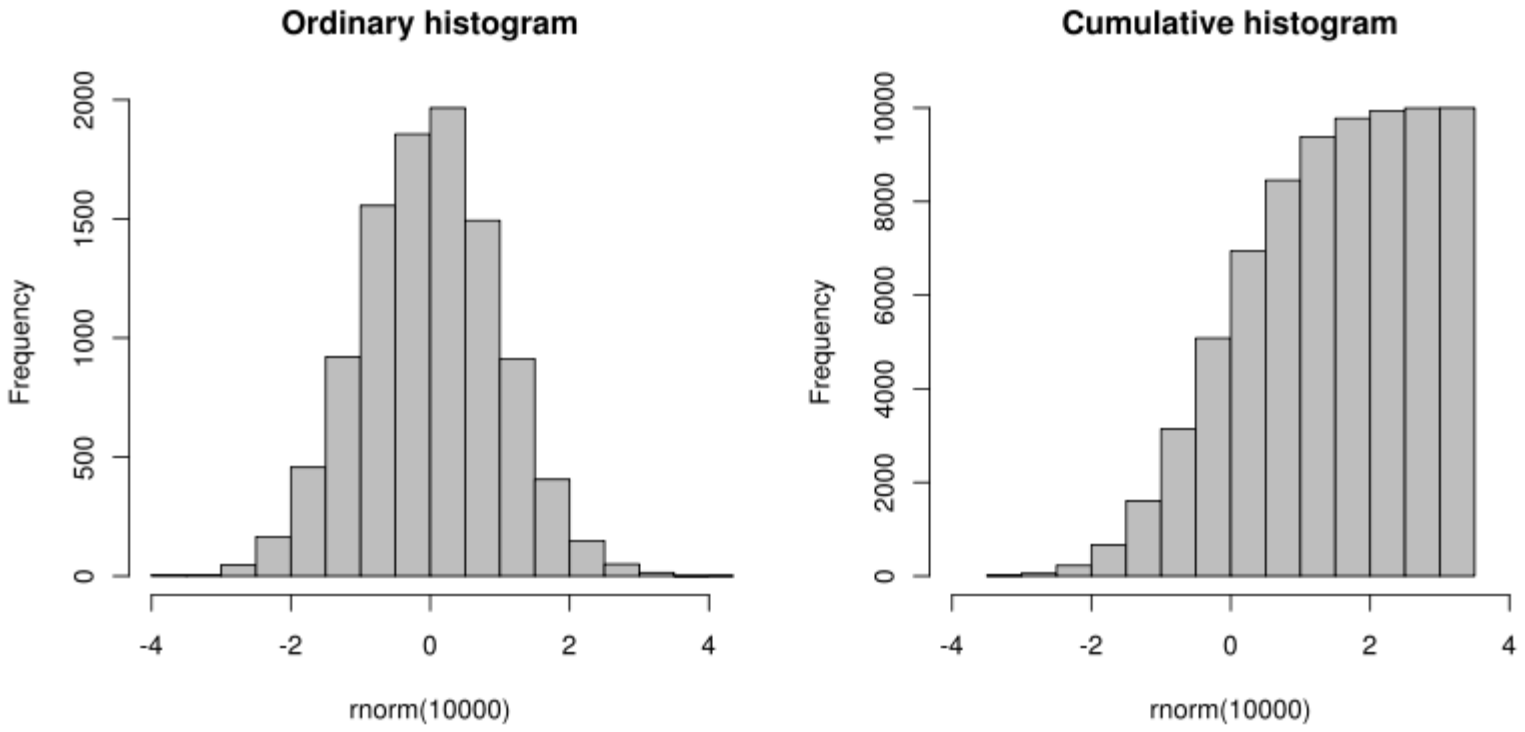
BASICS ON DISTRIBUTIONS

Histograms

Consider an experiment in which different outcomes are possible (ex. Dice tossing) . The probability of all the outcomes can be represented in an histogram



Distributions

Discrete distribution

- Probabilities are described with distributions
  - Discrete variables: data are valued on a discrete set: $x \in I=\{x_i, i \subseteq N\}$
    - Probability distribution: $f(x_i)$ = probability for the value x_i
    - Cumulative distribution: $F(x_i)$ = probability for a value $\leq x_i$
    - Normalization: $\sum_{i \in I} f(x_i) = 1$

$$F(x_{HIGHEST}) = 1$$

Examples

Consider a coin with a tail probability equal to p

Flip a coin three times.
Plot the probability distribution for the number of tails

you have to find the probability of getting anyone of the 4 possible result
P(t0)=(1-p)^3
p(T1)=3P(1-p)^2
p(t2)=3*P^2*(1-p)
P(t3)=P^3

Flip a coin until the first tail appears.
Plot the probability distribution for the number of flips

P(1)=p
P(2)=(1-p)p
p(3)=(1-p)^2*P
P(n)=(1-p)^(n-1)*p
the process stop when we get the first tail

Empirical Distributions

- Experimental data are represented with normalized histograms
  - Discrete variables: data are valued on a discrete set: $x \in I = \{x_i, i \subseteq \mathbb{N}\}$
    - Probability distribution: $f(x_i) = N(x=x_i)/N_{tot}$
    - Cumulative distribution: $F(x_i) = N(x \leq x_i)/N_{tot}$
    - Normalization: $\sum_{i \in I} f(x_i) = 1$

Characterization: Mode

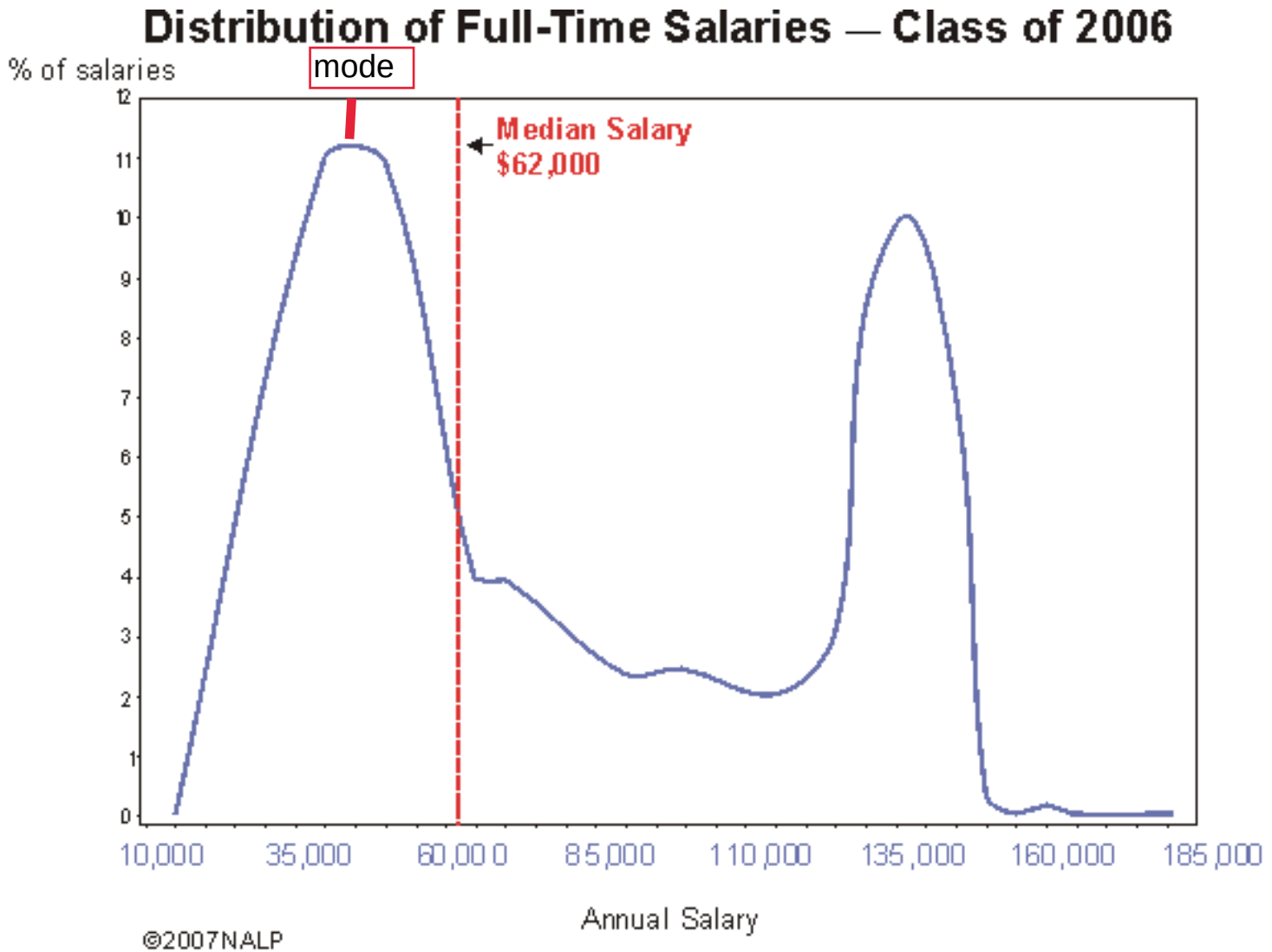
- The mode is the value that occurs the most frequently in a probability distribution.
- For empirical distributions, the mode is the value that most frequently occurs in a data set.
- More than one mode can be present

