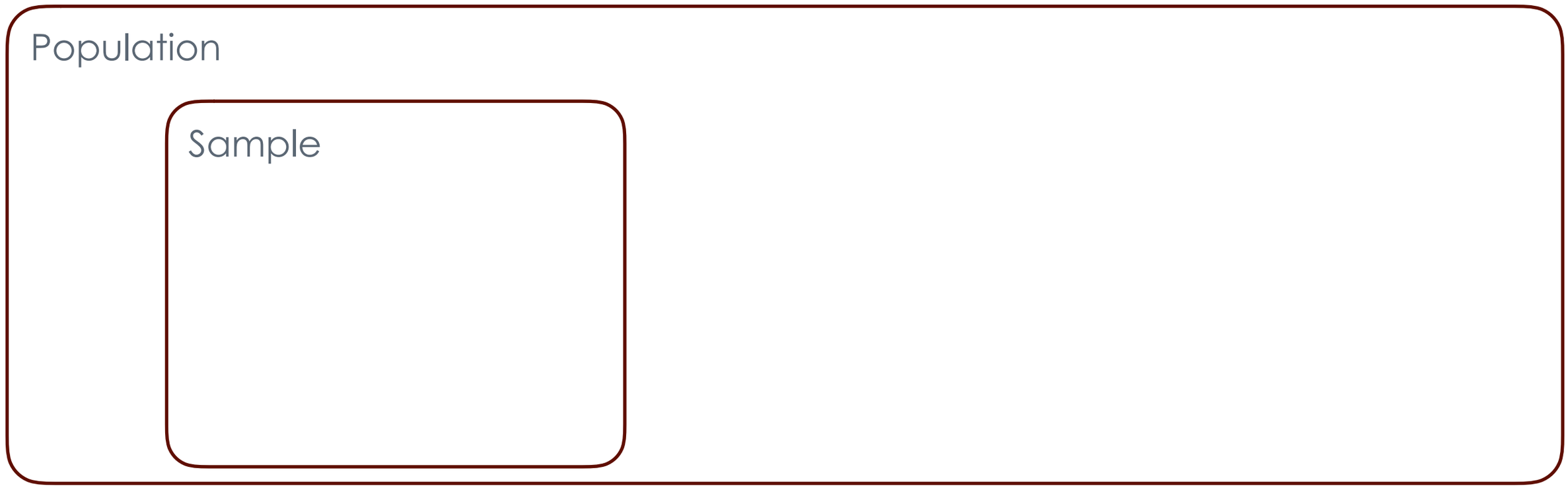
Much Ado About Sampling

CHIDATA

CHIDATA

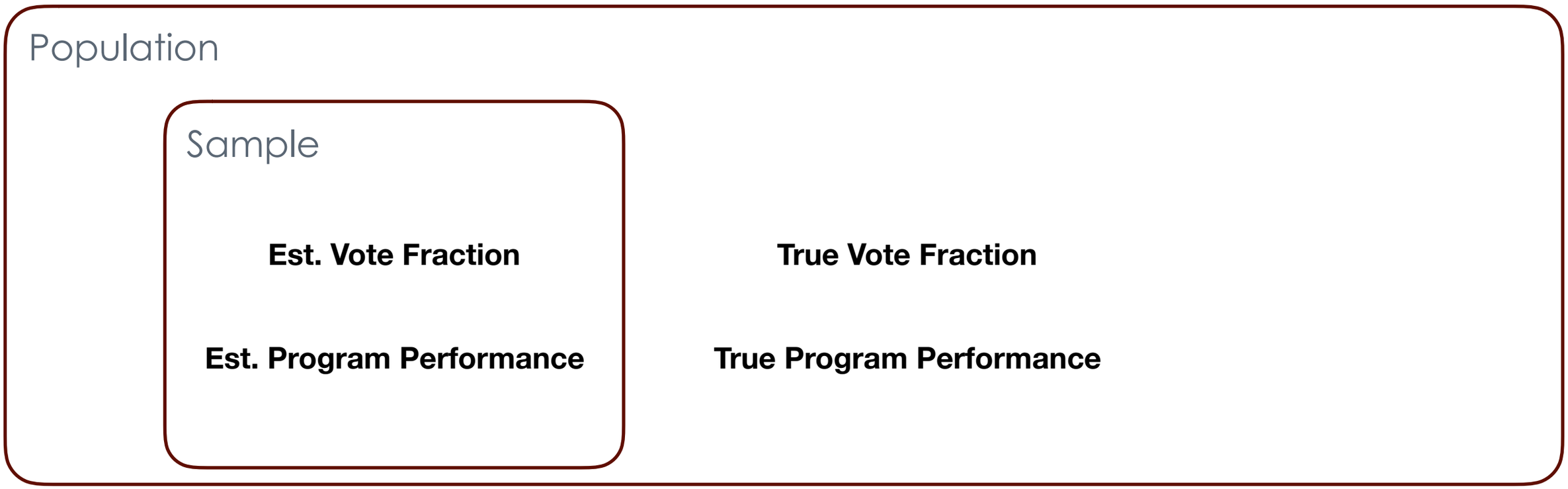
Recap



You have a subset of relevant facts called a **sample.**

The sample is derived from a closed-world called a **population.**

Recap



How well do “statistics” calculated on the sample align with the true values calculated over the population?

CHIDATA

Recap: Sampling Processes

Two axes: distribution and independence

Uniform v.s. Non Uniform: Does every element of the population have an equal probability of appearing in the sample.

Independent v.s. Dependent: Does observing one sample inform you about the likelihood of another sample.

CHIDATA

Today: Sampling Processes

Classical Sampling with replacement

Uniform v.s. Non Uniform: Does every element of the population have an equal probability of appearing in the sample.

Independent v.s. Dependent: Does observing one sample inform you about the likelihood of another sample.

