Class 3a: Review of concepts in Probability and Statistics

Business Forecasting

Roadmap

Last set of classes

- Types of data
- How to describe data
    - With visualizations
    - With summary statistics

This set of classes

- How to evaluate estimators
- How to build confidence intervals
- How to test hypotheses

Motivating Example

1. You run a bunch of Airbnbs
2. Should you invest more in cleaning?
3. Can you get higher price if your cleanliness score exceeds 4.5?
4. Get a sample of listings and compare the price of
    - Those with cleanliness score below 4.5 (dirty)
    - and above 4.5 (clean)

Show 5 entries

id	review_scores_cleanliness	price	clean
40032982	3	1023	Dirty
21962322	4.5	4500	Dirty
41841538	4.5	380	Dirty
624813934659858771	3.4	1350	Dirty
47030021	4.29	684	Dirty

Showing 1 to 5 of 200 entries

Previous 1 2 3 4 5 … 40 Next

Motivating example

In statistical language:

- Population: Entire group we want to learn about, impossible to assess directly
  - All listings of Airbnb in Mexico City
  - Ideally we would like to know the entire distribution of prices
- Parameters: Number describing a characteristic of the population
  - We want to know mean price of clean μ_c and dirty μ_d apartments
- Sample: Part of the population we have data for
  - We have a sample of 200 listings
- Goal: What we want to learn about the population?
  - Is $\mu_c > \mu_d$? If yes, by how much?
  - But we do not know μ_c and μ_d
  - We will try to guess it using an estimator and a random IID sample

What is a random sample?

- At random: A sample is random if each member of the population (each listing) has an equal chance of being selected. This process of selecting is called *drawing* from a population or a sample.
- Random Variable: P_i:
  - Random variable describing the observation i. Before drawing the sample, we don't know its value: it could be any price from the distribution.
- Random Sample is a collection of random variables $\{P_1, P_2, \ldots, P_n\}$
- Observed Value: p_i:
  - Once we observe a specific outcome for the random variable, it becomes a realized value, or p_i. It's no longer a random variable but a constant from our sample.

Before Drawing the Sample

Random Variables P_i (Before Drawing)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Selected Listings IDs								
Realized Values p_i (After Drawing)								

Random Samples — Draw 1

Each time we draw a sample, the random variables P_i get realized into concrete values p_i.

Sample 1

Random Variables P_i (Before Drawing)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Selected Listings IDs	8451	9015	8161	9085	8268	1622	1933	3947
Realized Values p_i (After Drawing)	120	150	800	200	1400	110	1800	900

Random Samples — Draw 2

A different random draw from the same population gives different realized values.

Sample 2

Random Variables P_i (Before Drawing)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Selected Listings IDs	3145	3773	6721	3373	2102	5365	4453	3621
Realized Values p_i (After Drawing)	260	420	500	2120	800	1450	120	809

Random Samples — Draw 3

Yet another draw — note how the point estimate changes every time.

Sample 3

Random Variables P_i (Before Drawing)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Selected Listings IDs	4971	2684	6331	3999	1995	4582	1478	1633
Realized Values p_i (After Drawing)	150	980	3450	220	120	853	2353	1244

What is a random sample?

- IID (Independent and Identically Distributed):
  - Independent: The selection of one unit (P_i) doesn't affect the selection of another (P_j)
  - Identically Distributed: All units P_i come from the same distribution.

Estimators

- Intuition
  - It's our method of guessing the parameter based on the data we have
  - A function of random variables in our sample $\hat{\theta} = f(P_1, P_2, \ldots, P_n)$
  - Given its random nature, we can analyze its statistical properties
  - Examples we have seen:
    - $\hat{\mu_c} = \bar{P} = f(P_1, P_2, \ldots, P_n) = \frac{\sum_n P_i}{n}$
    - $s_c = g(P_1, P_2, \ldots, P_n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(P_i - \bar{P})^2}$
  - It cannot contain any unknown quantities (like σ or μ_p)
- Point Estimate:
  - A single number computed from the realized sample data $\{p_1, p_2, \ldots p_n\}$
    - $\bar{p} = f(p_1, p_2, \ldots, p_n) = \frac{\sum_n p_i}{n}$
    - No longer random

Example: Estimator

- Suppose we want to know average price of the apartment in Mexico City, but we don't have data for the whole population.
- We take a sample of 8 listings and calculate the average price.

Before Drawing the Sample

Random Variables P_i (Before Drawing)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Selected Listings IDs								
Realized Values p_i (After Drawing)								

Estimator: $\hat{\mu} = \frac{P_1+P_2+P_3+P_4+P_5+P_6+P_7+P_8}{8}$

Example: Estimator

- Suppose we want to know average price of the apartment in Mexico City, but we don't have data for the whole population.
- We take a sample of 8 listings and calculate the average price.

After Drawing the Sample (Sample 1)

Random Variables P_i (Before Drawing)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Selected Listings IDs	8451	9015	8161	9085	8268	1622	1933	3947
Realized Values p_i (After Drawing)	120	150	800	200	1400	110	1800	900

Estimator: $\hat{\mu} = \frac{P_1+P_2+P_3+P_4+P_5+P_6+P_7+P_8}{8}$

Point estimate: $\frac{p_1+p_2+p_3+p_4+p_5+p_6+p_7+p_8}{8} = 685$

Example: Estimator

- Suppose we want to know average price of the apartment in Mexico City, but we don't have data for the whole population.
- We take a sample of 8 listings and calculate the average price.

