Sampling Distributions
Limit Theorems

Point estimators are RVs

An estimator (MLE or otherwise) is an RV and a statistic (a function of the sample)

The estimator for the Bernoulli parameter is the sample mean $\overline{X_n} = \frac{\sum_i X_i}{n}$

Let’s talk about the distribution of this RV

Desired properties of estimators

- Unbiased: $E\left[\widehat{\Theta}\right] = \theta$
    - Not always true for MLE estimator

- Consistent: $\Theta_n \xrightarrow{p} \theta$
    - True under mild conditions for MLE estimators

- Low Mean Squared Error
    - $E\left[\left(\widehat{\Theta} - \theta\right)^2\right] = \mathrm{Var}\left(\widehat{\Theta}\right) + Bias\left(\widehat{\Theta}\right)^2$ is low.

Today

The sample mean as an estimator for the population mean

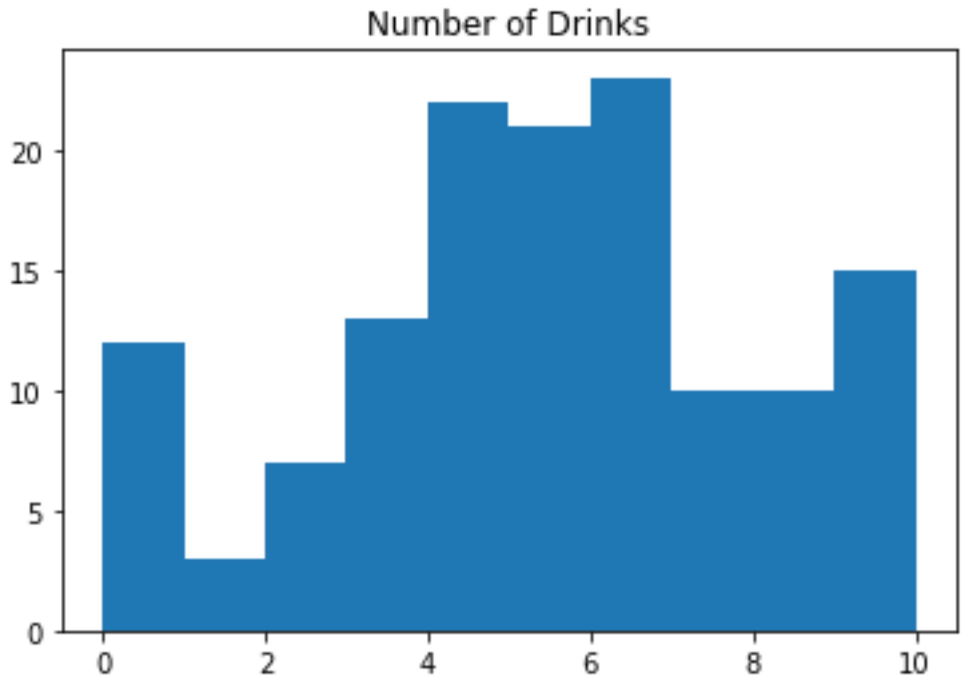
$$\overline{X_n} = \frac{\sum_i X_i}{n}$$

Let’s poll!

- We want to poll how many drinks University students drink every week.
- Go to https://polyhedron.math.uoc.gr/2223/moodle/mod/data/view.php?id=428
- Fill in the number of drinks you drink per week.

Practice

The following histogram shows the distribution of number of drinks a group of college students drink per week. We will assume that this is our population of interest. If we randomly select observations from this data set, which values are most likely to be selected, which are least likely?