CHIDATA

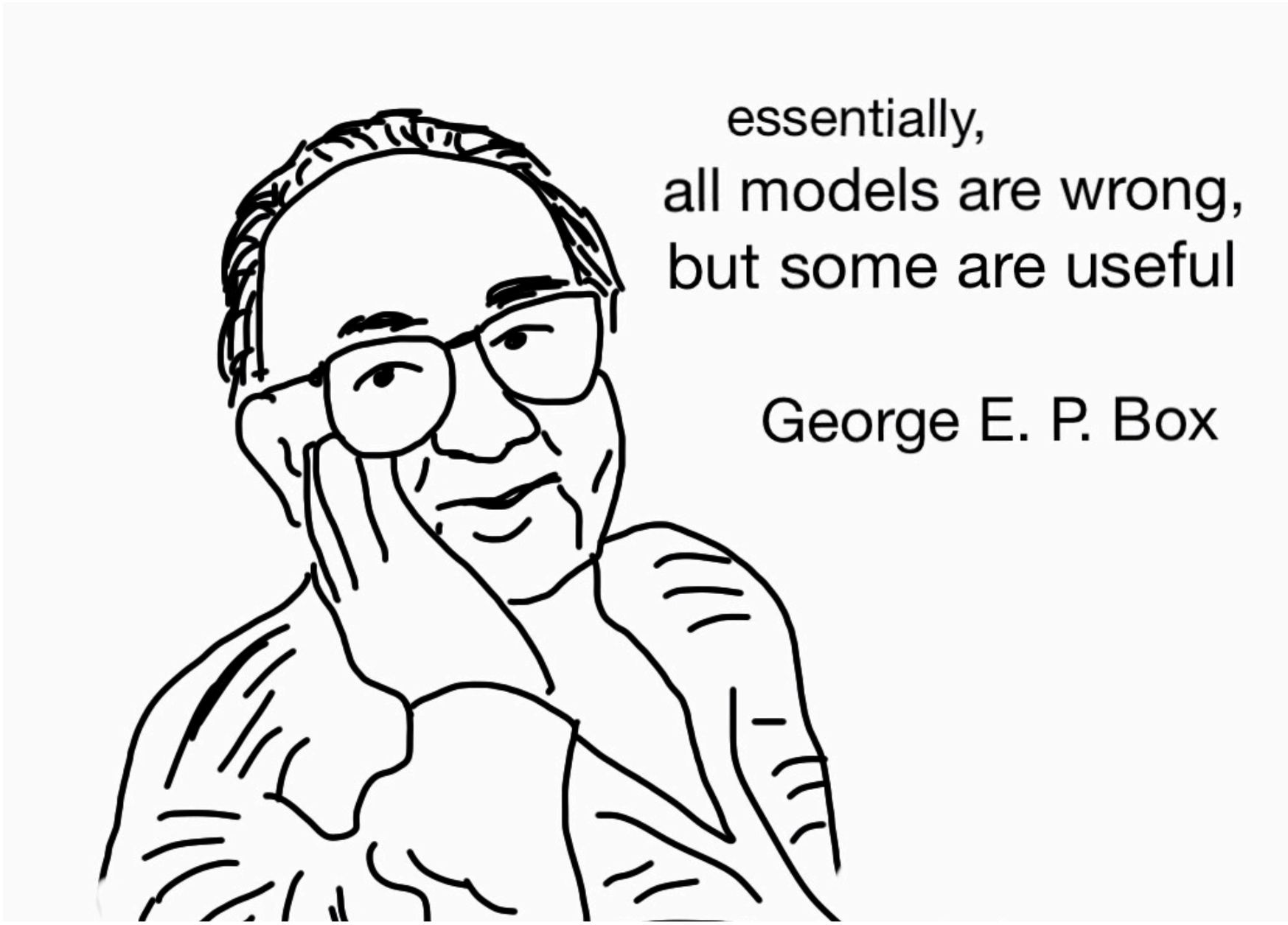
Uniform Sampling With Replacement

Population

```
For 1…K:
 Pick an element from the population with equal prob
```

CHIDATA

Uniform Sampling With Replacement



CHIDATA

Practical Experiments?

Real world processes often don't give you duplicates:

- Clinical trials
- Opinion polls
- Natural experiments

"Simple" Random Sampling: *A reasonable approximation if the samples are mostly independent and the sample size is small compared to the population.*

CHIDATA

Probability Every Sampled Elem is Unique?

Population

K : Size of the sample
N : Size of the population

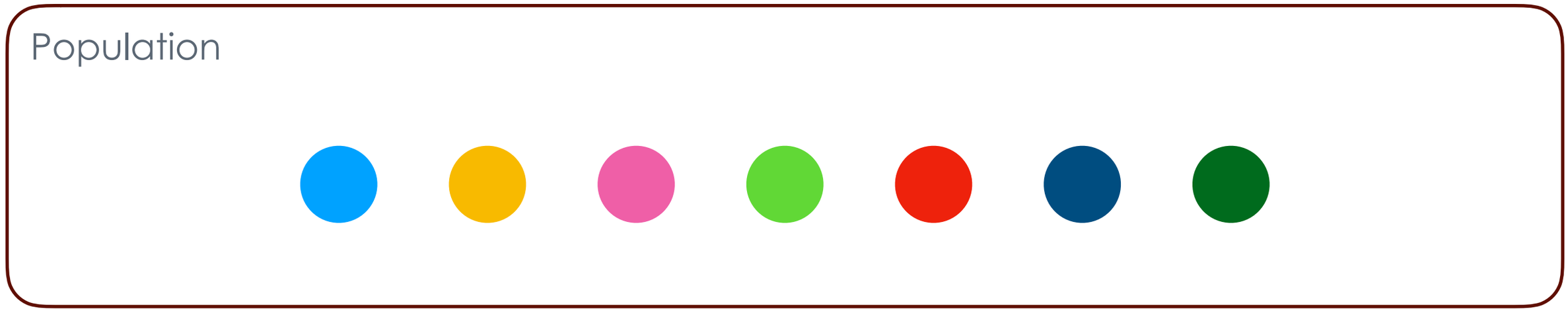
$$P(\text{unique}) = 1 \cdot \frac{N-1}{N} \cdot \frac{N-2}{N} \cdot \ldots \cdot \frac{N-(K-1)}{N}$$

$$P(\text{unique}) \geq (\frac{N-K}{N})^K$$

Prob that it's indistinguishable from sampling w/o replacement

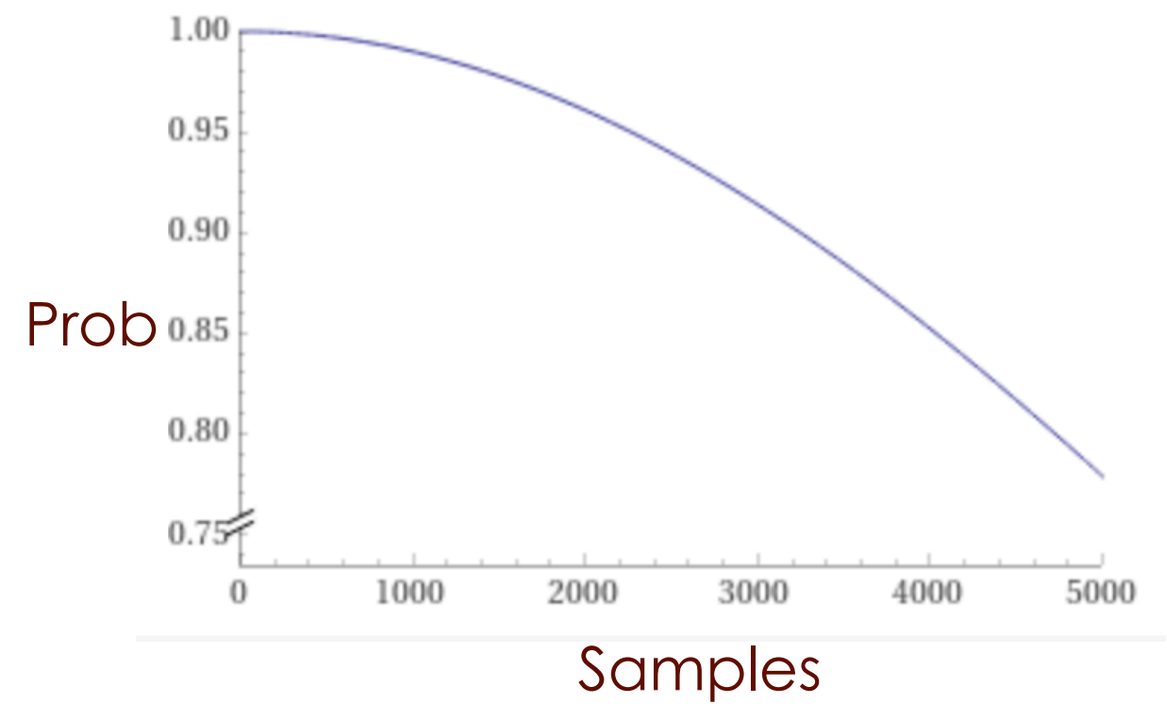
All elements will prob be unique is the sample is small relative to the population

