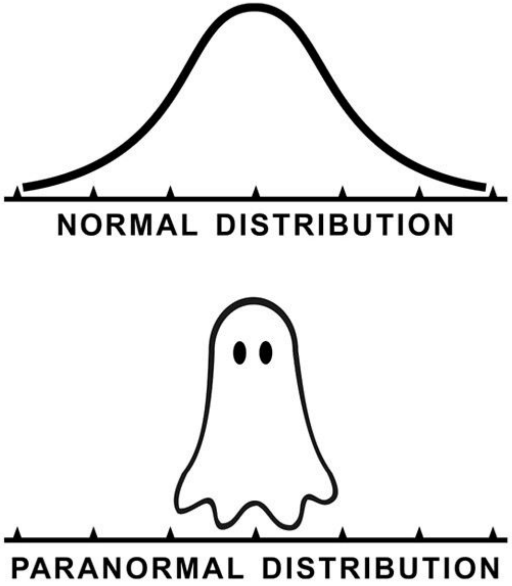
Sample Statistics and The Central Limit Theorem

LECTURE 17



CSCI 3022

Maribeth Oscamou

Content credit: Acknowledgments

Course Logistics: 7th Week At A Glance

Mon 2/26	Tues 2/27	Wed 2/28	Thurs 2/29	Fri 3/1
Attend & Participate in Class	(Optional): Attend Notebook Discussion with our TA (5-6pm Zoom)	Attend & Participate in Class	HW 6 Due 11:59pm	Attend & Participate in Class NO QUIZ!

- Finish Lesson 16:
    - Sampling Bias
    - Random Samples
- Parameters vs Statistics
- Sampling Distributions of Statistics
- Central Limit Theorem

Today's Roadmap

CSCI 3022

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.

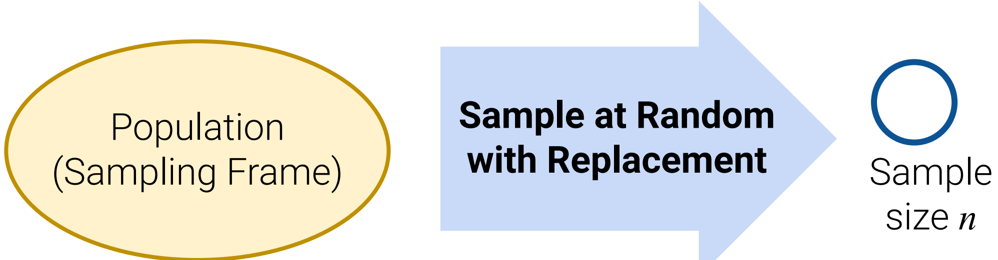
However, in Data Science, we often collect **samples**.

- 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 data points are **INDEPENDENT and IDENTICALLY DISTRIBUTED (IID)**

We can safely make this assumption anytime we sample at random with replacement (OR when we use a simple random sample and our sample size < 10% of the population size)



A Random Sample With Replacement is a Set of IID Random Variables

Population
(Sampling Frame)

