Linear Models: Comparing Variables

Stony Brook University
CSE545, Fall 2017

Statistical Preliminaries

Random Variables

Random Variables

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

Random Variables

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

Example: $\boldsymbol{\Omega}$ = 5 coin tosses = $\{\langle HHHHH\rangle, \langle HHHHT\rangle, \langle HHHTH\rangle, \langle HHHTH\rangle \ldots\}$

We may just care about how many tails? Thus,

$X(\langle HHHHH\rangle) = 0$

$X(\langle HHHTH\rangle) = 1$

$X(\langle TTTHT\rangle) = 4$

$X(\langle HTTTT\rangle) = 4$

X only has 6 possible values: 0, 1, 2, 3, 4, 5

What is the probability that we end up with $k = 4$ tails?

$$\mathbf{P}(X = k) := \mathbf{P}(\ \{\omega : X(\omega) = k\}\ ) \qquad \text{where } \omega \in \boldsymbol{\Omega}$$

Random Variables

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = 5 coin tosses = {<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>…}

We may just care about how many tails? Thus,

X(<HHHHH>) = 0

X(<HHHTH>) = 1

X(<TTTHT>) = 4

X(<HTTTT>) = 4

X only has 6 possible values: 0, 1, 2, 3, 4, 5

What is the probability that we end up with k = 4 tails?

$\mathbf{P}(X = k) := \mathbf{P}(\ \{\omega : X(\omega) = k\}\ )$ where $\omega \in \mathbf{\Omega}$

X(ω) = 4 for 5 out of 32 sets in **Ω**. Thus, assuming a fair coin, $\mathbf{P}(X = 4) = 5/32$

(Not a "variable", but a function that we end up notating a lot like a variable)

Random Variables

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = 5 coin tosses = {**<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>**…}

We may just care about how many tails? Thus,

X(<HHHHH>) = 0

X(<HHHTH>) = 1

X(<TTTHT>) = 4

X(<HTTTT>) = 4

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

X only has 6 possible values: 0, 1, 2, 3, 4, 5

What is the probability that we end up with $k = 4$ tails?

$\mathbf{P}(X = k) := \mathbf{P}(\ \{\omega : X(\omega) = k\}\ )$ where $\omega \in \mathbf{\Omega}$

$X(\omega) = 4$ for 5 out of 32 sets in **Ω**. Thus, assuming a fair coin, $\mathbf{P}(X = 4) = 5/32$

(Not a "variable", but a function that we end up notating a lot like a variable)

Random Variables

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

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

$\mathbf{X}$ is a *discrete random variable* if it takes only a countable number of values.

Random Variables

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

Example: **Ω** = inches of snowfall = $[0, \infty) \subseteq \mathbb{R}$

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

X amount of inches in a snowstorm

$\mathbf{X}(\omega) = \omega$

What is the probability we receive (at least) a *inches?*

$\mathbf{P}(X \geq a) := \mathbf{P}(\ \{\omega : X(\omega) \geq a\}\ )$

What is the probability we receive between a *and* b *inches?*

$\mathbf{P}(a \leq X \leq b) := \mathbf{P}(\ \{\omega : a \leq X(\omega) \leq b\}\ )$

Random Variables

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

Example: $\mathbf{\Omega}$ = inches of snowfall = $[0, \infty) \subseteq \mathbb{R}$

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

X amount of inches in a snowstorm

$\mathbf{X}(\omega) = \omega$

$\mathbf{P}(X = i) := 0$, for all $i \in \mathbf{\Omega}$

(probability of receiving exactly i inches of snowfall is zero)

What is the probability we receive (at least) a *inches?*

$\mathbf{P}(X \geq a) := \mathbf{P}(\ \{\omega : X(\omega) \geq a\}\ )$

What is the probability we receive between a *and* b *inches?*

$\mathbf{P}(a \leq X \leq b) := \mathbf{P}(\ \{\omega : a \leq X(\omega) \leq b\}\ )$

Random Variables

X: A mapping from **Ω** to ℝ that describes the question we care about in practice.

Example: **Ω** = inches of snowfall = $[0, \infty) \subseteq \mathbb{R}$

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

X amount of inches in a snowstorm

$\mathbf{X}(\omega) = \omega$

$\mathbf{P}(X = i) := 0$, for all $i \in \mathbf{\Omega}$

(probability of receiving <u>exactly</u> i inches of snowfall is zero)

How to model?

s?

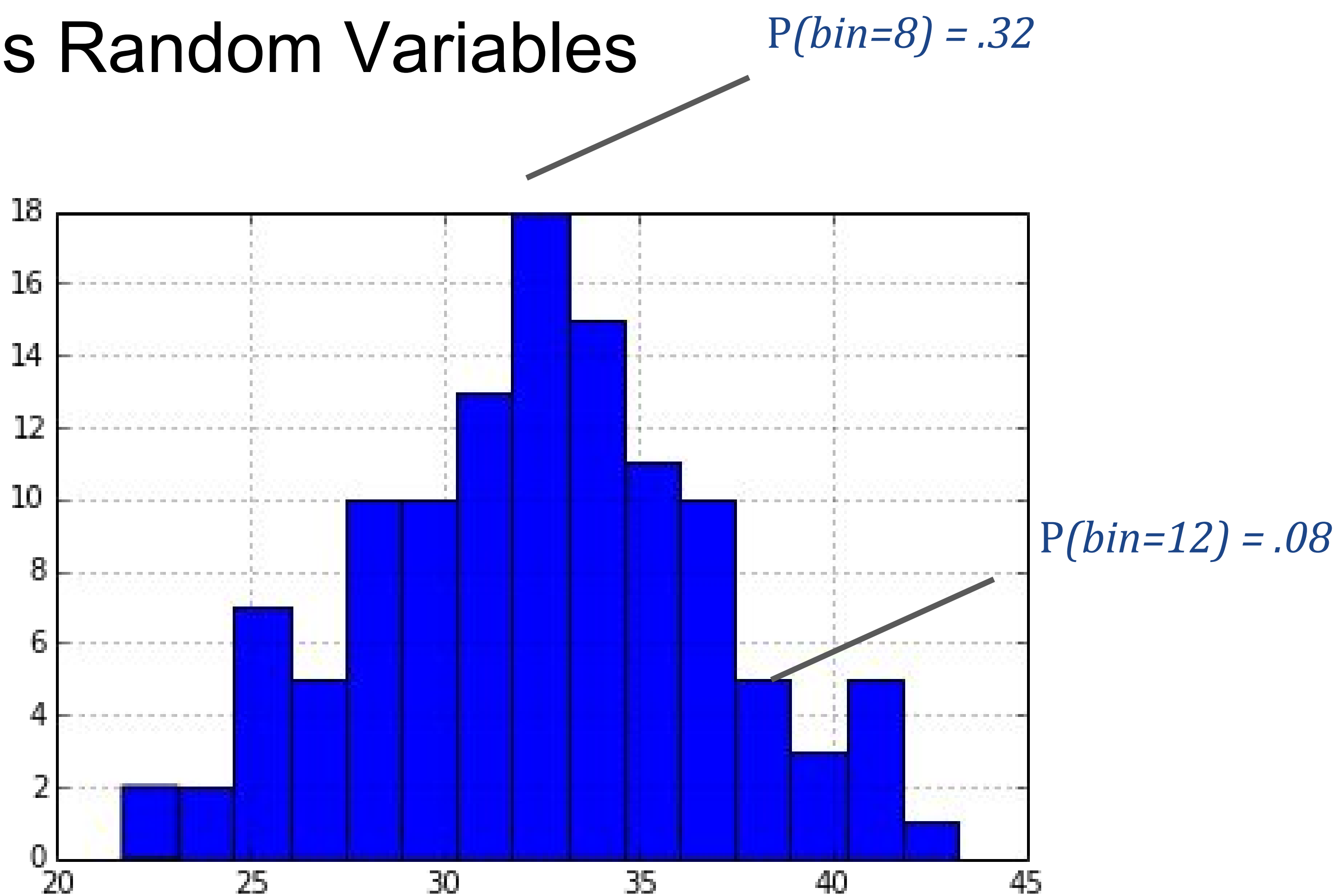
inches?

Continuous Random Variables

Discretize them!
(group into discrete bins)

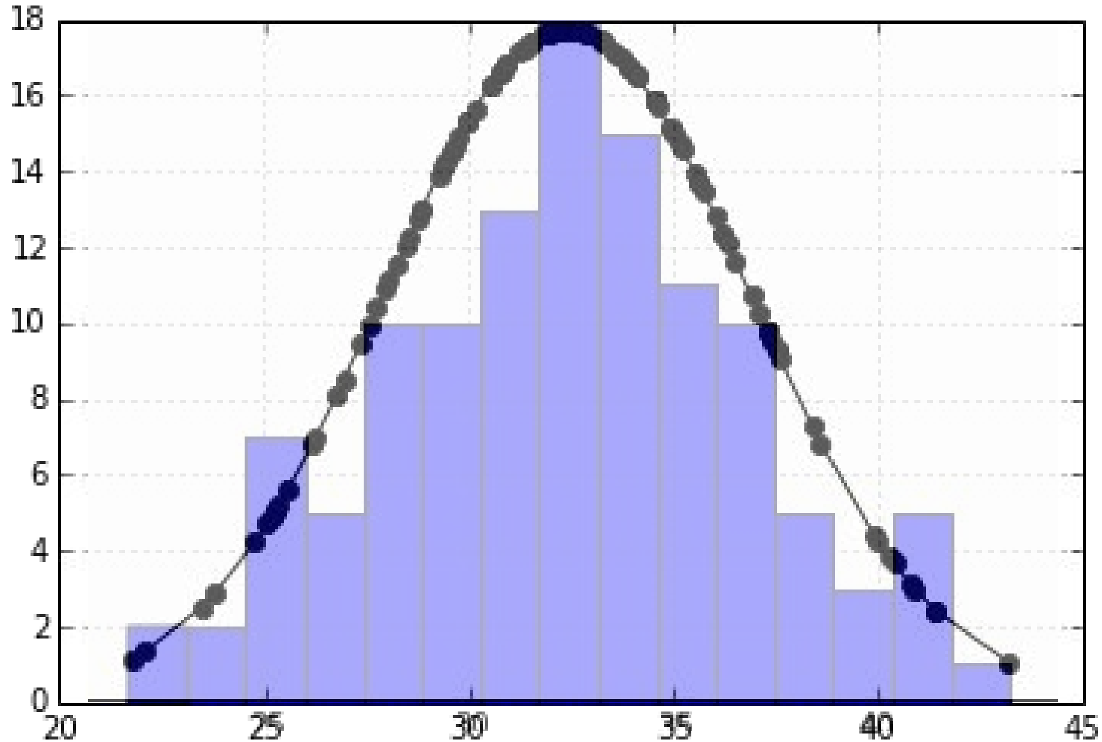
How to model?

Continuous Random Variables



But aren't we throwing away information?

