Class 3a: Review of concepts in Probability and Statistics

Business Forecasting

Roadmap

Last set of classess

- 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 hypothesis

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)								

What is a random 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.

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

What is a random 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.

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

What is a random 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.

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

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 $\bar{p}_c = 1245.43$ MXN
- This is our point estimate
- Can 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 it's sampling distribution

Estimators

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

https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm

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 a many,many,many samples of apartments and calculate 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
- Lets look at a couple of estimators and check if they are biased or not

Example 1: Estimator = 570

Expectation

- Consider some random variable X_i with unknown mean $E(X_i) = \mu$
- We want to estimate this mean
- The estimator: $\hat{\theta}_1 = 570$
- Expected Value: $E(\hat{\theta}_1) = 570$
- Bias: $E(\hat{\theta}_1) - \mu \neq 0$ if $\mu \neq 570$ (biased)

Example 2: Estimator = X_i

Expectation

- Consider some random variable X_i with unknown mean $E(X_i) = \mu$
- We want to estimate this mean
- The estimator: $\hat{\theta}_2 = X_i$
- Expected Value: $E(\hat{\theta}_2) = E(X_i) = \mu$
- Bias: $E(\hat{\theta}_2) - \mu = 0$ (unbiased)
- Is it a good estimator?

Example 3: 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: $\hat{\theta}_3 = \frac{3X_1+X_2}{5}$
- Expected Value: $E(\hat{\theta}_3) = \frac{3}{5}E(X_1) + \frac{1}{5}E(X_2) = \frac{3}{5}\mu + \frac{1}{5}\mu = \frac{4}{5}\mu$
- Bias: $E(\hat{\theta}_3) - \mu = \frac{4}{5}\mu - \mu = -\frac{1}{5}\mu$ (biased)

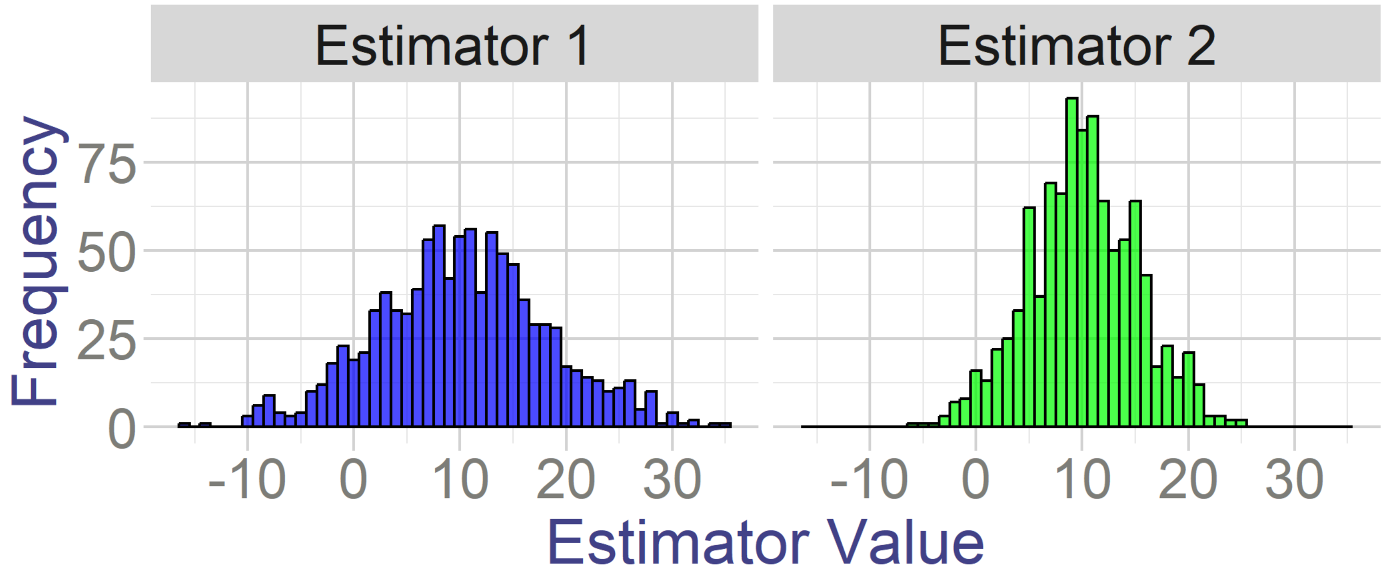
Example 4: Estimator = $\frac{\sum X_i}{n}$

