Data Mining

CS57300
Purdue University

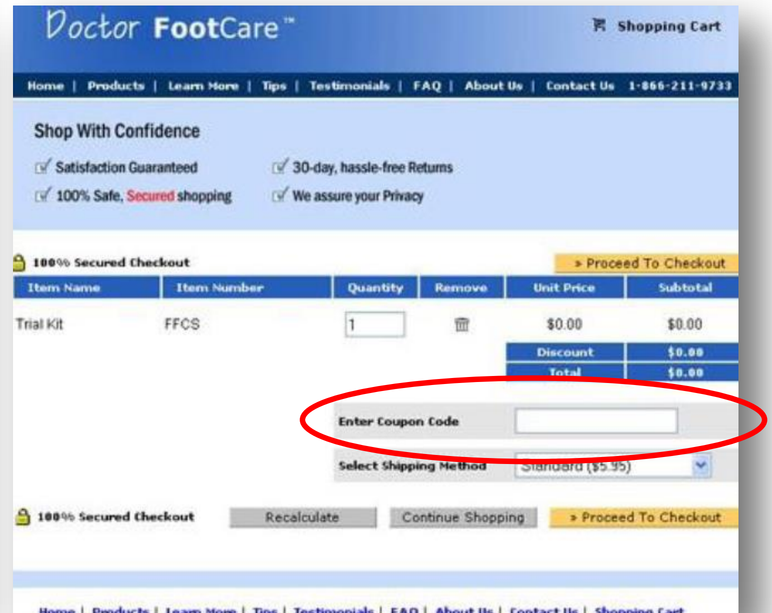
March 20, 2018

A/B Testing

- Select 50% users to see headline A
  - Unlimited Clean Energy: Cold Fusion has Arrived
- Select 50% users to see headline B
  - Wedding War
- Do people click more on headline A or B?

A/B Testing on Websites

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

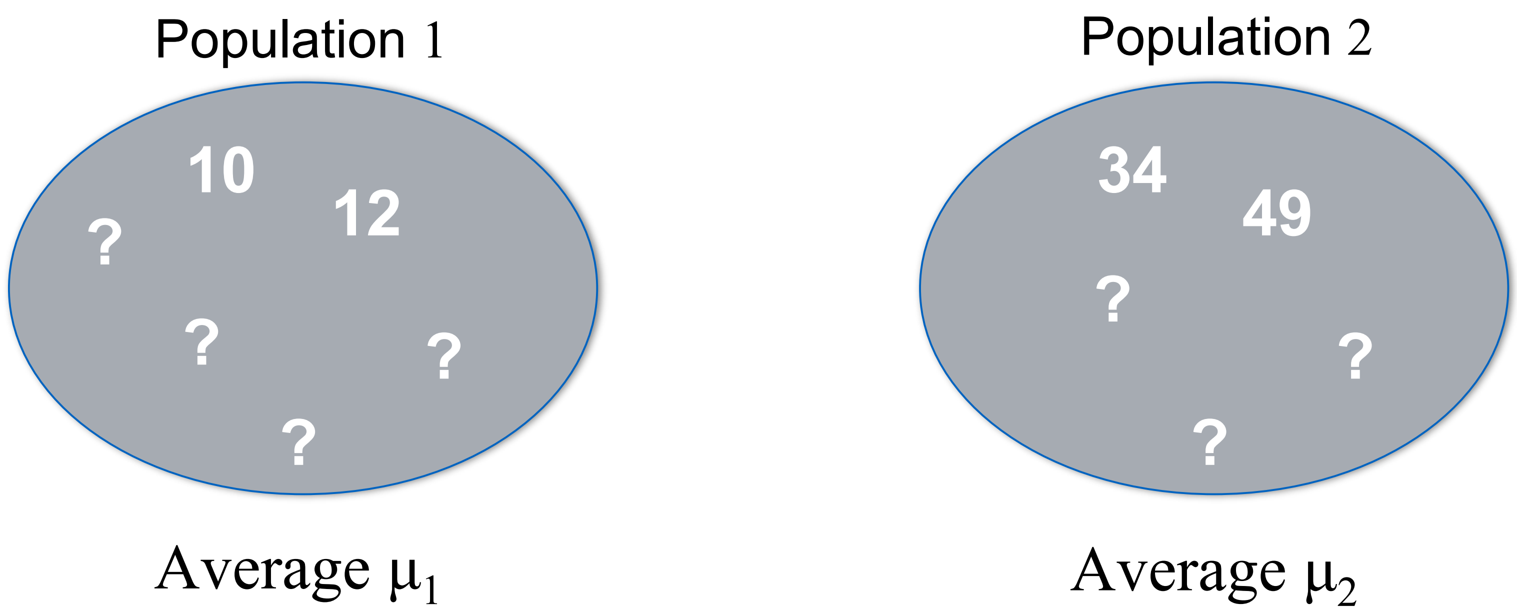


Kumar et al. 2009

A B

- 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]