Sample at Random
with Replacement

Sample size n

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 $X_1, X_2, \ldots, X_n$

A Random Sample With Replacement 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
```

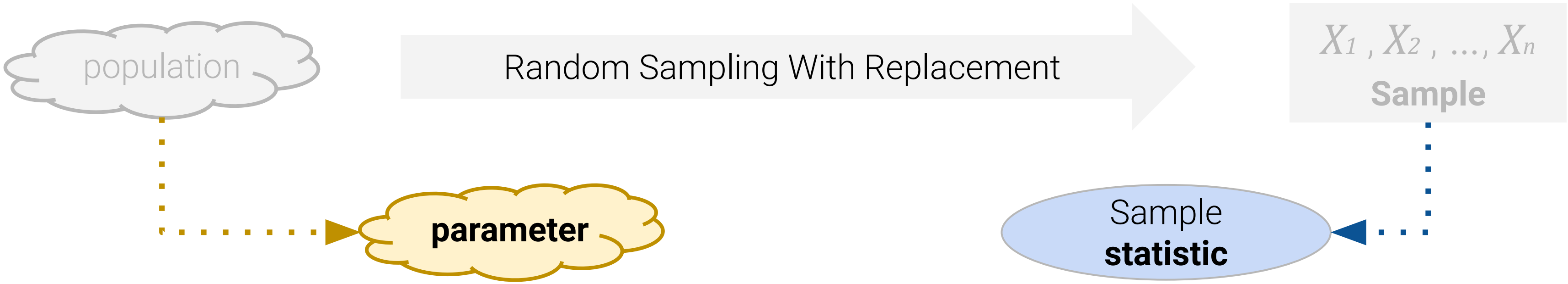
Sample $X_1, X_2, \ldots, X_n$

[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**.



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**.

Statistics are **random variables**!

[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

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

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

	Parameter		Statistic
Mean	μ	mu	$\bar{x}$
Proportion	p		$\hat{p}$
Std. Dev.	σ	sigma	s
Correlation	ρ	rho	r

[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

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

parameter

Estimate

Sample **statistic**

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. We want to know HOW off is it?

Sampling Distributions of Statistics

CSCI 3022

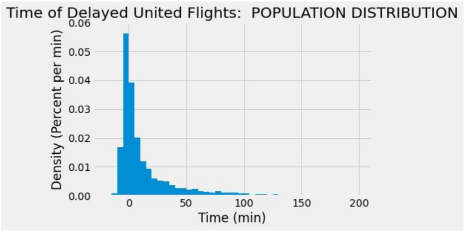
- Parameters vs Statistics
- **Sampling Distributions of Statistics**
- Central Limit Theorem

Understanding Distribution Terminology

Population Distribution

The underlying distribution

(usually unknown)



parameter

Population median = 2

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

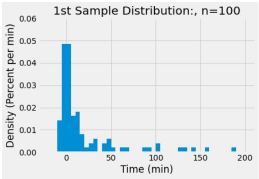
`df.sample(n, replace=True)`

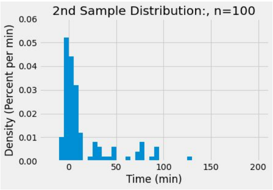
`df.sample(n, replace=True)` [documentation]

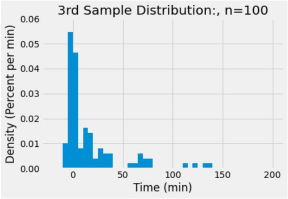
`df.sample(n, replace=True)`

Sample Distribution(s)

(aka the histogram of your data)







...

Sampling Distribution of a Statistic

Calculate medians from all possible samples and create histogram

Sample **statistic**

1st Sample Median= 3

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

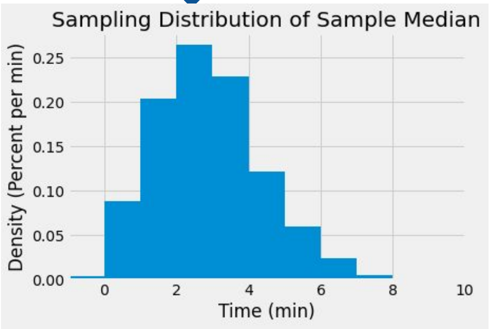
Sample **statistic**

