Lecture 15
The Central Limit Theorem
Sampling Distributions of $\bar{x}$ and $\hat{p}$

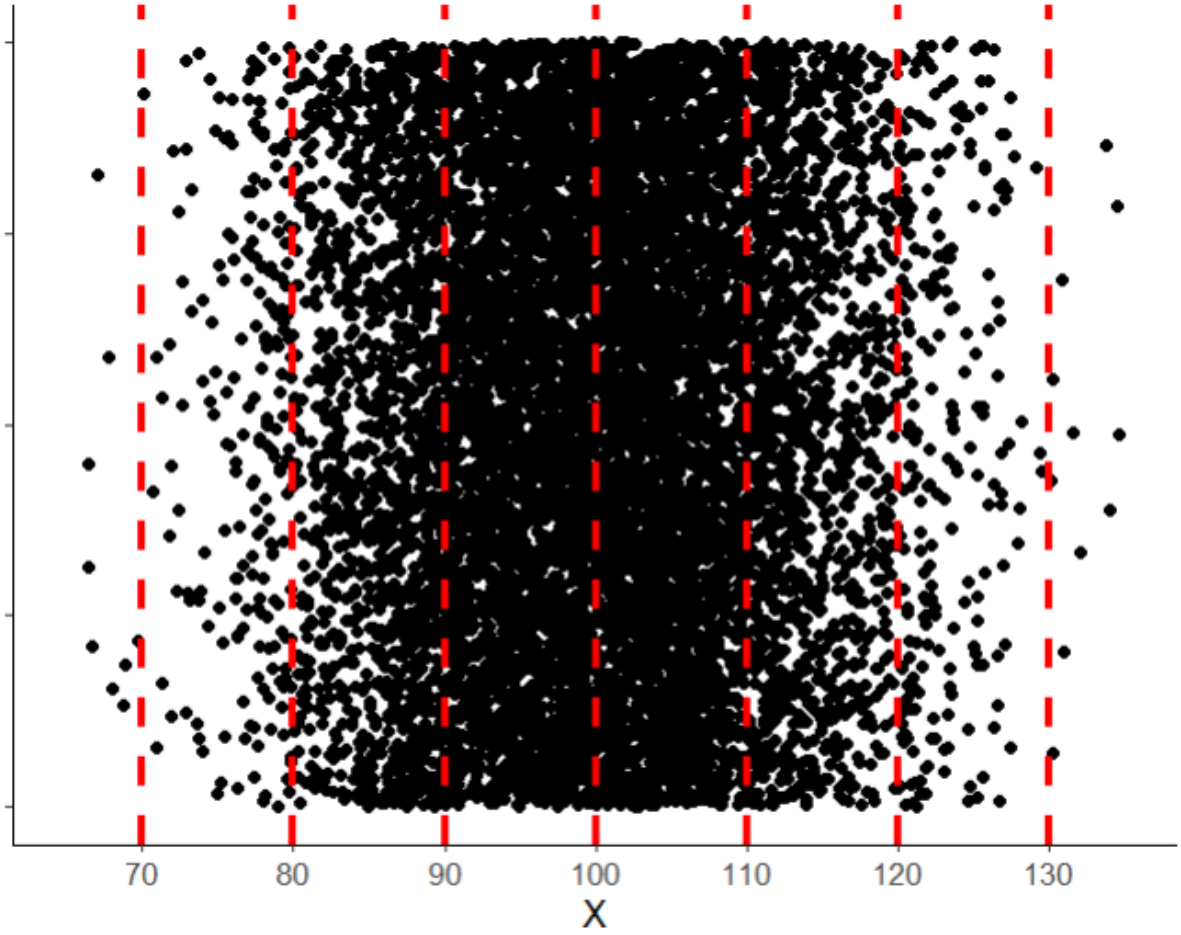
Mean and variance of Continuous Random Variables

- The mean μ of continuous random variable **cannot** (usually) be defined or computed without calculus, but it is the "balance point" of its probability distribution.

