Sampling

Joe

Introduction

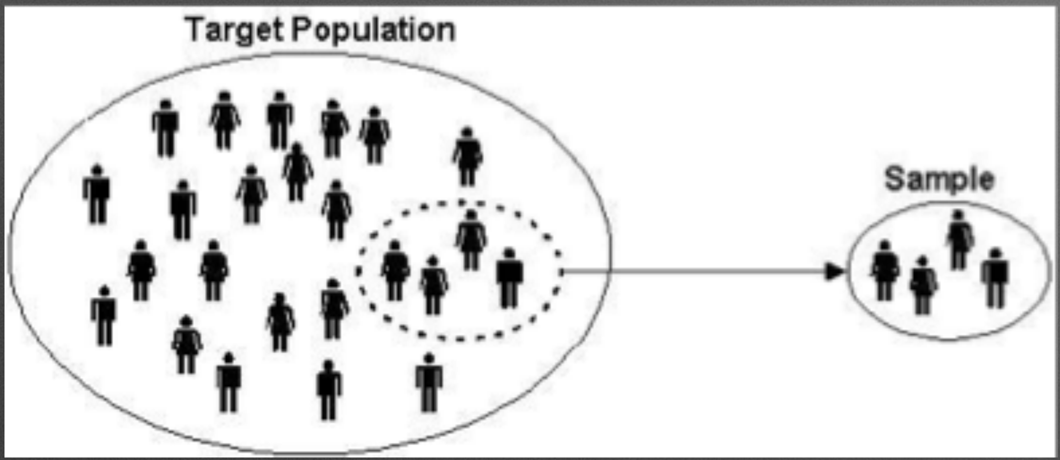
Session Objective

1. Enumerate the ways that samples can introduce bias into an analysis

2. Use sampling and the central limit theorem to measure the mean of several distributions

Sampling Fundamentals

Sampling



The collection of data is associated with costs of time, effort, and money.

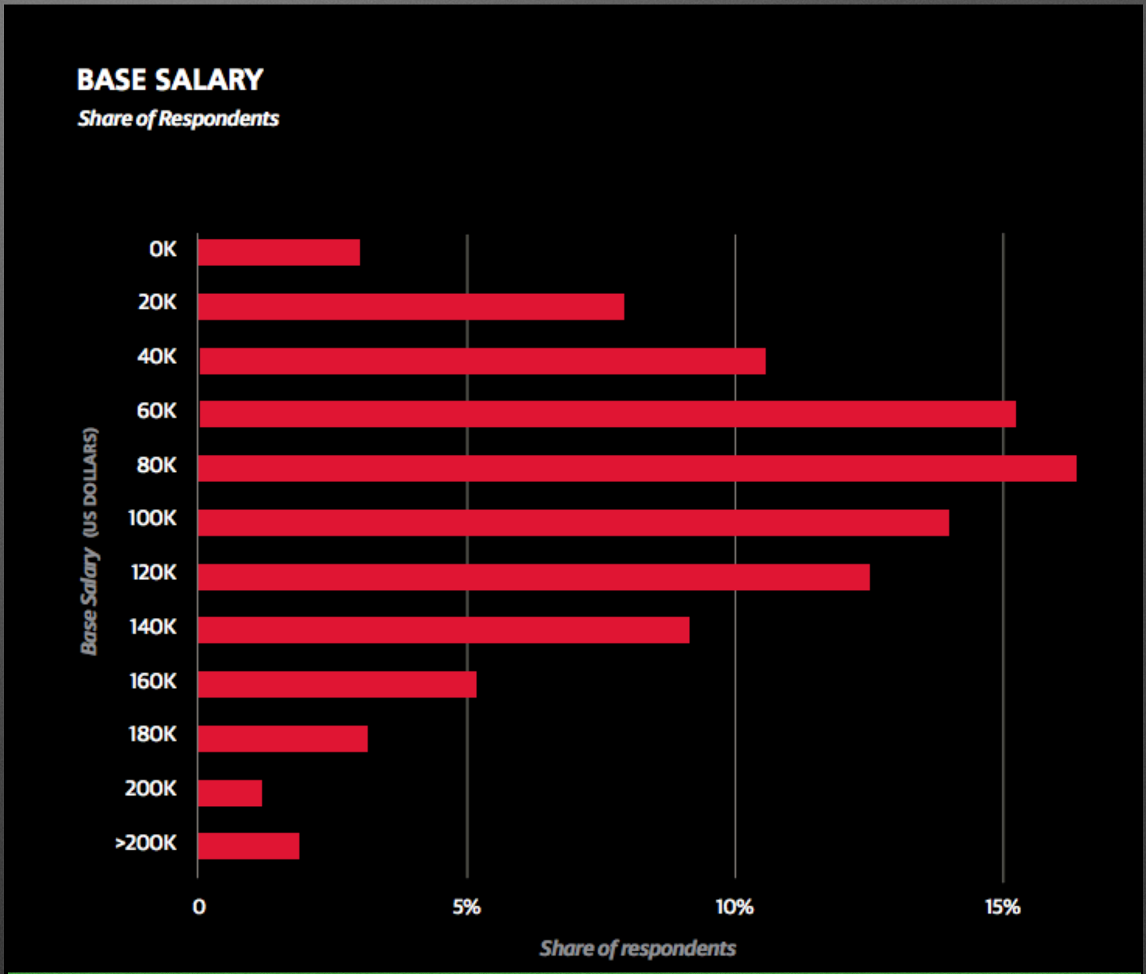
Samples are taken to garner information about a population, and statistics are used to infer properties of the population based on the sample.

