

# Stochastic Modelling and Processes

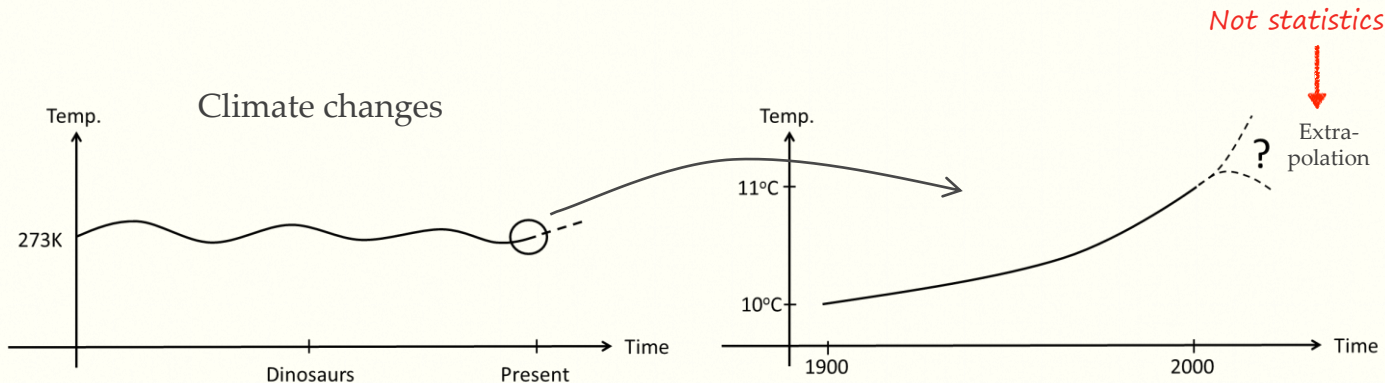
Point Estimation and  
sampling

This is a “PowerPoint” presentation



# Descriptive Statistics

- **Descriptive statistics** summarizes data from a sample using parameters such as the mean or standard deviation.
- **Descriptive statistics** also includes tables, graphs, etc.



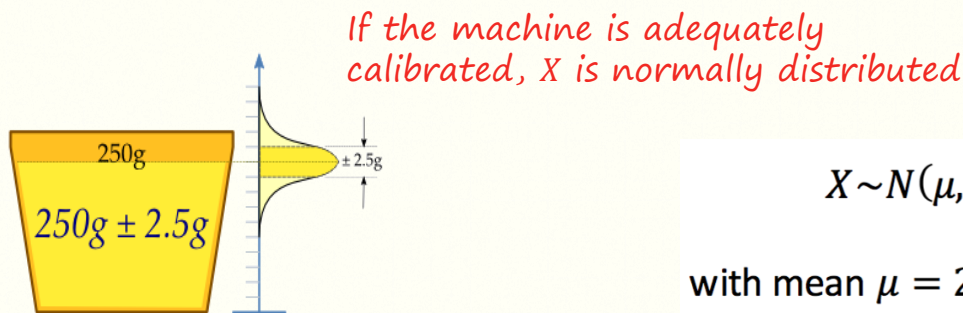
# Inferential Statistics

- *Inferential statistics* infers predictions about a larger population than the sample represents.
- It uses patterns in the sample data to draw inferences about the population represented, accounting for randomness.
- These inferences may take the form of:
  - answering yes/no questions about the data (*hypothesis testing*)
  - estimating numerical characteristics of the data (*estimation*)
  - describing associations within the data (*correlation*)
  - modeling relationships within the data (*regression analysis*).



# Cup Example

- ❖ A machine fills cups with a liquid, the content of the cups is 250 grams of liquid.
- ❖ The machine cannot fill with exactly 250 grams, the content added to individual cups shows some variation, and is considered a random variable,  $X$ .



$$X \sim N(\mu, \sigma^2)$$

with mean  $\mu = 250$  g and  
standard deviation  $\sigma = 2.5$  g

# Cup Example

*ONE sample of the population!*


- To determine if the machine is adequately calibrated, a sample of  $n = 25$  cups of liquid is chosen at random and the cups are weighed.
- The resulting measured masses of liquid are  $X_1, X_2, \dots, X_{25}$ , a random sample from  $X$ . *Population — All cups for all times*
- To get an impression of the population mean ( $\mu$ ), we use the average (or sample mean) as an estimate:

*Sample mean is NOT the expected value (true mean)!*

*$\hat{\mu}$  means an estimator of the true population value*  $\longrightarrow \hat{\mu} = \frac{1}{25} \sum_{i=1}^{25} X_i = 250.2 \text{ g}$

- Is the machine adequately calibrated?