Characterization: Median

split the probability in two half so 50% to get a result on either side

- a **median** is described as the numeric value separating the higher half of a sample, a population, or a probability distribution, from the lower half. The *median* of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one. If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values.

Characterization: Median



Averages (The Median)

The median is the middle value of a set of data once the data has been ordered.

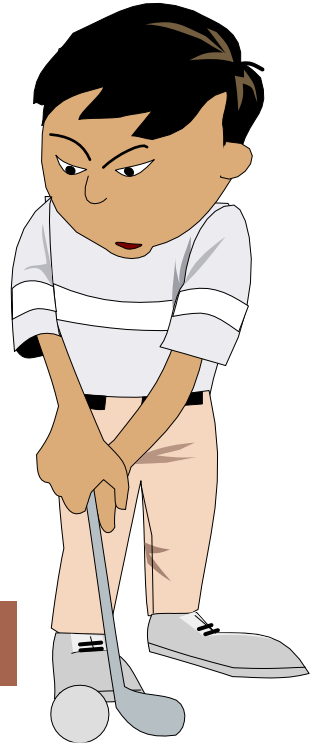
Example 1. Robert hit 11 balls at Grimsby driving range. The recorded distances of his drives, measured in yards, are given below. Find the median distance for his drives.

85, 125, 130, 65, 100, 70, 75, 50, 140, 95, 70

50, 65, 70, 70, 75, 85, 95, 100, 125, 130, 140

Single middle value

Ordered data



Median drive = 85 yards

Averages (The Median)

The median Algorithm.

1 Sort the data

2 If the number of data is odd:

Take the middle value

Else:

Take the average between the two central values

Measuring the Spread with Median

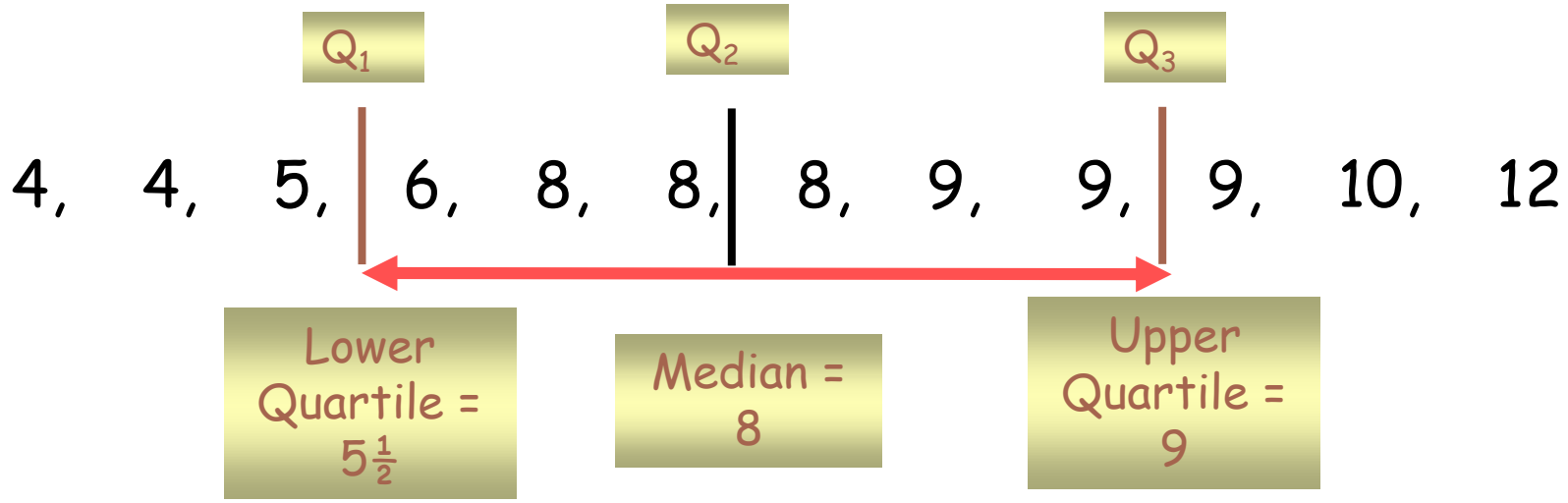
Finding the median, quartiles and inter-quartile range.

quartiles=median of the medians

Example 1: Find the median and quartiles for the data below.

12, 6, 4, 9, 8, 4, 9, 8, 5, 9, 8, 10

Order the data



Inter-Quartile Range = 9 - 5½ = 3½

Expected Value

For a given function g(X) defined over discrete random variables with probability distribution p(X) the expected value of a discrete variable is computed as:

$$E\left[g(X)\right] = \sum_{i \in I} g(x_i)\, p\left(x_i\right)$$

Variance and Moments of a Random Variable

- The 1th moment of a random variable X is $E[X]$
- **Definition**
  - The k th moment of a random variable X is $E[X^k]$
- **Definition**
  - The Variance of a random variable X is defined as

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2.$$

mean of the squares minus the square of the mean

  - The standard deviation of a random variable X is

$$\sigma[\mathrm{X}] = \sqrt{\mathrm{Var}[\mathrm{X}]}.$$

Characterization: Mean

If X is a random variable with probability distribution p(X)

each value* ìthe number of times you got it summed together divaided by the total number of data

- Mean

$$\mu = E[X] = \sum_{i \in I} x_i \, p(x_i)$$

E=expected value(weighted mean)

- For empirical distributions

$$\mu = E[X] = \sum_{i \in I} x_i \frac{N(x_i)}{N_{TOT}} = \frac{\sum_{j \in Data} x_j}{N_{TOT}}$$

Mean and Variance

- Variance

$$Var(X) = E\left[(X-\mu)^2\right] = E\left[X^2 - 2X\mu + \mu^2\right] =$$

$$= E\left[X^2\right] - 2\mu \cdot E\left[X\right] + \mu^2 = E\left[X^2\right] - \mu^2$$

Properties of Expected Values

If X and Y are random variables,

1) $E[X+Y]=E[X]+E[Y]$
2) $E[X+c]=E[X]+c$
3) $E[aX+bY]=aE[X]+bE[Y]$

For example, if $E[X]=5$ and $E[Y]=6$, then

$E[X+5]=5+5=10$

$E[2X+5]=2*5+5=15$

$E[3X+2Y]=3*5+2*6=27$

Prove 1), 2) and 3)

Variance and Moments of a Random Variable

- Proof 1:

$$E[X+Y]=\sum_{i,j}(x_i+y_j)\,p(x_i,y_j)=$$

$$=\sum_i x_i\sum_j p(x_i,y_j)+\sum_j y_j\sum_i p(x_i,y_j)=$$

$$=\sum_i x_i\,p(x_i)+\sum_j y_j\,p(y_j)=E[X]+E[Y]$$

Variance and Moments of a Random Variable

- **Theorem**
  - If X and Y are two independent random variables, then

$$\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y].$$

Variance and Moments of a Random Variable

- Proof:

$$E[X \cdot Y] = \sum_{i,j} x_i y_j p(x_i, y_j) =$$

Only if the events are independent

$$= \sum_{i,j} x_i y_j p(x_i) p(y_j) =$$

$$= \sum_{i} x_i p(x_i) \sum_{i} y_j p(y_j) = E[X] \cdot E[Y]$$

Properties of Expected Values and Variances NB

1. Var(X+Y)=Var(X)+Var(Y), ***if X and Y are independent.***
2. Var(aX)=(a^2)Var(X)
3. Var(X+c)=Var(X)
4. Var(aX+bY)=(a^2)Var(X)+(b^2)Var(Y), ***if X and Y are independent.***

Variance and Moments of a Random Variable

- **Definition** is the variance for two variable
  - The covariance of two random variable X and Y is

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])] =$$
$$= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] =$$
$$= E[XY] - E[X]E[Y]$$

Variance and Moments of a Random Variable

- **Definition**
  - The covariance of two random variable *X* and *Y* is

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

- **Theorem**
  - For any two random variables *X* and *Y*.

