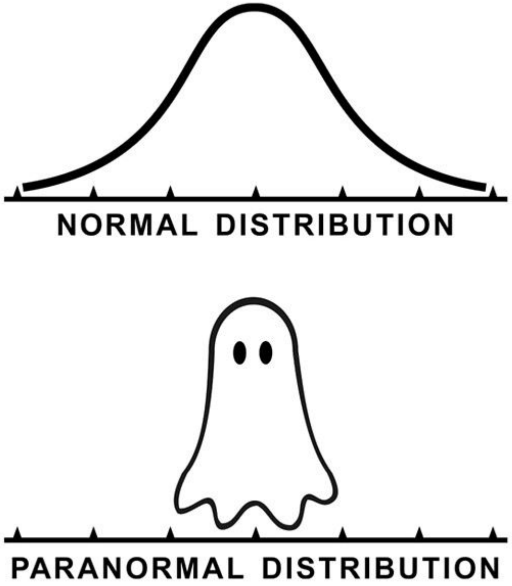
Sample Statistics and The Central Limit Theorem

LECTURE 19



CSCI 3022

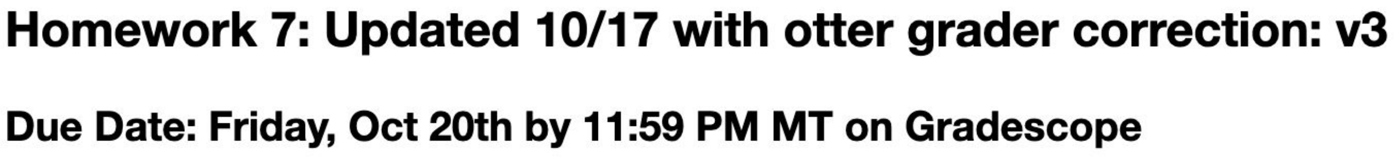
Maribeth Oscamou

Content credit: Acknowledgments

Announcements

No quiz this Friday

HW 7: Make sure you have the updated version (reclick on the assignment from Canvas). It



Homework 7: Updated 10/17 with otter grader correction: v3

Due Date: Friday, Oct 20th by 11:59 PM MT on Gradescope

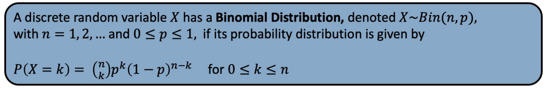
Probability vs Empirical Distributions

Any random variable has a distribution:

Probability (Theoretical) Distribution

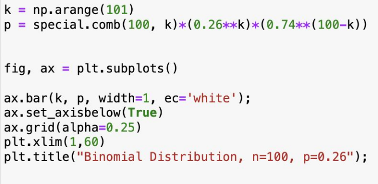
These are the distributions of random variables. We have focused on some common ones in the past few weeks:

(Recall from HW 6:)

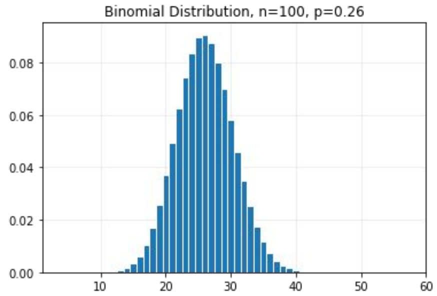


A discrete random variable X has a **Binomial Distribution,** denoted $X \sim Bin(n, p)$, with $n = 1, 2, \ldots$ and $0 \leq p \leq 1$, if its probability distribution is given by

$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$



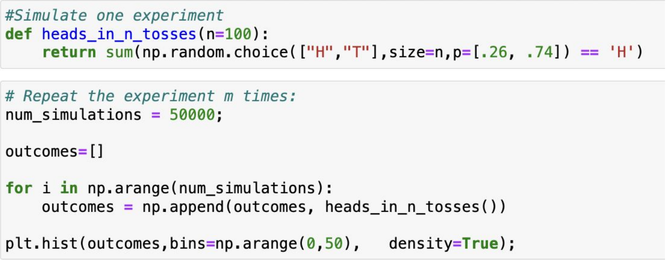
```
k = np.arange(101)
p = special.comb(100, k)*(0.26**k)*(0.74**(100-k))


fig, ax = plt.subplots()

ax.bar(k, p, width=1, ec='white');
ax.set_axisbelow(True)
ax.grid(alpha=0.25)
plt.xlim(1,60)
plt.title("Binomial Distribution, n=100, p=0.26");
```



- <u>Empirical (Simulated) Distribution:</u>

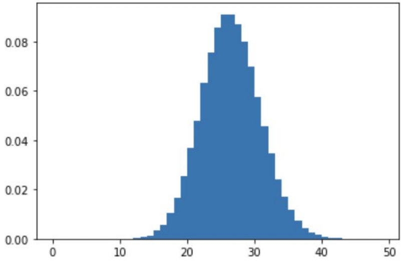
  based on simulations/observations
- Observations can be from **repetitions of an experiment**
  - All observed unique values
  - The proportion of times each value appears

(Recall from lec 12:)



```
#Simulate one experiment
def heads_in_n_tosses(n=100):
    return sum(np.random.choice(["H","T"],size=n,p=[.26, .74]) == 'H')
```

```
# Repeat the experiment m times:
num_simulations = 50000;

outcomes=[]

for i in np.arange(num_simulations):
    outcomes = np.append(outcomes, heads_in_n_tosses())

plt.hist(outcomes,bins=np.arange(0,50),   density=True);
```



