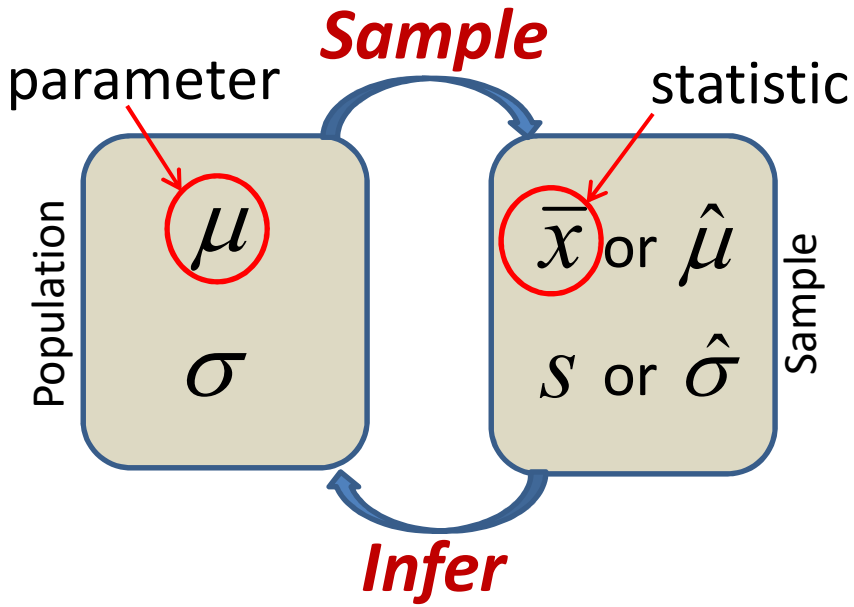
DA503 Applied Statistics

Lecture 04

Point Estimation

Population vs Sample

- Summary: Inferential statistics estimates population parameters from a random sample to draw conclusions and make better decisions.



As different samples will give rise to different values for $\bar{x}$, There will be a sampling error:

$$Sampling\ error = \bar{x} - \mu$$

- What's measured from a sample is called a **statistic**.
- The same thing measured from a population is called a **parameter.**
- We're using the **statistic** from a sample to estimate the corresponding population **parameter**.

Sampling methods

- Terminology
  - Population: Consists of all possible observations relating to a given phenomenon
  - Sample: A part of the population
- **Systematic random sampling**
  - Randomly select a starting number and an interval
  - Example: Population: 100, sample: 10

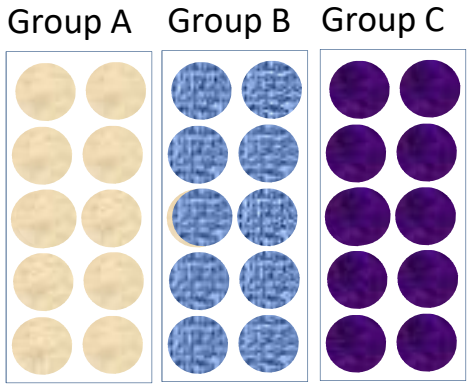
    Starting #: 14, interval: 8 => sample={14,22,30,38,46,54,62,70,78,86}
- **Simple random sampling**
  - The entire process of sampling is done in a single step with each subject drawn independently (like lottery)

Sampling methods – cont'd

- **Cluster sampling**
  - Gather sample population and then select groups of clusters, For each cluster select individual subjects by either systematic random or simple random sampling.
  - Example: Geographical clusters. Divide the entire population of Turkey into cities for a study of the academic performance of students in Turkey.
- **Stratified sampling**
  - Entire population is divided into subgroups, then subjects are randomly selected from subgroups in a proportional way.

Sampling methods – cont'd

- **Stratified sampling** (cont'd)
  - Groups (age, gender, religion, nationality, socioeconomic status, etc) must be non-overlapping



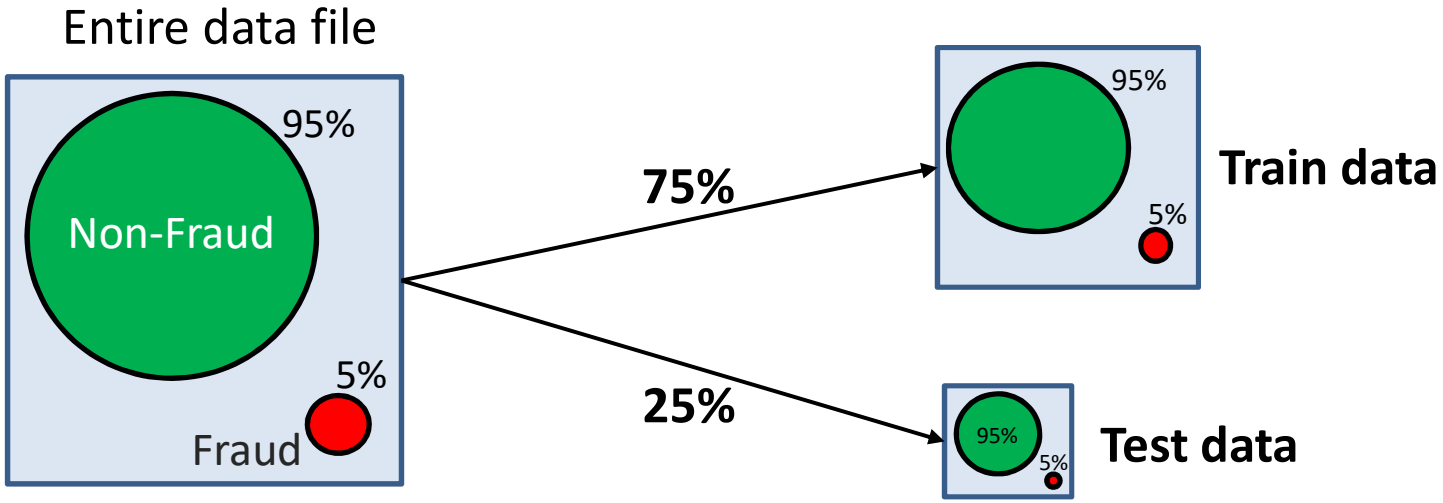
a) Proportionate
statistical random sampling

Groups	A	B	C
Population	100	200	300
Sampling fraction	¼	¼	¼
Sample size	25	50	75

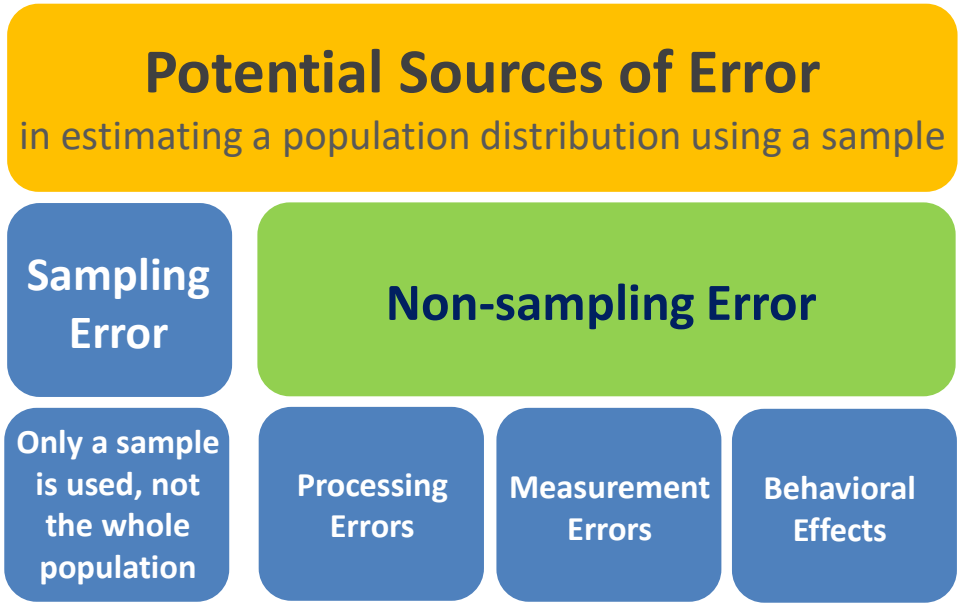
b) Disproportionate stratified random sampling (different sampling fractions)

Sampling methods – cont'd

- **Stratified sampling (example)**
  - Suppose you have data composed of Fraud (5%) and non-Fraud (95%) transactions.
  - You want to split this data file as train (75%) and test (25%) data sets for a Machine Learning model.
  - Stratified sampling ensures that we have the same ratio of Fraud/non-Fraud cases in both train and test files



Errors involved



Sampling error: Arises in a data collection process as a result of taking a sample from a population rather than using the population (quantified by standard error). Creates the difference between estimate and the true (but unknown) value of a population parameter. Generally due to:

- the size of the sample
- variability within the population
- sample design (stratification, clustering etc.)

Processing errors: Could arise during the data processing (data cleaning/capture & editing)

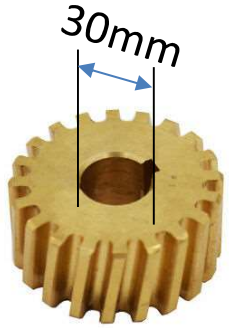
Measurement errors: Arises from due to imperfections in the way data is collected, e.g. poor wording or misleading/ambigious questions, faulty assumptions, scale issues, etc.

How about **Sampling bias?**

Behavioral effects: Respondent bias (refusal to answer questions or giving incorrect answers due to privacy issues), Non-response error (partially-filled questionnaires or rejected ones)

A word on sample size

- You have 2 plants: same technology, same processes but different amounts of production



A product with good specs: **30 ± 0.45** mm

- This technology + process (assuming μ=30 and σ=0.2) produce 2.45% defective items.
- Whenever # of defective items per day > 3% => **Alert**!
- By the end of 3 months, which plant you would expect to have more flagged days?
  a) Small plant
  b) Big plant
  c) The same number on average

22%
30%
48%

What did students say?

