Statistical Preliminaries

Stony Brook University
CSE545, Fall 2016

Random Variables

X: A mapping from Ω 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 = $\{<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH> \ldots\}$

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?

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

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 $\mathbf{\Omega}$ to $\mathbb{R}$ that describes the question we care about in practice.

Example: $\mathbf{\Omega}$ = 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 $\mathbf{\Omega}$. 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 $\mathbf{\Omega}$ 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

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

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 <u>exactly</u> 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, Revisited

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 $i \in \boldsymbol{\Omega}$

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

How to model?

s?

inches?

Continuous Random Variables

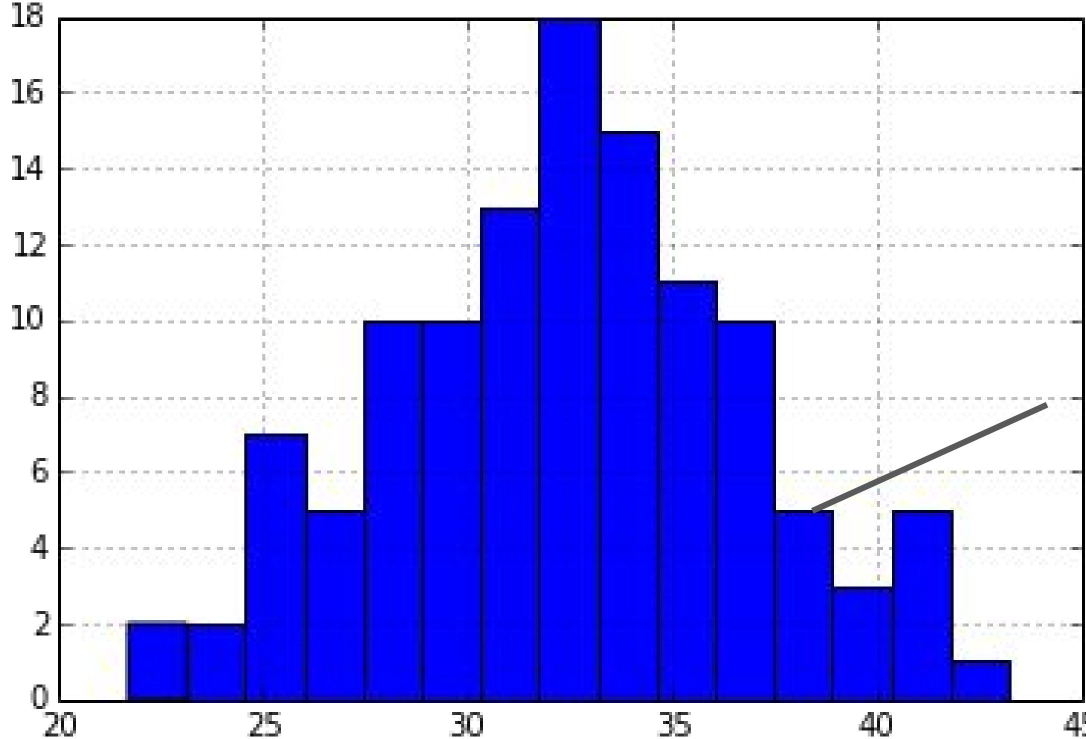
Discretize them!

(group into discrete bins)

How to model?