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}(X,Y).$$

Variance and Moments of a Random Variable

- Proof:

directly proof by $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$

$$\mathrm{Var}[X+Y] = \mathrm{E}[(X+Y-\mathrm{E}[X+Y])^2]$$

$$= \mathrm{E}[(X+Y-\mathrm{E}[X]-\mathrm{E}[Y])^2]$$

$$= \mathrm{E}[(X-\mathrm{E}[X])^2 + (Y-\mathrm{E}[Y])^2 + 2(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])$$

$$= \mathrm{E}[(X-\mathrm{E}[X])^2 + \mathrm{E}[(-\mathrm{E}[Y])^2 + 2\mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]$$

$$= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}(X,Y)$$

Covariance and Variance

- **Corollary**
  - If *X* and *Y* are independent random variables, then

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

(already proved)

→ THEN

$$\mathrm{Cov}(X,Y) = 0$$

Binomial distribution

Compute the probability of obtaining exactly 3 tails tossing 5 coins. These are all the possibilities

TTTTT	HTTTT	HHTTT	TTHHH	THHHH	HHHHH
	THTTT	HTHTT	THTHH	HTHHH	
	TTHTT	HTTHT	THHTH	HHTHH	
	TTTHT	HTTTH	THHHT	HHHTH	
	TTTTH	THHTT	HTTHH	HHHHT	
		THTHT	HTHTH		
		THTTH	HTHHT		
		TTHHT	HHTTH		
		TTHTH	HHTHT		
		TTTHH	HHHTT		

f(5)=10/32

And what if the probabilities of head and tail are different?

Binomial distribution

How can we count without enumerating?

Total number of configurations: 5 events, with two possibilities each → $2^5=32$

Number of configurations with two heads:
1) choose 3 of the 5 positions: 5 x (5-1) x (5-2) possibilities
2) the order of choice is irrelevant: there are 6 possible permutations of 3 objects (3!)
→5 x 4 x 3/ 6 =5!/(3! X 2!) = 10

And what if the probabilities of head and tail are different?

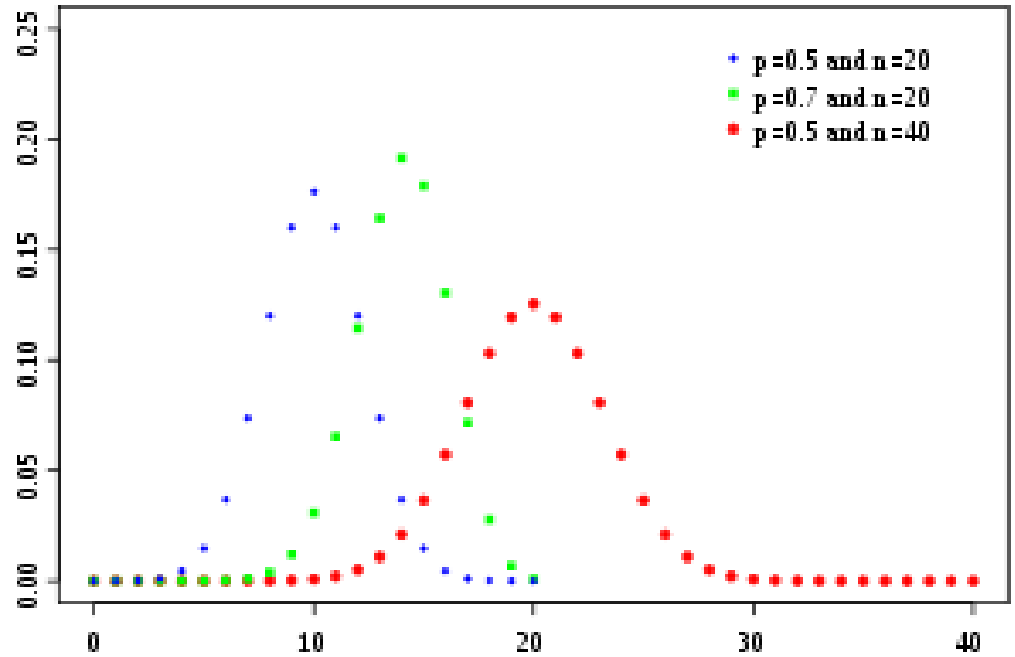
Binomial distribution

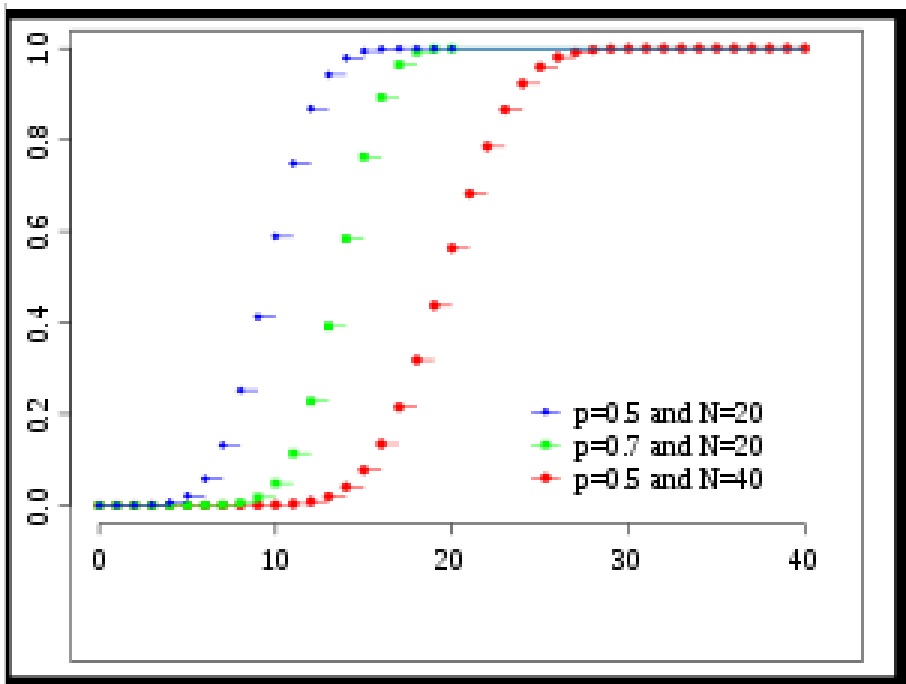
Probability of having k positive events out of n independent experiments.
The probability of a positive event is p

P=probability of single positive event

$$f(k;n,p) = \Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial coefficent=n!
\K!(n-k)!





