Inference 1: Hypothesis testing

Last updated: September 12, 2017

Statistically significant = unlikely due to pure chance

Descriptive statistics = Tell me what you see

eg: First class's survival rate is **62%**, **1.5 times higher** than second class, and **2.5 times higher** than third class and crew.

Inferential statistics = Is what you see significant?

eg: Is the difference in survival rates statistically significant (unlikely due to chance alone) ?

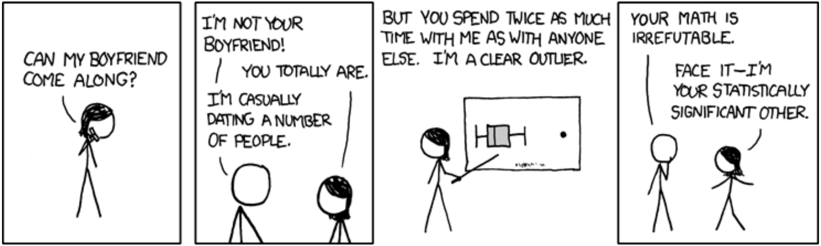
Statistically significant = unlikely due to pure chance

Descriptive statistics = Tell me what you see

eg: First class's survival rate is **62%**, **1.5 times higher** than second class, and **2.5 times higher** than third class and crew.

Inferential statistics = Is what you see significant?

eg: Is the difference in survival rates statistically significant (unlikely due to chance alone) ?



How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see

1. State the question: what is ...?
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Inferential statistics = Is what you see significant?

1. State the question: is ... significant or not?
2. Choose an appropriate statistical test / model
3. Do the test / fit the model
4. Report findings
5. Criticize the data AND the methods used

How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see

1. State the question: what is ...?
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

- **No inference without descriptions!**
- **Both are hard to do well**
- **Both are widely used AND abused**

Inferential statistics = Is what you see significant?

1. State the question: is ... significant or not?
2. Choose an appropriate statistical test / model
3. Do the test / fit the model
4. Report findings
5. Criticize the data AND the methods used

How statisticians think: Hypothesis testing

Hypothesis testing = statistical proof by contradiction

How statisticians think: Hypothesis testing

Hypothesis testing = statistical proof by contradiction

Proof by contradiction

- Assume H_0 is right.
- Observe some event E from data.
- Deduce logically that if H_0 is true, the chance of seeing E is zero .
- Conclude that H_0 is wrong.

How statisticians think: Hypothesis testing

Hypothesis testing = statistical proof by contradiction

Proof by ~~contradiction~~ hypothesis testing

- Assume H_0 is right.
- Observe some event E from data.
- Deduce logically that if H_0 is true, the chance of seeing E is ~~zero~~ tiny.
- Conclude that H_0 is most likely wrong.

How statisticians think: Hypothesis testing

Hypothesis testing = statistical proof by contradiction

Proof by ~~contradiction~~ hypothesis testing

- Assume H_0 is right.
- Observe some event E from data.
- Deduce logically that if H_0 is true, the chance of seeing E is ~~zero~~ tiny.
- Conclude that H_0 is most likely wrong.

Terminologies:

- H_0: *null hypothesis*
- H_A: *alternative hypothesis*
- Probability of E under H_0: *p-value*
- Often $E = \{S \geq s\}$ for some random variable S. This S is the *test statistic*. s is the value of S observed from data.

How statisticians think: Hypothesis testing

Hypothesis testing = statistical proof by contradiction

Proof by ~~contradiction~~ hypothesis testing

- Assume H_0 is right.
- Observe some event E from data.
- Deduce logically that if H_0 is true, the chance of seeing E is ~~zero~~ tiny.
- Conclude that H_0 is most likely wrong.

Terminologies:

- H_0: *null hypothesis* (common sense BEFORE data arrived)
- H_A: *alternative hypothesis*
- Probability of E under H_0: *p-value*
- Often $E = \{S \geq s\}$ for some random variable S. This S is the *test statistic*. s is the value of S observed from data.

How statisticians think: Hypothesis testing

Hypothesis testing = statistical proof by contradiction

Proof by ~~contradiction~~ hypothesis testing

- Assume H_0 is right.
- Observe some event E from data.
- Deduce logically that if H_0 is true, the chance of seeing E is ~~zero~~ tiny.
- Conclude that H_0 is most likely wrong.

Terminologies:

- H_0: *null hypothesis* (common sense BEFORE data arrived)
- H_A: *alternative hypothesis* (common sense if we end up rejecting H_0)
- Probability of E under H_0: *p-value*
- Often $E = \{S \geq s\}$ for some random variable S. This S is the *test statistic*. s is the value of S observed from data.