Testing Hypotheses over Two Populations



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

The two-sample t-test

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

PS: Same as alien identification problem: we don't know how to model "average is different"

t-Test (Independent Samples)

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

vectors

$$\boldsymbol{X}^{(1)} = \text{random variable of population 1 values}$$

$$\boldsymbol{X}^{(2)} = \text{random variable of population 2 values}$$

Two hypotheses:

population 1 average

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

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

In the t-test we make the following assumptions

- The averages $\bar{\boldsymbol{X}}^{(1)}$ and $\bar{\boldsymbol{X}}^{(2)}$ 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

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

Empirical standard deviation (formula later)

t-Statistics p-value

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

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

- What is the p-value?

$$\bar{x}^{(i)} = \text{empirical average of population } i$$

Random variables

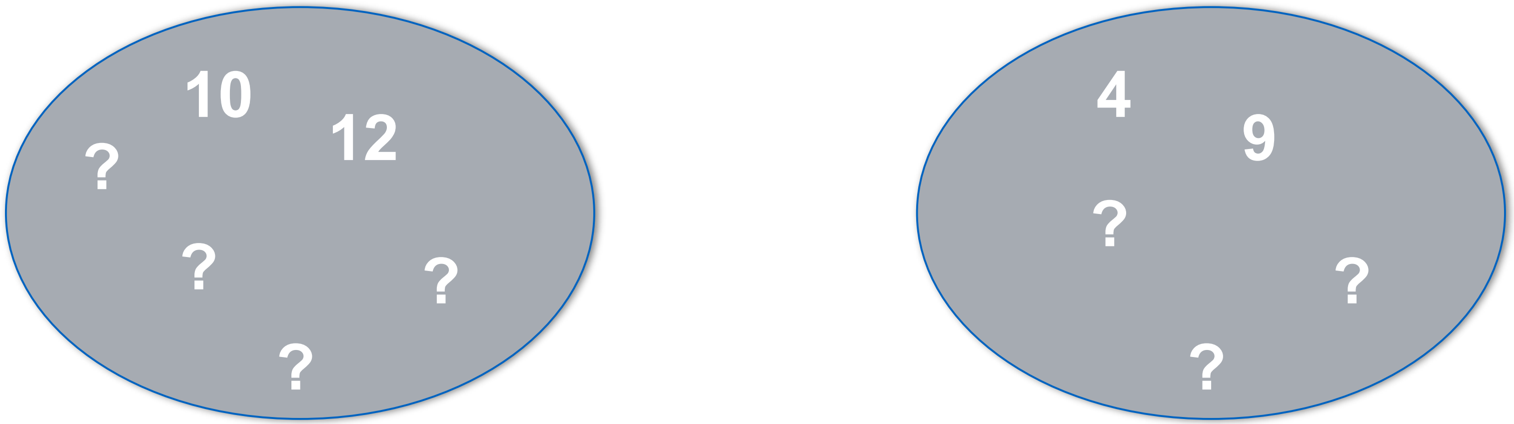
$$P[\bar{X}^{(1,n_1)} - \bar{X}^{(2,n_2)} > \bar{x}^{(1,n_1)} - \bar{x}^{(2)} | H_0] = p$$

- Can we test H_1?

$$\cancel{P[\bar{X}^{(1,n_1)} - \bar{X}^{(2,n_2)} > \bar{x}^{(1,n_1)} - \bar{x}^{(2)} | H_1] = 1 - p?}$$

- Can we ever directly accept hypothesis H_1 ?
  - No, we can't test H_1, we can only reject H_0 in favor of H_1

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
$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 < d$	1
$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 > d$	1

Less Obvious Applications

- E.g. software updates
  - Perform incremental A/B testing before rolling ANY big system change on a website that should have no effect on users
    (even if users don't directly see the change)
    - 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 50% of visitors (machines)
    - If at any time H_0 is rejected, stop the roll out
    - Must account for testing multiple hypotheses (next class)
      (more precisely, this is **sequential analysis**)

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
    - E.g.: H_0: $\mu = \mu_0$ and H_1: $\mu = \mu_1$
    - P[Data | H_0] / P[Data | H_1]

