

Large Scale Hypothesis Testing

CSE545 - Spring 2020
Stony Brook University

H. Andrew Schwartz

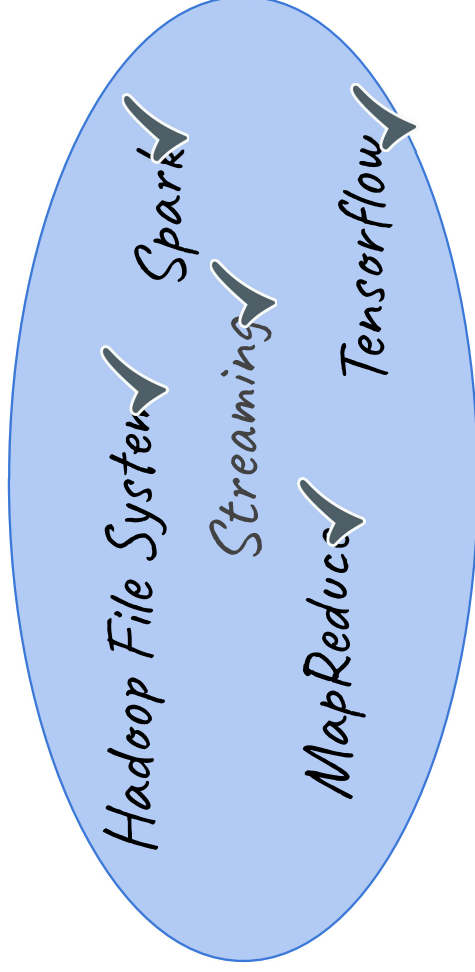


Big Data Analytics, The Class

Goal: Generalizations
A model or summarization of the data.



Data Frameworks



Algorithms and Analyses

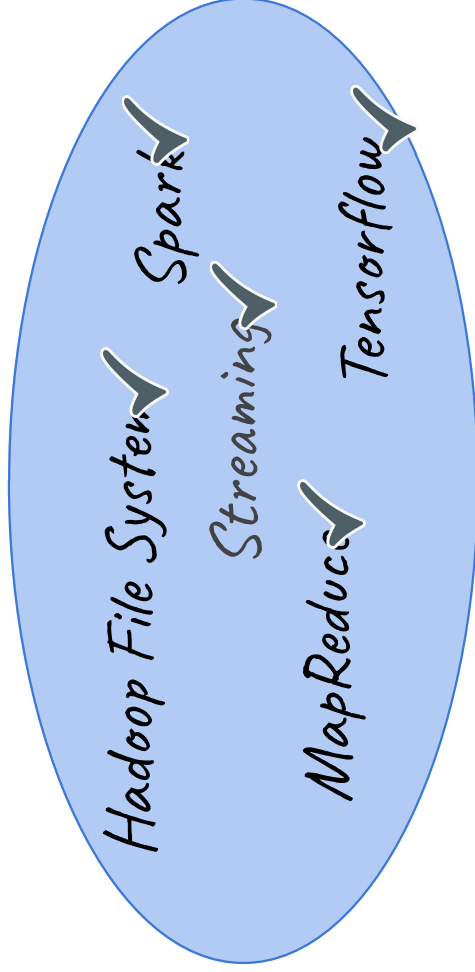


Big Data Analytics, The Class

Goal: Generalizations
A model or summarization of the data.



Data Frameworks



Algorithms and Analyses



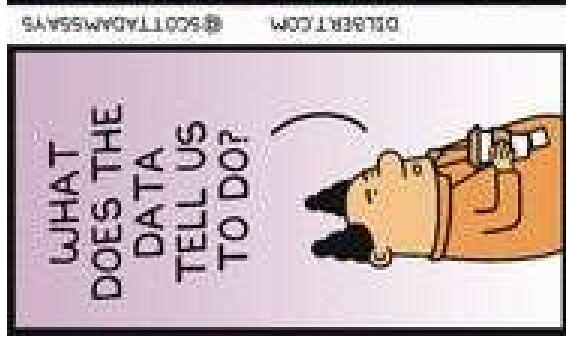
Goal of Data Science

Goal: Generalizations

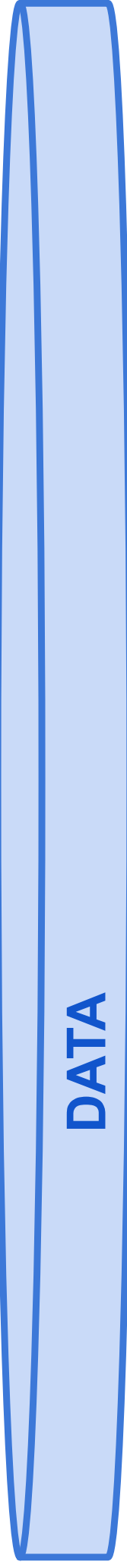
A *model* or *summarization* of the data.

The Data Whisperer

Goal: Generalizations
A *model* or *summarization* of the data.



Goal of Data Science



Goal: Generalizations
A *model* or *summarization* of the data.



Data-driven (evidence-based) decision

Goal of Data Science

DATA

Goal: Generalizations
A *model* or *summarization* of the data.

Discrete Finding(s)
 \mathcal{F} is (likely) True

Data-driven (evidence-based) **decision**

Goal of Data Science

DATA

Goal: Generalizations
A *model* or *summarization* of the data.

Discrete Finding(s)
 \mathcal{F} is (likely) True

Data-driven (evidence-based) **decision**

Blue cell phones cases are selling the most.

The ResImageGenNet model is most accurate.

Those >70 have a greater mortality rate from the viral infection.

Goal of Data Science

DATA

Goal: Generalizations
A *model* or *summarization* of the data.

Discrete Finding(s)
 F is (likely) True

Data-driven (evidence-based) **decision**

Hypotheses!
Potential findings -- to be tested
for happenstance.

Blue cell phones cases are
selling the most.

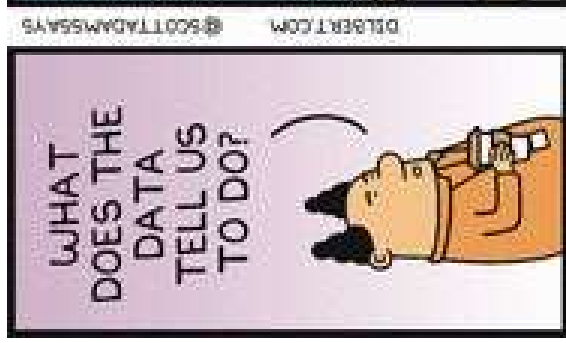
The ResImageGenNet
model is most accurate.

Those >70 have a greater
mortality rate from the
viral infection.

The Data Whisperer

Hypotheses!
Potential findings -- to be tested
for happenstance.

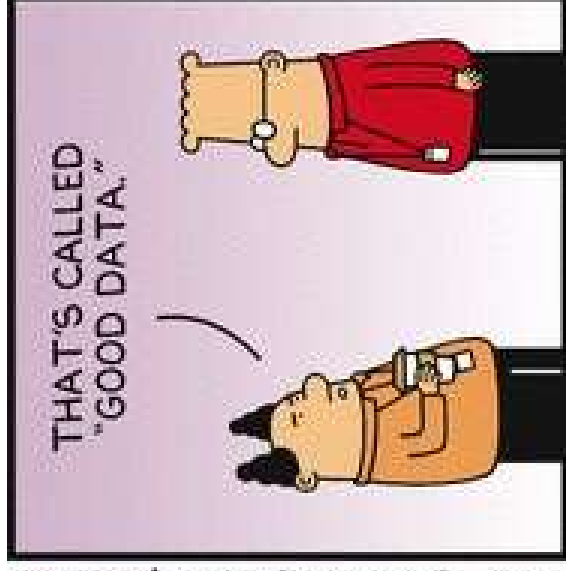
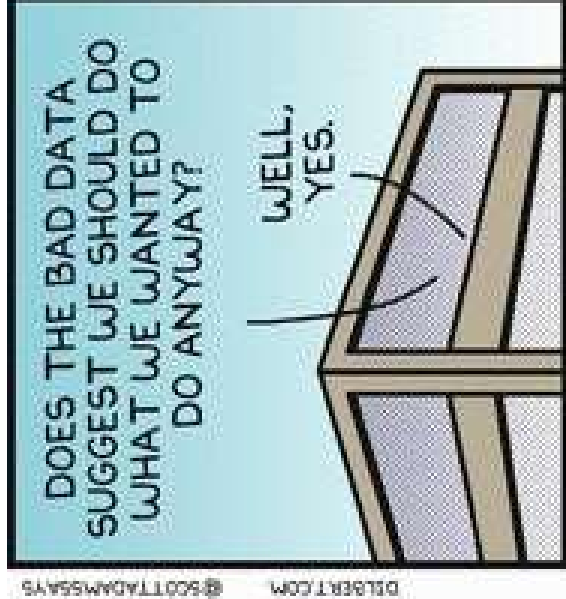
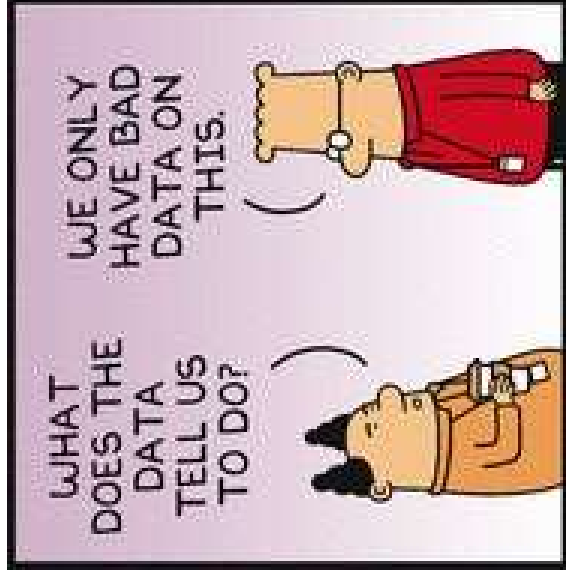
Goal: Generalizations
A *model* or *summarization* of the data.



The Data Whisperer

Hypotheses!
Potential findings -- to be tested
for happenstance.

Goal: Generalizations
A *model* or *summarization* of the data.



Hypothesis Testing

Also known as... “Don’t be Dilbert’s Boss!”

Hypothesis -- something one asserts to be true.

Hypothesis Testing

Also known as... “Don’t be Dilbert’s Boss!”

Hypothesis -- something one asserts to be true.

Formally, two types:

H_0 : *null hypothesis* -- some “default” value; “null”: nothing changes

H_1 : the *alternative* -- the opposite of the null => a change or difference

Hypothesis Testing

H_0 : *null hypothesis* -- some “default” value; “null”: nothing changes

H_1 : the alternative -- the opposite of the null => a change or difference

Goal: Make sure what we observed was unlikely to happen by chance.

Thus, we want to know:

Given null, what is the probability of the observation or worse

Hypothesis Testing

H_0 : *null hypothesis* -- some “default” value; “null”: nothing changes

H_1 : the alternative -- the opposite of the null => a change or difference

Goal: Make sure what we observed was unlikely to happen by chance.

Thus, we want to know:

Given null, what is the probability of the observation or worse?

-> If low enough, then we “reject the null (H_0) in favor of H_1 .”

Hypothesis Testing

H_0 : null hypothesis -- some “default” value; “null”: nothing changes

H_1 : the alternative -- the opposite of the null => a change or difference

Goal: Make sure what we observed was unlikely to happen by chance.

Thus, we want to know:

Given null, what is the probability of the observation or worse?

-> If low enough, then we “reject the null (H_0) in favor of H_1 .”

H_0 : *The blue case is not selling more than average.*

The Hypothesis Test “Algorithm”

observations (i.e. data)

Input: H_0 , obs, α

level of significance

Output: decision

H_0 : The blue case is not selling more than average.

The Hypothesis Test “Algorithm”

observations (i.e. data)

level of significance

Input: H_0 , obs, α

probability of what we observed or worse (i.e. more extreme)

$$p(x \geq \text{obs} \mid H_0) < \alpha$$

Output: decision

H_0 : The blue case is not selling more than average.

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

```
if  $p(x > \text{obs} \mid H_0) < \alpha$ :  
  decision = “Reject  $H_0$ !”  
else:  
  decision = “Accept  $H_0$ .”  
Output: decision
```

H_0 : The blue case is not selling more than average.