After Drawing the Sample (Sample 2)

Random Variables P_i (Before Drawing)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Selected Listings IDs	3145	3773	6721	3373	2102	5365	4453	3621
Realized Values p_i (After Drawing)	260	420	500	2120	800	1450	120	809

Estimator: $\hat{\mu} = \frac{P_1+P_2+P_3+P_4+P_5+P_6+P_7+P_8}{8}$

Point estimate: $\frac{p_1+p_2+p_3+p_4+p_5+p_6+p_7+p_8}{8} = 809.875$

Example: Estimator

- Suppose we want to know average price of the apartment in Mexico City, but we don't have data for the whole population.
- We take a sample of 8 listings and calculate the average price.

After Drawing the Sample (Sample 3)

Random Variables P_i (Before Drawing)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Selected Listings IDs	4971	2684	6331	3999	1995	4582	1478	1633
Realized Values p_i (After Drawing)	150	980	3450	220	120	853	2353	1244

Estimator: $\hat{\mu} = \frac{P_1+P_2+P_3+P_4+P_5+P_6+P_7+P_8}{8}$

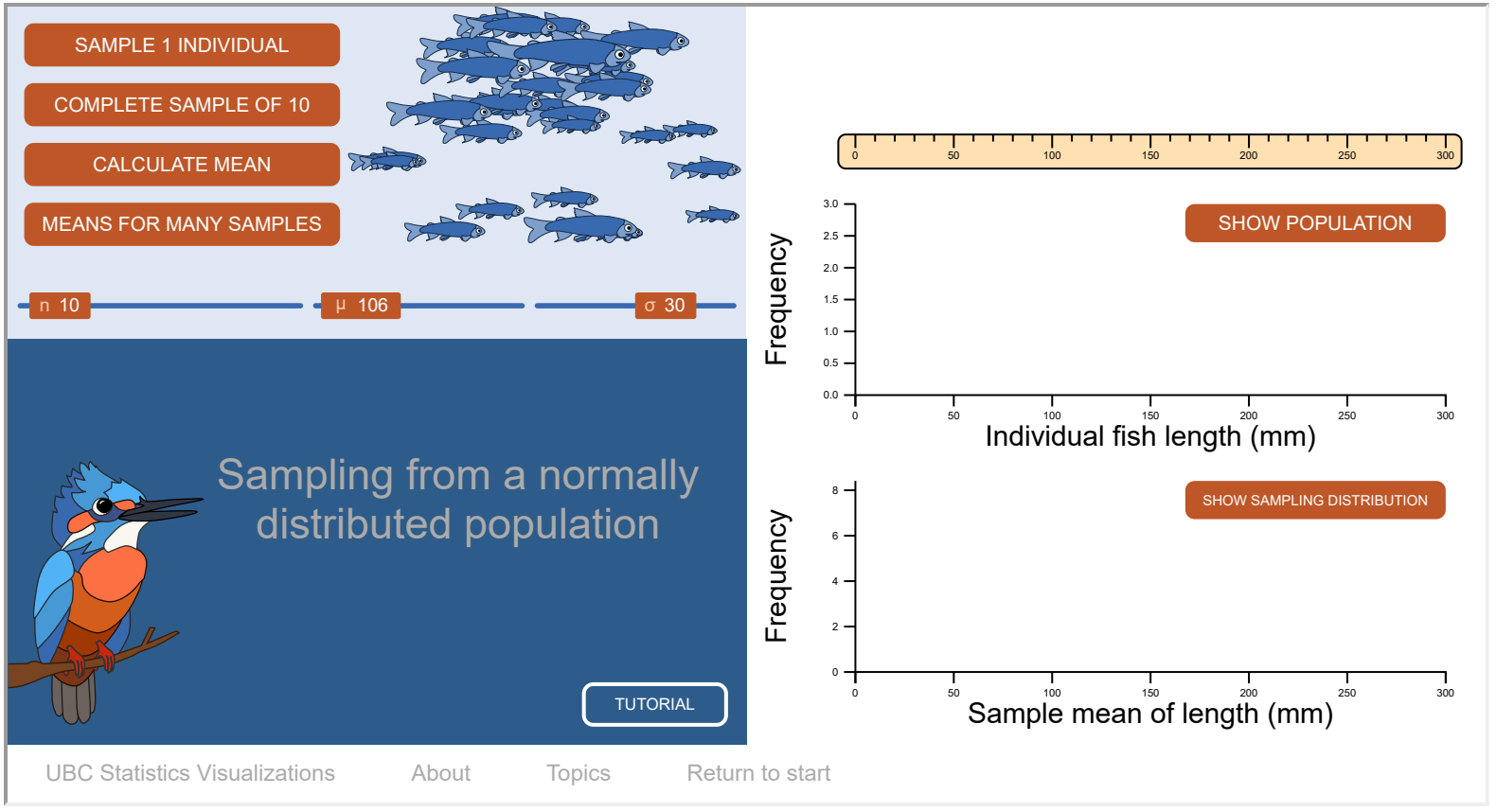
Point estimate: $\frac{p_1+p_2+p_3+p_4+p_5+p_6+p_7+p_8}{8} = 1171.25$

Estimators

- The mean price in our sample is $\bar{p}_c$ = NaN MXN
- This is our point estimate
- We can't really say how close this one number (point estimate) is to the true mean price in Mexico City without knowing the population
- But we can say how good our method of guessing (estimator) is by looking at its sampling distribution

Estimators

- Sampling distribution is the distribution of the estimator calculated from multiple random samples drawn from the same population.