# Cup Example

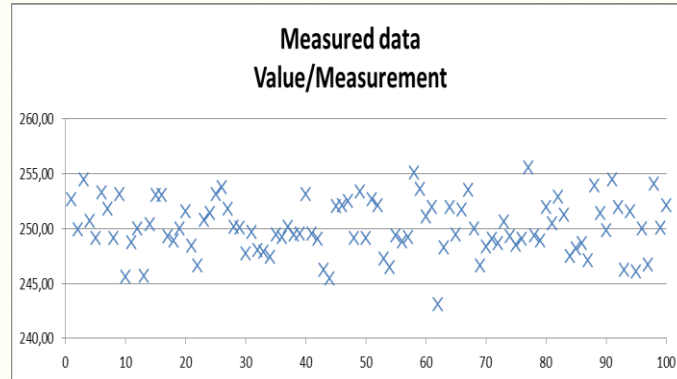
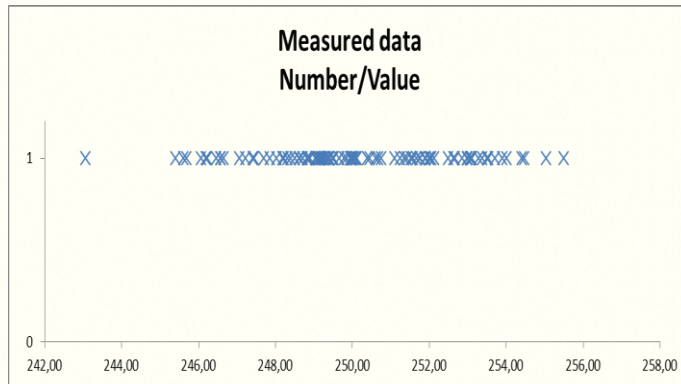
- What would we typically observe, if the machine is indeed adequately calibrated?
- Perform 100 simulations of the cup filling experiment, where in each experiment we assume that the machine is adequately calibrated.  100 samples of the population – fx. in Python
- The result of each experiment is a sample mean,  $\hat{\mu}$ , which is calculated by first drawing 25 random samples  $(X_1, X_2, \dots, X_{25})$  from a normal distribution with mean  $\mu = 250$  grams and standard deviation  $\sigma = 2.5$  grams.



# Histograms

- Presentation of data

252,6	248,7	248,4	249,7	249,5	252,7	251,9	249,1	250,5	254,4
249,9	250,0	246,6	248,0	249,0	252,1	243,1	248,6	252,9	251,9
254,5	245,7	250,8	247,9	246,2	247,2	248,2	250,6	251,2	246,2
250,7	250,4	251,4	247,4	245,4	246,5	251,9	249,3	247,4	251,5
249,1	253,0	253,1	249,4	252,0	249,4	249,4	248,5	248,2	246,1
253,3	253,0	253,7	249,2	252,1	248,8	251,7	249,1	248,6	249,9
251,8	249,3	251,8	250,1	252,5	249,2	253,5	255,5	247,1	246,7
249,1	248,9	250,1	249,4	249,1	255,0	250,0	249,3	253,9	254,0
253,1	250,0	250,0	249,5	253,4	253,5	246,6	248,9	251,3	250,0
245,6	251,6	247,7	253,1	249,1	251,1	248,3	251,9	249,8	252,1