Continuous Random Variables



Continuous Random Variables

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 *fx* such that:

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

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

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

Continuous Random Variables

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 *fx* such that:

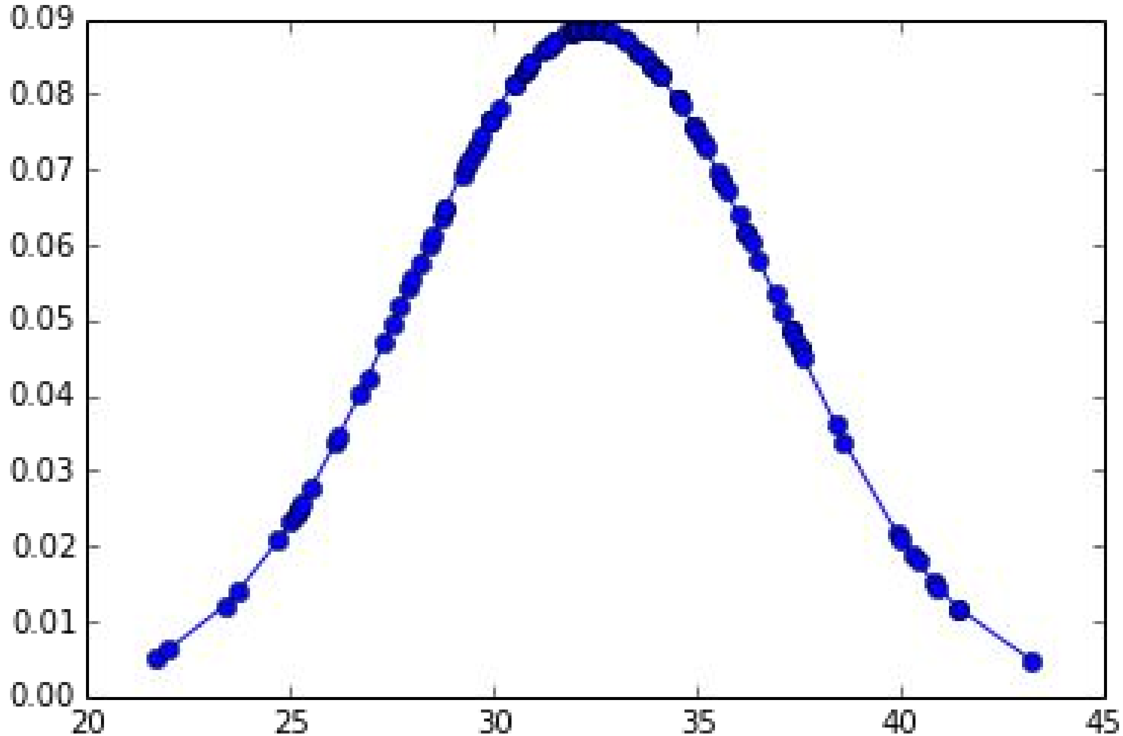
$$f_X(x) \geq 0, \text{for all } x \in X,$$

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

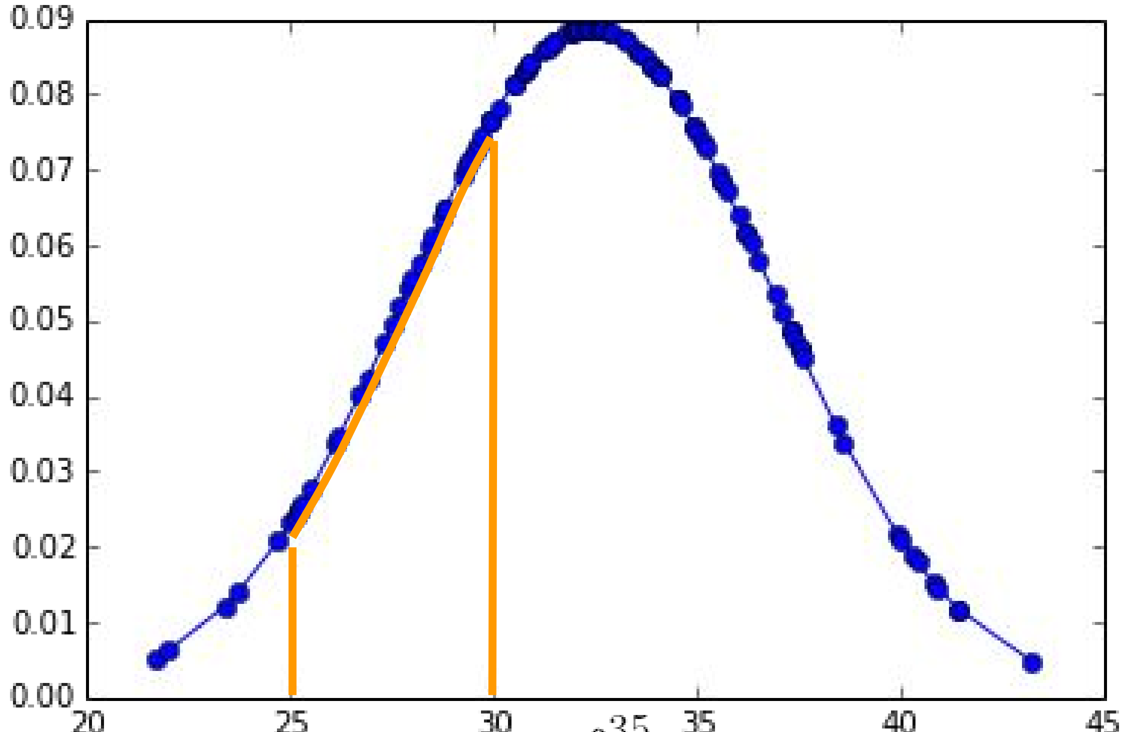
$$\mathrm{P}(a < X < b) = \int_a^b f_X(x)dx$$

***fx* : “probability density function” (pdf)**

Continuous Random Variables



Continuous Random Variables

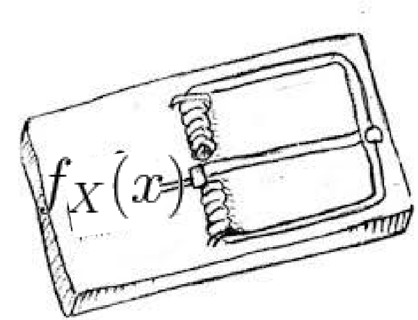


$$P(25 < X < 35) = \int_{25}^{35} fx(x)dx$$

Continuous Random Variables

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



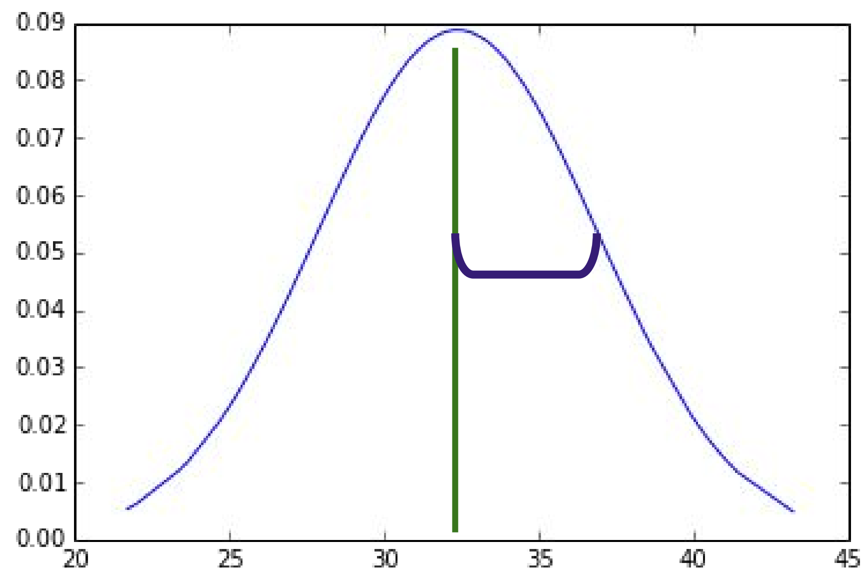
Continuous Random Variables

A Common Probability Density Function

Continuous Random Variables

Common *pdfs*: Normal(μ, σ^2)

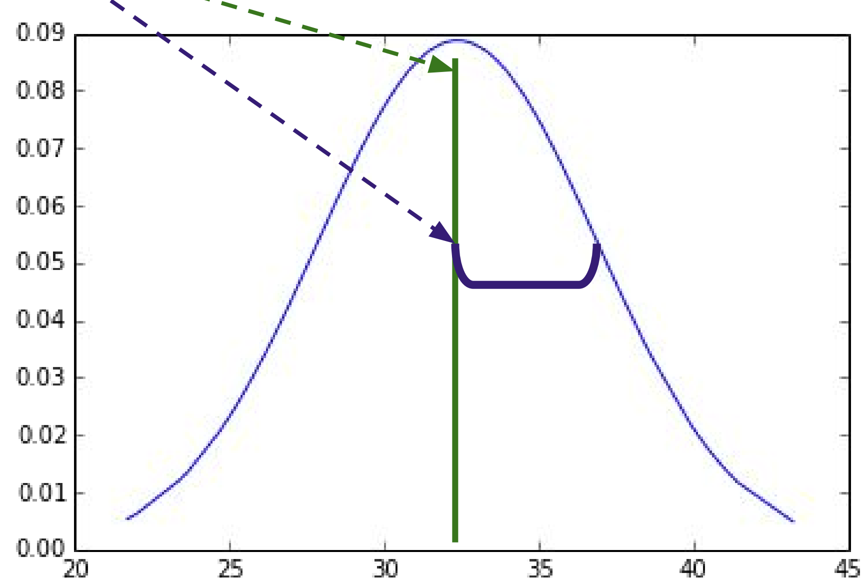
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Continuous Random Variables

Common *pdfs*: Normal(μ, σ^2)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



μ: mean (or "center")

= expectation

σ^2: variance,

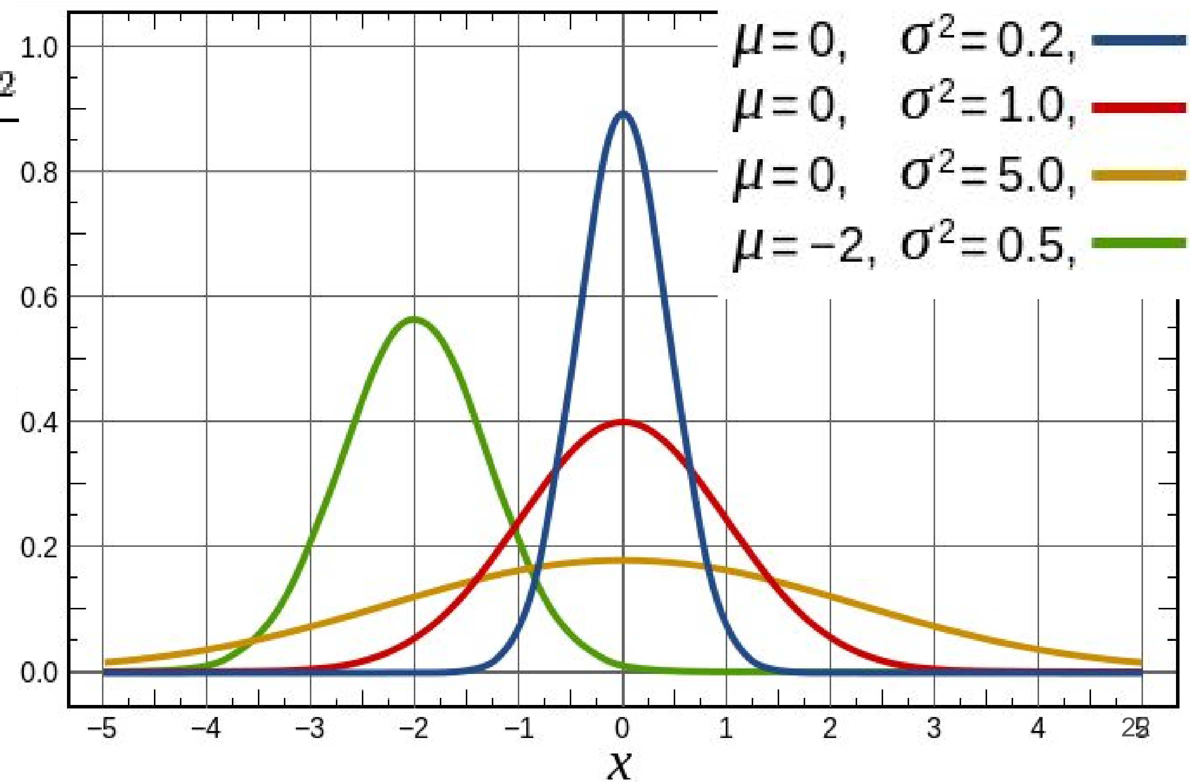
σ: standard deviation

Continuous Random Variables

Common *pdf*s: Normal(μ, σ^2)