Continuous Random Variables

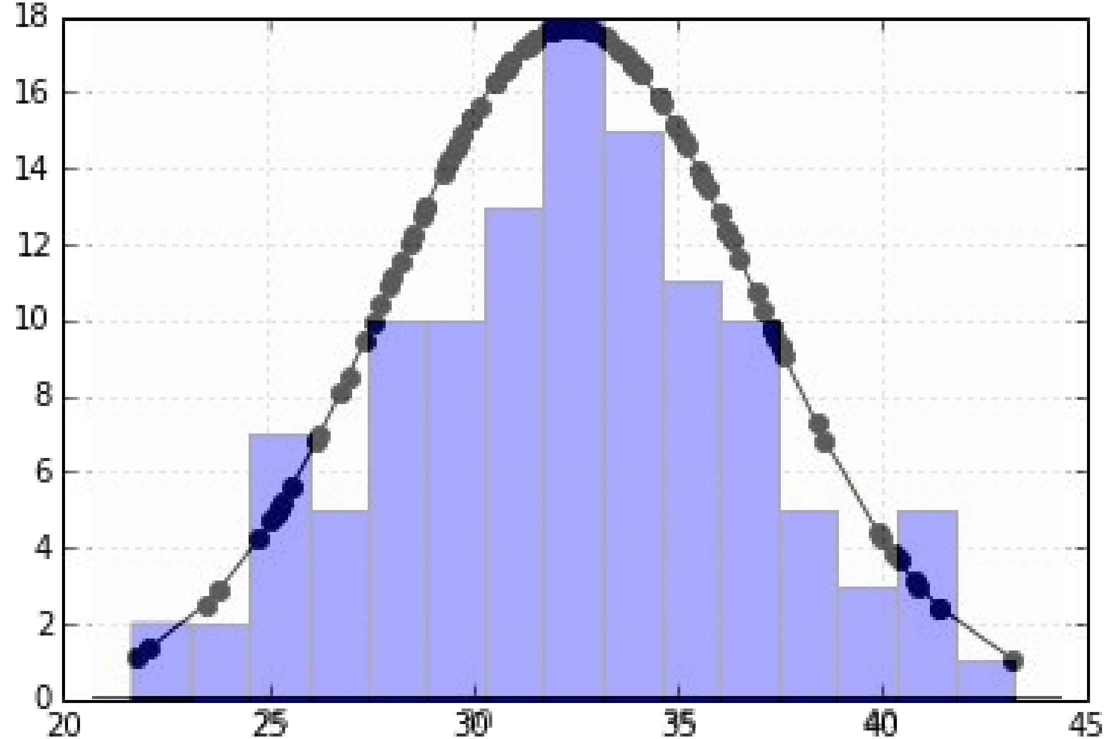
$\mathrm{P}(bin{=}8) = .32$



$\mathrm{P}(bin{=}12) = .08$

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:

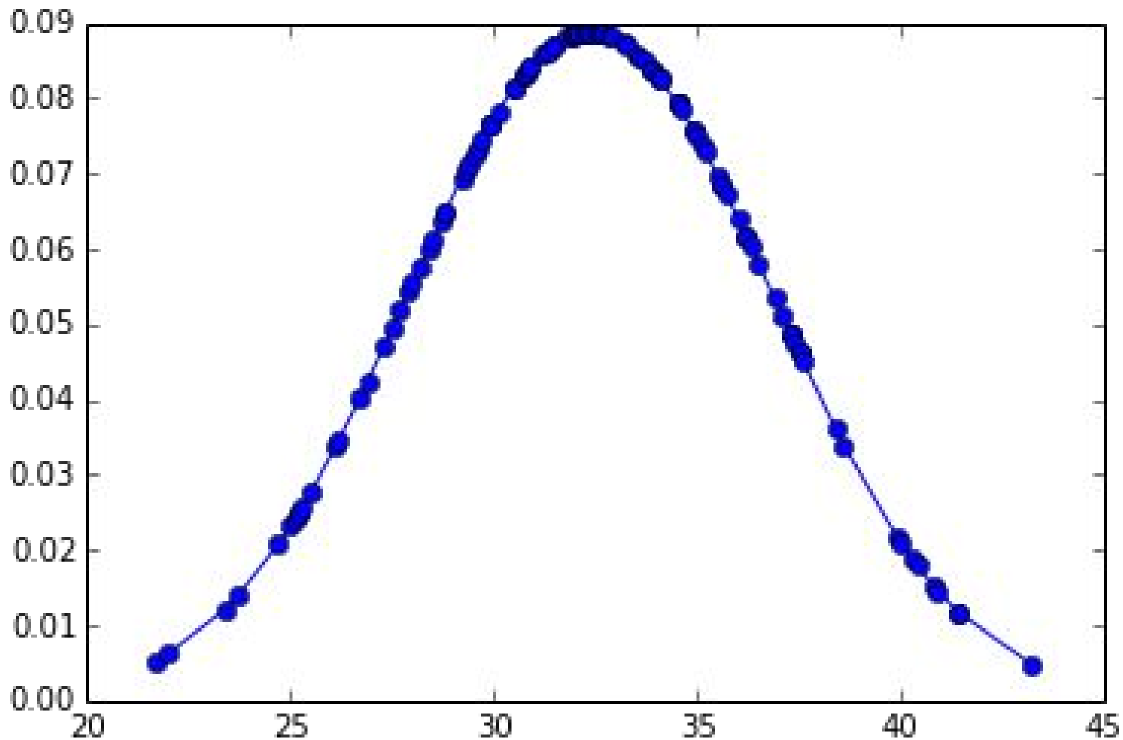
$$f_X(x) \geq 0, \text{for all } x \in X,$$

