intro to data science with probability & statistics

CSCI 3022

Oct 12, 2018

1. The Central Limit Theorem

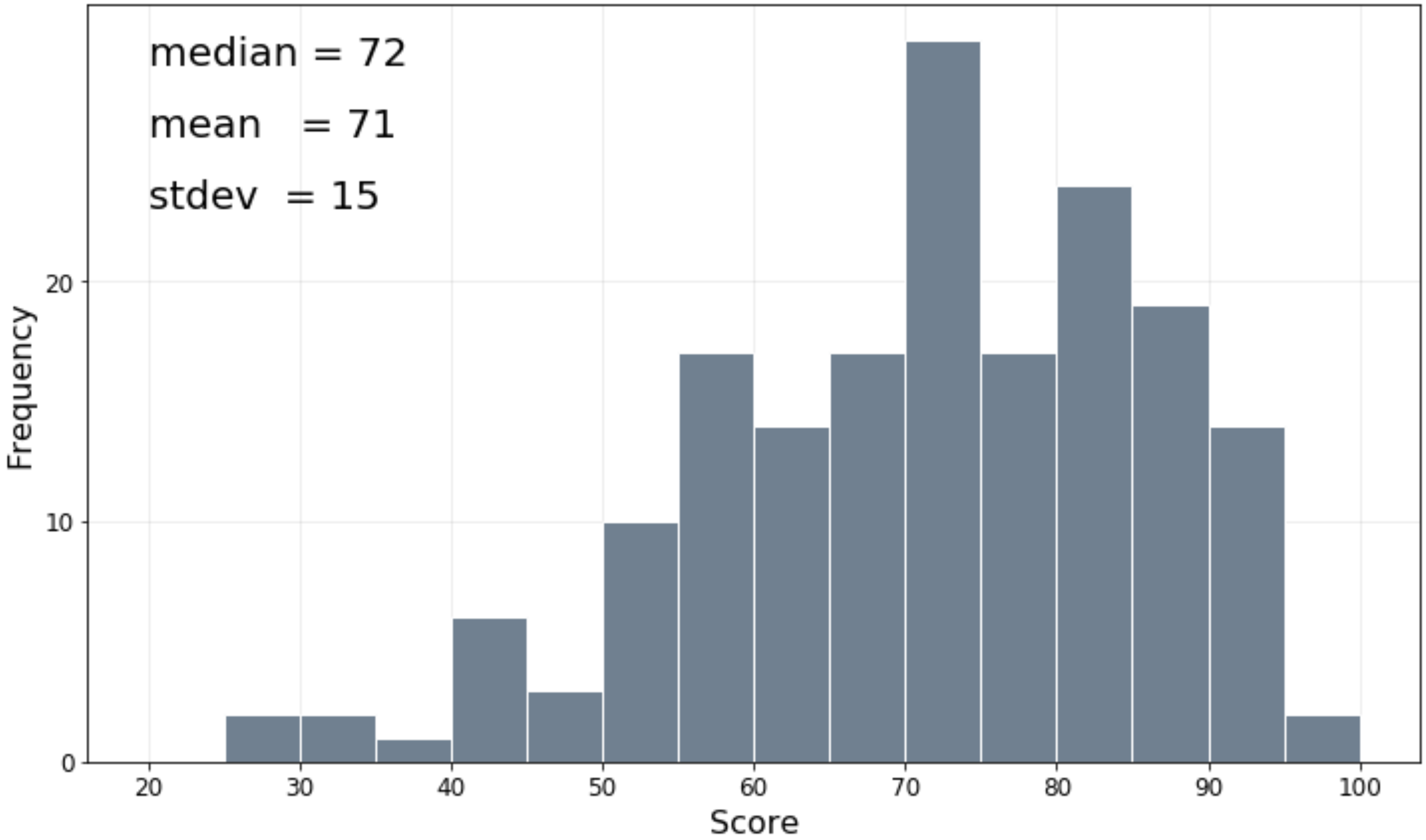
Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- **Office Hrs** as usual today: 4 to 5, Fleming 417.