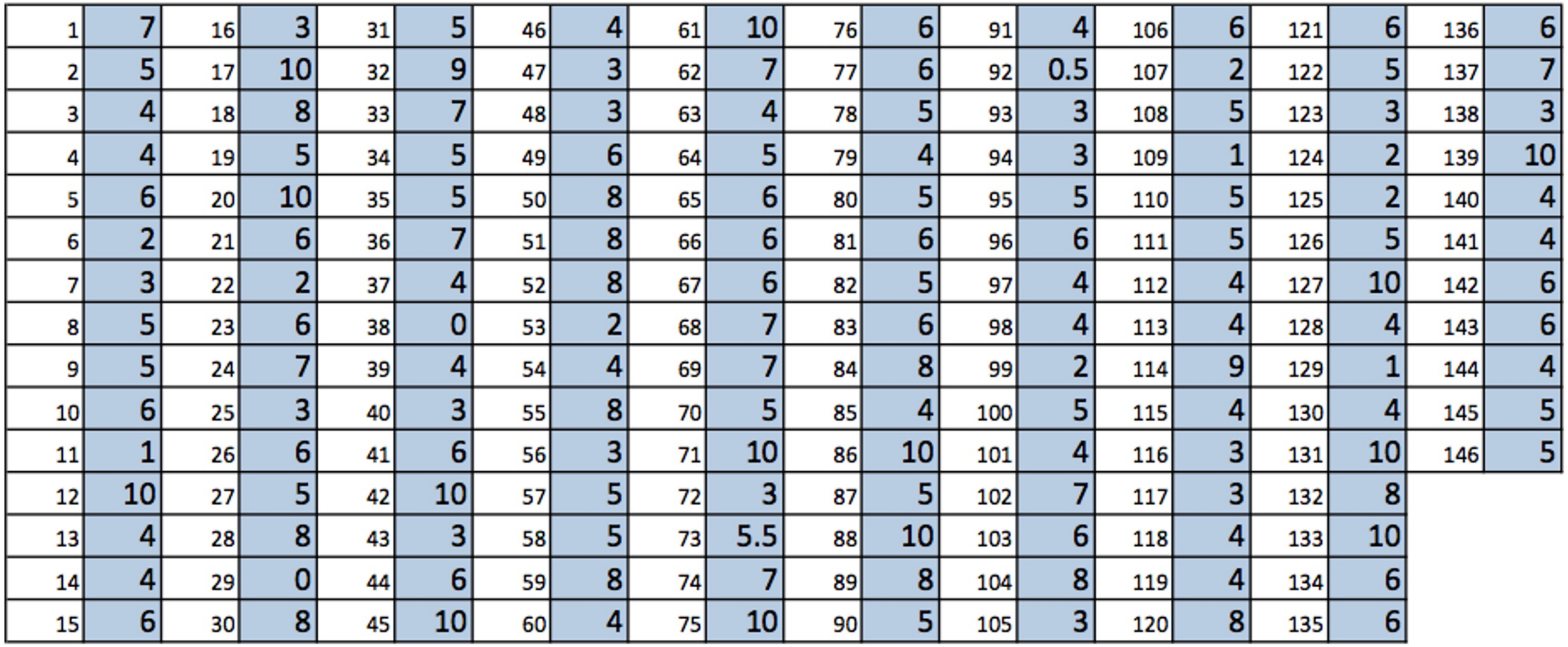


True mean of the distribution: 5.037

Standard deviation of the distribution: 2.7

Suppose that you don't have access to the population data. In order to estimate the average number of drinks college students drink per week, you might sample from the population and use your sample mean as the best guess for the unknown population mean.