Probability Every Sampled Elem is Unique?




```
K : Size of the sample
N : Size of the population
```

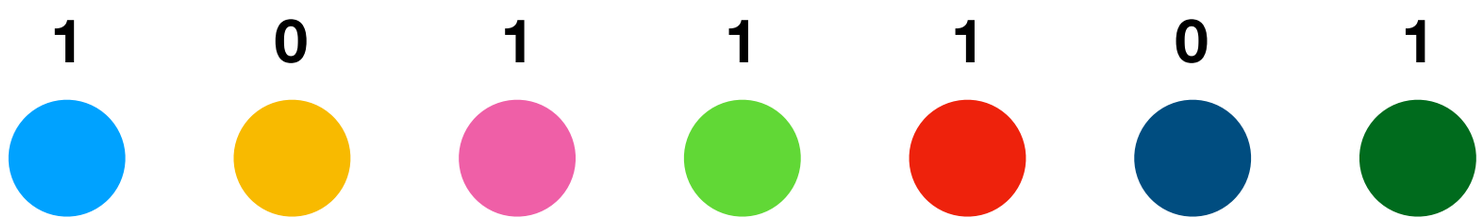
Curve for typical voting population



CHIDATA

Populations Define Numeric Distributions

Yes/No Vote

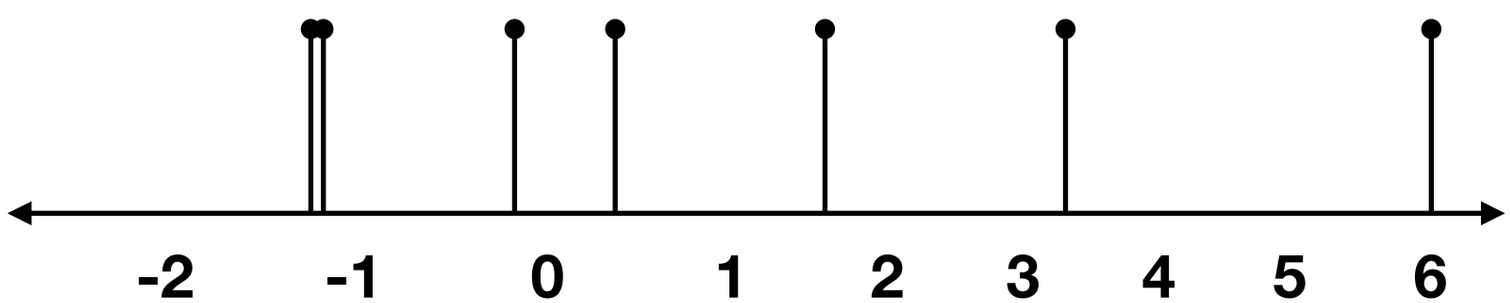


Probability of numeric value if you draw uniformly at random

Populations Define Numeric Distributions

Weight loss after treatment

Even though the values are continuous, the "distribution" is still discrete for a **finite** population!