LearnR: Panels A, B, C — Explore sampling distributions of SD, IQR, and 99th percentile

Expectation of an estimator

- A good estimator should be unbiased:

$$E[\hat{\theta}] = \theta$$

- Where θ is some parameter and $\hat{\theta}$ is its estimator
- This should be true for any value of θ
- The sampling distribution should be centered at the parameter's value
- Intuitively, on average the estimator should give us the parameter's value
- When I take many, many, many samples of apartments and calculate the mean price in each sample
    - The average of these means should be super close to the true mean price in Mexico City

$$Bias(\hat{\theta}) = E[\hat{\theta}] - \theta$$

- Bias of an estimator is a difference between its expectation and the parameter
- Let's look at a couple of estimators and check if they are biased or not

Example 1: Estimator = X_i (a single observation)

Expectation

- Consider some random variable X_i with unknown mean $E(X_i) = \mu$
- We want to estimate this mean
- The estimator: $\hat{\theta}_1 = X_i$
- Expected Value: $E(\hat{\theta}_1) = E(X_i) = \mu$
- Bias: $E(\hat{\theta}_1) - \mu = 0$ (unbiased)
- Is it a good estimator? (We'll come back to this)

> **LearnR:** Example 1 — Simulate the sampling distribution of a single observation

Example 2: Estimator = $(3X_1 + X_2)/5$

Expectation

- Consider some random variable X_i with unknown mean $E(X_i) = \mu$
- We want to estimate this mean
- The estimator: $\hat{\theta}_2 = \frac{3X_1 + X_2}{5}$
- Expected Value: $E(\hat{\theta}_2) = \frac{3}{5}E(X_1) + \frac{1}{5}E(X_2) = \frac{3}{5}\mu + \frac{1}{5}\mu = \frac{4}{5}\mu$
- Bias: $E(\hat{\theta}_2) - \mu = \frac{4}{5}\mu - \mu = -\frac{1}{5}\mu$ (biased)

LearnR: Example 2 — Simulate the biased estimator and observe the shift

Example 3: Estimator = $\frac{\sum X_i}{n}$ (sample mean)

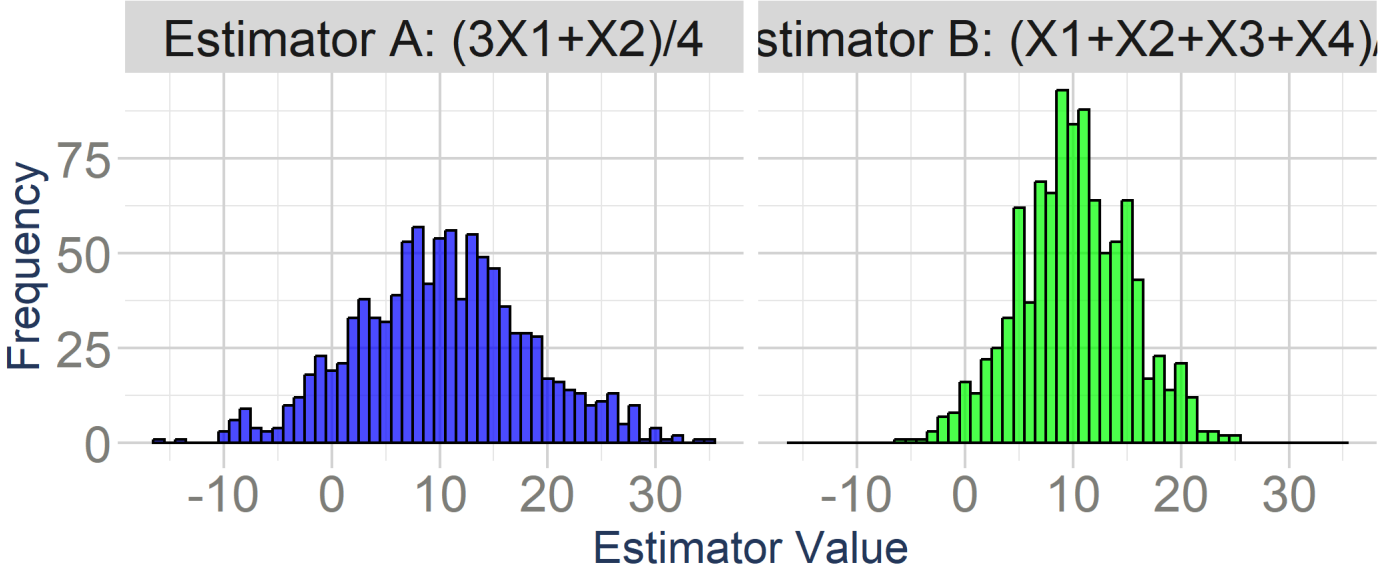
Expectation

- Consider some random variable X_i with unknown mean $E(X_i) = \mu$
- We want to estimate this mean
- The estimator: $\hat{\theta}_3 = \bar{X} = \frac{\sum_n X_i}{n}$
- Expected Value: $E(\hat{\theta}_3) = E\left(\frac{\sum_n X_i}{n}\right) = \frac{\sum_n E(X_i)}{n} = \frac{\sum_n \mu}{n} = \mu$
- Bias: $E(\hat{\theta}_3) - \mu = 0$ (unbiased)