Midterm Results



Last time on CSCI 3022

- **Def**: A continuous random variable has a normal (or Gaussian) distribution with parameters μ and σ^2 if its probability density function is given by the following. We say $\underline{X \sim N(\mu, \sigma^2)}$

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

- **Proposition**: If X is a normally distributed random variable with mean μ and standard deviation σ , then Z is a standard normal distribution if

Box-Muller

$$Z = \frac{X - \mu}{\sigma} \quad \text{or} \quad X = \sigma Z + \mu$$

- **Fact**: If Z is a standard normal random variable, then we can compute probabilities using the standard normal CDF

$$P(Z \leq z) = \int_{-\infty}^{z} f(x)dx = \underline{\Phi(z)}$$

Motivating example

- Soon, we'll be talking about *statistical inference* where we'll try to infer (learn) things about the true mean of a population using sample datasets

- **Examples**:
  - CU AERO mean GPA? // sample 30 students
  - Do all zebras have stripes? // sample 50 zebras
  - ...

Random samples

- The random variables $X_1, X_2, \ldots, X_n$ are said to form a [redacted] random sample of size *n* if:
  - all X_k $k = 1, 2, \ldots, n$ are independent
  - all X_k $k = 1, 2, \ldots, n$ have the same distribution (identical)
- We say that these X_k's are

I. I. D.

↑ indep

↑ identically

↖ distributed

Estimators and their distributions

- We use **estimators** to summarize our i.i.d. sample

- **Examples**:

1. $\bar{x}$ sample mean → use to estimate true mean μ

2. $\hat{p}$ sample proportion → use to estimate true proportion p

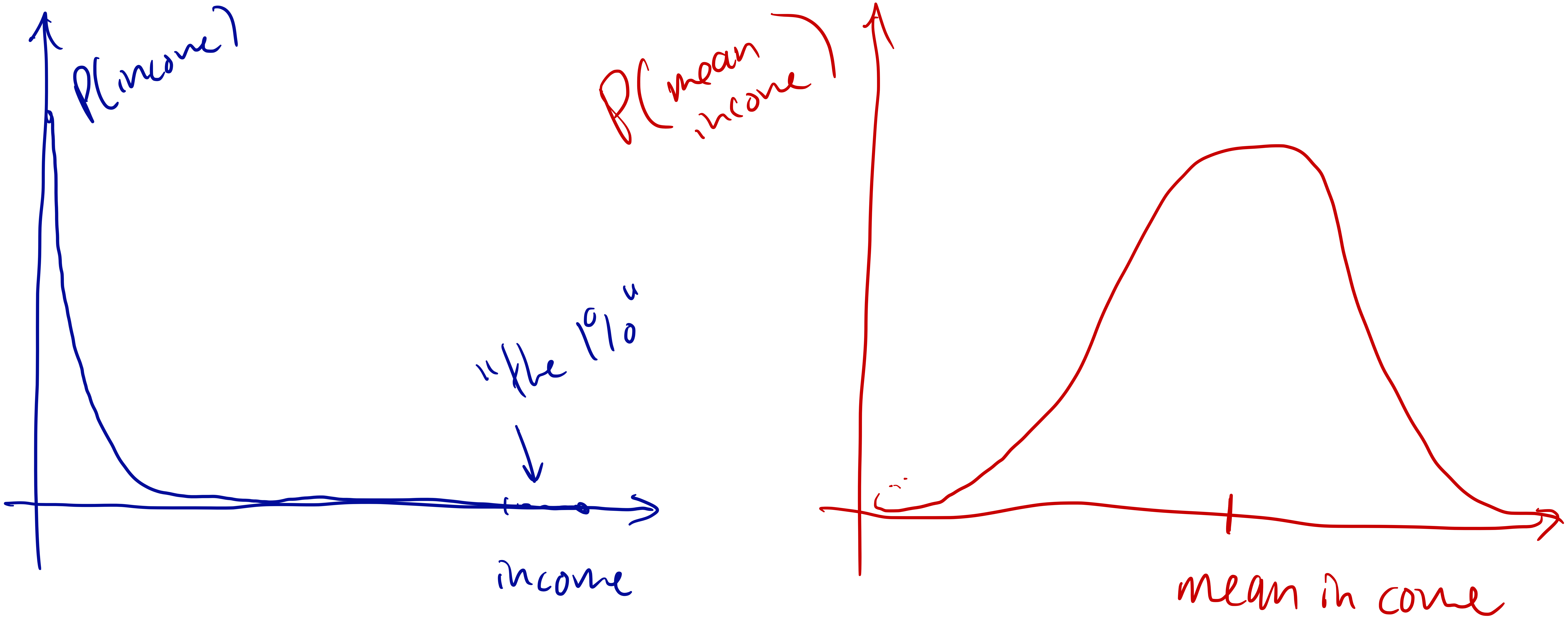
3. s^2 sample variance → use to estimate true variance σ^2

Estimators and their distributions

- We use **estimators** to summarize our i.i.d. sample
- Any estimator, including the **sample mean**, $\bar{X}$, is a random variable. Why? Because it's based on a random sample.
- This means that $\bar{X}$ has a distribution of its own, which is referred to as the **sampling distribution of the sample mean**.
- The sampling distribution depends on:
  - What is the underlying (true) distribution?
  - What is the number of samples, n?
  - Method of sampling?