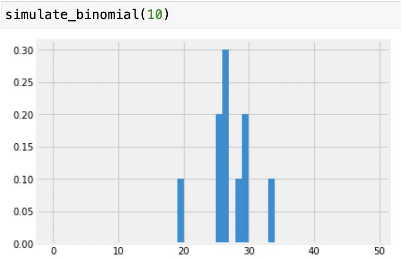
Law of Averages / Law of Large Numbers

If a chance experiment is **repeated many times**,

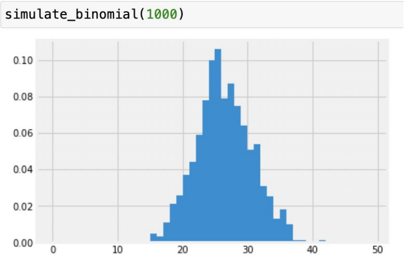
independently and under the **same conditions**,

then the **Empirical Distribution** gets closer to the **Probability Distribution.**

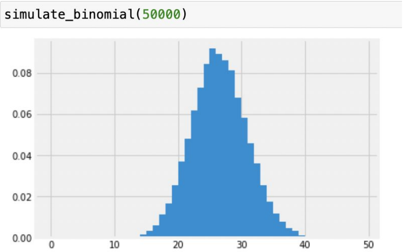
An experiment consists of flipping a coin 100 times and counting the number of heads, where the probability of heads is 0.26. You repeat this experiment and plot the distribution for the number of heads:



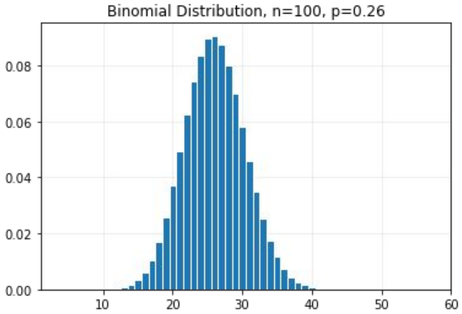
```
simulate_binomial(10)
```



```
simulate_binomial(1000)
```



```
simulate_binomial(50000)
```



As you increase the number of times you do this experiment, the empirical distribution gets closer to the theoretical probability distribution

Empirical Distribution of a Sample

If the **sample size is large**, then

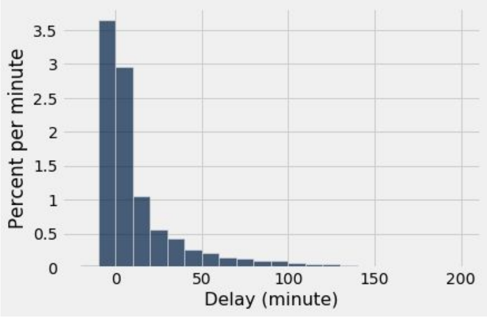
the empirical distribution of a random sample with replacement

resembles the probability distribution of the population with high probability.

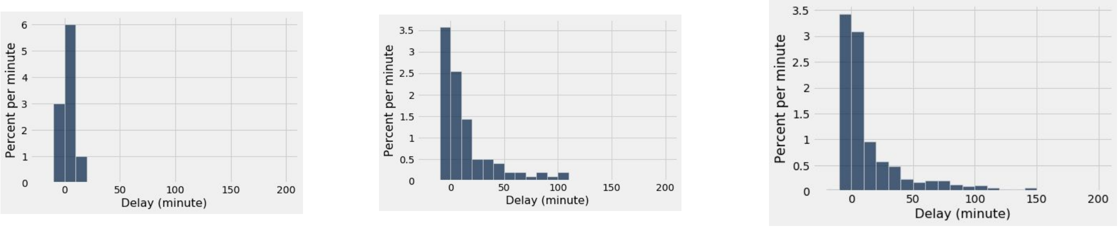
Recall:

A **random sample with replacement** is a sample drawn **uniformly** at random **with** replacement.

Population (Theoretical Probability) Distribution



Empirical distribution of random sample of size n with replacement



n=10

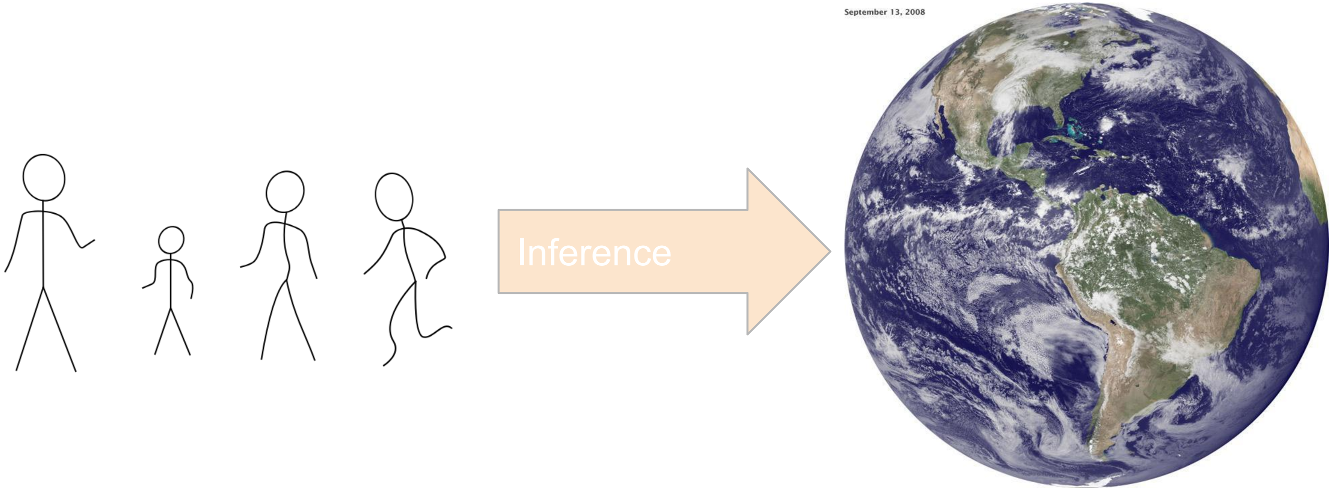
n=100

n=1000

Inference

Sample Data

Population



Inference

Inference

- Statistical Inference:

