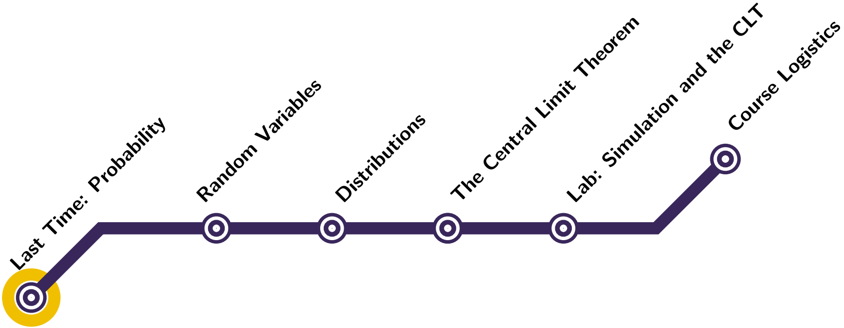
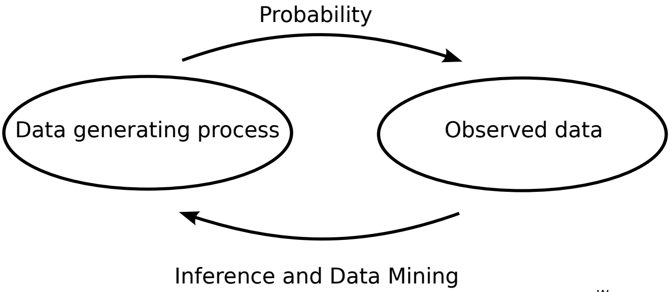
Today's Roadmap



Probability and Statistics



Wasserman, 2013

Axioms of Probability

Definition: A function Pr that assigns a real number $Pr(A)$ to each event A is a probability distribution if it satisfies the following three axioms:

Axiom 1: $Pr(A) \geq 0$, for every A.
Axiom 2: $Pr(S) = 1$.
Axiom 3: If $A_1, A_2, \ldots$ are disjoint then,

$$Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} Pr(A_i)$$

.

Axioms of Probability

Definition: A function Pr that assigns a real number $Pr(A)$ to each event A is a probability distribution if it satisfies the following three axioms:

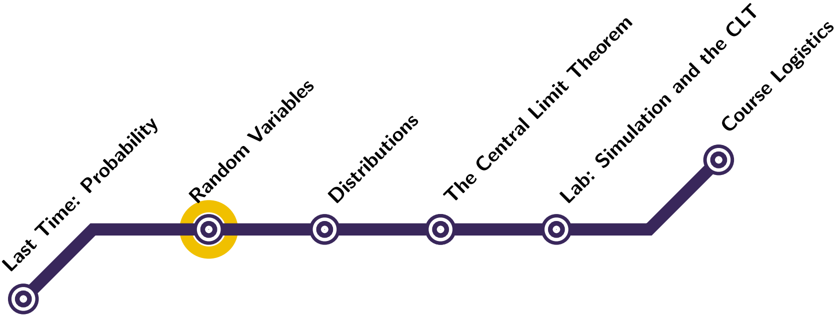
Axiom 1: $Pr(A) \geq 0$, for every A.
Axiom 2: $Pr(S) = 1$.
Axiom 3: If $A_1, A_2, \ldots$ are disjoint then,

$$Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} Pr(A_i)$$

.

List of all possible outcomes with their associated probabilities.

Today's Roadmap



Introduction to Random Variables

- Consider some experiment for which the sample space is S
- A real-valued function that is defined on S is called a random variable
- A random variable X is a random process with a numerical outcome.

Random Variables

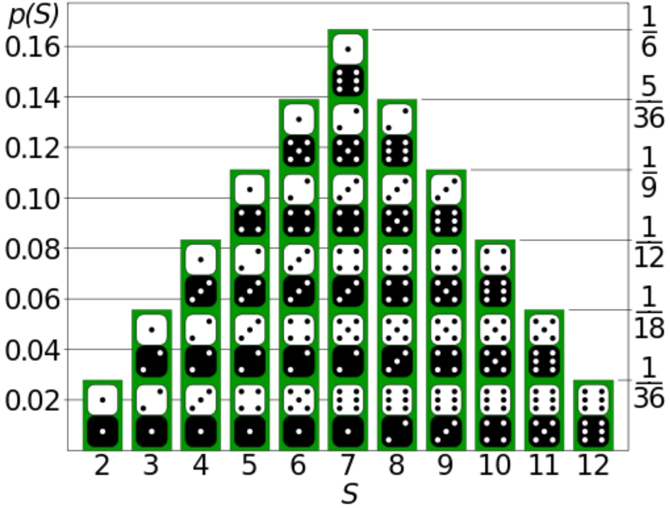
- We can determine a probability distribution for the possible values of a random variable X