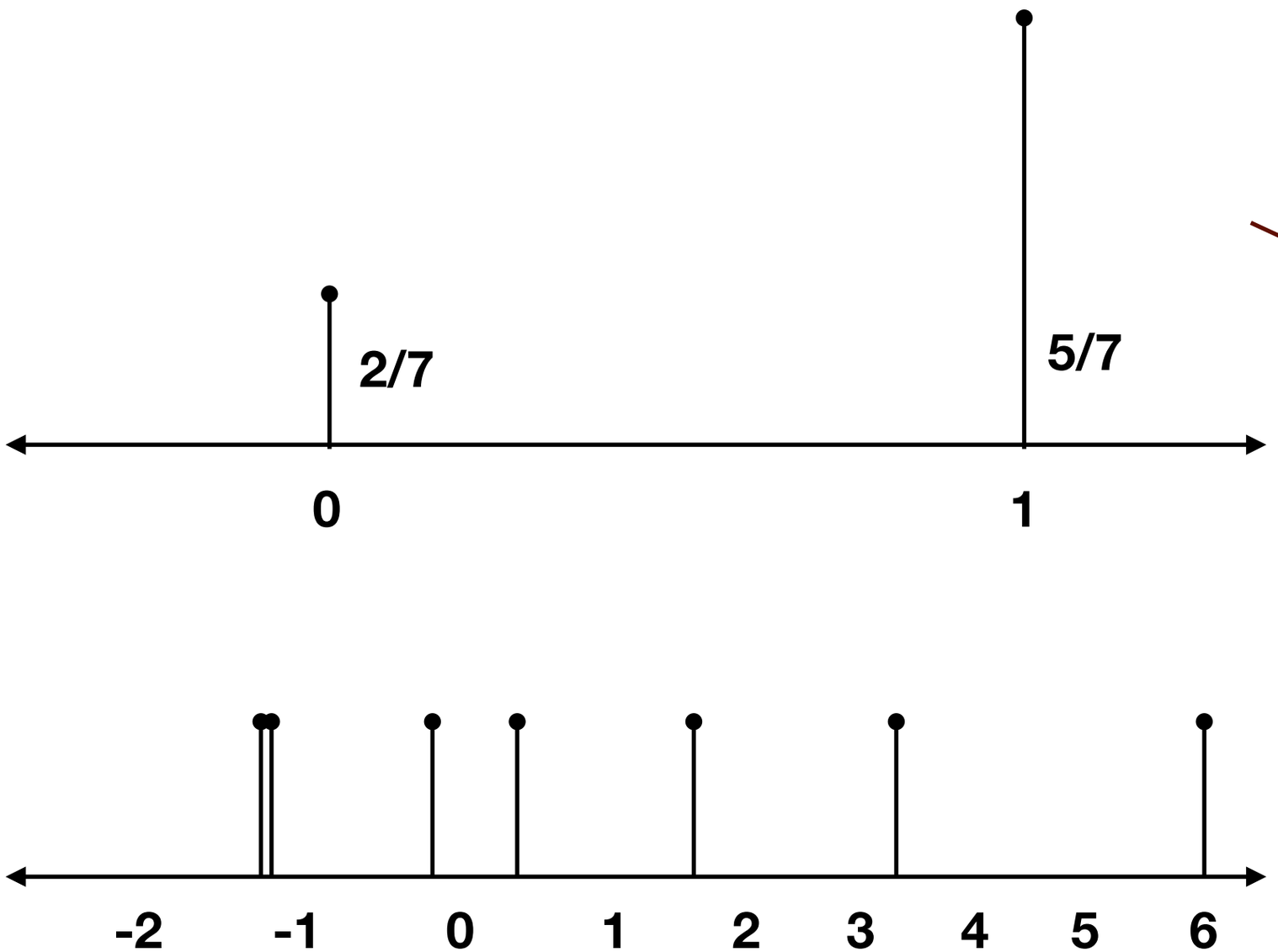


Probability of numeric value if you draw uniformly at random

CHIDATA

Populations Define Numeric Distributions

Probability of numeric value if you draw uniformly at random

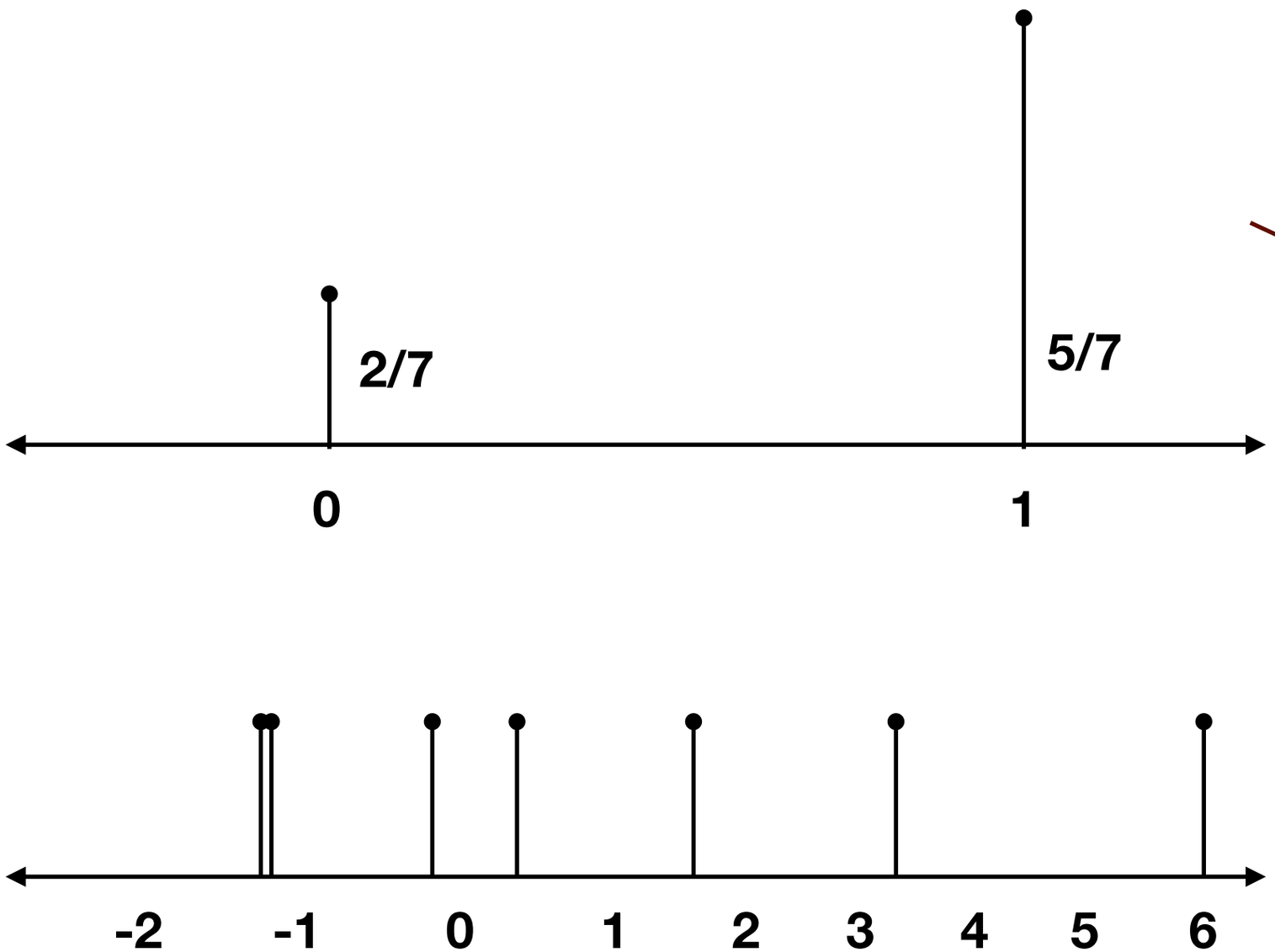


Each “draw” is a random variable

CHIDATA

Populations Define Numeric Distributions

Probability of numeric value if you draw uniformly at random



X is the value
of a randomly chosen sample

CHIDATA

Uniform Sampling With Replacement

Population

```
# Code()
For I in range(K):
  X[i] = random.choice(popl).value
```

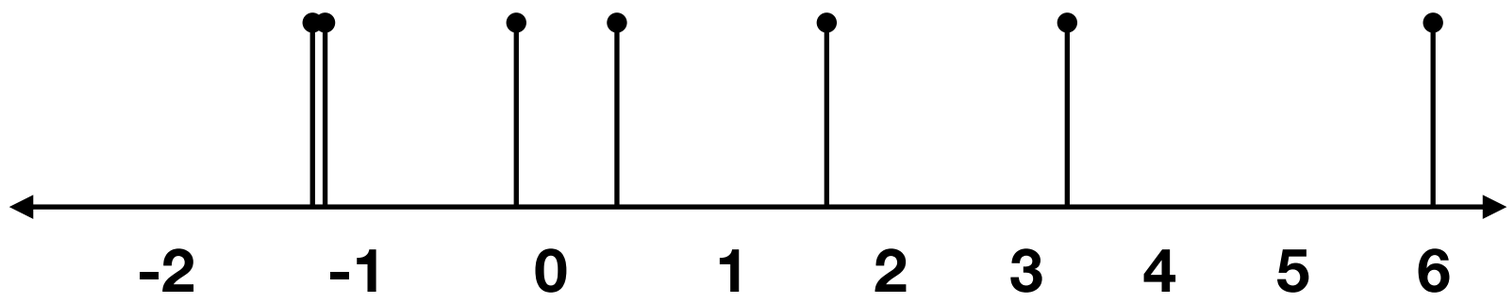
Math

$\{X_i\}_{i=1}^{K} \quad i.i.d$ “independent and identically distributed”

Mathematical Properties

What is the expected value?

$$\mathbf{E}[X_i] = \sum_{x=-\infty}^{\infty} x \cdot P(X_i = x)$$



1.6*1/7 + 3.4*1/7 + -0.15*1/7 + -1.2*1/7 + 6.0*1/7 + 0.1*1/7 + -1.1 *1/7

$$\mathbf{E}[X_i] = \frac{1}{N} \sum_{p \in P} p = AVG(Pop)$$

CHIDATA

Suppose we calculated an average of the sample

$$\frac{1}{K}\sum_{i=1}^{K} X_i = MEAN(Samp)$$

Some notes…

The result is a “function” of K random variables so is itself a random variable.