Distribution of the Sample Mean

- What does the distribution of the sample mean actually look like?
- For example, does it look like the distribution that it's sampling? Not rly.



Distribution of the Sample Mean

- What does the distribution of the sample mean actually look like?
- **Proposition**: Let $X_1, X_2, \dots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. Then for any *n*,

variance is same as the i.i.d. samples variance, but divided by n.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

estimates of mean from samples

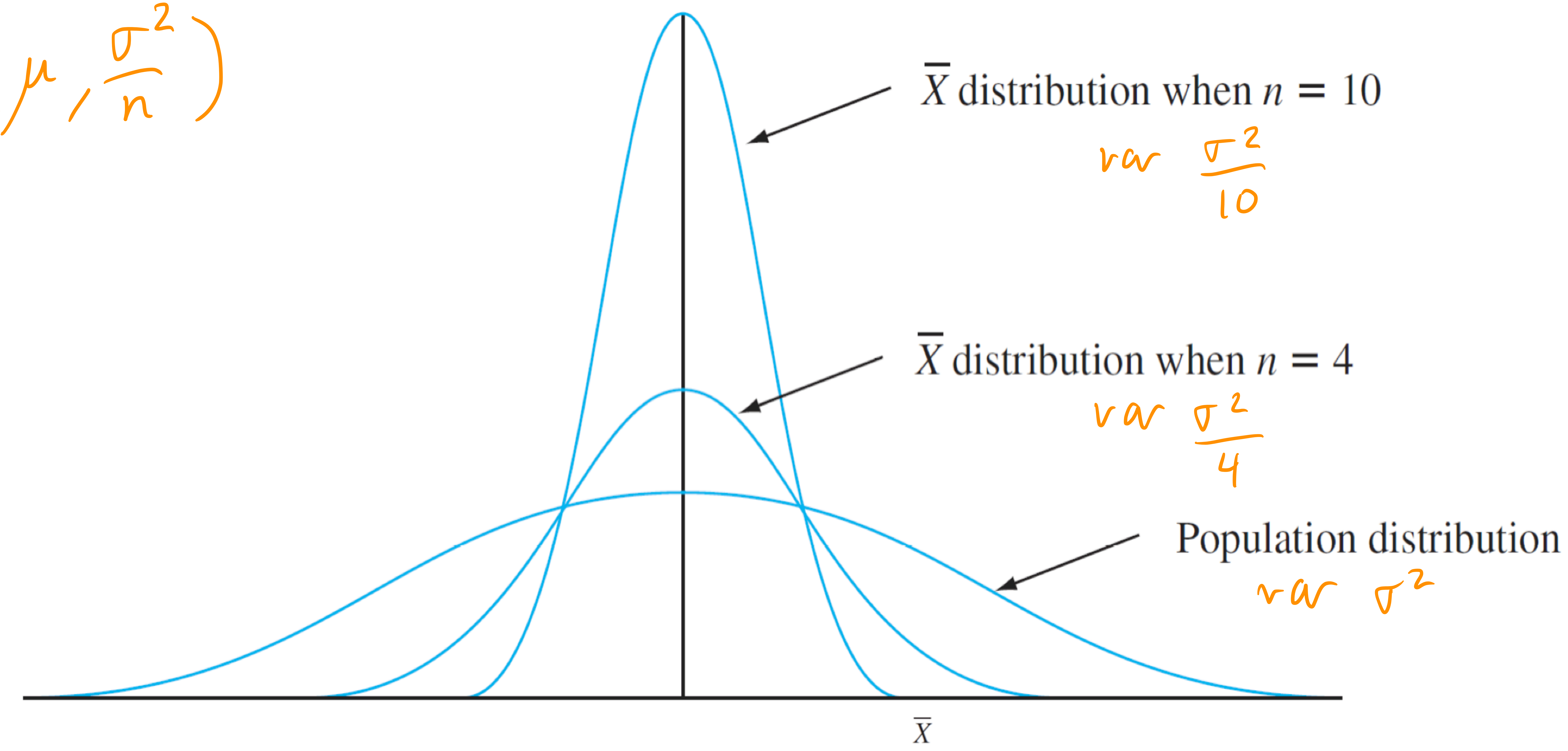
same mean as true, underlying R.V. (i.i.d samples)

- We know everything there is to know about the distribution of the sample mean when the population distribution is normal!