How statisticians think: Hypothesis testing

Hypothesis testing = statistical proof by contradiction

Proof by ~~contradiction~~ hypothesis testing

- Assume H_0 is right.
- Observe some event E from data.
- Deduce logically that if H_0 is true, the chance of seeing E is ~~zero~~ tiny.
- Conclude that H_0 is most likely wrong.

Terminologies:

- H_0: *null hypothesis* (common sense BEFORE data arrived)
- H_A: *alternative hypothesis* (common sense if we end up rejecting H_0)
- Probability of E under H_0: *p-value* (smaller = stronger contradiction)
- Often $E = \{S \geq s\}$ for some random variable S. This S is the *test statistic*. s is the value of S observed from data.

How statisticians think: Hypothesis testing

Hypothesis testing = statistical proof by contradiction

Proof by ~~contradiction~~ hypothesis testing

- Assume H_0 is right.
- Observe some event E from data.
- Deduce logically that if H_0 is true, the chance of seeing E is ~~zero~~ tiny.
- Conclude that H_0 is most likely wrong.

Terminologies:

- H_0: *null hypothesis* (common sense BEFORE data arrived)
- H_A: *alternative hypothesis* (common sense if we end up rejecting H_0)
- Probability of E under H_0: *p-value* (smaller = stronger contradiction)
- Often $E = \{S \geq s\}$ for some random variable S. This S is the *test statistic*. s is the value of S observed from data. (hardest part: choose the right S)

Which test to use

The analogue of the question ‘Which graph to use?’ in descriptive analysis.

Like descriptives: **depends on the data**.
Different from descriptives:

- many more tests
- S (test stat) depends on H_A (alternative)
- a test is only valid if the data fit the assumptions
- justifying the assumptions = CRUCIAL.

Which test to use

The analogue of the question 'Which graph to use?' in descriptive analysis.

Like descriptives: **depends on the data**.
Different from descriptives:

- many more tests
- S (test stat) depends on H_A (alternative)
- a test is only valid if the data fit the assumptions
- justifying the assumptions = CRUCIAL.

Have you heard of any statistical tests?

Which test to use

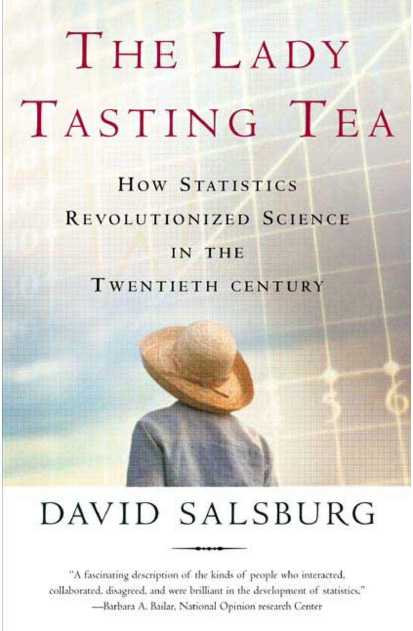
The analogue of the question ‘Which graph to use?’ in descriptive analysis.

Like descriptives: **depends on the data**.
Different from descriptives:

- many more tests
- S (test stat) depends on H_A (alternative)
- a test is only valid if the data fit the assumptions
- justifying the assumptions = CRUCIAL.

Tests we learn	num	cat	cat. vs cat.	num. vs cat.
Fisher's exact			✓	
permutation (randomization)	✓	✓	✓	✓
chi-square			✓	
z-test for proportions		✓		
t-test	✓			
two-sample t-test				✓
ANOVA				✓

Classical Fisher's exact test



Classical Fisher's exact test

In 1935, Dr. Bristol, a British biologist, insisted that she can taste the difference of milk before tea vs tea before milk. Fisher was skeptical, so he conducted the following experiment. Experiment:

- Take 8 cups of tea. 4 are milk before tea, 4 are tea before milk. Present them to Dr. Bristol in a random order.
- Dr. Bristol tastes all the cups, and then give her guess.
- Record the number of correct guesses.

Results: Dr. Bristol correctly guessed all 8.
Question: did she get lucky, or could she really tell the difference?

Classical Fisher's exact test

In 1935, Dr. Bristol, a British biologist, insisted that she can taste the difference of milk before tea vs tea before milk. Fisher was skeptical, so he conducted the following experiment. Experiment:

- Take 8 cups of tea. 4 are milk before tea, 4 are tea before milk. Present them to Dr. Bristol in a random order.
- Dr. Bristol tastes all the cups, and then give her guess.
- Record the number of correct guesses.