There is uncertainty about which particular sample you got.

CHIDATA

Suppose we calculated an average of the sample

$$Z = \frac{1}{K}\sum_{i=1}^{K} X_i$$ //"Sample Mean"

What is the expected value?

$$\mathbf{E}[Z] = \mathbf{E}[\frac{1}{K}\sum_{i=1}^{K} X_i] = \frac{1}{K}\mathbf{E}[\sum_{i=1}^{K} X_i]$$ //K is a constant

$$\mathbf{E}[Z] = \frac{1}{K}\mathbf{E}[\sum_{i=1}^{K} X_i] = \frac{1}{K}\sum_{i=1}^{K}\mathbf{E}[X_i]$$ // Linearity

$$\mathbf{E}[Z] = \frac{1}{K}\sum_{i=1}^{K}\mathbf{E}[X_i] = E[X_i] = AVG(Pop)$$

CHIDATA

Suppose we calculated an average of the sample

$$Z = \frac{1}{K}\sum_{i=1}^{K} X_i$$ //"Sample Mean"

What is the expected value?

The expected value of a sample mean is the population mean!

*The sample mean is **an estimator** for the population average.*

If you play the lottery infinite times, you win in expectation

Such estimates are called "unbiased", but that doesn't necessarily mean they are accurate!

$$\mu = AVG(Pop)$$

$$\epsilon = Z - \mu = \frac{1}{K}(\sum_{i=1}^{K} X_i) - \mu$$

"Error" in Estimation

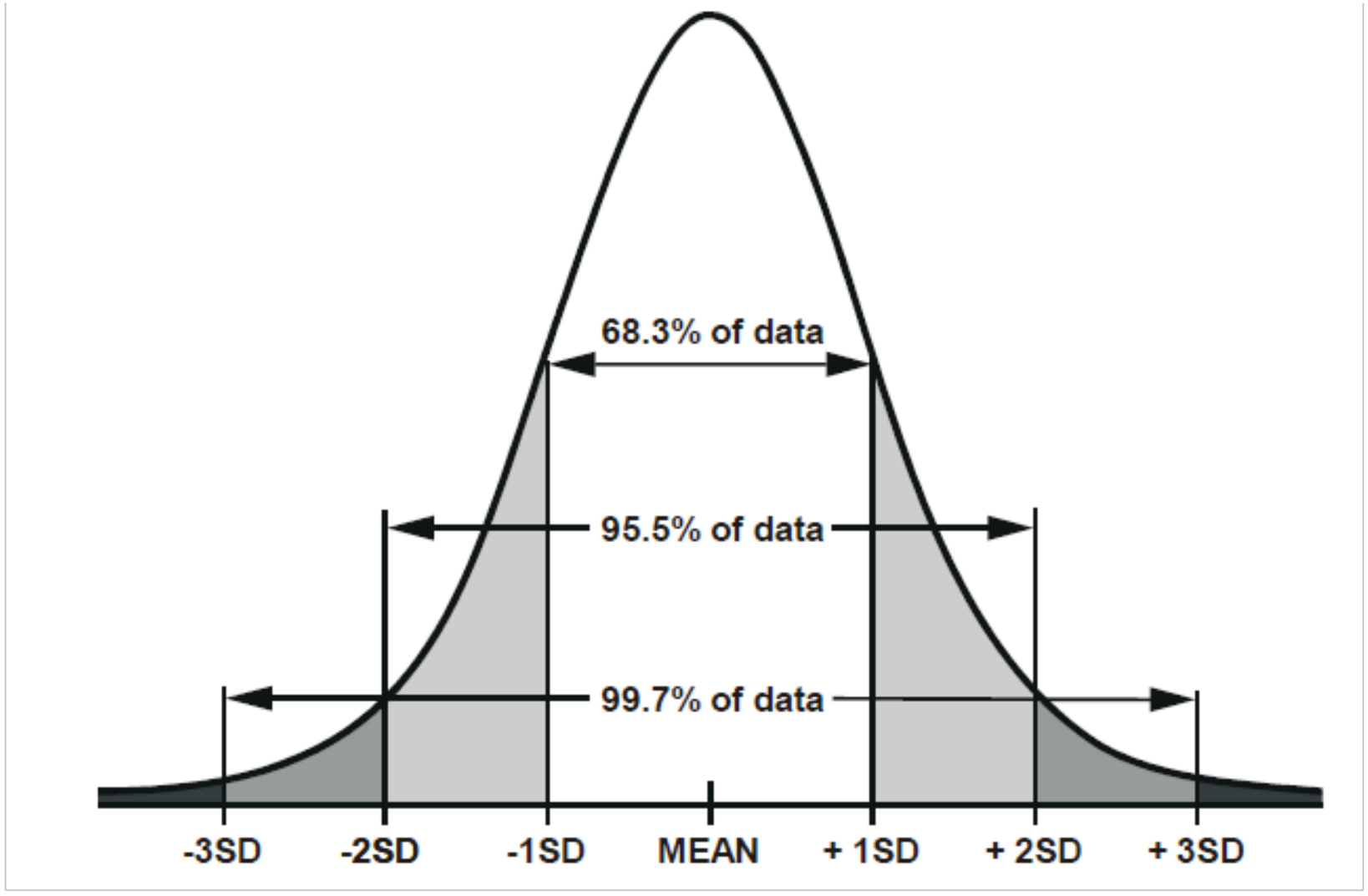
$$\mathbf{E}[\epsilon] = 0$$

What does the distribution of epsilon look like?

What is the probability I get a really misleading sample?

