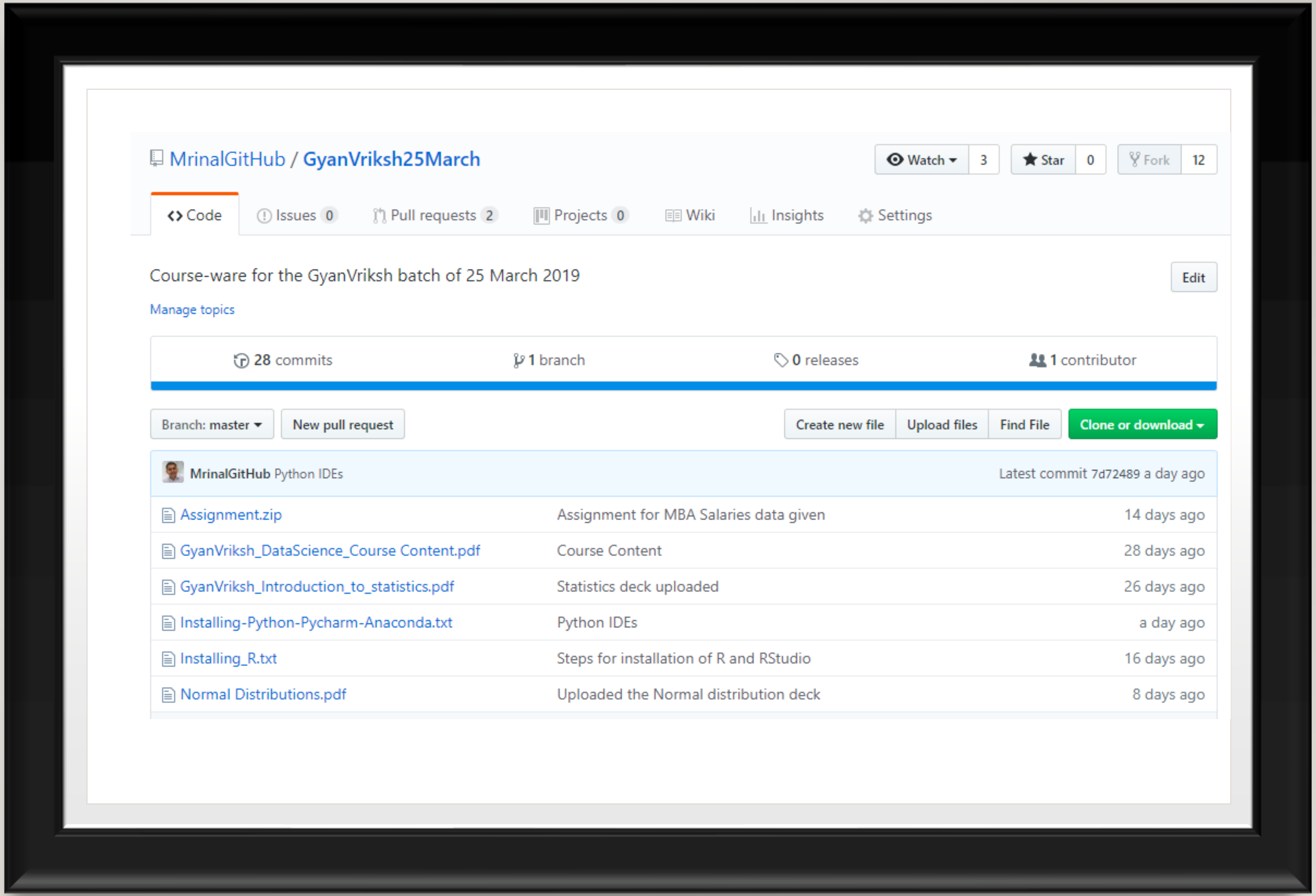
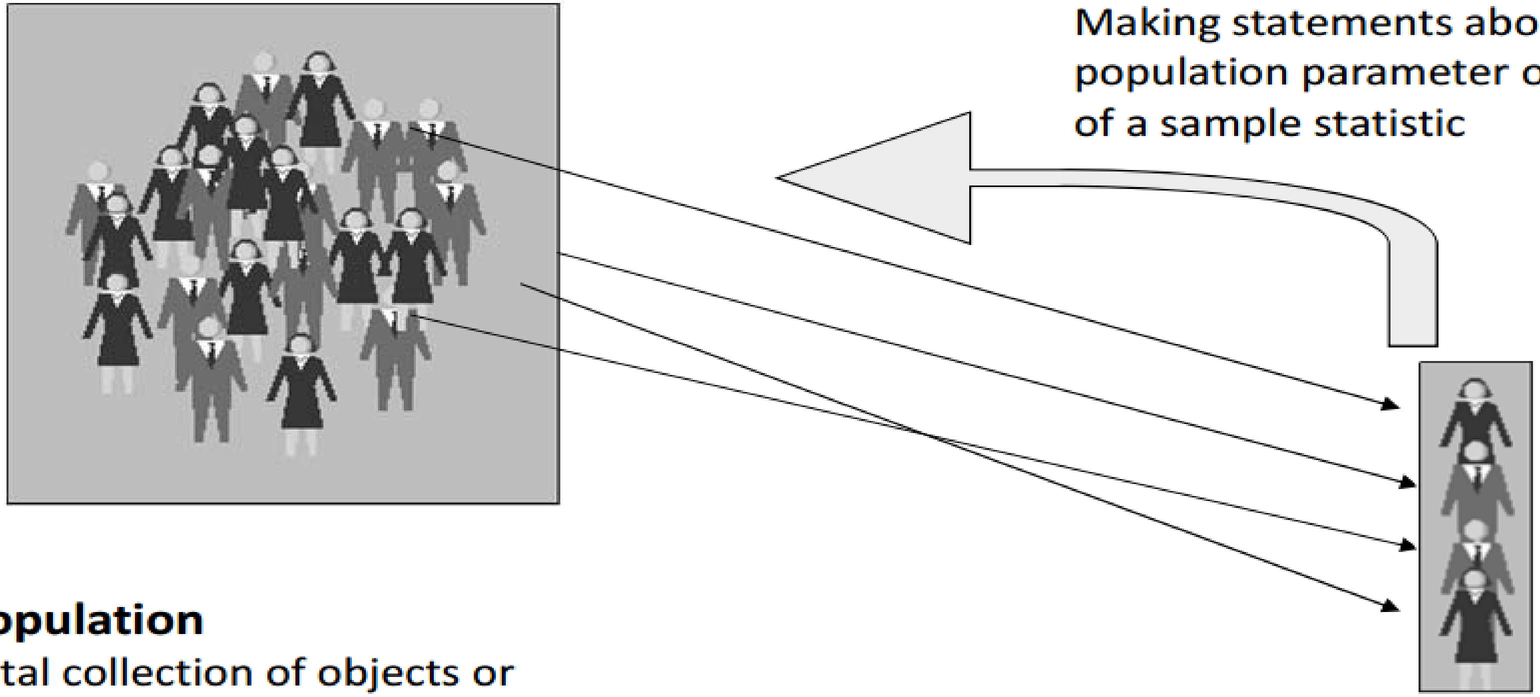
SAMPLING DISTRIBUTIONS AND THE CENTRAL LIMIT THEOREM



