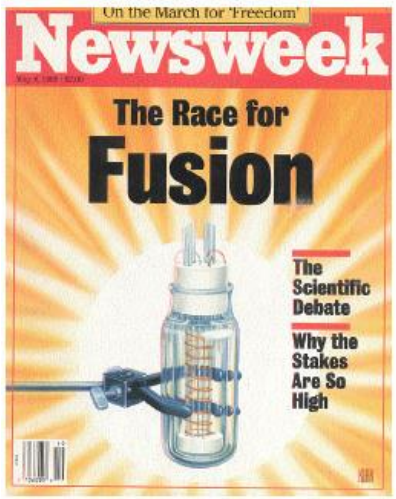
Data Mining

CS57300
Purdue University

March 22, 2018

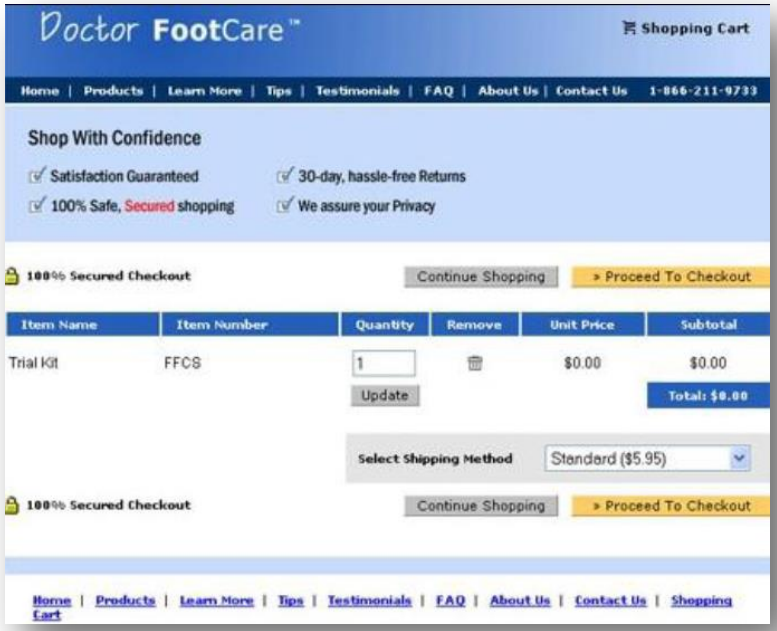
Hypothesis Testing

- 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?