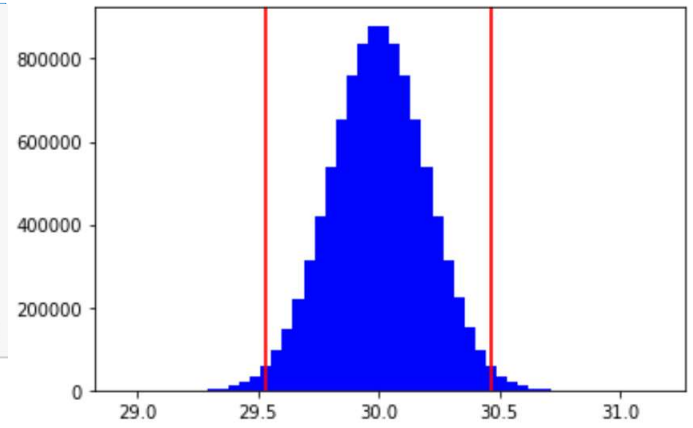
A word on sample size – cont'd

- It is intuitive that the larger our sample size, the more confident we are that it is representative of the entire population. In a larger sample, we are likely to capture the natural variation and diversity in the data. As the sample size increases, sample means converge to the population mean.
- Statistically speaking, **smaller the sample size, greater the likelihood of seeing an outlier**. Thus you'll see more extreme values in small samples. So caution is required when extrapolating from small samples.
- Fortunately, there is a point when increasing the sample size offers no more statistical benefits.
- A good read: Small Sample Size Paradox

  https://clinicaltrialist.com/2017/11/13/small-sample-size-paradox

A word on sample size – cont'd

```
NL = 10000 ; NS = 100
print('Ratio of extremes in:')
print('Large sample   Small sample')
print('---------------------------')
for i in range(20):
    smp1 = stats.norm.rvs(loc=30, scale=0.2, size=NL)
    smp2 = stats.norm.rvs(loc=30, scale=0.2, size=NS)
    cL = ((smp1 > 30.45) | (smp1 < 29.55)).sum()
    cS = ((smp2 > 30.45) | (smp2 < 29.55)).sum()
    print("{0:.2%}".format(cL/NL), ' '*8, "{0:.2%}".format(cS/NS))
```




```
Ratio of extremes in:
Large sample   Small sample
---------------------------
2.69%          2.00%
2.50%          5.00%
2.18%          1.00%
2.47%          5.00%
2.65%          4.00%
2.40%          2.00%
2.52%          2.00%
2.59%          5.00%
2.49%          0.00%
2.46%          5.00%
2.46%          1.00%
2.30%          1.00%
2.54%          6.00%
2.61%          3.00%
2.33%          5.00%
2.55%          3.00%
2.43%          1.00%
2.70%          4.00%
2.37%          1.00%
2.39%          7.00%
```

20 days

In 90 days …

```
NL = 10000 ; NS = 100
countL = 0 ; countS = 0
for i in range(90):
    smp1 = stats.norm.rvs(loc=30, scale=0.2, size=NL)
    smp2 = stats.norm.rvs(loc=30, scale=0.2, size=NS)
    if(100*((smp1 > 30.45) | (smp1 < 29.55)).sum()/NL) >= 3 : countL += 1
    if(100*((smp2 > 30.45) | (smp2 < 29.55)).sum()/NS) >= 3 : countS += 1
print(countL,'alert flags in 90 days for the Large Factory ')
print(countS,'alert flags in 90 days for the Small Factory ')
```

```
1 alert flags in 90 days for the Large Factory
38 alert flags in 90 days for the Small Factory
```

```
NL = 10000 ; NS = 100
countL = 0 ; countS = 0
for i in range(90):
    smp1 = stats.norm.rvs(loc=30, scale=0.2, size=NL)
    smp2 = stats.norm.rvs(loc=30, scale=0.2, size=NS)
    if(100*((smp1 > 30.45) | (smp1 < 29.55)).sum()/NL) >= 3 : countL += 1
    if(100*((smp2 > 30.45) | (smp2 < 29.55)).sum()/NS) >= 3 : countS += 1
print(countL,'alert flags in 90 days for the Large Factory ')
print(countS,'alert flags in 90 days for the Small Factory ')
```

```
0 alert flags in 90 days for the Large Factory
42 alert flags in 90 days for the Small Factory
```

Keep in mind that your run will yield different results, but the messge will remain the same!

Point estimators and point estimates

- A **point estimate** consists of a single value. Suppose we have an unknown population parameter such as population mean and population proportion.

  => Mean number of days it rains in Istanbul in April

  => Proportion of elderly above 65 with a smart phone

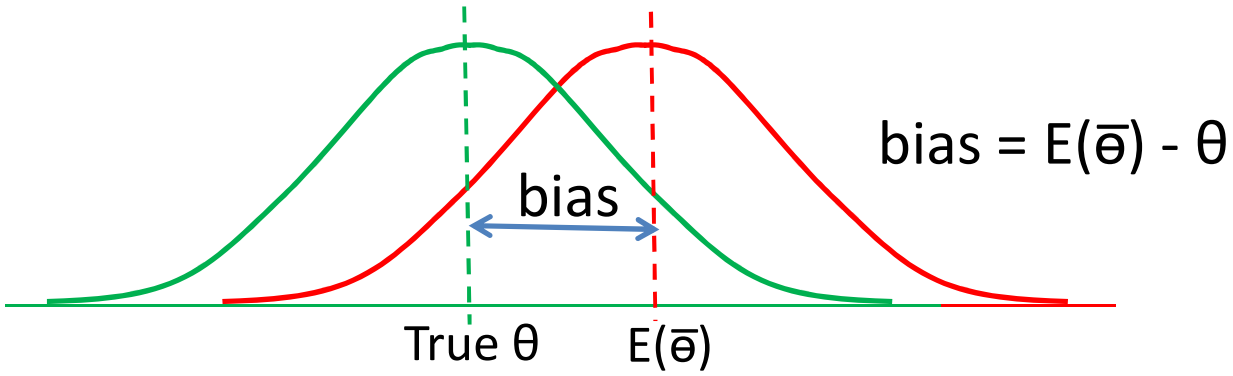
- To estimate these parameters, we use a random sample of size n from the population.
- An estimate is the value obtained when the observations x_i have been substituted into a formula (**estimator**) such as computing the mean or proportion.
- **Estimators**:
  - **Point estimator**: A single number (statistic) is calculated to estimate the parameter
  - **Interval estimator**: 2 numbers are calculated to create an interval within which the parameter is expected to lie (later)

Desirable properties of point estimators

- A good estimator should:
  - be **unbiased**
    - Doesn't systematically overestimate or underestimate the target parameter
  - have **small variance**
    - Sampling distribution of the estimator has a small spread
  - be **efficient**
    - Has minimum Mean Square Error (MSE) among all competitors
  - be **consistent**
    - Statistic converges (in probability) to population parameter as the sample size gets larger

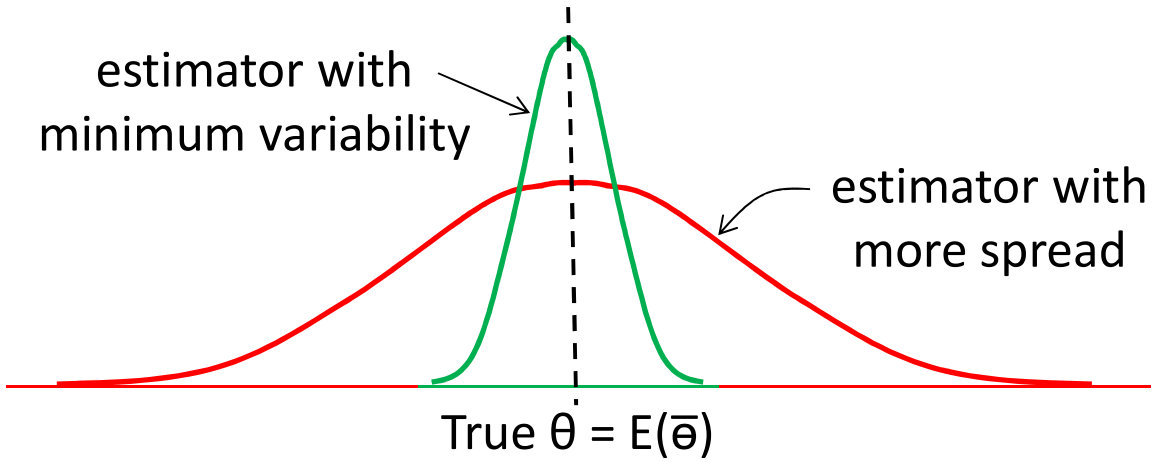
Unbiased

- Distribution centers correctly
  - The mean of the sampling distributions of sample estimators is the same as the population parameter.
  - We want the mean of the means from each sample to be the same as the mean of the population. The estimator doesn't routinely over or under-estimate the population parameter (Central Limit Theorem – later).
  - The sample mean is a minimal variance unbiased estimator for the population mean.