2nd Sample Median= 3

Sample **statistic**

3rd Sample Median= 2

...



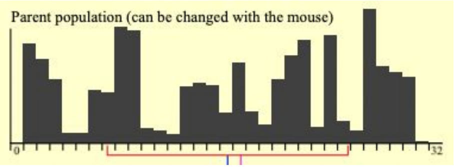
Demo:https://onlinestatbook.com/stat_sim/sampling_dist/

Sampling Distribution of the SAMPLE MEAN

Population Distribution

The underlying distribution

(usually unknown)



parameter

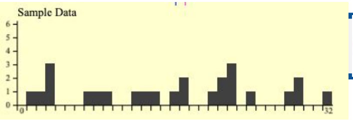
Population mean = 16.06

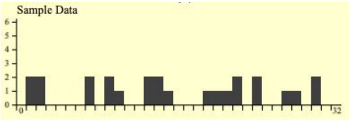
Population std = 9.22

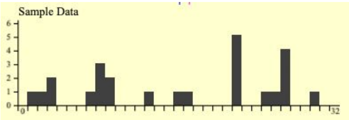
Parameters are **numbers** i.e., fixed values

Sample Distribution(s)

(aka the histogram of your data)







Sample **statistic**

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

Sample **statistic**

Sample **statistic**

Sampling Distribution of a Statistic

Calculate MEAN from all possible samples and create histogram

1st Sample mean= 16

2nd Sample mean=15

3rd Sample mean = 16.5

...

What does the sampling distribution of the sample mean look like?

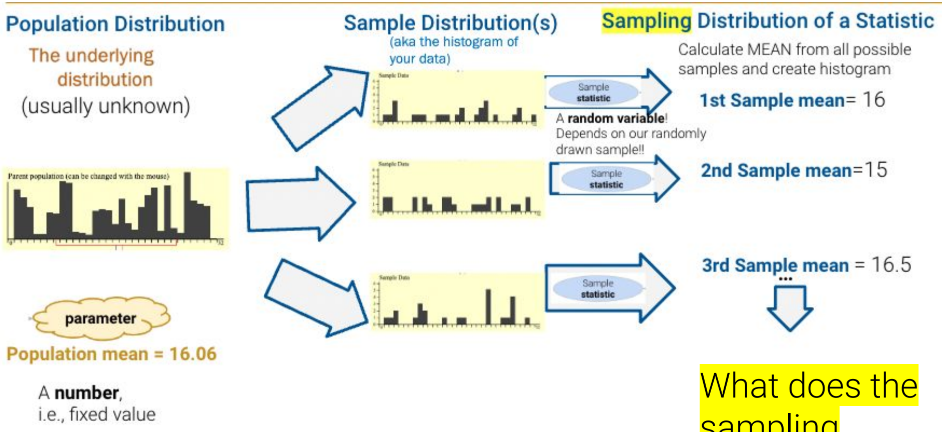
Sampling Distribution of the SAMPLE MEAN

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

We want to know HOW off is it?

To answer this we need to know:

1. What distribution does the sample mean have?
2. What is the expected value of the sample mean?
3. What is the variance of the sample mean?
4. What is the **standard error** of the sample mean?



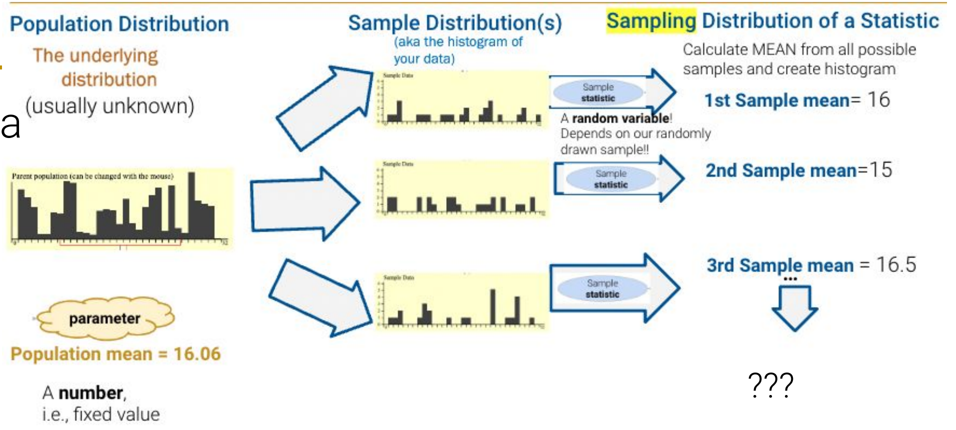
What does the sampling distribution of the sample mean look like?

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

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



Thus the sample mean is a RANDOM VARIABLE

1. What distribution does the sample mean have?
2. What is the expected value of the sample mean? $$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}(n\mu) = \boxed{\mu}$$