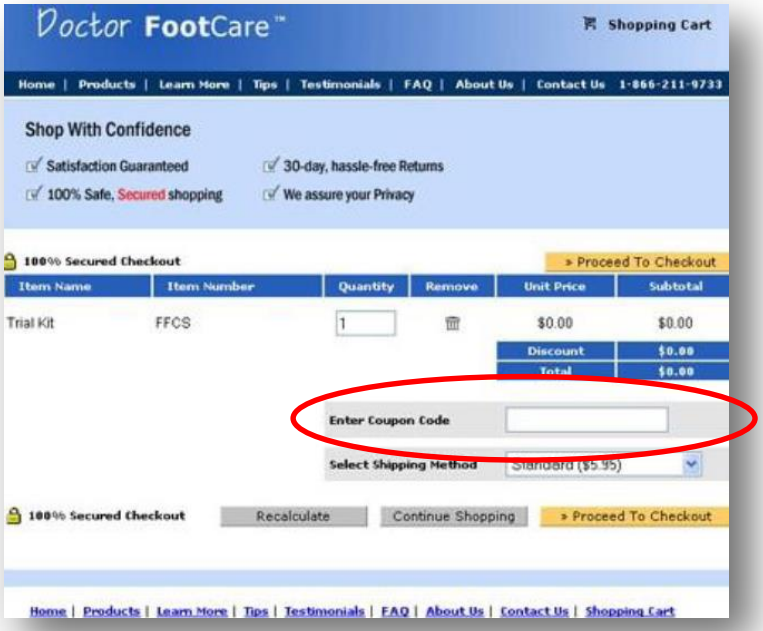


Hypothesis Testing, Because Reality is not Easy

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



A



B

Kumar et al. 2009

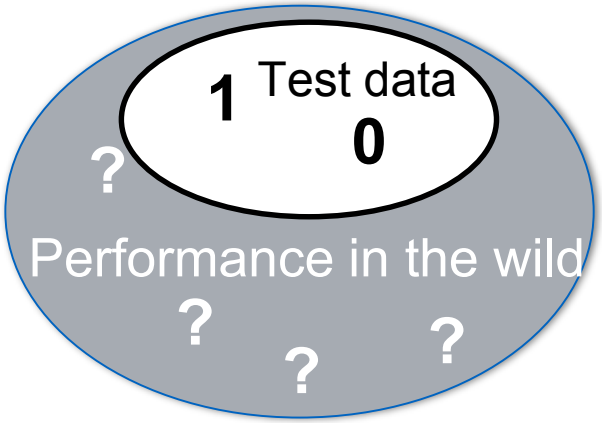
- When "upgraded" from the A to B the site lost 90% of their revenue
- Why? "*There maybe discount coupons out there that I do not have. The price may be too high. I should try to find these coupons.*" [Kumar et al. 2009]

Testing Hypotheses over Two Populations

- "Is my classifier better than existing classifiers in the literature?"

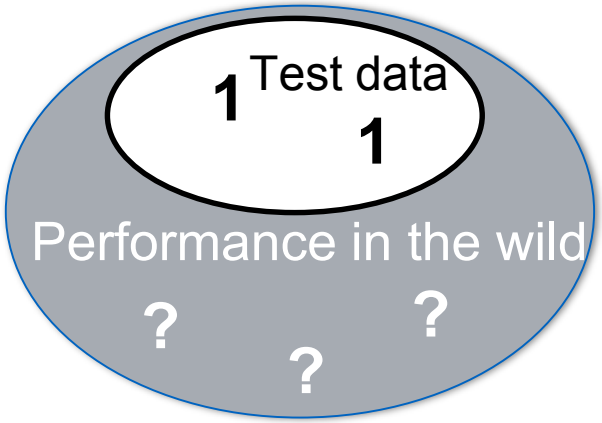
- Why?
  - "I got higher accuracy and tested against all existing methods"
  - 0 – correctly classified test example, 1 – incorrectly classified test example

"My Classifier" Accuracy



True Average μ_1 error

"Competing Classifier" Accuracy



True Average μ_2

Student's Hypothesis: $\mu_1 < \mu_2$

Replication Crisis in Science (John Oliver)

Close to Home

- Machine Learning is somewhat going through a replication crisis
  - Depends on topic, some topics more prone to errors than others

  Examples:

  - Lucic et al.(2017) conducted a large-scale empirical comparison of generative adversarial networks methods and found that most of them reach similar scores with sufficient hyperparameter optimization.
    - Hyperparameters: Neural net number of layers, no. neurons, batch sizes, learning rates
  - Henderson et al. (2017) show they beat a host of sequence-to-sequence methods in the Penn Treebank dataset simply by doing better hyperparameter tuning on the baseline LSTM
  - Henderson et al. (2017) reviewed reproducibility in deep reinforcement learning and found significant variability between baseline implementations across recent work.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. arXiv preprint arXiv:1709.06560, 2017.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. arXiv preprint arXiv:1711.10337, 2017.

Statistical Analysis: Hypothesis testing

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

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

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

- What is the p-value?

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

Random variables

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

Interpretation:
Under the assumption that averages are the same

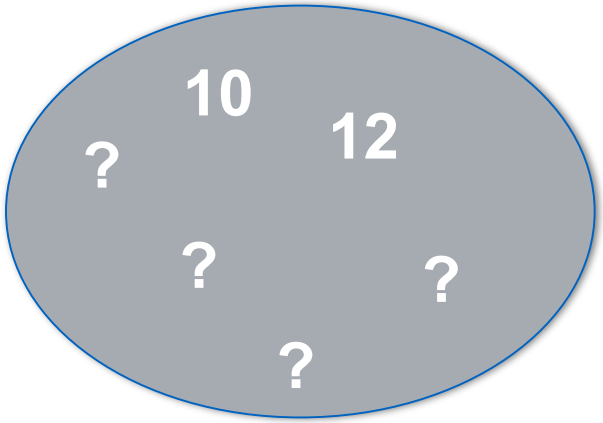
- Can we test H_1?