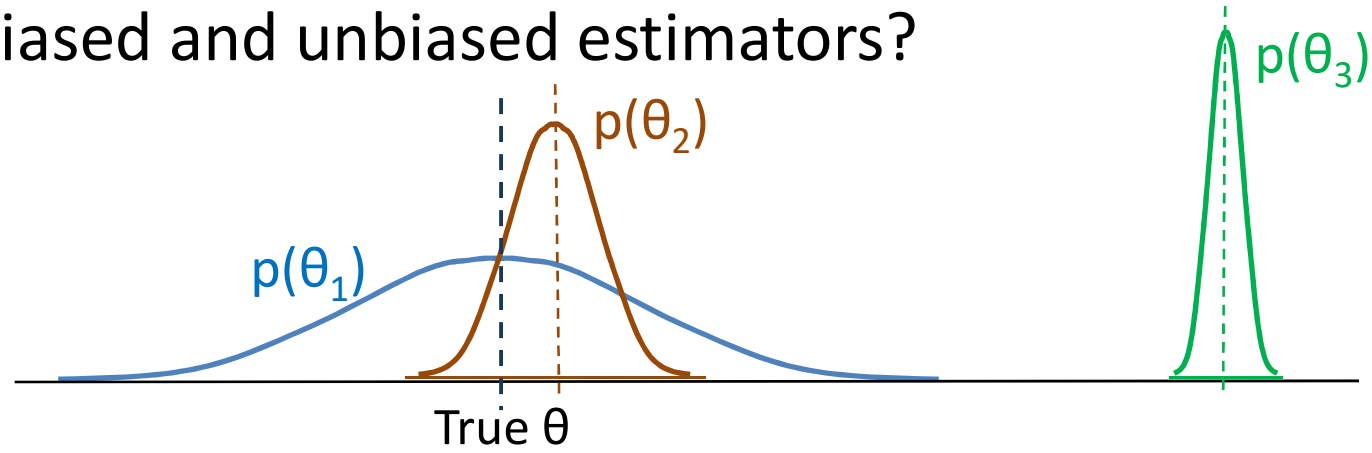
Minimum variance

- Distribution with small variance
  - We want it to have small variability when we repeat the experiment over and over again.
  - Of all the possible unbiased estimators, we want the one with the smallest variance in the sampling dist.
  - Suppose A and B are both unbiased estimators of a population parameter. A is said to be a more efficient estimator than B if the variance of A is smaller than that of B.



Efficiency

- In comparing unbiased estimators, we choose the one with minimum variance. What if we're comparing both biased and unbiased estimators?



θ_2 is the estimator with the best combination of small bias and variance. An estimator is said to be efficient if its MSE is minimum among all competitors:

- Mean squared error (MSE): $MSE = E(\overline{\theta} - \theta)^2$

$MSE = Bias(\overline{\theta})^2 + Var(\overline{\theta})$ where $Bias(\overline{\theta}) = E(\overline{\theta}) - \theta$

$$MSE = Bias^2 + \sigma^2$$

For proof, see: https://en.wikipedia.org/wiki/Mean_squared_error

MSE derived

- Proof of the expression for MSE (may skip this)
  - Variance for X (a random variable) is written as:

$$Var(X) = E(X^2) - [E(X)]^2 \qquad \text{So let} \quad X = \bar{\theta} - \theta$$

(See Appendix I)

random variable

true (fixed) value = parameter

$$Var(\bar{\theta} - \theta) = E[(\bar{\theta} - \theta)^2] - [E(\bar{\theta} - \theta)]^2$$

$$[\ \mathbf{I}\ ] \qquad [\ \mathbf{II}\ ] \qquad [\ \mathbf{III}\ ]$$

$$[\ \mathbf{I}\ ] \qquad Var(\bar{\theta} - \theta) = Var(\bar{\theta})$$

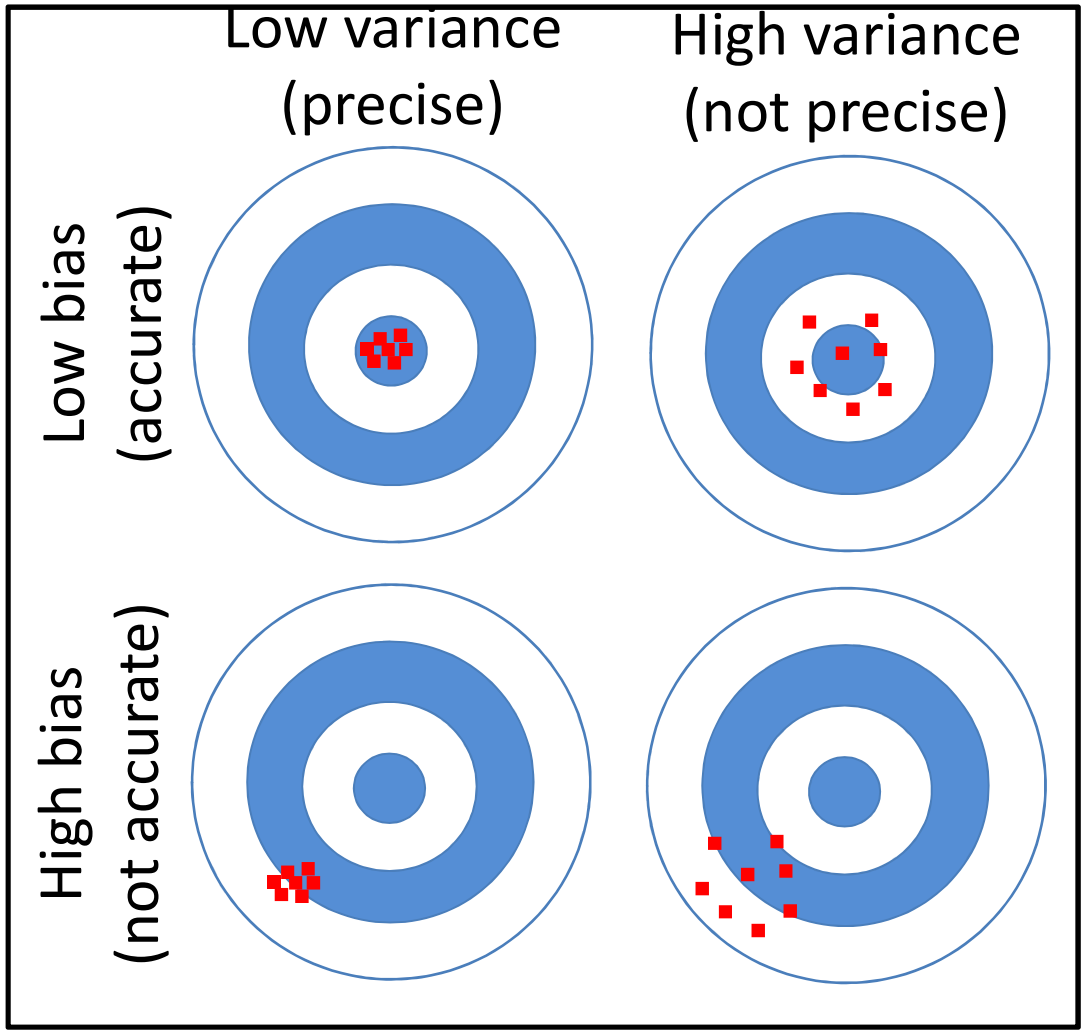
$$[\ \mathbf{II}\ ] \qquad E[(\bar{\theta} - \theta)^2] = MSE(\bar{\theta})$$

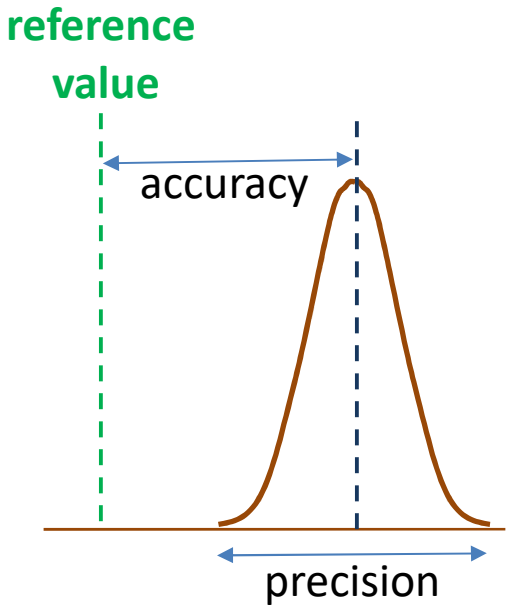
$$[\ \mathbf{III}\ ] \qquad [E(\bar{\theta} - \theta)]^2 = [E(\bar{\theta}) - E(\theta)]^2 = \underbrace{[E(\bar{\theta}) - \theta]^2}_{bias^2(\bar{\theta})}$$

$$Var(\bar{\theta}) = MSE(\bar{\theta}) - bias^2(\bar{\theta})$$

$$\boxed{MSE(\bar{\theta}) = Var(\bar{\theta}) + bias^2(\bar{\theta})}$$

A graphical depiction of bias and variance





What is a good estimator?

- Suppose you're given a set of data points drawn randomly from a normal distribution. **What is the point estimate for the population mean**?

  {-0.441, 1.774, -0.101, -1.138, 2.975, -2.138}

- One reasonable choice is to use the sample mean $\bar{x}$ as an estimate of μ : $\bar{x} = 0.155$
- Is sample mean the best choice?
- A crucial question: Does the estimator (sample mean) differ from the parameter in a systematic manner? Are there any biases?
- Can we answer this via a simulation?

What is a good estimator? – cont'd

- From a normally dist. population with a mean of $\mu = 4$ and std. dev. of $\sigma = 4$, we randomly select 10 numbers:

```
x = np.random.normal(4,4,10)
x.mean()
2.95238626757
```

Python code

much lower than 4!
is this systematic?

- Let's run a simulation in which we select 10 numbers 1000 times and look at the estimates:

```
sum = 0
for i in range(1000):
      x =  np.random.normal(4,4,10)
      sum += x.mean()
print(sum/1000)   => 4.00571914945
```

Python code

- So, the estimate of the sample mean $\bar{x}$ does not systematically over- or under-estimates the parameter. Hence, **$\bar{x}$** is an **unbiased estimate** of the parameter **μ**.