  Making conclusions based on data in **random samples**

- **Example**:

  fixed

  Use the data to guess the value of an *unknown number*

  depends on the **random** sample

  Create an **estimate** of the unknown quantity

From Populations to Samples

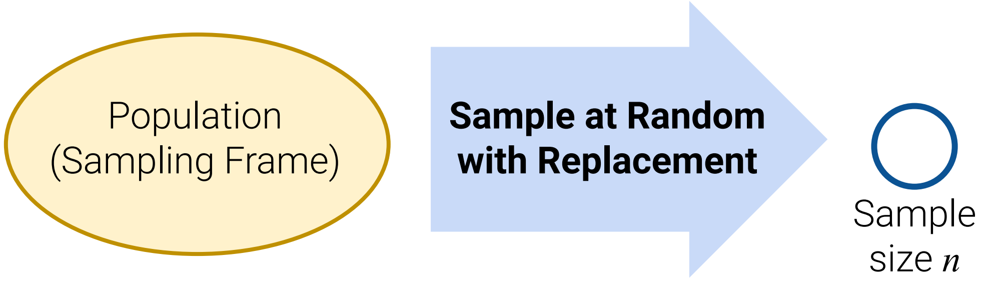
We've talked extensively about **populations**:

- If we know the **distribution of a random variable**, we can reliably compute expectation, variance, functions of the random variable, etc.

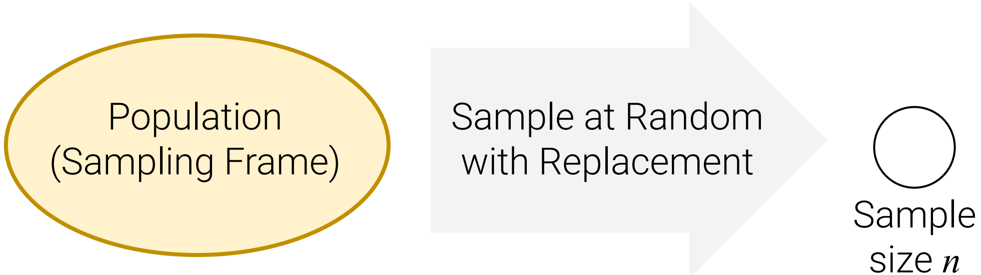
However, in Data Science, we often collect **samples**.

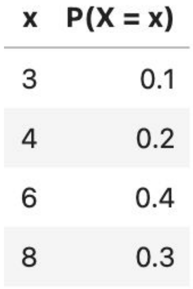
- We don't know the distribution of our population.
- We'd like to use the distribution of your sample to estimate/infer properties of the population.

The **big assumption** we make in modeling/inference: Our random sample datapoints are **IID**.



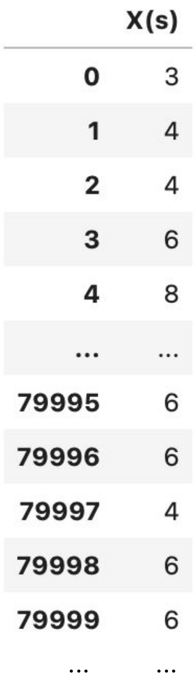
The Sample is a Set of IID Random Variables





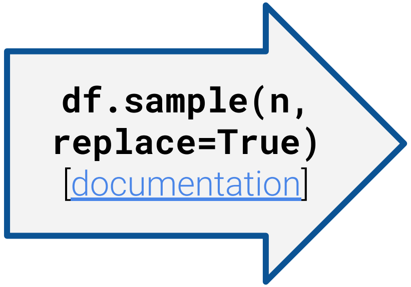
x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

or



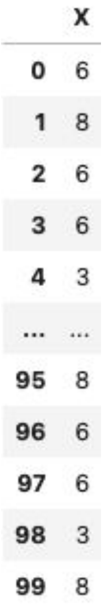
	X(s)
0	3
1	4
2	4
3	6
4	8
...	...
79995	6
79996	6
79997	4
79998	6
79999	6
...	...

Population
(really large N)



```
df.sample(n,
replace=True)
```

[documentation]



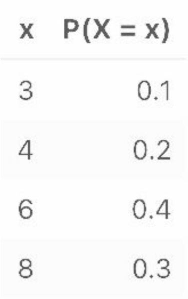
	x
0	6
1	8
2	6
3	6
4	3
...	...
95	8
96	6
97	6
98	3
99	8

Each observation in our sample is a **Random Variable** drawn **IID** from our population distribution.