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

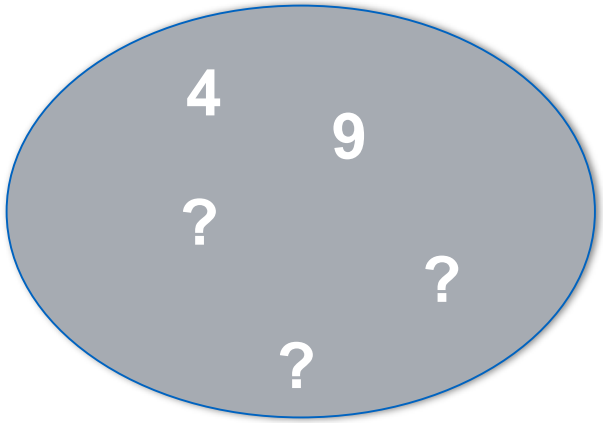
Interpretation:
Under the assumption that averages are different

- 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
  - Why? Because if H_1 does not tell us how different the averages are, how can we compute the probability?

Two Sample Tests (Fisher)



True Population Average: μ_1
Empirical Average over Test Data: $\overline{x}_1$



True Population Average: μ_2
Empirical Average over Test Data: $\overline{x}_2$

Null hypothesis H_0	Alternative hypothesis H_1	No. Tails
$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 \neq d$	2
$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 > d \quad (\overline{x}_1 - \overline{x}_2 > d)$	1
$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 < d \quad (\overline{x}_1 - \overline{x}_2 < d)$	1

Types of Hypothesis Tests

- Fisher's test
  - Test can only reject H_0 (we **never** accept a hypothesis)
  - H_0 is unlikelyin 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 $x^{(i)}$

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

and

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

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

(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)

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

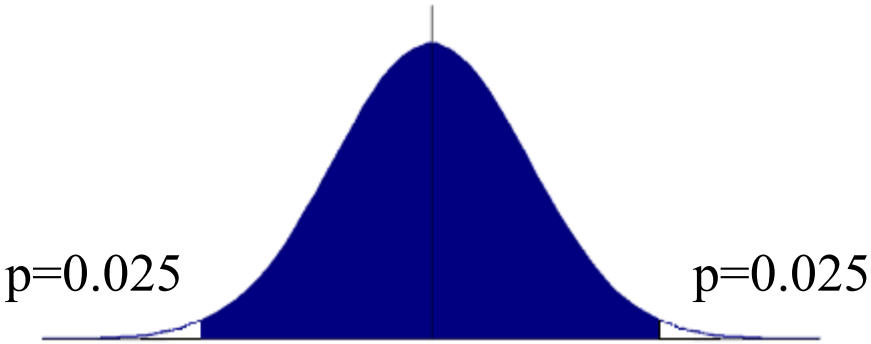
where d is the Null hypothesis difference.

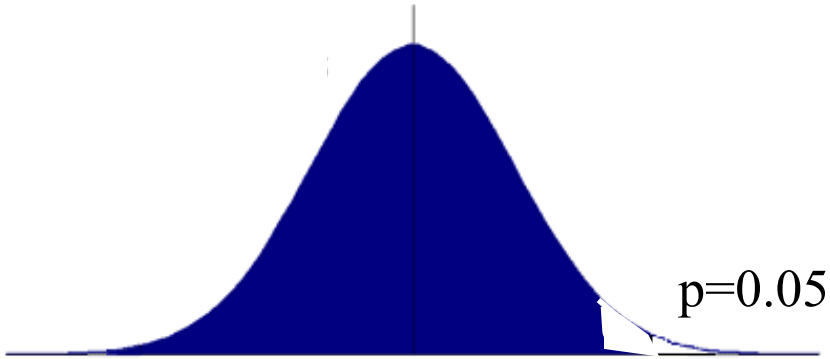
4) Compute p-value (depends on H_1)

- $p = P[T_{DF} < -|t_d|] + P[T_{DF} > |t_d|]$ (Two-Tailed Test H_1: $\mu_1 - \mu_2 \neq d$)
- $p = P[T_{DF} > t_d]$ (One-Tailed Test for H_1 : $\mu_1 - \mu_2 > d$ (& $\bar{x}_1 - \bar{x}_2 > d$))
- 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 $Var[X_k^{(i)}] = \sigma_i^2 < \infty$

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

(here ∞ is with respect to n_i)

- CLT: If we have enough independent observations with relative small variance (wrt number of observations) 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 (wrong class / correct class) we can use exact Bernoulli random variables rather than the Normal approximation
- Normal approximation generally OK for large enough number of examples (> 30)

What about
false positives and
false negatives
of a test?