3. What is the variance of the sample mean? $$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\underbrace{\left(\sum_{i=1}^{n}\mathrm{Var}(X_i)\right)}_{\text{IID} \to \mathrm{Cov}(X_i, X_j) = 0} = \frac{1}{n^2}\left(n\sigma^2\right) = \boxed{\frac{\sigma^2}{n}}$$
4. What is the **standard error** of the sample mean? $$\mathrm{SD}(\bar{X}_n) = \boxed{\frac{\sigma}{\sqrt{n}}}$$

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

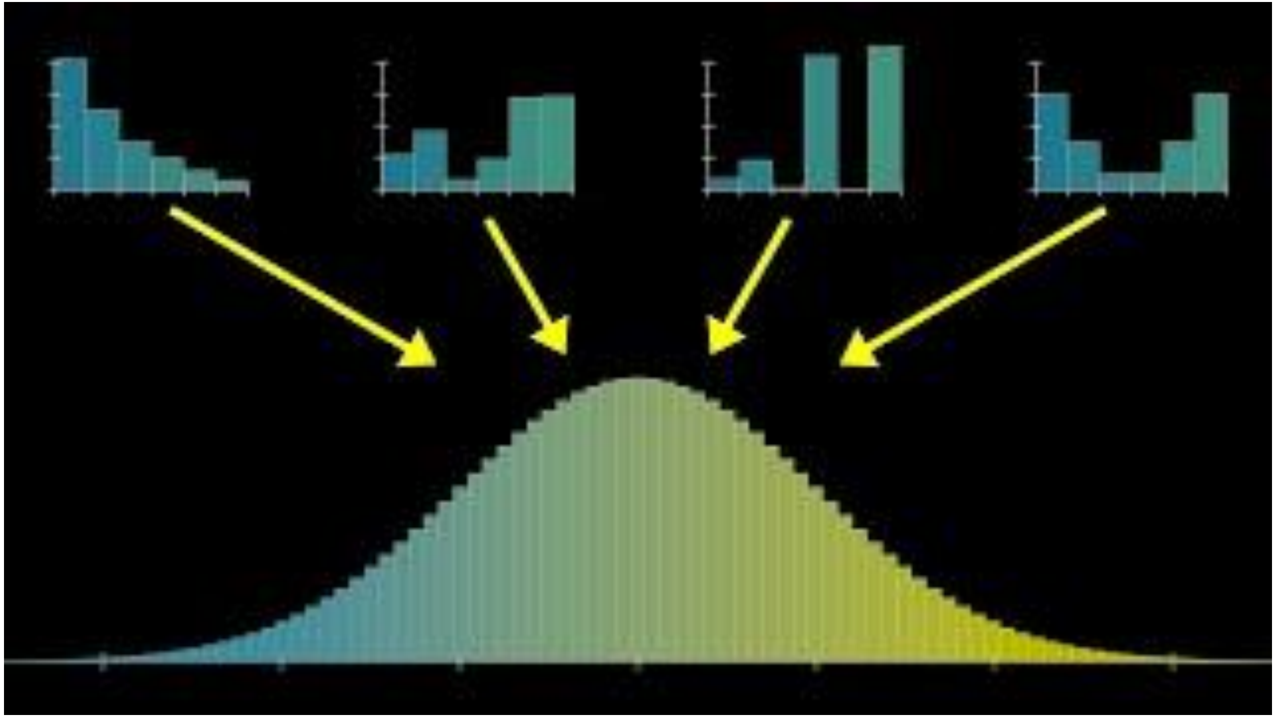
The Central Limit Theorem

CSCI 3022

- Parameters vs Statistics
- Sampling Distributions of Statistics
- **Central Limit Theorem**

Demo

Central Limit Theorem



3Blue1Brown Video:

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

The Central Limit Theorem (CLT)

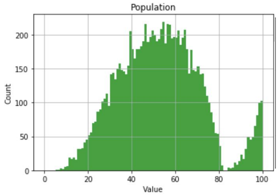
No matter what population you are drawing from:

Let $X_1, X_2, \ldots, X_n$ iid, where $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$. As $n \to \infty$:

Sampling Distribution of the Statistic:

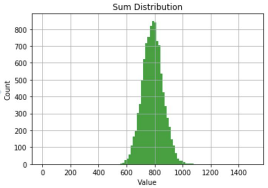
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

Sum of iid RVs



Distribution of X_i

Sample of size 15, sum values

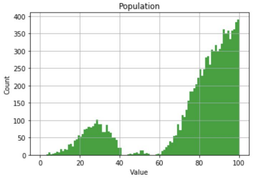


Distribution of $\sum_{i=1}^{15} X_i$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Average of iid RVs (sample mean)

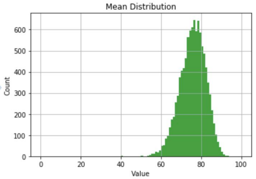
(so also works with sample proportions!)



Distribution of X_i

Sample of size 15, average values

(sample mean)



Distribution of $\frac{1}{15}\sum_{i=1}^{15} X_i$

Any theorem that provides the rough sampling 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!

How Large Is “Large”?

No matter what population you are drawing from:

Consider an IID sample X_1 , X_2 , ..., X_n

drawn from a population with **mean μ and SD σ**.

If n is large, the probability distribution

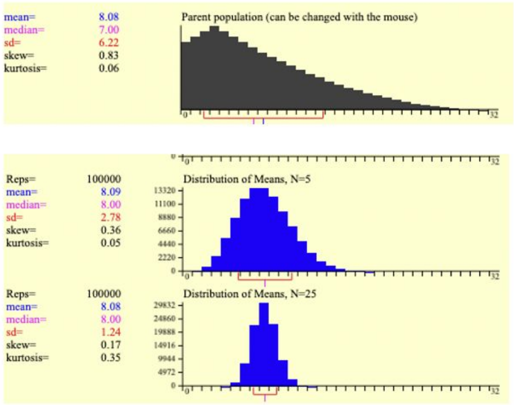
of the **sample mean** is **roughly normal**:

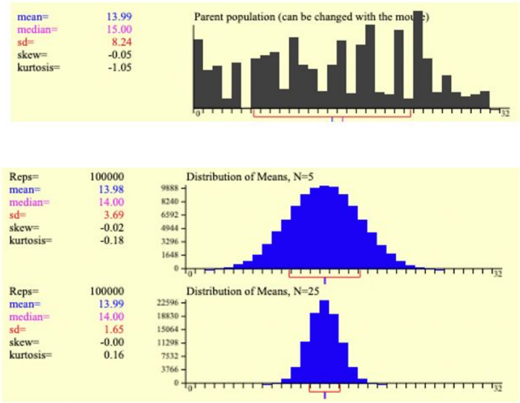
As $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