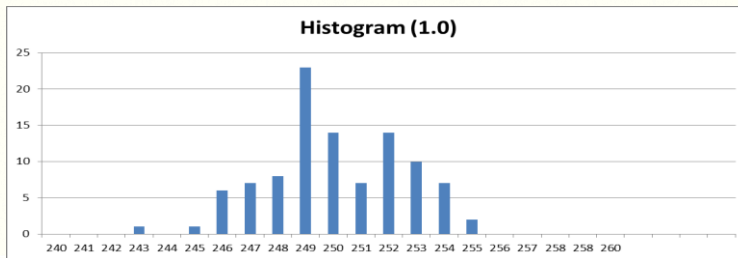
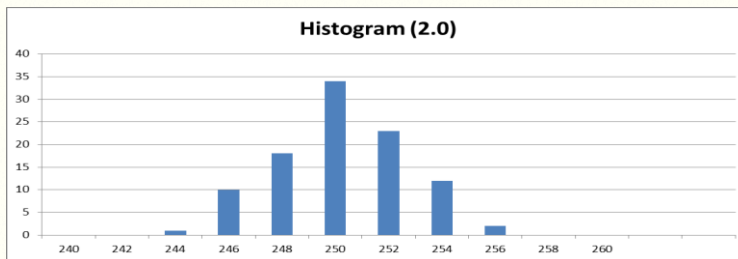
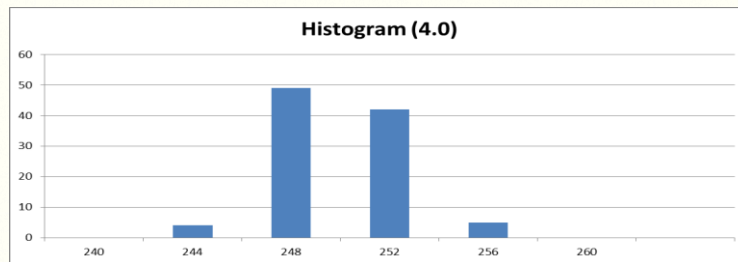




# Histograms

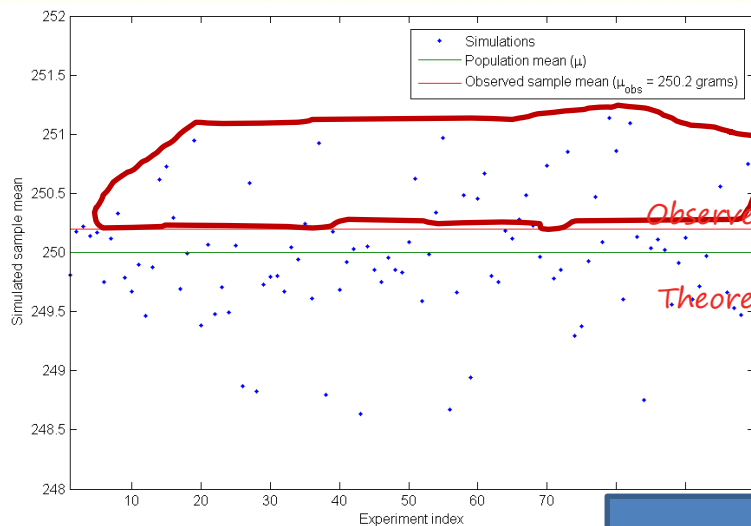
- Grouping of data

Interval	Center	Freq
239-241	240	0
241-243	242	0
243-245	244	1
245-247	246	10
247-249	248	18
249-251	250	34
251-253	252	23
253-255	254	12
255-257	256	2
257-259	258	0
259-261	260	0



# Cup Example

- What would we **typically** observe, if the machine is adequately calibrated?



28 of the 100 simulated values are more extreme than our observation.

Observed value

Theoretical value

**Conclusion:** If the machine is adequately calibrated, then due to random variation, we will often observe a sample mean (or average) that is more extreme than the actually observed value of 250.2 g.

# Cup Example

- A loose interpretation of this result is that – just by random variation in  $X$  – there is a 28% chance of observing a sample mean that is larger than what we have observed (250.2 g), even though the machine is adequately calibrated.
- This means that observing a sample mean of 250.2 g or larger, when the machine is adequately calibrated, is a relatively common event.
- Hence, the 0.2 gram deviation from the hypothesized mean ( $\mu = 250$  g) can be explained by random variation in  $X$ , and **we conclude that the machine is adequately calibrated.**