Hypothesis Test Possible Outcomes

Errors:
Variable: $R = 1 - H_0$ hypothesis rejected, $R = 0 - H_0$ hypothesis not rejected

$P[\, R = 1 \mid H_0 \,]$ - Reject H_0 given H_0 is true

$P[\, R = 1 \mid \text{not } H_0 \,]$ - - 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

$P[\, R = 0 \mid H_0 \,]$	Type I error (false positive) $P[\, R = 1 \mid H_0 \,]$
Type II error (false negative) $P[\, R = 0 \mid \text{not } H_0 \,]$	$P[\, R = 1 \mid \text{not } H_0 \,]$

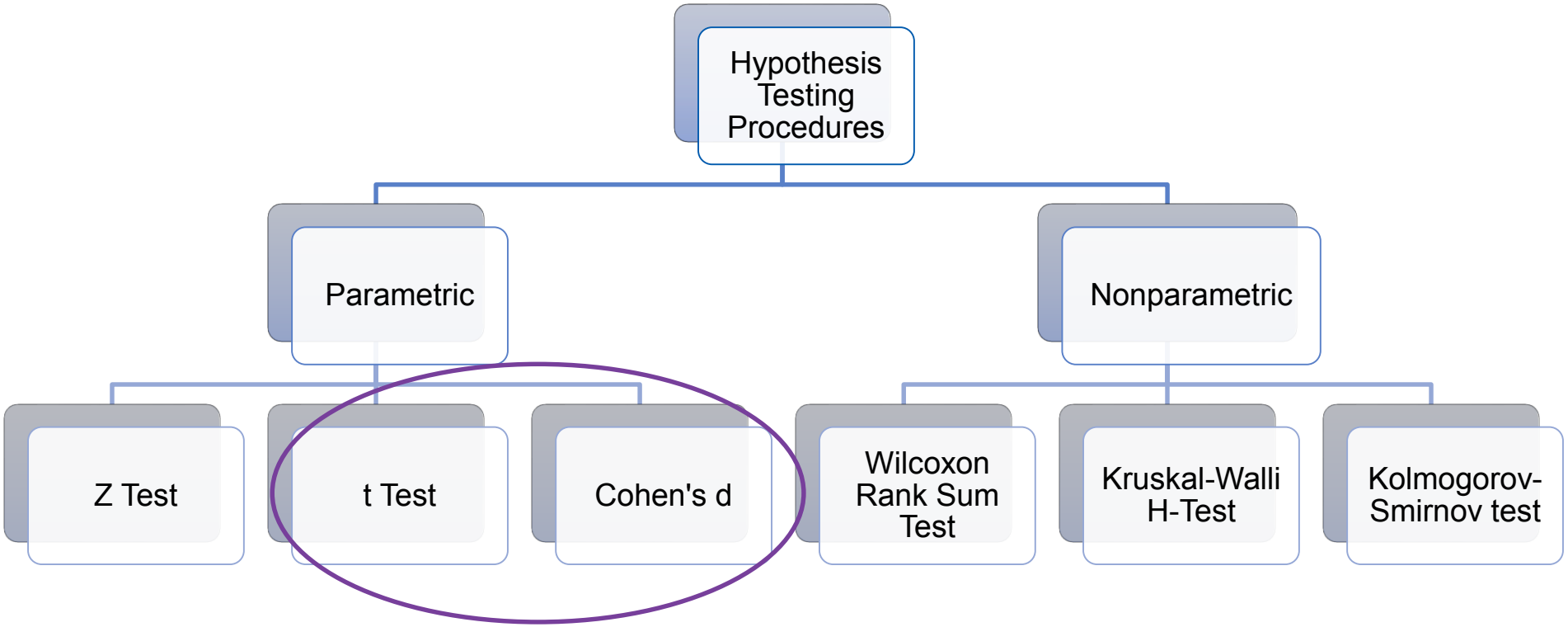
Statistical Power

$$\text{Power} = P[\, R = 1 \mid \text{not } H_0 \,]$$

- 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 $|\mu_1 - \mu_2| = \Delta$:
  - $P[R = 1 \mid |\mu_1 - \mu_2| = \Delta] = 0.8$
    where $R = \mathbf{1}\{P[x^{(1)}, x^{(2)} \mid \mu_1 - \mu_2 = 0] < 0.05\}$
    which gives