The mean of a sample of a given population is a function of randomly distributed variables. As such <u>for a random sample, properties such as the mean and variance are themselves random variables</u>!

Random Sampling

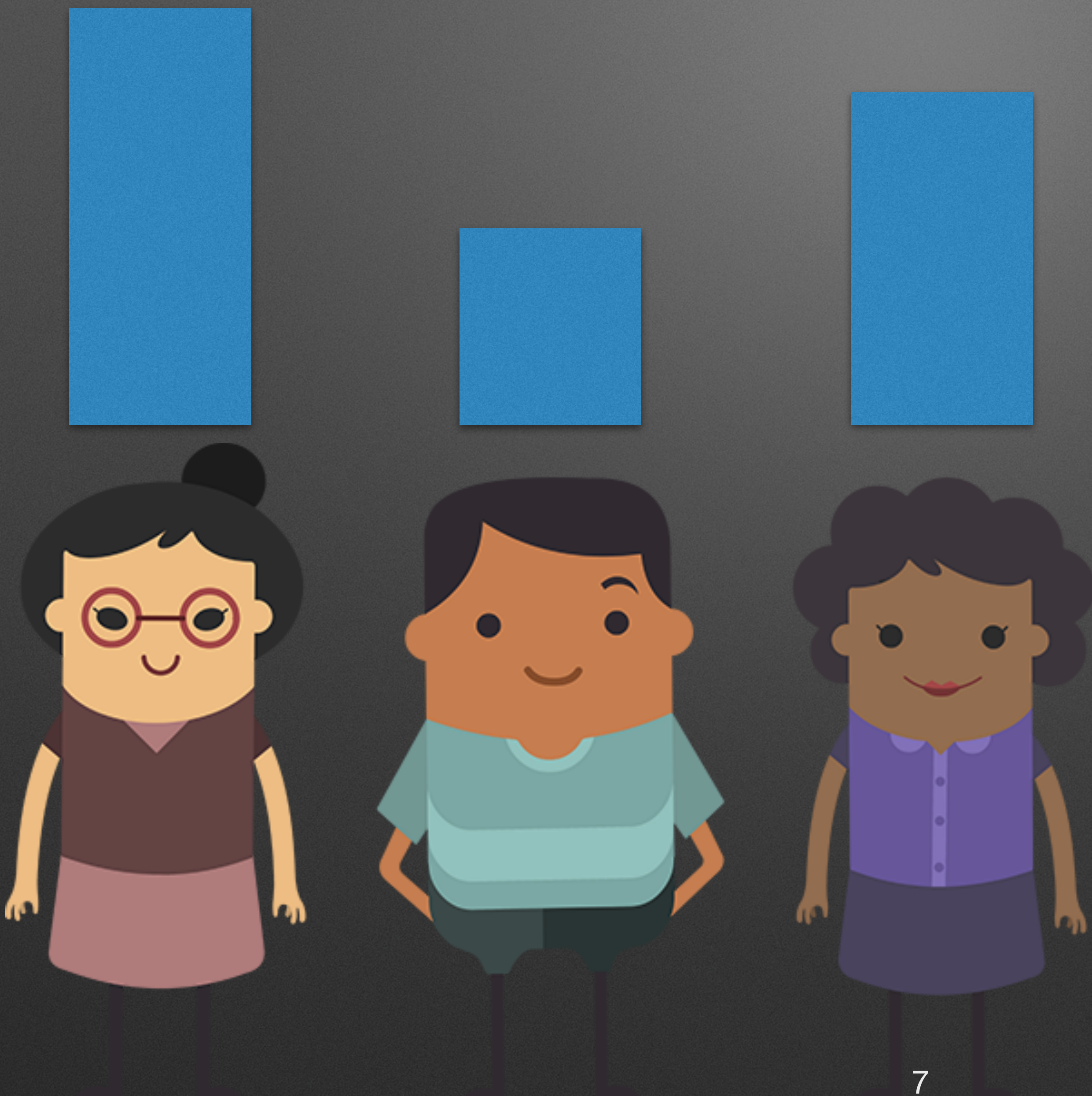
Random Sampling aims to have representative samples of the distribution