  
*it is likely that*

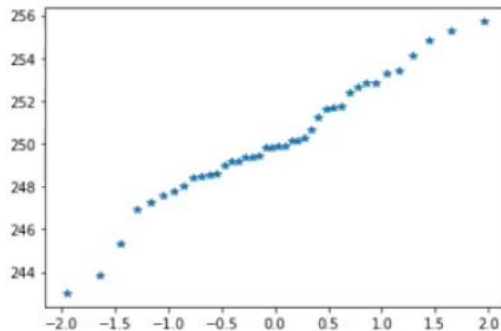
# Probability Plot

- `stats.probplot(data, dist=stats.norm, plot=plt)`

```
In [34]: import scipy.stats as stats
import numpy as np
import matplotlib.pyplot as plt

n = 40
prob = np.linspace(1/40, 1-1/40, 40)
ideal = stats.norm.ppf(prob, scale=1, loc=0)
X = np.random.randn(40)*2.5+250
X.sort()
plt.plot(ideal, X, '*')
```

Out[34]: [matplotlib.lines.Line2D at 0x2e42e56ecc8]





# Population

*Key term*

- **Definition 1 - Population**

- In statistics, a **population** is a complete set of items that share at least one property in common that is the subject of a statistical analysis.
- Populations can be diverse topics such as “all persons living in a country” or “every atom composing a crystal”.
- In the cup filling example, the population is “all cups filled by the machine”.

*Key term*

## ▪ **Definition 2 - Sample**

- A statistical sample is a subset drawn from the population to represent the population in a statistical analysis.
- If a sample is chosen properly, characteristics of the entire population that the sample is drawn from can be inferred from corresponding characteristics of the sample.
- We will denote the sample  $X_1, X_2, \dots, X_n$ , where the samples are iid random variables with a given probability distribution.

## ▪ **Definition 3 – Statistical model**

- A random sample  $X$  and its PDF,  $f_X(x; \theta)$ , where  $\theta$  is the parameter of the PDF.
- The parameter,  $\theta$ , is in general a vector and often unknown.
- If  $f_X(x; \theta)$  is a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , then  $\theta = [\mu, \sigma^2]$ .
- The purpose of inferential statistics is to infer knowledge about the unknown parameter(s),  $\theta$ .



- **Definition 4 – Statistic**

- A *statistic* is a random variable that is a function of the random sample,  $X$ , but not a function of unknown parameters,  $\theta$ .

- A commonly used statistic is the average or sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$



- **Definition 5 – Estimator**

- An estimator,  $\hat{\theta}(X)$ , is a statistic used to estimate the unknown parameter  $\theta$  of a random sample,  $X$ .
- For notational convenience, we will often write  $\hat{\theta}$  instead of  $\hat{\theta}(X)$ .
- We have already seen an example of an estimator, namely the sample mean

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$

- which is an estimator of the (true) population mean,  $\mu$ , of a normally distributed random variable,  $X$ .

# Point Estimation

- A **point estimate** is a reasonable value of a population parameter.
- $X_1, X_2, \dots, X_n$  are **random variables**.
- Functions of these random variables,  $\bar{x}$  and  $s^2$ , are also random variables called **statistics**.
- Statistics have their unique distributions which are called **sampling distributions**.

# Some Parameters & Their Statistics

Estimation problems occur frequently in engineering. We often need to estimate

- The mean  $\mu$  of a single population
- The variance  $\sigma^2$  (or standard deviation  $\sigma$ ) of a single population
- The proportion  $p$  of items in a population that belong to a class of interest
- The difference in means of two populations,  $\mu_1 - \mu_2$
- The difference in two population proportions,  $p_1 - p_2$

Reasonable point estimates of these parameters are as follows:

- For  $\mu$ , the estimate is  $\hat{\mu} = \bar{x}$ , the sample mean.
- For  $\sigma^2$ , the estimate is  $\hat{\sigma}^2 = s^2$ , the sample variance.
- For  $p$ , the estimate is  $\hat{p} = x/n$ , the sample proportion, where  $x$  is the number of items in a random sample of size  $n$  that belong to the class of interest.
- For  $\mu_1 - \mu_2$ , the estimate is  $\hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2$ , the difference between the sample means of two independent random samples.
- For  $p_1 - p_2$ , the estimate is  $\hat{p}_1 - \hat{p}_2$ , the difference between two sample proportions computed from two independent random samples.

- There could be choices for the point estimator of a parameter.
- To estimate the mean of a population, we could choose the:
  - Sample mean.
  - Sample median.
  - Average of the largest & smallest observations in the sample.
  - Etc.