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

*Conditional is sometimes evaluated indirectly by first finding the “critical value”
of some measurement such that:
if measurement > critical_value then $p(\text{obs}/H_0) < \alpha$*

```
if  $p(x > \text{obs} \mid H_0) < \alpha$ :  
  decision = “Reject  $H_0$ !”  
else:  
  decision = “Accept  $H_0$ .”  
Output: decision
```

H_0 : The blue case is not selling more than average.

The Hypothesis Test “Algorithm”

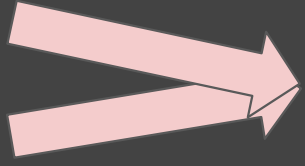
Input: H_0 , obs, α

```
if  $p(x > \text{obs} \mid H_0) < \alpha$ :  
  decision = “Reject  $H_0$ !”  
else:  
  decision = “Accept  $H_0$ .”  
Output: decision
```

H_0 : The blue case is not selling more than average.

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α



Need to estimate

What is the distribution of values we would expect if the null was true?
-- the “null distribution”

```
if  $p(x > \text{obs} \mid H_0) < \alpha$ :  
  decision = “Reject  $H_0$ !”  
else:  
  decision = “Accept  $H_0$ .”  
Output: decision
```

H_0 : The blue case is not selling more than average.

Probability Distributions: Review

X: A mapping from Ω to \mathbb{R} that describes the question we care about in practice.

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X is a *discrete random variable* if it takes only a countable number of values.

Probability Distributions: Review

X: A mapping from Ω to \mathbb{R} that describes the question we care about in practice.
↓

“sample space”, set of all possible outcomes.

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X is a *discrete random variable* if it takes only a countable number of values.

Probability Distributions: Review

X : A mapping from Ω to \mathbb{R} that describes the question we care about in practice.
↓

“sample space”, set of all possible outcomes.

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X is a *discrete random variable* if it takes only a countable number of values.

Error of RedImageGenNet Classifier

Amount of sales of a blue case

Continuous Distributions

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X is a *continuous random variable* if there exists a function f_X such that:

$$f_X(x) \geq 0, \text{ for all } x \in X,$$

$$\int_{-\infty}^{\infty} f_X(x) dx = 1, \text{ and}$$

$$P(a < X < b) = \int_a^b f_X(x) dx$$

Continuous Distributions

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X is a *continuous random variable* if there exists a function f_X such that:

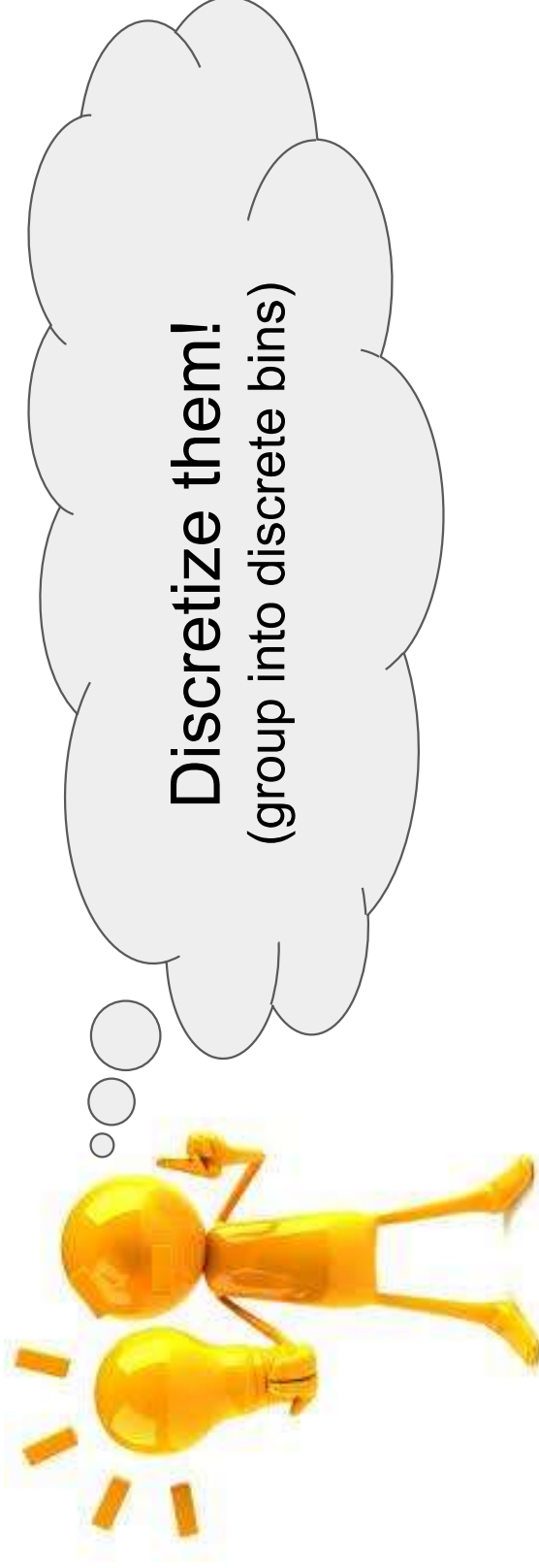
$$f_X(x) \geq 0, \text{ for all } x \in X,$$

$$\int_{-\infty}^{\infty} f_X(x) dx = 1, \text{ and}$$

$$P(a < X < b) = \int_a^b f_X(x) dx$$

f_X : “probability density function” (pdf)

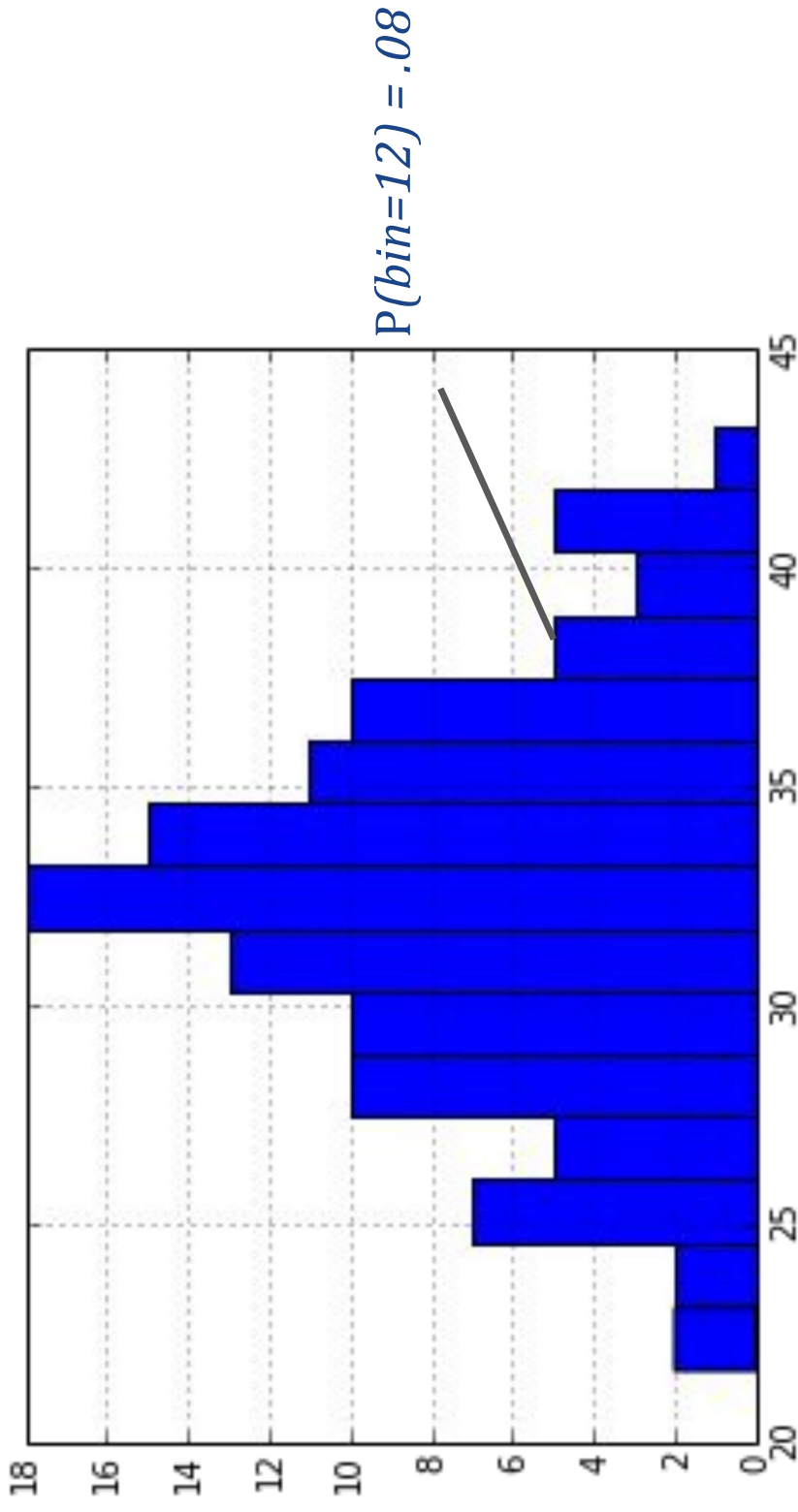
Continuous Distributions



How to model?