Distribution of the Sample Mean

- If the population is normally distributed, then:

$\tilde{X} \sim N(\mu, \frac{\sigma^2}{n})$



Distribution of the Sample Mean

- What if the population is *not* normally distributed?

The Central Limit Theorem

- What if the population is *not* normally distributed?
- **Important**: When the population distribution is non-normal, averaging produces a distribution more bell-shaped than the one being sampled.
- A reasonable assumption is that *if n is large*, a suitable normal curve will well-approximate the actual distribution of the sample mean.
- **The Central Limit Theorem**: Let $X_1, X_2, \ldots, X_n$ be i.i.d. draws from some distribution. Then as n becomes large

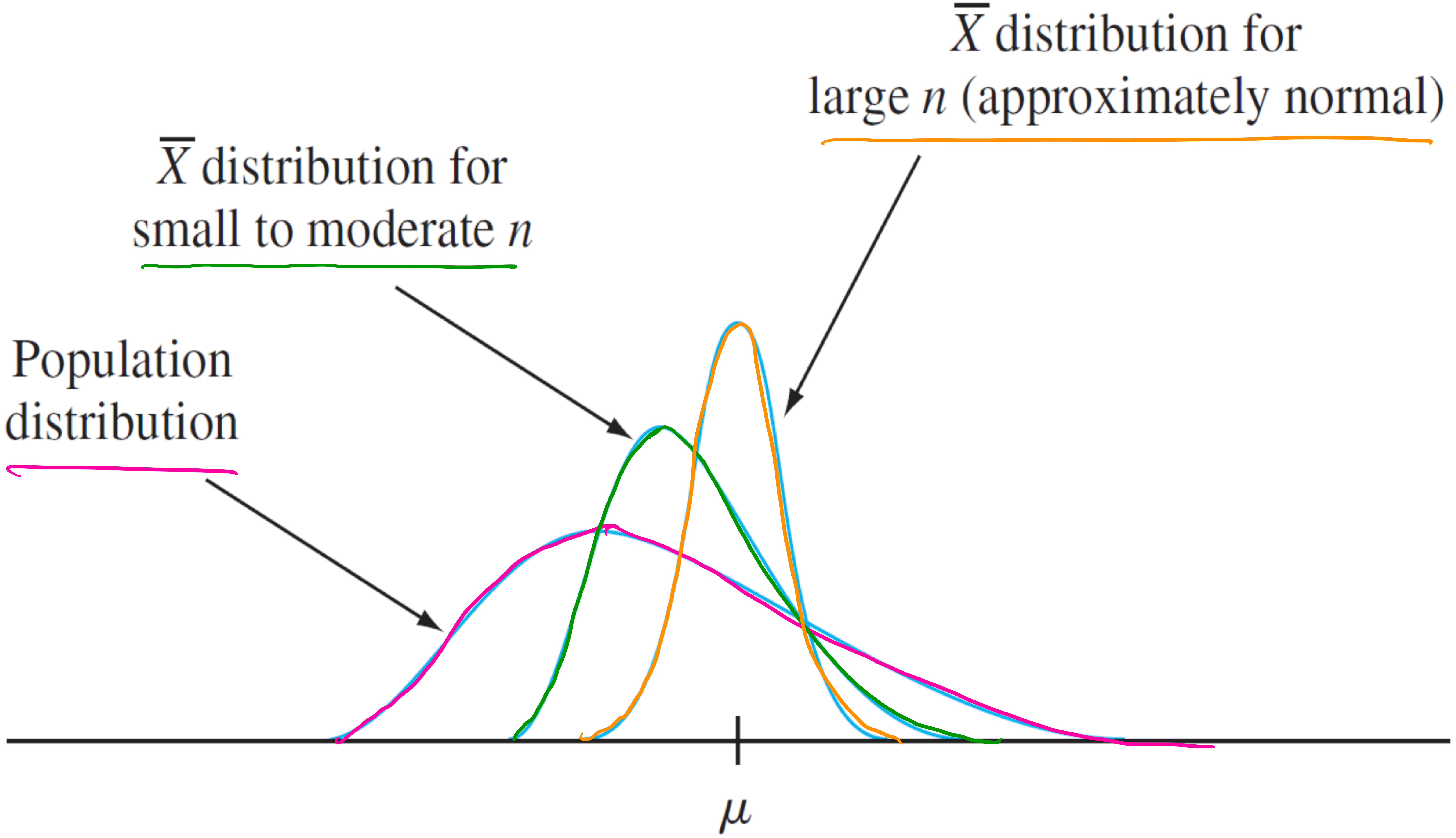
$$\bar{X} \sim N\left(\overset{\text{true mean}}{\mu}, \overset{\frac{\text{true variance}}{n}}{\frac{\sigma^2}{n}}\right)$$

so cool!