Expectation

- Consider some random variable X_i with unknown mean $E(X_i) = \mu$
- We want to estimate this mean
- The estimator: $\hat{\theta}_4 = \frac{\sum_n X_i}{n}$
- Expected Value: $E(\hat{\theta}_4) = E(\frac{\sum_n X_i}{n}) = \frac{\sum_n E(X_i)}{n} = \frac{\sum_n E(\mu)}{n} = \mu$
- Bias: $E(\hat{\theta}_4) - \mu = 0$ (unbiased)

Variance of the estimator

- Good estimator is unbiased
- But how do we choose among unbiased estimator?
  - Suppose we sample 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 1: $\hat{\mu}_1 = (3X_1 + X_2)/4$
  - Estimator 2: $\hat{\mu}_2 = (X_1 + X_2 + X_3 + X_4)/4$
  - An estimator is more **efficient** if it has a smaller variance



Variance of the estimator

- Variance of an estimator is defined as:

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

- We want the estimator to have low variance!
- Estimator with the lower variance is more efficient
- In the example above

$$var(\hat{\mu}_1) = var(\frac{3X_1 + X_2}{4}) > var(\frac{X_1 + X_2 + X_3 + X_4}{4}) = var(\hat{\mu}_2)$$

- Relative efficiency of the two estimators is the ratio of their variances

$$Eff_{\hat{\mu}_1,\hat{\mu}_2} = \frac{var(\frac{3X_1+X_2}{4})}{var(\frac{X_1+X_2+X_3+X_4}{4})} = \frac{\frac{10}{16}}{\frac{4}{16}} = \frac{5}{2}$$

Variance of estimators

Example 1: Estimator = 570

- $Var(\hat{\theta}_1) = E[(\hat{\theta}_1 - E[\hat{\theta}_1])^2] = E[(570 - E[570])^2] = 0$

  Example 2: Estimator = X_i

- $Var(\hat{\theta}_2) = E[(X_i - \mu)^2] = \sigma^2$

  Example 4: Estimator = $\frac{\sum X_i}{n}$

- $Var(\hat{\theta}_4) = E\left[\left(\frac{\sum X_i}{n} - \mu\right)^2\right] = \frac{\sigma^2}{n}$

  Example 3: Estimator = $\frac{3X_1+X_2}{4}$

- $Var(\hat{\theta}_4) = E\left[((3X_1 + X_2)/4 - \mu)^2\right] = \frac{10\sigma^2}{16}$

Side note

- 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
- What is $E(X_1 + X_2)$
- What is $var(X_1 + X_2)$?
- What about $var(X_1 - X_2)$?

Biased Estimator = $s_b^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$

- Consider the estimator: $\hat{\theta}_6 = s_b^2$
- We are trying to estimate σ^2

$$E[\hat{\theta}_6] = E[s_b^2] = E[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}] = \frac{(n-1)\sigma^2}{n}$$

- So:

$$Bias(\hat{\theta}_6) = E[\hat{\theta}_6] - \sigma^2 = -\frac{\sigma^2}{n}$$

- We are underestimating the variance
- The sample variance estimator (divided by $\frac{1}{n-1}$) is unbiased:

$$E[\hat{\theta}_7] = E[s^2] = E[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}] = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

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:

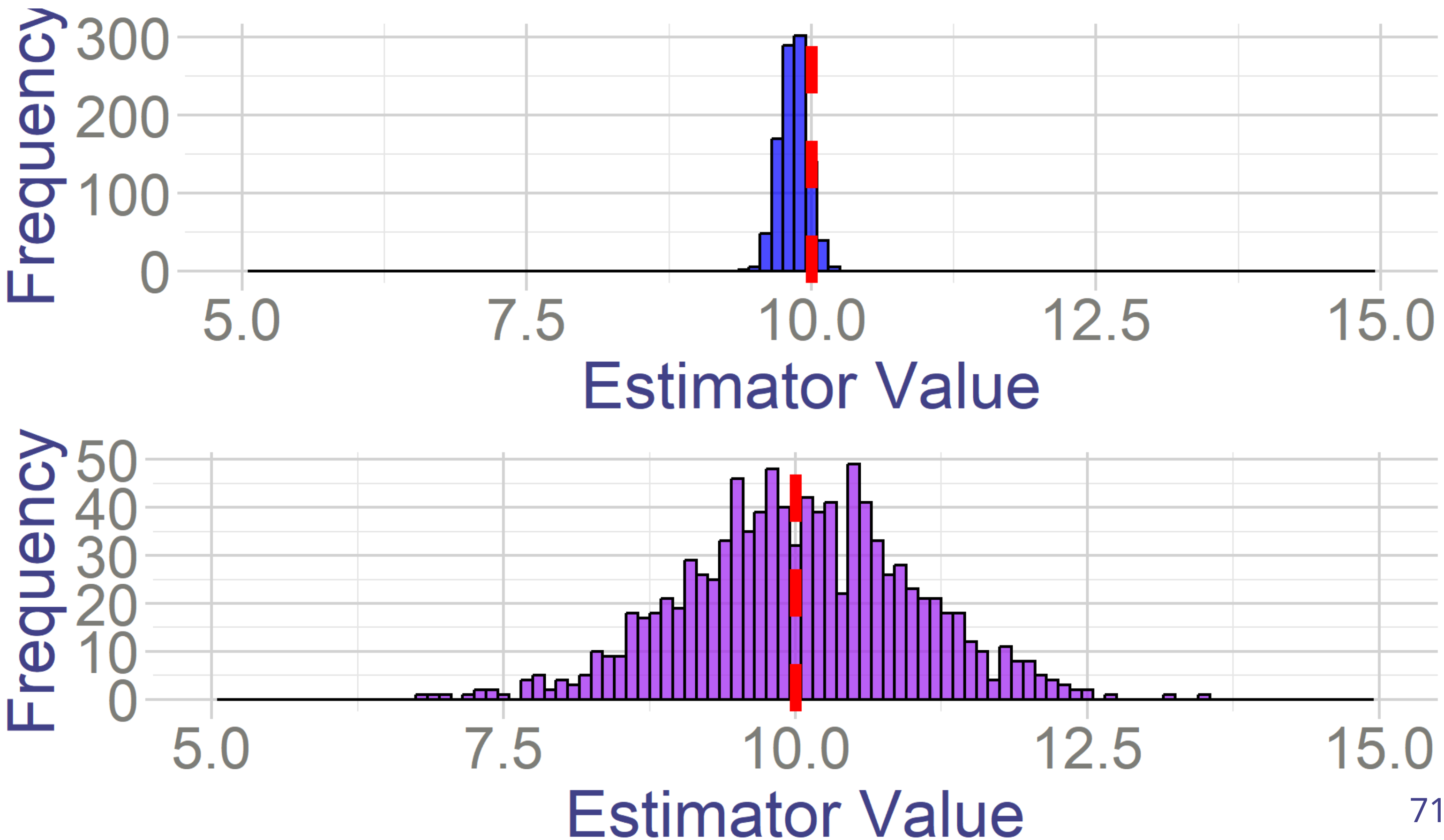
$$\begin{aligned}
MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\
&= E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\
&= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2] \\
&= E[(\hat{\theta} - E(\hat{\theta}))^2]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

- If estimator is unbiased, then

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

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?



Mean Squared Error of sample mean (optional)

- $\frac{3X_1+X_2}{4}$ is worse than $\frac{X_1+X_2}{2}$?
- Both estimators have the form of $\hat{\theta} = \sum_n c_i X_i$ with $n = 2$
  - They have different weights c_i or in vector form $\mathbf{c} = \{c_1, c_2, .. c_n\}$, with $\sum_i c_i = 1$
  - Sample mean is the best because for any n and $\mathbf{c}$ such that $\sum_i c_i = 1$:

$$argmin_{\mathbf{c}} \underbrace{E[(\sum_n c_i X_i - \mu)^2]}_{MSE} = \{\frac{1}{n}, \frac{1}{n}, .. \frac{1}{n}\}$$

Mean Squared Error of sample mean (optional)

And hence

$$\underbrace{min_{\mathbf{c}} E[(\sum_n c_i X_i - \mu)^2]}_{MSE} = E[(\frac{\sum_n X_i}{n} - \mu)^2]$$

- That is, for any estimator of μ of the form $\hat{\theta} = \sum_n c_i X_i$, sample mean has the lowest MSE!
  - Having different c_i than $\frac{1}{n}$ would increase the MSE

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 statistcs!