Credit: Wikipedia

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



μ: mean (or "center")

= expectation

σ^2: variance,

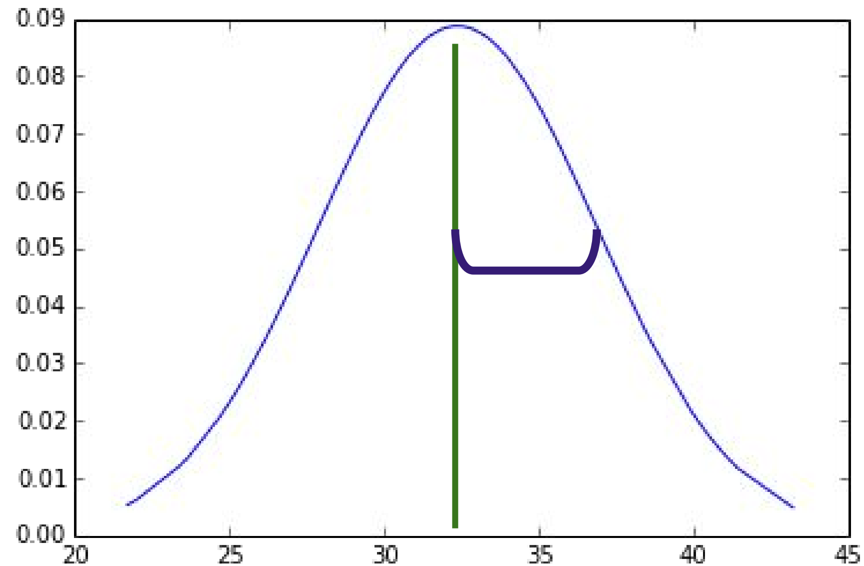
σ: standard deviation

Continuous Random Variables

Common *pdfs*: Normal(μ, σ^2)

$X \sim \text{Normal}(\mu, \sigma^2)$, examples:

- height
- intelligence/ability
- **measurement error**
- averages (or sum) of lots of random variables



Continuous Random Variables

Common *pdf*s: Normal(0, 1) ("standard normal")

How to "standardize" any normal distribution:

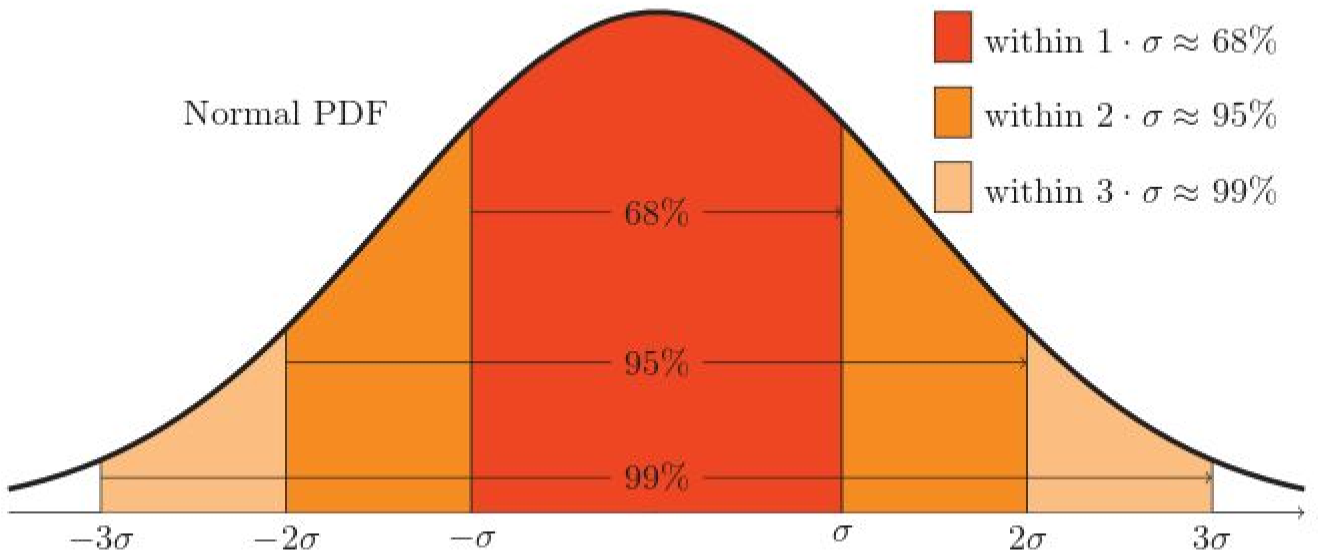
- subtract the mean, μ (aka "mean centering")
- divide by the standard deviation, σ

$z = (x - \mu) / \sigma$, (aka "z score")

Continuous Random Variables

Common *pdfs*: Normal(0, 1)

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



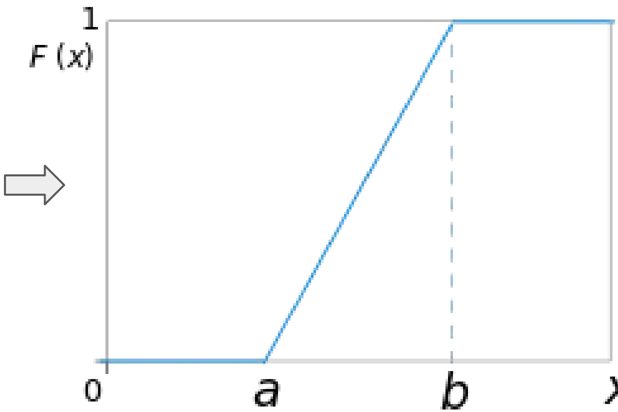
Credit: MIT Open Courseware: Probability and Statistics

Cumulative Distribution Function

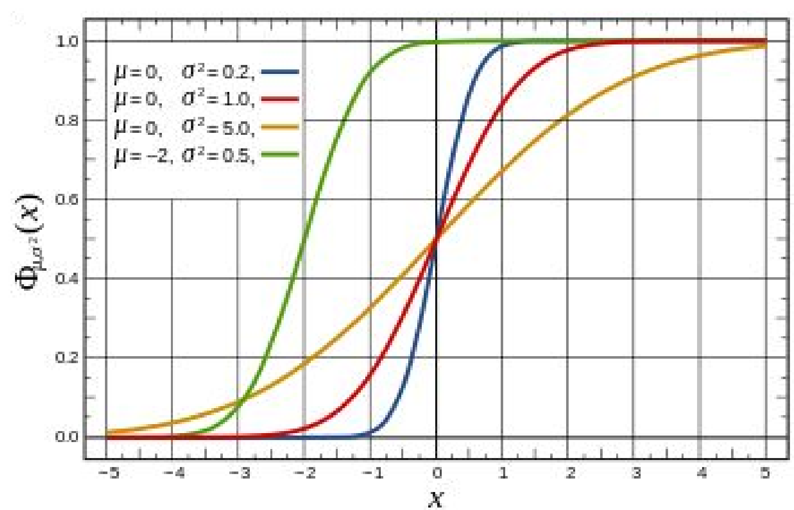
For a given random variable X, the *cumulative distribution function* (CDF), $F_X: \mathbb{R} \to [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \leq x)$$

Uniform



Normal

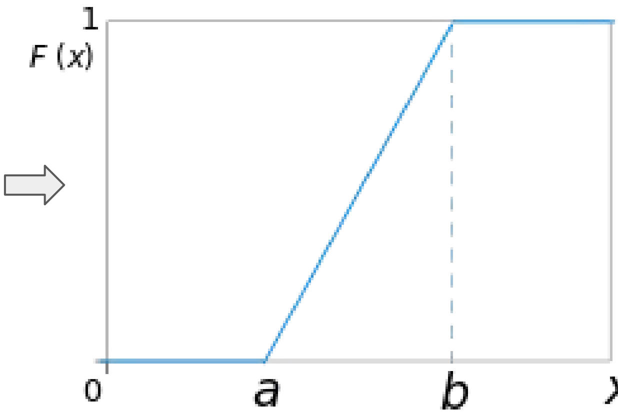


Cumulative Distribution Function

For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

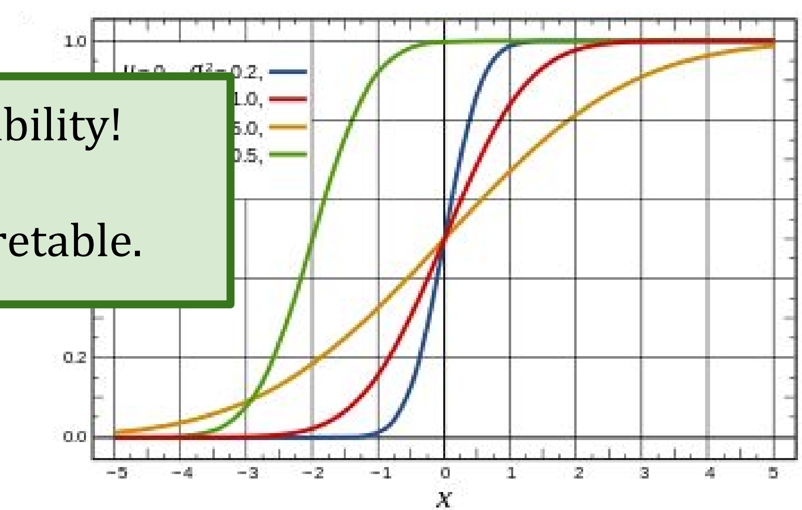
$$F_X(x) = \mathrm{P}(X \leq x)$$

Uniform



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

Con: Not intuitively interpretable.



Random Variables, Revisited

X: A mapping from **Ω** to 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.