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

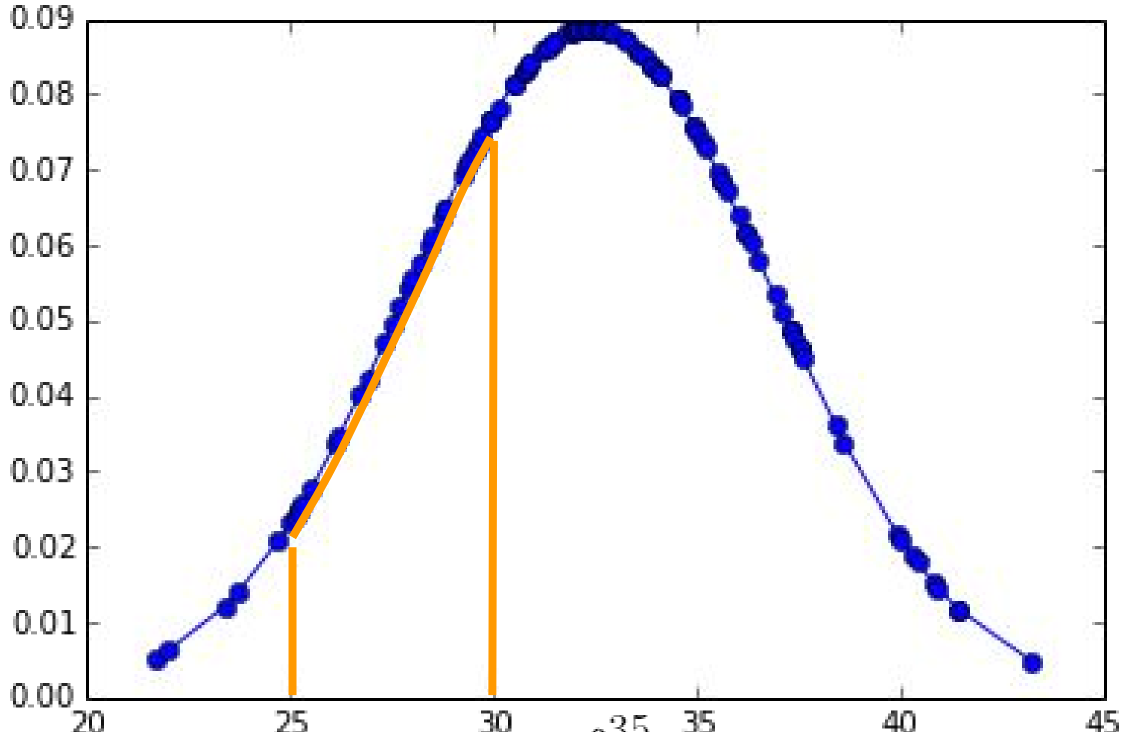
$$\mathrm{P}(a < X < b) = \int_a^b f_X(x)dx$$