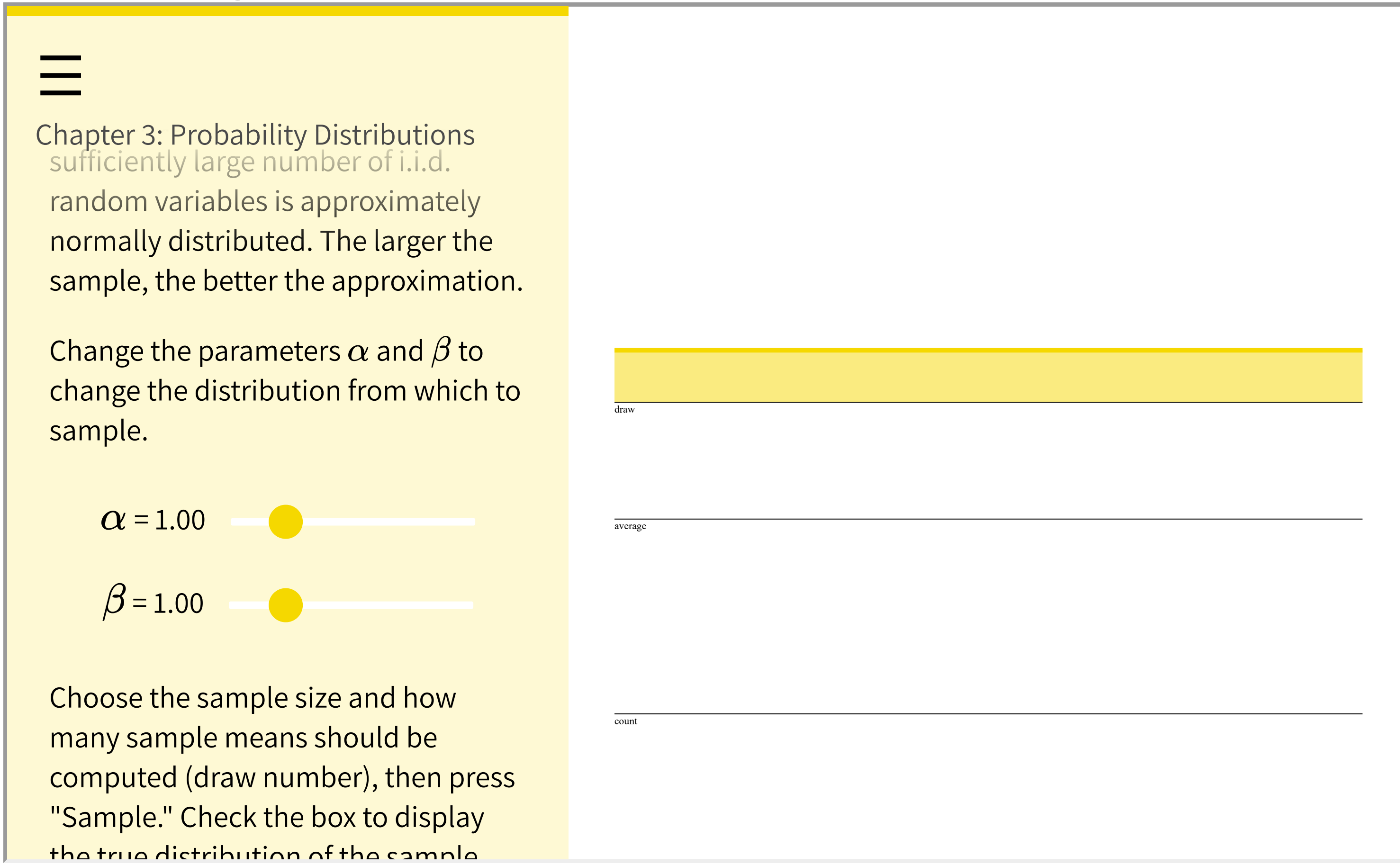
Central Limit Theorem

- 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**:

$$S_n \sim \mathcal{N}(n\mu, \underbrace{\sqrt{n}\sigma}_{st.dev.}) \qquad \text{and} \qquad \bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}) \qquad \text{and} \qquad \bar{Z}_n \sim \mathcal{N}(0, 1)$$

- In large samples, sample mean is normally distributed with mean μ and st. dev. $\frac{\sigma}{\sqrt{n}}$

