesian test
  - Compute probability $P[H_0 \mid Data]$ and compare against $P[H_1 \mid Data]$
  - More precisely, compare $P[H_0 \mid Data]/P[H_1 \mid Data]$

About the Usefulness of Priors in Life

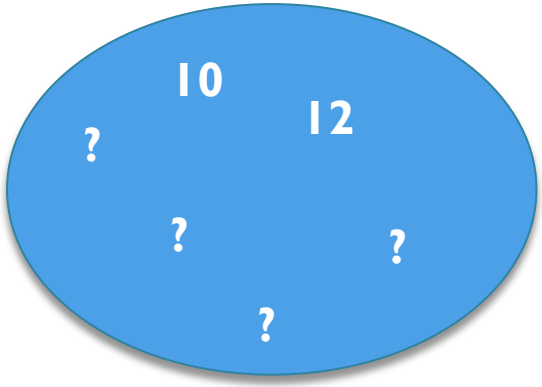
Bayesian Hypothesis Tests Are Very Useful (given good priors)

- Are there aliens visiting Earth?
- 77% of U.S. adults believe there are signs that aliens have visited Earth (National Geographic Poll 2014)

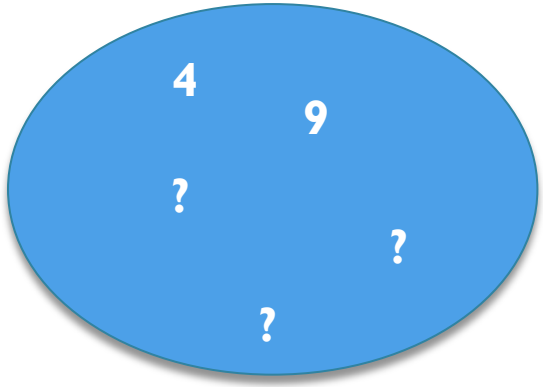
Richard Feynman,
American physicist,
Nobel Laureate in Physics
(1918-1988),
The Character of Physical Law,
Cornell University
Messenger Lectures (1964)

Back to Fisher's test (no priors)

Two Sample Tests (Fisher)



Average μ_1



Average μ_2

Null hypothesis H_0	Alternative hypothesis H_1	No. Tails
$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 \neq d$	2
$\mu_1 - \mu_2 \geq d$	$\mu_1 - \mu_2 < d$	1
$\mu_1 - \mu_2 \leq d$	$\mu_1 - \mu_2 > d$	1

How to Compute Two-sample t-test (1)

1) Compute the empirical standard error

$$\mathrm{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Sample variance of $x^{(1)}$

Number of observations in $x^{(1)}$

where,

$$s_i^2 = \frac{1}{n_i} \sum_{k=1}^{n_i} (x_k^{(i)} - \bar{x}^{(i)})^2$$

and

$$\bar{x}_i = \sum_{m=1}^{n_i} x_m^{(i)}$$

How to Compute Two-sample t-test (2)

2) Compute the degrees of freedom

$$\mathrm{DF} = \left\lfloor \frac{\left(\sigma_1^2/n_1 + \sigma_2^2/n_2\right)^2}{(\sigma_1^2/n_1)^2/(n_1-1) + (\sigma_2^2/n_2)^2/(n_2-1)} \right\rfloor$$

3) Compute test statistic (t-score, also known as Welsh's t)

$$t_d = \frac{(\bar{x}_1 - \bar{x}_2) - d}{\mathrm{SE}}$$

where d is the Null hypothesis difference

$$\mathrm{p} = \mathrm{P}[\mathrm{T}_{\mathrm{DF}} < -|\mathrm{t_d}|] + \mathrm{P}[\mathrm{T}_{\mathrm{DF}} > |\mathrm{t_d}|] \quad \text{(Two-Tailed Test } \mu_1 - \mu_2 = \mathrm{d})$$

$$p = \int_{d \geq 0} P[T_{\mathrm{DF}} > t_d]dd \quad \text{(One-Tailed Test for } \mathrm{H}_0 : \mu_1 - \mu_2 \geq 0)$$

What is the distribution of t_d?

- I don't know (majority answer)
- I don't know (the true answer)

Some assumptions about $\mathbf{x}_1$ and $\mathbf{x}_2$

- $x^{(1)} = [x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)}]$
- $x^{(2)} = [x_1^{(2)}, x_2^{(2)}, \ldots, x_{n_2}^{(2)}]$

- Observations of $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent and identically distributed (i.i.d.)