Results: Dr. Bristol correctly guessed all 8.
Question: did she get lucky, or could she really tell the difference?

Paraphrase: is 8/8 statistically significant?

Classical Fisher's exact test

Proof by hypothesis testing:

- Assume H_0 (common sense before data)
- State H_A (alternative theory)
- Choose S. Observe $S = s$ from data
- Compute p-value $= \mathbb{P}(S \geq s | H_0)$

The tea experiment

- H_0:
- H_A:
- S:

Classical Fisher's exact test

Proof by hypothesis testing:

- Assume H_0 (common sense before data)
- State H_A (alternative theory)
- Choose S. Observe $S = s$ from data
- Compute p-value $=$ $\mathbb{P}(S \geq s | H_0)$

The tea experiment

- H_0: Dr. Bristol guessed randomly
- H_A:
- S:

Classical Fisher's exact test

Proof by hypothesis testing:

- Assume H_0 (common sense before data)
- State H_A (alternative theory)
- Choose S. Observe $S = s$ from data
- Compute p-value $= \mathbb{P}(S \geq s | H_0)$

The tea experiment

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol was **more** correct than pure chance
- S:

Classical Fisher's exact test

Proof by hypothesis testing:

- Assume H_0 (common sense before data)
- State H_A (alternative theory)
- Choose S. Observe $S = s$ from data
- Compute p-value $= \mathbb{P}(S \geq s | H_0)$

The tea experiment

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol was **more** correct than pure chance
- S: Number of correct cups

Classical Fisher's exact test

Proof by hypothesis testing:

- Assume H_0 (common sense before data)
- State H_A (alternative theory)
- Choose S. Observe $S = s$ from data
- Compute p-value $= \mathbb{P}(S \geq s | H_0)$

The tea experiment

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol was **more** correct than pure chance
- S: Number of correct cups
- Compute: $\mathbb{P}(S \geq 8 | H_0)$

Classical Fisher's exact test

Guessed randomly = chose 4 cups uniformly at random out of 8 and declare that they are all milk before tea.

Ways to choose 4 cups at random:
Ways to choose all 4 cups correctly: .

Classical Fisher's exact test

Guessed randomly = chose 4 cups uniformly at random out of 8 and declare that they are all milk before tea.

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups correctly: .

Classical Fisher's exact test

Guessed randomly = chose 4 cups uniformly at random out of 8 and declare that they are all milk before tea.

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups correctly: 1.

Classical Fisher's exact test

Guessed randomly = chose 4 cups uniformly at random out of 8 and declare that they are all milk before tea.

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups correctly: 1.
Assumed: chose uniformly at random. Then,

$$\mathbb{P}(S \geq 8 | H_0) = 1/70 \approx 1.4\%$$

Conclusion?

Classical Fisher's exact test

Guessed randomly = chose 4 cups uniformly at random out of 8 and declare that they are all milk before tea.

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups correctly: 1.
Assumed: chose uniformly at random. Then,

$$\mathbb{P}(S \geq 8 | H_0) = 1/70 \approx 1.4\%$$

Conclusion? If H_0, then we will observe $S \geq s$ with probability 1.4%. This is pretty small, so we conclude that most likely, H_A is better than H_0 at explaining the data .

Classical Fisher's exact test

Guessed randomly = chose 4 cups uniformly at random out of 8 and declare that they are all milk before tea.

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups correctly: 1.
Assumed: chose uniformly at random. Then,

$$\mathbb{P}(S \geq 8 | H_0) = 1/70 \approx 1.4\%$$

Conclusion? If H_0 **(Dr. Bristol were to guess at random)**, then we will observe $S \geq s$ **(that she got all 8 cups correct)** with probability 1.4%. This is pretty small, so we conclude that most likely, H_A is better than H_0 at explaining the data **(Dr. Bristol performed better than random guessing)**.

What does it mean when a test 'fail'?

What if the lady got zero cups right?

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups wrong: .

.

What does it mean when a test 'fail'?

What if the lady got zero cups right?

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups wrong: 1.

.

What does it mean when a test 'fail'?

What if the lady got zero cups right?

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups wrong: 1.
Then,

$$\mathbb{P}(S \geq 0 | H_0) = 1.$$

Conclusion?

.

What does it mean when a test 'fail'?

What if the lady got zero cups right?

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups wrong: 1.
Then,

$$\mathbb{P}(S \geq 0 | H_0) = 1.$$

Conclusion?
If H_0, then we will observe $S \geq s$ with probability 1. This is pretty large, so we conclude that most likely, H_0 is better than H_A at explaining the data .