Binomial distribution (mean)

$$\mu = E[k] = \sum_{k \in N} k f(k) = \boxed{NP}$$

$$= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} =$$

$$= \sum_{k=1}^{n} \frac{(n-1)!\,n}{(k-1)!(n+1-k-1)!} p\, p^{(k-1)} (1-p)^{(n-k)} =$$

Define j=k-1

$$= np \sum_{j=0}^{n-1} \frac{(n-1)!}{(j)!(n-1-j)!} p^j (1-p)^{(n-1-j)} = np$$

Binomial distribution (variance)

$$Var\,(k) = E\left[(k-\mu)^2\right] = np(1-p)$$

Variance and Moments of a Random Variable

- **Example**
  - **Variance of a Binomial Random Varaible**
    - The variance of a binomial random variable *X* with parameters n and p can be determined directly by computing $E[X^2]$.

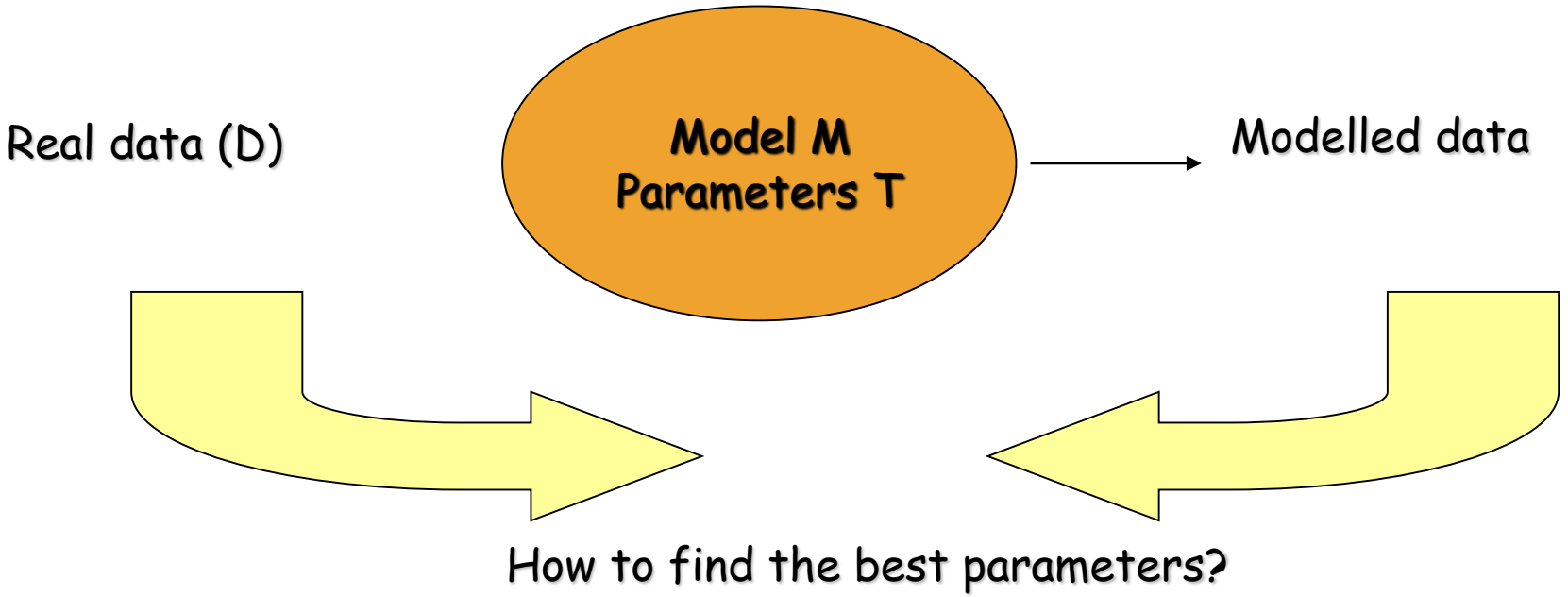
$$E[X^2] = \sum_{j}^{n} C_j^n p^j (1-p)^{n-j} j^2 = n(n-1)p^2 + np$$

$$\mathrm{Var}[X] = E[X^2] - (E[X])^2 = n(n-1)p^2 + np - (np)^2$$

$$= np(1-p)$$

Evaluation of parameters from empirical data

Generally speaking, a parametric model M aims to reproduce a set of known data



We serch to a P the maximize the similarity with our data

Maximum likelihood

$$T^* = \underset{T}{\operatorname{argmax}}\, P(D|T,M) =$$

$$T^* = \underset{T}{\operatorname{argmax}}\, \log(P(D|T,M))$$

D=data, M= model, T=model parameters

Evaluation of p

- Given n independent experiments, with k positive events, how to evaluate the best p?

- MAXIMUM LIKELIHOOD definition:
  - Best p → p that maximises the total probability of the experiment results

- Example
  - Which is the p that maximises the outcome of 10010001100111100010 ?

Evaluation of p

$$f(k \mid n, p) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

evaluating the best P to get our data

$$\left.\frac{df}{dp}\right|_{p=p_{ML}} = 0$$

maximum of likelihood because the derivate=0

$$\Rightarrow k{p_{ML}}^{k-1}(1-p_{ML})^{n-k} - {p_{ML}}^{k}(n-k)(1-p_{ML})^{(n-k-1)} = 0$$

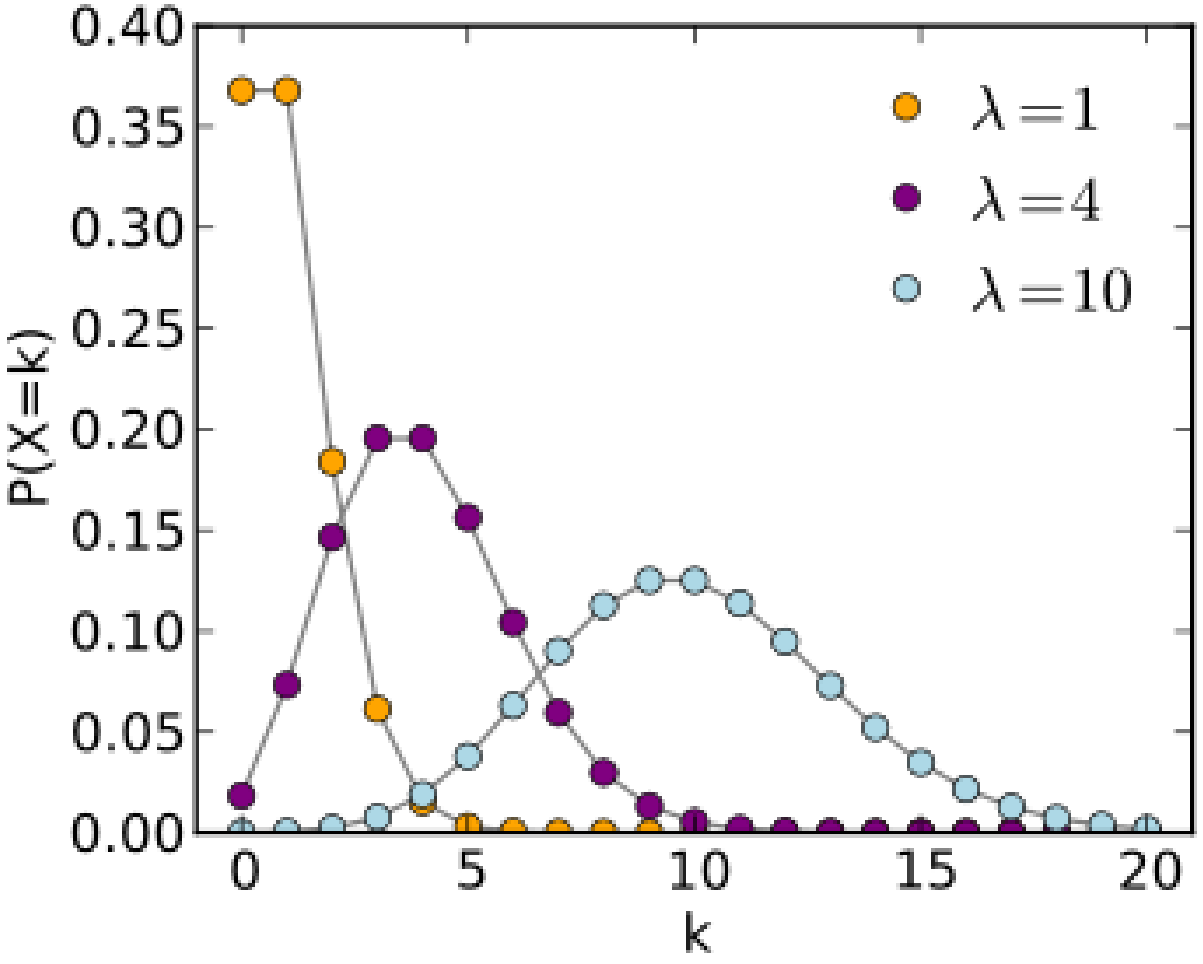
$$k(1-p_{ML}) - p_{ML}(n-k) = 0$$

$$p_{ML} = \frac{k}{n}$$

Poisson distribution

the **Poisson distribution** is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate (λ) and independently of the time since the last event. (The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.)