Random Variables

- We can determine a probability distribution for the possible values of a random variable X
- The collection of all these probabilities is the distribution of X

Random Variables

Probability distribution for the sum of two dice:



Allows for computation of probabilities of events.

Distribution Functions

- Given a random variable X we define the cumulative distribution function or CDF is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

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

Distribution Functions

- Given a random variable X we define the cumulative distribution function or CDF is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

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

- The CDF effectively contains all the information about the random variable.

Types of Random Variables

- Discrete: finite or countable list of possible values; probability mass function (pmf) assigns a probability to each value

$$f_X(x) = Pr(X = x)$$

- Continuous: taking any numerical value in an interval; probability density function (pdf) assigns a probabilities to intervals

$$Pr(a \leq X \leq b) = \int_a^b f_X(x)dx$$

Discrete Random Variables

- We say X has a discrete distribution or that X is a discrete random variable if X can take only a countable number k of different values $x_1, \ldots, x_k$

Discrete Random Variables

- We say X has a discrete distribution or that X is a discrete random variable if X can take only a countable number k of different values $x_1, \ldots, x_k$
- If a random variable X has a discrete distribution, the probability function or probability mass function is defined as the function f such that for every real number x,

$$f(x) = Pr(X = x)$$

Discrete Random Variables Example

Suppose the value of a random variable X is equally likely to be each of k integers $1, 2, \ldots, k$.

Discrete Random Variables Example

Suppose the value of a random variable X is equally likely to be each of k integers $1, 2, \ldots, k$.

What is the probability mass function (pmf) of X?

Discrete Random Variables Example

Suppose the value of a random variable X is equally likely to be each of k integers $1, 2, \ldots, k$.

What is the probability mass function (pmf) of X?

$$f(x) = \begin{cases} \frac{1}{k} & x = 1, 2, \ldots, k \\ 0 & otherwise \end{cases}$$

Discrete Random Variables Example

Suppose the value of a random variable X is equally likely to be each of k integers $1, 2, \ldots, k$.

What is the probability mass function (pmf) of X?

$$f(x) = \begin{cases} \frac{1}{k} & x = 1, 2, \ldots, k \\ 0 & otherwise \end{cases}$$

This is discrete distribution is called the uniform distribution on the integers.

Continuous Random Variables

- A random variable has a continuous distribution if there exists a nonnegative function f defined on the real that such that for every subset A of the real line the probability that X takes on a value in A is the integral of f over the set A.

Continuous Random Variables

- A random variable has a continuous distribution if there exists a nonnegative function f defined on the real that such that for every subset A of the real line the probability that X takes on a value in A is the integral of f over the set A.
- We will primarily be concerned with sets that are intervals:

$$Pr(a < X < b) = \int_a^b f(x)dx$$

- The function f is called the probability density function (pdf) of X
- f must satisfy the following $f(x) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Continuous Random Variables Example

Suppose that X has pdf

$$f(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Continuous Random Variables Example

Suppose that X has pdf

$$f(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

X is said to have a Uniform(0,1) distribution, this situation captures the idea of choosing a point at random between 0 and 1.

Expectation of Random Variables

- The expectation or mean of a random variable X is the average value of X

  Definition: The expected value of X is defined to be

  $$\mathrm{E}(X) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

- We often refer to the expected value of a random variable as μ

Expectation Example

Suppose we flip a fair coin two times. Let X be the number of heads. What is the expected value of X?

Expectation Example

Suppose we flip a fair coin two times. Let X be the number of heads. What is the expected value of X?

$$\mathrm{E}(X) = \sum_x x f(x) = (0 \times f(0)) + (1 \times f(1)) + (2 \times f(2))$$

Expectation Example

Suppose we flip a fair coin two times. Let X be the number of heads. What is the expected value of X?