What is a good estimator? – cont'd

- How about variance for the same data set?
- For variance σ^2, we have two choices:

$$S_n^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / n \quad \text{and} \quad S_{n-1}^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$$

```
var_biased = 0
var_unbiased = 0
for i in range(1000):
    x = np.random.normal(4,4,10)
    var_biased  += np.sum((x-x.mean())**2)/10
    var_unbiased += np.sum((x-x.mean())**2)/9
print(var_biased/1000)
14.3956043972
print(var_unbiased/1000)
15.9951159969
```

biased

unbiased

divide by **n**

divide by **n-1**

Python code

- (n-1) is used to make this estimator variance unbiased, especially for small sample sizes (Bessel correction)

Concept of degrees of freedom

- Many statistics are made up of many random variables, say n of them. But if there are k linear constraints on the n random variables, only n−k of them are actually random, i.e., the statistic has n-k degrees of freedom.
- The estimator of variance for a sample:

$$S_{n-1}^{\ 2} = \sum_{i=1}^{n} (X_i - \overline{X})^2 / (n-1)$$

- İt's made up of n random variables, X_1, X_2 , ..., X_n. X_1 is random meaning that it can be any number. Same for X_2, X_3 and all others up to X_{n-1}.
- X_n, however, isn't random at all. There is only one possible value for X_n, it has to have a value that makes the sample mean equal $\bar{X}$.
- So S^2 is made of n-1 random variables (one being redundant), i.e., has n-1 degrees of freedom.

What is a good estimator? – cont'd

- Sample variance has a tendency to under-estimate the population variance. So the population variance with the (*n-1*) adjustment is an unbiased estimator.

$$E[\sigma^2 - S_{biased}{}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}\left((x^2{}_i - 2x_i\mu + \mu^2) - (x_i{}^2 - 2x_i\bar{x} + \bar{x}^2)\right)\right]$$

Proof

$$= E\left[\mu^2 - \bar{x}^2 + \frac{1}{n}\sum_{i=1}^{n}\left(2x_i(\bar{x}-\mu)\right)\right] = E[\mu^2 - 2\bar{x}\mu + \bar{x}^2]$$

$$= E[(\bar{x}-\mu)^2] = Var(\bar{x}) = \frac{\sigma^2}{n} > 0$$

So $S_{biased}{}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 / n$ underestimates the variance.

- So, the expected value of the estimators will be:

$$E[S_{biased}{}^2] = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

$$S_{unbiased}{}^2 = \frac{n}{n-1}S_{biased}{}^2$$

This is why S^2 with n-1 is an unbiased estimator.

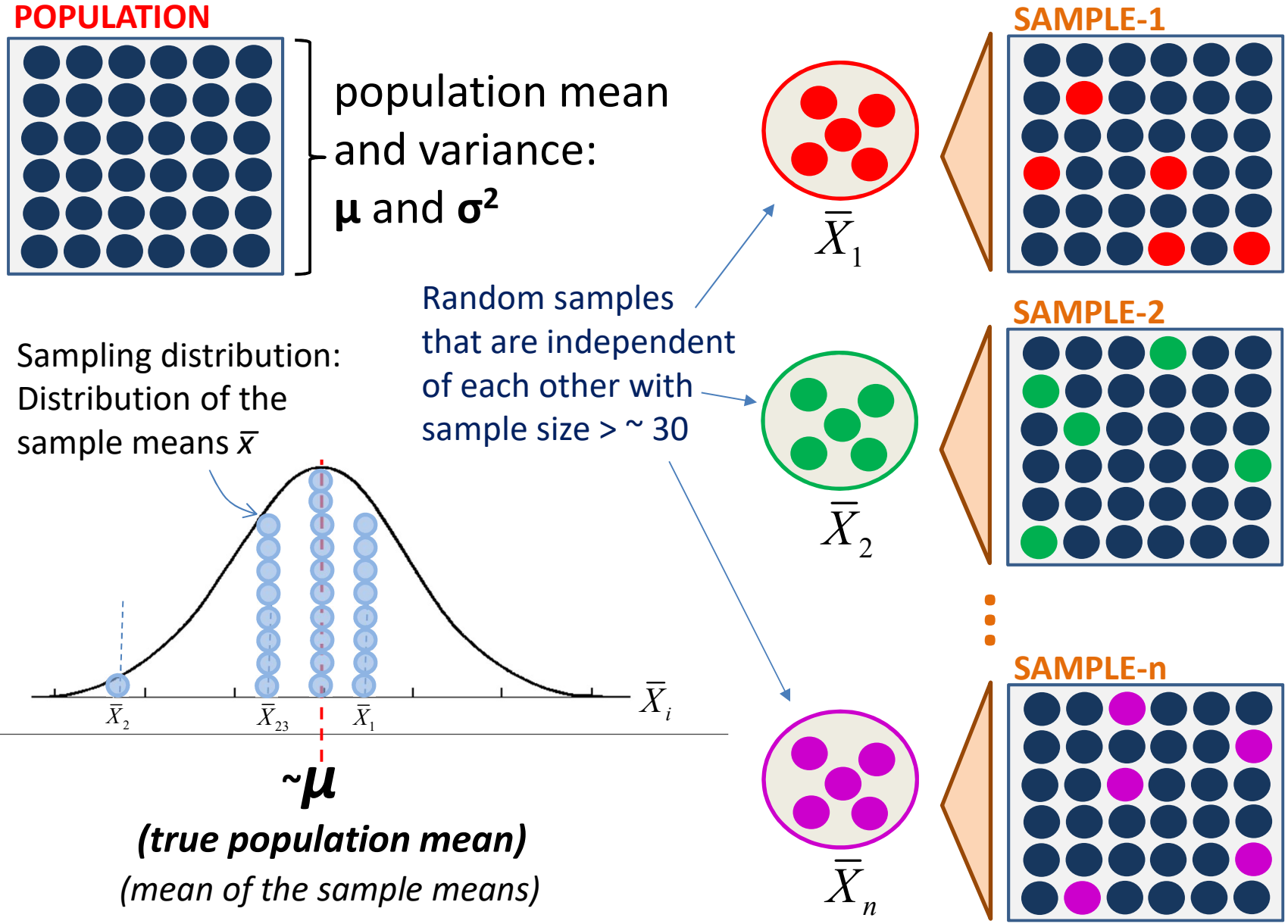
Population vs Sample

- When the population size **N** is large, we take a sample of **n** observations and compute the mean & variance.
- Given the random sample $X_1, X_2, \ldots, X_n$:

	Population (parameter)	Sample (statistic)
Mean	$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$	$s^2_{biased} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$ $s^2_{unbiased} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$

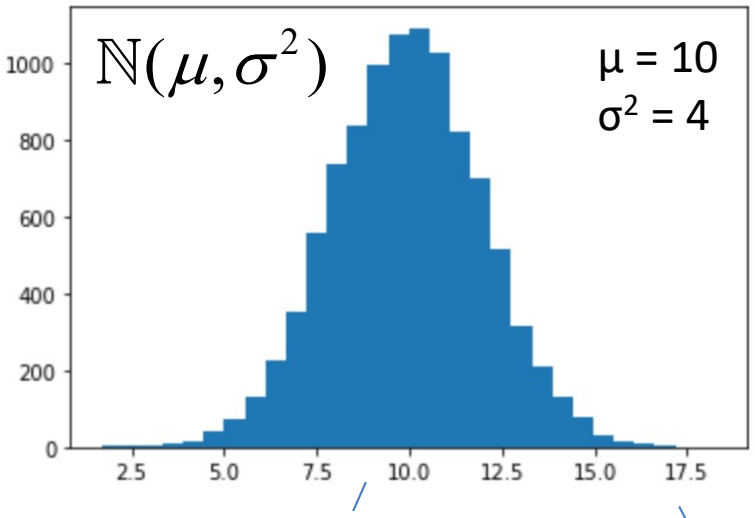
unbiased estimators

Central Limit Theorem (CLT)

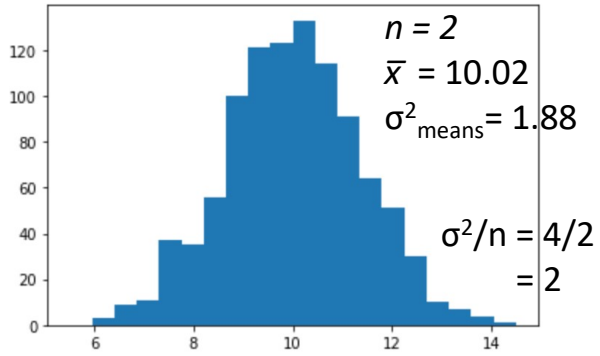


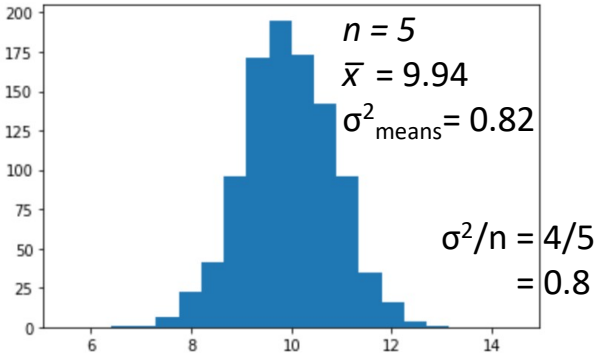
Sampling distribution of the sample mean $\bar{X}$