Poisson distribution



$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$E[k] = \lambda$$

$$Var[k] = \lambda$$

probability to have (Xaxis) event in a unit of time given the Lambda frequence of it

Continuous Distributions

- Data are described with distributions
  - Continuous variables: data are valued on a continuous interval: $x \in I \subseteq \Re$
    - Cumulative probability: $F(x) = N(x \leq X)/N_{tot}$
    - Prob($A < x \leq B$) = $F(B)-F(A) = \int_A^B f(x)dx$ (cumulative)
    - Probability density function (limit) $f(x) = \frac{dF(x)}{dx}$
    - Normalization: $\int_I f(x)dx = 1$

Mean and Variance

- Mean

$$\mu = E[x] = \int_{x \in I} x \cdot f(x) dx$$

- Variance

$$Var[x] = E\left[(x-\mu)^2\right] = \int_{x \in I} (x-\mu)^2 \cdot f(x) dx$$

The Uniform Probability Distribution

- Uniform Probability Density Function

$$f(x) = 1/(b - a) \text{ for } a \leq x \leq b$$
$$= 0 \text{ elsewhere}$$

where

a = smallest value the variable can assume

b = largest value the variable can assume

The probability of the continuous random variable assuming a specific value is 0.

$P(x=x_1) = 0$

Example: Buffet

Customers are charged for the amount of salad they take. Sampling suggests that the amount of salad taken is uniformly distributed between 5 ounces and 15 ounces.

- Probability Density Function

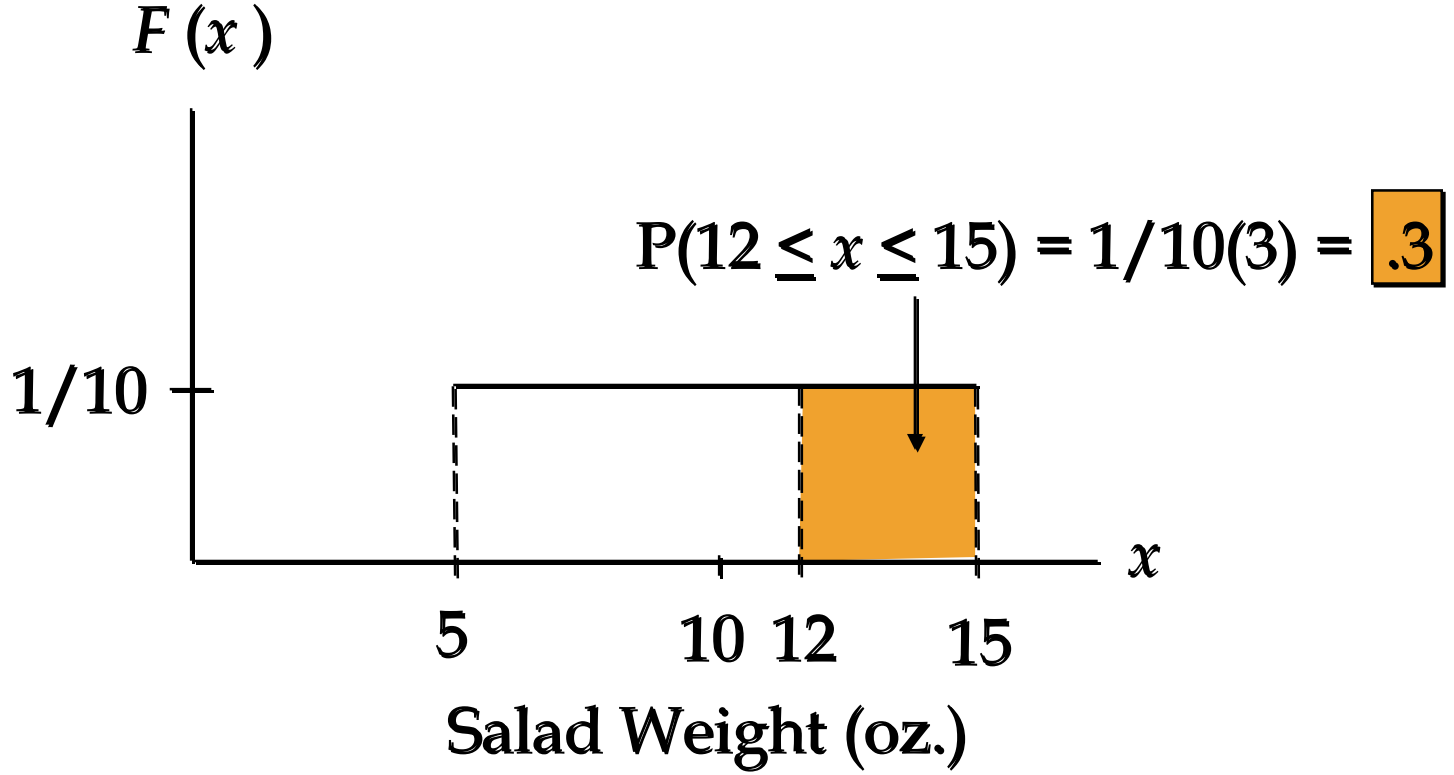
$$f(x) = 1/10 \text{ for } 5 \leq x \leq 15$$
$$= 0 \text{ elsewhere}$$

where

x = salad plate filling weight

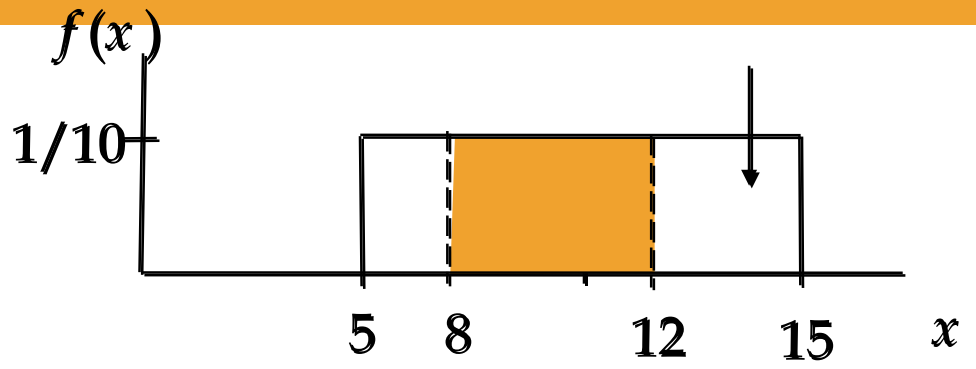
Example: Buffet

What is the probability that a customer will take between 12 and 15 ounces of salad?