What does it mean when a test 'fail'?

What if the lady got zero cups right?

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups wrong: 1.
Then,

$$\mathbb{P}(S \geq 0 | H_0) = 1.$$

Conclusion?
If H_0 **(Dr. Bristol were to guess at random)**, then we will observe $S \geq s$ **(that she got 0 or more cups correct)** with probability 1. This is pretty large, so we conclude that most likely, H_0 is better than H_A at explaining the data **Dr. Bristol did no better than guessing at random**.

What does it mean when a test 'fail'?

What if the lady got zero cups right?

Ways to choose 4 cups at random: $\binom{8}{4} = 70$.
Ways to choose all 4 cups wrong: 1.
Then,

$$\mathbb{P}(S \geq 0 | H_0) = 1.$$

Conclusion?
If H_0 **(Dr. Bristol were to guess at random)**, then we will observe $S \geq s$ **(that she got 0 or more cups correct)** with probability 1. This is pretty large, so we conclude that most likely, H_0 is better than H_A at explaining the data **Dr. Bristol did no better than guessing at random**.
Test fail = evidence does not support H_A better than H_0
$\neq H_0$ is true.

The other alternative

What if Fisher suspected that Dr. Bristol is **extra bad** at guessing? (not extra good)?

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol was ~~more~~ **less** correct than pure chance
- S:

The other alternative

What if Fisher suspected that Dr. Bristol is **extra bad** at guessing? (not extra good)?

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol was ~~more~~ **less** correct than pure chance
- S: Number of **in**correct cups

The other alternative

What if Fisher suspected that Dr. Bristol is **extra bad** at guessing? (not extra good)?

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol was ~~more~~ **less** correct than pure chance
- S: Number of **in**correct cups

In this case, if she got all 8 cups right:

$$\mathbb{P}(incorrect \geq 0 | H_0) = 1.$$

ie: no evidence to reject H_0 (guess random) in favor of the alternative (tea fool).

The other alternative

What if Fisher suspected that Dr. Bristol is **extra bad** at guessing? (not extra good)?

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol was ~~more~~ **less** correct than pure chance
- S: Number of **in**correct cups

In this case, if she got all 8 cups right:

$$\mathbb{P}(incorrect \geq 0|H_0) = 1.$$

ie: no evidence to reject H_0 (guess random) in favor of the alternative (tea fool). Does NOT affirm that she guessed randomly.

The other alternative

What if Fisher suspected that Dr. Bristol is **extra bad** at guessing? (not extra good)?

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol was ~~more~~ **less** correct than pure chance
- S: Number of **in**correct cups

In this case, if she got all 8 cups right:

$$\mathbb{P}(incorrect \geq 0 | H_0) = 1.$$

ie: no evidence to reject H_0 (guess random) in favor of the alternative (tea fool). Does NOT affirm that she guessed randomly.
If she got all 8 cups wrong:

$$\mathbb{P}(S \geq 8 | H_0) = 1/70.$$

She is likely a tea fool!

Two-sided test

What if there is no reason to assume that Dr. Bristol is extra bad or extra good?

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol did NOT guess randomly
- S:

Two-sided test

What if there is no reason to assume that Dr. Bristol is extra bad or extra good?

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol did NOT guess randomly
- S: $|$Number of correct - number of incorrect$|$

Two-sided test

What if there is no reason to assume that Dr. Bristol is extra bad or extra good?

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol did NOT guess randomly
- S: $|$Number of correct - number of incorrect$|$

If she guessed it all wrong OR she guessed it all correct, then in both cases, $S = 8$. Then, $\mathbb{P}(S \geq 8 | H_0) = 2/70$.

Two-sided test

What if there is no reason to assume that Dr. Bristol is extra bad or extra good?

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol did NOT guess randomly
- S: $|$Number of correct - number of incorrect$|$

If she guessed it all wrong OR she guessed it all correct, then in both cases, $S = 8$. Then, $\mathbb{P}(S \geq 8|H_0) = 2/70$.

Why is it called two-sided test?

Two-sided test

What if there is no reason to assume that Dr. Bristol is extra bad or extra good?

- H_0: Dr. Bristol guessed randomly
- H_A: Dr. Bristol did NOT guess randomly
- S: $|$Number of correct - number of incorrect$|$

If she guessed it all wrong OR she guessed it all correct, then in both cases, $S = 8$. Then, $\mathbb{P}(S \geq 8|H_0) = 2/70$.

Why is it called two-sided test?