# Maximum-likelihood Estimator

- Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model.
- For a given distribution with the parameter  $\hat{\theta}$  and the pdf  $f_X(x|\hat{\theta})$  the likelihood function is:

$$L(\hat{\theta}) = f_X(x|\hat{\theta})$$

- Important:
  - The likelihood function  $L(\hat{\theta})$  is a function of the parameter estimate  $\hat{\theta}$  and not  $x$
  - The likelihood function  $L(\hat{\theta})$  is the probability density of observing  $x$  if the true parameter is  $\hat{\theta}$
- **The optimum (MLE) estimate of the parameter is the one that maximizes  $L(\hat{\theta})$ .**



# Maximum-likelihood Estimate of the Mean

- Given independent observations,  $x_1, x_2, \dots, x_n$ , the likelihood function given the true variance ( $\sigma^2$ ) and some choice of mean ( $\hat{\mu}$ ) is

*Likelihood*

$$L(\hat{\mu}) = f(x_1, x_2, \dots, x_n | \hat{\mu}, \sigma^2) = f(x_1) \cdot f(x_2) \cdot \dots \cdot f(x_n) = \prod_{i=1}^n f(x_i)$$

*i.i.d. data*

*Multiplication*

- where

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x_i - \hat{\mu})^2 / \sigma^2}$$

*in Gaussian data*

- Inserting, we get

$$L(\hat{\mu}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x_i - \hat{\mu})^2 / \sigma^2}$$

# Maximum-Likelihood Estimate of the Mean

- We have,

*i.i.d. Gaussian data*

$$L(\hat{\mu}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x_i - \hat{\mu})^2 / \sigma^2}$$

- Ignoring constants and taking the logarithm, we see that the optimum choice of  $\hat{\mu}$  must maximize

$$-\sum_{i=1}^n (x_i - \hat{\mu})^2$$

*Called: the log-likelihood*

- Differentiating with respect to  $\hat{\mu}$  and setting equal to zero, we get the desired result

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

# Maximum-likelihood Estimator for Binomial Data

- For the binomial distribution the likelihood function is:

$$L(\hat{p}) = f_{\text{bino}}(x|\hat{p}) = \binom{n}{x} \hat{p}^x (1 - \hat{p})^{n-x}$$

where  $f_{\text{bino}}(x|\hat{p})$  denotes the binomial pdf, and  $x$  is the observed number of successes out of  $n$  trials.

- When maximizing, we can ignore the binomial coefficient  $\binom{n}{x}$ , because it does not depend on  $\hat{p}$ .
- Since the logarithm is a monoton function, the choice of  $\hat{p}$  that maximizes  $L(\hat{p})$  also maximizes  $\log(L(\hat{p}))$
- So, the optimum  $\hat{p}$  maximizes:  $\log(\hat{p}^x (1 - \hat{p})^{n-x}) = x \cdot \log(\hat{p}) + (n - x) \cdot \log(1 - \hat{p})$
- Differentiate with  $\hat{p}$  and setting equal to zero:  $\frac{x}{\hat{p}} - \frac{n-x}{1-\hat{p}} = 0 \Rightarrow \hat{p} = \frac{x}{n}$
- Hence, the parameter estimate from the  $n$  trials, is exactly the maximum-likelihood estimate.



# Unbiased Estimator

Key term

- **Definition 6 – Unbiased estimator**

- An estimator,  $\hat{\theta}$ , is unbiased if

$$E[\hat{\theta}] = \theta$$

When an estimator is unbiased, the bias is zero; that is,  $E(\hat{\Theta}) - \theta = 0$ .

- i.e., if it's expected value is equal to the true value of the unknown parameter being estimated.

- The sample mean

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$

- is an unbiased estimator of the population mean,  $\mu$ , of a normally distributed random variable,  $X$ .

- The sample succesrate  $\hat{p} = \frac{x}{n}$  is also an unbiased estimator:

$$E[\hat{p}] = E\left[\frac{x}{n}\right] = \frac{1}{n} \cdot E[x] = \frac{1}{n} \cdot np = p$$



# The Sample Variance

Key term

- Let  $X_1, X_2, \dots, X_n$  be i.i.d. samples of a random variable  $X$  with mean  $\mu$  and variance  $\sigma^2$ . Then the maximum-likelihood estimate of the variance is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- This is a biased estimator

$$E[\hat{\sigma}^2] = \sigma^2 - \frac{1}{n} \sigma^2 \neq \sigma^2$$

- The unbiased estimate of the variance is

*n-1 degrees  
of freedom*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- We refer to this unbiased estimator as the *sample variance* or *empirical variance*.