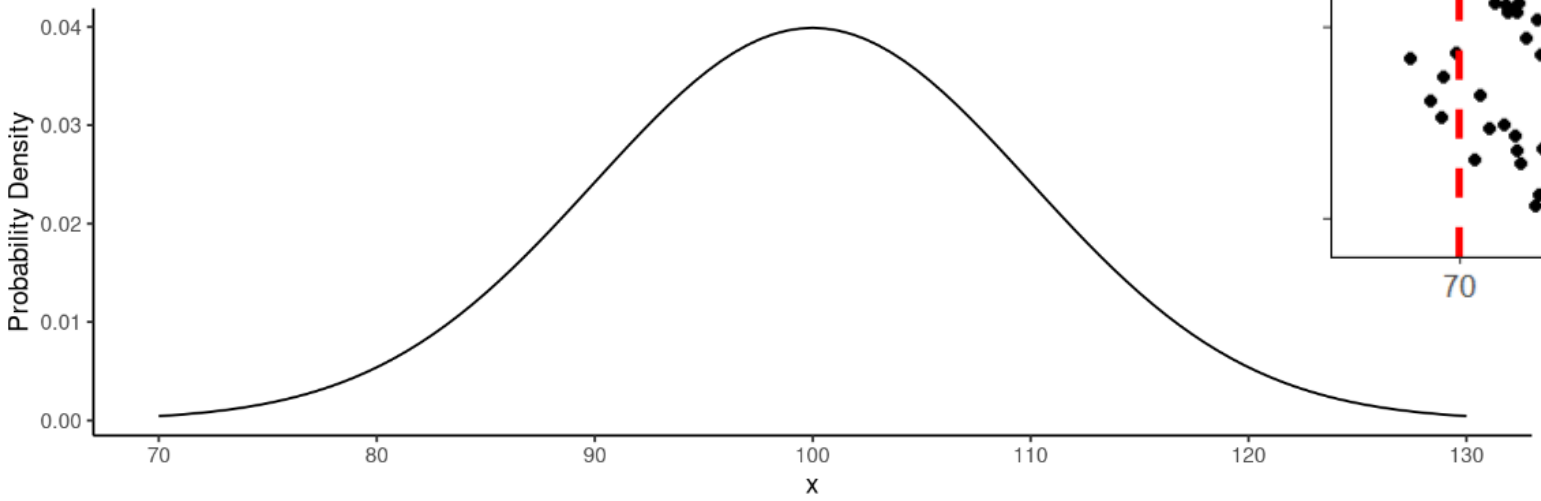
- The standard deviation σ of a continuous random variable **cannot** (usually) be defined or computed without calculus, but it measures the "spread" of the probability distribution.

Continuous Distributions: A few extra points

- Unlike discrete distributions, the height of the curve **DOES NOT** denote the probability of a given value
- The y-axis of a continuous distribution is the probability density
- Because a continuous distribution covers an infinite number of possible outcomes, the **probability of observing any particular value is zero!**



Example: Normal probability distribution with mean $\mu = 100$ and standard deviation $\sigma = 10$.