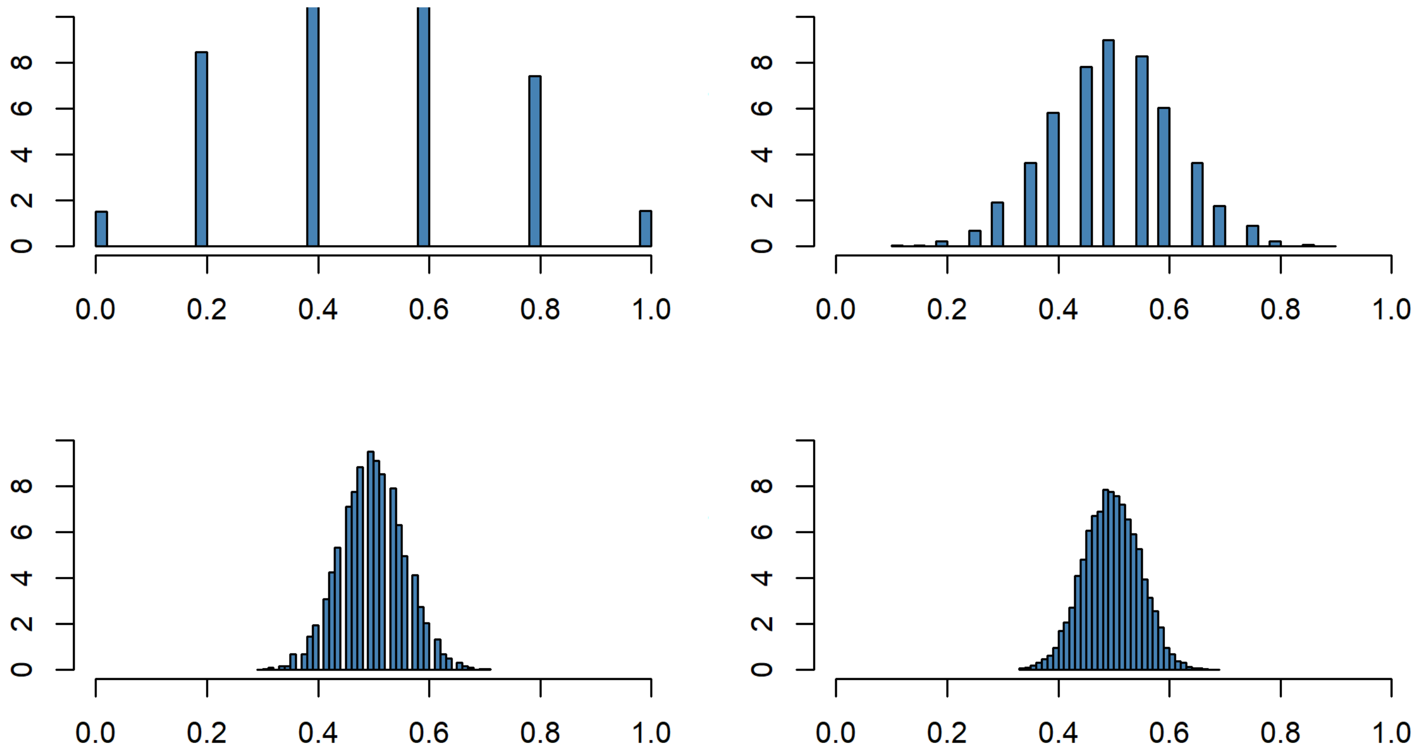
- The original distribution of X_i does not matter (but outliers make convergence longer)
- Larger **n**, tighter distribution around the mean
- Smaller σ , tighter distribution around the mean



Source: [https://seeing-theory.brown.edu/probability-distributions/index.html#section3)

What if it's a discrete variable?

- Let $X_i \sim \text{Bernoulli}(p = 0.5)$. Here is the distribution of $\bar{X}_n$:



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

Central Limit Theorem

What happens if some assumptions are not respected?

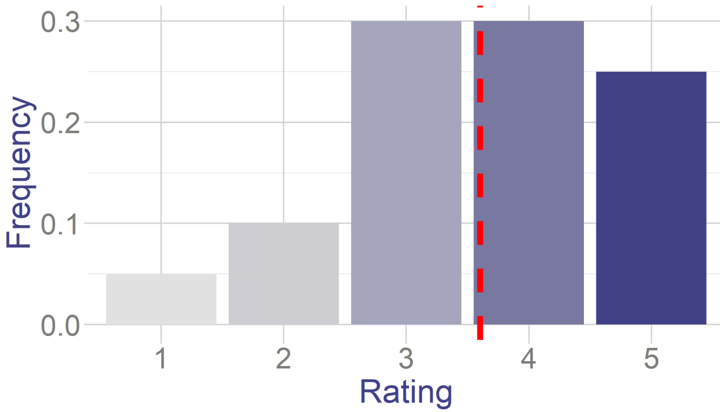
- Random draws means that each member of the population has equal chance of being selected
- Keep in mind that some values occur more often in the population than others
- More members with this value - higher chance of this value being sampled