Learning objectives

- What is statistical inference?
- How to (and how not to) choose a sample?
- What are sample statistics and their properties?
- What is the central limit theorem and how is it useful?

Statistical Inference



Statistical Inference
Making statements about a population parameter on the basis of a sample statistic

Population
Total collection of objects or people to be studied (or set of all information of interest to the decision maker)

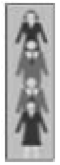
Sample
Subset of a population

Typical Pitfalls in Sampling

- Collecting data only from volunteers (voluntary response sample)
  - e.g. online reviews (yelp.com, maps.google.com, tripadvisor.com)
- Picking easily available respondents (convenience sample)
  - e.g. choosing to survey in In-Orbit mall
- A high rate of non-response (more than 70%)
  - e.g. CEO / CIO surveys on some industry trends

Sample statistics and population parameters

- A sample statistic is a characteristic of the sample
- Some sample statistics might be used as a point estimate for a population parameter
- We use different notations to distinguish between the two groups of numbers



Population Parameter		Sample Statistic
μ	Mean	$\bar{x}$
σ^2	Variance	s^2
π	Proportion	p

Selecting a Simple Random Sample (SRS)

- Unbiased: Each unit has equal chance of being chosen in the sample
- Independent: Selection of one unit has no influence on selection of other units
- SRS is a gold standard against which all other samples are measured

Selecting the Sampling Frame

- Sampling frame is simply a list of items from which to draw a sample
- Does the sampling frame represent the population?
  - e.g. Literary Digest vs. George Gallup polls
- The available list may differ from desired list
  - e.g. we don't have list of customers who did not buy from a store
- Sometimes, no comprehensive sampling frame exists
  - e.g. when forecasting for the future. Thus a comprehensive list of acceptances of credit card offers does not exist yet

Example

- What is the average work experience of all participants of the BA course?

	Sample 1	Sample 2
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
Sample Mean		

Central Limit Theorem (CLT) & the distribution of the sample mean

- The distribution of the sample mean
  - will be normal when the distribution of data in the population is normal
  - will be approximately normal even if the distribution of data in the population is not normal, if the sample size is "fairly large"

- Mean ($\bar{X}$)= μ (the same as the population mean of the raw data)

- Standard deviation (X)= $\frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation and n is the sample size
  - This is referred to as Standard Error of the Mean

Activity: http://www.socr.ucla.edu/htmls/SOCR_Experiments.html

CLT is Valid When…

- Each data point in the sample is independent of the other
- The sample size is large enough

Suppose salaries at a very large corporation have a mean of \$62,000 and a standard deviation of \$32,000.

If a single employee is randomly selected, what is the probability their salary exceeds \$66,000?

Suppose salaries at a very large corporation have a mean of \$62,000 and a standard deviation of \$32,000.

If 100 employees are randomly selected, what is the probability their average salary exceeds \$66,000?

Example and Resource material:

T-table sample link : http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm

Refer page 24:
https://www.saylor.org/site/wp-content/uploads/2012/03/Introductory-Business-Statistics.pdf

Summary of Session

- Statistical inference is the process of making probabilistic inferences about population parameters based on sample statistics
- Simple random sample is the gold standard and it requires a sampling frame that represent the population and a randomization device
- Sample statistics are random variables because they vary across samples drawn from the same population. They can be used as point estimates of the population parameter.
- Central limit theorem states that, no matter what the population distribution is, the sample mean ($\bar{X}$) is normally distributed with mean (μ) and standard error $\left(\frac{\sigma}{\sqrt{n}}\right)$, approximately