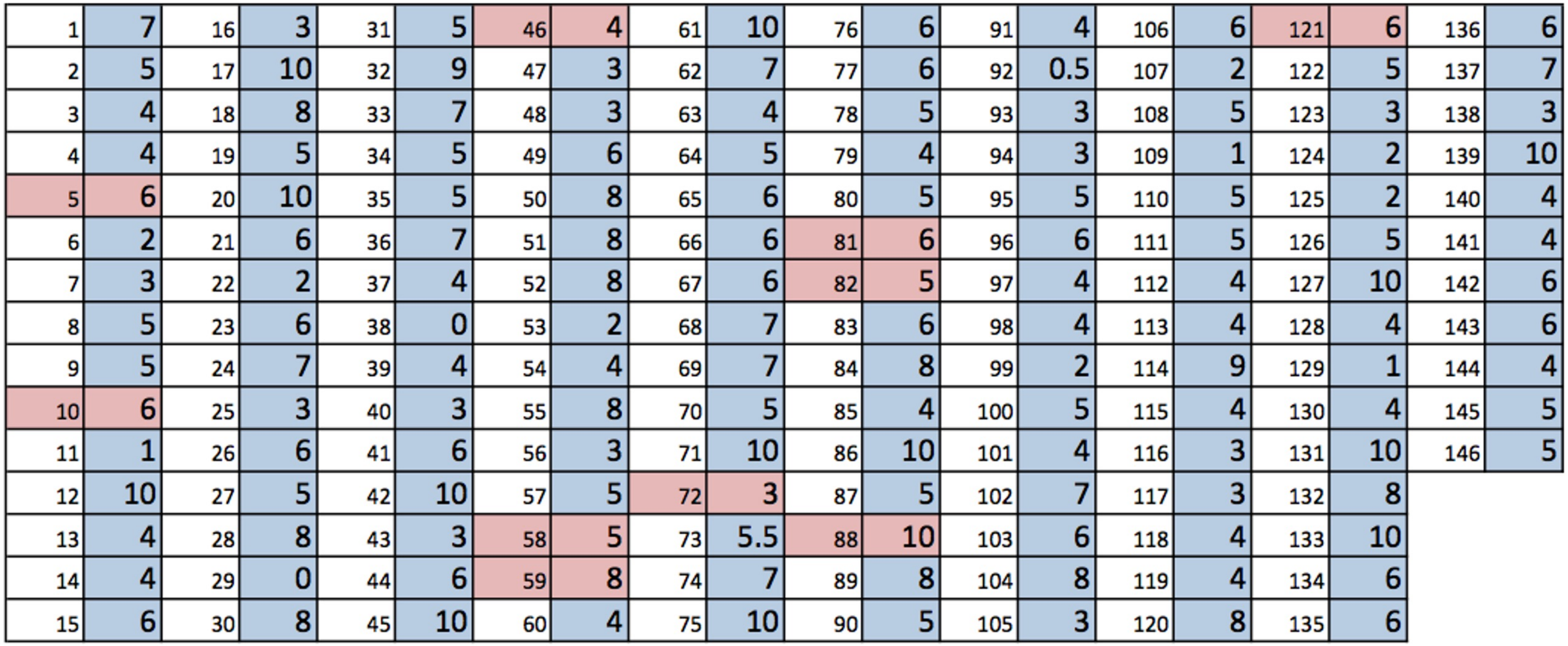
- Sample, with replacement, ten students from the population, and record the number of drinks they drink per week.
- Find the sample mean.
- Plot the distribution of the sample averages obtained by members of the class.



1	7	16	3	31	5	46	4	61	10	76	6	91	4	106	6	121	6	136	6
2	5	17	10	32	9	47	3	62	7	77	6	92	0.5	107	2	122	5	137	7
3	4	18	8	33	7	48	3	63	4	78	5	93	3	108	5	123	3	138	3
4	4	19	5	34	5	49	6	64	5	79	4	94	3	109	1	124	2	139	10
5	6	20	10	35	5	50	8	65	6	80	5	95	5	110	5	125	2	140	4
6	2	21	6	36	7	51	8	66	6	81	6	96	6	111	5	126	5	141	4
7	3	22	2	37	4	52	8	67	6	82	5	97	4	112	4	127	10	142	6
8	5	23	6	38	0	53	2	68	7	83	6	98	4	113	4	128	4	143	6
9	5	24	7	39	4	54	4	69	7	84	8	99	2	114	9	129	1	144	4
10	6	25	3	40	3	55	8	70	5	85	4	100	5	115	4	130	4	145	5
11	1	26	6	41	6	56	3	71	10	86	10	101	4	116	3	131	10	146	5
12	10	27	5	42	10	57	5	72	3	87	5	102	7	117	3	132	8		
13	4	28	8	43	3	58	5	73	5.5	88	10	103	6	118	4	133	10		
14	4	29	0	44	6	59	8	74	7	89	8	104	8	119	4	134	6		
15	6	30	8	45	10	60	4	75	10	90	5	105	3	120	8	135	6		