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

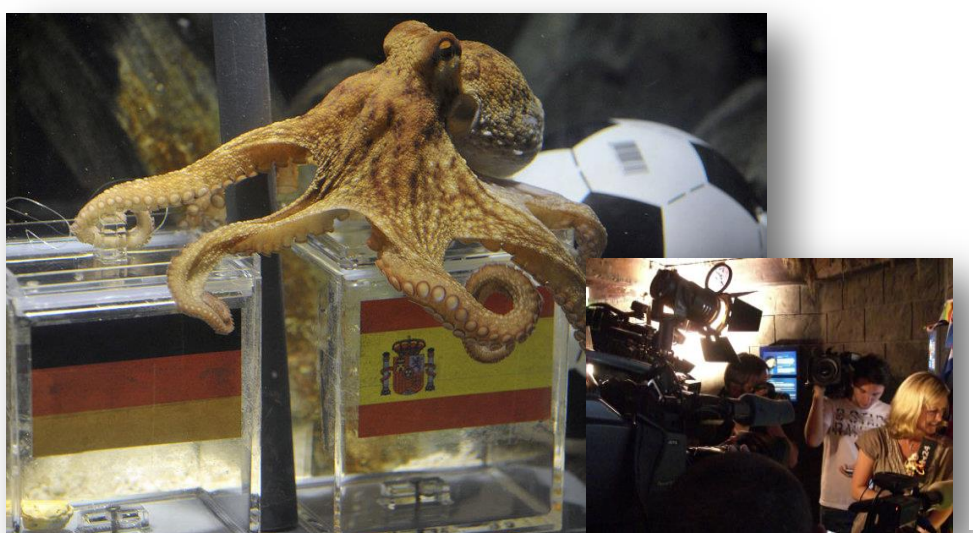
More Broadly: Hypothesis Testing Procedures



Multiple Hypothesis Testing

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

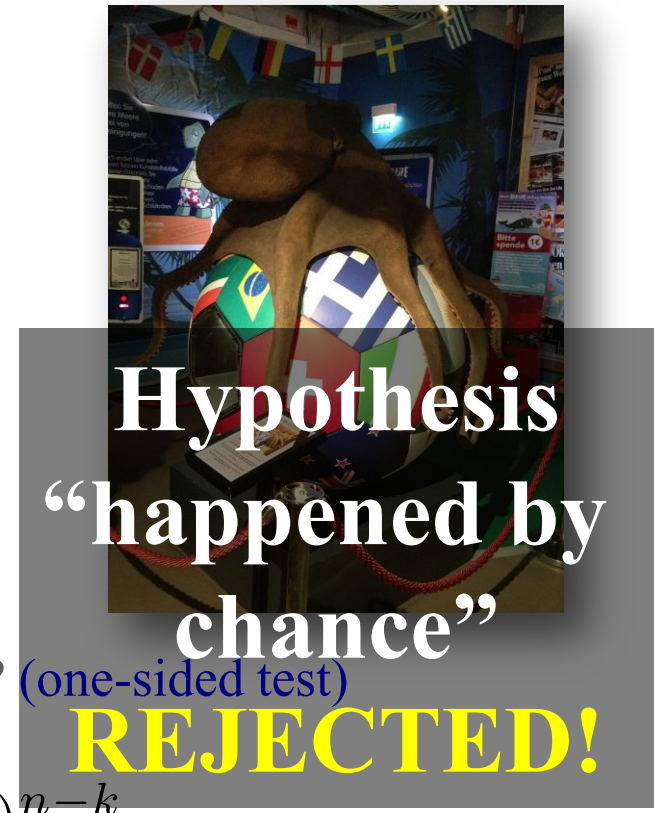
- Random variable (i.i.d.)

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

- Variable of interest:

$$Y_{13} = \sum_{i=1}^{13} X_i$$

- What is the Null Hypothesis?
  - Paul is not an animal oracle
  - Mathematical definition?
    - $H_0 := P[X_i = 1] = p = 0.5$



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

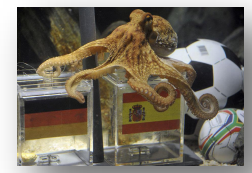
$$\Rightarrow P[Y_{13} = k|H_0] = \binom{13}{k} 0.5^k (1-0.5)^{n-k}$$

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

Anything Wrong in our Hypothesis Test?