Connecting Back to Cumulative Distributions

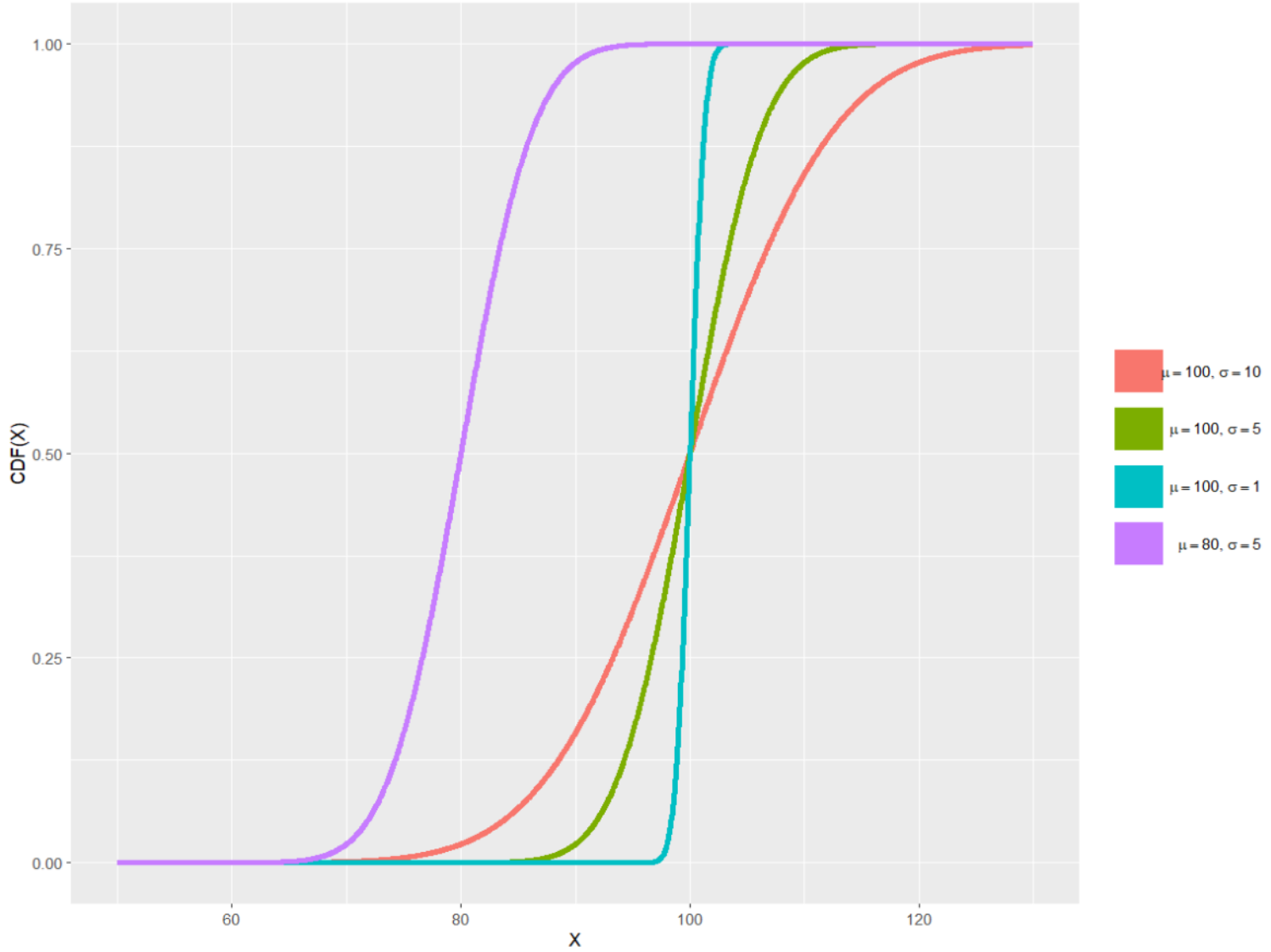
- Review:

  The **cumulative distribution** of a variable gives the proportion of observations that at less than or equal to a certain value

- The **cumulative distribution of random variable** X is a function which gives the probability that X is less than or equal to some value x