- Bayesian test
    - Compute probability P[H_0 | Data] and compare against P[H_1 | Data]
    - More precisely, test P[H_0 | Data] / P[H_1 | Data]
        - >1 implies H_0 is more likely
        - <1 implies H_1 is more likely
    - Neyman-Pearson's test = Bayes factor when H_0 and H_1 have same priors

Back to Fisher's test
(no priors)

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

1) Compute the pooled empirical standard error

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

where,

Sample variance of $\mathrm{x}^{(\mathrm{i})}$

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

and

Number of observations in $\mathrm{x}^{(\mathrm{i})}$

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

(assumes both populations have equal variance)

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)

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

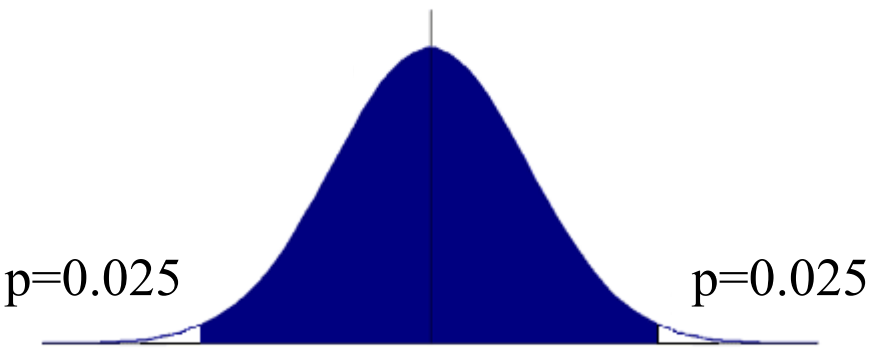
where d is the Null hypothesis difference.

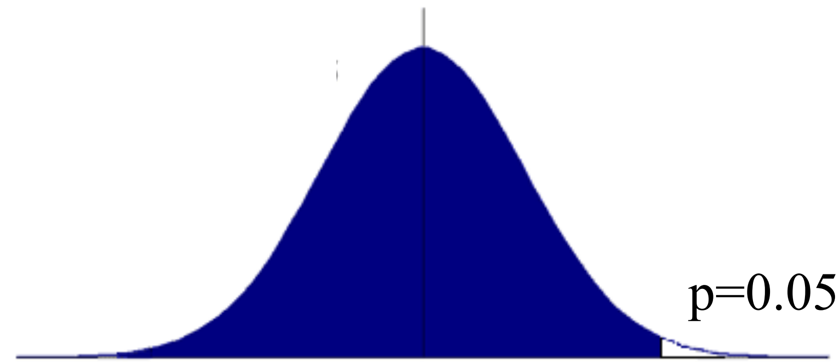
4) Compute p-value (depends on H_1)

- $p = P[T_{DF} < -|t_d|] + P[T_{DF} > |t_d|]$ (Two-Tailed Test H_1: $\mu_1 - \mu_2 \neq d$)
- $p = P[T_{DF} > t_d]$ (One-Tailed Test for $H_1 : \mu_1 - \mu_2 > d$)
- Important: H_0 is always $\mu_1 - \mu_2 = d$ even when $H_1 : \mu_1 - \mu_2 > d$!!
  Testing H_0: $\mu_1 - \mu_2 \leq d$ is harder and "has same power" as H_0: $\mu_1 - \mu_2 = d$

Rejecting H_0 in favor of H_1

- Back to step 4 of slide 16:





4) Compute p-value (depends on H_1)

$p = P[T_{DF} < -|t_d|] + P[T_{DF} > |t_d|]$ (Two-Tailed Test H_1: $\mu_1 - \mu_2 \neq d$)

$p = P[T_{DF} > t_d]$ (One-Tailed Test for $H_1 : \mu_1 - \mu_2 > d$)

Reject H_0 with 95% confidence if $p < 0.05$

Some assumptions about X_1 and X_2

- $\boldsymbol{X}^{(1)} = [\boldsymbol{X}_1^{(1)}, \boldsymbol{X}_2^{(1)}, \ldots, \boldsymbol{X}_{n_1}^{(1)}]$
- $\boldsymbol{X}^{(2)} = [\boldsymbol{X}_1^{(2)}, \boldsymbol{X}_2^{(2)}, \ldots, \boldsymbol{X}_{n_2}^{(2)}]$

- Observations of X_1 and X_2 are independent and identically distributed (i.i.d.)