Hypothesis Test as Random Variable

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

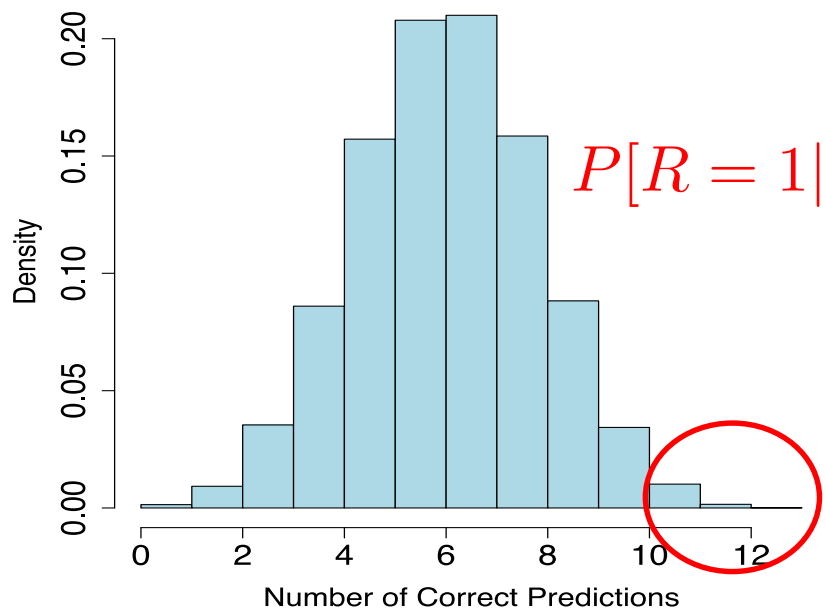


- R is random variable that defines if hypothesis is rejected

$$\text{if } P[Y_{13} \geq k | H_0] < 0.05 \text{ then } R = 1; \text{ otherwise } R = 0$$

k correct predictions by animal

Binomial Distribution (p=0.5)

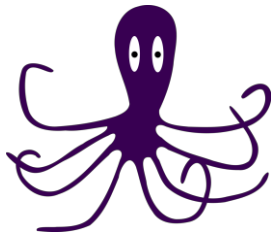


$$P[R = 1 | H_0] = P[Y_{13} \geq 10 | H_0] = 0.046$$

Testing Multiple Hypotheses

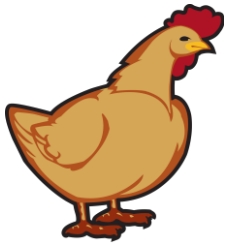
Familywise Error

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

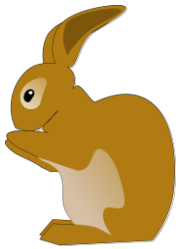


Paul

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)

$$1 - \underbrace{(1 - P[R = 1|H_0])^4}_{} = 0.17$$

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

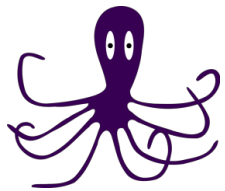
Example

Bob is developing a new caching policy for high-performance databases. For this research Bob secured a real-world dataset containing N user query requests to a large real-world database. Bob tells Alice he is now ready to publish his research. His method incurs a 10% less cache misses than the competing state-of-the-art method, a very good result that will change the industry. Bob complains to Alice that this has been very hard work, he has tried over 100 different caching algorithms for his research. Alice took CS57300 and is skeptical about Bob's results.

What is the problem with Bob's claim?

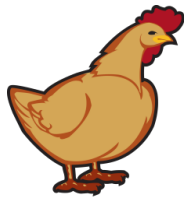
Bonferoni's correction

- 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 α=0.05
  - Bonferoni's corrected significance level α'=0.05/4 = 0.0125
  - Hypothesis test: "Paul is not an animal oracle"

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

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 >> 1)
- False Discovery Rate:
  - Controls:
  - Greater statistical power at expense of more false positives
  - Order p-values of all m tests: recall p-value is related to P[R=1 | H_0]
  - Holm's Method:
    - $\tilde{p}_i = \min((m-i+1)p_i, 1) \qquad p_1 \leq p_2 \leq \cdots \leq p_m$
    - Reject if adjusted p-value < α
  - Benjamini-Hochberg method:
    - Reject j null hypothesis if

$$p_j \leq \alpha \frac{j}{m}$$

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.