$$F_X(x) = P(X < x)$$

Cumulative probability is calculated from left to right

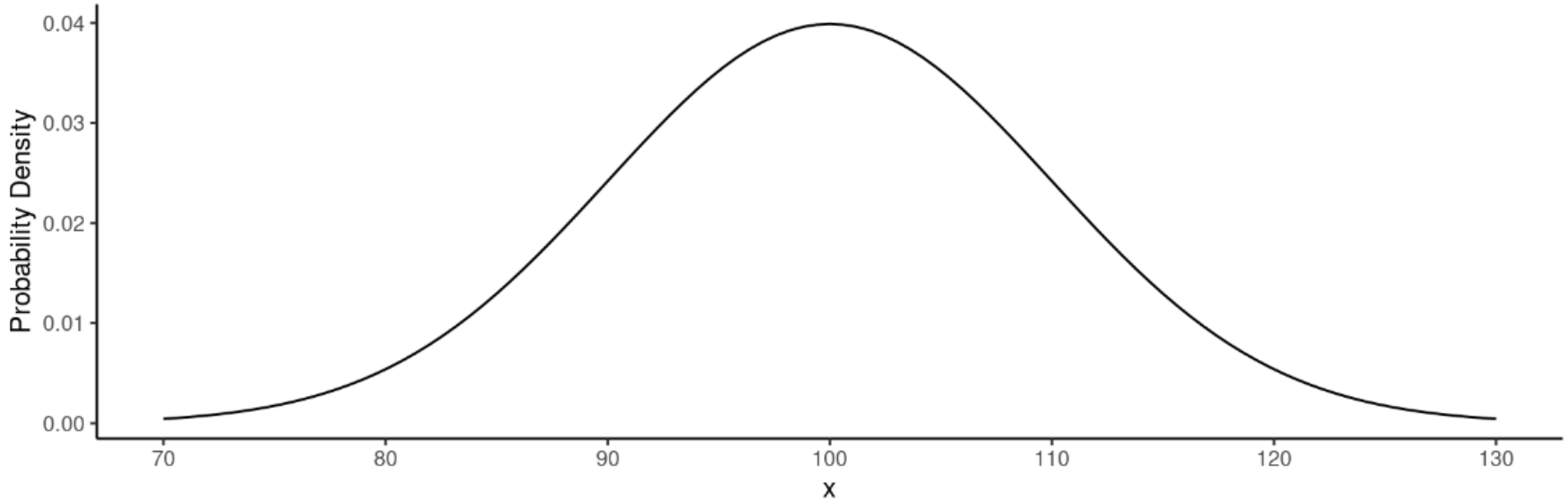


The Normal Distribution

- One important family of continuous probability distributions is the **normal distribution**.
- PDF normal

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

Example: Normal probability distribution with mean $\mu = 100$ and standard deviation $\sigma = 10$.



Tips For Finding Probabilities From Continuous Distributions

- Tips for finding probabilities from continuous distributions
- As we point out before, for a continuous distribution $P(X = x) = 0$

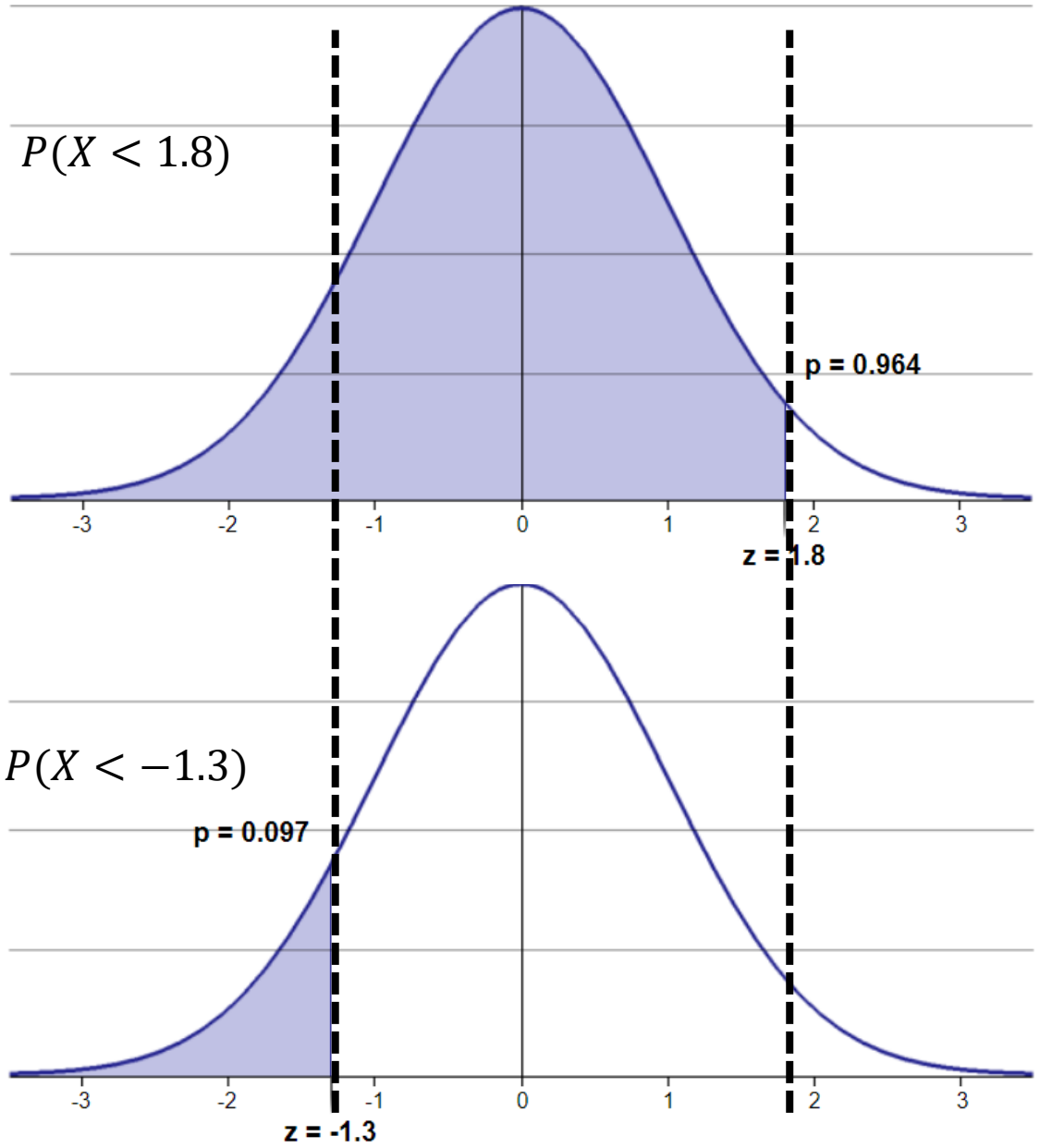
- So we typically deal with finding probabilities of X falling in some interval