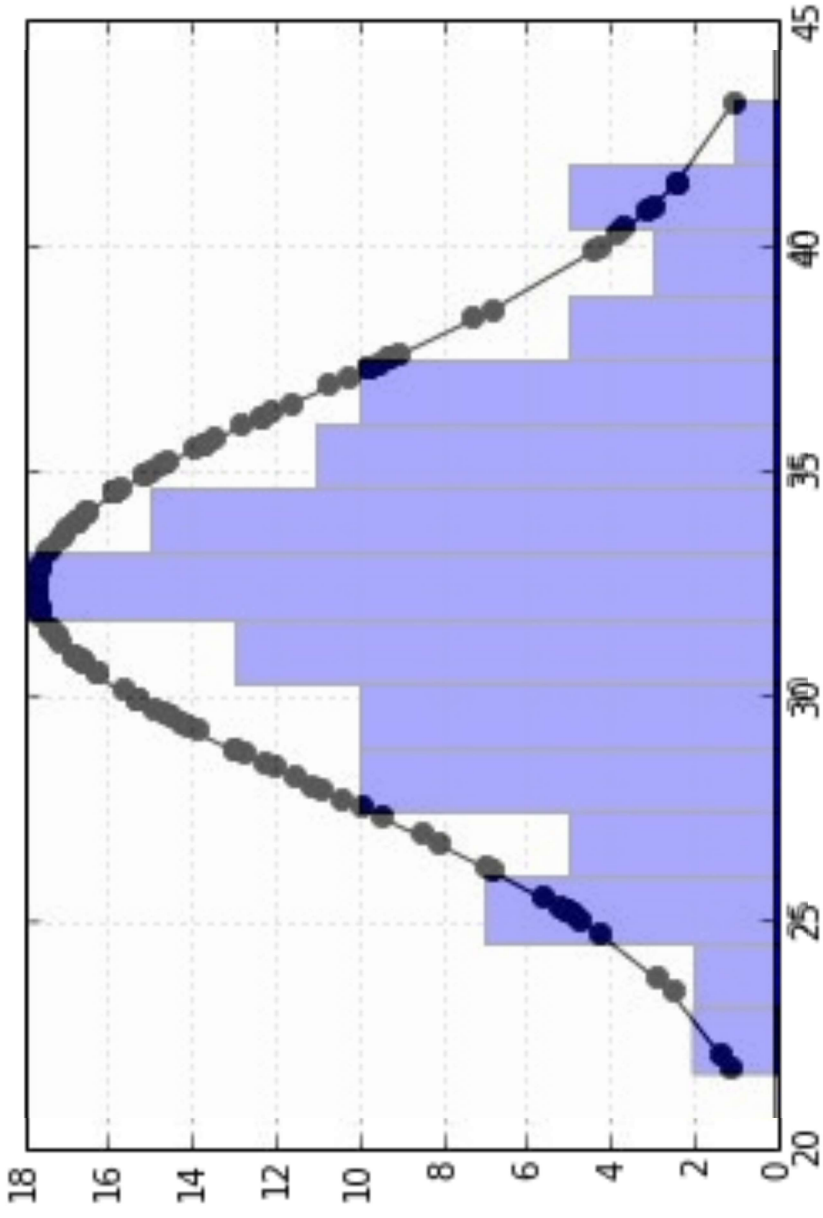
Continuous Distributions

$P(\text{bin}=8) = .32$

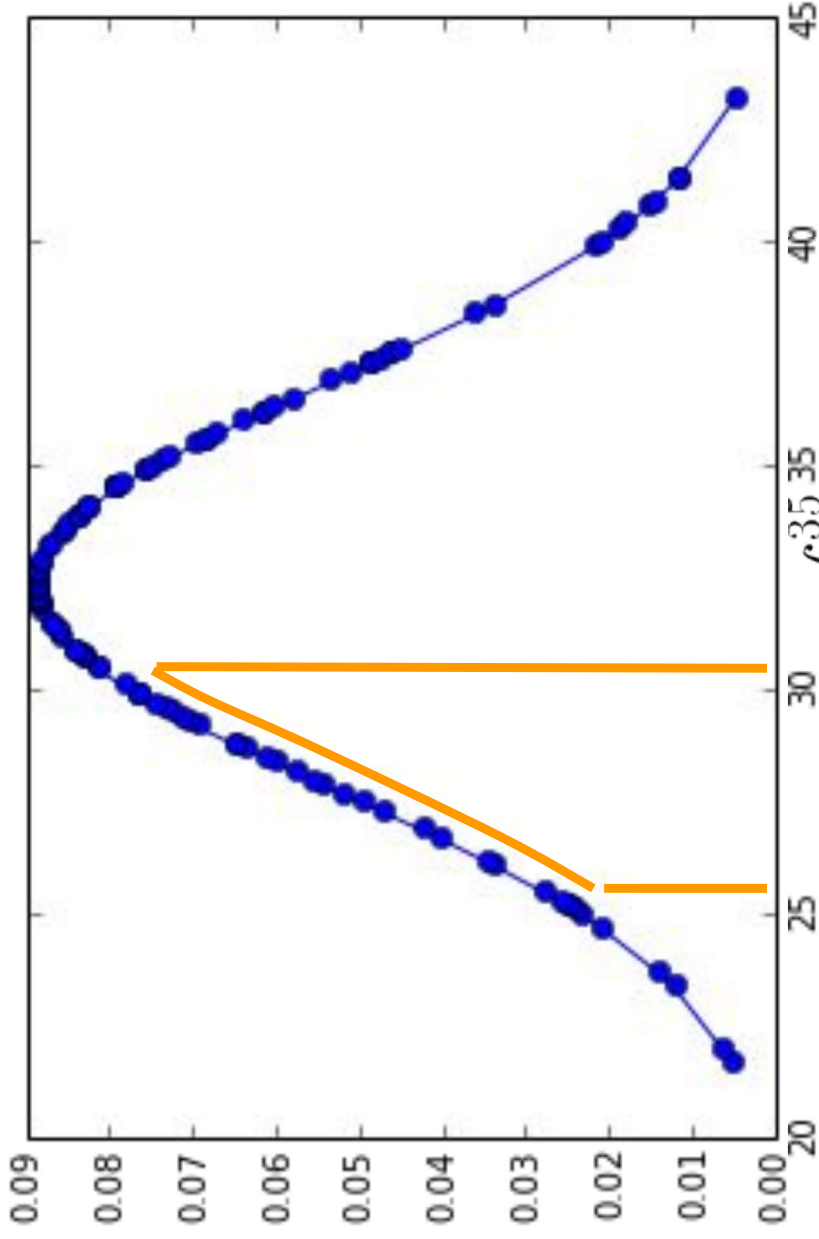


But aren't we throwing away information?

Continuous Distributions



Continuous Distributions

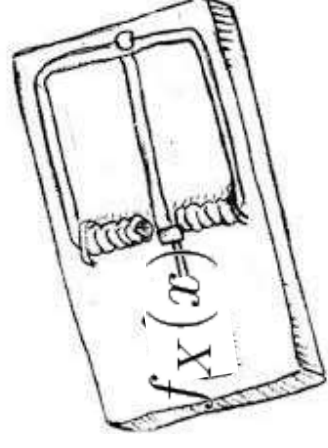


$$P(25 < X < 35) = \int_{25}^{35} f(x) dx$$

Continuous Distributions

Common Trap

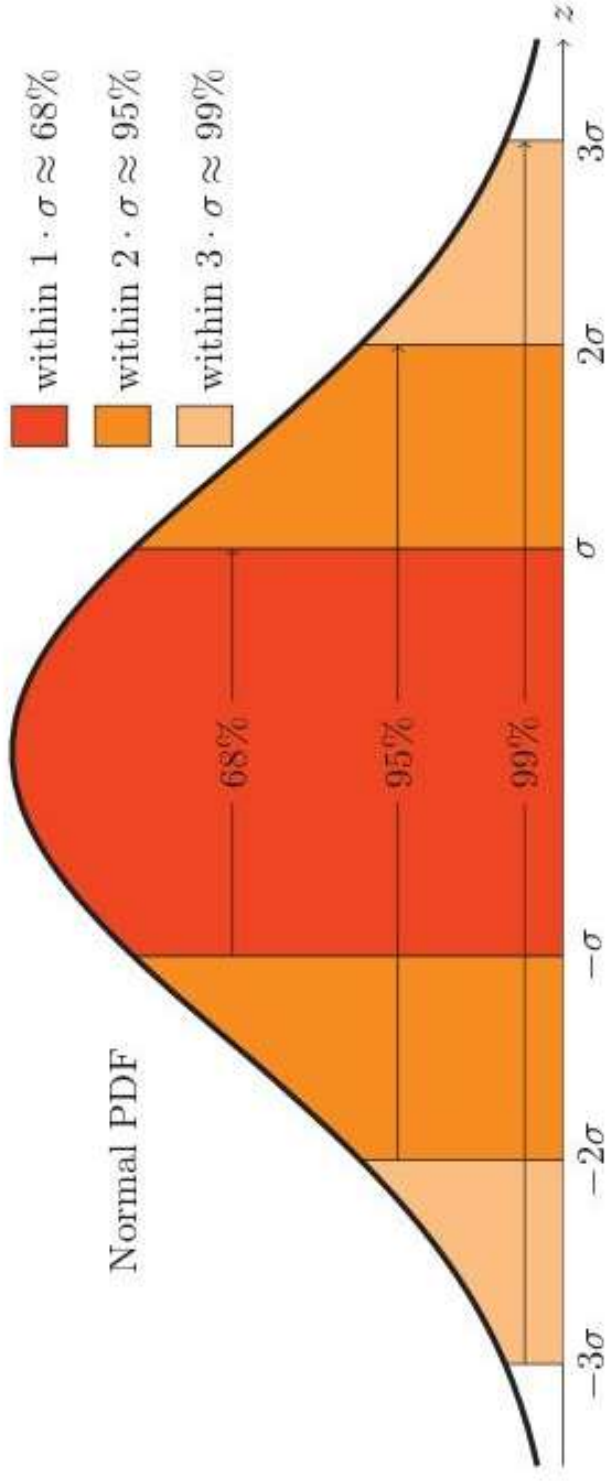
- $f_X(x)$ does not yield a probability
 - $\int_a^b f_X(x)dx$ does
 - x may be anything (\mathbb{R})
 - thus, $f_X(x)$ may be > 1



Continuous Distributions

Common pdfs: Normal(0, 1)

$$P(-1 \leq Z \leq 1) \approx .68, \quad P(-2 \leq Z \leq 2) \approx .95, \quad P(-3 \leq Z \leq 3) \approx .99$$



Continuous Distributions

Common *pdfs*: Normal(0, 1) (“standard normal”)

How to “standardize” any normal distribution:

1. subtract the mean, μ (aka “mean centering”)
2. divide by the standard deviation, σ

$$z = (x - \mu) / \sigma, \text{ (aka “z score”)}$$

Probability Distributions: Review

X : A mapping from Ω to \mathbb{R} that describes the question we care about in practice.
↓

“sample space”, set of all possible outcomes.

X is a *continuous random variable* if it can take on an infinite number of values between any two given values.

X is a *discrete random variable* if it takes only a countable number of values.

Error of RedImageGenNet Classifier

Amount of sales of a blue case

Discrete Random Variables

For a given discrete random variable X , probability mass function (pmf),

$f_X: \mathbb{R} \rightarrow [0, 1]$, is defined by:

$$f_X(x) = P(X = x)$$

X is a discrete random variable if it takes only a countable number of values.

Amount of sales of a blue case

Was a single sale a blue case: $\{0, 1\}$

Discrete Random Variables

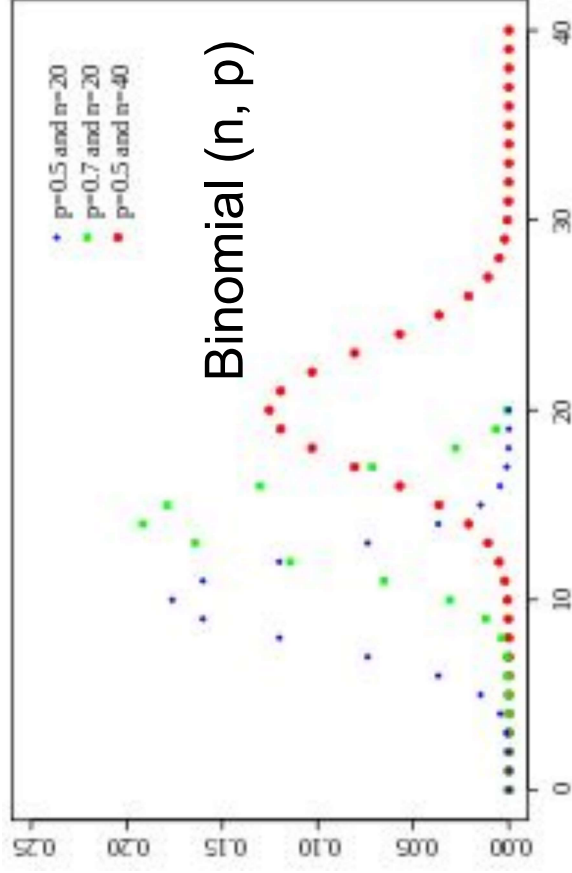
For a given discrete random variable X , probability mass function (pmf), $f_X: \mathbb{R} \rightarrow [0, 1]$, is defined by:

$$f_X(x) = P(X = x)$$

X is a discrete random variable if it takes only a countable number of values.

Amount of sales of a blue case

Was a single sale a blue case: $\{0, 1\}$



Discrete Random Variables

For a given random variable X , the *cumulative distribution function (CDF)*,

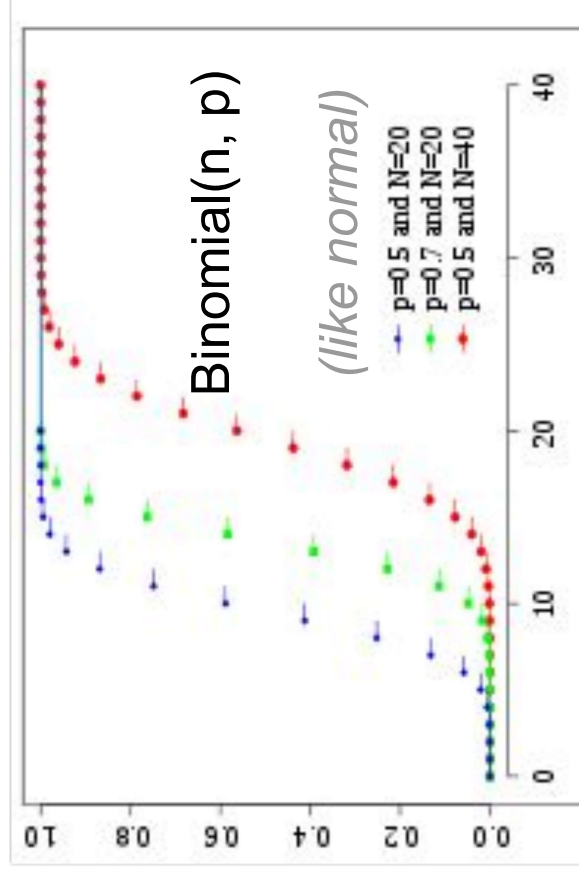
$F_X: \mathbb{R} \rightarrow [0, 1]$, is defined by:

$$F_X(x) = P(X \leq x)$$

X is a *discrete random variable* if it takes only a countable number of values.