Example

List of random numbers: 59, 121, 88, 46, 58, 72, 82, 81, 5, 10

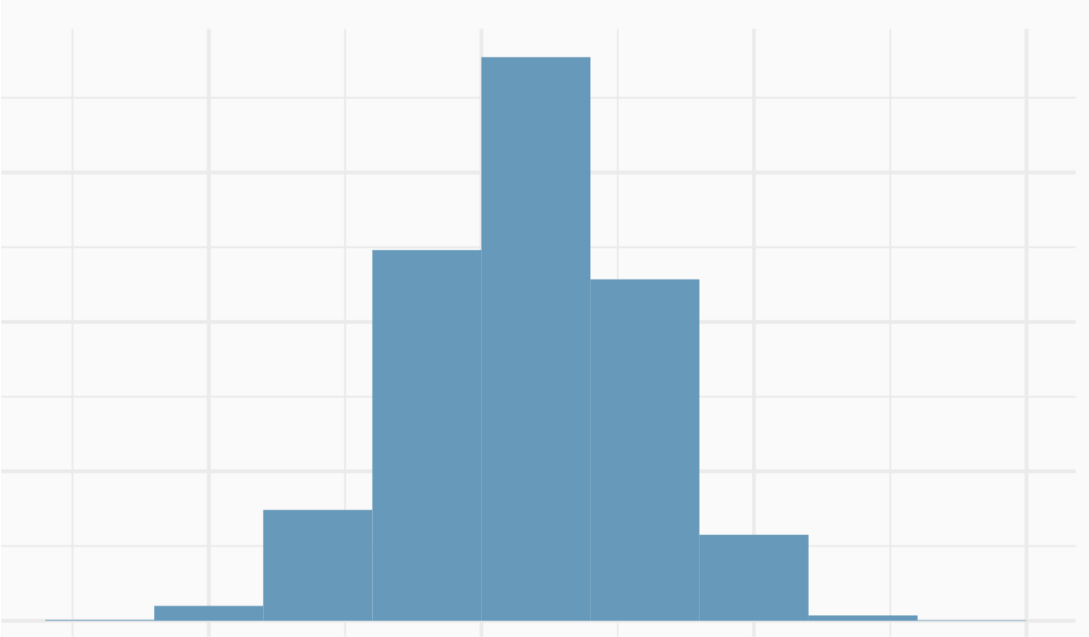


1	7	16	3	31	5	46	4	61	10	76	6	91	4	106	6	121	6	136	6
2	5	17	10	32	9	47	3	62	7	77	6	92	0.5	107	2	122	5	137	7
3	4	18	8	33	7	48	3	63	4	78	5	93	3	108	5	123	3	138	3
4	4	19	5	34	5	49	6	64	5	79	4	94	3	109	1	124	2	139	10
5	6	20	10	35	5	50	8	65	6	80	5	95	5	110	5	125	2	140	4
6	2	21	6	36	7	51	8	66	6	81	6	96	6	111	5	126	5	141	4
7	3	22	2	37	4	52	8	67	6	82	5	97	4	112	4	127	10	142	6
8	5	23	6	38	0	53	2	68	7	83	6	98	4	113	4	128	4	143	6
9	5	24	7	39	4	54	4	69	7	84	8	99	2	114	9	129	1	144	4
10	6	25	3	40	3	55	8	70	5	85	4	100	5	115	4	130	4	145	5
11	1	26	6	41	6	56	3	71	10	86	10	101	4	116	3	131	10	146	5
12	10	27	5	42	10	57	5	72	3	87	5	102	7	117	3	132	8		
13	4	28	8	43	3	58	5	73	5.5	88	10	103	6	118	4	133	10		
14	4	29	0	44	6	59	8	74	7	89	8	104	8	119	4	134	6		
15	6	30	8	45	10	60	4	75	10	90	5	105	3	120	8	135	6		

Sample mean: (8+6+10+4+5+3+5+6+6+6) / 10 = 5.9

Sampling distribution

Suppose you were to repeat this process many times and plot the results. What you just constructed is called a sampling distribution.



Sampling distribution

What you just constructed is called a *sampling distribution*.

Sampling distribution

What you just constructed is called a *sampling distribution*.

What is the shape and center of this distribution?

Sampling Distribution of the Sample mean

- Sample mean:
- $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$

- What is the expectation of the sample mean?

- What is the variance of the sample mean?

- What is the distribution of the sample mean?

Sampling Distribution of the Sample mean

- Sample mean:
- $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$

- What is the expectation of the sample mean?
  - μ
- What is the variance of the sample mean?
  - $\frac{\sigma^2}{n}$

- What is the distribution of the sample mean?