Central Limit Theorem

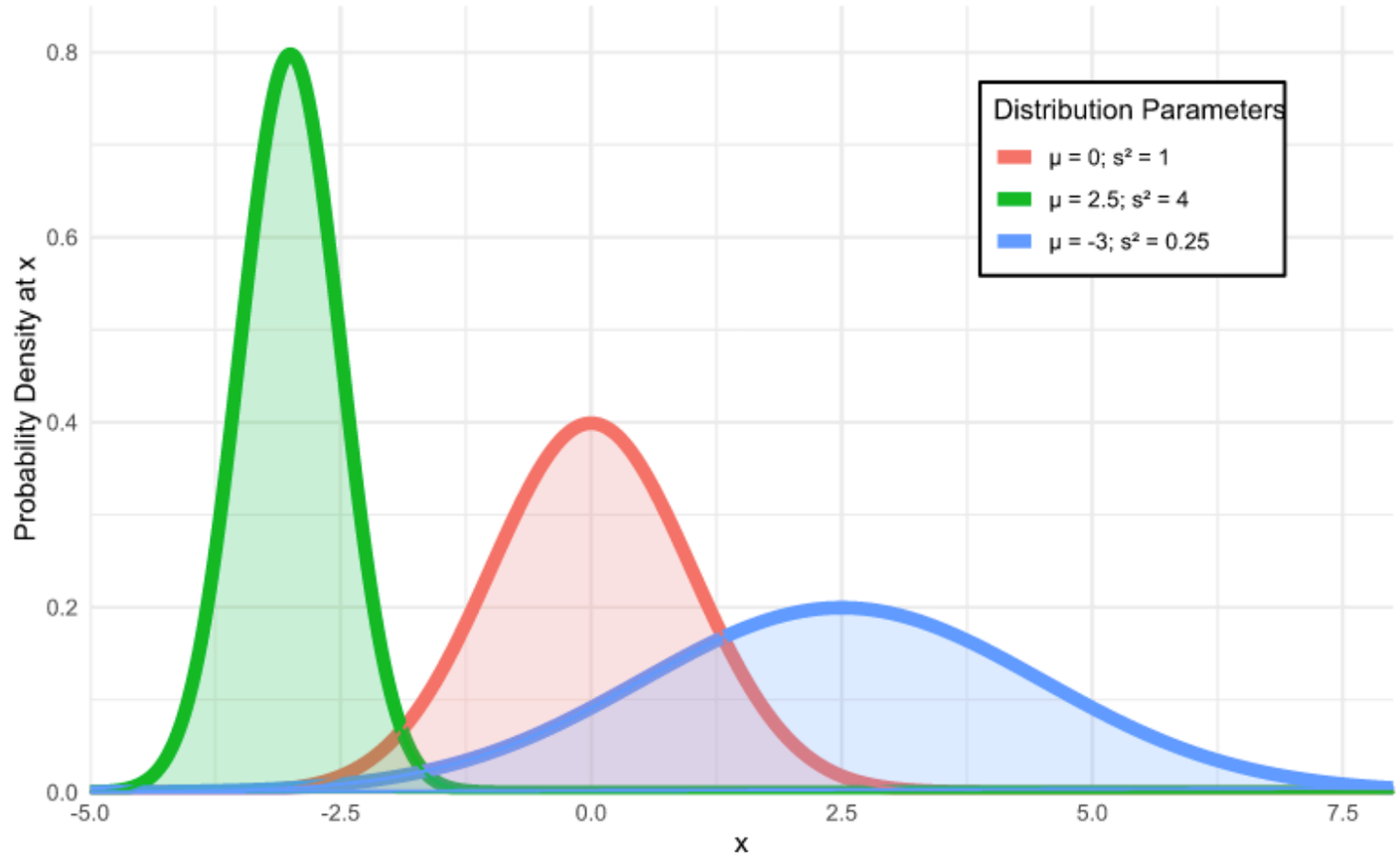
The error in a sample mean estimate is (approximately) normally distributed



Aside on normal distributions

Mean Variance

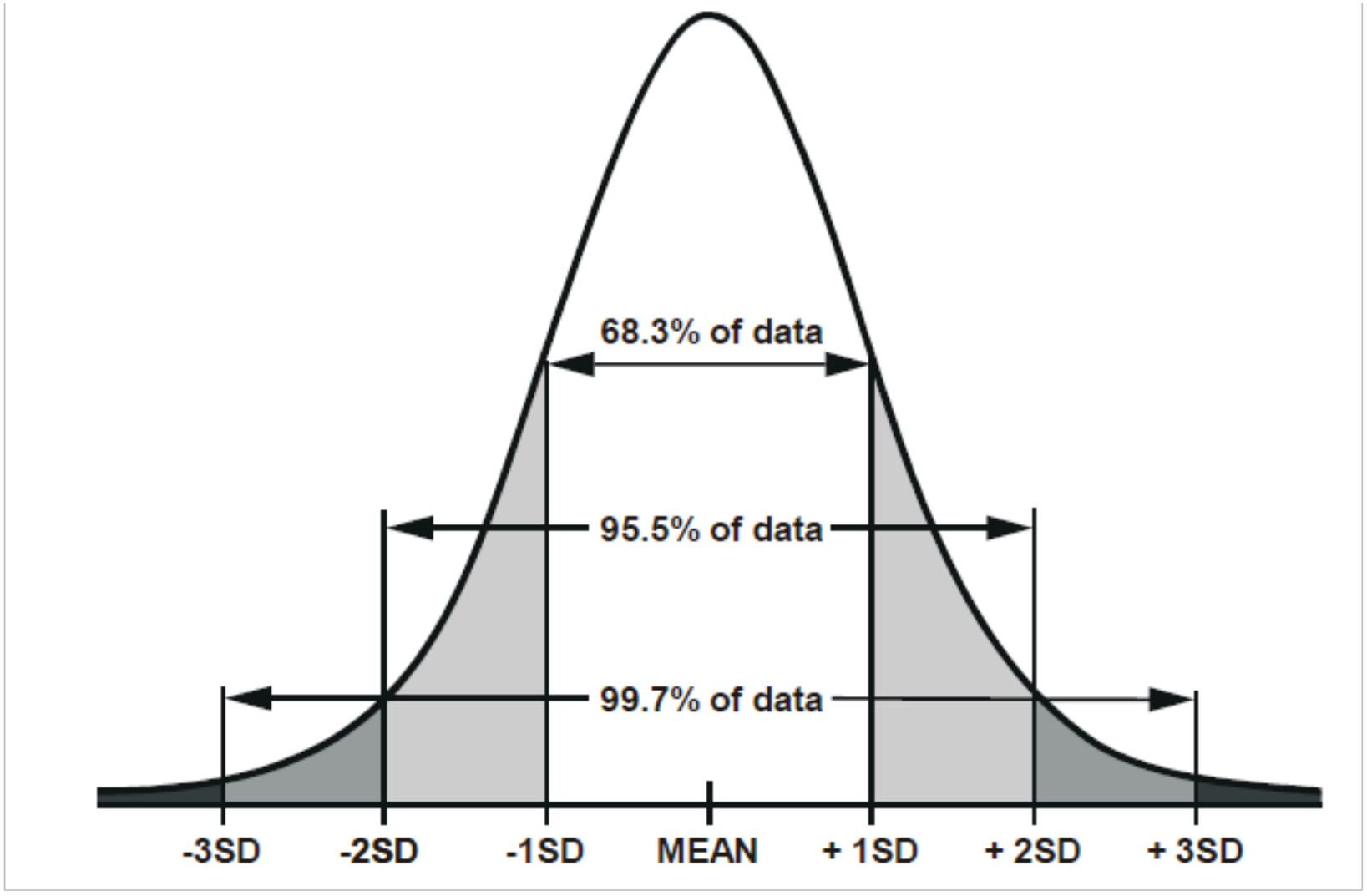
$$N(\mu, \sigma^2) = P(x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Aside on normal distributions

2 SDs (sigmas) roughly 95% probability

2.5 SDs (sigmas) roughly 99% probability



Central Limit Theorem

For i.i.d X_i,

$$\lim_{K\to\infty}(\frac{1}{K}\sum_{i=1}^{K}X_i) - \mathbf{E}[X_i] = N(0, \frac{Var[X_i]}{K})$$