fx **: "probability density function" (pdf)**

Continuous Random Variables



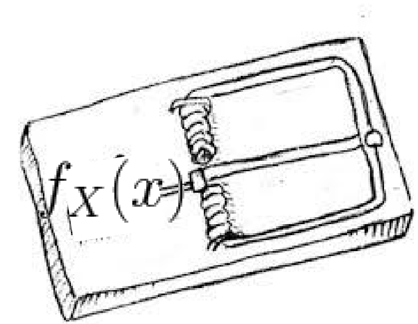
Continuous Random Variables



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

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

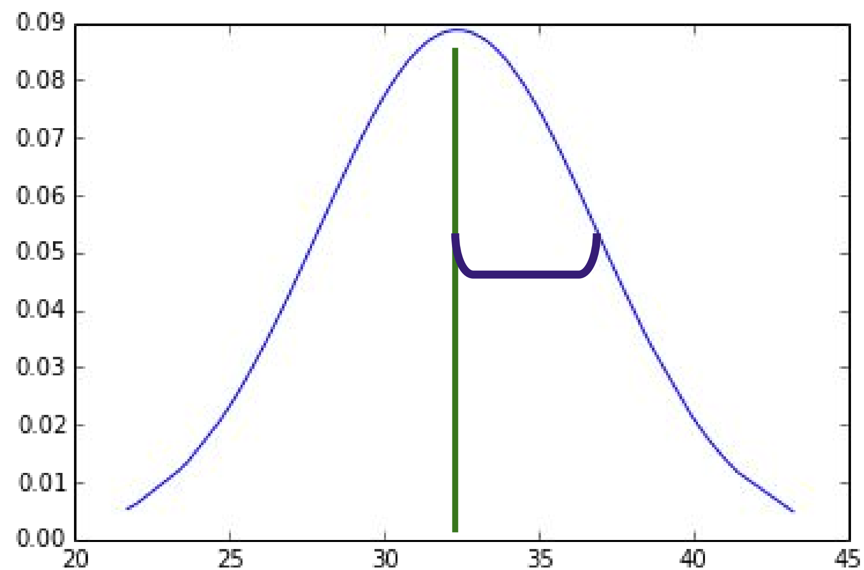
Continuous Random Variables

A Common Probability Density Function

Continuous Random Variables

Common *pdfs*: Normal(μ, σ^2)

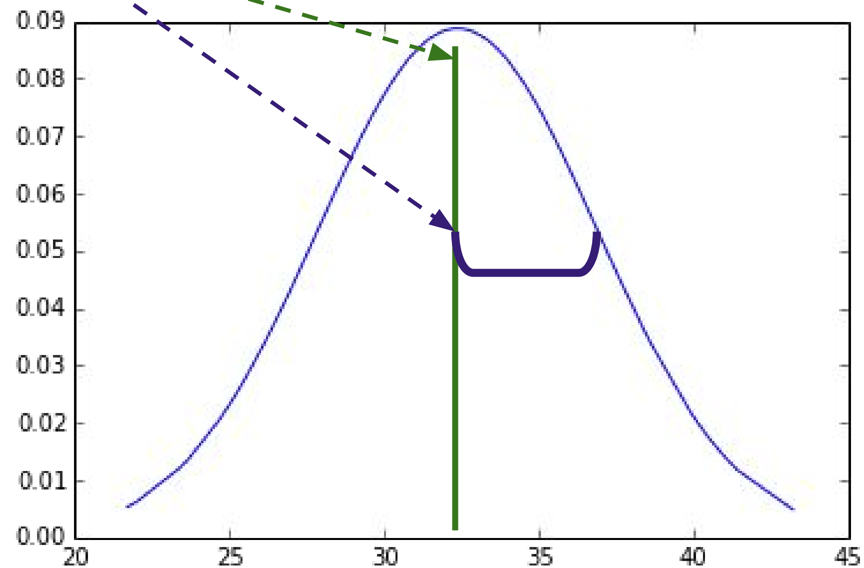
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Continuous Random Variables

Common *pdf*s: Normal(μ, σ²)

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



μ: mean (or "center")

= expectation

σ²: variance,

σ: standard deviation

Continuous Random Variables

Common *pdf*s: Normal(μ, σ^2)

Credit: Wikipedia

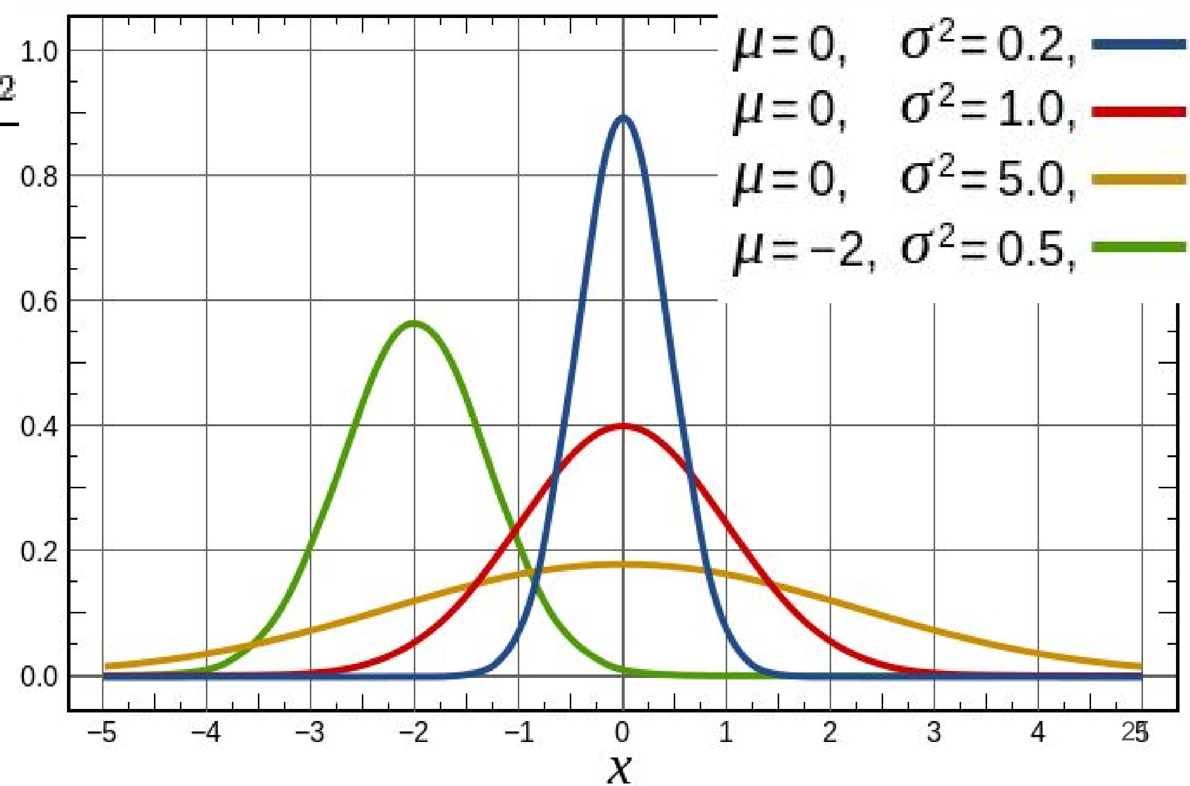
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ: mean (or "center")