# Standard Error: Reporting a Point Estimate

## Standard Error of an Estimator

The **standard error** of an estimator  $\hat{\Theta}$  is its standard deviation given by  $\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$ . If the standard error involves unknown parameters that can be estimated, substitution of those values into  $\sigma_{\hat{\Theta}}$  produces an **estimated standard error**, denoted by  $\hat{\sigma}_{\hat{\Theta}}$ .

Suppose that we are sampling from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Now the distribution of  $\bar{X}$  is normal with mean  $\mu$  and variance  $\sigma^2/n$ , so the **standard error** of  $\bar{X}$  is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

If we did not know  $\sigma$  but substituted the sample standard deviation  $S$  into the preceding equation, the **estimated standard error** of  $\bar{X}$  would be

$$SE(\bar{X}) = \hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}$$

# Central Limit Theorem

- Let  $X_1, X_2, \dots, X_n$  be i.i.d. samples of a random variable  $X$  with mean  $\mu$  and variance  $\sigma^2$ .
- Then, as  $n \rightarrow \infty$ , the sample mean ( $\bar{X}$ ) becomes normally distributed with a mean that is equal to the population mean ( $\mu$ ) and a variance that is scaled by  $1/n$ :

*Sample mean*

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

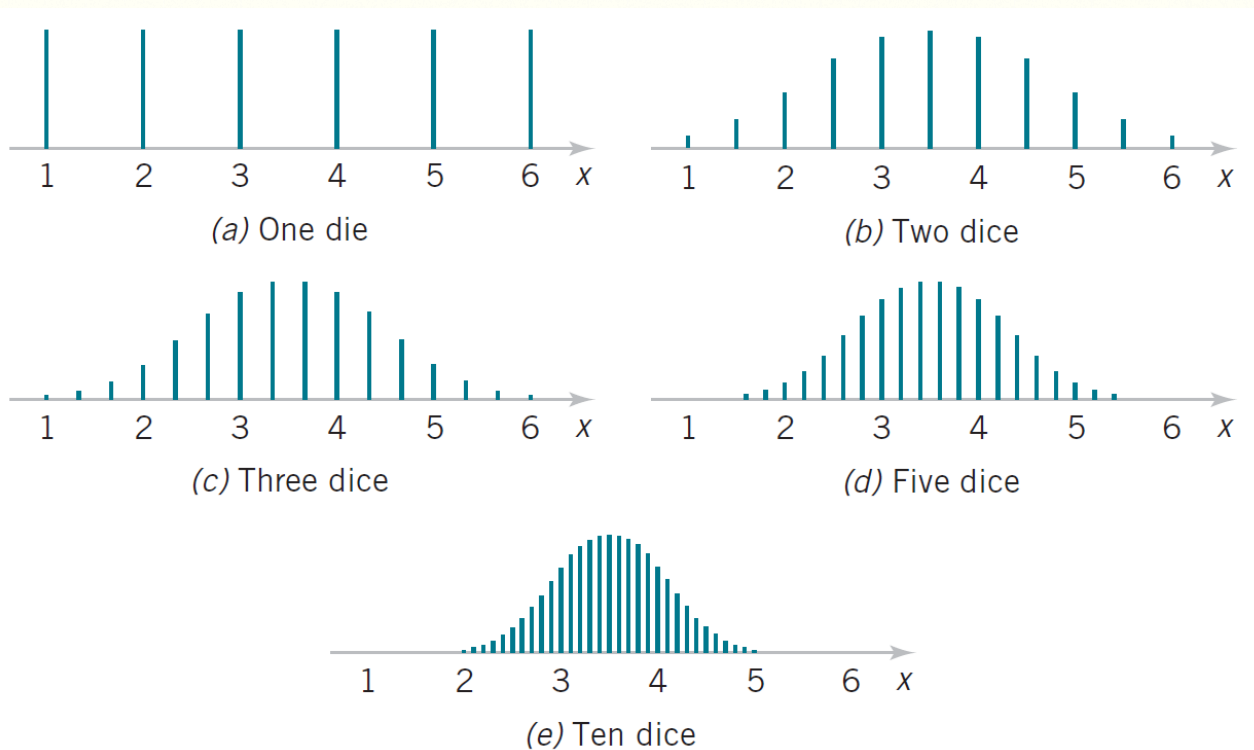
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- Note that  $X$  can have any distribution, i.e., it is *not* required to be normally distributed.
- Although the exact number is subject of debate, it is common practice to require that the number of samples ( $n$ ) should be 30 or larger in order to apply the CLT:

$$n \geq 30$$

*...most important number in statistics*

# Central Limit Theorem



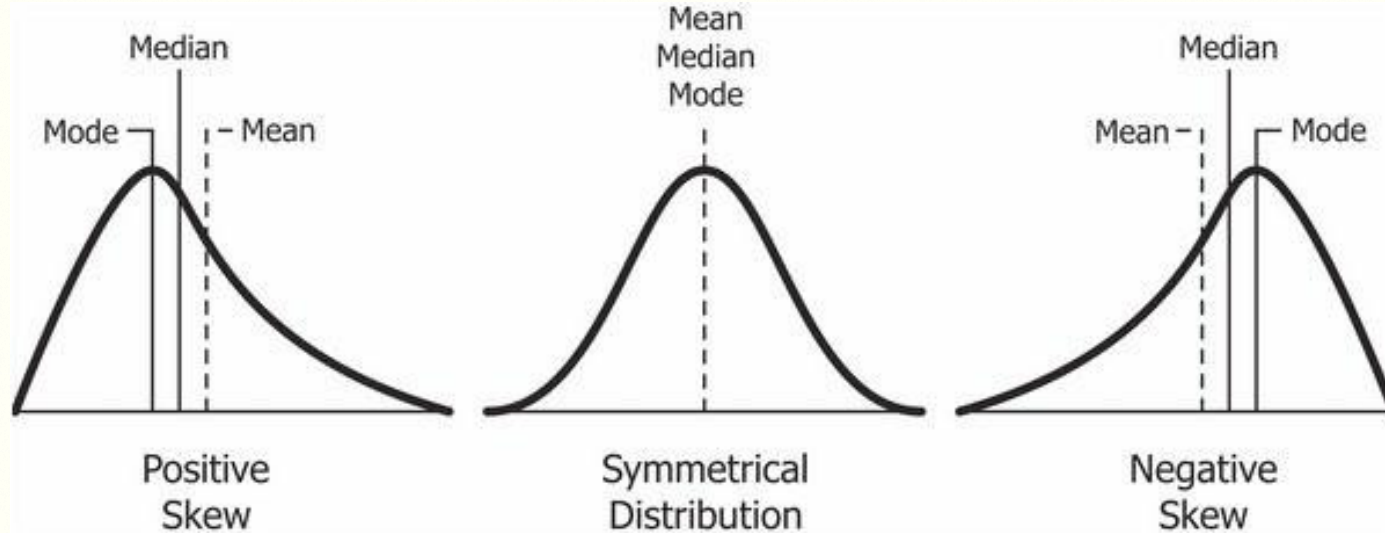


# Skewness and kurtosis

**Positive (right) Skewed:**  $SkewnessCoeff > 0.5$

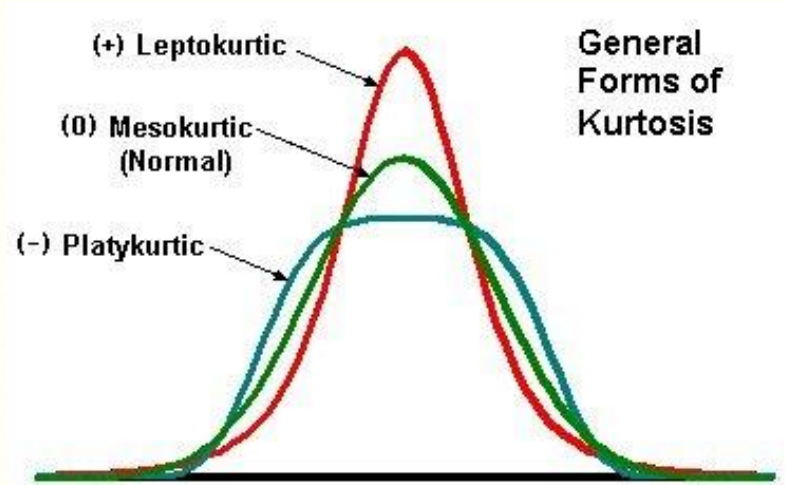
**Negative (left) Skewed:**  $SkewnessCoeff < -0.5$