Particularly when humans are involved, this is incredibly difficult



Source: O'Reilly Data Science Salary Survey - is there any reason to be suspicious of these results?

Pseudorandom Sampling



Pseudorandom sampling is the process where samples are deliberately shaped so their distribution matches know distributions:

1. Scaling
2. Selective Sampling

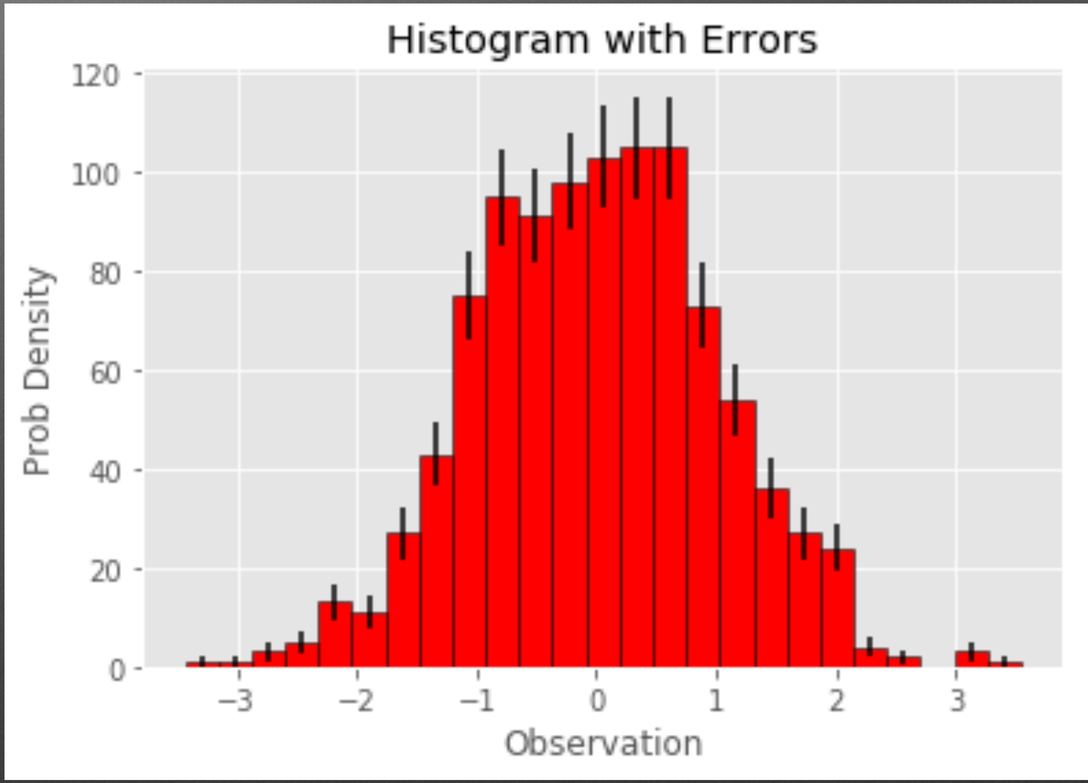
An important note - if the sample size is large w.r.t. the population size, a sample is no longer independent.