Discrete Random Variables

For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \to [0, 1]$, is defined by:

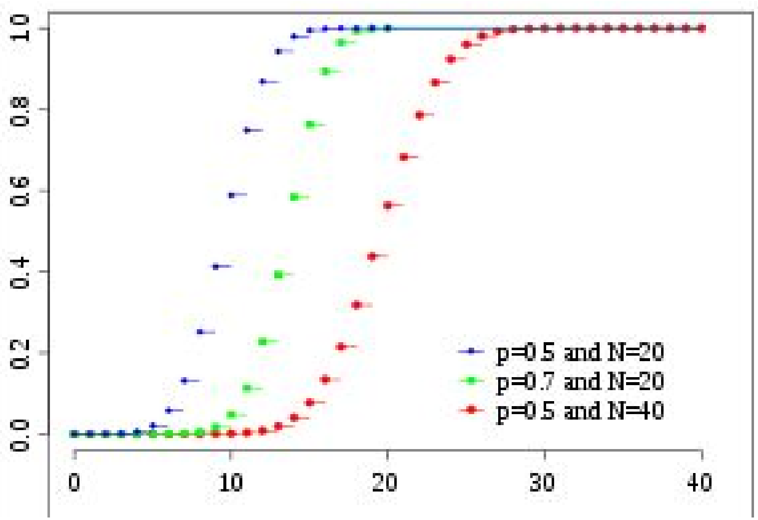
$$F_X(x) = \mathrm{P}(X \leq x)$$

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

Discrete Random Variables

For a given random variable X, the *cumulative distribution function* (CDF), *Fx*: $\mathbb{R} \rightarrow [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \leq x)$$

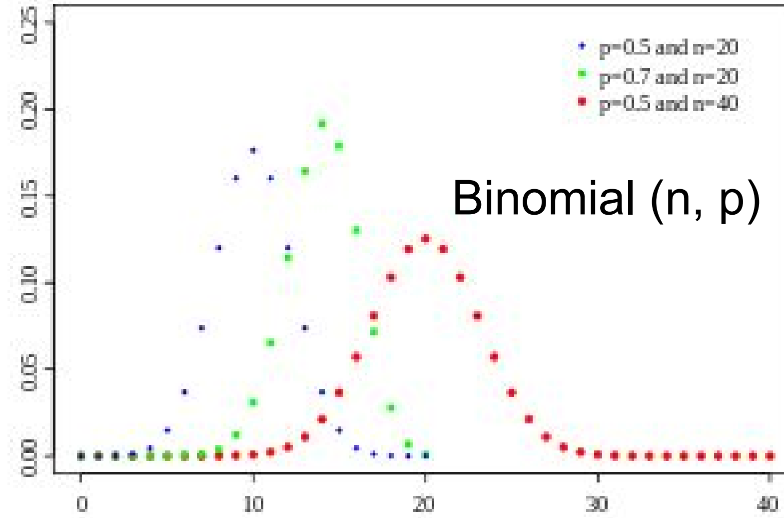


Binomial (n, p)

(like normal)

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

Discrete Random Variables



For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \leq x)$$

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

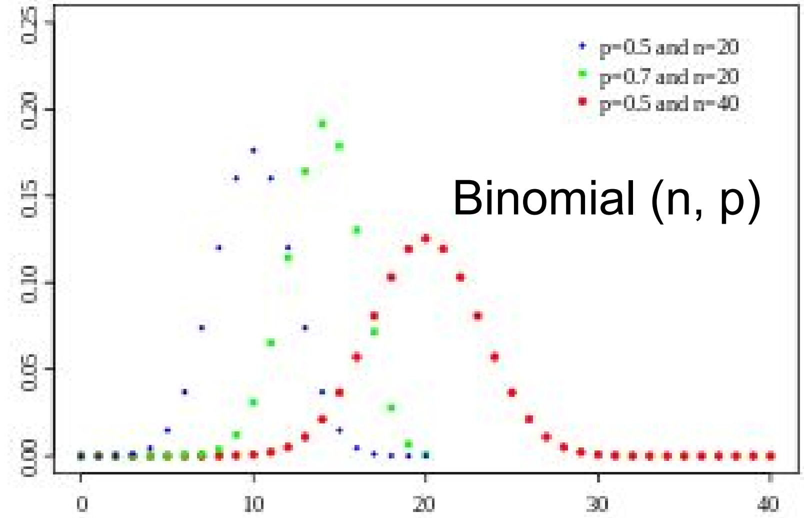
$$f_X(x) = \mathrm{P}(X = x)$$

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

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

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

Discrete Random Variables



Two Common **Discrete** Random Variables

- Binomial(n, p)

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x},\ \text{if } 0 \leq x \leq n \ (0 \text{ otherwise})$$

example: number of heads after n coin flips (p, probability of heads)

- Bernoulli(p) = Binomial(1, p)

example: one trial of success or failure

Hypothesis Testing

Hypothesis -- something one asserts to be true.

Classical Approach:

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

Hypothesis -- something one asserts to be true.

Classical Approach:

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

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

Goal: Use probability to determine if we can:

"reject the null" (H_0) in favor of H_1.

"There is less than a 5% chance that the null is true"
(i.e. 95% chance that alternative is true).

Hypothesis Testing

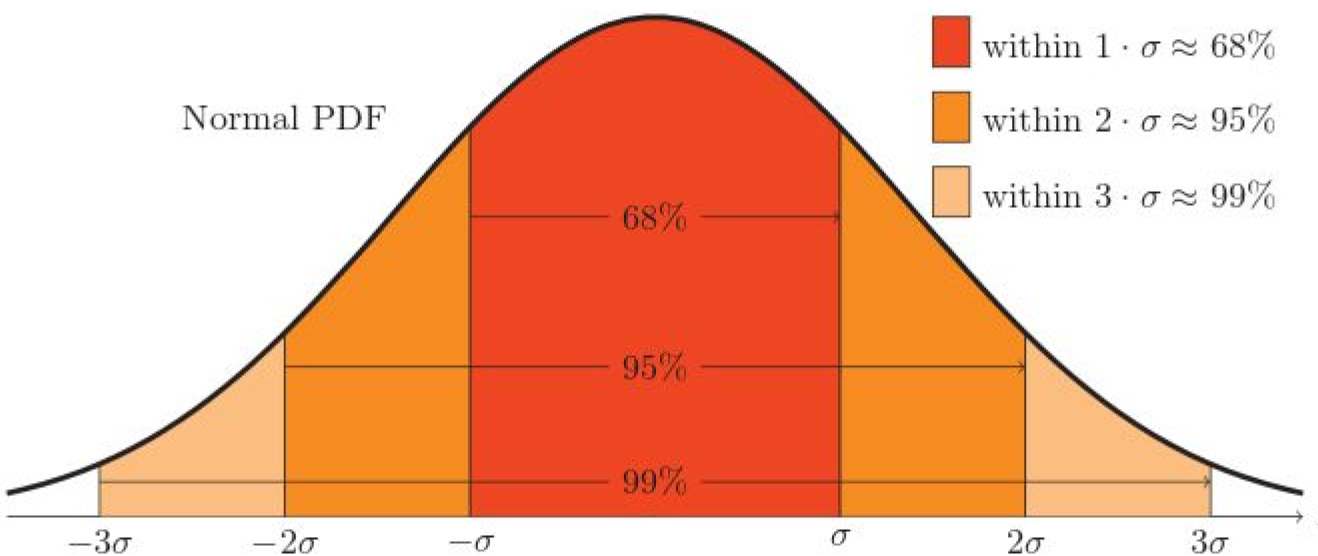
Example: Hypothesize a coin is biased.

H_0: the coin is not biased
(i.e. flipping n times results in a Binomial(n, 0.5))

H_1: the coin is biased (i.e. flipping n times results in a Binomial(n, 0.5))

Hypothesis Testing

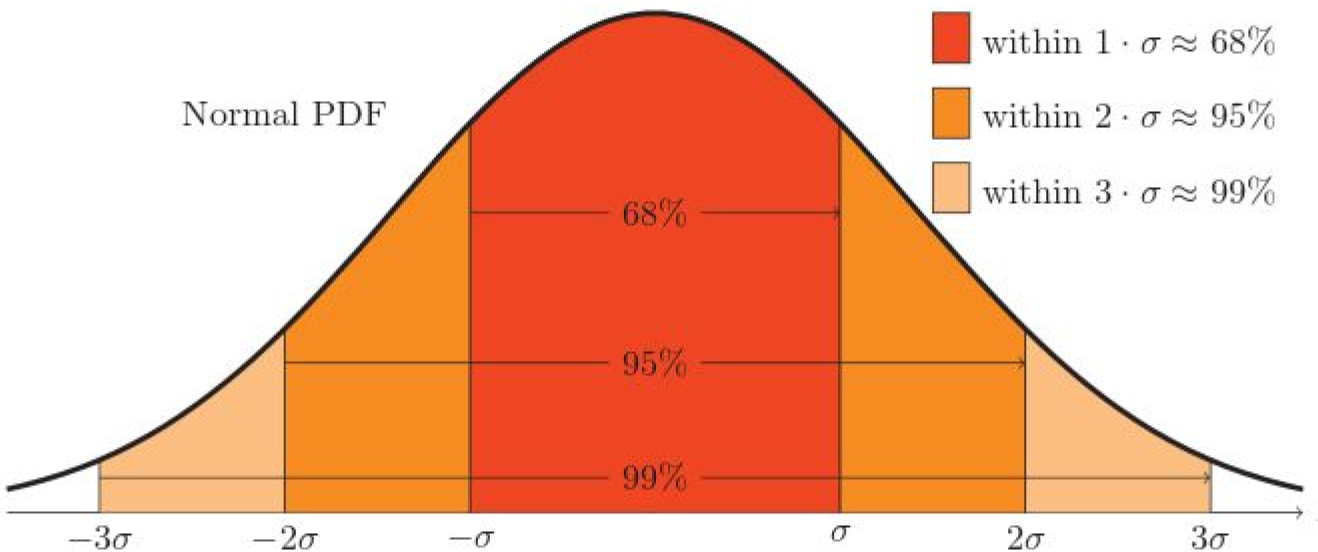
More formally: Let X be a random variable and let R be the range of X. $R_{reject} \subset R$ is the *rejection region.* If $X \in R_{reject}$ then we reject the null.



Hypothesis Testing

More formally: Let X be a random variable and let R be the range of X. $R_{reject} \subset R$ is the *rejection region.* If $X \in R_{reject}$ then we reject the null.

alpha : size of rejection region (e.g. 0.05, 0.01, .001)



Hypothesis Testing

More formally: Let X be a random variable and let R be the range of X. $R_{reject} \subset R$ is the *rejection region.* If $X \in R_{reject}$ then we reject the null.

alpha : size of rejection region (e.g. 0.05, 0.01, .001)

In the biased coin example,

if n = 1000, then then $R_{reject} = [0, 469] \cup [531, 1000]$

Hypothesis Testing

Important logical question:

Does failure to reject the null mean the null is true?

Hypothesis Testing

Important logical question:

Does failure to reject the null mean the null is true?

Thought experiment: If we have infinite data, can the null ever be true?