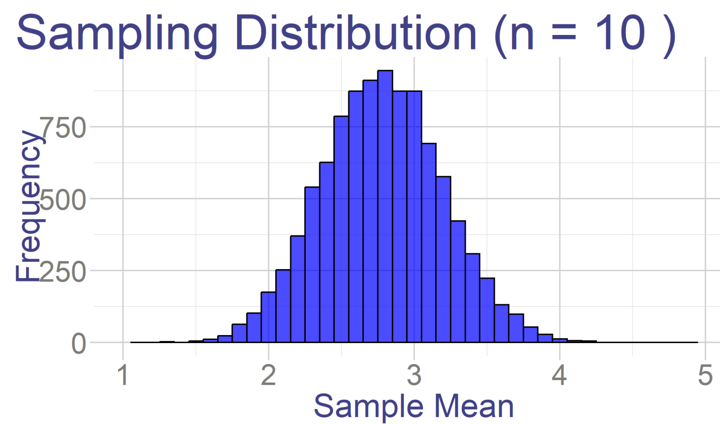
Example

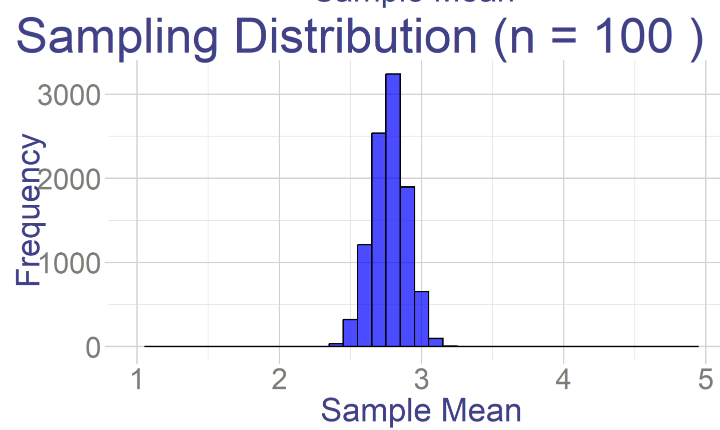
- Imagine you are evaluating a new skincare product to determine how people like it (on scale 1-5)
- However, you can only access online reviews
- The mean rating you calculated is 2.5
- Is it low because people don't like or because of other reason?

Central Limit Theorem

Suppose that this is the true distribution:







- But people who post online are more likely to be unhappy
- Suppose you are twice more likely to post if your rating is 1 or 2
- **Sample is not at random** from the population of customers
- Sampling distribution of the mean would look like this:

It's not centered at the correct value, no matter n !

Central Limit Theorem

Example 2 What happens if some assumptions are not respected?

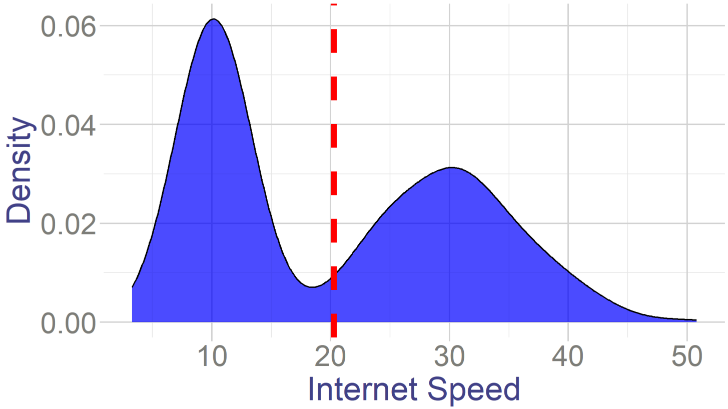
- IID means one draw does not change likelihood of other draws

Example

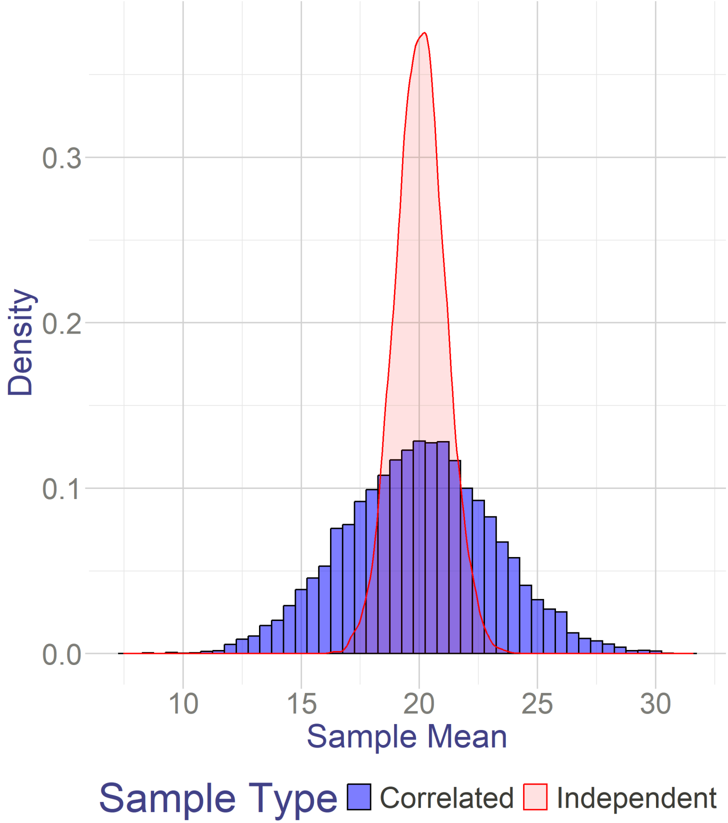
- Suppose you want to learn what's an average speed of internet in CDMX
- You choose at random the first apartment to measure the speed
- For the rest of the observations, you stay in the same building and measure at neighbors apartments

Central Limit Theorem

Suppose that this is the true distribution of speed:



- Speed across neighbors in the same building is likely correlated
- Observations are **not independent**
- Sampling distribution of the mean would look like this:



Variance is wider than implied by CLT!