Counting Errors

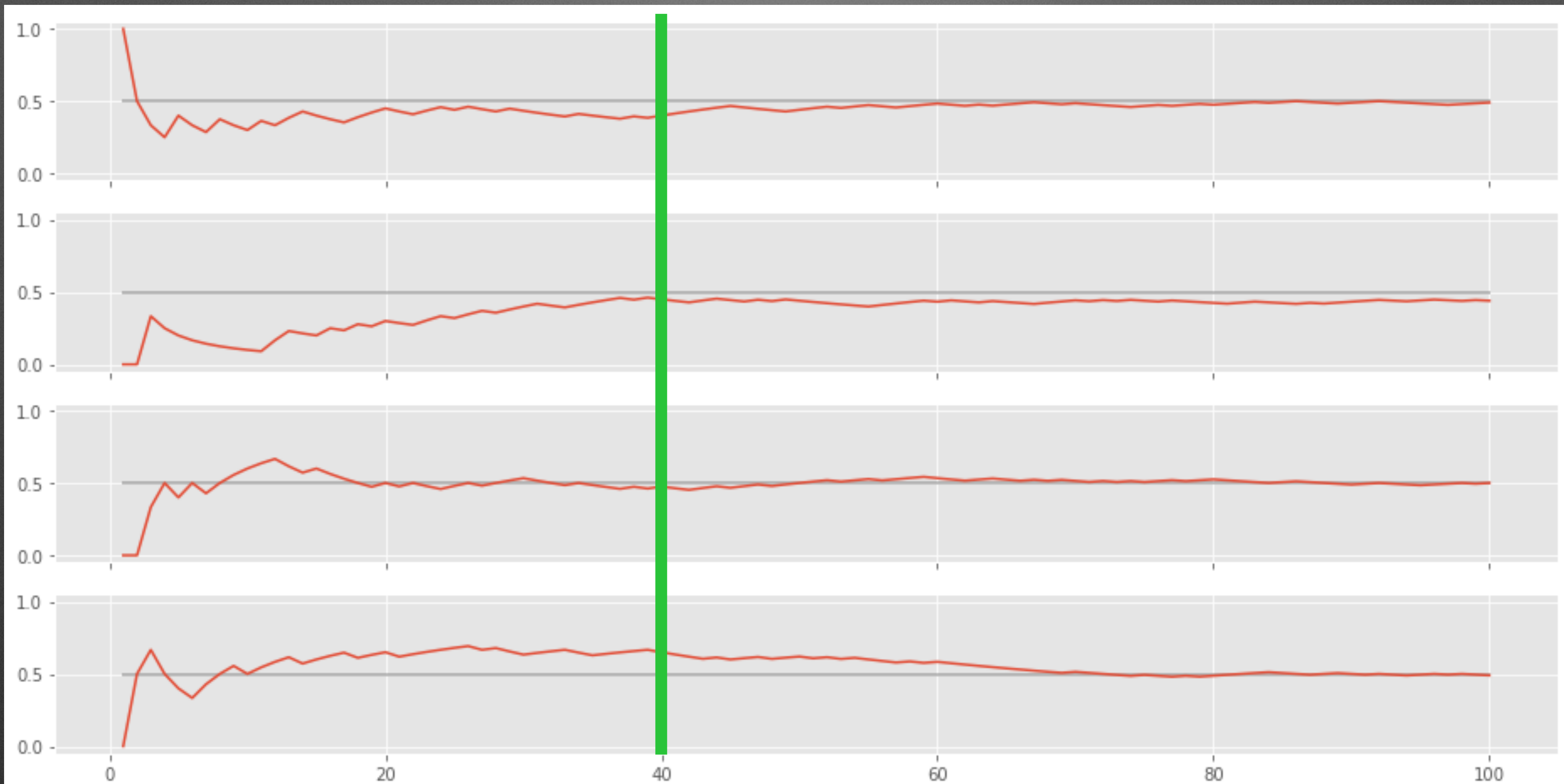
Experiments using histograms are referred to as 'counting experiments', and the in/out of bin error can be modeled as a binomial process. Hence :