Amount of sales of a blue case

Was a single sale a blue case: $\{0, 1\}$

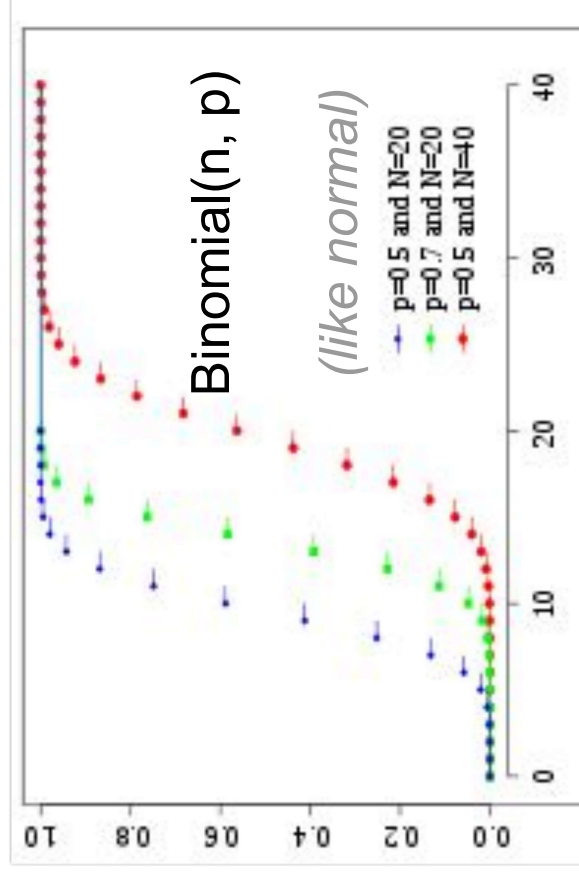


Discrete Random Variables

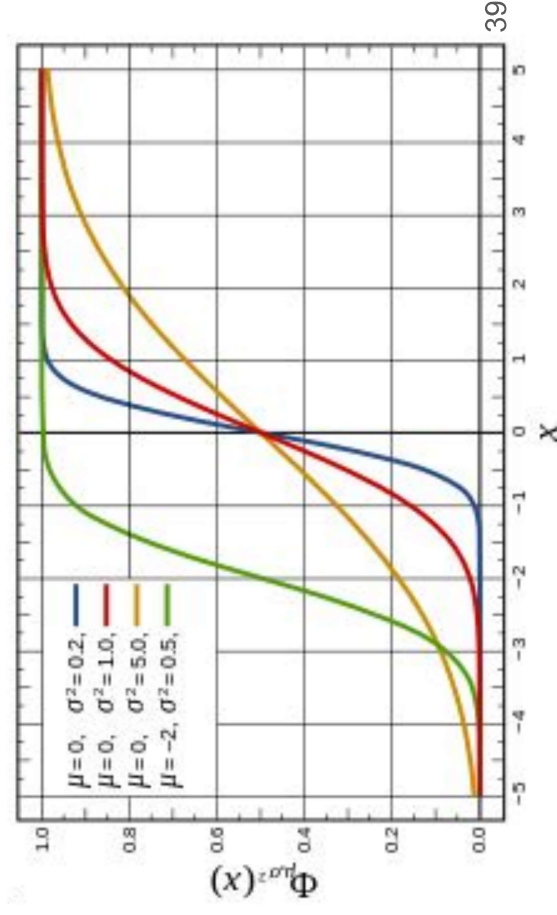
For a given random variable X , the *cumulative distribution function (CDF)*,

$F_X: \mathbb{R} \rightarrow [0, 1]$, is defined by:

$$F_X(x) = P(X \leq x)$$



Normal



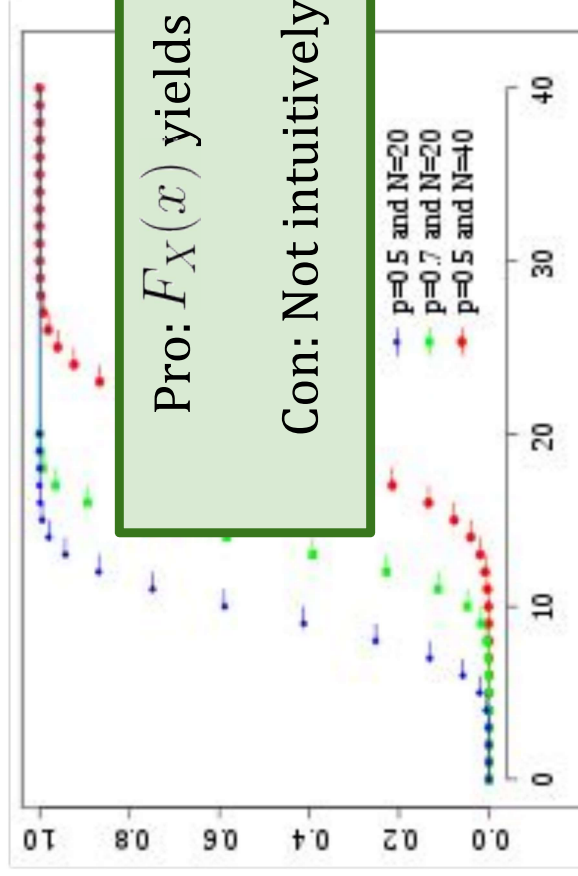
Discrete Random Variables

For a given random variable X , the *cumulative distribution function (CDF)*,

$F_X: \mathbb{R} \rightarrow [0, 1]$, is defined by:

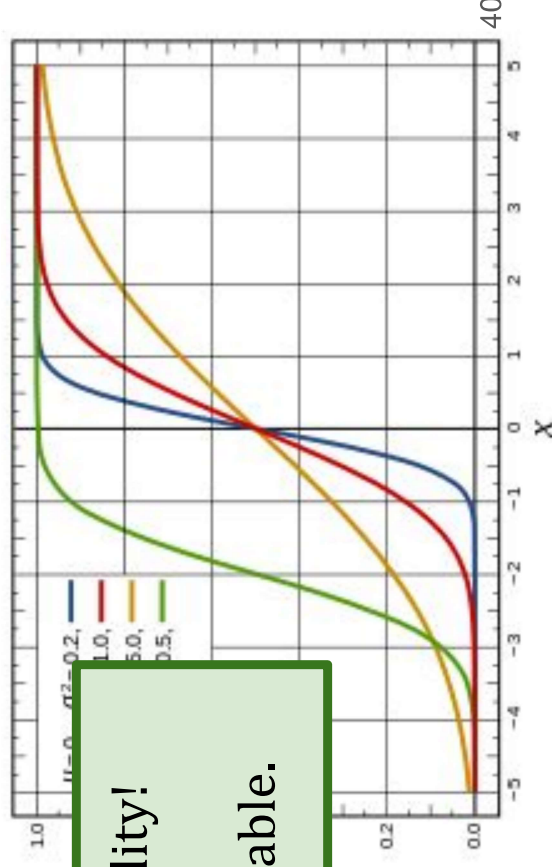
$$F_X(x) = P(X \leq x)$$

Normal



Pro: $F_X(x)$ yields a probability!

Con: Not intuitively interpretable.



Discrete RVs

For a given random variable X , the *cumulative distribution function* (CDF),

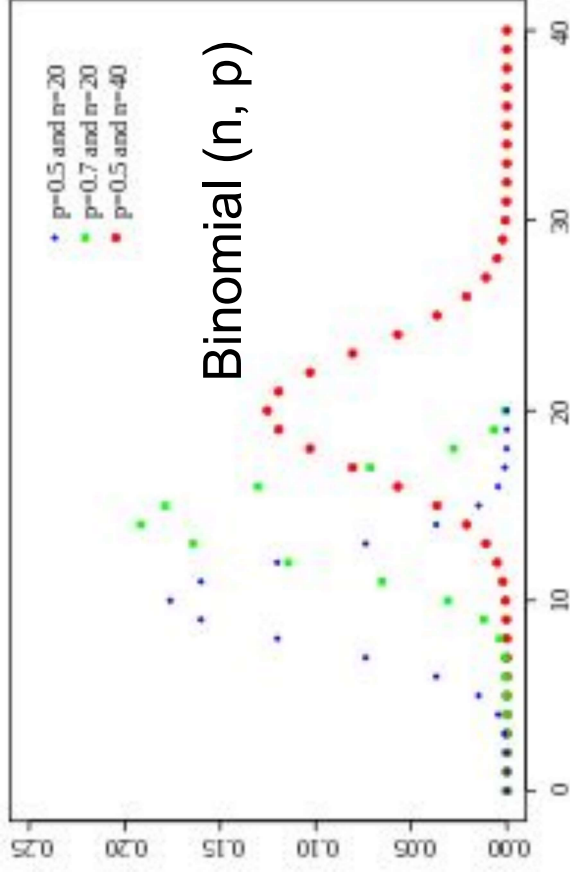
$F_X: \mathbb{R} \rightarrow [0, 1]$, is defined by:

$$F_X(x) = P(X \leq x)$$

For a given *discrete* random variable X , *probability mass function* (pmf),

$f_X: \mathbb{R} \rightarrow [0, 1]$, is defined by:

$$f_X(x) = P(X = x)$$



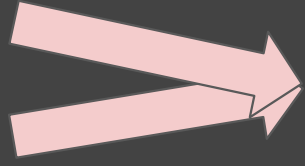
X is a *discrete random variable* if it takes only a countable number of values.

$$\sum_i f_X(x) = 1$$

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x)$$

The Hypothesis Test “Algorithm”

Input: H_0 , observations, α



Need to estimate

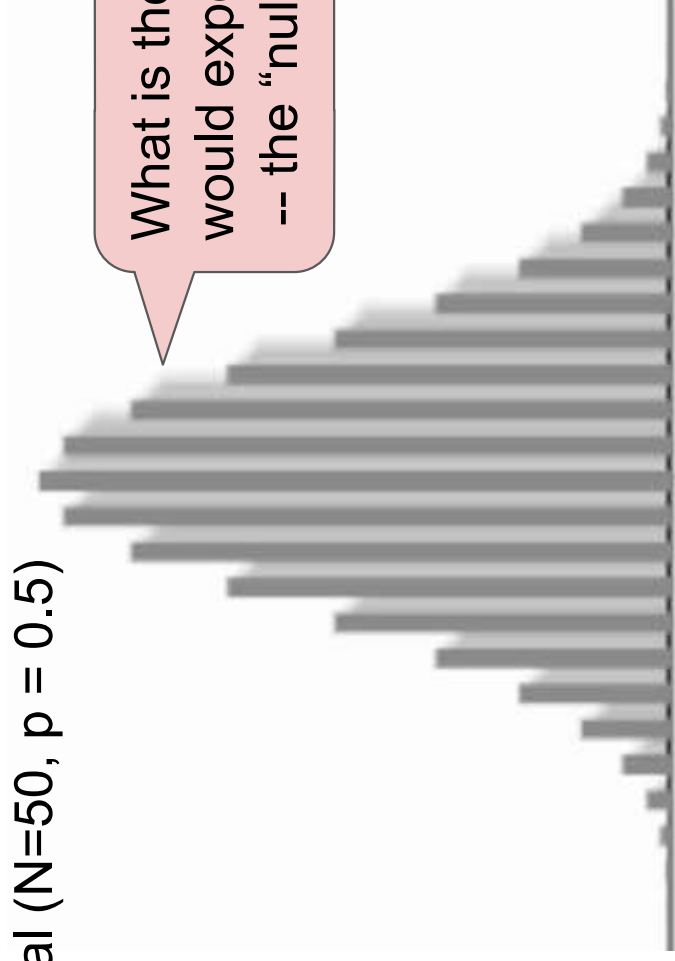
What is the distribution of values we would expect if the null was true?
-- the “null distribution”

```
if  $p(x > \text{obs} \mid H_0) < \alpha$ :  
  decision = “Reject  $H_0$ !”  
else:  
  decision = “Accept  $H_0$ .”  
Output: decision
```

H_0 : The blue case is not selling more than average.