The Uniform Probability Distribution

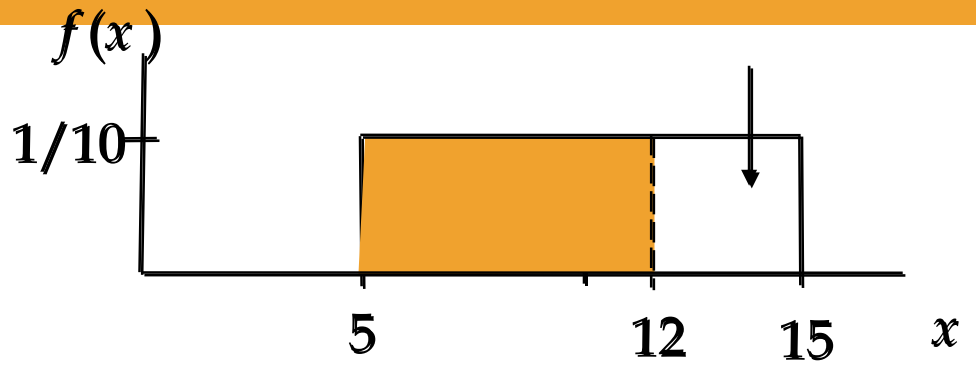
$P(8 \leq x \leq 12) = ?$



$P(8 \leq x \leq 12) = (1/10)(12-8) = .4$

The Uniform Probability Distribution

$P(0 \leq x \leq 12) = ?$



$P(0 \leq x \leq 12) = P(5 \leq x \leq 12) =$
$= (1/10)(12-5) = .7$

The Uniform Probability Distribution

- Uniform Probability Density Function

$$f(x) = 1/(b - a) \quad \text{for } a \leq x \leq b$$
$$= 0 \quad \text{elsewhere}$$

- Expected Value of x

$$E(x) = (a + b)/2$$

- Variance of x

$$\text{Var}(x) = (b - a)^2/12$$

where

a = smallest value the variable can assume

b = largest value the variable can assume

Prove it!!

Normal distribution

The normal distribution is considered the most "basic" continuous probability distribution. Specifically, by the central limit theorem, under certain conditions, the sum of a number of random variables with finite means and variances approaches a normal distribution as the number of variables increases.

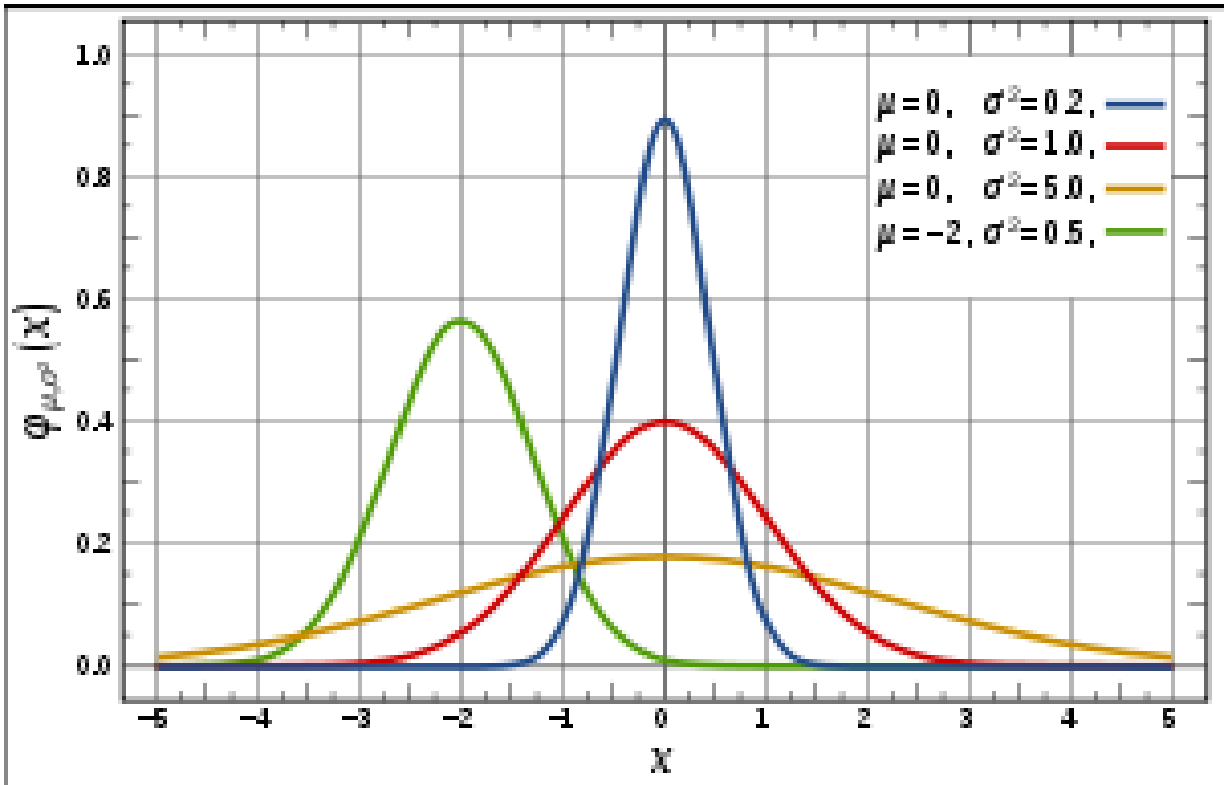
For this reason, the normal distribution is commonly encountered in practice, and is used throughout statistics, natural sciences, and social sciences as a simple model for complex phenomena. For example, the observational error in an experiment is usually assumed to follow a normal distribution, and the propagation of uncertainty is computed using this assumption.

Normal distribution

$$f(x) = \mathrm{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{1}{2}}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

μ : mean

σ^2 : variance



Normal distribution:continuous

$$f(x) = \mathrm{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{1}{2}}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

normalize

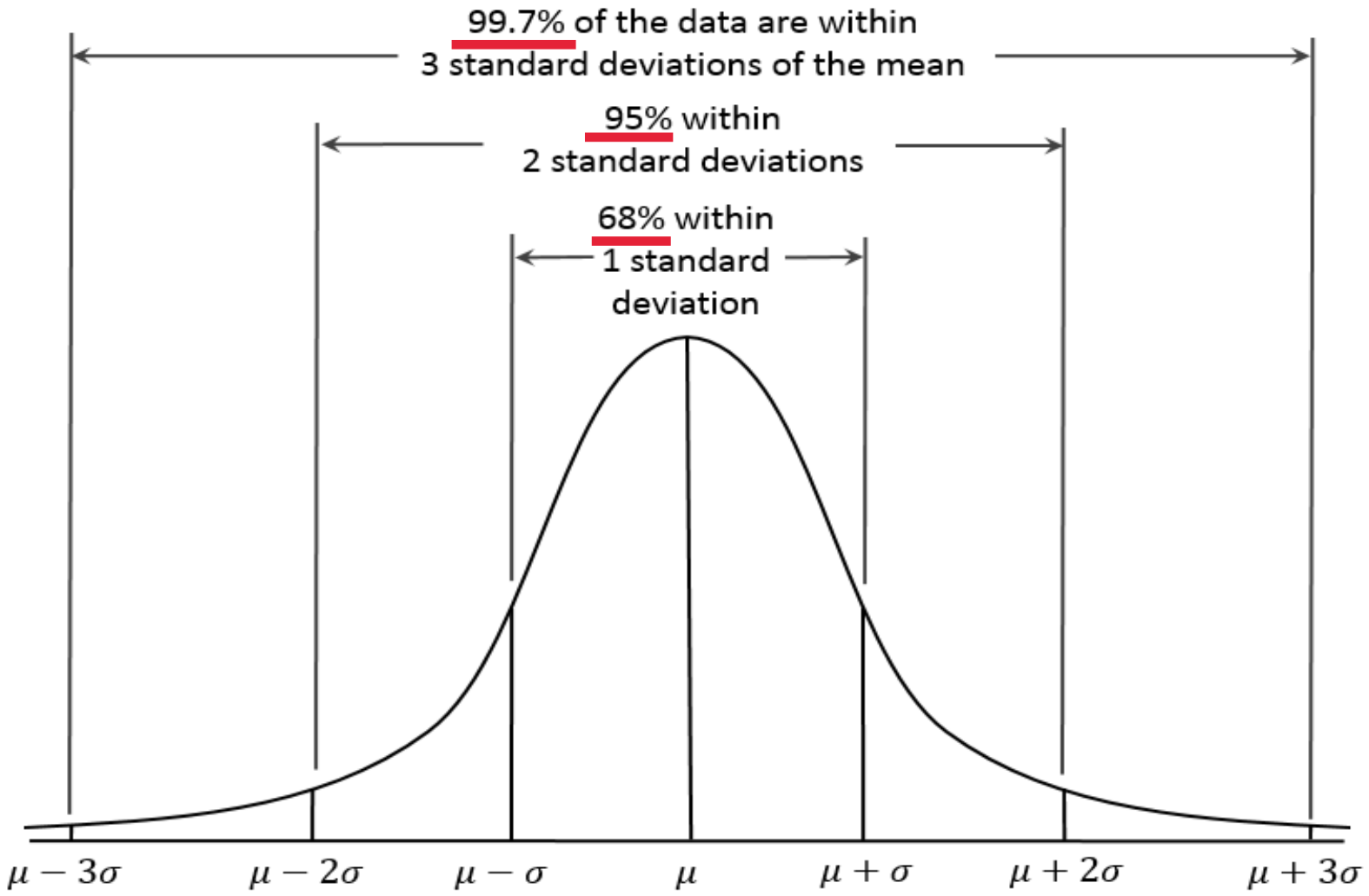
μ : mean

σ^2 : variance

$$E[x] = \int x\mathrm{N}\left(x \mid \mu, \sigma^2\right)dx = \mu$$

$$E[x^2] = \int x^2\mathrm{N}\left(x \mid \mu, \sigma^2\right)dx = \mu^2 + \sigma^2$$