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	correct decision	Type II error

(Orloff & Bloom, 2014)

Power

significance level ("p-value") = P(type I error) = **P(Reject H_0 | H_0)**
(probability we are incorrect)

power = 1 - P(type II error) = **P(Reject H_0 | H_1)**
(probability we are correct)

	H_0	H_A
Reject H_0	**P(Reject H_0 | H_0)**	**P(Reject H_0 | H_1)**

		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

If alpha = .05, and I run 40 variables through significance tests, then, by chance, how many are likely to be significant?

Multi-test Correction

How to fix?

Multi-test Correction

How to fix?

What if all tests are independent?
=> "Bonferroni Correction" (α/m)

Better Alternative: False Discovery Rate
(Bejamini Hochberg)

Statistical Considerations in 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

(http://simplystatistics.org/2014/05/22/10-things-statistics-taught-us-about-big-data-analysis/)

Measures for Comparing Random Variables

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

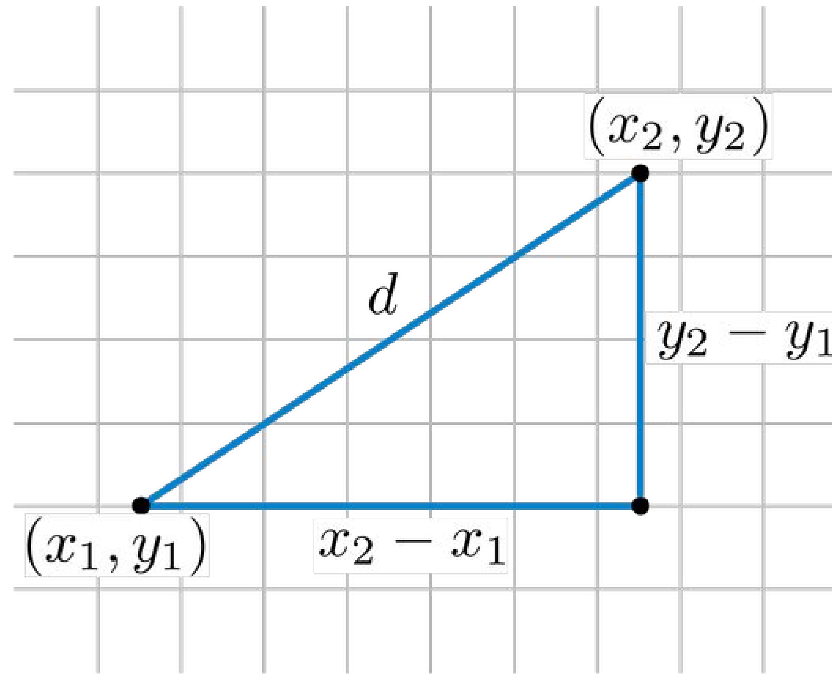
Distance Metrics

Typical properties of a distance metric, *d*:

$d(x, x) = 0$

$d(x, y) = d(y, x)$

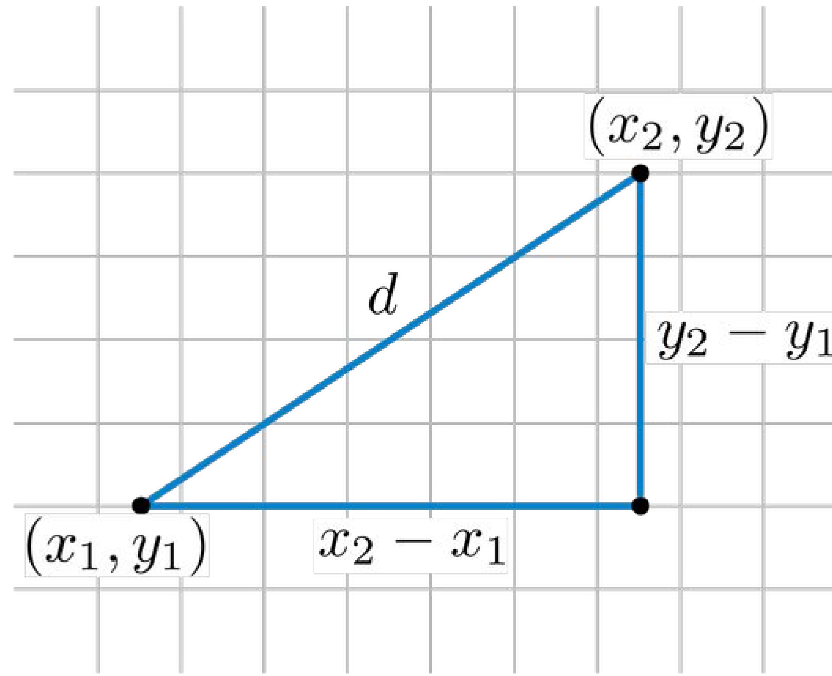
$d(x, y) \leq d(x,z) + d(z,y)$



(http://rosalind.info/glossary/euclidean-distance/)

Distance Metrics

- Jaccard Distance (1 - JS)
- Euclidean Distance
- Cosine Distance
- Edit Distance
- Hamming Distance



(http://rosalind.info/glossary/euclidean-distance/)

Distance Metrics

- Jaccard Distance (1 - JS)
- Euclidean Distance $distance(X,Y) = \sqrt{\sum_{i}^{n}(x_i - y_i)^2}$ ("L2 Norm")
- Cosine Distance
- Edit Distance
- Hamming Distance

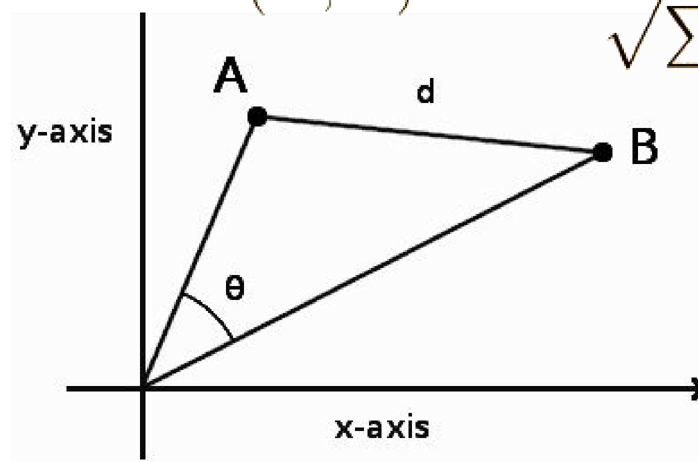
(http://rosalind.info/glossary/euclidean-distance/)

Distance Metrics

- Jaccard Distance (1 - JS)
- Euclidean Distance $distance(X,Y) = \sqrt{\sum_{i}^{n}(x_i - y_i)^2}$ ("L2 Norm")
- Cosine Distance

$$distance(X,Y) = 1 - \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$$

- Edit Distance
- Hamming Distance



(http://rosalind.info/glossary/euclidean-distance/)

Measures for Comparing Random Variables

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

Linear Regression

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: $r(x) = \mathrm{E}(Y|X = x)$

goal: estimate function r

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

Linear Regression

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: $r(x) = \mathrm{E}(Y|X = x)$

goal: estimate the function *r*

Linear Regression (univariate version): $r(x) = \beta_0 + \beta_1 x$

goal: find β_0, β_1 such that $r(x) \approx \mathrm{E}(Y|X = x)$

Linear Regression

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

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

more precisely

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

Linear Regression

intercept slope error

Simple Linear Regression

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

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

expected variance

Linear Regression

intercept slope error

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

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

expected variance

Estimated intercept and slope

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

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

Linear Regression

intercept slope error

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

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

expected variance

Estimated intercept and slope

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

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

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

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

Linear Regression

via Gradient Descent

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

Repeat until convergence:

Calculate all $\hat{Y}_i$

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

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

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

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

Linear Regression

via Gradient Descent

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

Repeat until convergence:

Calculate all $\hat{Y}_i$

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

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

Learning rate

Based on derivative of *RSS*

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

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

Linear Regression

via Gradient Descent

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

Repeat until convergence:

Calculate all $\hat{Y}_i$

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

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

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

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

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

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

Correlation

$$r = r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)$$

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

$$Cov(X,Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$$
$$= \mathbf{E}\left((X-\bar{X})(Y-\bar{Y})\right)$$

Correlation

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

via Direct Estimates (normal equations)

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

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

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!*

Measures for Comparing Random Variables

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

Multiple Linear Regression

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

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:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

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:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

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: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$

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

$$Y_i = \sum^{m} \beta_j X_{ij} + \epsilon_i$$

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 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: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$

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

Or in vector notation across all i: $Y = X\beta + \epsilon$

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

Estimating β :

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

To test for significance of individual coefficient, *j*:

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

Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

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

To test for significance of individual coefficient, *j*:

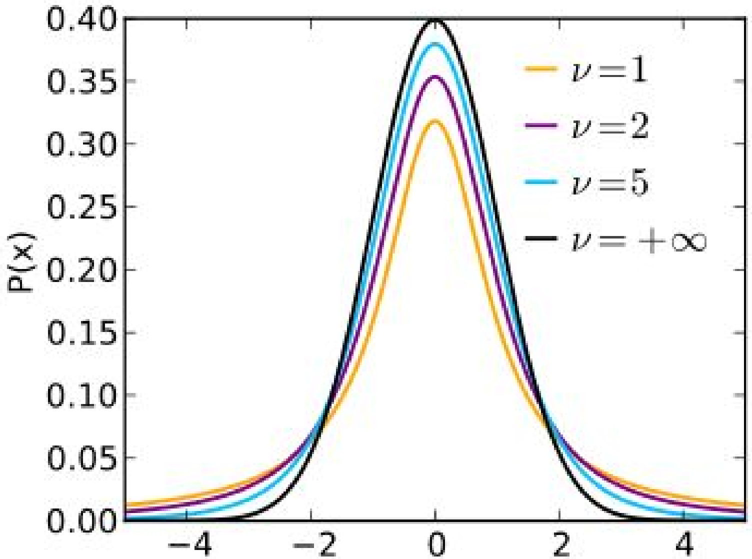
$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}}}$$

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:



$\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$

$$\frac{t}{}$$

To test for significance of individual coefficient, *j*:

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

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: $(df = v)$

Hypothesis Testing

Important logical question:

Does failure to reject the null mean the null is true?

Thought experiment: If we have infinite data, can the null ever be true?

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	correct decision	Type II error

(Orloff & Bloom, 2014)

Power

significance level ("p-value") = P(type I error) = **P(Reject H_0 | H_0)**
(probability we are incorrect)

power = 1 - P(type II error) = **P(Reject H_0 | H_1)**
(probability we are correct)

	H_0	H_A
Reject H_0	**P(Reject H_0 | H_0)**	**P(Reject H_0 | H_1)**

		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

If alpha = .05, and I run 40 variables through significance tests, then, by chance, how many are likely to be significant?

Multi-test Correction

How to fix?

Multi-test Correction

How to fix?

What if all tests are independent?
=> “Bonferroni Correction” (α/m)

Better Alternative: False Discovery Rate
(Bejamini Hochberg)

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$

Note: this is a probability here.

In simple linear regression we wanted an expectation:

$$r(x) = \mathrm{E}(Y | X = x)$$

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \underbrace{\mathbf{P}(Y_i = 1 | X = x)} = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$

Note: this is a probability here.