Sample $X_1, X_2, \ldots, X_n$

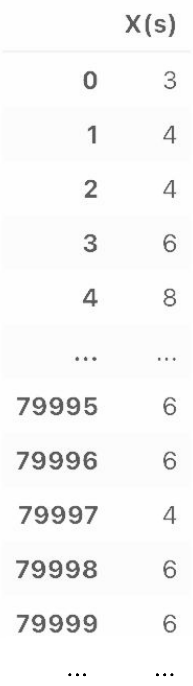
The Sample is a Set of IID Random Variables

Population
(Sampling Frame)



x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

or



	X(s)
0	3
1	4
2	4
3	6
4	8
...	...
79995	6
79996	6
79997	4
79998	6
79999	6
...	...

$E[X] = 5.9$

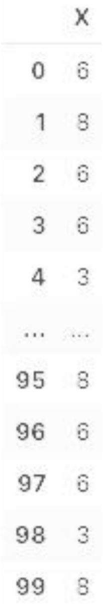
Population Mean
A **number**,
i.e., fixed value

μ

Sample at Random
with Replacement

```
df.sample(n,
replace=True)
```
[documentation]

Sample
size n



	x
0	6
1	8
2	6
3	6
4	3
...	...
95	8
96	6
97	6
98	3
99	8

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

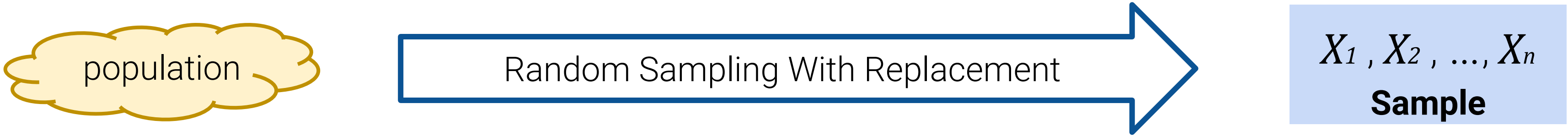
Sample Mean
A **random variable**!
Depends on our randomly
drawn sample!!

```
np.mean(…) = 5.71
```

Sample $X_1, X_2, \ldots, X_n$

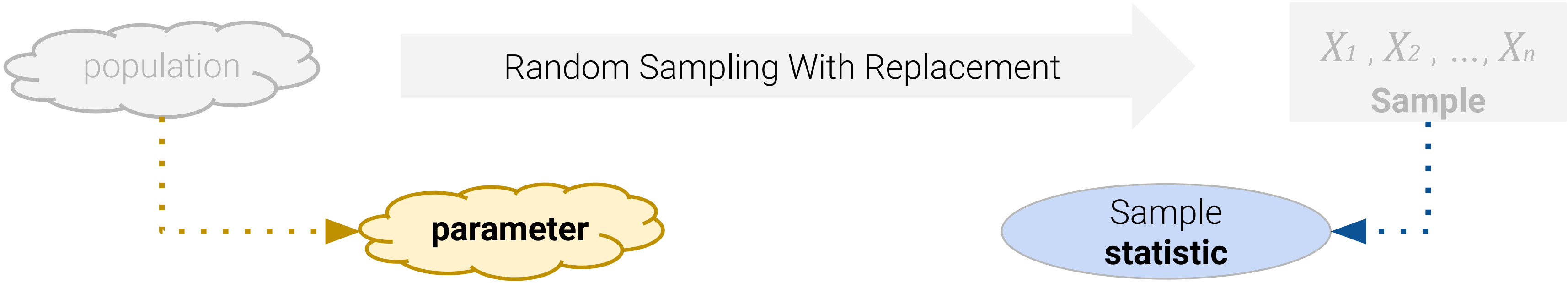
[Terminology] Parameters, Statistics, and Estimators

Inference is all about **drawing conclusions** about **population parameters**, given only a **random sample**.



[Terminology] Parameters, Statistics, and Estimators

Inference is all about **drawing conclusions** about **population parameters**, given only a **random sample**.



population

Random Sampling With Replacement

$X_1, X_2, \ldots, X_n$

Sample

parameter

Sample **statistic**

A **parameter** is a numerical function of the **population**.

Population mean

μ

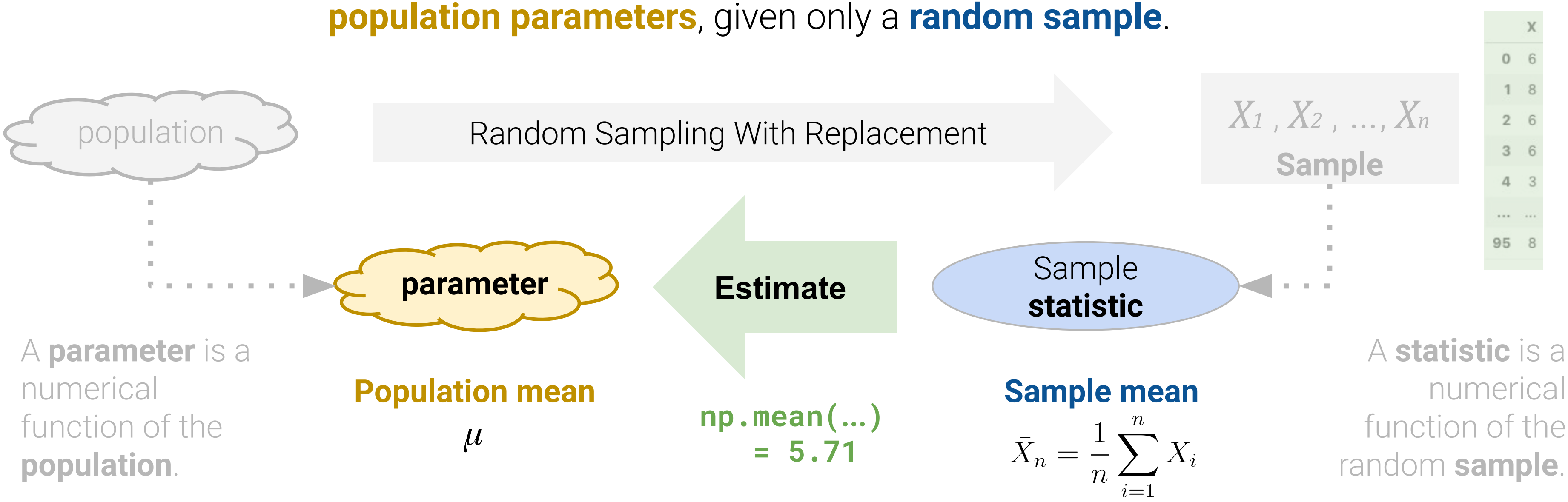
Sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