What does this mean for us?

$$\epsilon \approx N(0, \frac{\sigma^2_{pop}}{K})$$

Variance of the population

Size of the sample

CHIDATA

Population Variance

Less data to accurately estimate the popl. average

Low-Variance Population

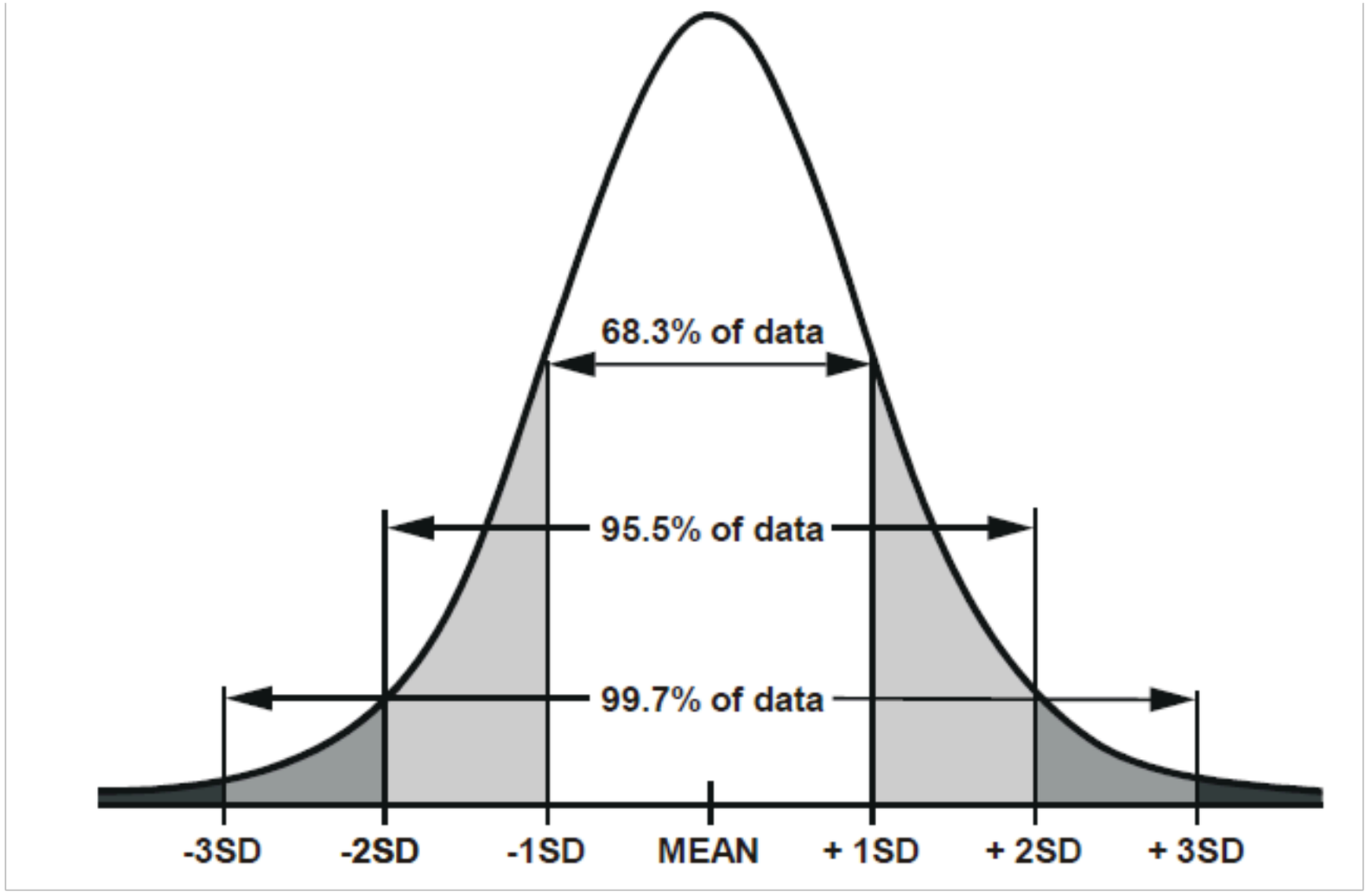
More data to accurately estimate the popl. average

High-Variance Population

“Confidence” Interval

Error within $\pm 2\frac{\sigma_{pop}}{\sqrt{K}}$ roughly 95% probability

Error within $\pm 2.5\frac{\sigma_{pop}}{\sqrt{K}}$ roughly 99% probability



CHIDATA

“Confidence” Interval

Error within $\pm 2\frac{\sigma_{pop}}{\sqrt{K}}$ roughly 95% probability

Error within $\pm 2.5\frac{\sigma_{pop}}{\sqrt{K}}$ roughly 99% probability

Measures the degree of confidence in a sample estimate

Two parts: **a range** and **a probability**

Chicken-and-Egg Problem?

How do we get the population variance?

$$\epsilon \approx N(0, \frac{\sigma^2_{pop}}{K})$$

Variance of the population

Size of the sample

If we know bounds on our population,
the variance is bounded as well.

$$X_i \in [a, b] \implies Var[X_i] \leq \frac{(b-a)^2}{4}$$

CHIDATA

Sample Mean Estimation Error

Draw a sample of size K from a population with a range of values in [a,b].

The difference between the sample mean and the population average is:

Error within $\boxed{\pm\frac{(b-a)}{\sqrt{K}}}$ with 95% probability

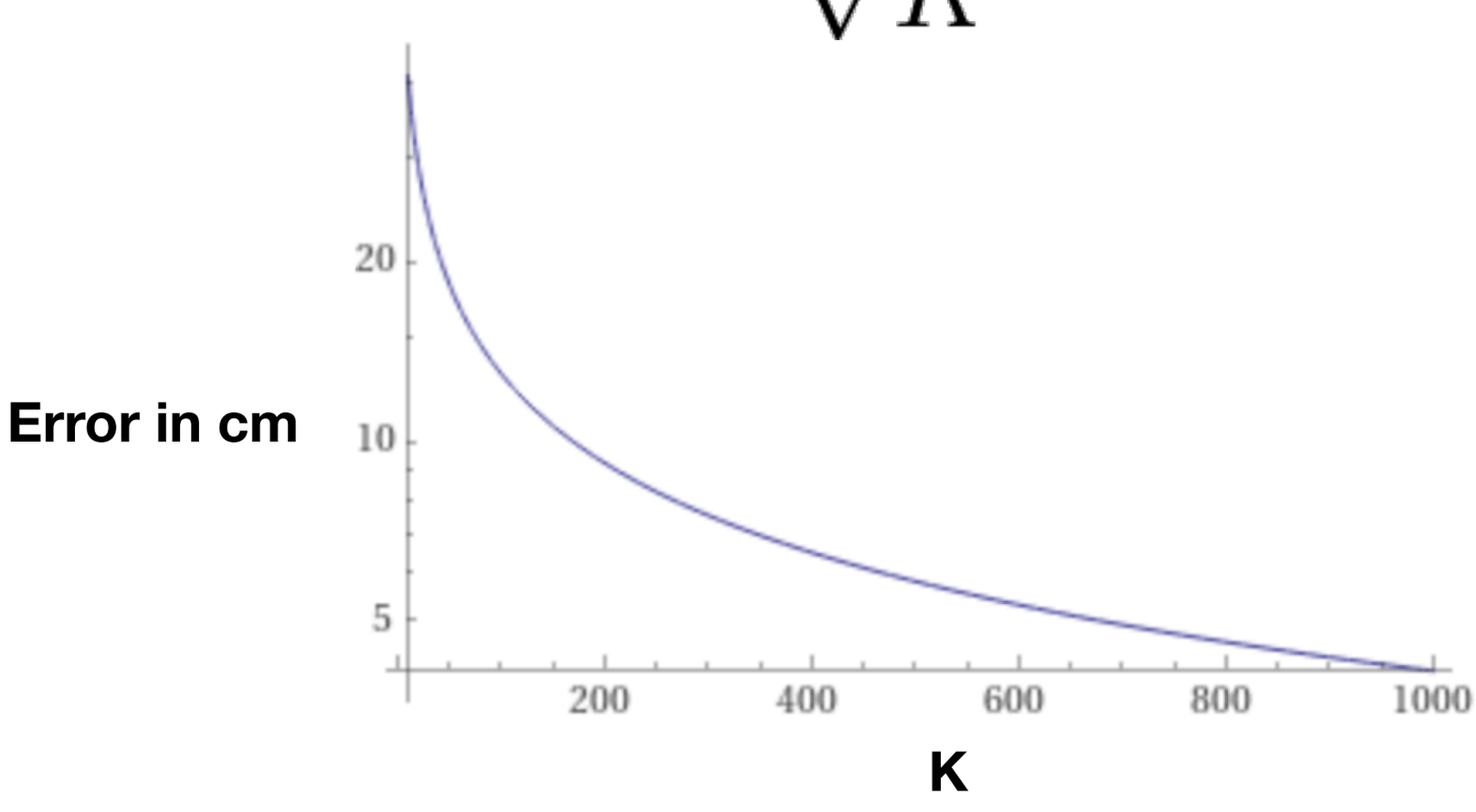
Error within $\pm 1.25\frac{(b-a)}{\sqrt{K}}$ with 99% probability

Some examples

Average height of students at U Chicago

a=100cm, b=230cm

$$\pm \frac{(b-a)}{\sqrt{K}}$$

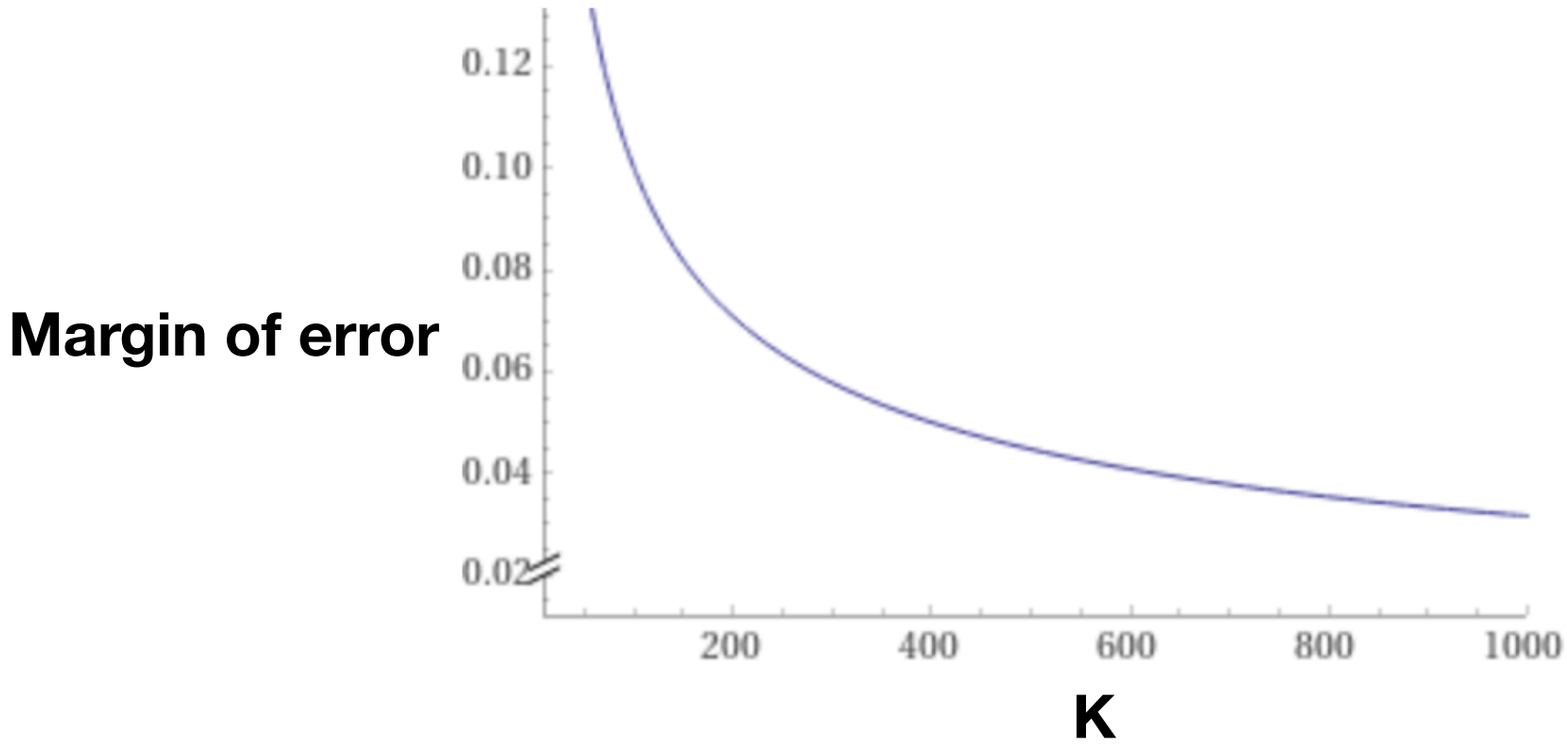


Some examples

Yes/No vote (or a 2-person election)

a=0, b=1

$$\pm \frac{(b-a)}{\sqrt{K}}$$



Some examples

Show that a treatment has a significant effect

a (lowest measured signal), b (highest), expected effect Δ

$$\frac{(b-a)}{\sqrt{K}} \leq \Delta$$

$$\frac{(b-a)^2}{\Delta^2} \leq K$$

Signal-to-Noise Ratio (actually it's inverse)

CHIDATA

Summary

Population

```
For 1…K:
  Pick an element from the population with equal prob
```

The difference between the sample mean and the population average is:

Error within $\pm\frac{(b-a)}{\sqrt{K}}$ with 95% probability

CHIDATA

Summary

Population

Simplest **model** for “data acquisition”

Sample a small amount of a very large population

Samples are mostly independent

Nice formulas for analysis