μ : Population mean
σ^2 : Population variance

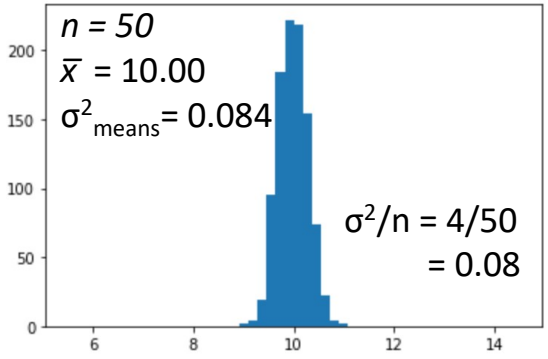


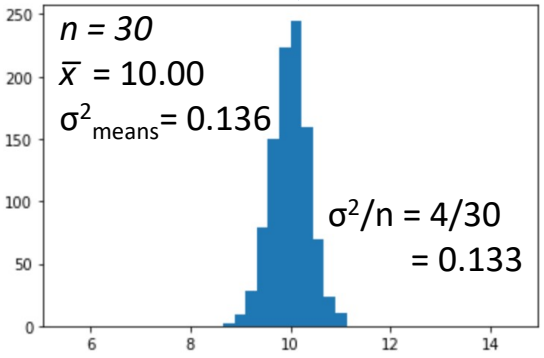
1000 iterations

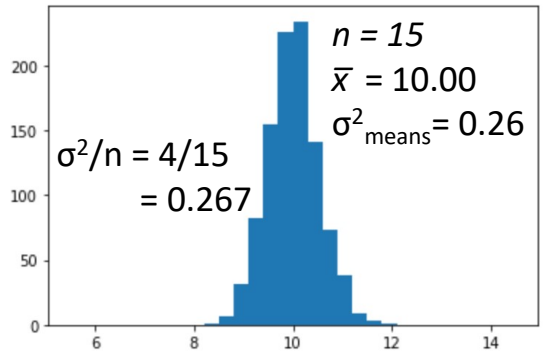




n : Sample size
$\bar{X}$: Mean of the sample means
σ^2_{means} : Std dev of sample means







Central Limit Theorem (CLT)

- Let $X_1, X_2, \ldots, X_n$ represent a random sample of size ***n*** selected from a population with a mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$.
- **CLT says**: The mean $\bar{X}$ of a sample with size ***n*** will follow approximately a **Normal distribution** with:
  - mean = $\boldsymbol{\mu}$, and
  - variance $\boldsymbol{\sigma^2/n}$
  - provided that **n is large** (for roughly $n \geq 30$)
- This is true regardless of the distribution of the original population.
- CLT enables us to make accurate probability statements. Knowing the sampling distribution for the mean $\bar{X}$, we could determine the probability that $\bar{X}$ would fall in any specified interval.

Why do we love Gaussian Distribution?

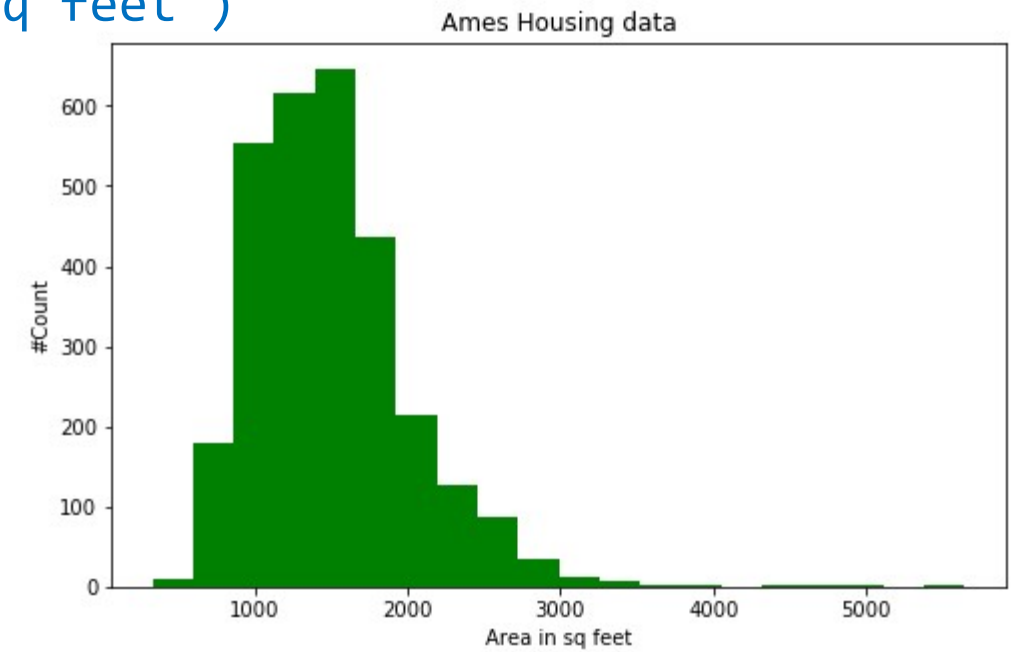
- So many of processes in nature and social sciences naturally follows the Gaussian distribution. Even when they don't, Gaussian gives the best model approximation for these processes (blood pressure, height, particle position in diffusion, measurement errors, etc.)
- Mathematical reason: Central Limit Theorem
  - When we add large number of independent random variables, irrespective of the original distribution of these variables, their normalized sum tends towards a Gaussian distribution
- Once a Gaussian, it's always a Gaussian!
  - Product of 2 Gaussians is a Gaussian, Sum of 2 independent Gaussian random variables is a Gaussian, Convolution of Gaussian with another Gaussian is a Gaussian, Fourier transform of Gaussian is a Gaussian, ...
- Simplicity
  - The entire distribution specified by 2 variables ($\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$)

CLT simulation

- Using internal area of houses (in sq.ft) from the data file ames.csv:

Python code

```
import numpy as np
import pandas as pd
df = pd.read_csv('ames.csv')
area = df['Gr.Liv.Area']
plt.hist(area, bins=20, color='g')
plt.title('Ames Housing data')
plt.xlabel('Area in sq feet')
plt.ylabel('#Count')
plt.show()
```

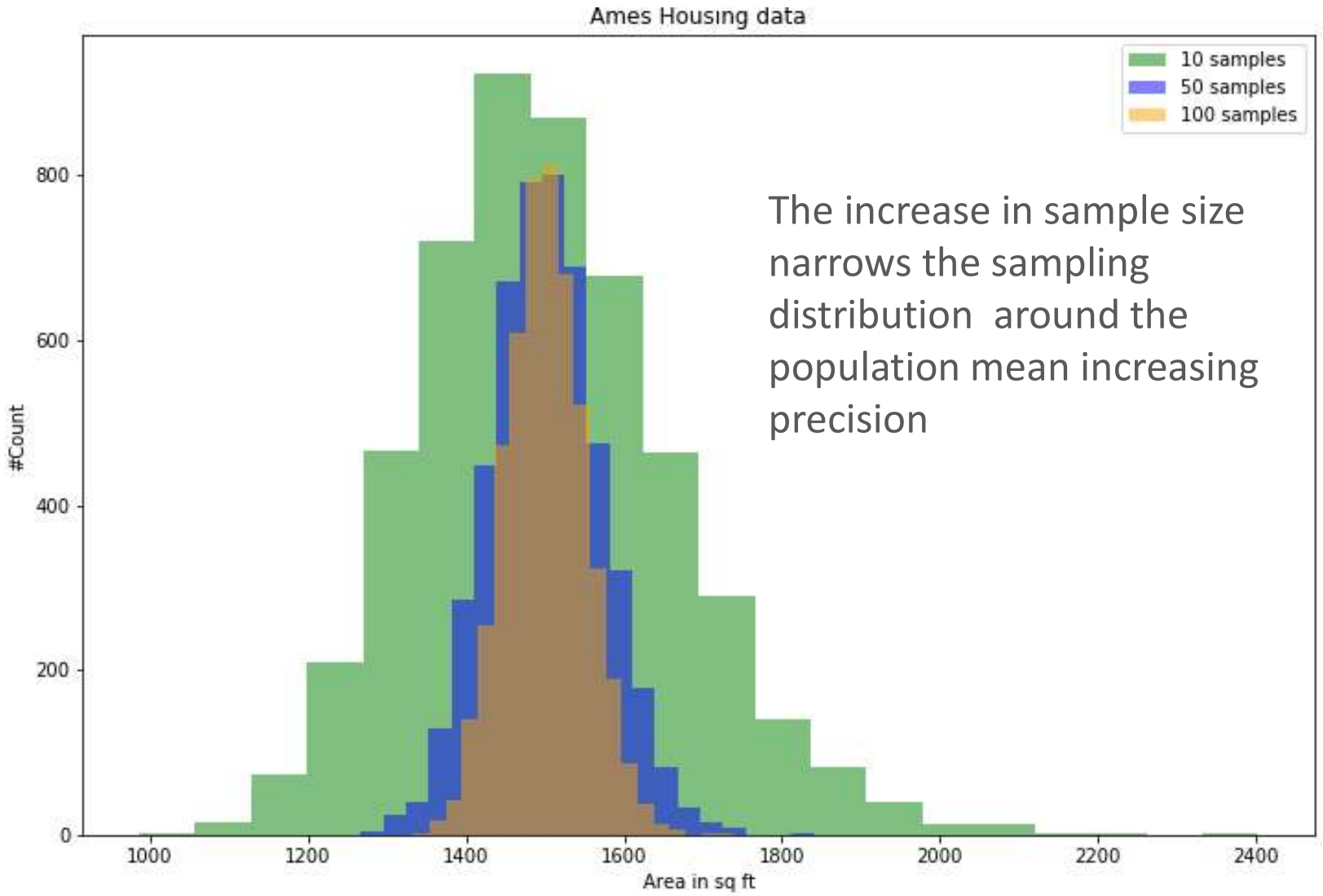


CLT simulation – cont'd

```
import random ; import numpy as np
means10 = [] ; means50 = [] ; means100 = []
for i in range(5000):
    samp = np.random.choice(area, size=10, replace=False)
    means10.append(np.mean(samp))
    samp = np.random.choice(area, size=50, replace=False)
    means50.append(np.mean(samp))
    samp = np.random.choice(area, size=100, replace=False)
    means100.append(np.mean(samp))
plt.hist(means10, bins=20, alpha=0.5, color='green',
      label='10 samples')
plt.hist(means50, bins=20, alpha=0.5,  color='blue',
      label='50 samples')
plt.hist(means100, bins=20, alpha=0.5, color='orange',
      label='100 samples')
plt.title('Ames Housing data')
plt.xlabel('Area in sq ft') ; plt.ylabel('#Count') ;
plt.legend(loc='upper right')
plt.show()
```

Python code

CLT simulation – cont'd



The increase in sample size narrows the sampling distribution around the population mean increasing precision

The mean of the distribution of $\bar{X}$

- Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution (population) with mean μ and variance σ^2. What is the expected value of the sample mean $\bar{X}$?