- Central Limit Theorem (Classical CLT)
  - If: $\mathbf{E}[\mathbf{x}_k^{(i)}] = \mu_i$ and $\mathbf{Var}[\mathbf{x}_k^{(i)}] = \sigma_i^2 < \infty$ (here ∞ is with respect to n_i)

$$\sqrt{n_i}\left(\left(\frac{1}{n_i}\sum_{k=1}^{n} x_k^{(i)}\right) - \mu_i\right) \xrightarrow{d} N(0, \sigma_i^2)$$

- More generally, the real CLT is about stable distributions

CLT: If we have enough independent observations with small variance we can approximate the distribution of their average with a normal distribution

* But we don't know the variance of $\mathbf{x}^{(1)}$ or $\mathbf{x}^{(2)}$

- $N(0, \sigma_i^2)$ approximation not too useful if we don't know σ_i^2
- We can estimate σ_i^2 with n_i observations of $N(0, \sigma_i^2)$
- But we cannot just plug-in estimate $\hat{\sigma}_i^2$ on the normal
  - It has some variability if $n_i < \infty$
  - $\hat{\sigma}_i^2$ is Chi-Squared distributed
  - The t-distribution is a convolution of the standard normal with a Chi-Square distribution to compute

$$t = \frac{\mu_i}{\sqrt{\hat{\sigma}_i^2 / \mathrm{DF}}}$$

For small samples we can use the Binomial distribution

- If results are 0 or 1 (buy, not buy) we can use Bernoulli random variables rather than the Normal approximation

What about
false positives and
false negatives
of a test?

Hypothesis Test Possible Outcomes

$P[H_0 \mid H_0]$	Type II error (false negative) $P[H_0 \mid \neg H_0]$
Type I error (false positive) $P[\neg H_0 \mid H_0]$	$P[\neg H_0 \mid \neg H_0]$

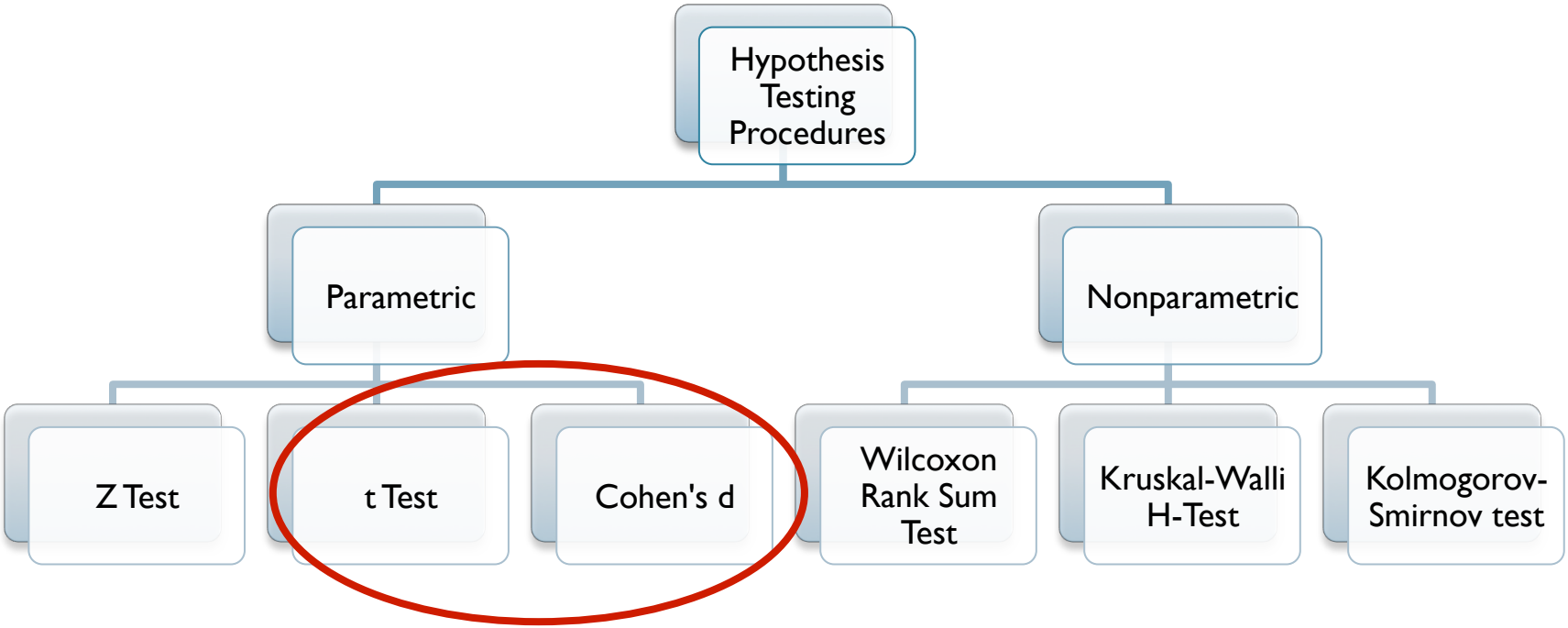
Statistical Power

$$\text{power} = P[\neg H_0 | \neg H_0]$$

- Statistical power is probability of rejecting H_0 when H_0 is indeed false
- Statistical Power ⇒ Number of Observations Needed
- Standard value is 0.80 but can go up to 0.95
- E.g.: H_0 is $\mu_1 - \mu_2 = 0$;
  - Define $n = n_1 = n_2$ such that statistical power is 0.8:
    $P[\text{Test Rejects} \mid |\mu_1 - \mu_2| > \Delta] = 0.8$
    where $\text{Test Rejects} = \mathbf{1}\{P[x^{(1)}, x^{(2)} \mid \mu_1 - \mu_2 = 0] < 0.05\}$
    which gives

$$n = \frac{16\sigma^2}{\Delta^2}$$

More Broadly: Hypothesis Testing Procedures



Parametric Test Procedures

- Tests Population Parameters (e.g. Mean)
- Distribution Assumptions (e.g. Normal distribution)
- Examples: Z Test, t-Test, χ^2 Test, F test

Effect Size

t-Test: Effect Size

t-Test tests only if the difference is zero or not.