- Central Limit Theorem (Classical CLT)
  - If: $E[X_k^{(i)}] = \mu_i$ and $Var[X_k^{(i)}] = \sigma_i^2 < \infty$

$$\sqrt{n_i}\left(\left(\frac{1}{n_i}\sum_{k=1}^{n} x_k^{(i)}\right) - \mu_i\right) \xrightarrow{d} N(0, \sigma_i^2) \quad \text{(here } \infty \text{ is with respect to } n_i\text{)}$$

- More generally, the real CLT is about stable distributions

Central Limit Theorem

- 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 $\mathrm{X}^{(1)}$ or $\mathrm{X}^{(2)}$

- $N(0, \sigma_i^2)$ 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

Errors:

$P[\neg H_0|H_0]$ - Reject H_0 given H_0 is true

$P[H_0|\neg H_0]$ - Accept H_0 given H_0 is false

In medicine our "goal" is to reject H_0
(drug, food has no effect / not sick), thus a "positive" result rejects H_0

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

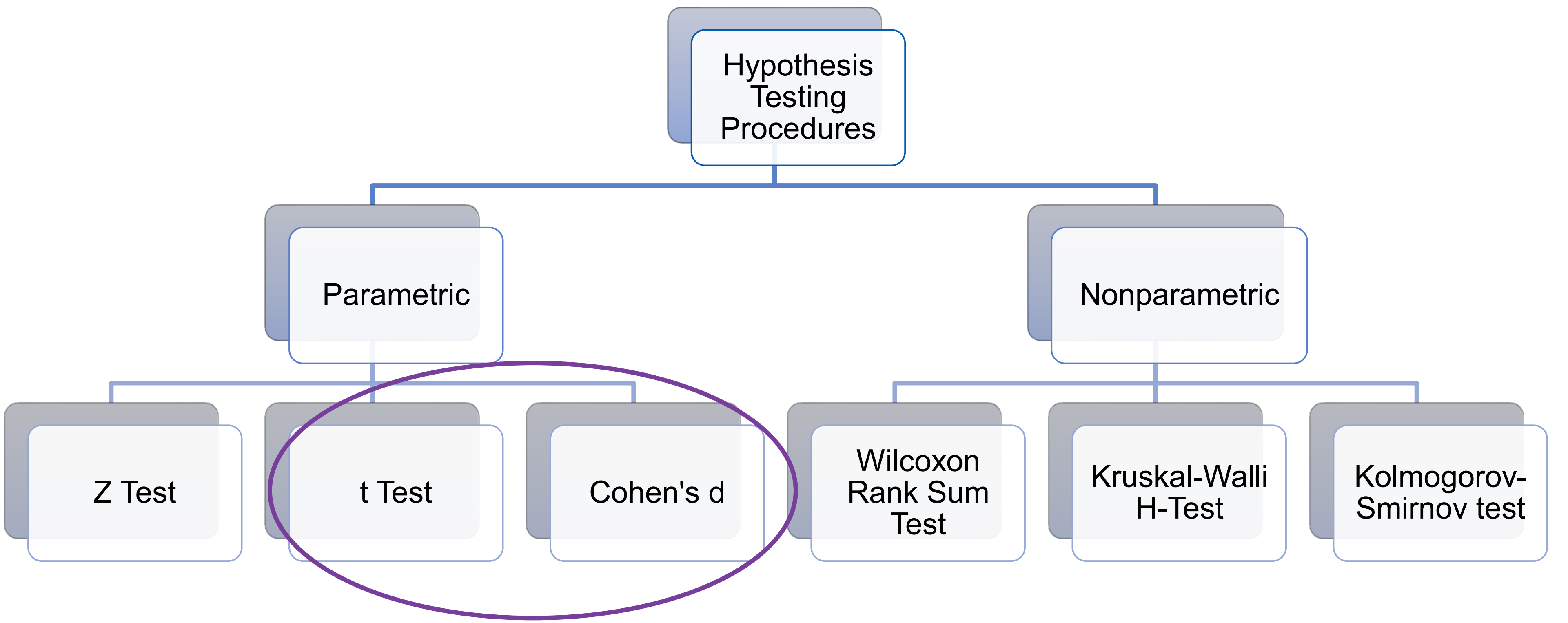
Statistical Power

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

- 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 $\mu_1 - \mu_2 = 0$, where μ_i = true average of population i
  - Define $n = n_1 = n_2$ such that statistical power is 0.8 under assumption $|\mu_1 - \mu_2| = \Delta$:
  - $P[\text{Test Rejects} \mid |\mu_1 - \mu_2| = \Delta] = 0.8$
    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

Testing Effect Sizes

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

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

Effect Size: Good practice

Cohen's d often used to complement t-test when reporting effect sizes

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

where S is the pooled variance

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

Important Warning

American Statistical Association Statement On Statistical Significance And p-values

1. p-values can indicate how incompatible the data are with a specified statistical model.