The mean of the sampling distribution of $\bar{X}$:

$$\bar{X} = \frac{1}{n}\left[X_1 + X_2 + \ldots + X_n\right]$$

Each observation X_i has the same population distribution with expectation μ, so the expected value of $\bar{X}$, which is also a RV :

$$E(\bar{X}) = E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{1}{n}\left[E(X_1) + E(X_2) + \ldots + E(X_n)\right]$$

See Appendix I

The X_i are identically distributed, which means they have the same mean μ. Therefore, replacing $E(X_i)$ with the alternative notation μ, we get:

$$E(\bar{X}) = \frac{1}{n}\left[\mu + \mu + \ldots + \mu\right] = \frac{1}{n}\left[n\mu\right] = \mu$$

The mean (expected value) of the sampling distribution of $\bar{X}$

The variance of the distribution of $\bar{X}$

- Variance of the sampling distribution of $\bar{X}$?

$$Var(\bar{X}) = Var\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = Var\left[\frac{1}{n}X_1 + \frac{1}{n}X_2 + \ldots + \frac{1}{n}X_1\right]$$

$$Var(\bar{X}) = \frac{1}{n^2}Var(X_1) + \frac{1}{n^2}Var(X_2) + \ldots + \frac{1}{n^2}Var(X_n)$$

See Appendix I

The X_i are identically distributed, which means they have the same variance σ^2. Therefore, replacing $Var(X_i)$ with the alternative notation σ^2, we get:

$$Var(\bar{X}) = \frac{1}{n^2}\left[\sigma^2 + \sigma^2 + \ldots + \sigma^2\right] = \frac{1}{n^2}\left[n\sigma^2\right] = \frac{\sigma^2}{n}$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$\sigma_{\bar{X}}$: standard deviation of the sample means ***(standard error of the sampling distribution** –SE)*

Central Limit Theorem (CLT) – revisited

- To re-iterate: Why is Central Limit Theorem important?
- If we know the population mean and standard deviation, we know the following will be true:

The distribution of means across repeated samples of size > 30 will be normal with a mean ($\bar{X}$) where

$$\overline{X} = \mu_{population}$$

Standard deviation ($\sigma_{\bar{x}}$) of the sample means (a.k.a. Standard Error of the means, **SE** or **SEM**):

$$\sigma_{\bar{x}} = \sigma_{population} / \sqrt{N}$$

$\bar{X}$ fluctuates around μ with a standard deviation of $\sigma_{\bar{X}}$

- Knowing what the distribution of the means looks like for a given population, we can take the mean from a single sample and compare it to the sampling distribution to assess the likelihood that our sample comes from the same population.
- i.e., we can test the hypothesis whether our sample represents a population distinct from the population.

Standard error – cont'd

- **Estimated SE** (standard error of the mean)
- Until now we've been sampling from a population whose **μ** and **σ** are known. In reality, we don't know the population parameters.
  We learned that $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, but when we take samples in real life, we almost never know σ.
- We can, however, use **s**, the standard deviation of the sample as the best estimate for **σ**. Remember, this is just an estimate for $\sigma_{\bar{x}}$, and it's called the **Standard Error of the statistic** (sample mean):

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$$

(a measure of uncertainty in the sample mean)

Standard error – cont'd

- Lower values of the standard error of the mean indicate more precise estimates of the population mean.
- Usually, a larger standard deviation will result in a larger standard error of the mean and a less precise estimate.
- A larger sample size will result in a smaller standard error of the mean and a more precise estimate.
- **Example:** Random sample of 312 delivery times yield a mean of 3.8 days with a standard deviation of 1.43 days:

$$SEM = S_{\bar{X}} = s/n^{1/2} = 1.43/(312)^{1/2} = 0.08 \text{ days}$$

Had you taken multiple random samples of the same size from the same population, standard deviation of those different sample means would be around 0.08 days (uncertainty due to random sampling is small).

Summary on SE/SEM

Standard Deviation (s)	Standard Error (SE)
Tells us how the sample data is distributed around the mean	Tells us how the sample mean is distributed
A large **s** tells us that some of the data points are quite far from the sample mean	A large **SE** tells us that the sample mean is far from representing the true population parameter (μ)
This is about the sample (**measures the variability within a single sample**)	This is about the sampling distribution (**estimates the variability between samples**)

- SE measures how far the sample mean is likely to be from the population mean (a measure of accuracy of a statistic calculated on a sample).
- In short, we use the SEM to determine how precisely the mean of the sample estimates the population mean.

Fitting distributions

- Fitting data to probability distributions:
  - Values of the parameters that yield the best fit between the model and the data (point estimates)
- Parametric inference
  - Specifying a priori a suitable distribution, then choosing the parameters that best fit this data set (e.g. mean and variance in a Normal distribution)
- Estimation
  - Finding estimates of the relevant parameters corresponding to the distribution that best represents the data (Methods of moments, Maximum likelihood estimation)

Fitting distributions – cont'd

- There are 4 steps in fitting distributions:
  1. Model choice (which family of distributions?)
     - Binomial, Normal, Gamma, etc.
     - Decision here is art as much as science (examine the sample distribution carefully, try a few others)
  2. Estimation of parameters
     - e.g., calculation of mean (μ) and variance (σ) if the distribution looks like Normal $N(\mu,\sigma^2)$
  3. Evaluate quality of fit (using visuals)
     - Plot the distribution with the estimated parameters and check the fit against the sample histogram
  4. Statistical tests for goodness of fit (will cover this later in the course)
     - For more rigorous reasonings

Methods of point estimation

- 2 well-known methods of point estimation:
  - **Method of Moments** (MoM)
    - Involves choosing population parameters by relating the sample moments (typically the sample mean and variance) to the theoretical moments of our chosen distribution
  - **Maximum Likelihood Method** (MLE)
    - Setting up a likelihood function corresponding to the distribution and finding the values that maximizes the function, which measures how likely it is to observe our given sample.

Method of Moments (MoM)

- Assume (x_1, x_2, … x_k) is a random sample taken from a distribution called f(x)
- Suppose there are **m** parameters to be solved $\theta=\theta_1,....\theta_m$
- k^{th} theoretical (population) moment of the distribution about the origin for k=1,2,... $: \mu_k = E(x^k) = \int_{-\infty}^{+\infty} x^k f(x)dx$
- k^{th} sample moment for k=1,2,... $: m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,k}$
- We equalize the first **m** population moments to the corresponding sample moments and solve the system of equations to find MoM estimates:

$$m_k = \mu_k \quad for \quad k = 1,2, ..., m$$

  - Then solve for parameters $\theta_1, \theta_2, ..., \theta_m$

Example

- A sample of 3 observations is collected from a continuous distribution with density f(x):

x1	x2	x3
0.4	0.7	0.9

$$f(x) = \theta x^{\theta-1} \quad for\ 0 < x < 1$$

- Estimate θ by using the **MoM**

$$\mu_1 = E(x^1) = \int_0^1 x f(x) dx = \int_0^1 x\,\theta\, x^{\theta-1} dx = \left. \frac{\theta\, x^{\theta+1}}{\theta+1} \right|_0^1 = \frac{\theta}{\theta+1}$$

$$m_1 = \frac{1}{3}\sum_{i=1}^{3} x_i^{\,1} = \frac{1}{3}(0.4+0.7+0.9) = \frac{2}{3}$$