- Rule of Thumb: $n \geq 30$

Distribution of the sample mean:

- If the population is *not* normally distributed



$\overline{X}$ distribution for large n (approximately normal)
$\overline{X}$ distribution for small to moderate n
Population distribution
μ

Examples:

- **Example 1**: A hardware store receives a shipment of bolts that are supposed to be 12cm long. The mean is indeed 12cm, and the standard deviation is 0.2cm. For quality control, the hardware store chooses 100 bolts at random to measure. They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97cm or greater than 12.04cm. Find the probability that the shipment is found satisfactory.

question

① Information about pop: $\mu = 12$ $\sigma = 0.2$

Information about sample: $n = 100$

② How is $\bar{X}$ distributed?

$$\bar{X} \overset{\text{due to C.L.T.}}{\sim} N\left(12, \frac{0.2^2}{100}\right)$$

Examples:

- **Example 1**: A hardware store receives a shipment of bolts that are supposed to be 12cm long. The mean is indeed 12cm, and the standard deviation is 0.2cm. For quality control, the hardware store chooses 100 bolts at random to measure. They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97cm or greater than 12.04cm. Find the probability that the shipment is found satisfactory.

$$\bar{X} \sim N\left(12, \frac{0.2^2}{100}\right)$$

③

What is $P(11.97 \le \bar{X} \le 12.04)$

Box-Muller: $Z = \dfrac{\bar{X} - 12}{\sqrt{\frac{0.2^2}{100}}} = \dfrac{\bar{X} - 12}{0.02}$

What is $P\left(\dfrac{11.97-12}{0.02} \le Z \le \dfrac{12.04-12}{0.02}\right)$

$$= P(-1.5 \le Z \le 2)$$

$$= \Phi(2) - \Phi(-1.5)$$

$$\boxed{= 0.91}$$

Examples:

- **Example 2**: Suppose you have a jar of lemon and banana jelly beans where it is known that the true proportion of lemon jelly beans is 0.5. You try to estimate the proportion of lemon beans by reaching in and drawing 50 jelly beans and testing them (by eating them). What is the probability that your sample is 75% or more lemon jelly beans?
- Note: this is a little different because we're estimating a *proportion*. What changes?

Population: $p = 0.5$

Sample: $n = 50$

Q: $P(X \geq 0.75) = 1 - P(X < 0.75)$

Proportions are a little different.

$$\bar{X} = \hat{p} = \frac{Bin(n,p)}{n}.$$

↑ we call this estimator

What is the variance of $\hat{p}$?

$$var(\hat{p}) = var\left(\frac{Bin(n,p)}{n}\right) = \frac{1}{n^2} var(Bin(n,p))$$

$$= \frac{1}{n^2} np(1-p) = \boxed{\frac{p(1-p)}{n}}$$

Examples:

$$\frac{0.5(1-0.5)}{50} = \frac{\frac{1}{2}\cdot\frac{1}{2}}{50} = \frac{1}{200} = 0.005$$

- **Example 2**: Suppose you have a jar of lemon and banana jelly beans where it is known that the true proportion of lemon jelly beans is 0.5. You try to estimate the proportion of lemon beans by reaching in and drawing 50 jelly beans and testing them (by eating them). What is the probability that your sample is 75% or more lemon jelly beans?

Pop: $p = 0.5$

Sample: $n = 50$

$\mathrm{var}(\hat{p}) = \frac{p(1-p)}{n}$

$Q: P(\hat{p} \geq 0.75) = 1 - P(\hat{p} < 0.75)$

$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) = N\left(0.5, \frac{1}{200}\right)$

Box Muller $Z = \frac{\hat{p} - 0.5}{\sqrt{\frac{1}{200}}}$

$Q: 1 - P\left(Z < \frac{0.75 - 0.5}{\sqrt{\frac{1}{200}}}\right) = \ldots = 1 - \Phi(\text{something})$

Problem-solving hints:

- **First**, identify the *population* and identify the *sample*.
- **Second**, is the problem about *means* or *proportions*?
- **Then**, we're off to the races using the CLT, the Box-Muller transform, and our standard normal distribution!

E Z Mode