The Hypothesis Test “Algorithm”

Binomial (N=50, p = 0.5)
PMF



What is the distribution of values we would expect if the null was true?
-- the “null distribution”

25

H_0 : The blue case is not selling more than average.
50 sales; 2 colors (blue and red); Thus, average would be 25 blue sales

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

null_dist = distribution of expected values under H_0

if $p(x \geq \text{obs} \mid H_0) < \alpha$:
 decision = “Reject H_0 !”

else:

 decision = “Accept H_0 .”

Output: decision

H_0 : The blue case is not selling more than average.

50 sales; 2 colors (blue and red); Thus, average would be 25 blue sales

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

null_dist = distribution of expected values under H_0

if $p(x \geq \text{obs} \mid H_0) < \alpha$:
decision = “Reject H_0 !”

else:

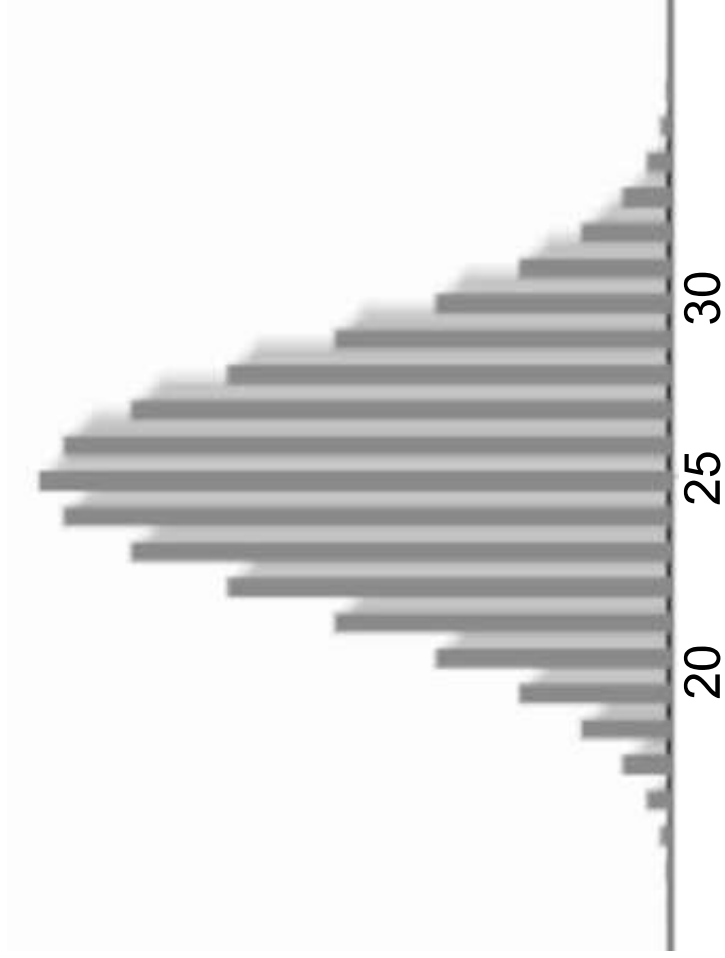
decision = “Accept H_0 .”

Output: decision

H_0 : The blue case is not selling more than average. Observed 32 blue sales

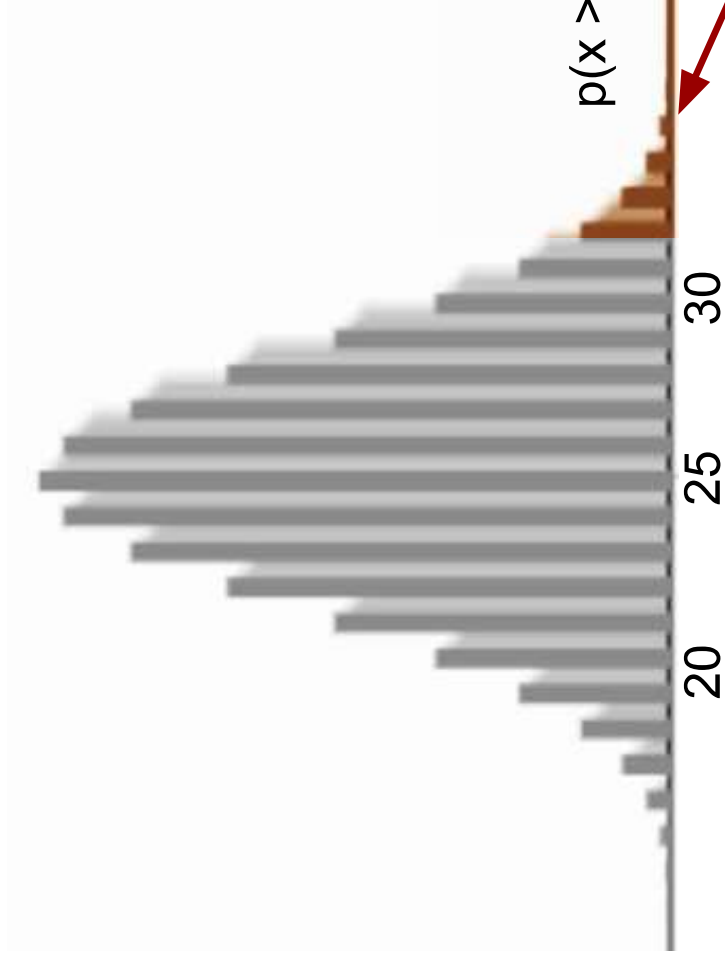
50 sales; 2 colors (blue and red); Thus, average would be 25 blue sales

The Hypothesis Test “Algorithm”



H_0 : The blue case is not selling more than average. Observed 32 blue sales
50 sales; 2 colors (blue and red); Thus, average would be 25 blue sales

The Hypothesis Test “Algorithm”



$$p(x \geq 32; \text{Binomial}(50, 0.5)) = 0.016$$

H_0 : The blue case is not selling more than average. **Observed 32 blue sales**
50 sales; 2 colors (blue and red); Thus, average would be 25 blue sales

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

null_dist = distribution of expected values under H_0

$p(x \geq \text{obs} \mid H_0) =$

if $p(x \geq \text{obs} \mid H_0) < \alpha$:

 decision = “Reject H_0 !”

else:

 decision = “Accept H_0 .”

Output: decision

H_0 : The blue case is not selling more than average. *Observed 36 blue sales*

50 sales; 2 colors (blue and red); Thus, average would be 25 blue sales

The Hypothesis Test “Algorithm”

```
Input:  $H_0$ , obs,  $\alpha$   
null_dist = distribution of expected values under  $H_0$   
 $p(x \geq \text{obs} \mid H_0) = \text{sum}(\text{pmf}(\text{null\_dist}, o) \text{ for } o \text{ in range}(\text{obs},))$   
if  $p(x \geq \text{obs} \mid H_0) < \alpha$ :  
    decision = “Reject  $H_0$ !”  
else:  
    decision = “Accept  $H_0$ .”  
Output: decision
```

H_0 : The blue case is not selling more than average. Observed 36 blue sales
50 sales; 2 colors (blue and red); Thus, average would be 20 blue sales

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

null_dist = distribution of expected values under H_0

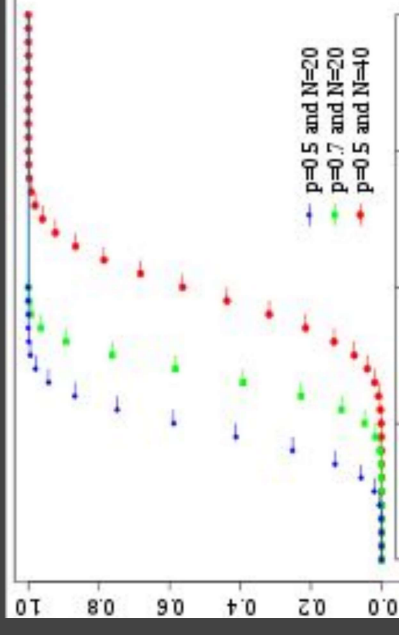
$p(x \geq \text{obs} \mid H_0) = 1 - \text{cdf}(\text{null_dist}, \text{obs})$

if $p(x \geq \text{obs} \mid H_0) < \alpha$:
decision = “Reject H_0 !”

else:

decision = “Accept H_0 .”

Output: decision



H_0 : The blue case is not selling more than average. Observed 36 blue sales

50 sales; 2 colors (blue and red); Thus, average would be 20 blue sales

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

null_dist = distribution of expected values under H_0

$p(x \leq \text{obs} \mid H_0) = \text{cdf}(\text{null_dist}, \text{obs})$

if $p(x \leq \text{obs} \mid H_0) < \alpha$:

 decision = “Reject H_0 !”

else:

 decision = “Accept H_0 .”

Output: decision

H_0 : The blue case is not selling less than average. Observed 36 blue sales

50 sales; 2 colors (blue and red); Thus, average would be 25 blue sales

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

null_dist = distribution of expected obs under H_0

$p(x \leq \text{obs} \mid H_0) = \text{cdf}(\text{null_dist}, \text{obs})$

if $p(x \leq \text{obs} \mid H_0) < \alpha$:

 decision = “Reject H_0 !”

else:

 decision = “Accept H_0 .”

Output: decision

H_0 : The blue case is not selling less than average. Observed 36 blue sales

50 sales; 2 colors (blue and red); Thus, average would be 25 blue sales

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

obs_ts = test_stat(obs)

null_dist = distribution of expected obs under H_0

$p(x \leq \text{obs} \mid H_0) = \text{cdf}(\text{null_dist}, \text{obs})$

if $p(x \leq \text{obs} \mid H_0) < \alpha$:

 decision = “Reject H_0 !”

else:

 decision = “Accept H_0 .”

Output: decision

H_0 : The blue case is not selling less than average. Observed 36 blue sales

50 sales; 2 colors (blue and red); Thus, average would be 25 blue sales

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

obs_ts = test_stat(obs)

null_dist = distribution of expected **test_stat** under H_0

p(x ≤ obs_ts | H_0) = cdf(null_dist, obs_ts)

if **p(x ≤ obs_ts | H_0) < α** :

decision = “Reject H_0 !”

else:

decision = “Accept H_0 .”

Output: decision

H_0 : The blue case is not selling less than average. Observed 36 blue sales

50 sales; 2 colors (blue and red); Thus, average would be 25 blue sales

Hypothesis Testing

Why?

Hypothesis Testing

Why?

A general framework for answering (yes/no) questions!

Hypothesis Testing

Why?

A general framework for answering (yes/no) questions!

- *Are height and baldness related?*
- *Is my deep predictive model better than the state of the art?*

Hypothesis Testing

Why?

A general framework for answering (yes/no) questions!

- *Are height and baldness related?*
- *Is my deep predictive model better than the state of the art?*
- *Is the heat index of a community related to poverty?*
- *Is the heat index of a community related to poverty controlling for education rates?*
- *Does my website receive a higher average number of monthly visitors?*

Hypothesis Testing

Failing to “reject the null” does not mean the null is true.

Why?

A general framework for answering (yes/**maybe**) questions!

- Are height and baldness related?
- Is my deep predictive model better than the state of the art?
- Is the heat index of a community related to poverty?
- Is the heat index of a community related to poverty **controlling for education rates?**
- Does my website receive a higher average number of monthly visitors?

Hypothesis Testing

Failing to “reject the null” does not mean the null is true. However, if the sample is large enough, it may be enough to say that the effect size (correlation, difference value, etc...) is not very meaningful.

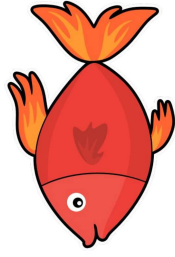
Why?

A general framework for answering (yes/**maybe**) questions!

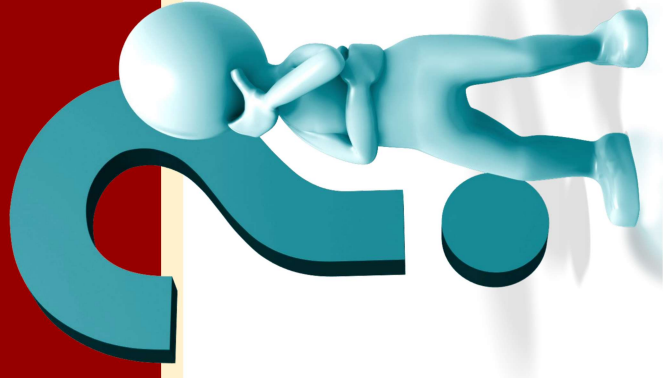
- Are height and baldness related?
- Is my deep predictive model better than the state of the art?
- Is the heat index of a community related to poverty?
- Is the heat index of a community related to poverty **controlling for education rates?**
- Does my website receive a higher average number of monthly visitors?