> **LearnR:** Example 3 — Compare the sample mean's spread to a single observation

Variance of the estimator

- A good estimator is unbiased
- But how do we choose among unbiased estimators?
  - Suppose we sample IID from $X \sim \mathcal{N}(\mu = 10, \sigma = 10)$
  - Imagine you don't know the mean is 10, and you try to estimate it:
  - Estimator A: $\hat{\mu}_A = (3X_1 + X_2)/4$ (note: unbiased since weights sum to 1)
  - Estimator B: $\hat{\mu}_B = (X_1 + X_2 + X_3 + X_4)/4$
  - An estimator is more **efficient** if it has a smaller variance



Variance of the estimator

- Variance of an estimator is defined as:

$$Var(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

- We want the estimator to have low variance!
- Estimator with the lower variance is more efficient
- Efficiency is a property of unbiased estimators
- In the example above

$$Var(\hat{\mu}_A) = Var\left(\frac{3X_1 + X_2}{4}\right) > Var\left(\frac{X_1 + X_2 + X_3 + X_4}{4}\right) = Var(\hat{\mu}_B)$$

- Relative efficiency of two (unbiased) estimators is the ratio of their variances

$$Eff_{\hat{\mu}_A, \hat{\mu}_B} = \frac{Var(\hat{\mu}_A)}{Var(\hat{\mu}_B)} = \frac{10\sigma^2/16}{4\sigma^2/16} = \frac{10}{4} = \frac{5}{2}$$

Variance: Examples

Example 1: $\hat{\theta}_1 = X_i$

- $Var(\hat{\theta}_1) = \sigma^2$

  Example 3: $\hat{\theta}_3 = \bar{X} = \frac{\sum X_i}{n}$

- $Var(\hat{\theta}_3) = \frac{\sigma^2}{n}$

  Estimator A: $\hat{\mu}_A = \frac{3X_1 + X_2}{4}$

- $Var(\hat{\mu}_A) = \frac{9\sigma^2 + \sigma^2}{16} = \frac{10\sigma^2}{16}$

Side note: what if variables are not independent?

- In all previous cases of estimators we assumed an independent sample
- Suppose that X_1 and X_2 are not independent
- Example: daily sales of two products in the same store
- $E(X_1 + X_2) = E(X_1) + E(X_2)$ (expectation of a sum is always the sum of expectations)
- $Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)$
- $Var(X_1 - X_2) = Var(X_1) + Var(X_2) - 2Cov(X_1, X_2)$

Mean Squared Error

Mean Squared Error (MSE) is a summary measure of how good an estimator is:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

- The lower MSE, the better the estimator
- It summarizes both the bias and the variance:

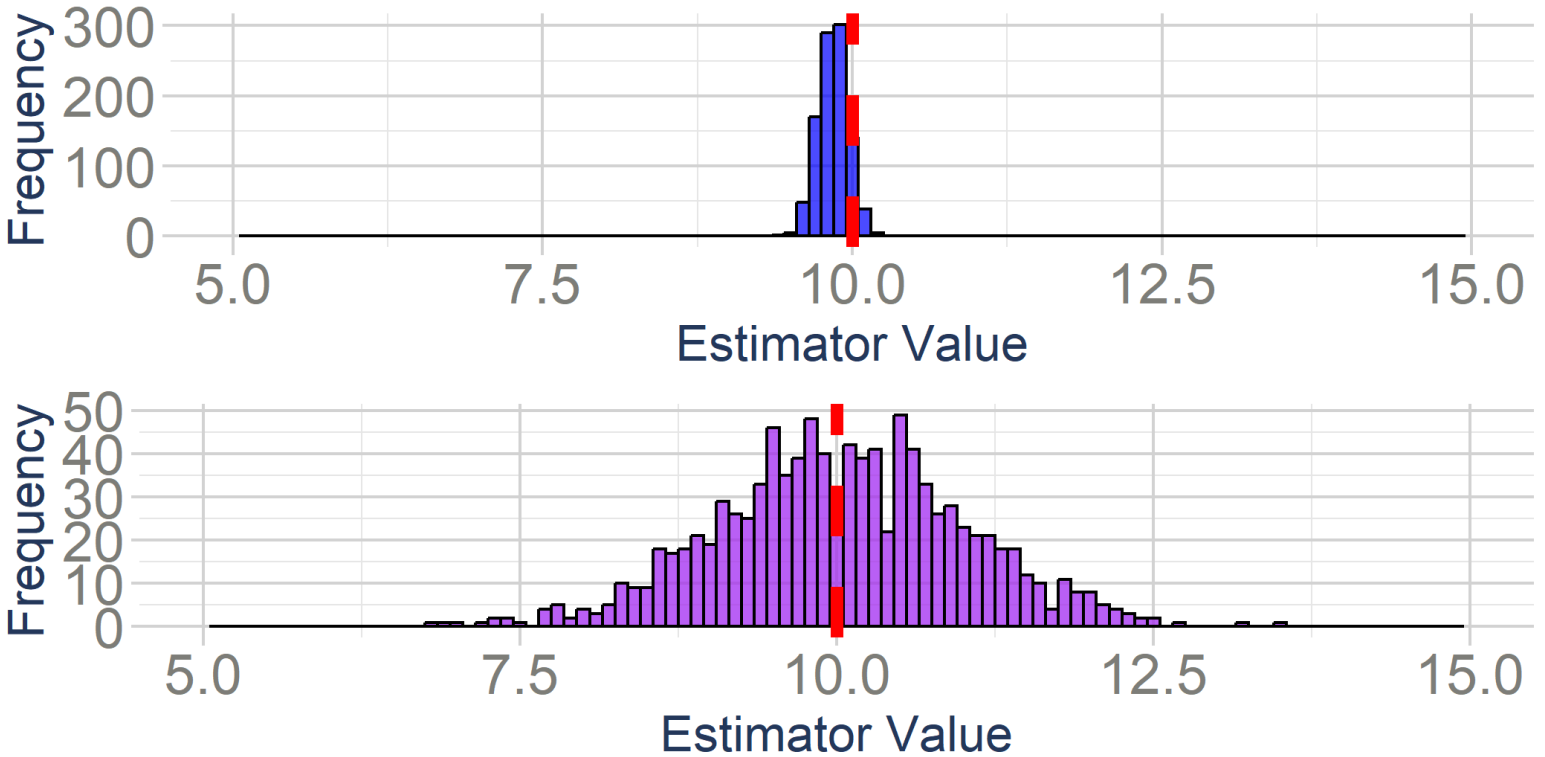
$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2] + 2(\underbrace{E[\hat{\theta} - E(\hat{\theta})]}_{=0})(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2 \\ &= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2 \end{aligned}$$