- H_A: she is better than random guess. Test: $correct - incorrect \geq s$
- H_A: she is worse than random guess. Test: $incorrect - correct \geq s$
- H_A: she is different from random guess. Test:

$$|S| \geq s = \{correct - incorrect \geq s\} \text{ or } \{incorrect - correct \geq s\}$$

Union of the previous two!

How to choose H_0 and H_A

- H_0: null hypothesis. Common sense before the data arrived.
- H_A: alternative hypothesis. Common sense if we knew H_0 were wrong BEFORE the data arrived.

One-sided or two-sided?

- Titanic's `sex` vs `survived`:
- Email's `spam` vs `inherit`:
- male/female vs (get A or not in this class):
- male/female vs (taller than 6ft or not):

How to choose H_0 and H_A

- H_0: null hypothesis. Common sense before the data arrived.
- H_A: alternative hypothesis. Common sense if we knew H_0 were wrong BEFORE the data arrived.

One-sided or two-sided?

- Titanic's `sex` vs `survived`: we heard strong rumors that females are rescued first. So H_A: females $>$ males is reasonable.
- Email's `spam` vs `inherit`:
- male/female vs (get A or not in this class):
- male/female vs (taller than 6ft or not):

How to choose H_0 and H_A

- H_0: null hypothesis. Common sense before the data arrived.
- H_A: alternative hypothesis. Common sense if we knew H_0 were wrong BEFORE the data arrived.

One-sided or two-sided?

- Titanic's `sex` vs `survived`: we heard strong rumors that females are rescued first. So H_A: females $>$ males is reasonable.
- Email's `spam` vs `inherit`: we have no reason to believe that spams contain more or less occurrence rate of the word "inherit". Two-sided H_A is reasonable.
- male/female vs (get A or not in this class):
- male/female vs (taller than 6ft or not):

How to choose H_0 and H_A

- H_0: null hypothesis. Common sense before the data arrived.
- H_A: alternative hypothesis. Common sense if we knew H_0 were wrong BEFORE the data arrived.

One-sided or two-sided?

- Titanic's `sex` vs `survived`: we heard strong rumors that females are rescued first. So H_A: females $>$ males is reasonable.
- Email's `spam` vs `inherit`: we have no reason to believe that spams contain more or less occurrence rate of the word "inherit". Two-sided H_A is reasonable.
- male/female vs (get A or not in this class): Two-sided
- male/female vs (taller than 6ft or not):

How to choose H_0 and H_A

- H_0: null hypothesis. Common sense before the data arrived.
- H_A: alternative hypothesis. Common sense if we knew H_0 were wrong BEFORE the data arrived.

One-sided or two-sided?

- Titanic's `sex` vs `survived`: we heard strong rumors that females are rescued first. So H_A: females $>$ males is reasonable.
- Email's `spam` vs `inherit`: we have no reason to believe that spams contain more or less occurrence rate of the word "inherit". Two-sided H_A is reasonable.
- male/female vs (get A or not in this class): Two-sided
- male/female vs (taller than 6ft or not): One-sided

How to choose H_0 and H_A

- H_0: null hypothesis. Common sense before the data arrived.
- H_A: alternative hypothesis. Common sense if we knew H_0 were wrong BEFORE the data arrived.

One-sided or two-sided?

- Titanic's `sex` vs `survived`: we heard strong rumors that females are rescued first. So H_A: females $>$ males is reasonable.
- Email's `spam` vs `inherit`: we have no reason to believe that spams contain more or less occurrence rate of the word "inherit". Two-sided H_A is reasonable.
- male/female vs (get A or not in this class): Two-sided
- male/female vs (taller than 6ft or not): One-sided

NEVER snoop the data before deciding on the test direction!

When should we reject H_0?

Problem: Reject vs accept = binary decision (yes/no).
p-value: a continuous number in $[0, 1]$. To produce a decision, need a cut-off α

$$p < \alpha \Rightarrow \textit{reject}.$$

This α is called the significance level of the yes/no hypothesis test.
Some common α: $0.05, 0.01$.

When should we reject H_0?

Problem: Reject vs accept = binary decision (yes/no).
p-value: a continuous number in $[0, 1]$. To produce a decision, need a cut-off α

$$p < \alpha \Rightarrow \textit{reject}.$$

This α is called the significance level of the yes/no hypothesis test.
Some common α: $0.05, 0.01$.

Another interpretation:
Significance level $= \mathbb{P}(\text{ reject } H_0 | H_0)$.
That is, $\alpha =$ probability that the test wrongly reject H_0!

WARNING: what NOT to do with p-values