$$\mu_1 = m_1 \quad => \quad \hat{\theta} = 2$$

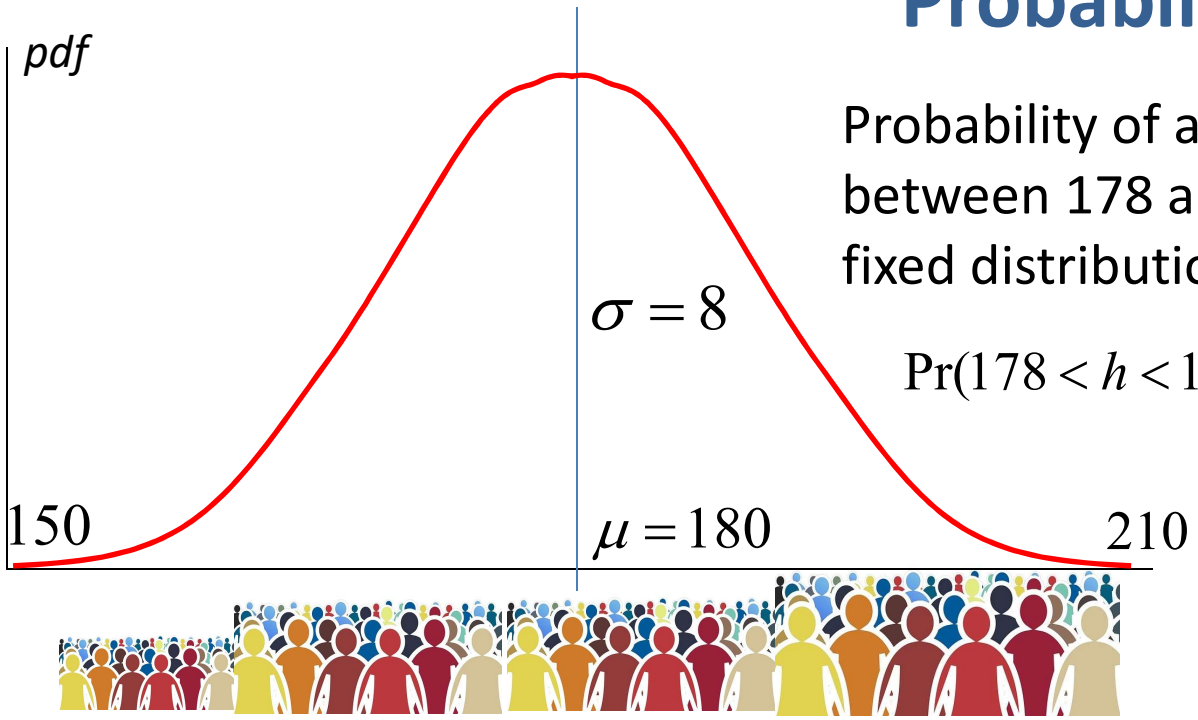
$$\boxed{f(x) = 2x \quad for\ 0 < x < 1}$$

Probability vs Likelihood

- Likelihood is a tool for summarizing the data's evidence about the unknown distribution.
- Why do we need this? This is the fundamental distinction between probability where we know about the population and statistics where we want to infer about population.
- Suppose we have a bag of red and blue marbles. Knowing what's inside the bag, you reach in and grab some marbles:
- **Probability** answers questions about what's likely to be in your hand given the contents of the bag.
- **Statistics** answers questions about what's likely to be in the bag given the contents of your hand.
- Probability density functions are good for answering questions of the first type, and likelihood functions are good for answering questions of the second type.

Ref: https://www.quora.com/What-is-the-difference-between-a-likelihood-function-and-a-probability-density-function

Probability vs Likelihood



Probability of a person whose height between 178 and 188 cm given the fixed distribution parameters μ and σ

$$\Pr(178 < h < 188 \mid \mu = 180, \sigma = 8) = 0.3$$

$\Pr(\text{data} \mid \text{distribution})$

variable

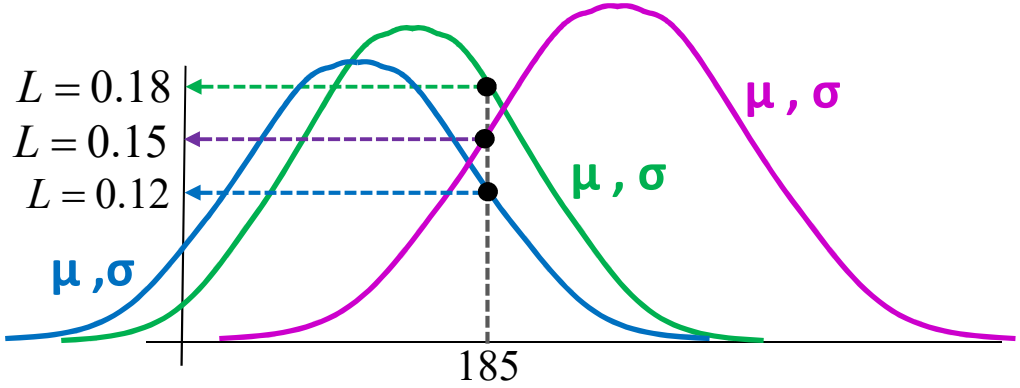
fixed/given

In likelihood we already have a person selected whose height is known (185), i.e., data is given/fixed, but the distribution is a variable.

$L(\text{distribution} \mid \text{data})$

variable

fixed/given



Maximum Likelihood Estimation (MLE)

- **Probability**: Area under a pdf for a given distribution with fixed model parameters.
- **f(**D | θ**)** = **f(**data|distribution**)** : **probability density of observed data given a model with fixed parameters**
- **Likelihood**: A measure of the extent to which a sample provides support for particular values of a parameter
- **L(**θ | D**)** = **L(**distribution|data**)** : **likelihood of model parameters taking certain values given the observed data**

f(D | θ) = L(θ | D) with different interpretations

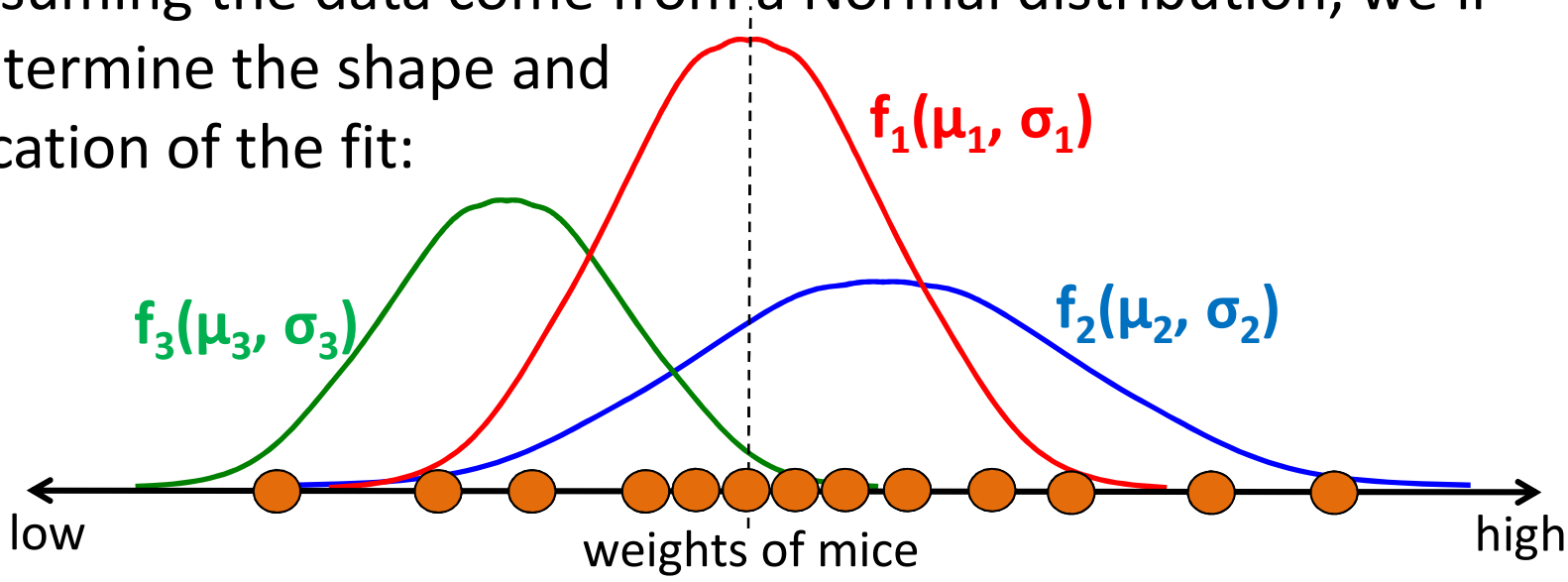
(About data : quantifies anticipation of outcome)

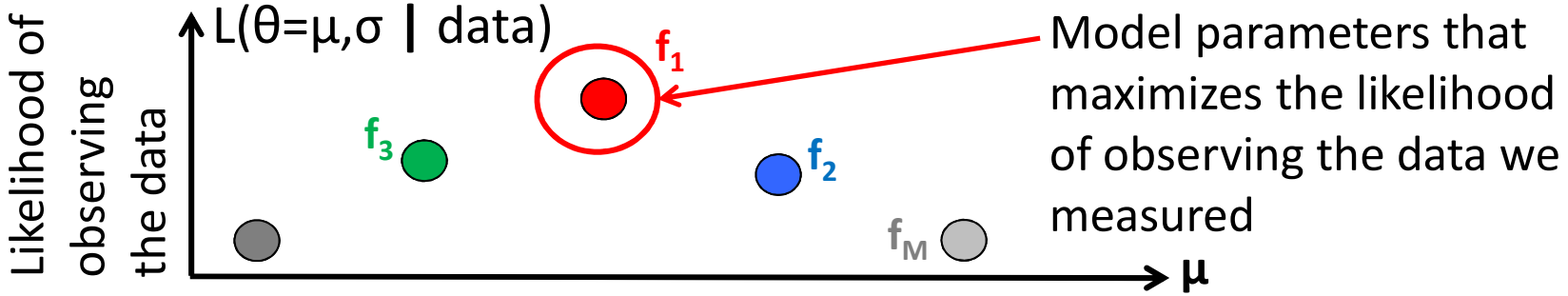
About parameters: quantifies trust in the model)

- MLE provides the parameter value(s) that make he observed sample the most likely one among all possible samples

Maximum Likelihood Estimation (MLE)

- MLE fits an optimal distribution to data and determines the values of the parameters of the model.
- **Example**: We have measurements for the weights of 13 mice. Assuming the data come from a Normal distribution, we'll determine the shape and location of the fit:





MLE – cont'd

- Is there a formal way to do this?
- Assume $f(x_1, x_2, ..., x_n \mid \theta)$ is the joint probability function for n random variables with sample values $x_1, x_2, ..., x_n$