- If estimator is unbiased, then

$$MSE(\hat{\theta}) = \mathrm{var}(\hat{\theta})$$

Trading Bias for Variance

- Suppose you want to estimate customer's income to know who to target.
- Red line shows the true value
- Which of the estimators would you prefer?



Sampling Distribution

- We know how to determine the mean and the variance of the estimator
- Can we say anything about the distribution of the estimator?
- In case of sample mean, yes!
- That's what Central Limit Theorem is about, the most exciting theorem in statistics!

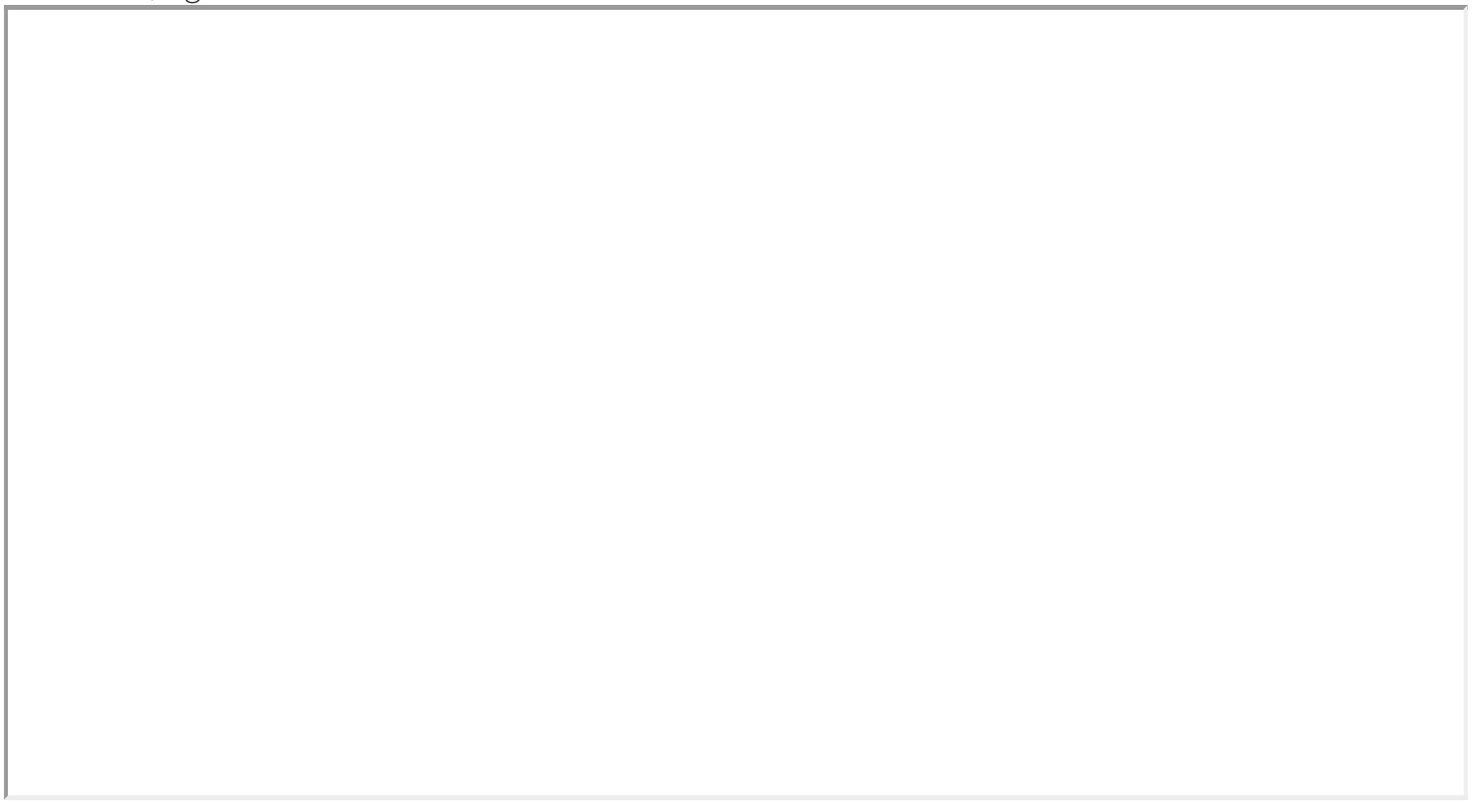
Central Limit Theorem

- Suppose $X_1, X_2, \ldots, X_n$ are i.i.d variables drawn at random from a distribution with mean μ and standard deviation σ
- Let $S_n = \sum_n X_n$.
  - Note that: $E[S_n] = n\mu$ and $st.\,dev.\,(S_n) = \sqrt{n}\sigma$
- Let $\bar{X}_n = \frac{\sum_n X_n}{n}$
  - Note that: $E[\bar{X}_n] = \mu$ and $st.\,dev.\,(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$
- Let $Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$
  - Note that: $E[Z_n] = 0$ and $st.\,dev.\,(Z_n) = 1$
- Central Limit Theorem says that for large n:

$$S_n \dot{\sim} \mathcal{N}\left(n\mu,\, \sqrt{n}\,\sigma\right), \quad \bar{X}_n \dot{\sim} \mathcal{N}\left(\mu,\, \frac{\sigma}{\sqrt{n}}\right), \quad Z_n \dot{\sim} \mathcal{N}(0,1)$$

- In large samples, $\bar{X}_n$ is approximately normal with mean μ and st.dev. $\frac{\sigma}{\sqrt{n}}$. As n grows, the approximation improves

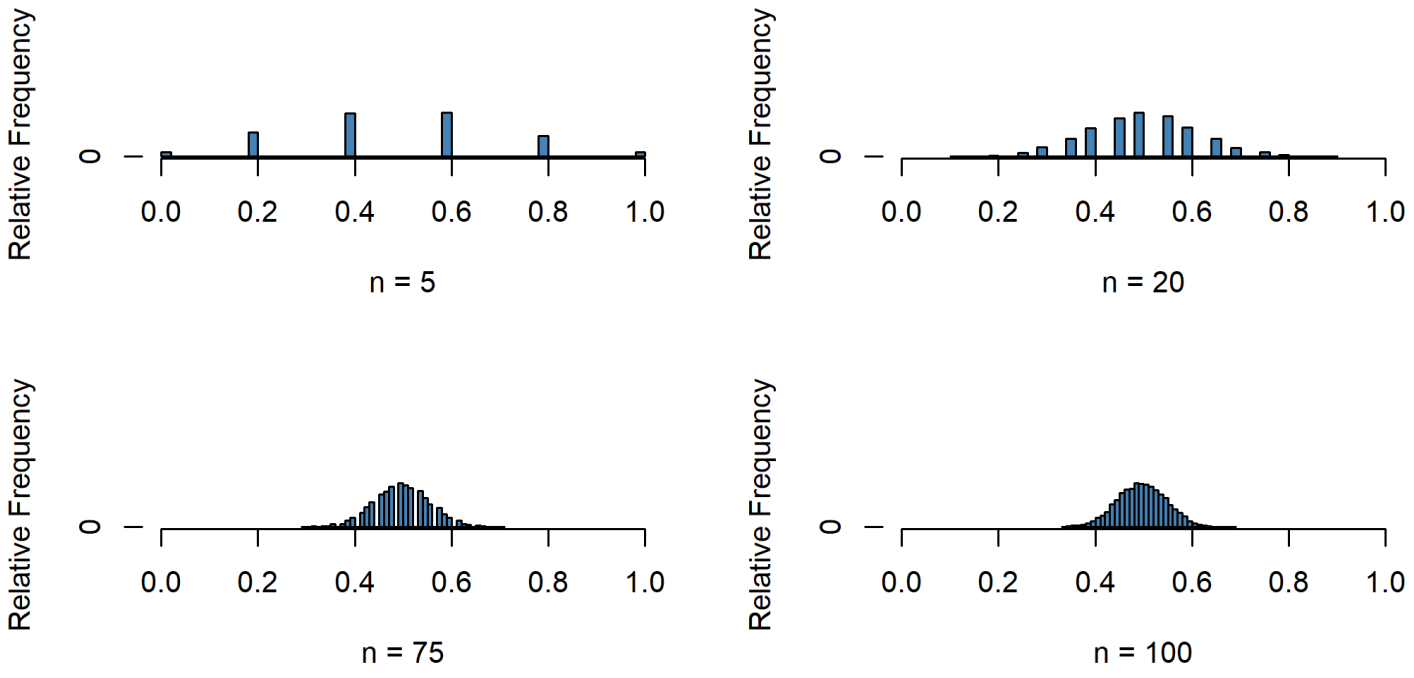
- The original distribution of X_i does not matter (but outliers make convergence longer)
- Larger n, tighter distribution around the mean
- Smaller σ , tighter distribution around the mean



Source: Seeing Theory

CLT also works for discrete variables!

- Let $X_i \sim \text{Bernoulli}(p = 0.5)$. Distribution of $\bar{X}_n$ for different sample sizes:



- What is the standard deviation?
- $\sigma_{\bar{X}} = \sqrt{var(\bar{X}_n)} = \frac{\sigma_X}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{0.5}{\sqrt{n}}$

Normal Distribution

Consider the event that a customer who opened the DiDi app will call the car. Suppose X and Y represent the events that a customer calls a car in Cancun (X) and Puerto Vallarta (Y) respectively.