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?

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

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

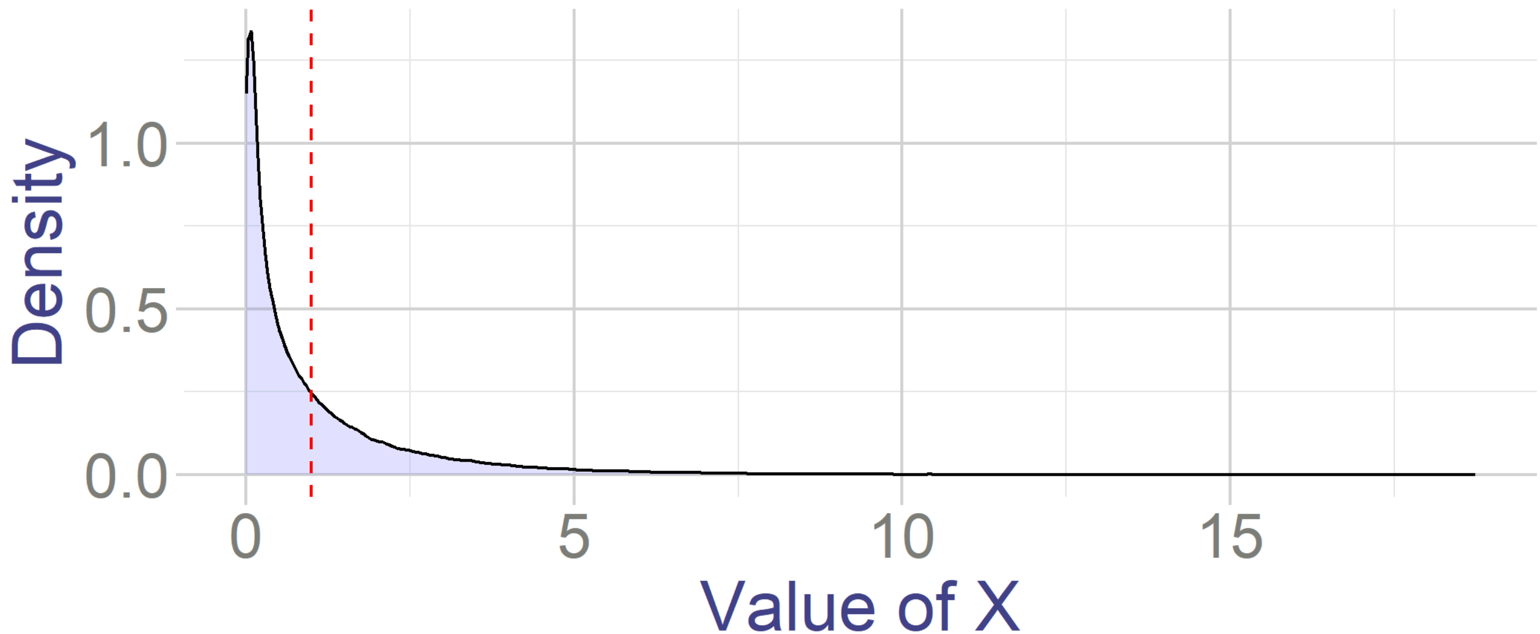
- But if I don't know σ how can I use it?
- We can use the sample standard deviation instead to estimate σ
- Since it is just an estimate, it adds uncertainty
- But if you have big sample, then you are really good at estimating standard deviation and the error is small
- So the distribution will still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

Standard deviation

- Great, sample means have normal distribution in large samples
- Can we say something about the standard deviation?
- If X_i is normal, then yes! Standard deviation will have **chi-square** distribution