Bonferroni's Cats

General Question: Which fish do cats like?



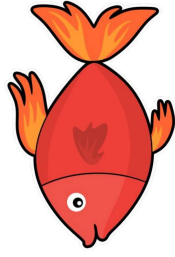
⋮



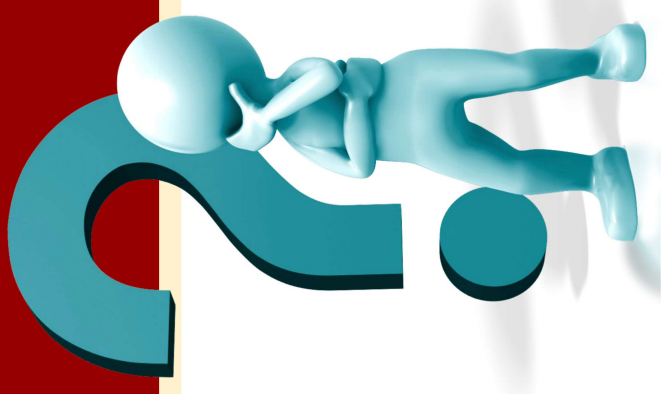
Bonferroni's Cats

General Question: Which fish do cats like?

$N = 100$ cats



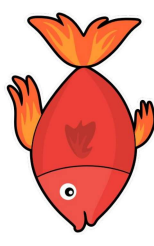
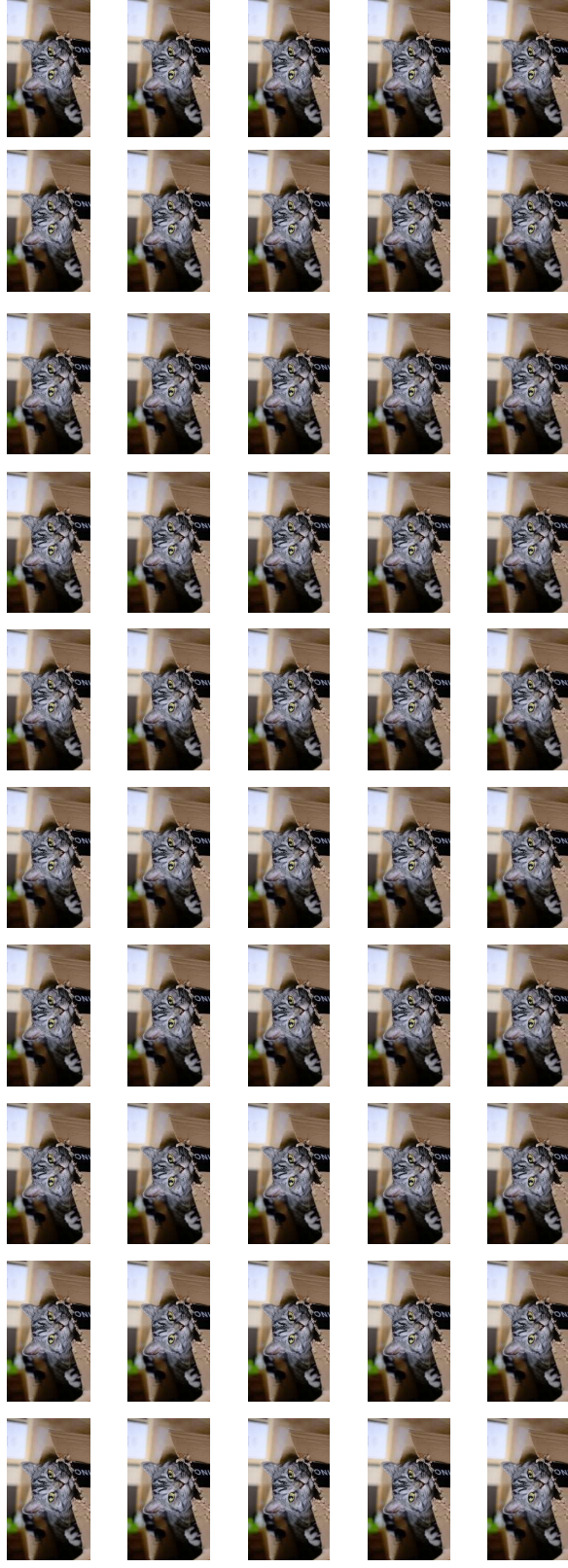
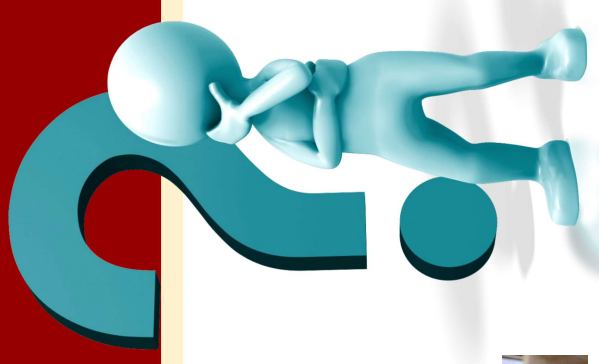
⋮



Bonferroni's Cats

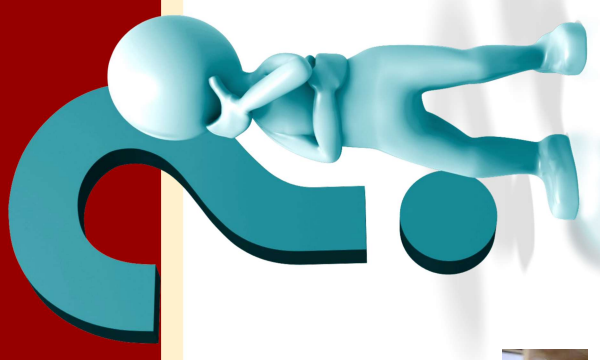
General Question: Which fish do cats like?

$N = 50$ cats



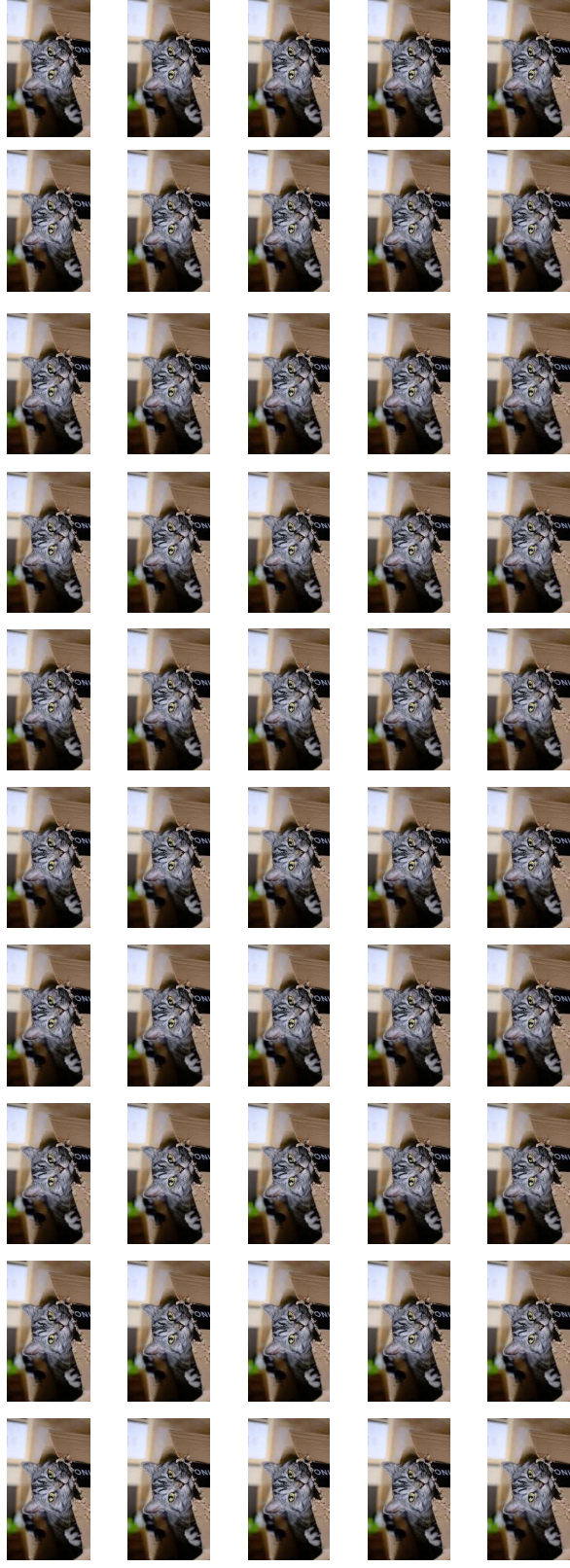
,
=

Bonferroni's Cats



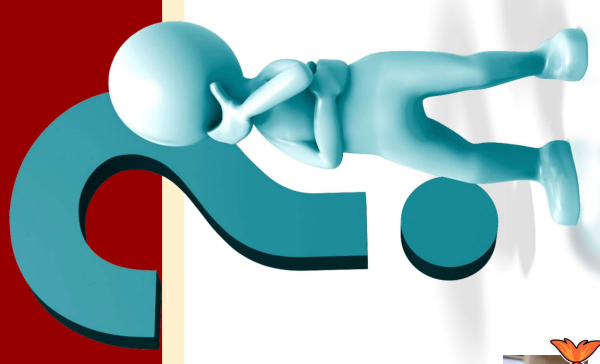
General Question: Which fish do cats like?

$N = 50$ cats



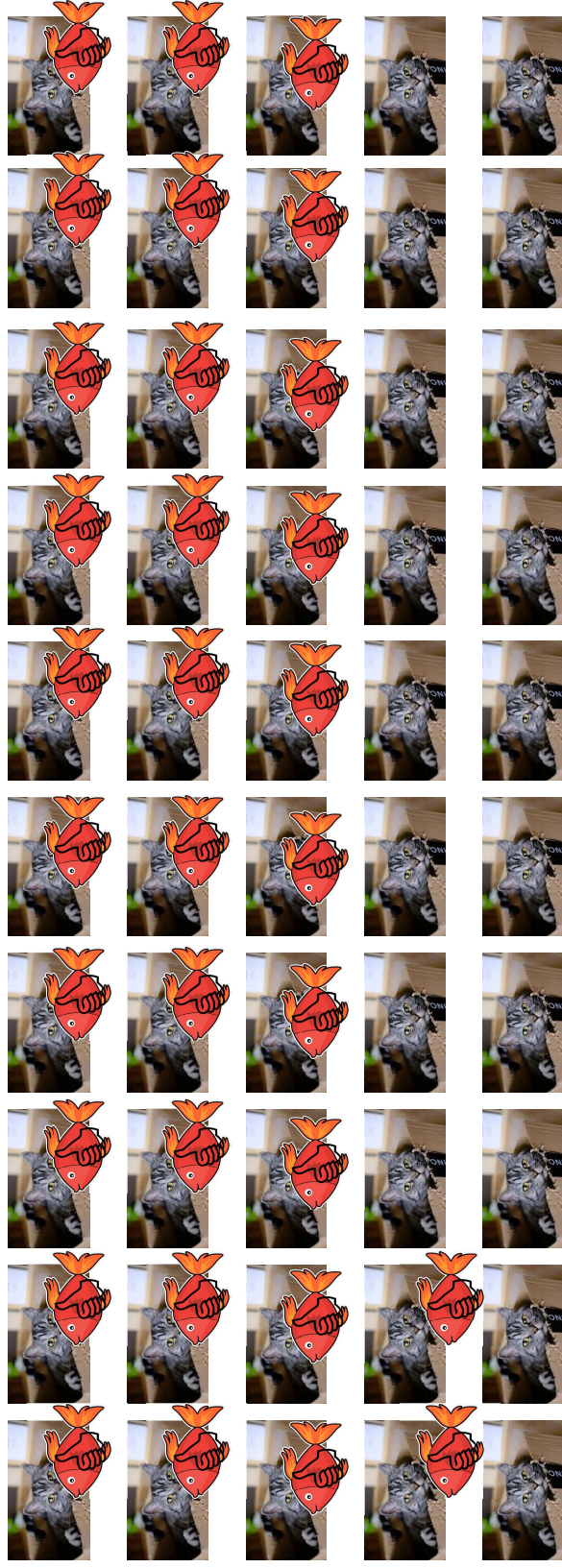
H_1 : Most cats like redfish. H_0 : Most cats don't like redfish.