- X and Y are Bernoulli variables with probabilities 0.4 and 0.6 respectively
- Suppose you have a random (iid) sample of 100 customers opening the app from Cancun and 80 from Puerto Vallarta.
- What is the probability that more than 100 people will call the car?

> **LearnR:** DiDi Simulator — Work through the problem analytically, then verify with simulation

Reminders

$$\text{If } X \sim \mathcal{N}(\mu, \sigma) \text{ and } c \text{ is a constant, then } X + c \sim \mathcal{N}(\mu + c, \sigma)$$

$$\text{If } X \sim \mathcal{N}(\mu, \sigma) \text{ and } c \text{ is a constant, then } cX \sim \mathcal{N}(c\mu, |c|\sigma)$$

$$\text{If } X \sim \mathcal{N}(\mu_1, \sigma_1) \text{ and } Y \sim \mathcal{N}(\mu_2, \sigma_2), \text{ then } X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$$

What if I don't know σ

- Suppose that sales in stores are normally distributed with mean 200 and with unknown variance
- I want to take a sample of 80 stores and I want to know the probability that the average sales in a sample will be greater than 220

$$P(\frac{\sum_{i=1}^{80} X_i}{80} > 220)$$

Ok, I know that according to central limit theorem

$$\frac{\sum_{i=1}^{80} X_i}{80} \sim N(200, \frac{\sigma}{\sqrt{80}})$$

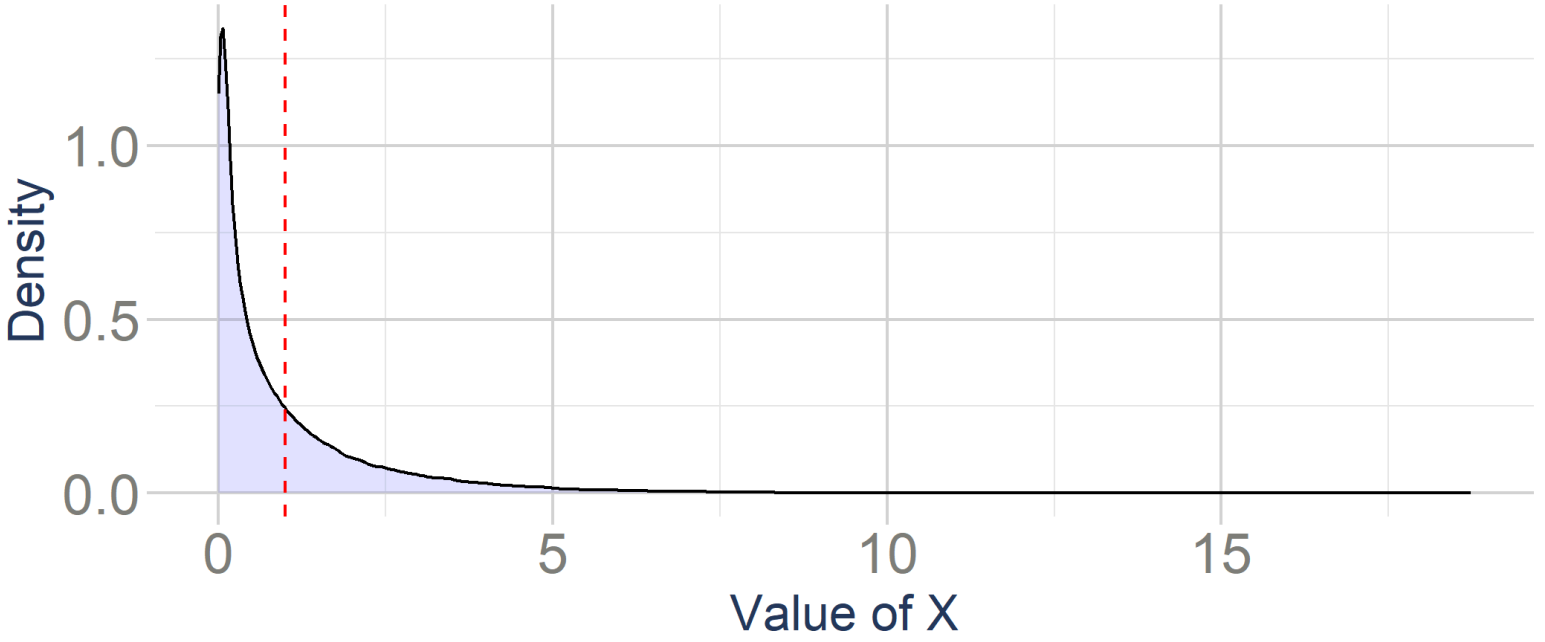
- But if I don't know σ how can I use it?
- We can use the sample standard deviation instead to estimate σ
- Since it is just an estimate, it adds uncertainty
- But if you have big sample, then you are really good at estimating standard deviation and the error is small
- So the distribution will still converge to normal, but you will need a bit more observations (say 50 rather than 40)

Sampling distribution of standard deviation

- Great, we now know sample means have a normal distribution in large samples
- Can we say something about the sampling distribution of the standard deviation?
- That is, if we take multiple samples and calculate the standard deviation of each sample, what will the distribution of these standard deviations look like?
- If X_i comes from a normal distribution, then yes! The sample variance is related to the chi-squared distribution
- This will be useful for constructing confidence intervals for the variance