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 and see if the bootstrapped distribution is bell-shaped.





TRUE or FALSE?

A). No matter what population you are drawing from, the sample distribution is roughly normal (for large enough n).

B). No matter what population you are drawing from, the sampling distribution of the sample mean is roughly normal (for large enough n)

C). No matter what population you are drawing from, the sampling distribution of the sample median is roughly normal (for large enough n)

D). If you are drawing from a Bernoulli distribution, the sampling distribution of the sample proportion is roughly normal (for large enough n)

Discussion Question

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

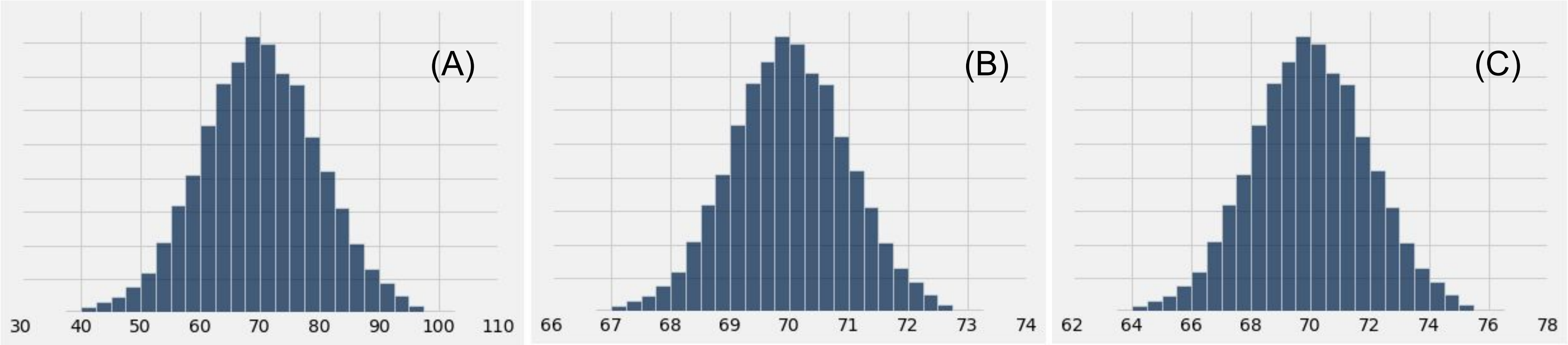
a). If a single employee is randomly selected, what is the probability that their salary exceeds \$175,000?

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

```
1-stats.norm.cdf(175000, 162000, 3200)
```

Practice Question

A population distribution has an average of 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 and why?



Practice 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?

Practice 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!

Appendix

CSCI 3022

Proof of CLT (beyond scope of CSCI 3022):

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{As } n \to \infty$$

The sum of n iid random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.

Proof:

- The Fourier Transform of a PDF is its characteristic function.
- Take the characteristic function of the probability mass of the sample distance from the mean, divided by standard deviation
- Show that this approaches an exponential function in the limit as $n \to \infty$: $f(x) = e^{-\frac{x^2}{2}}$
- This function is in turn the characteristic function of the Standard Normal, $Z \sim \mathcal{N}(0,1)$.