A **statistic** is a numerical function of the random **sample**.

[Terminology] Parameters, Statistics, and Estimators

Inference is all about **drawing conclusions** about **population parameters**, given only a **random sample**.



	x
0	6
1	8
2	6
3	6
4	3
...	...
95	8

A **parameter** is a numerical function of the **population**.

Population mean

μ

`np.mean(...)`
`= 5.71`

Sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

A **statistic** is a numerical function of the random **sample**.

We can then use the sample statistic as an **estimator** of the true population parameter.

Since our **sample is random**, our statistic (which we use as our estimator) could have been different.

Example: When we use the sample mean to estimate the population mean, our estimator is almost always going to be somewhat off.

[Terminology] Parameters, Statistics, and Estimators

Inference is all about **drawing conclusions** about **population parameters**, given only a **random sample**.

A **parameter** is a numerical function of the **population**.

A **statistic** is a numerical function of the random **sample**.

Wait: Statistics are Random variables? Then they have their own distributions!

Example of a sample statistic: Median of a sample
The median of a sample is a random variable (because different samples lead to different medians)
The sample median has its own distribution!

DEMO: https://onlinestatbook.com/stat_sim/sampling_dist/

- You have only one random sample, and it has only one median.
- But the sample could have come out differently.
- And then the sample median might have been different.
- So there are many possible sample medians.

Definition: Standard Error of a Statistic

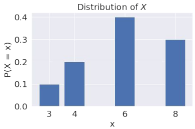
Example of a sample statistic: Median of a sample
The median of a sample is a random variable (because different samples lead to different medians)
The sample median has its own distribution!

Definition: The **standard error** of a statistic is the standard deviation of the sampling distribution of that statistic.

Data Generation Process: Estimating a Value

One View: Randomly draw a random sample, then compute the statistic for that sample.

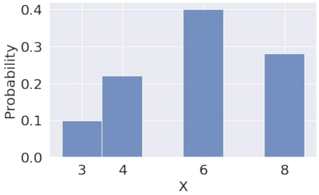
population



$X_1, X_2, \ldots, X_n$

Sample

	X
0	6
1	8
2	6
3	6
4	3
...	...
95	8

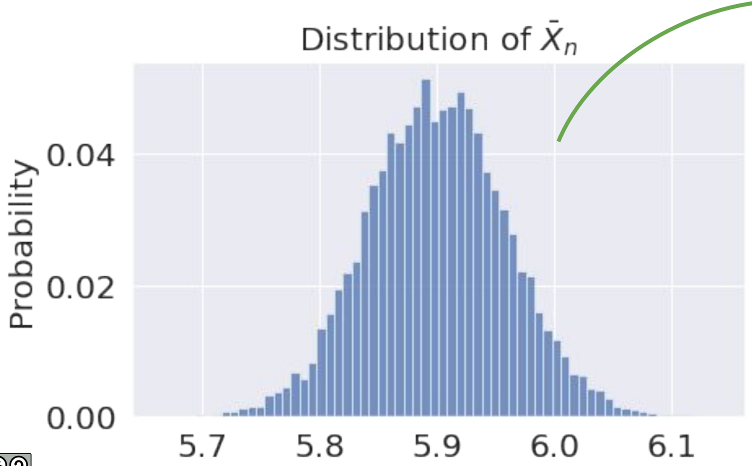


Sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

`np.mean(...) = 5.82`

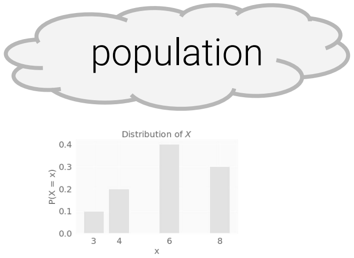
Another View: Randomly draw from the distribution of the statistic (generated from all possible samples).



5.82

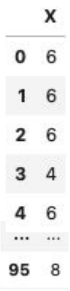
We often use the 2nd view because we want to interpret the estimator directly, and not the random sample.