= expectation

σ^2: variance,

σ: standard deviation

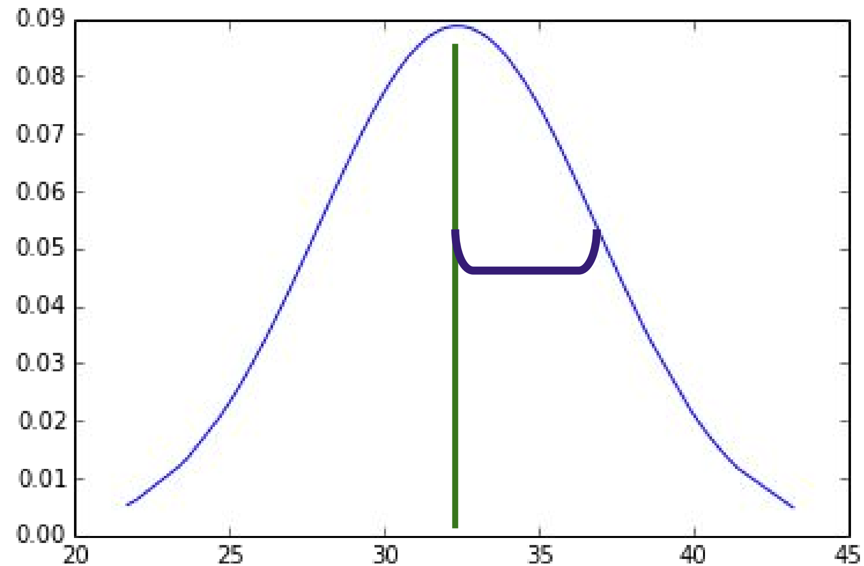


Continuous Random Variables

Common *pdfs*: $\text{Normal}(\mu, \sigma^2)$

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

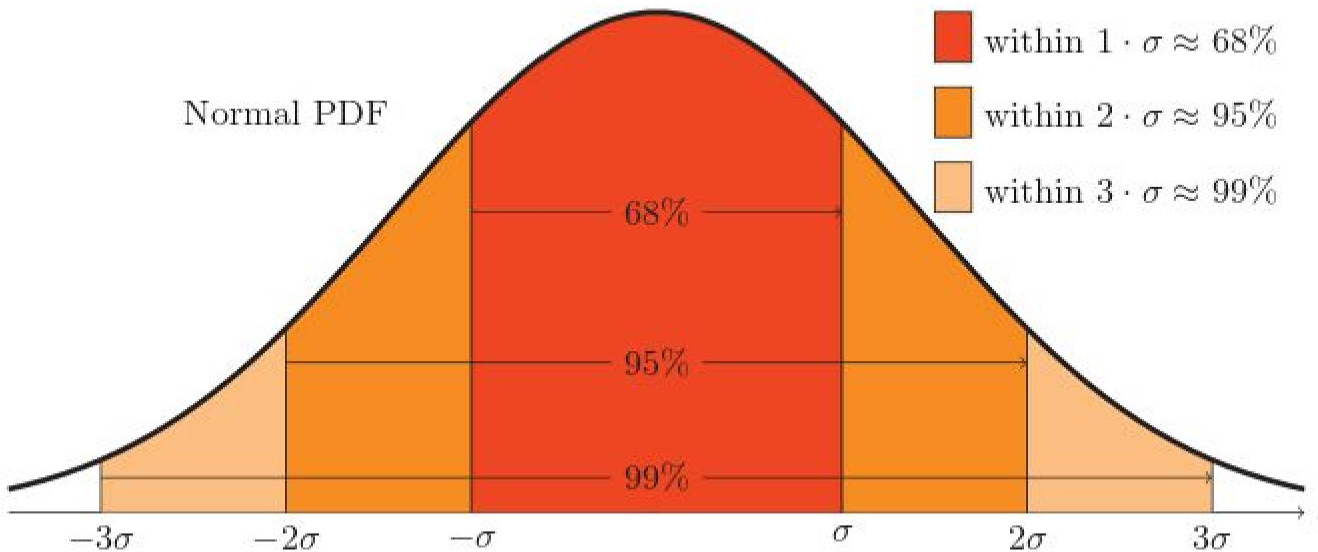
- subtract the mean, μ (aka "mean centering")
- divide by the standard deviation, σ