Practice

- Draw the distribution for the roll of one die.
- Draw the distribution for the mean of rolls of two dice.

Sequence of Random Variables

- A sequence of random variables is a sequence of functions.
- Example: Toss a fair coin once. Define

$$X_n = \begin{cases} \dfrac{1}{n+1}, & if\ heads \\ 1, & otherwise \end{cases}$$

a. Find the PMF and CDF of X_n for $n = 1, 2, 3, \ldots$

b. As $n \to \infty$, what does $F(X_n)$ look like?

c. Are they independent?

Sequence of Random Variables

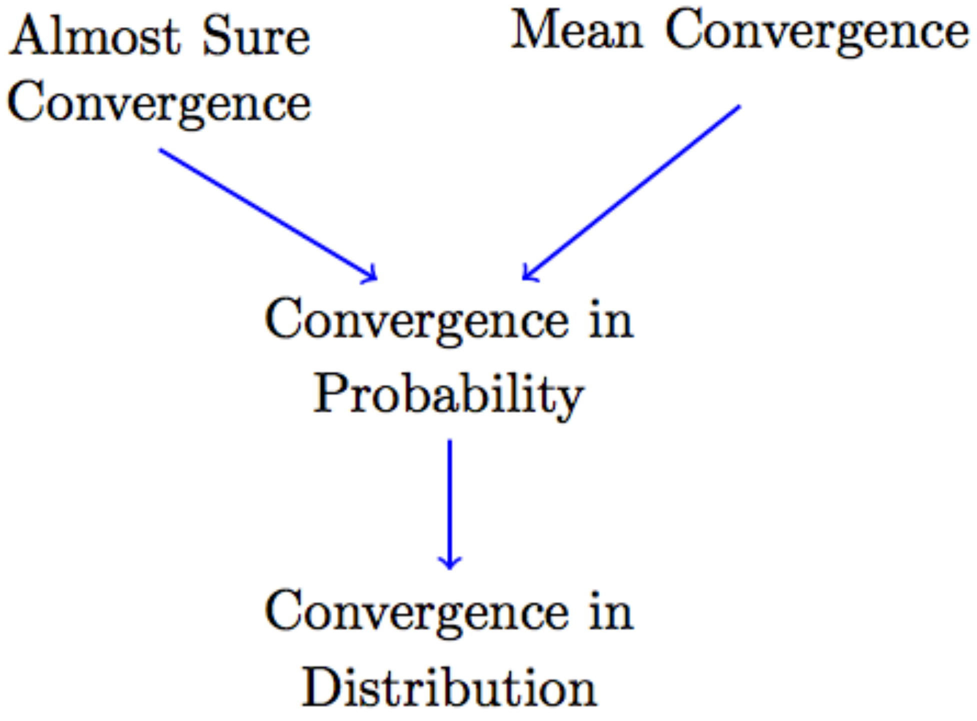
- A sequence of random variables is a sequence of functions.

- Example: Toss a fair coin forever. Define
- $X_n = \begin{cases} 1, & if\ n - th\ toss\ is\ heads \\ 0, & otherwise \end{cases}$

a. Find the PMF and CDF of X_n for $n = 1, 2, 3, \ldots$

b. Are they independent?

Convergence of RVs



Almost Sure Convergence

Mean Convergence

Convergence in Probability

Convergence in Distribution

Convergence In Probability

Arithmetic Convergence	Convergence in Probability
Sequence a_n of numbers converges to number l	Sequence X_n of random variables converges to random variable X
$\lim_{n\to\infty} a_n = l \text{ or } a_n \to l$	$X_n \overset{p}{\to} X$
a_n gets arbitrarily close to l	The probability distribution of $\lvert X_n - X\rvert$ gets more concentrated around 0.
$\forall\epsilon > 0, \exists n_o\colon \forall n > n_0 \quad \lvert a_n - l\rvert < \epsilon$	$\forall\epsilon > 0, \lim_{n\to\infty} P(\lvert X_n - X\rvert < \epsilon) = 1$

Convergence in distribution

- A sequence of random variables $X_1, X_2, X_3, \cdots$ converges in distribution to a random variable X, shown by $X_n \overset{d}{\rightarrow} X$, then

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x),$$

for all x at which $F_X(x)$ is continuous.