In simple linear regression we wanted an expectation:

$$r(x) = \mathrm{E}(Y | X = x)$$

(i.e. if $p > 0.5$ we can confidently predict $Y_i = 1$)

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$

$$logit(p_i) = log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}$$

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$

$$logit(p_i) = log\left(\frac{p_i}{\boxed{1 - p_i}}\right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}$$

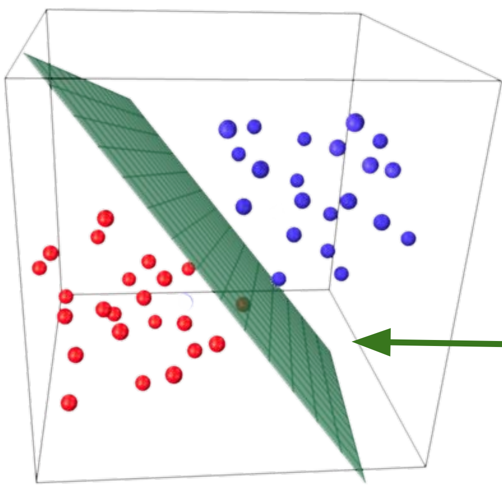
$P(Y_i = 0 \mid X = x)$

Thus, 0 is class 0 and 1 is class 1.

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$



$$logit(p_i) = log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^{m} \boxed{\beta_j x_{ij}}$$

We're still learning a linear *separating hyperplane*, but fitting it to a *logit* outcome.

(https://www.linkedin.com/pulse/predicting-outcomes-pr obabilities-logistic-regression-konstantinidis/)

Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$

$$logit(p_i) = log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}$$

To estimate β, one can use *reweighted least squares:*

(Wasserman, 2005; Li, 2010)

set $\hat{\beta}_0 = ... = \hat{\beta}_m = 0$ (remember to include an intercept)

1. Calculate p_i and let W be a diagonal matrix where element$(i, i) = p_i(1 - p_i)$.
2. Set $z_i = logit(p_i) + \frac{Y_i - p_i}{p_i(1 - p_i)} = X\hat{\beta} + \frac{Y_i - p_i}{p_i(1 - p_i)}$
3. Set $\hat{\beta} = (X^T W X)^{-1} X^T W z$ //weighted lin. reg. of Z on Y.
4. Repeat from 1 until $\hat{\beta}$ converges.

Uses of linear and logistic regression

1. Testing the relationship between variables given other variables. β is an "effect size" -- a score for the magnitude of the relationship; can be tested for significance.

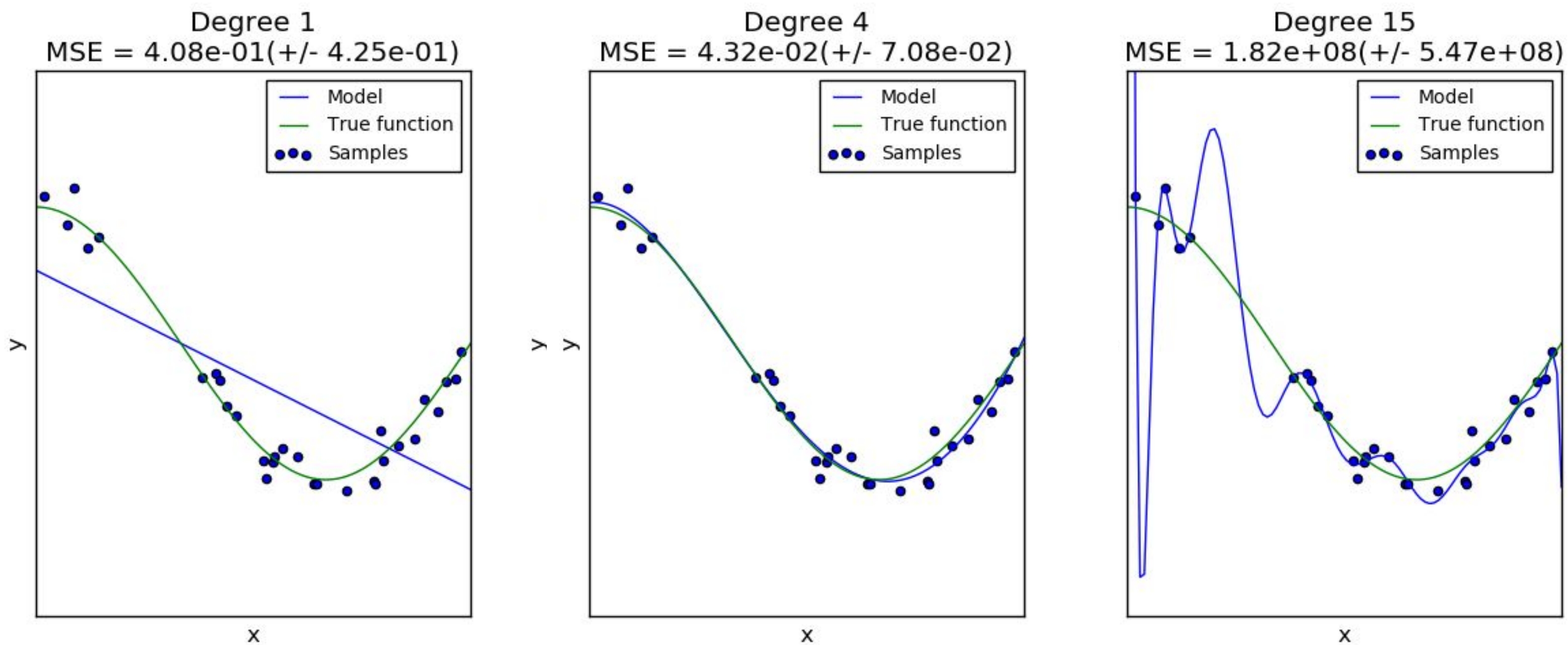
2. Building a predictive model that generalizes to new data. $\hat{Y}$ is an estimate value of Y given X.

Uses of linear and logistic regression

1. Testing the relationship between variables given other variables. β is an "effect size" -- a score for the magnitude of the relationship; can be tested for significance.

2. Building a predictive model that generalizes to new data. $\hat{Y}$ is an estimate value of Y given X.
   However, unless $|X| <<< observatations$ then the model might "overfit".

Overfitting (1-d non-linear example)



Underfit

High Bias

Overfit

High Variance

(image credit: Scikit-learn; in practice data are rarely this clear)

Overfitting (5-d linear example)

$Y \quad = \qquad\qquad X$

1
1
0
0
1

0.5	0	0.6	1	0	0.25
0	0.5	0.3	0	0	0
0	0	1	1	1	0.5
0	0	0	0	1	1
0.25	1	1.25	1	0.1	2

Overfitting (5-d linear example)

$Y \quad = \qquad\qquad X$

1		0.5	0	0.6	1	0	0.25
1		0	0.5	0.3	0	0	0
0		0	0	1	1	1	0.5
0		0	0	0	0	1	1
1		0.25	1	1.25	1	0.1	2

$$logit(Y) = 1.2 + -63*X_1 + 179*X_2 + 71*X_3 + 18*X_4 + -59*X_5 + 19*X_6$$

Overfitting (5-d linear example)

Do we really think we found something generalizable?

$Y \quad = \qquad X$

1		0.5	0	0.6	1	0	0.25
1		0	0.5	0.3	0	0	0
0		0	0	1	1	1	0.5
0		0	0	0	0	1	1
1		0.25	1	1.25	1	0.1	2

$$logit(Y) = 1.2 + -63*X_1 + 179*X_2 + 71*X_3 + 18*X_4 + -59*X_5 + 19*X_6$$

Overfitting (2-d linear example)

Do we really think we found something generalizable?