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

Continuous Random Variables

Common *pdfs*: Normal(0, 1)

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



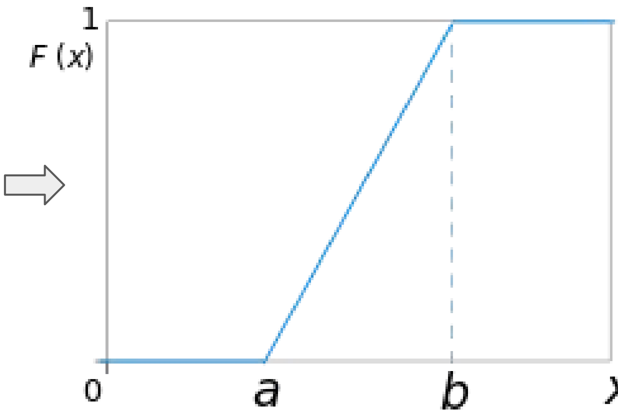
Credit: MIT Open Courseware: Probability and Statistics

Cumulative Distribution Function

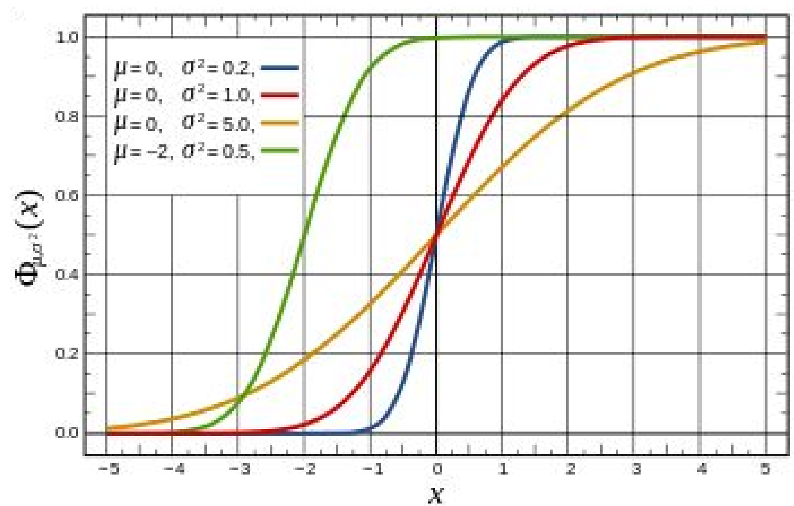
For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