$$\sigma_k^2 = Np_k(1-p_k)$$

$\sigma_k = \sqrt{n_k}$ is valid for $n_k > 10$



Law of Large Numbers



Look at the sample mean for flipping a fair coin. The law of large numbers stats that for sufficiently large samples, the sample mean converges to the population mean

Central Limit Theorem

Central Limit Theorem

Recall - the mean of a sample is a randomly distributed variable.

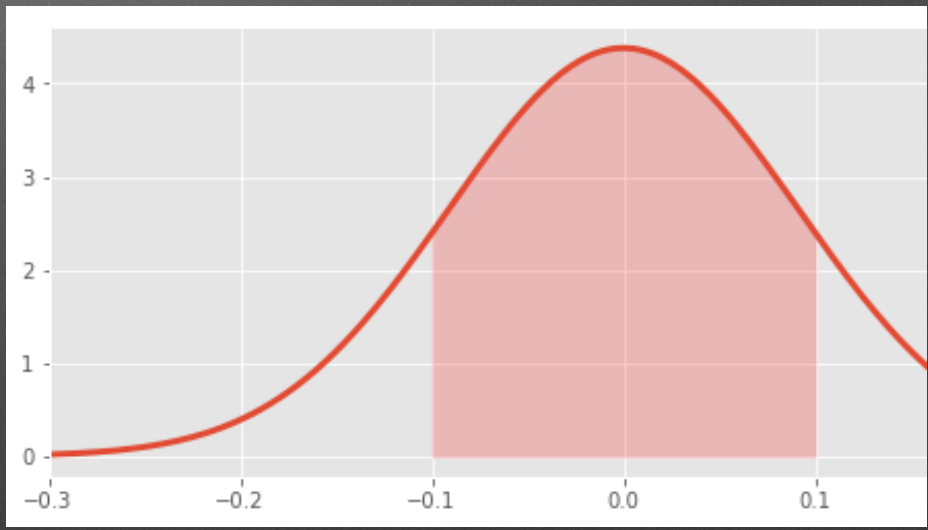
There will be some difference between the sample mean, and the population mean; for sufficiently large samples (typically $n>30$) a sample will be normally distributed about the population mean, even if the population is not normally distributed.

This profound observation is a cornerstone of statistics, let's hop into a notebook to have a look.

Confidence Intervals

Confidence Intervals

Suppose you want your partner to meet you after the morning lecture, and you want to describe an interval in time when you anticipate being done.