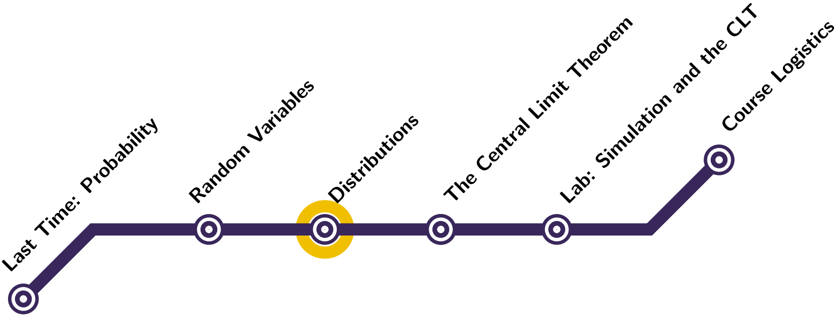
$$\mathrm{E}(X) = \sum_x x f(x) = (0 \times f(0)) + (1 \times f(1)) + (2 \times f(2))$$

$$(0 \times 1/4) + (1 \times 1/2) + (2 \times 1/4) = 1$$

Random Variables in Practice

Any numerical quantity that does not have a fixed value and has a distribution function associated with it becomes a RV.

Today's Roadmap



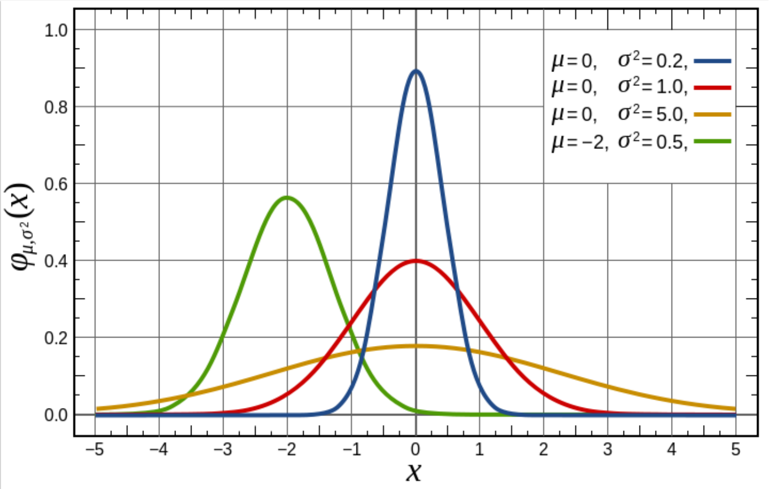
Distributions of Random Variables

- Normal distribution: symmetric, unimodal
- Bernoulli distribution: success/failure
- Geometic distribution: success on nth trial
- Binomial distribution: k successes in n trials
- Neg. binomial distribution: kth success on nth trial
- Poisson distribution: k rare events

The Normal Distribution

- Many variables are nearly normal, none are exactly normal

Normal Distribution Example



The Normal (Gaussian) Distribution

- A random variable X has a normal distribution with parameters μ and σ^2 if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- We write $X \sim \mathrm{N}(\mu, \sigma^2)$

The Normal Distribution: Z scores

- The Z score of an observation is the number of standard deviations it falls above/below the mean.

$$Z = \frac{x - \mu}{\sigma}$$

- Use this idea to find the probability for a particular observation - how likely/extreme is that observation?

Z scores: Self Test

Based on a sample of 100 men, the distribution of heights of male adults between ages 20 and 62 in the US is nearly normal with a mean of 70 inches and a s.d. of 3.3 in.

Mike in 5'7" and Jim is 6'4".

- What is Mike's height percentile?
- What is Jim's percentile?
- What is the probability that a randomly chosen male in class is taller than Jim?

Hint: Use the `pnorm()` function in R.

Z scores: Self Test

- What is Mike's height percentile?

Z scores: Self Test

- What is Mike's height percentile?

$$Z_{Mike} = \frac{67 - 70}{3.3} = -.91 \Rightarrow 0.1814$$

Z scores: Self Test

- What is Mike's height percentile?

$$Z_{Mike} = \frac{67 - 70}{3.3} = -.91 \Rightarrow 0.1814$$

- What is Jim's percentile?