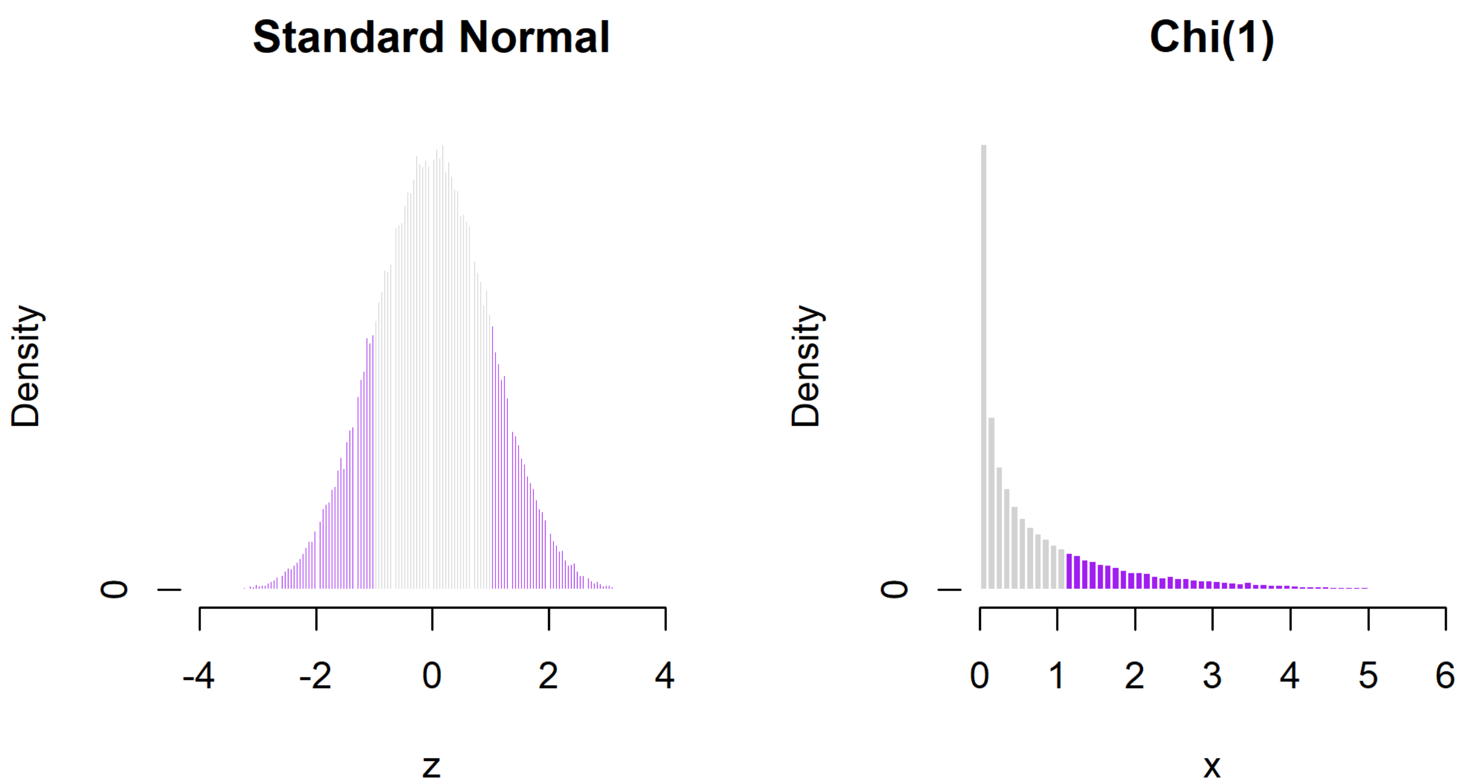
From Normal to Chi-Square

- 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 $var(X) = 2$



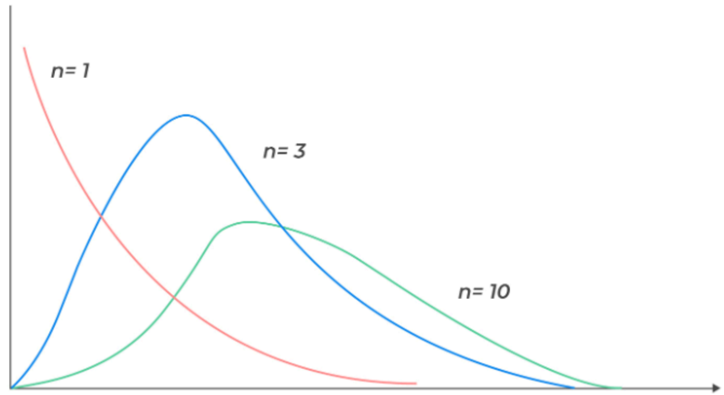
Visualizing the Connection

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



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 the shapes converges to normal with large n?
- Because of CLT - it's sum of random variables

Exercises:

- Review Exercises:
  - PDF 3: 1,2,3,4,6,7(b),9,10,11,12,13,14,15,16
- Homeworks
  - Lista 00.1: 6,7,8,9,10,11,12,13,14,15
  - Lista 00.2: 1,2,3,4,5,6,7,8,9,10,11,12,16,17,18,19,