**Symmetrical:**  $-0.5 < SkewnessCoeff < 0.5$



$$SkewnessCoeff = \frac{Mean - Mode}{standard\ deviation}$$

# Skewness and kurtosis



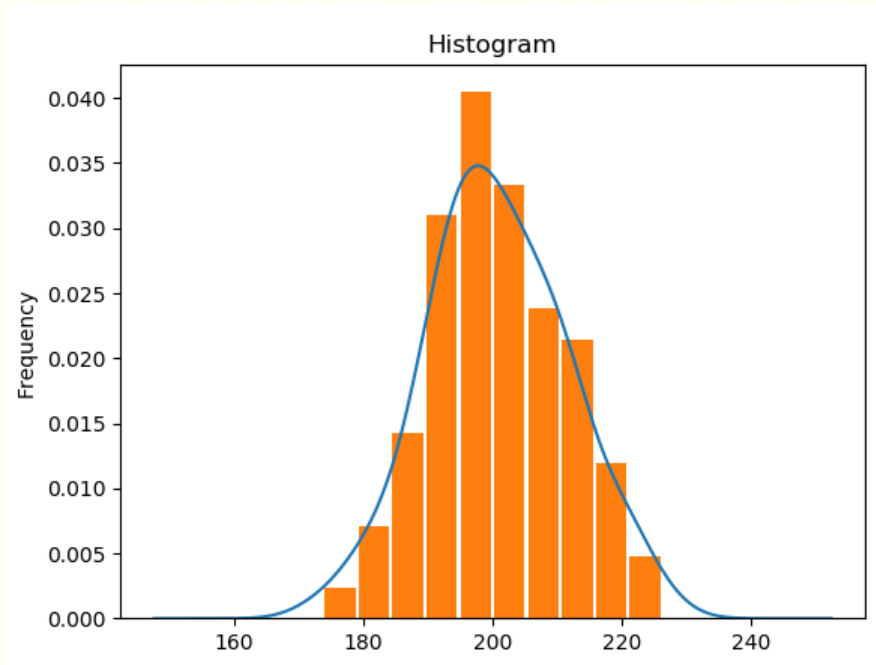
Mesokurtic:  $-0.5 < \text{kurtosis} < 0.5$

Platykurtic:  $\text{kurtosis} < -0.5$













Leptokurtic:  $\text{kurtosis} > 0.5$

# Kernel Density Estimate

- A kernel density estimation (KDE) is a way to estimate the probability density function (PDF) of the random variable that “underlies” our sample. KDE is a means of data smoothing



File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

          Code  

In [19]:

```
import numpy as np
from scipy import stats

X = np.random.randn(45)*3+6
print(X)

print(stats.skew(X))
print(stats.kurtosis(X))
```

3.56074839	8.31262024	5.66618403	9.29599304	3.25619483	6.66260732
10.87918384	5.49379615	7.3149702	6.9694492	5.3164123	6.53059108
4.40343382	1.76805045	9.32697166	6.39677529	3.61218036	4.01757276
6.71753634	8.70419848	8.24769716	4.37149665	4.2311889	6.83194732
7.21489326	3.37343837	5.04044427	3.6395873	9.21214272	5.76428683
6.65843516	1.32236591	5.83910126	6.82359435	9.53551964	4.59246859
3.37967734	9.72919267	6.79789771	5.34118628	6.47524681	5.89308625
9.72975818	3.78691562	6.81084406]			
0.07441329266320276					
-0.5748628923165424					



# Mean Squared Error of an Estimator

The mean squared error is equal to the variance of the estimator plus the squared bias

## **Mean Squared Error of an Estimator**

The **mean squared error of an estimator**  $\hat{\Theta}$  of the parameter  $\theta$  is defined as

$$\text{MSE}(\hat{\Theta}) = E(\hat{\Theta} - \theta)^2 \quad (7.8)$$

The mean squared error can be rewritten as follows:

$$\begin{aligned} \text{MSE}(\hat{\Theta}) &= E[\hat{\Theta} - E(\hat{\Theta})]^2 + [\theta - E(\hat{\Theta})]^2 \\ &= V(\hat{\Theta}) + (\text{bias})^2 \end{aligned}$$