Z scores: Self Test

- What is Mike's height percentile?

$$Z_{Mike} = \frac{67 - 70}{3.3} = -.91 \Rightarrow 0.1814$$

- What is Jim's percentile?

$$Z_{Jim} = \frac{76 - 70}{3.3} = 1.82 \Rightarrow 0.9656$$

Z scores: Self Test

- What is Mike's height percentile?

$$Z_{Mike} = \frac{67-70}{3.3} = -.91 \Rightarrow 0.1814$$

- What is Jim's percentile?

$$Z_{Jim} = \frac{76-70}{3.3} = 1.82 \Rightarrow 0.9656$$

- What is the probability that a randomly measure male in class is taller than Jim?

Z scores: Self Test

- What is Mike's height percentile?

$$Z_{Mike} = \frac{67 - 70}{3.3} = -.91 \Rightarrow 0.1814$$

- What is Jim's percentile?

$$Z_{Jim} = \frac{76 - 70}{3.3} = 1.82 \Rightarrow 0.9656$$

- What is the probability that a randomly measure male in class is taller than Jim?

$$1 - .9656 = 0.344$$

The Standard Normal Distribution

- If $\mu = 0$ and $\sigma^2 = 1$ we say that X has a standard normal distribution
- We use Z to denote a random variable with a standard normal distribution
- The standard normal distribution comes up so frequently we have special symbols to denote its pdf $\phi(z)$ and cdf $\Phi(z)$

Important Facts about Normal Distribution

There is no closed form expression for $\Phi(z)$, but we can rely on these facts to calculate probabilities.

1. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $Z = \frac{(X-\mu)}{\sigma} \sim \mathrm{N}(0,1)$
2. If $Z \sim \mathrm{N}(0,1)$, then $X = \mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$

Important Facts about Normal Distribution

There is no closed form expression for $\Phi(z)$, but we can rely on these facts to calculate probabilities.

1. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $Z = \frac{(X-\mu)}{\sigma} \sim \mathrm{N}(0,1)$
2. If $Z \sim \mathrm{N}(0,1)$, then $X = \mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$

$$Pr(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Normal Distribution Example

Suppose that $X \sim \mathrm{N}(3, 5)$. Find $Pr(X > 1)$.

Normal Distribution Example

Suppose that $X \sim \mathrm{N}(3, 5)$. Find $Pr(X > 1)$.

$$Pr(X > 1) = 1 - Pr(X < 1)$$

Normal Distribution Example

Suppose that $X \sim \mathrm{N}(3,5)$. Find $Pr(X > 1)$.

$$Pr(X > 1) = 1 - Pr(X < 1)$$

$$= 1 - Pr\left(Z < \frac{1-3}{\sqrt{5}}\right) = 1 - \Phi(-0.8944) = 0.81$$

Evaluating the Normal Approximation

- Plot the data!
- Normal probability plots in R: `qqplot()`

Foundations for Statistical Inference

- We carry out an experiment and get a random sample of the underlying population
- Data are the values in the sample
- Our aim is to infer the population probability distribution (parameters) from the data we observe in the sample
- Our assumption is that that samples behave approximately as the population

Foundations for Statistical Inference

- We carry out an experiment and get a random sample of the underlying population
- Data are the values in the sample
- Our aim is to infer the population probability distribution (parameters) from the data we observe in the sample
- Our assumption is that that samples behave approximately as the population

This is statistical inference!

Statistical Inference: Point Estimates

- In many situations we want to estimated the population mean based on a sample.
- What should we do?

Statistical Inference: Point Estimates

- In many situations we want to estimated the population mean based on a sample.
- What should we do?
- Take the sample mean!

Statistical Inference: Point Estimates

- In many situations we want to estimated the population mean based on a sample.
- What should we do?
- Take the sample mean!
- The sample mean $\overline{x}$ is called a point estimate of the population mean.
- It is our simple best guess at the population mean.

Statistical Inference: Point Estimates

- Suppose we take another sample. Will our estimate be the same?

Statistical Inference: Point Estimates

- Suppose we take another sample. Will our estimate be the same?
- Estimates vary, this is called sampling variation

Statistical Inference: Point Estimates

- Suppose we take another sample. Will our estimate be the same?
- Estimates vary, this is called sampling variation
- Our estimate may be close to the true parameter but not exactly equal.