  e.g $P(X < x)$, $P(X > x)$, or $P(a < X < b)$

Since most probability tables and software compute the probabilities using the cumulative distribution function we can use the following rules:

- $P(X > x) = 1 - P(X \leq x)$
- $P(a < X < b) = P(X < \mathrm{b}) - P(X < a)$

$P(-1.3 < X < 1.8)$?



Computing Probabilities From A Normal Distribution

- If we want to find the probability of a given value that we know follows a normal probability distribution we must first find its z-score

$$z = \frac{x - \mu}{\sigma} \sim N(0,1) \quad \text{if } X \sim N(\mu, \sigma)$$

- We can use a probability table for the standard normal distribution or use software such as http://www.statdistributions.com/normal/ or the app in the course website to compute the probabilities based on z-scores.

Examples:

Using the Z-table in the course website find the following probabilities. (you can use the app in the course website to check your answers)

$X \sim N(\mu = 100, \sigma = 5)$:

- $P(X < 90)$
- $P(X > 85)$
- $P(90 \leq X \leq 110)$

Examples:

Using the Z-table in the course website find the following probabilities. (you can use the app in the course website to check your answers)

$X \sim N(\mu = 20, \sigma = 10)$:

- $P(X \leq 5)$
- $P(X \geq 45)$
- $P(5 < X < 15)$

Three Types of Distributions:

- **Population Distribution** – the probability distribution of a single observation of a random variable – shows the possible outcomes of the single observation and their probabilities.
  - Its properties are described by unknown parameters such as p or μ, σ^2, σ

- **Data Distribution** – This is the distribution(s) of variable(s) in our sample based on the observations we sampled.
  - Its properties are described by statistics such as $\bar{x}$ or $\hat{p}, s^2, s$
  - The data distribution of a variable will converge to the population distribution as $n \to N$

- **Sampling Distribution** – This is the distribution(s) of statistic(s) computed from the observations in the sample. This distribution arises from repeatedly sampling from the same population and computing statistics from those samples. It tells us how close a given estimate is to the true population parameter it is estimating (sampling error).
  - Its properties are described by the properties of the population distribution and sample size n

Central Limit Theorem

- The central limit theorem gives us some nice guarantees about the shape of the distribution of a statistic

Definition: if $X_1, X_2, \ldots X_n$ are independent and identically distributed random variables (all have the same distribution) such that

$E[X_i] = \mu$ and $E[X_i - \mu]^2 = \sigma^2 < \infty$ (have finite variance)

Then,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

Where $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ and where $\xrightarrow{d}$ denotes convergence in distribution

(in layman's terms) the **central limit theorem** states that as the sample size increases the *shape* of a sampling distribution of $\bar{x}$ will "approach" that of a normal distribution

Sampling Distributions of $\bar{x}$ and $\hat{p}$

The sampling distribution of the mean

- The mean of $\bar{x}$ is: μ - the population mean
- The standard deviation of $\bar{x}$ is: $\sigma/\sqrt{n}$

The sampling distribution of the sample proportion

- The mean of $\hat{p}$ is: p – the population proportion
- The standard deviation of $\hat{p}$ is: $\sqrt{\frac{p(1-p)}{n}}$

California Gubernatorial Election

- Election polling is one of the few cases where we know p - the true proportion of voters (either voting for one candidate or another) - because all the votes are counted.