If We Drew a Different Sample, We'd Get A Different Estimator

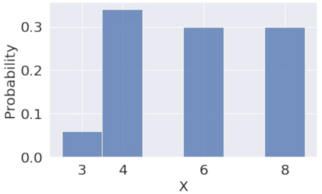


$X_1 , X_2 , \ldots, X_n$

Sample



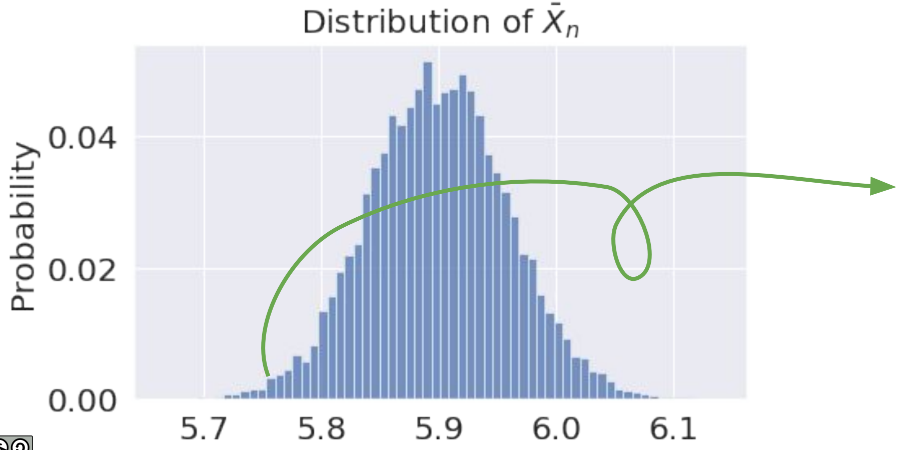
	X
0	6
1	6
2	6
3	4
4	6
...	...
95	8



Sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

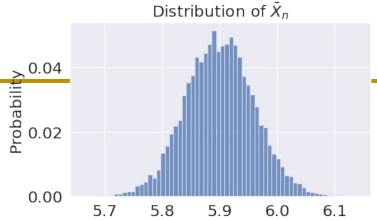
```
np.mean(...) = 5.74
```



5.74

The value of our estimator is a function of the random sample. The estimator is therefore also random.

Probability vs Empirical Distributions



- Values of a statistic vary because of **random samples**

- Probability (Sampling) Distribution of **a statistic:**
  - All possible values of the statistic,
  - and all the corresponding probabilities

- Often challenging to calculate analytically
  - Either have to do the math (may not be possible…)
  - Or generate all possible samples and calculate the statistic based on each sample (lots of compute!)

- Empirical distribution of a statistic:
  - Based on **simulated values** of the statistic
  - Consists of all the **observed values** of the statistic,
  - and the **proportion of times** each value appeared

- Good approximation to the probability distribution of the statistic
  - if the number of repetitions in the simulation is large

Sample Averages

- DEMO: https://onlinestatbook.com/stat_sim/sampling_dist/

Sample Averages

- The Central Limit Theorem describes how the normal distribution (a bell-shaped curve) is connected to random **sample sums** and **averages.**
- We care about sample averages because they estimate population averages.

Highly recommended video explaining the Central Limit Theorem:

https://www.youtube.com/watch?v=zeJD6dqJ5lo

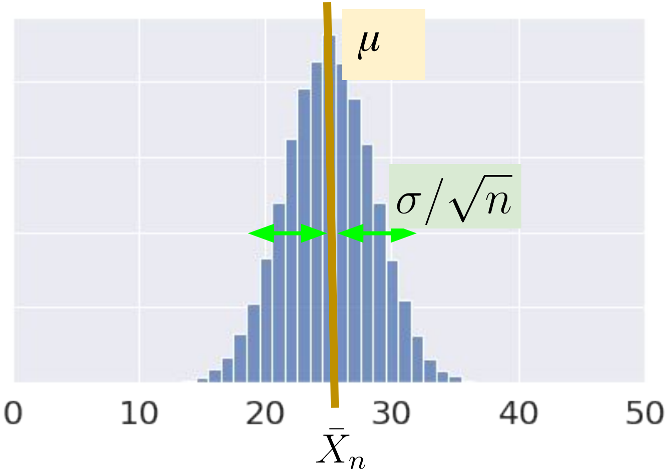
The Central Limit Theorem

No matter what population you are drawing from:

If an IID sample of size n is large,

the probability distribution of the **sample mean**

is $\underbrace{\textbf{roughly normal}}$ with $\underbrace{\text{mean } \mu \text{ and SD } \sigma/\sqrt{n}}_{\text{(pop mean } \mu\text{, pop SD } \sigma\text{)}}$



Any theorem that provides the rough distribution of a statistic
and **doesn't need the distribution of the population** is valuable to data scientists.

- Because we rarely know a lot about the population!

For a more in-depth demo: https://onlinestatbook.com/stat_sim/sampling_dist/

Properties of the Sample Mean

Consider an IID sample $X_1, X_2, \ldots, X_n$ drawn from a numerical population with **mean μ and SD σ**.

Define the **sample mean**:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Expectation:

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}(n\mu) = \mu$$

Variance/Standard Deviation:

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \underbrace{\frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{Var}(X_i)\right)}_{\text{IID} \rightarrow \mathrm{Cov}(X_i, X_j) = 0} = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}$$

$$\mathrm{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

Called the **standard error** of the sample mean

How Large Is "Large"?

No matter what population you are drawing from:

If an IID **sample of size n is large**,

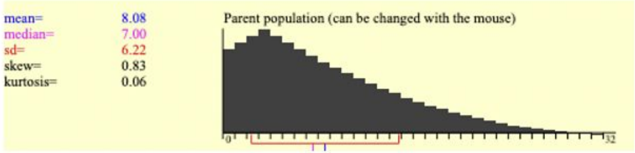
the probability distribution of the sample mean

is **roughly normal** with mean μ and SD $\dfrac{\sigma}{\sqrt{n}}$.