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

Uniform



Normal

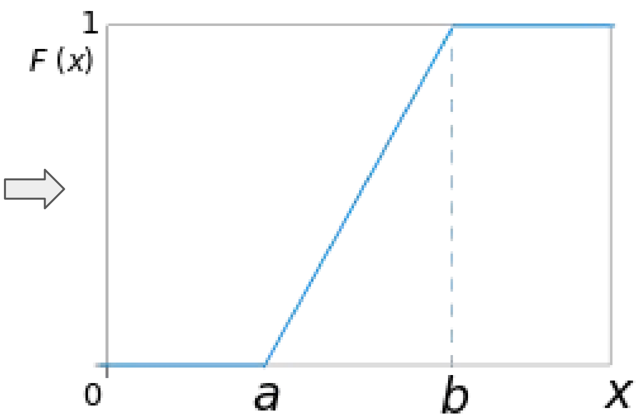


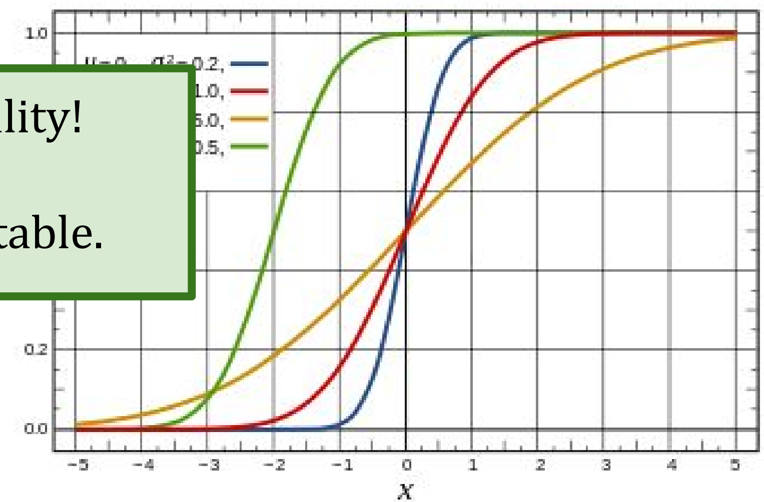
Cumulative Distribution Function

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

$\mathbf{X}$: 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.

Discrete Random Variables

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

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

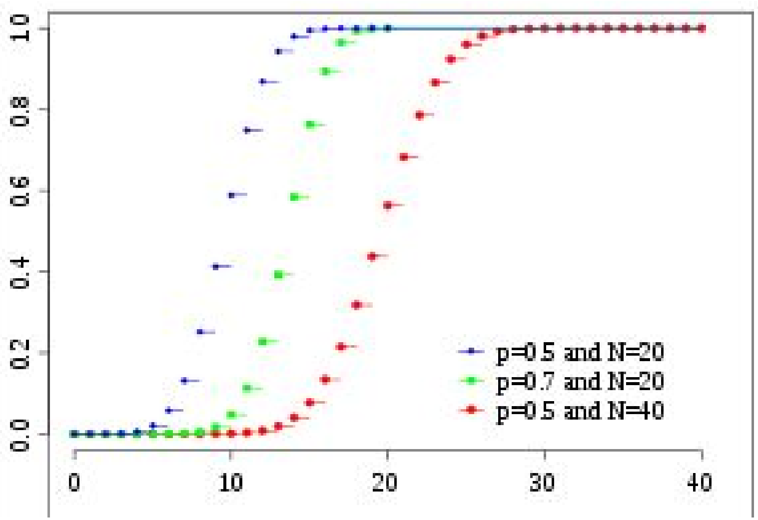
X is a *discrete random variable* if it takes only a countable number of values.

Discrete Random Variables

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

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

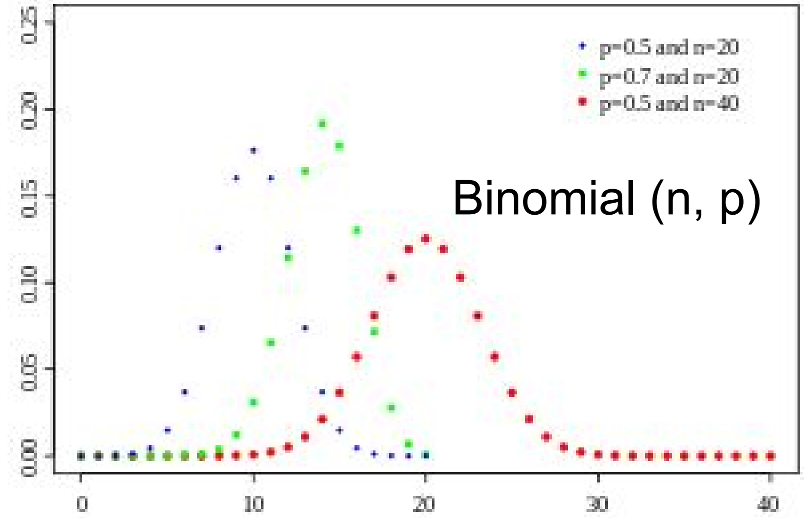
X is a *discrete random variable* if it takes only a countable number of values.