2. p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4. Proper inference requires full reporting and transparency.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Bayesian Approach

Bayesian Approach

- Probability of hypothesis given data

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

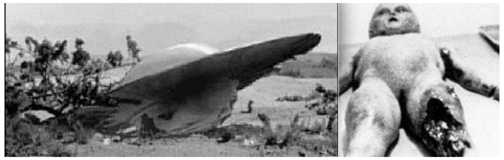
- The **Bayes factor**

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

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

Bayesian Hypothesis Tests Need Assumptions

- Have Aliens visited Earth and government keeping secret?
  - 21% of U.S. voters say a UFO crashed in Roswell, NM in 1947 and the US government covers it up
  - Priors:
    - H_0 : At least 21% of U.S. voters are irrational, will believe in alien story without evidence
      - $P[H_0] = 10^{10}/(10^{10}+1)$ [Ribeiro's prior]
      - $P[H_0] \sim \text{Beta}(10^{10},1)$ [Prior can also be a random variable, better models uncertainty]
    - H_1 : Aliens can travel faster than the speed of light and, despite that, can't drive and are easily captured by humans.
      - Because either H_0 or H_1 must be true: $P[H_1]=1- P[H_0]$



- What is the data?
- Data:
  - 15% of U.S. voters say the government or the media adds mind-controlling technology to TV broadcast signals (a.k.a., the Tinfoil Hat crowd)
  - 20% of U.S. voters believe there is a link between childhood vaccines and autism, despite scientific evidence there is no such link
  - 15% of U.S. voters think the medical industry and the pharmaceutical industry "create" new diseases to make money (Ebola, Zika,…)
  - 14% of U.S. voters say the CIA was instrumental in creating the crack cocaine epidemic
- Bayesian Disadvantage : Often hard to define $P[\text{Data} \mid H_1]$
- Bayesian Advantage: Prior helps encode your uncertainty and beliefs about the world

Source: Public Policy Polling, 2013

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

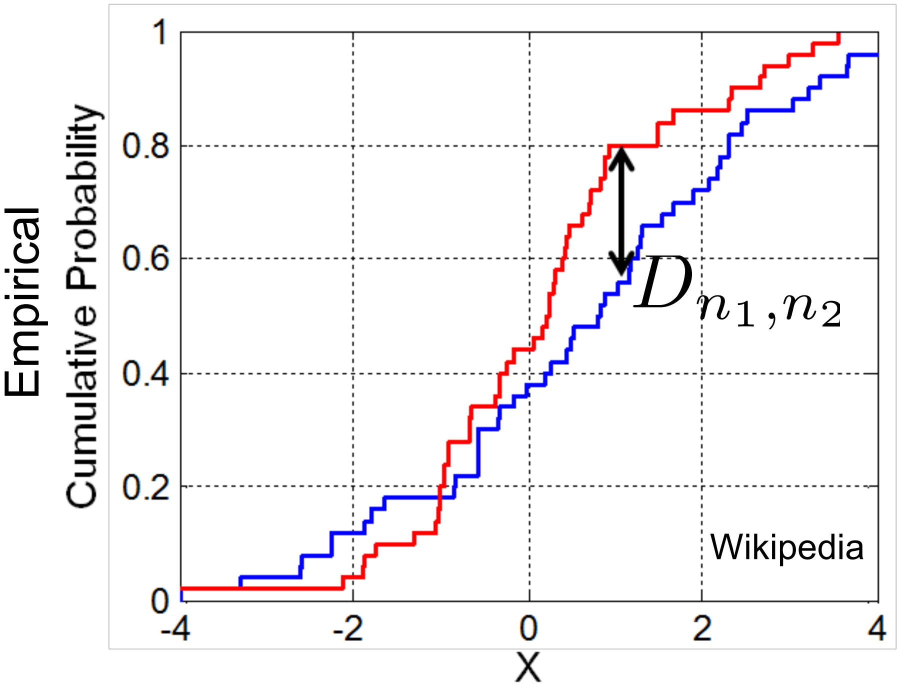
Nonparametric Testing of Distributions

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

Sample size correc

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

Confidence interval fac



The K-S test is less sensitive when the differences between curves is greatest at the beginning or the end of

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.

Unsettled Tuesday; Wednesday, fair, cooler; moderate southerly winds, becoming variable.

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 on Gigantic Ship That Was to Surpass All Others.

The admission that the Titanic, the biggest steamship in the world, had been sunk by an iceberg and had gone to the bottom of the Atlantic, probably carrying more than 1,400 of her passengers and crew with her, was made

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.

BOSTON, April 15.—A wireless message picked up late to-night, relayed from the Olympic, says

- 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?