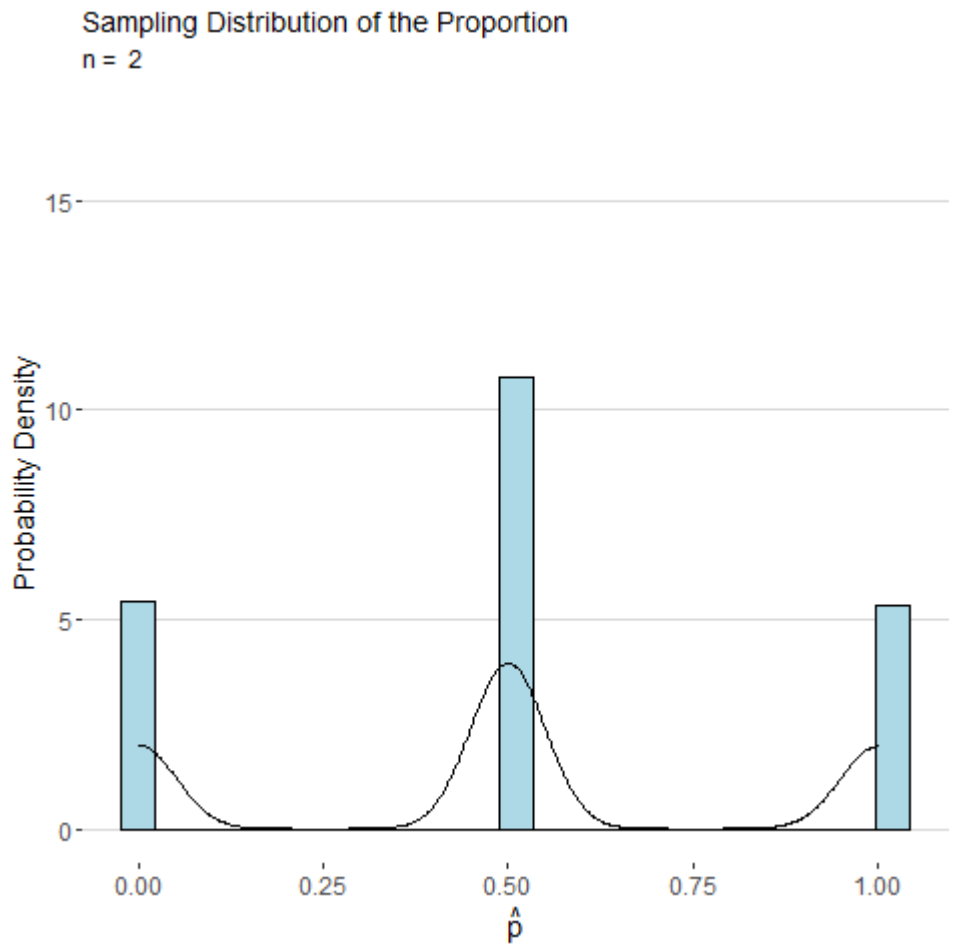
- From our example in week two about the California race for governor, the true population proportion of voters who cast a vote for Democrat Jerry Brown was 54.8% while the sample proportion measured from 3,889 voter interviews was 53.1%.
- What are the mean and standard deviation of $\hat{p}$?