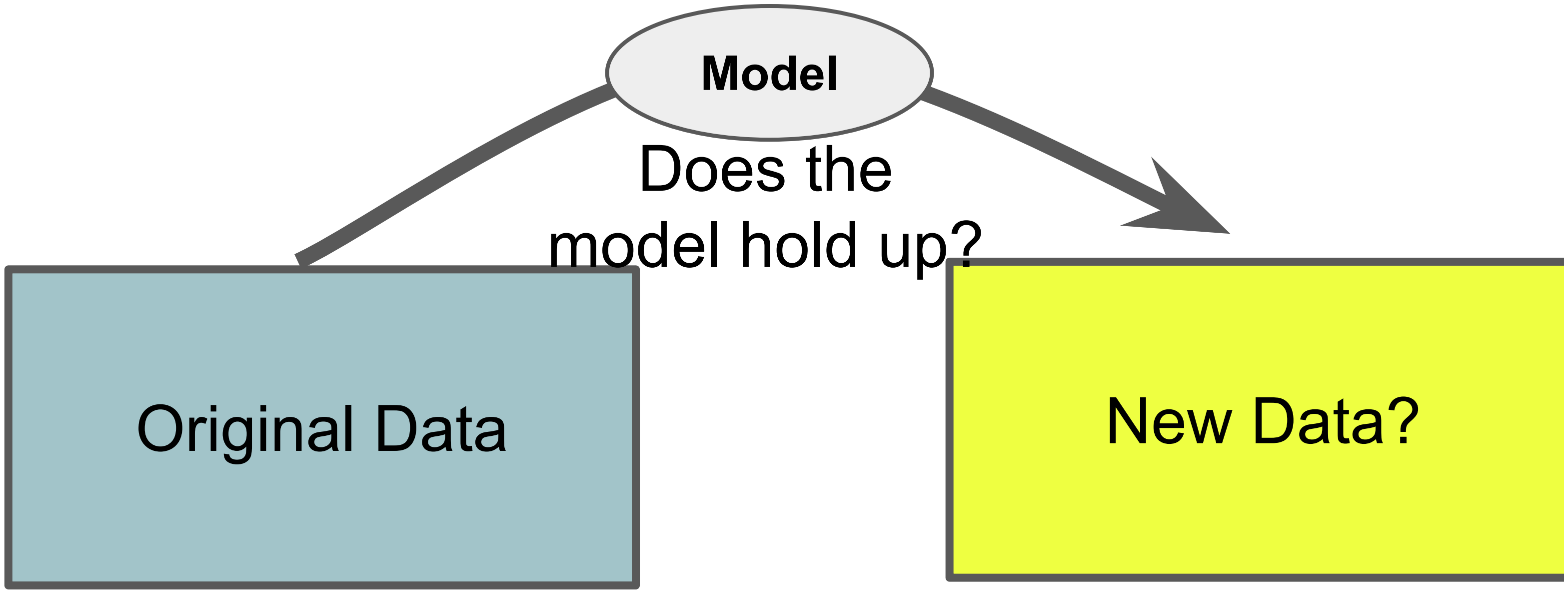
$Y \quad = \qquad X$

Y	X_1	X_2
1	0.5	0
1	0	0.5
0	0	0
0	0	0
1	0.25	1

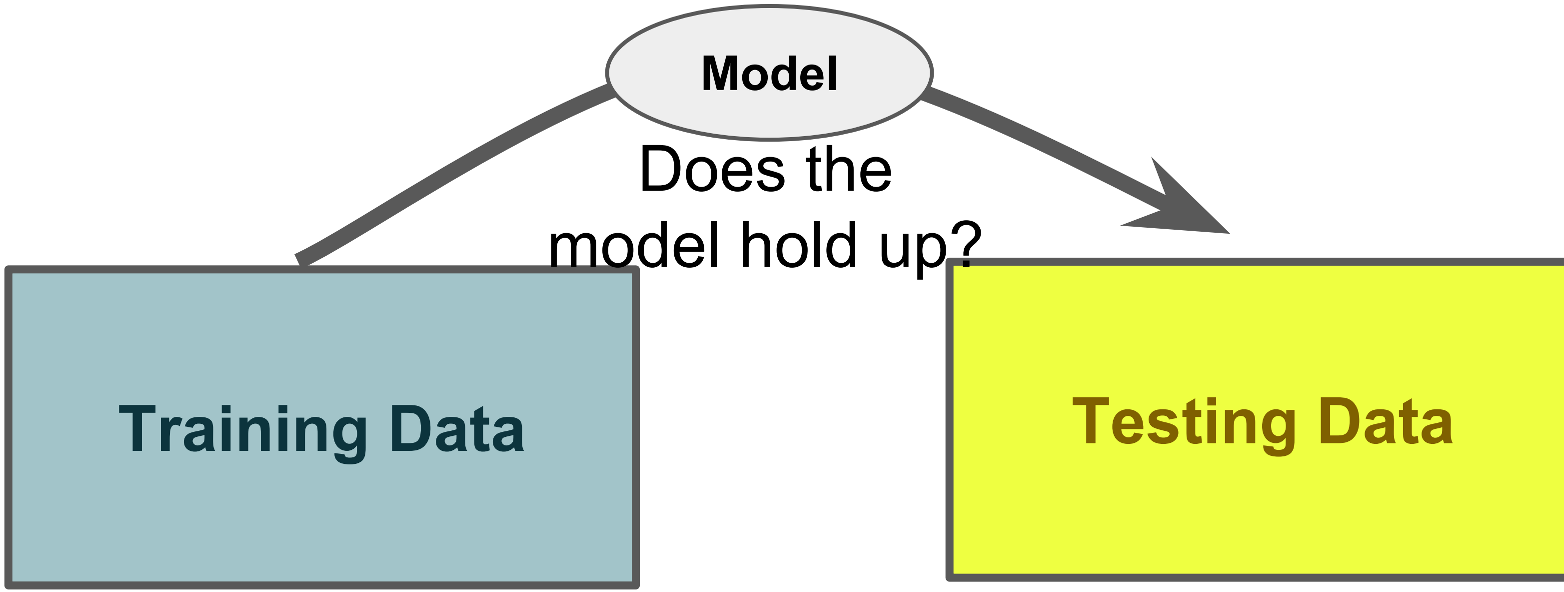
What if only 2 predictors?

$$logit(Y) = 0 + 2*X_1 + 2*X_2$$

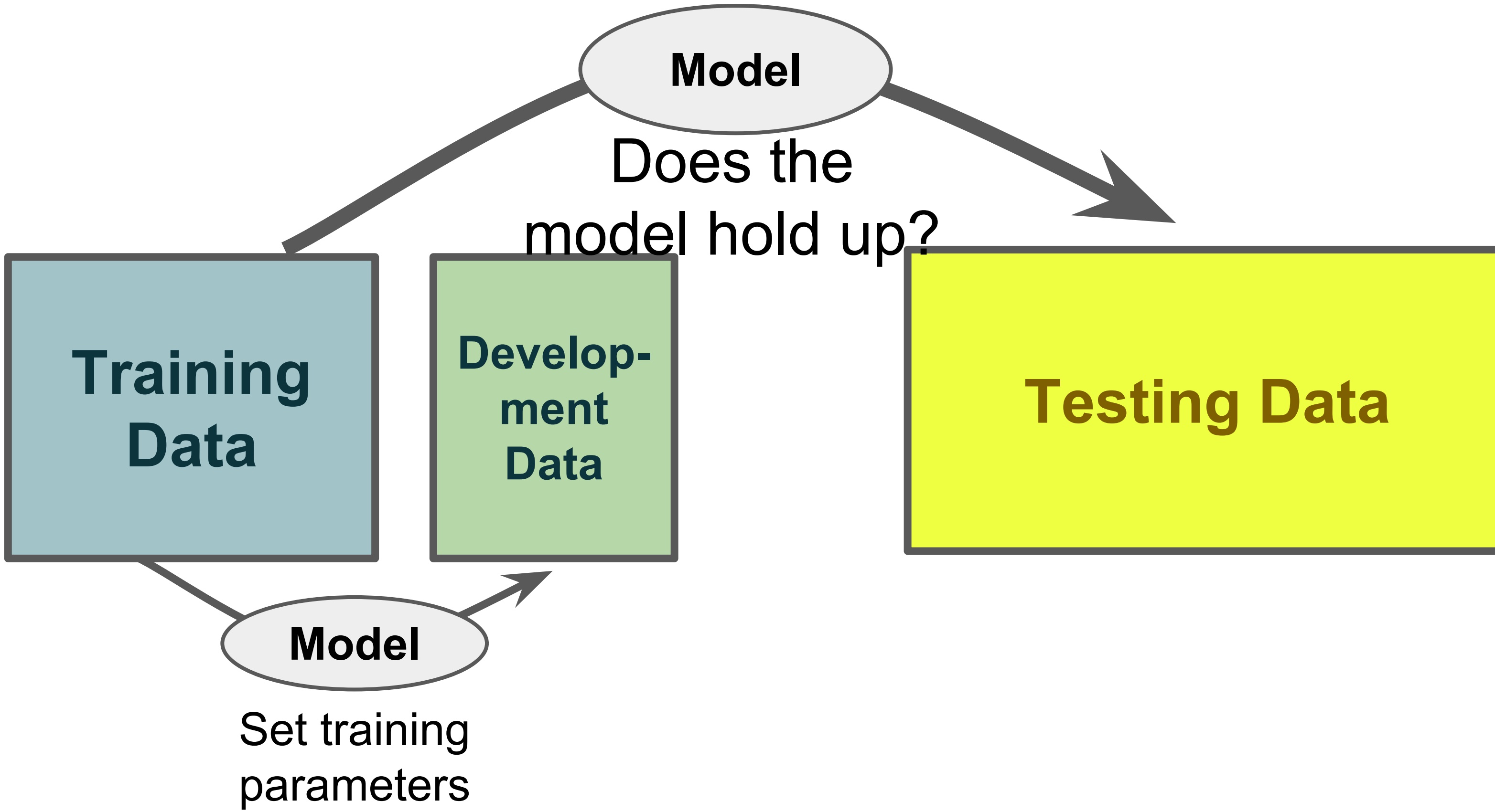
Common Goal: Generalize to new data



Common Goal: Generalize to new data



Common Goal: Generalize to new data



Feature Selection / Subset Selection

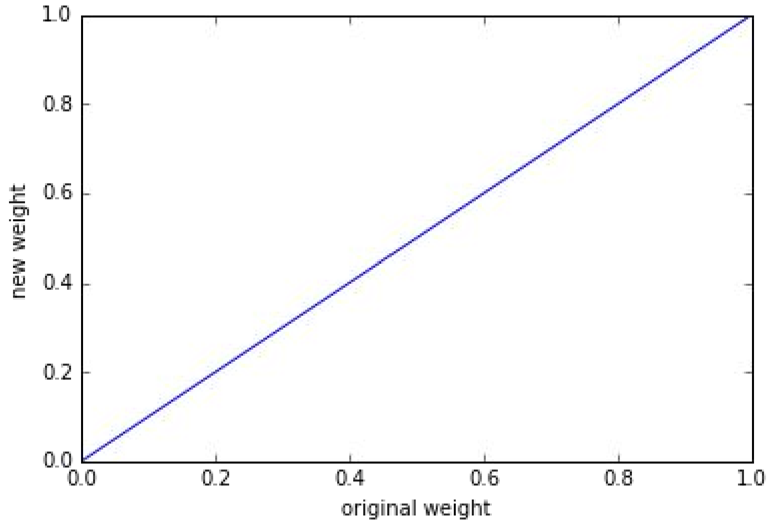
(bad) solution to overfit problem

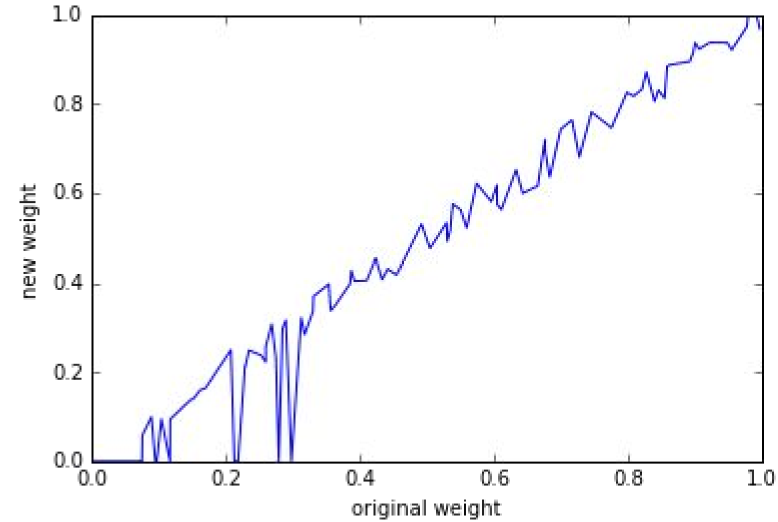
Use less features based on Forward Stepwise Selection:

- start with current_model just has the intercept (mean)
  remaining_predictors = all_predictors

```
 for i in range(k):
   #find best p to add to current_model:
   for p in remaining_prepdictors
      refit current_model with p
       #add best p, based on RSS_p to current_model
   #remove p from remaining predictors
```

Regularization (Shrinkage)





No selection (weight=beta)

forward stepwise

Why just keep or discard features?