Bonferroni's Cats



General Question: Which fish do cats like?

$N = 50$ cats; 32 like redfish; $p = 0.016$



H_1 : Most cats like redfish. H_0 : Most cats don't like redfish.

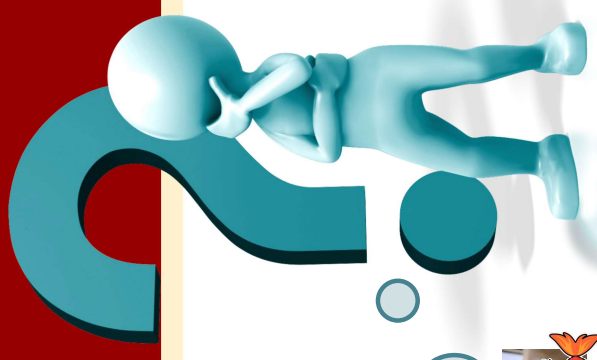
Bonferroni's Cats

General Question: Which fish do cats like?

$N = 50$ cats; 32 like redfish; $p = 0.016$

Now suppose instead of just redfish, you wanted to ask the same question for 10 kinds of fish: $H_{1,1}$: Most cats like redfish; $H_{1,2}$: Most cats like bluefish; $H_{1,3}$: Most cats like orangefish; ... with $\alpha = 0.05$, can you still conclude most cats like redfish?

H_1 : Most cats like redfish. H_2 : Most cats don't like redfish.



Bonferroni's Cats

General Question: Which fish do cats like?

$N = 50$ cats; 32 like redfish; $p = 0.016$

Now suppose instead of just redfish, you wanted to ask the same question for 10

kinds of fish: $H_{1,1}$: Most cats like redfish;

$H_{1,2}$: Most cats like bluefish; $H_{1,3}$: Most

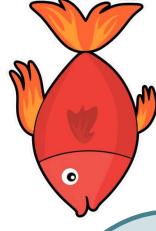
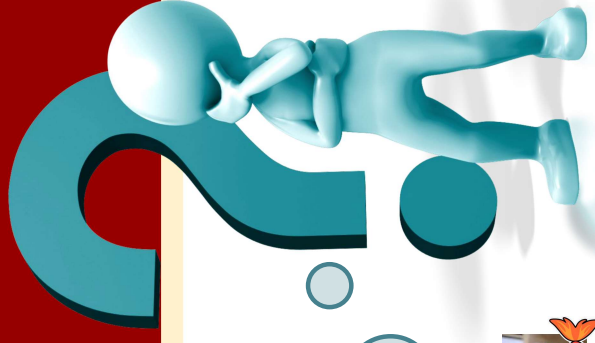
cats like orangefish; ... with $\alpha = 0.05$, can

you still conclude most cats like redfish?

hint: $P(1 \text{ sig}) = 1 - P(\text{no sig}) = 1 - (1 - 0.05)^{10} = 0.40$

don't like redfish.

H_1 : Most cats



Bonferroni's Cats

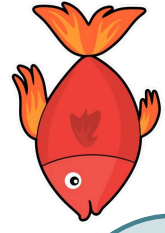
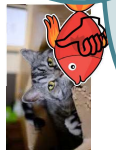
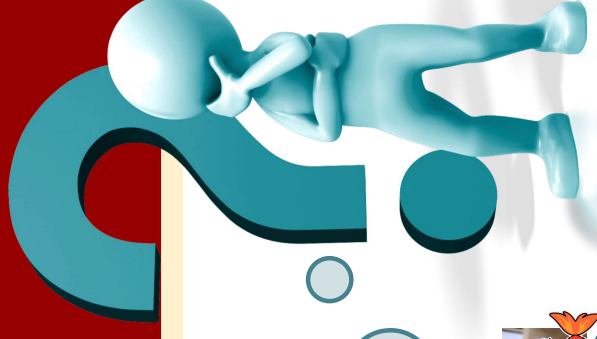
General Question: Which fish do cats like?

$N = 50$ cats; *32 like redfish*; $p = 0.016$

$\alpha = 0.05$ -- probability threshold for happening upon the result even if it really doesn't exist.

What is the probability we happen upon once in ten times?

H₁: Most cats don't like redfish.



Bonferroni's Cats

General Question: Which fish do cats like?

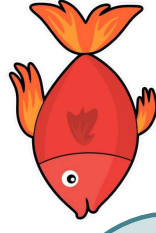
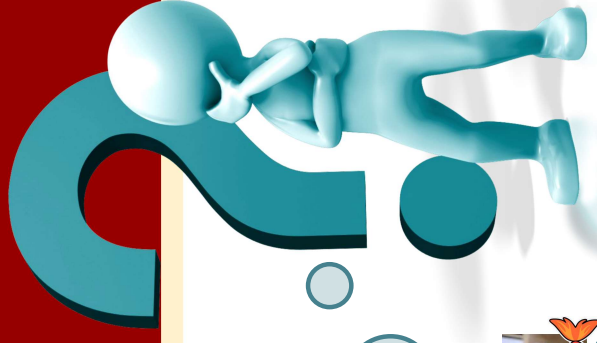
$N = 50$ cats; *32 like redfish*; $p = 0.016$

$\alpha = 0.05$ -- probability threshold for happening upon the result even if it really doesn't exist.

What is the probability we happen upon once in ten times?

$$1 - p(\text{not happening upon the result}) = 1 - (1 - .05)^{10} \\ = 1 - 0.599 = .4$$

H₁: Most cats don't like redfish.



Bonferroni's Cats

General Question: Which fish do cats like?

$N = 50$ cats; *32 like redfish*; $p = 0.016$

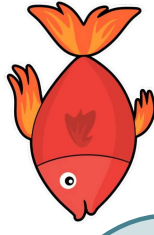
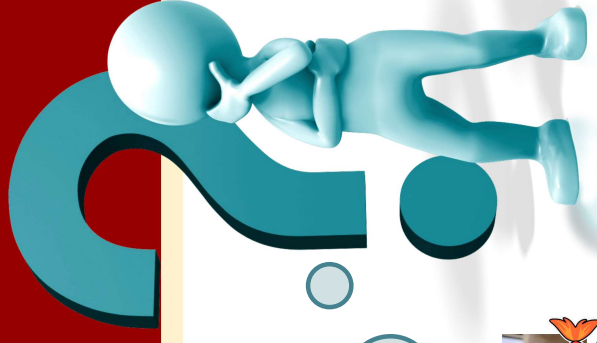
$\alpha = 0.05$ -- probability threshold for happening upon the result even if it really doesn't exist.

What is the probability we happen upon once in ten times?

$$1 - p(\text{not happening upon the result}) = 1 - (1 - .05)^{10} \\ = 1 - 0.599$$

How to fix?

H_1 : *Most cats don't like redfish.*



Bonferroni's Cats

General Question: Which fish do cats like?

$N = 50$ cats; *32 like redfish*; $p = 0.016$

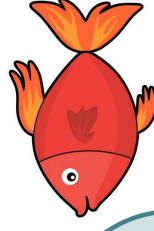
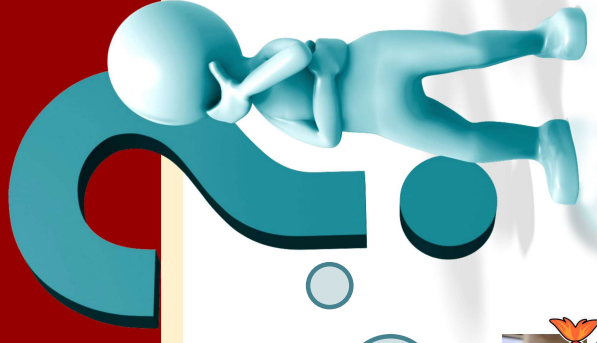
$\alpha = 0.05$ -- probability threshold for happening upon the result even if it really doesn't exist.

What is the probability we happen upon once in ten times?

$$1 - p(\text{not happening upon the result}) = 1 - (1 - .05)^{10} = 0.599$$

How to fix? $1 - (1 - \text{adjust}(.05))^{10} < .05$

H_1 : *Most cats don't like redfish.*



Bonferroni's Cats

General Question: Which fish do cats like?

$N = 50$ cats; 32 like redfish; $p = 0.016$

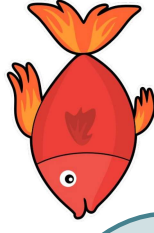
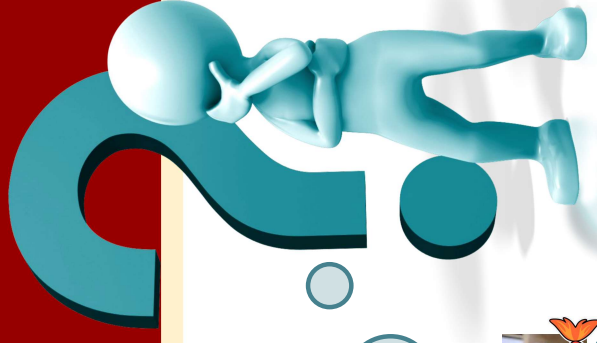
$\alpha = 0.05$ -- probability threshold for happening upon the result even if it really doesn't exist.

What is the probability we happen upon once in ten times?

$$1 - p(\text{not happening upon the result}) = 1 - (1 - .05)^{10} = 0.599$$

$$\text{How to fix? } 1 - (1 - (.05/10))^{10} = .488$$

H_1 : Most cats don't like redfish.



Bonferroni's Cats

General Question: Which fish do cats like?

$N = 50$ cats; 32 like redfish; $p = 0.016$

$\alpha = 0.05$ -- probability threshold for happening upon the result even if it really doesn't exist.

What is the Bonferroni correction:

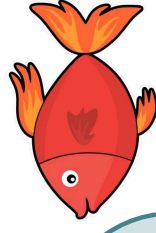
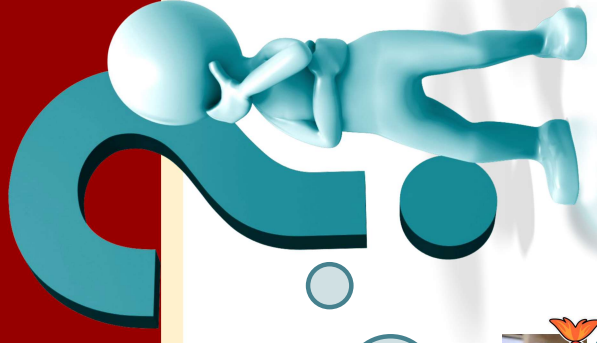
$$1 - p(n) = 1 - (1 - \alpha / |h|)^n$$

How often will I see a redfish once in ten times?

$$\begin{aligned} &= 1 - (1 - .05)^{10} \\ &= 1 - 0.599 \end{aligned}$$

$$\text{How to fix? } 1 - (1 - (.05/10))^{10} = .0488$$

H_1 : Most cats like redfish. *don't like redfish.*



Multi-test Correction

Type I, Type II Errors

	True state of nature	
	H_0	H_A
Our decision	Reject H_0	Type I error correct decision
	'Accept' H_0	Type II error correct decision

(Orloff & Bloom, 2014)

Multi-test Correction

significance level (“p-value”) = $P(\text{type I error}) = P(\text{Reject } H_0 \mid H_0)$
(probability we are incorrect)

	H_0	H_A
<u>Reject H_0</u>	$P(\text{Reject } H_0 \mid H_0)$	

	True state of nature	
	H_0	H_A
Our decision	Reject H_0	Type I error correct decision
	‘Accept’ H_0	correct decision Type II error

(Orloff & Bloom, 2014)

Multi-test Correction

significance level (“p-value”) = $P(\text{type I error}) = P(\text{Reject } H_0 \mid H_0)$
(probability we are incorrect)

power = $1 - P(\text{type II error}) = P(\text{Reject } H_0 \mid H_1)$
(probability we are correct)