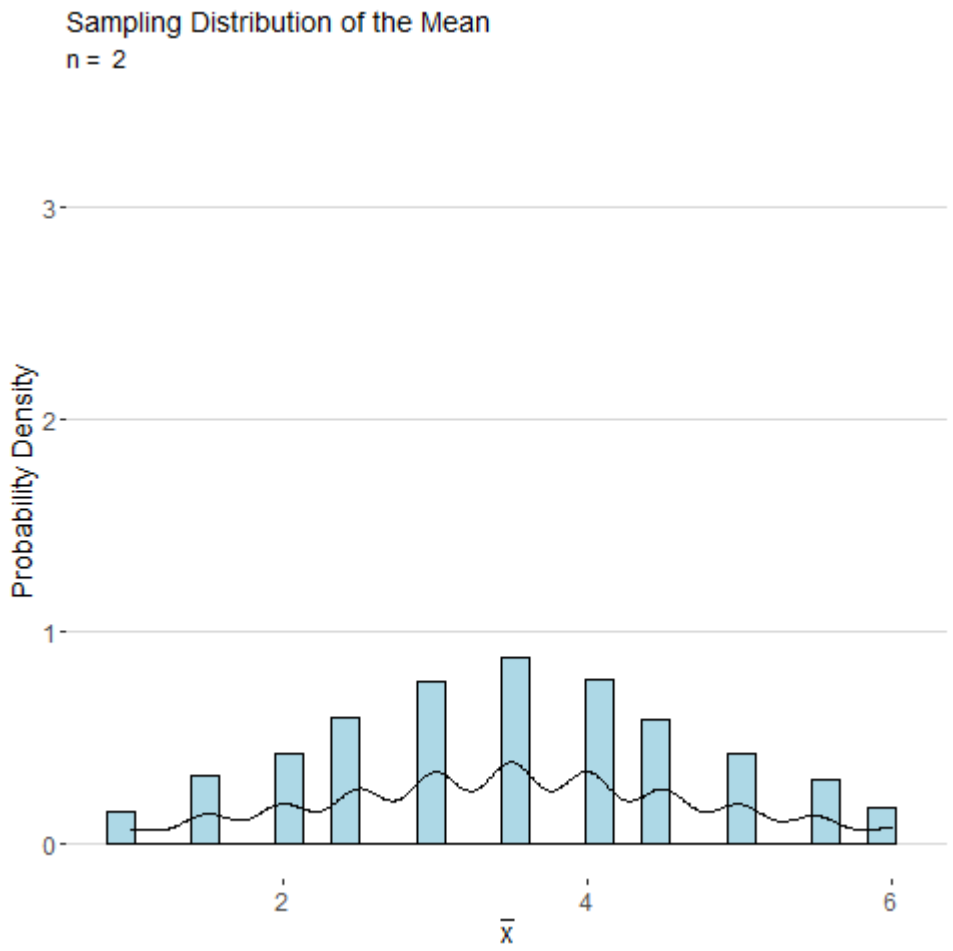
$$\text{mean of } \hat{p} = 0.548$$

$$\text{SD of } \hat{p} = \sqrt{\frac{0.531 \times (1 - 0.531)}{3889}} = \sqrt{6.4e^{-5}} = 0.008$$

- Why is the standard deviation so small?

Central Limit Theorem





Applying The CLT

- Recall that the **Empirical Rule** tells us how observations are distributed for approximately symmetric bell-shaped (normal) distributions.

- Since 95% of observations in a normal distribution fall within 2 standard deviations of the mean.

- Adapting the rule for probability distributions means that there is a 95% probability that random variable will fall within ∓ 2 standard deviations of the mean of the distribution.

Applying The CLT

The probability that $\bar{x}$ will be between $\mu - {}^{2\sigma}/_{\sqrt{n}}$ and $\mu + {}^{2\sigma}/_{\sqrt{n}}$ is approximately 0.95

- The probability that $\hat{p}$ will be between $p - 2\sqrt{{}^{p(1-p)}/_{n}}$ and $p + 2\sqrt{{}^{p(1-p)}/_{n}}$ is approximately 0.95

Estimation

Estimation is a type of statistical inference where we use our statistic to estimate a parameter

- We can use $\bar{x}$ (the means of a sample of n observations) to estimate the mean of single observation (i.e μ)
- We can use s (the standard deviation of the observations a sample of n observations) to estimate the standard deviation of a single observation (i.e σ)
- We can use $\hat{p}$ (the proportion of observations that are a "success" in a sample of n observations) to estimate the probability of success (i.e p)

Some Technical Points

Note that parameters μ, p, σ have a couple of interpretations.

- The first is that they are the properties of the population distribution.
- In a survey with a finite number of observations, these parameters are also properties of the set of all observations in the population
- Ex. μ is the mean for the probability distribution of a single observation but it is also the mean of the set of all observations in the population – we can interpret it either way

- We use the sampling distribution of our statistics to determine how effective they are at estimating the parameters of interest.
  - Both $\bar{x}$ and p are **unbiased** estimators
  - A **standard error** is the standard deviation of a statistic
  - The central limit theorem implies that (unless n is very small) the shape of the sampling distributions of $\bar{x}$ and p are approximately normal

The above properties rely on some technical assumptions about how the data are collected which we will talk about in a few lectures

Example

- Recall from the tiger trout example on Wednesday that the probability of catching a tiger trout in a single cast was 5%. Suppose a fisherman makes 450 casts in an afternoon and marks any time he catches a tiger trout as a success. Compute the interval for which the probability of $\hat{p}$ is approximately 0.95

- $n = 450$
- $p = 0.05$
- $P\left(p - 2\sqrt{{p(1-p)}/{n}} < \quad \hat{p} \quad < p + 2\sqrt{{p(1-p)}/{n}}\right) = 0.95$
- $SD = 0.01$

$$[0.05 \ - 2 \times 0.01, 0.05 + 2 \times 0.01]$$

$$P(0.03 < \hat{p} < 0.07) = 0.95$$