(where pop mean is μ, pop SD is σ)

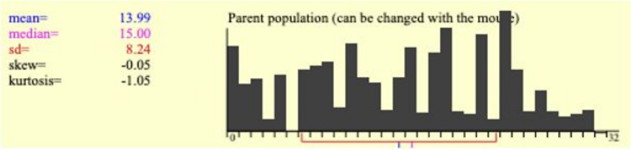
How large does n have to be for the normal approximation to be good?

- …It depends on the shape of the distribution of the population…
- Common rule of thumb: **n > 30**.
- If population is **roughly symmetric and unimodal**/uniform, could need as few as **n = 20**.
- If population is very skewed, ***you will need bigger n.***
- If in doubt, you can use a technique called bootstrapping (which we'll learn in a couple of weeks) and see if the bootstrapped distribution is bell-shaped.



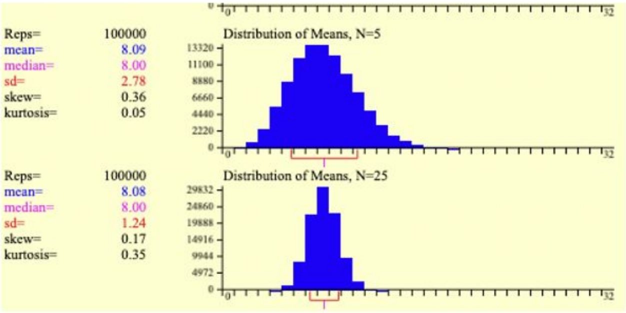
```
mean=        8.08
median=      7.00
sd=          6.22
skew=        0.83
kurtosis=    0.06
```

Parent population (can be changed with the mouse)



```
mean=        13.99
median=      15.00
sd=           8.24
skew=        -0.05
kurtosis=    -1.05
```

Parent population (can be changed with the mouse)

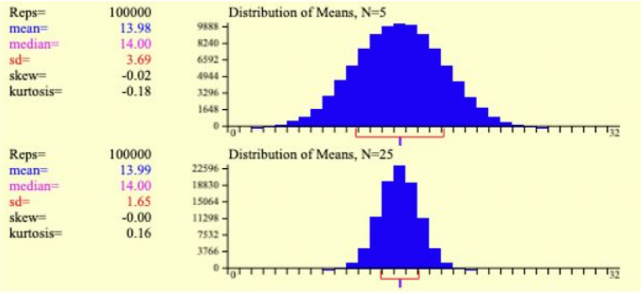


```
Reps=        100000
mean=          8.09
median=        8.00
sd=            2.78
skew=          0.36
kurtosis=      0.05
```

Distribution of Means, N=5

```
Reps=        100000
mean=          8.08
median=        8.00
sd=            1.24
skew=          0.17
kurtosis=      0.35
```

Distribution of Means, N=25



```
Reps=        100000
mean=         13.98
median=       14.00
sd=            3.69
skew=         -0.02
kurtosis=     -0.18
```

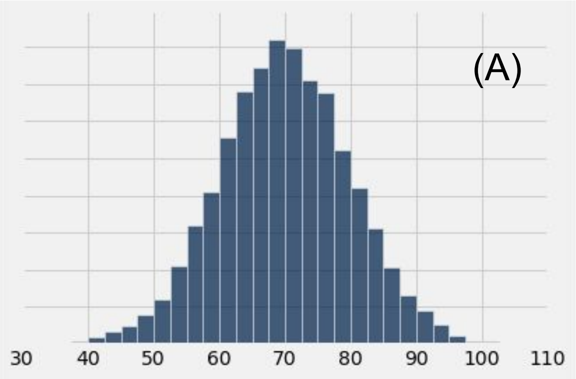
Distribution of Means, N=5

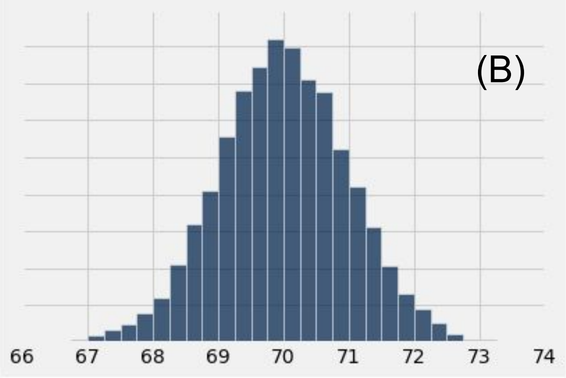
```
Reps=        100000
mean=         13.99
median=       14.00
sd=            1.65
skew=         -0.00
kurtosis=      0.16
```

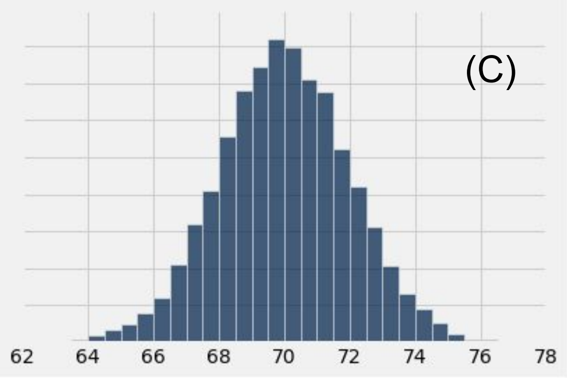
Distribution of Means, N=25

Discussion Question

A population has average 70 and SD 10. One of the histograms below is the distribution of the averages of 10,000 random samples of size 100 drawn from the population. Which one?