	H_0	H_A
Reject H_0	$P(\text{Reject } H_0 \mid H_0)$	$P(\text{Reject } H_0 \mid H_A)$

	True state of nature	
	H_0	H_A
Our decision	Reject H_0	Type I error correct decision
	‘Accept’ H_0	correct decision Type II error

(Orloff & Bloom, 2014)

Multi-test Correction

FWER: Family-wise error rate (Bonferroni Corrects)

The probability of making ≥ 1 type 1 error.

$$FWER = Pr(\text{type1s} > 0) = 1 - Pr(\text{type1s} = 0) = 1 - (1 - \alpha)^m$$

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Accept' H_0	correct decision	Type II error

(Orloff & Bloom, 2014)

Multi-test Correction

FWER: Family-wise error rate (Bonferroni Corrects)

The probability of making ≥ 1 type 1 error.

$$FWER = Pr(\text{type1s} > 0) = 1 - Pr(\text{type1s} = 0) = 1 - (1 - \alpha)^m$$

$$1 - (1 - (.05/10))^{10} = .488$$

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Accept' H_0	correct decision	Type II error

(Orloff & Bloom, 2014)

Multi-test Correction

FWER: Family-wise error rate (Bonferroni corrects)

The probability of making ≥ 1 type 1 error.

$$FWER = Pr(\text{type1s} > 0) = 1 - Pr(\text{type1s} = 0) = 1 - (1 - \alpha)^m$$

FDR: False discovery rate (Benjamini-Hochberg corrects)
type1s / (type1s + correctRejects)

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Accept' H_0	correct decision	Type II error

(Orloff & Bloom, 2014)

Multi-test Correction

FWER: Family-wise error rate (Bonferroni corrects)

The probability of making ≥ 1 type 1 error.

$$FWER = Pr(\text{type1s} > 0) = 1 - Pr(\text{type1s} = 0) = 1 - (1 - \alpha)^m$$

FDR: False discovery rate (Benjamini-Hochberg corrects)
type1s / (type1s + correctRejects)

Proportion of false positives among *all* significant results.

The Hypothesis Test “Algorithm”

Input: H_0 , obs, α

obs_ts = test_stat(obs)

null_dist = distribution of expected test_stat under H_0

$p(x \leq \text{obs_ts} \mid H_0) = \text{cdf}(\text{null_dist}, \text{obs_ts})$

if $p(x \leq \text{obs_ts} \mid H_0) < \alpha$:

 decision = “Reject H_0 !”

else:

 decision = “Accept H_0 .”

Output: decision

The Multi-test "Algorithm"

Input: $H_{\theta}s$, obs, α

```
decisions = []
```

```
 $\alpha = \alpha / \text{len}(H_{\theta}s)$ 
```

```
for  $H_{\theta}$  in  $H_{\theta}s$ 
```

```
    obs_ts = test_stat(obs)
```

```
    null_dist = distribution of expected test_stat under  $H_{\theta}$ 
```

```
     $p(x \leq \text{obs\_ts} \mid H_{\theta}) = \text{cdf}(\text{null\_dist}, \text{obs\_ts})$ 
```

```
    if  $p(x \leq \text{obs\_ts} \mid H_{\theta}) < \alpha$ :
```

```
        decisions.append("Reject  $H_{\theta}$ !")
```

```
    else:
```

```
        decisions.append("Accept  $H_{\theta}$ .")
```

Output: decisions

Statistical Considerations for Big Data

1. Average multiple models (ensemble techniques)
2. Correct for multiple tests (Bonferonni's Principle)
3. Smooth data
4. "Plot" data (or figure out a way to look at a lot of it "raw")
5. Interact with data
6. Know your "real" sample size
7. Correlation is not causation
8. Define metrics for success (set a baseline)
9. Share code and data
10. The problem should drive solution

Comparing Variables

- **Linear Regression**
- Pearson Product-Moment Correlation
- Multiple Linear Regression
- (Multiple) Logistic Regression
- Ridge Regression (L2 Penalized)
- Lasso Regression (L1 Penalized)

Comparing Variables

Finding a linear function based on X to best yield Y .

X = “covariate” = “feature” = “predictor” = “regressor” = “independent variable”

Y = “response variable” = “outcome” = “dependent variable”

Regression: $r(x) = E(Y | X = x)$

goal: estimate function r

The expected value of Y , given that the random variable X is equal to some specific value, x .

Comparing Variables

Finding a linear function based on X to best yield Y .

X = “covariate” = “feature” = “predictor” = “regressor” = “independent variable”

Y = “response variable” = “outcome” = “dependent variable”

Regression: $r(x) = E(Y|X = x)$

goal: estimate the function r

Linear Regression (univariate version): $r(x) = \beta_0 + \beta_1 x$
goal: find β_0, β_1 such that $r(x) \approx E(Y|X = x)$

Comparing Variables

Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\mathbf{E}(\epsilon_i|X_i) = 0$ and $\mathbf{V}(\epsilon_i|X_i) = \sigma^2$

more precisely

$$r(x) = \beta_0 + \beta_1 x$$

Comparing Variables

Simple Linear Regression $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

intercept slope error

where $\mathbf{E}(\epsilon_i|X_i) = 0$ and $\mathbf{V}(\epsilon_i|X_i) = \sigma^2$

expected variance

Linear Regression

via Gradient Descent

Start with $\hat{\beta}_0 = \hat{\beta}_1 = 0$

Repeat until convergence:

Calculate all \hat{Y}_i

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha \left(\sum_{i=1}^n \hat{Y}_i - Y_i \right)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha \left(\sum_{i=1}^n X_i (\hat{Y}_i - Y_i) \right)$$

via Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the residual sum of squares:

$$RSS = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}((X - \bar{X})(Y - \bar{Y})) \end{aligned}$$

via Direct Estimates
(normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}((X - \bar{X})(Y - \bar{Y})) \end{aligned}$$

Correlation

$$\begin{aligned} r = r_{X,Y} &= \frac{\text{Cov}(X, Y)}{s_X s_Y} \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \end{aligned}$$

via Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}((X - \bar{X})(Y - \bar{Y})) \end{aligned}$$

Correlation

$$\begin{aligned} r = r_{X,Y} &= \frac{\text{Cov}(X, Y)}{s_X s_Y} \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \end{aligned}$$

via Direct Estimates
(normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

If one standardizes X and Y (i.e. subtract the mean and divide by the standard deviation) before running linear regression, then:
 $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = r$ --- i.e. $\hat{\beta}_1$ is the Pearson correlation!

Comparing Variables

- **Linear Regression**
- **Pearson Product-Moment Correlation**
- Multiple Linear Regression
- (Multiple) Logistic Regression
- Ridge Regression (L2 Penalized)
- Lasso Regression (L1 Penalized)

Comparing Variables

- Linear Regression
- Pearson Product-Moment Correlation
- **Multiple Linear Regression**
- (Multiple) Logistic Regression
- Ridge Regression (L2 Penalized)
- Lasso Regression (L1 Penalized)

Multiple Linear Regression

$$\text{Simple Linear Regression} \quad Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $E(\epsilon_i|X_i) = 0$ and $V(\epsilon_i|X_i) = \sigma^2$

— expected variance

Estimated intercept and slope

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \quad \hat{Y}_i = \hat{r}(X_i)$$

Residual: $\hat{\epsilon}_i = Y_i - \hat{Y}_i$

Multiple Linear Regression

Suppose we have multiple X that we'd like to fit to Y at once:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$$

If we include and $X_{oi} = 1$ for all i (i.e. adding the intercept to X), then we can

say:

$$Y_i = \sum_{j=0}^m \beta_j X_{ij} + \epsilon_i$$

Multiple Linear Regression

Suppose we have multiple X that we'd like to fit to Y at once:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$$

If we include and $X_{oi} = 1$ for all i , then we can say:

$$Y_i = \sum_{j=0}^m \beta_j X_{ij} + \epsilon_i$$

Or in vector notation across all i :

$$Y = X\beta + \epsilon$$

where β and ϵ are vectors and X is a matrix.

Multiple Linear Regression

Suppose we have multiple X that we'd like to fit to Y at once:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$$

If we include and $X_{oi} = 1$ for all i , then we can say:

$$Y_i = \sum_{j=0}^m \beta_j X_{ij} + \epsilon_i$$

Or in vector notation across all i :

$$Y = X\beta + \epsilon$$

where β and ϵ are vectors and X is a matrix.

Estimating β :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Multiple Linear Regression

Suppose we have multiple independent variables that we'd like to fit to our dependent variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$

If we include and $X_{oi} = 1$ for all i . Then we can say:

$$Y_i = \sum_{j=1}^m \beta_j X_{ij} + \epsilon_i$$

To test for significance of individual coefficient, j :

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\frac{s^2}{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}}}$$

Or in vector notation

$$\text{across all } i: Y = X\beta + \epsilon$$

Where β and ϵ are vectors and X is a matrix.

Estimating β :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Multiple Linear Regression

Suppose we have multiple independent variables that we'd like to fit to our dependent variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$

If we include and $X_{0i} = 1$ for all i . Then we can say:

To test for significance of individual coefficient, j :

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\frac{s^2}{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}}}$$

Or in vector notation

$$\text{across all } i: Y = X\beta + \epsilon$$

Where β and ϵ are vectors and X is a matrix.

Estimating β :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$$

RSS

$$s^2 = \frac{\text{RSS}}{df}$$

df

To test for significance of individual coefficient, j :

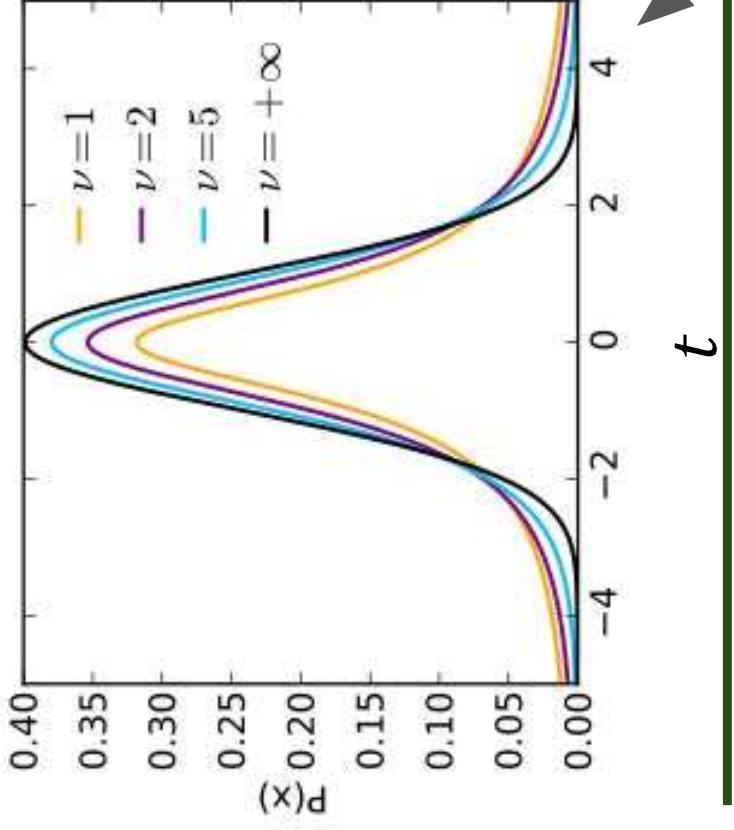
$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2 s^2}}$$

T-Test for significance of hypothesis:

- 1) Calculate t
- 2) Calculate degrees of freedom:

$$df = N - (m+1)$$

- 3) Check probability in a t distribution:



To test for significance of individual coefficient, j :

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2 s^2}}$$

$$\beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \epsilon_i$$

T-Test for significance of hypothesis:

- 1) Calculate t
- 2) Calculate degrees of freedom:

$$df = N - (m+1)$$

- 3) Check probability in a t distribution: ($df = \nu$)

Large-Scale Hypothesis Testing

- Findings and Uncertainty
- Hypothesis Testing
- Bonferroni's Cats
- Multi-test Corrections
 - Family-wise Error Rate
 - False-Discovery Rate
- Correlation Metrics
 - Effect Size
 - Significance