From Normal to Chi-Square

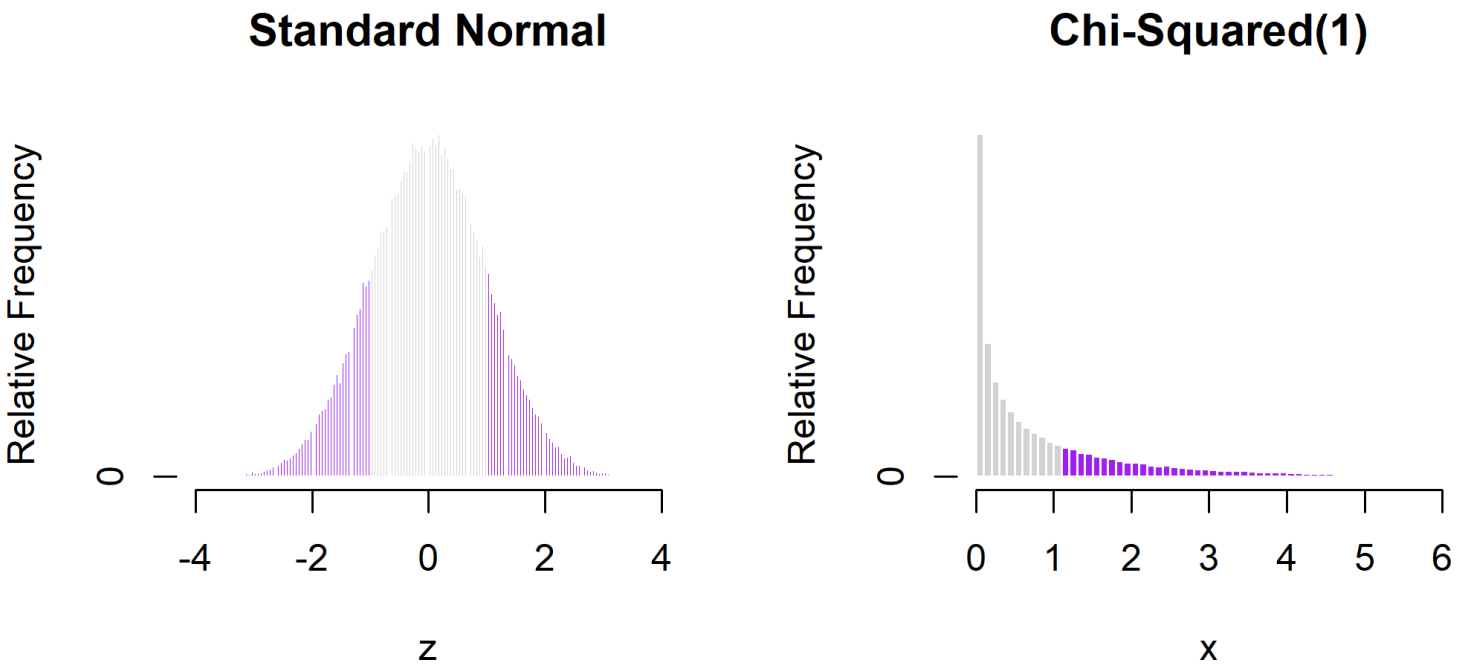
- We start with the standard random normal distribution N(0, 1).
- The transformation $X = Z^2$ gives rise to the Chi-Square distribution with 1 degree of freedom $\chi^2(1)$.
- The expectation of $\chi^2(1)$ is $E[X] = E[Z^2] = Var(Z) + E[Z]^2 = Var(Z) = 1$
- The variance of $\chi^2(1)$ is $var(X) = 2$



Visualizing the Connection

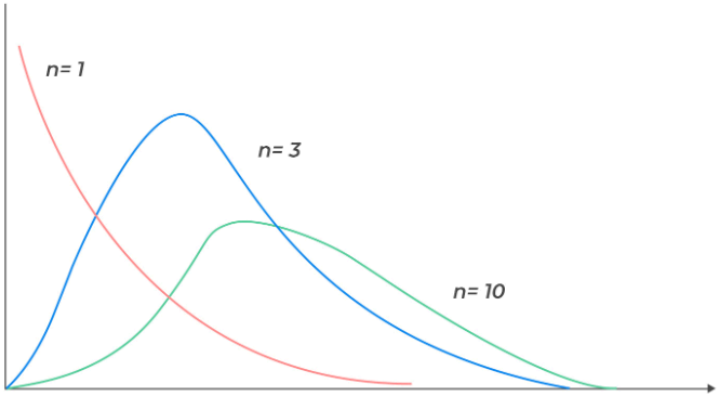
- The shaded areas represent probability that $X = Z^2 > 1$ where $X \sim \chi^2(1)$ and $Z \sim N(0, 1)$
- Shaded parts have the same area in both graphs

LearnR: Interactive Normal vs Chi-Square — adjust the threshold and see both distributions update



Chi-Square and the Sum of Random Normals

- More generally, sum of n iid squared standard normal variables is distributed as Chi-Square with n degrees of freedom
- $\sum_n Z^2 \sim \chi^2(n)$
- The expectation of $\chi^2(n)$ is $E[X(n)] = E[\sum_n Z_i^2] = \sum_n Var(Z_i) = n$
- The variance of $\chi^2(n)$ is $var(X) = 2n$



- Why does the shape converge to normal with large n?
- Because of CLT - it's sum of random variables