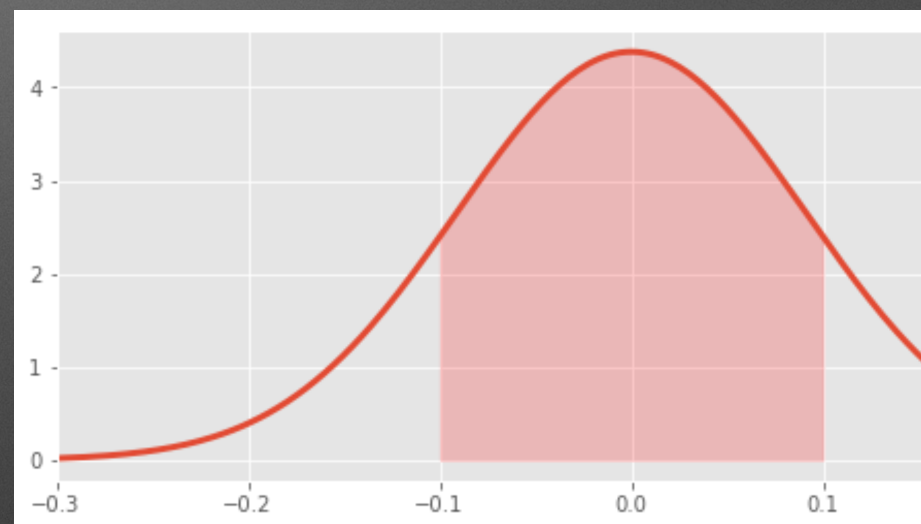


Confidence Intervals

The CLT allows us to estimate probability regions by modeling averages and their error bars

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

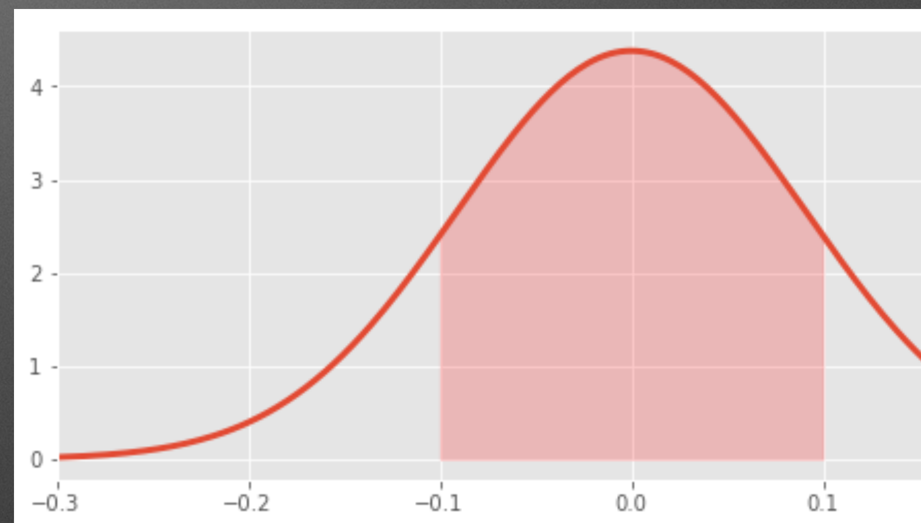
$$\bar{x} - \mu \sim N\left(0, \frac{\sigma}{\sqrt{n}}\right)$$



$$Pr\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Confidence Intervals

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \longleftrightarrow \bar{x} - \mu \sim N\left(0, \frac{\sigma}{\sqrt{n}}\right)$$



Those correction factors come from the integral of the normal distribution to the 95% level. Use a different correction for a different significance level.