Normal distribution



Normal distribution in standard form

Let x be a Normally distributed random variable with mean μ and variance σ^2

Introduce the variable **z = (x-μ)/σ**

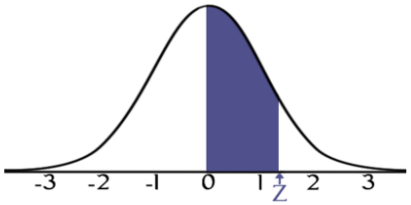
Z=(x-mean)
\sqr(variance)

$$E[z] = E\left[\frac{x-\mu}{\sigma}\right] = \frac{E[x]-\mu}{\sigma} = 0$$

$$E[z^2] = E\left[\frac{(x-\mu)^2}{\sigma^2}\right] = \frac{E[x^2]-\mu^2}{\sigma^2} = 1$$

$$Var[z] = E[z^2] - E[z]^2$$

z is a Normally distributed random variable with mean 0 and variance 1



STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and *z* standard deviations above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and z is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Heights of females are normally distributed with a mean of 63.6 inches and a standard deviation of 2.5 inches. What is the probability that a female selected at random will be taller than 70 inches ?

Z=(x-63.6)\2.5 X=70
Z=2.56
look on the table to see the area under a 2.56 Z
A=0.4948 it is the chance to be between 0 and 63.6 and so 0,5-0,4948 is the chance to be higer than 70

Standard variable:

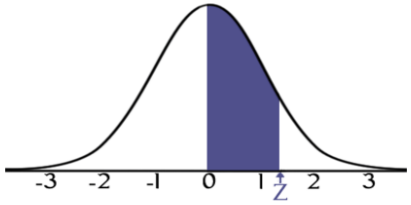
$z = (x-63.6)/2.5$

For $x = 70 \rightarrow z = 2.56$

Prob [0-2.56] = 0.4948

Prob [z>2.56] =

=0.5 -0.4948=0.0052



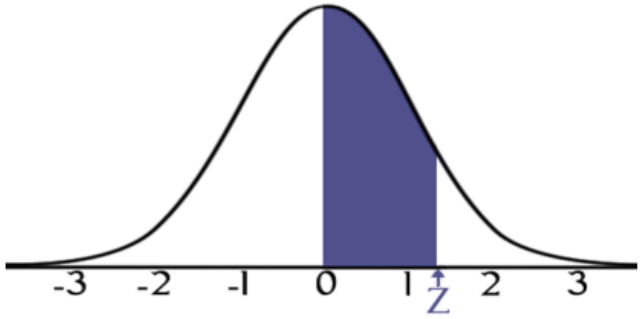
STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and *z* standard deviations above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and *z* is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

- The SAT scores are normally distributed with a mean of 500 and a standard deviation of 150 points. If an admissions office wanted to accept students who were in the top 40%, what would the cut-off scores be?

- Top 40%: given the standardized variable
  - The cutoff is the value of the standard table giving an area equal to 0.1 (between 0 and the cutoff)



STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and *z* standard deviations above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and z is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224

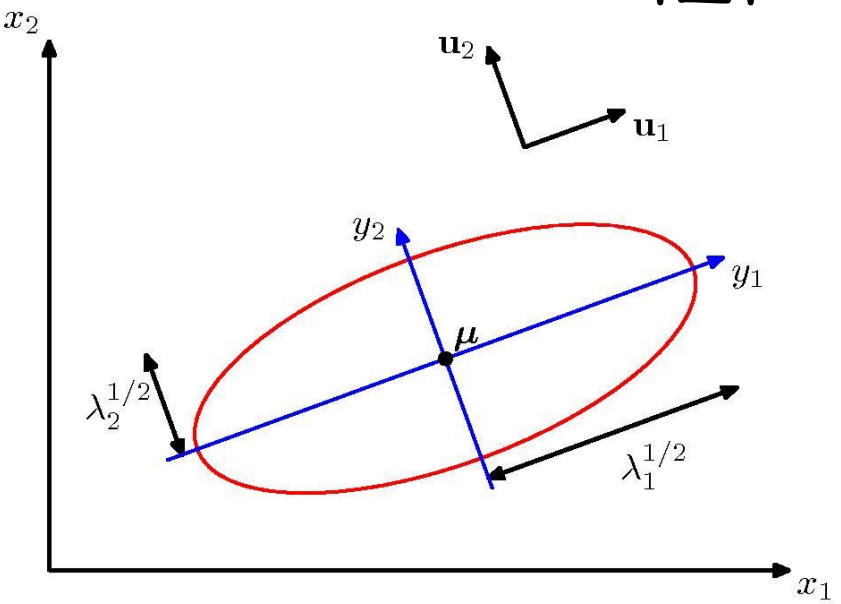
$0.26 = \text{cutoff}(z) = [\text{cutoff}(x) - \mu]/\sigma$

$\text{cutoff}(x) = 150 * 0.26 + 500 = 539$

Gaussian distribution in a D-dimensional space

$$\mathrm{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

where μ is a D-valued vector (means) and Σ is a DxD symmetric matrix (covariance matrix), with determinant |Σ|



Estimating Mean and Variance from sampled data

- Consider a set of n independent and identically distributed data, following a normal distribution with mean μ and variance σ^2

$x_1, x_2 .., x_n = X$

How can we estimate mean and variance?

Example:

X=0.2;0.18;0.24;0.14;0.05;0.26;0.15;0.29,0.49

Estimating mean and Variance from sampled data

MAXIMUM LIKELIHOOD:

Estimating the parameters as to maximize the probability for the sampled values

$$\left(\mu_{ML}, \sigma^2_{ML}\right) = \arg\max_{\mu,\sigma^2} f(x_1, x_2..x_n \mid \mu, \sigma^2)$$

ML Mean and Variance

- The probability for the sampled data is

$$f\left(X \mid \mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\left(2\pi\sigma^2\right)^{\frac{1}{2}}} \cdot e^{\frac{-\left(x_n-\mu\right)^2}{2\sigma^2}}$$

- It has to be maximised over the variables μ and σ^2

ML Mean and Variance

- Since logarithm is a monotonic function

$$\arg\max_{\mu,\sigma}\left[f(X|\mu,\sigma^2)\right] = \arg\max_{\mu,\sigma}\left[\ln f\,(X|\mu,\sigma^2)\right] =$$

$$= \arg\max_{\mu,\sigma}\left[-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

ML Mean and Variance

$$\begin{cases} \dfrac{\partial}{\partial \mu}\left[-\dfrac{n}{2}\ln\left(\sigma^2\right)-\dfrac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right]\Bigg|_{\mu_{ML},\sigma^2{}_{ML}} = 0 \\ \dfrac{\partial}{\partial \sigma^2}\left[-\dfrac{n}{2}\ln\left(\sigma^2\right)-\dfrac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right]_{\mu_{ML},\sigma^2{}_{ML}} = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n}\left(x_n-\mu_{ML}\right)=0 \\ -\dfrac{n}{2\sigma^2{}_{ML}}+\dfrac{1}{2\left(\sigma^2{}_{ML}\right)^2}\sum_{i=1}^{n}\left(x_i-\mu_{ML}\right)^2=0 \end{cases}$$

ML Mean and Variance

$$\begin{cases} \mu_{ML} = \dfrac{1}{n}\sum_{i=1}^{n} x_i \\ \sigma^2{}_{ML} = \dfrac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_{ML}\right)^2 \end{cases}$$

Are these the expressions that you usually apply?