Weak Law of Large Numbers

- Draw n samples from a distribution with $E[X] = \mu$
- Let $\overline{\mathrm{X}}_{\mathrm{n}}$ be the sample mean
- The sample means for $\mathrm{n} = 1, 2, \ldots$, form a sequence of random variables.

- Weak law of large numbers:

$$\overline{\mathrm{X}}_{\mathrm{n}} \xrightarrow{p} \mu$$

Sampling Distribution of the Sample mean

- Sample mean:
- $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$

- What is the expectation of the sample mean?
  - μ
- What is the variance of the sample mean?
  - $\frac{\sigma^2}{n}$

- What is the distribution of the sample mean?
  - Increasingly concentrated around μ

Central Limit Theorem

- $S_n = \sum_{i=1}^{n} X_i$, mean $n\mu$, variance $n\sigma^2$.
- $\bar{X}_n = \frac{S_n}{n}$, mean μ variance $\frac{\sigma^2}{n}$.
- $\frac{S_n}{\sqrt{n}}$, mean $\mu\sqrt{n}$, variance σ^2
- $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$, mean 0 , variance 1 .

Central Limit Theorem

- Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with expected value $E[X_i\,] = \mu < \infty$ and variance $0 < Var(X_i) = \sigma^2 < \infty$.

- Then, the random variable $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$ converges in distribution to the standard normal random variable as n goes to infinity:

  - $\lim_{n \to \infty} P(Z_n \leq x) = \Phi(x)$, for all $x \in \mathbb{R}$

Sampling Distribution of the Sample mean

- Sample mean:
- $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$

- What is the expectation of the sample mean?
  - μ
- What is the variance of the sample mean?
  - $\frac{\sigma^2}{n}$

- What is the distribution of the sample mean?
  - Increasingly concentrated around μ
  - Normal as $n \to \infty$

CLT - conditions

Certain conditions must be met for the CLT to apply:

Independence

Sampled observations must be independent. This is difficult to verify, but is more likely if

- random sampling/assignment is used, and
- if sampling without replacement, n < 10% of the population.

Sample size

CLT is an asymptotic result, but actually holds for low sample sizes.

How low?