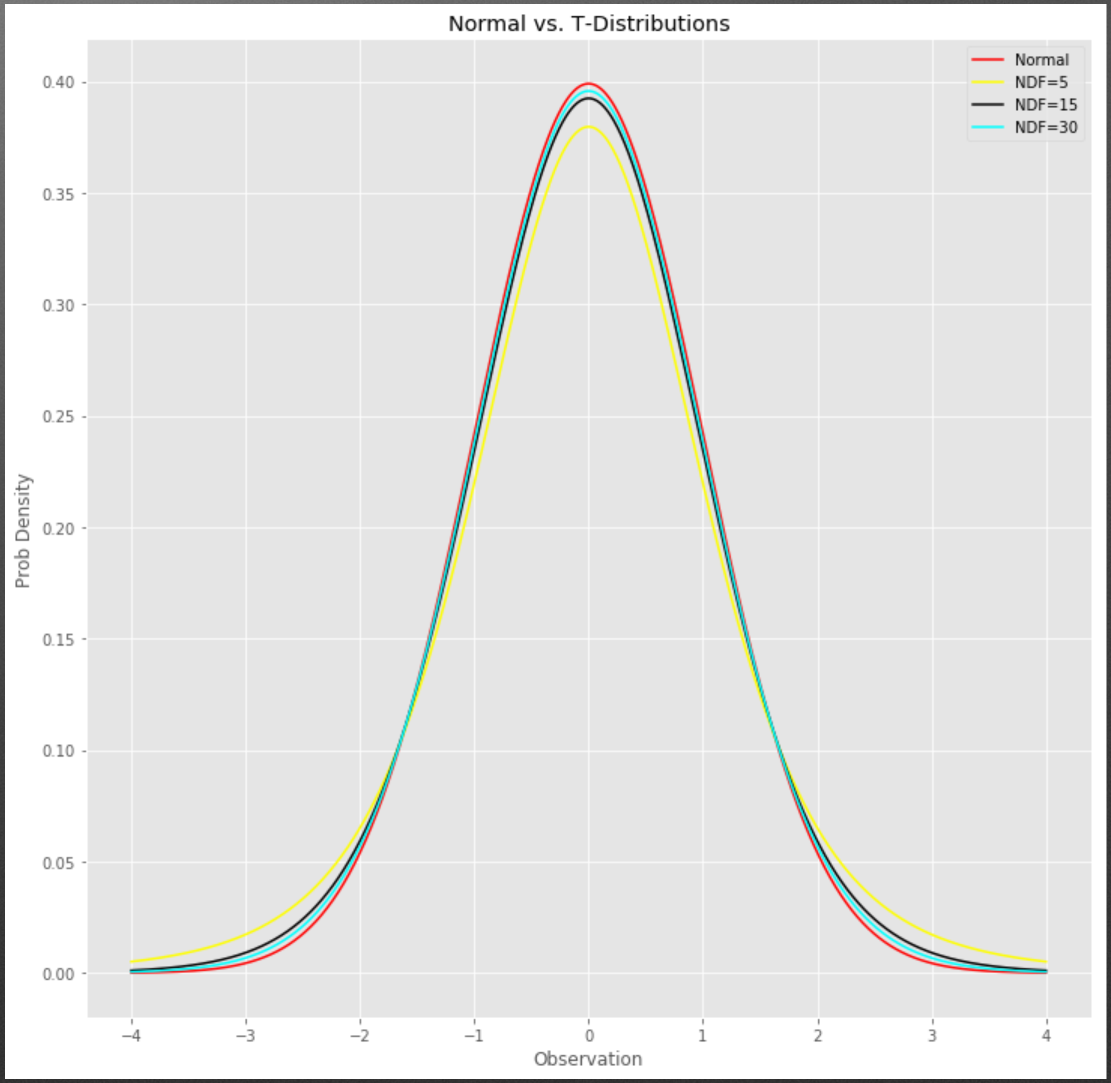
$$Pr\left(\bar{X} - \boxed{1.96}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + \boxed{1.96}\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

```
In [12]: 2*(1-stats.norm.cdf(1.959))
Out[12]: 0.05011278678472197 4
```

T-Distribution

CLT says that for large N, samples are ‘approximately’ normal.

For low N samples, we must use the ‘Student T’ distribution instead, which becomes <u>roughly</u> normal at n=30.