Sampling Distribution

The sampling distribution represents the distribution of the point estimates based on sample of a fixed size from a certain population.

Sampling Distribution

- How can we determine how confident we are in our model/estimate/decision?

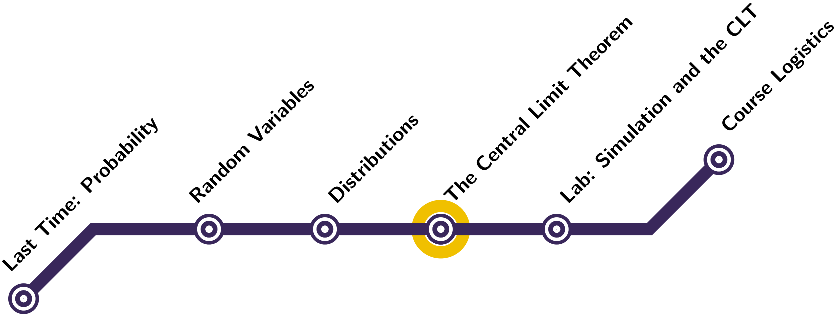
Sampling Distribution

- How can we determine how confident we are in our model/estimate/decision?

- By understanding the sampling distribution!

Break

Today's Roadmap



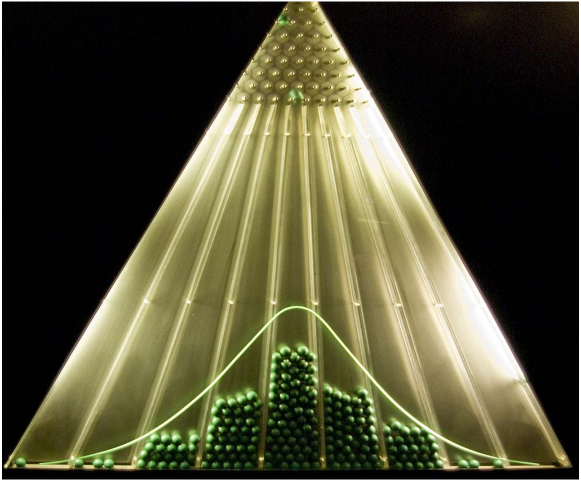
The Central Limit Theorem

The distribution of $\overline{x}$ is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes.

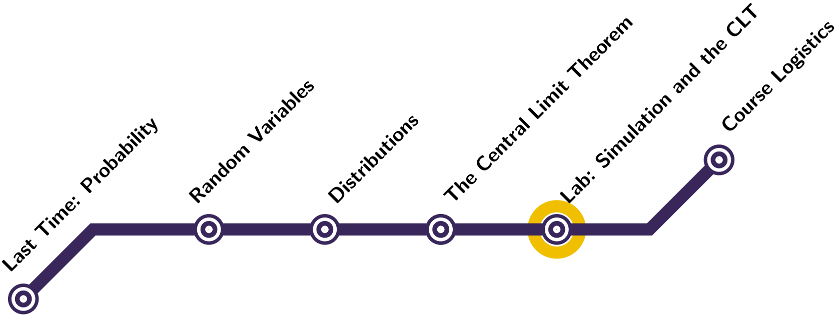
The Central Limit Theorem

- The CLT is about the sampling distribution for the population mean!

The Central Limit Theorem Intuition

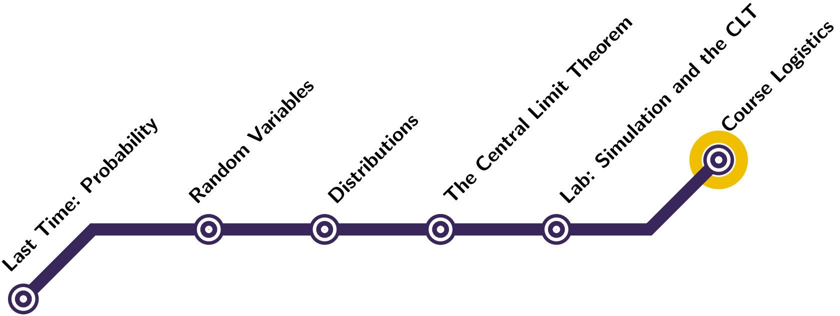


Today's Roadmap



Lab: Simulation and the CLT

Today's Roadmap



Questions?