Prob of an event **x** given model parameters **θ** : **p(x | θ)**

Knowing parameters => Prediction of outcome

Likelihood of the parameters **θ** given data **x** : **L(θ | x)**

Observation of data => Estimation of parameters

- L is a function of θ for fixed sample values
- If $x_1, x_2, ... x_n$ are discrete iid random variables with probability function $p(x \mid \theta)$, then the likelihood function L is given by:

$$L(\theta \mid x) = L(\theta \mid x_1, x_2 \ldots x_n)$$

Via independence

$$= L(\theta \mid x_1) L(\theta \mid x_2) \ldots L(\theta \mid x_n)$$

MLE – cont'd

- This could be re-written as:

$$L(\theta \mid x) = L(\theta \mid x_1, x_2 \ldots x_n) = \prod_{i=1}^{n} L(\theta \mid x_i) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

- For the continuous case of f(x | θ): $L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta)$
- MLE estimators are those values of the parameters that maximize L(θ) with respect to parameter θ. One way to do the maximization is to take the derivative. Converting the "multiplication" operator to "summation" is convenient by defining a log-likelihood:

$$\ell = \ln\left[L(\theta)\right] = \ln\left(\prod_{i=1}^{n} f(x_i \mid \theta)\right) = \sum_{i=1}^{n} \ln\left[f(x_i \mid \theta)\right]$$

- To find the model parameters θ that maximizes ln[L], solve for: $\frac{\partial \ell}{\partial \theta} = 0$

MLE – cont'd

- **Example: MLE for the Normal distribution**
- Suppose we have a sample that we want to represent with a Normal distribution **N(μ, σ)**.
- The mean **μ** and variance **σ^2** are the parameters that need to be estimated. The likelihood function is

$$L(\mu,\sigma^2 \mid x_1,x_2,\ldots x_n) = \prod_{i=1}^{n} f_X(x_i \mid \mu,\sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

- Taking the natural log of the likelihood function, we get the log-likelihood:

$$\ell = \ln(L) = \ln\left((2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\right)$$

$$= \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

MLE – cont'd

- **Example: MLE for the Normal distribution (cont'd)**
- First order conditions for the maximization problem becomes:

$$\ell = \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

- First, derivative of log-likelihood with respect to **μ** :

$$\frac{\partial \ell}{\partial \mu} = \frac{2\mu}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \;\; => \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\mu = \sum_{i=1}^{n} x_i - n\mu = 0$$

$$\mu_{\mathbf{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Then, derivative of log-likelihood with respect to **σ** :

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu^2) = 0 \;\; => -n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu^2) = 0$$

$$\sigma_{\mathbf{MLE}}{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu^2)$$

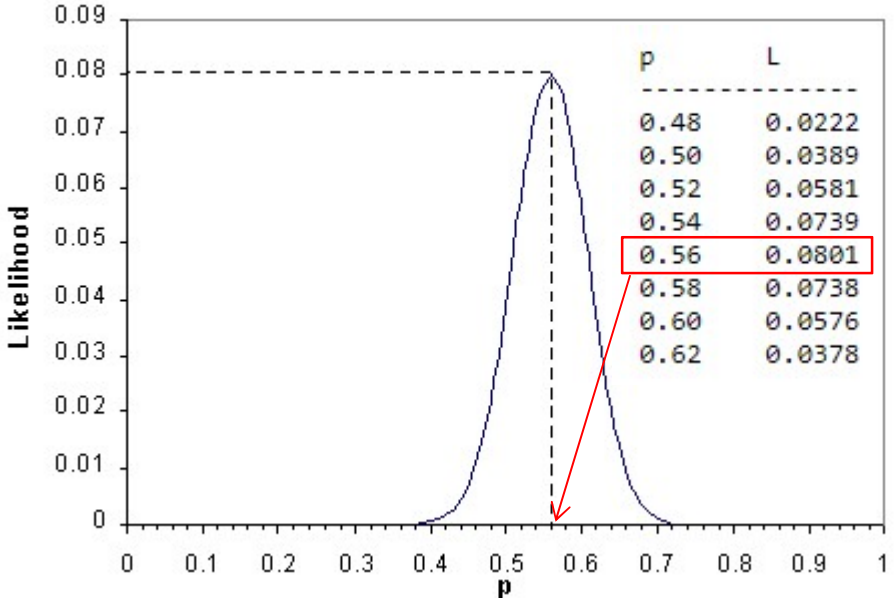
MLE example I

- We flip a coin 100 times and observe 56 Heads. Instead of assuming p=0.5, we want to find the **MLE for the probability p of heads in a single toss**.
- Value for p that makes the observed data most likely?
- Likelihood based on the Binomial model for the probability of 56 heads in 100 tosses:

$$L(p \mid data) = \binom{100}{56} p^{56}(1-p)^{44}$$

- Tabulate the likelihood for different parameter values to find the maximum likelihood estimate of *p*:

Likelihood surface for a range of possible values for *p*



MLE example I – cont'd

- Given data, the MLE for the parameter p is the value of p that maximizes the likelihood of data L(p|data). The MLE is the value of p for which the data is most likely:

$$\frac{dL}{dp} = \binom{100}{56}\left[56p^{55}(1-p)^{44} - 44p^{56}(1-p)^{43}\right] = 0$$

$$\Rightarrow 56p^{55}(1-p)^{44} = 44p^{56}(1-p)^{43}$$

$$\Rightarrow 100p = 56 \quad \Rightarrow \quad \hat{p} = \frac{56}{100} = 0.56$$

MLE for p is exactly the fraction of heads we have in our data!

- The best estimate for *p* from any one sample is clearly the proportion of heads observed in that sample. In a similar way, the best estimate for the population mean will always be the sample mean.
- Surely noone will use MLE for such a simple problem, but the point is, not all problems are this simple!

MLE example II

- Back to our previous problem (this time using **MLE**):

x1	x2	x3
0.4	0.7	0.9

$$f(x) = \theta x^{\theta-1} \quad for\ 0 < x < 1$$

- The likelihood function:

$$L(\theta, x) = \left[\theta x_1^{\theta-1}\right]\left[\theta x_2^{\theta-1}\right]\ldots\left[\theta x_3^{\theta-1}\right] = \theta^3 (x_1 x_2 x_3)^{\theta-1} = \theta^3 \prod_{i=1}^{3} x_i^{\theta-1}$$

- Take log of both sides for easy manipulation

$$\ln(L) = \ell = 3\ln\theta + (\theta-1)\sum_{i=1}^{3}\ln(x_i) \quad \text{then} \quad \frac{\partial \ell}{\partial \theta} = \frac{3}{\theta} + \sum_{i=1}^{3}\ln(x_i) = 0$$

$$\hat{\theta} = \frac{-3}{\sum_{i=1}^{3}\ln(x_i)} = \frac{-3}{\ln(0.4)+\ln(0.7)+\ln(0.9)} = \frac{3}{1.378} = 2.18$$

MLE solution

$$f(x) = 2.18\, x^{1.18} \quad for\ 0 < x < 1$$

MoM solution

$$f(x) = 2x \quad for\ 0 < x < 1$$

MoM vs MLE – Pros and Cons

- **MoM**
  - Simple and intuitive, easy to compute and always works
  - Consistent
  - Usually not the best estimators available (in achieving MSE)
  - May not be unique => For Poisson dist: $E(x_i) = Var(x_i) = \lambda$, you get two different estimators from the 1^{st} and 2^{nd} moments
  - Sometimes they could be meaningless
- **MLE**
  - When sample size is large (> 30), MLE is unbiased, normally distributed, consistent and efficient (minimum MSE)
  - Can be highly biased for small samples
  - May not always have a closed-form solution (need numerical solutions: e.g. Expectation Maximization)
  - Can be sensitive to starting values (no global optimum)

Problems with point estimators

- Point estimator is a single number that estimates the population parameter and varies from sample to sample. So it's almost guaranteed that we'll miss the actual true mean for the population (by some but hopefully not much).
- So we don't have a measure of how certain or confident we are that we actually got the true mean. We need a confidence interval in which the estimator has a likelihood of hitting the true mean, such as:

> We're 90% confident that our estimate interval $(\bar{x}-\epsilon, \bar{x}+\epsilon)$ includes the true unknown (but fixed) value of population mean μ, where ϵ is some margin of error.

- More on "confidence intervals" in the next lecture.

Properties of Expected values and variance for the sum of Random Variables:

$$E(a) = a \qquad E(bX) = bE(X) \qquad E(a+X) = a + E(X)$$

$$E(a+bX+cX^2) = a + bE(X) + cE(X^2)$$

$$Var(X) = \sigma^2 = E[(x-\mu)^2] = E[x^2 - 2x\mu + \mu^2] = E(x^2) - \mu^2$$

$$Var(aX+bY) = E[(aX+bY) - (a\mu_x + b\mu_y)]^2 = E[a(X-\mu_x) + b(Y-\mu_y)]^2$$

$$Var(aX+bY) = E[a^2(X-\mu_x)^2 + 2ab(X-\mu_x)(Y-\mu_y) + b^2(Y-\mu_x)^2]$$

$$Var(aX+bY) = a^2 \underbrace{E(x-\mu_x)^2}_{Var(X)} + \underbrace{2abE[(X-\mu_x)(Y-\mu_y)]}_{Cov(X,Y)} + b^2 \underbrace{E(y-\mu_y)^2}_{Var(Y)}$$

$$Var(aX+bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$$

0, if X and Y are independent

$$Var(aX+bY) = a^2Var(X) + b^2Var(Y)$$

$$Var(X+Y) = Var(X) + Var(Y)$$

$$Var(X-Y) = Var(X) + Var(Y)$$