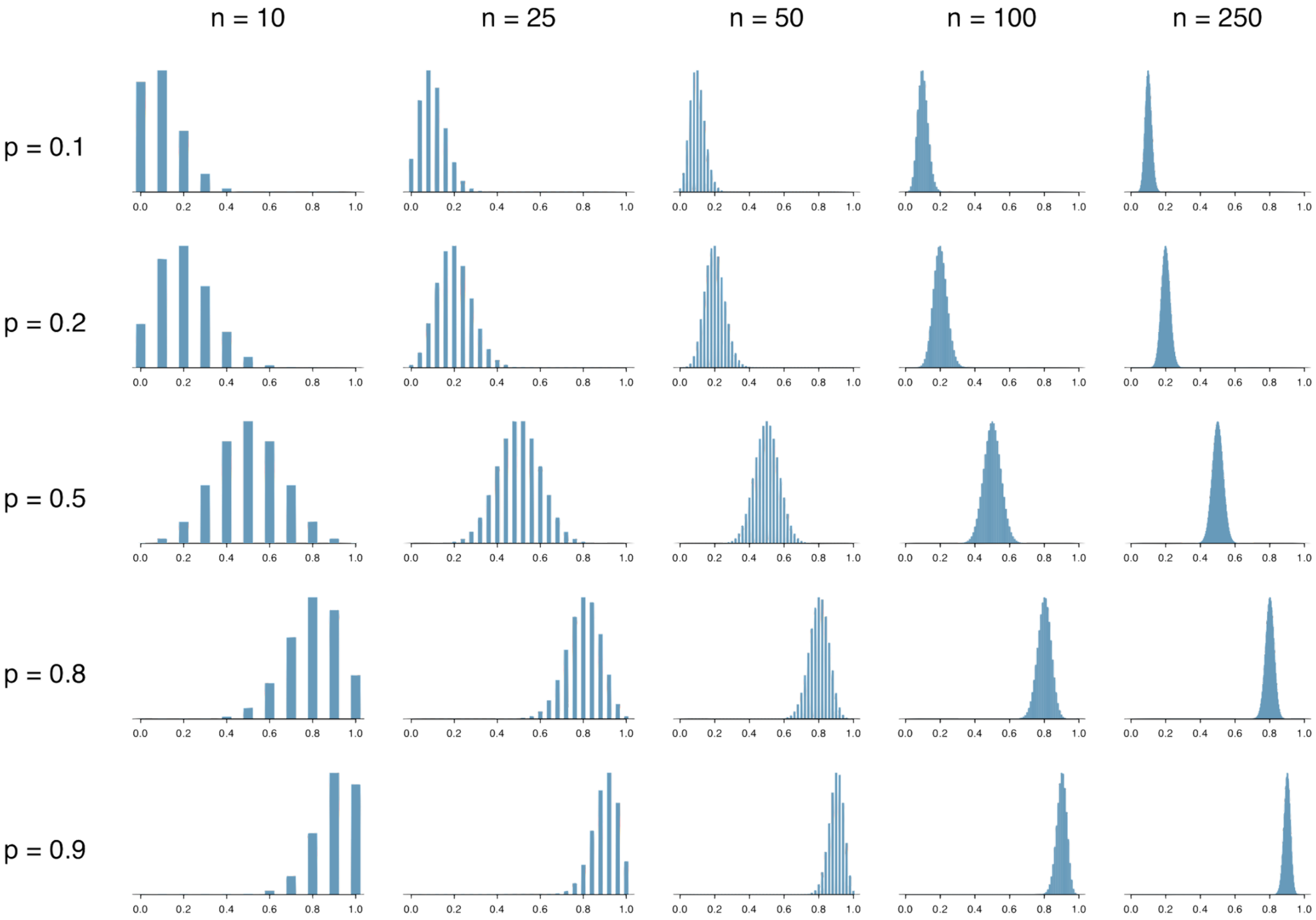
Special Case: Normal Distribution

- Sample mean:
- $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$

- What is the expectation of the sample mean?
  - μ
- What is the variance of the sample mean?
  - $\frac{\sigma^2}{n}$

- What is the distribution of the sample mean?
  - Increasingly concentrated around μ
  - Normal as $n \to \infty$

Sampling Distributions of Bernoulli Estimators



CLT - conditions

Certain conditions must be met for the CLT to apply:

Independence: Sampled observations must be independent. This is difficult to verify, but is more likely if

- random sampling / assignment is used, and
- if sampling without replacement, n < 10% of the population.

Sample size / skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large.

- the more skewed the population distribution, the larger sample size we need for the CLT to apply
- for moderately skewed distributions n > 30 is a widely used rule of thumb

.

CLT - Applications

The lengths of pregnancies are normally distributed with a mean of 268 days and a standard deviation of 15 days.

a. If one pregnant woman is randomly selected, find the probability that her length of pregnancy is less than 260 days.

CLT - Applications

The lengths of pregnancies are normally distributed with a mean of 268 days and a standard deviation of 15 days.

a. If one pregnant woman is randomly selected, find the probability that her length of pregnancy is less than 260 days.

b. If 25 pregnant women are randomly selected, find the probability that their lengths of pregnancy have a mean that is less than 260 days.

CLT - Applications

The lengths of pregnancies are normally distributed with a mean of 268 days and a standard deviation of 15 days.

a. If one pregnant woman is randomly selected, find the probability that her length of pregnancy is less than 260 days.

b. If 25 pregnant women are randomly selected, find the probability that their lengths of pregnancy have a mean that is less than 260 days.

c. If the 25 women do have a mean of less than 260 days, and you know that they were put on a special diet just before becoming pregnant, does it appear that the diet has an effect on the length of pregnancy, and should the medical supervisors be concerned?

CLT - Applications

The lengths of pregnancies are normally distributed with a mean of 268 days and a standard deviation of 15 days.

a. If one pregnant woman is randomly selected, find the probability that her length of pregnancy is less than 260 days.

b. If 25 pregnant women are put on a special diet just before they become pregnant, find the probability that their lengths of pregnancy have a mean that is less than 260 days (assuming that the diet has no effect).

c. If the 25 women do have a mean of less than 260 days, does it appear that the diet has an effect on the length of pregnancy, and should the medical supervisors be concerned?