What about effect size?

Cohen's d

$$d = \frac{\bar{x}^{(1)} - \bar{x}^{(2)}}{\mathrm{S}}$$

where S is the pooled variance

$$S = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Bayesian Approach

Bayesian Approach

- Probability of hypothesis given data

$$P[H_0|x^{(1)}, x^{(2)}]$$

- The **Bayes factor**

$$K = \frac{P[x^{(1)}, x^{(2)}|H_0]}{P[x^{(1)}, x^{(2)}|H_1]}$$

- Reject H_0 if $\mathbf{K\ P[H_0]/P[H_1]}$ is less than some value

Next Class:

Non-parametric Tests
Independence Tests
Testing Multiple Hypotheses
Sequential Analysis

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

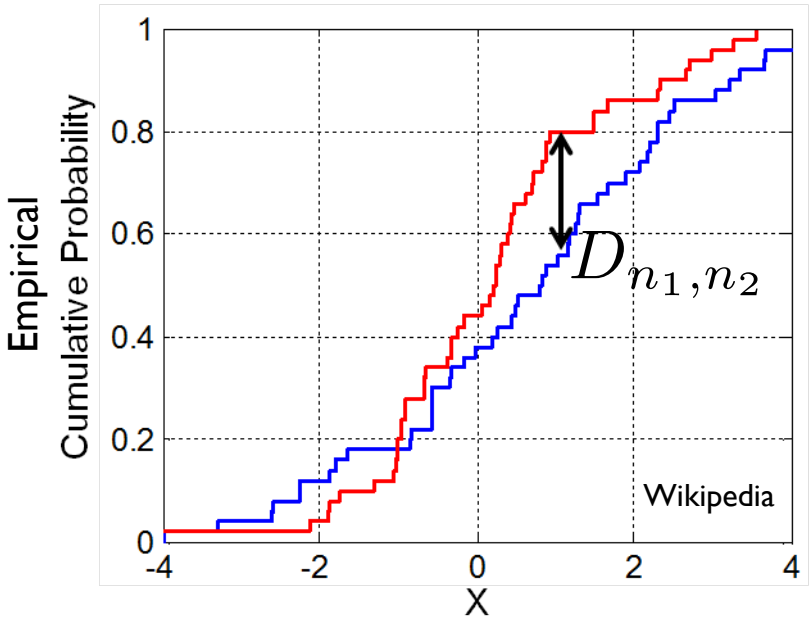
Nonparametric Testing of Distributions

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

Sample size correction

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

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. Tygert, 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 can have 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

- State the hypotheses:
  - H_0 ?
  - H_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('low', 'med', 'large')

data

chisq.test(data)
```

Reject H_0 ($p<0.05$) means …

Deciding Headlines

Revisiting *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

"All the News That's Fit to Print."

The New York Times.

THE WEATHER.

NEW YORK, TUESDAY, APRIL 16, 1912.—TWENTY-FOUR PAGES. ONE CENT

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG; 866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH; ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

Col. Astor and Bride, Isidor Straus and Wife, and Maj. Butt Aboard.

"RULE OF SEA" FOLLOWED

Women and Children Put Over in Lifeboats and Are Supposed to be Safe on Carpathia.

PICKED UP AFTER 8 HOURS

Vincent Astor Calls at White Star Office for News of His Father and Leaves Weeping.

FRANKLIN HOPEFUL ALL DAY

Manager of the Line Insisted Titanic Was Unsinkable Even After She Had Gone Down.

HEAD OF THE LINE ABOARD

J. Bruce Ismay Making First Trip in Gigantic Ship That Was to Surpass All Others.

Biggest Liner Plunges to the Bottom at 2:20 A. M.

RESCUERS THERE TOO LATE

Except to Pick Up the Few Hundreds Who Took to the Lifeboats.

WOMEN AND CHILDREN FIRST

Cunarder Carpathia Rushing to New York with the Survivors.

SEA SEARCH FOR OTHERS

The California Stands By on Chance of Picking Up Other Boats or Rafts.

OLYMPIC SENDS THE NEWS

Only Ship to Flash Wireless Messages to Shore After the Disaster.

LATER REPORT SAVES 866.

- Assign half the readers to headline A and half to headline B?
  - Yes?
  - No?
  - Which test to use?

What happens A is MUCH better than B?

Sequential Analysis (Sequential Hypothesis Test)

- How to stop experiment early if hypothesis seems true
  - Stopping criteria often needs to be decided before experiment starts