Regularization (L2, Ridge Regression)

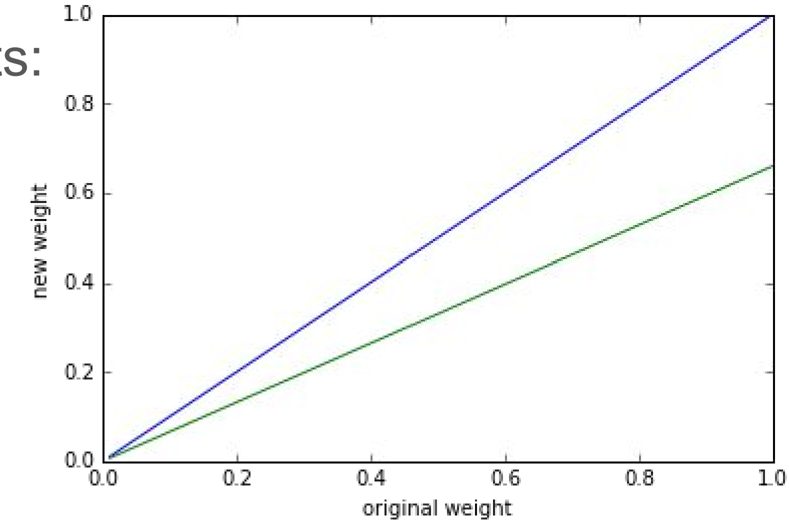
Idea: Impose a penalty on size of weights:

Ordinary least squares objective:

$$\hat{\beta} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2\}$$

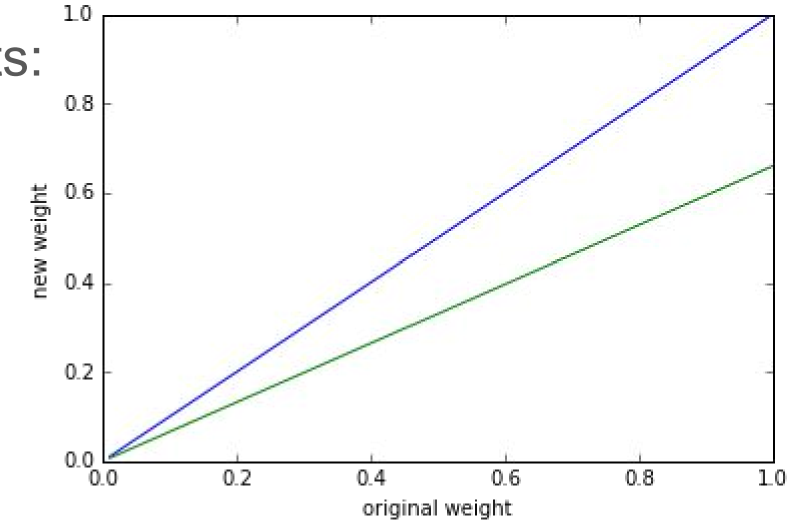
Ridge regression:

$$\hat{\beta}^{ridge} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{m} \beta_j^2\}$$



Regularization (L2, Ridge Regression)

Idea: Impose a penalty on size of weights:



Ordinary least squares objective:

$$\hat{\beta} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2\}$$

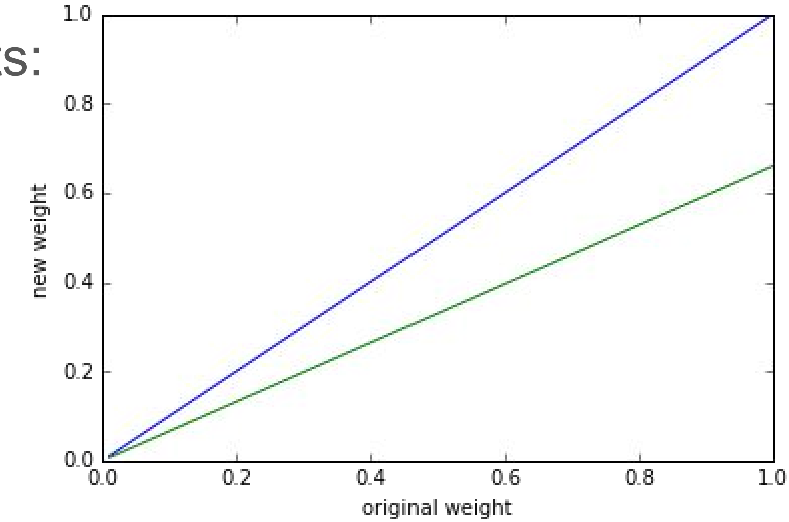
Ridge regression:

$$\hat{\beta}^{ridge} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{m} \beta_j^2\}$$

$$\lambda||\beta||_2^2$$

Regularization (L2, Ridge Regression)

Idea: Impose a penalty on size of weights:



Ordinary least squares objective:

$$\hat{\beta} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2\}$$

Ridge regression:

$$\hat{\beta}^{ridge} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{m} \beta_j^2\}$$

$$\lambda||\beta||_2^2$$

In Matrix Form:

$$\mathrm{RSS}(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^Ty$$

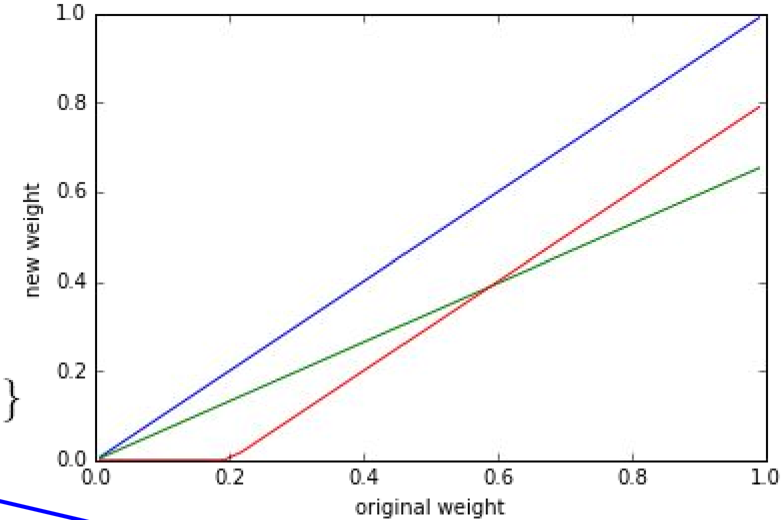
I: m x m identity matrix

Regularization (L1, The "Lasso")

Idea: Impose a penalty and zero-out some weights

The Lasso Objective:

$$\hat{\beta}^{lasso} = argmin_{\beta}\{\frac{1}{2}\sum_{i=1}^{N}(Y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{m}|\beta_j|\}$$



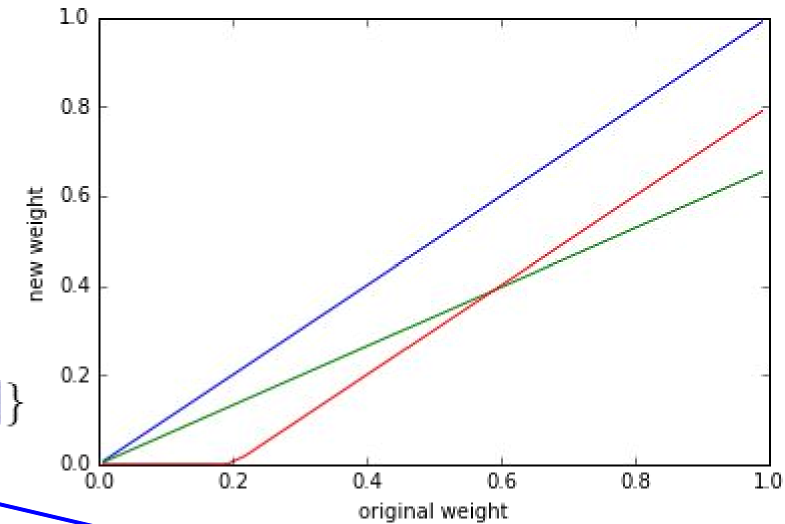
$$\lambda||\beta||_1$$

Regularization (L1, The "Lasso")

Idea: Impose a penalty and zero-out some weights

The Lasso Objective:

$$\hat{\beta}^{lasso} = argmin_{\beta}\{\frac{1}{2}\sum_{i=1}^{N}(Y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{m} |\beta_j|\}$$

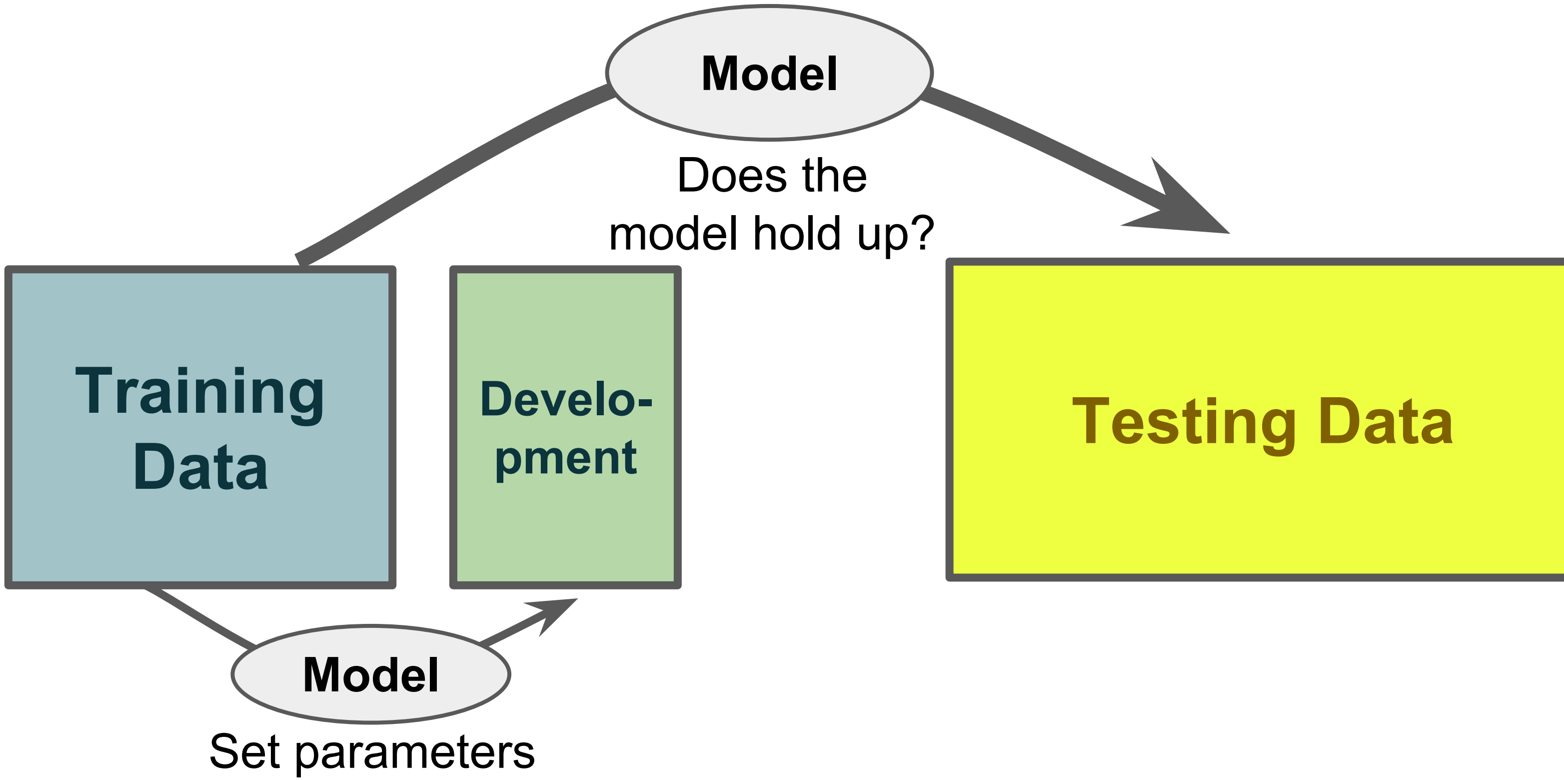


$$\lambda||\beta||_1$$

No closed form matrix solution, but often solved with coordinate descent.

Application: p ≅ n or p >> n (p: features; n: observations)

Common Goal: Generalize to new data



N-Fold Cross-Validation

Goal: Decent estimate of model accuracy