Discussion Question

Suppose salaries at a very large corporation have a mean of \$162,000 and a standard deviation of \$32,000.

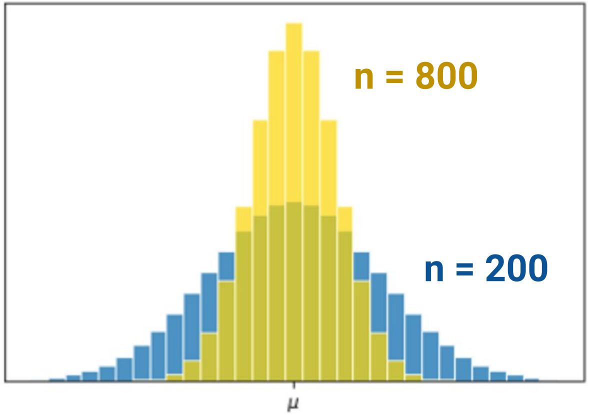
a). If a single employee is randomly selected, what is the probability that their salary exceeds \$175,000?

b). If 100 employees are randomly sampled, what is the probability that their average salary exceeds \$175,000?

Using the Sample Mean to Estimate the Population Mean

Our goal is often to **estimate** some characteristic of a population.

- Example: average height of Cal undergraduates.
- We typically can collect a **single sample**. It has just one average.
- Since that sample was random, it *could have* come out differently.



We should consider the **average value and spread** of all possible sample means, and how it scales with the sample size n.

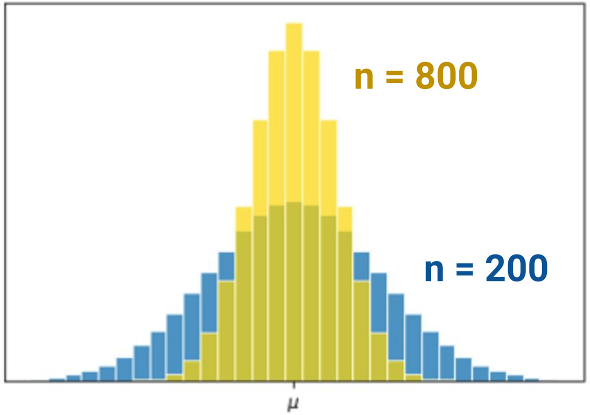
$$\mathbb{E}[\bar{X}_n] = \mu$$

$$\mathrm{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

Using the Sample Mean to Estimate the Population Mean

Our goal is often to **estimate** some characteristic of a population.

- Example: average height of Cal undergraduates.
- We typically can collect a **single sample**. It has just one average.
- Since that sample was random, it *could have* come out differently.



We should consider the **average value and spread** of all possible sample means, and what this means for how big n should be.

$$\mathbb{E}[\bar{X}_n] = \mu$$

$$\text{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

For every sample size, the expected value of the sample mean is the population mean.

We call the sample mean an **unbiased estimator** of the population mean.

We Have Finally Formalized the Distinction Between Estimators and Parameters

An **unbiased estimator** means that if we were to repeat this sampling process many times, the expected value of our estimates should be equal to the true values we are trying to estimate.

Ex: The sample mean is an unbiased estimator for the population mean, since

$$\mathbb{E}[\bar{X}_n] = \mu$$

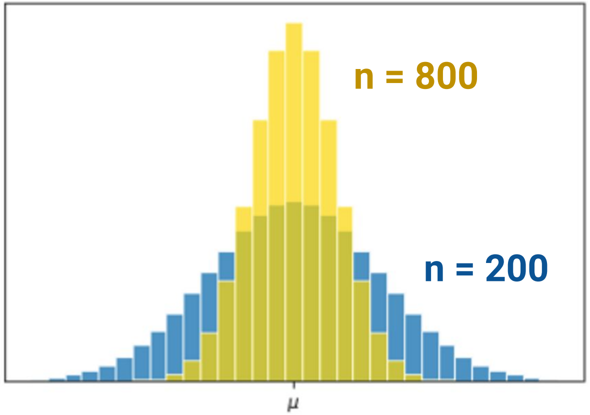
For every sample size, the expected value of the sample mean is the population mean.

We call the sample mean an **unbiased estimator** of the population mean.

Using the Sample Mean to Estimate the Population Mean

Our goal is often to **estimate** some characteristic of a population.

- Example: average height of Cal undergraduates.
- We typically can collect a **single sample**. It has just one average.
- Since that sample was random, it *could have* come out differently.



We should consider the **average value and spread** of all possible sample means, and what this means for how big n should be.

$$\mathbb{E}[\bar{X}_n] = \mu$$

$$\text{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

For every sample size, the expected value of the sample mean is the population mean.

We call the sample mean an **unbiased estimator** of the population mean.

Square root law If you increase the sample size by a factor, the SD decreases by the square root of the factor.

The sample mean is more likely to be close to the population mean if we have a larger sample size.

Discussion Question

A hardware store receives a shipment of 10,000 bolts that are supposed to be 12 cm long.
The mean is indeed 12 cm, and the standard deviation is 0.2 cm.

What is the mean and standard deviation of the ***average length of bolts in 100 randomly chosen*** bolts at this hardware store?

Discussion Question

A hardware store receives a shipment of bolts that are supposed to be 12 cm long. The mean is indeed 12 cm, and the standard deviation is 0.2 cm. For quality control, the hardware store chooses 100 bolts at random to measure.

They will declare the shipment defective and return it to the manufacturer if the average length of 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment is found **satisfactory**.

Have a Normal Day!