Binomial (n, p)

(like normal)

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:

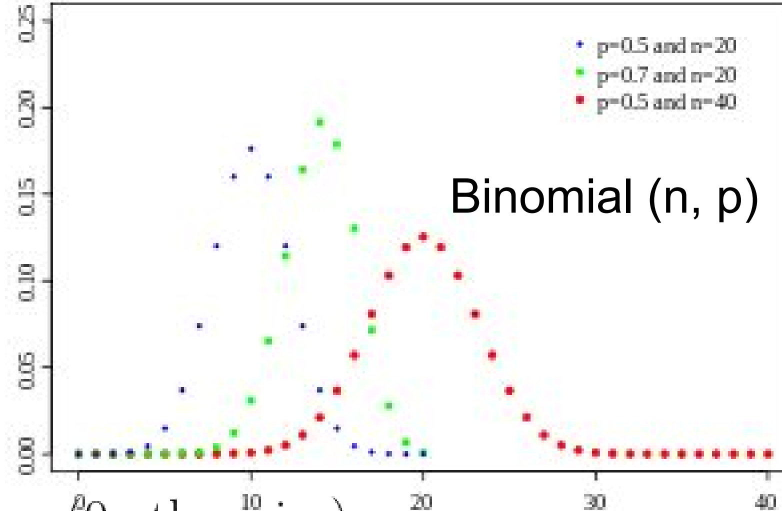
$$f_X(x) = \mathrm{P}(X = x)$$

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



Binomial (n, p)

Two Common **Discrete** Random Variables

- Binomial(n, p)

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

  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 a 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 a 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% alternative is true).

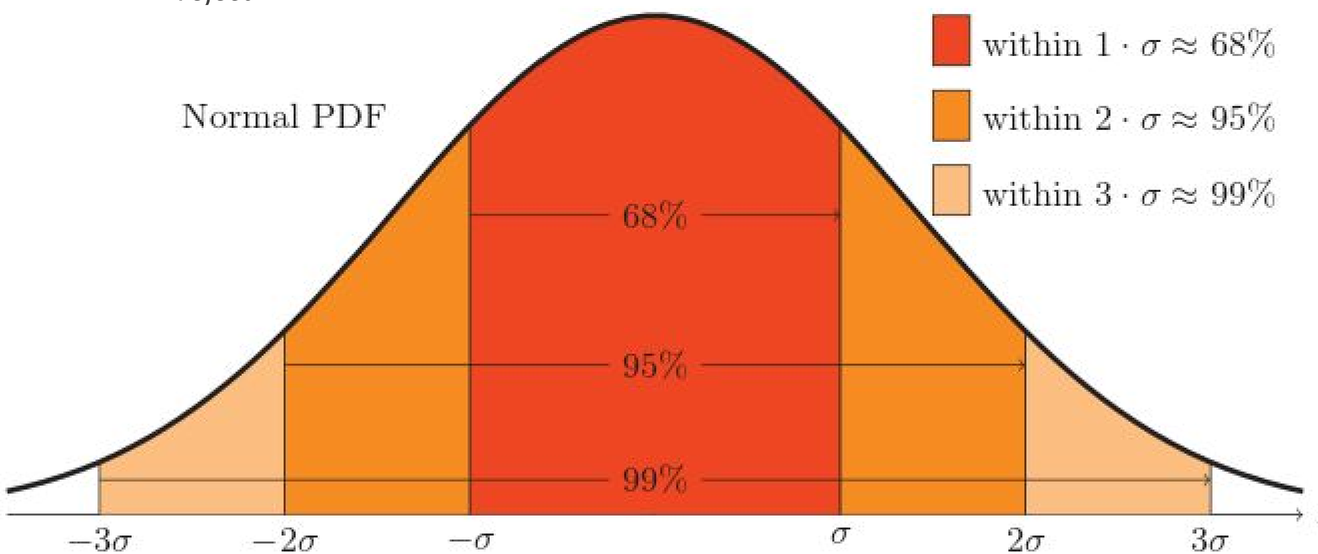
Example: Hypothesize a coin is biased.
H_0: the coin is not biased (i.e. flipping n times results in a Binomial(n, 0.5))

Hypothesis Testing

H_0: *null hypothesis* -- some "default" value (usually that one's hypothesis is false)

H_1: the alternative -- usually that one's "hypothesis" is true

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

H_0: null hypothesis -- some "default" value (usually that one's hypothesis is false)

H_1: the alternative -- usually that one's "hypothesis" is true

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.

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

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))

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	Reject H_0	Type I error	correct decision
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)**

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/)