Boot Strapping

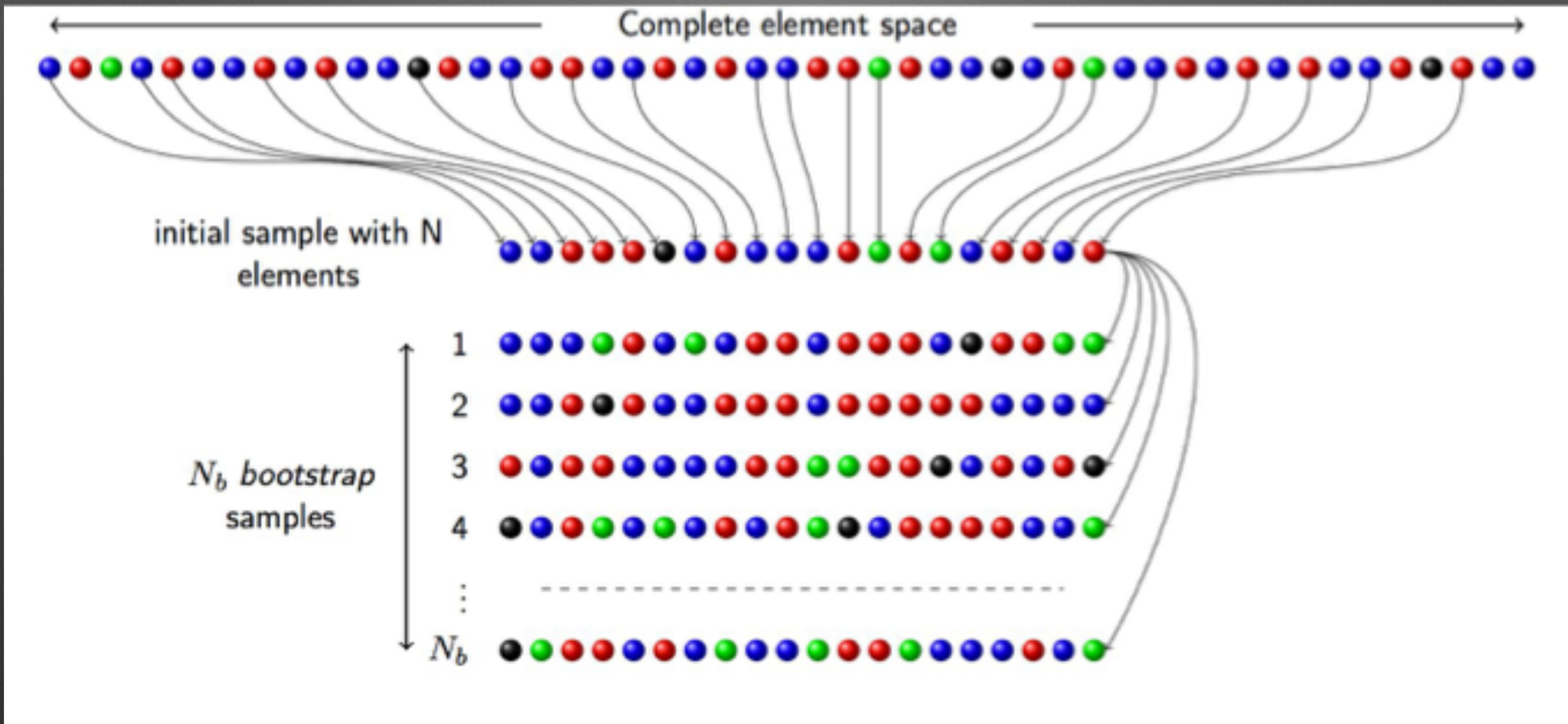
Bootstrapping

Bootstrapping is a method used in statistics and machine learning to increase the predictive power of a sample.

Procedure:

1. Start with a sample of size n
2. Sample from your dataset with replacement to create one bootstrap sample of size n
3. Repeat B times
4. Each bootstrap sample can then be used as a separate dataset for estimation using the CLT

Bootstrapping Visualized



Bootstrapping Best Practices

Use bootstrapping when :

1. The theoretical distribution is complicated
2. The sample size is too small
3. Favor accuracy over computational cost

Session Objective

1. Enumerate the ways that samples can introduce bias into an analysis

2. Use sampling and the central limit theorem to measure the mean of several distributions