The μ_{ML} is an unbiased estimation

- Consider a set of data generated starting from a normal distribution N(μ, σ²)the estimation

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

Is an unbiased estimate of μ since

$$E[\mu_{ML}] =$$

$$\int\int \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \cdot \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \cdot e^{\frac{-(x_1-\mu)^2}{2\sigma^2}} \cdot e^{\frac{-(x_2-\mu)^2}{2\sigma^2}} \cdot \ldots \cdot e^{\frac{-(x_n-\mu)^2}{2\sigma^2}} dx_1 dx_2 dx_n =$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \mu$$

The σ_{ML}^2 is a biased estimation

$\sigma_{ML}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - M)^2$ Is a biased estimate of σ^2 since

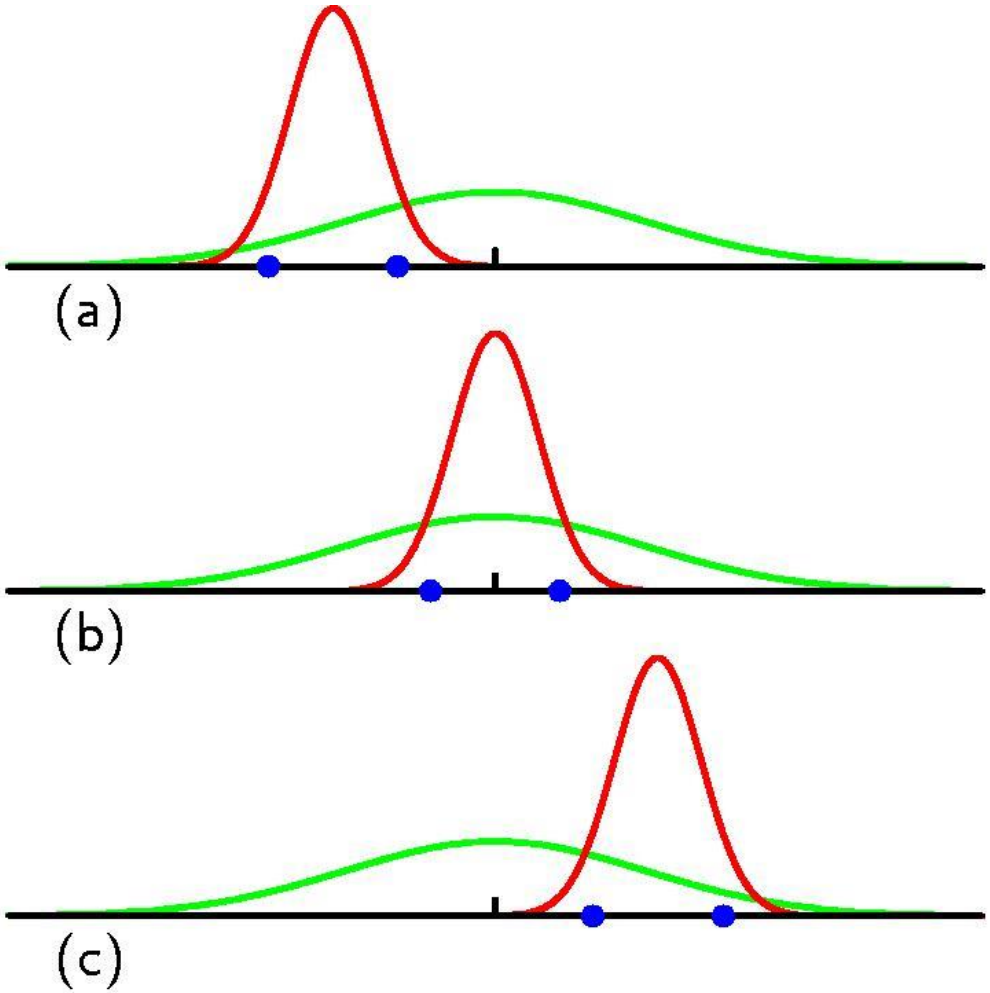
$$E\left[\sigma_{ML}^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2\right] =$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}x_i^{\,2} - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_i x_j + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}x_j x_k\right] =$$

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

$$= \mu^2 + \sigma^2 - \mu^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

σ_{ML}^2 is underestimating σ^2



Green: "Real" distribution
Red: Sampled distributions

σ_{ML}^2 is evaluated with respect to the sample mean

Unbiased estimation of μ and σ^2

$$M = \mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$S^2 = \frac{n}{n-1}\sigma^2{}_{ML} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - M\right)^2$$

Distribution of the sample mean

- Consider different sets of sampled data X^k (e.g. different levels of expression of genes in *n* different populations of individuals)

A sample mean for each set can be defined

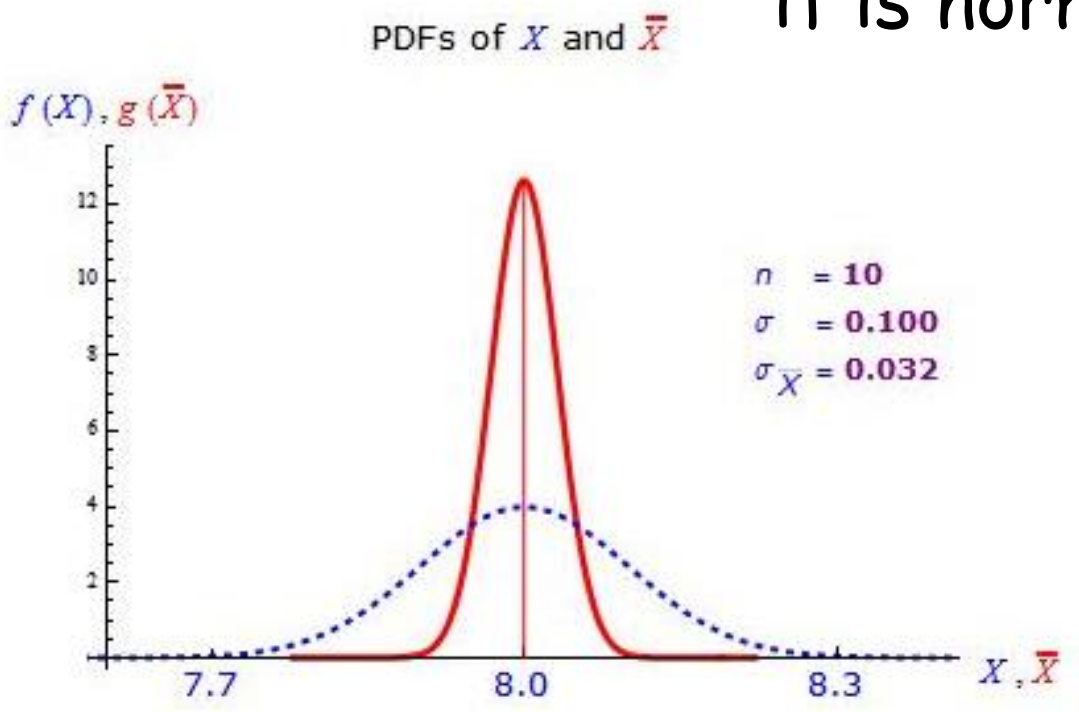
$$M^{(k)} = \frac{1}{n^{(k)}} \sum_{i=1}^{n^{(k)}} x_i^{(k)}$$

How are they distributed?

Distribution of the sample means

$$M^{(k)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(k)}$$

M is a linear combination of independent identically distributed normal variables: it is normally distributed



Distribution of the sample means

$$E[M] = \mu$$

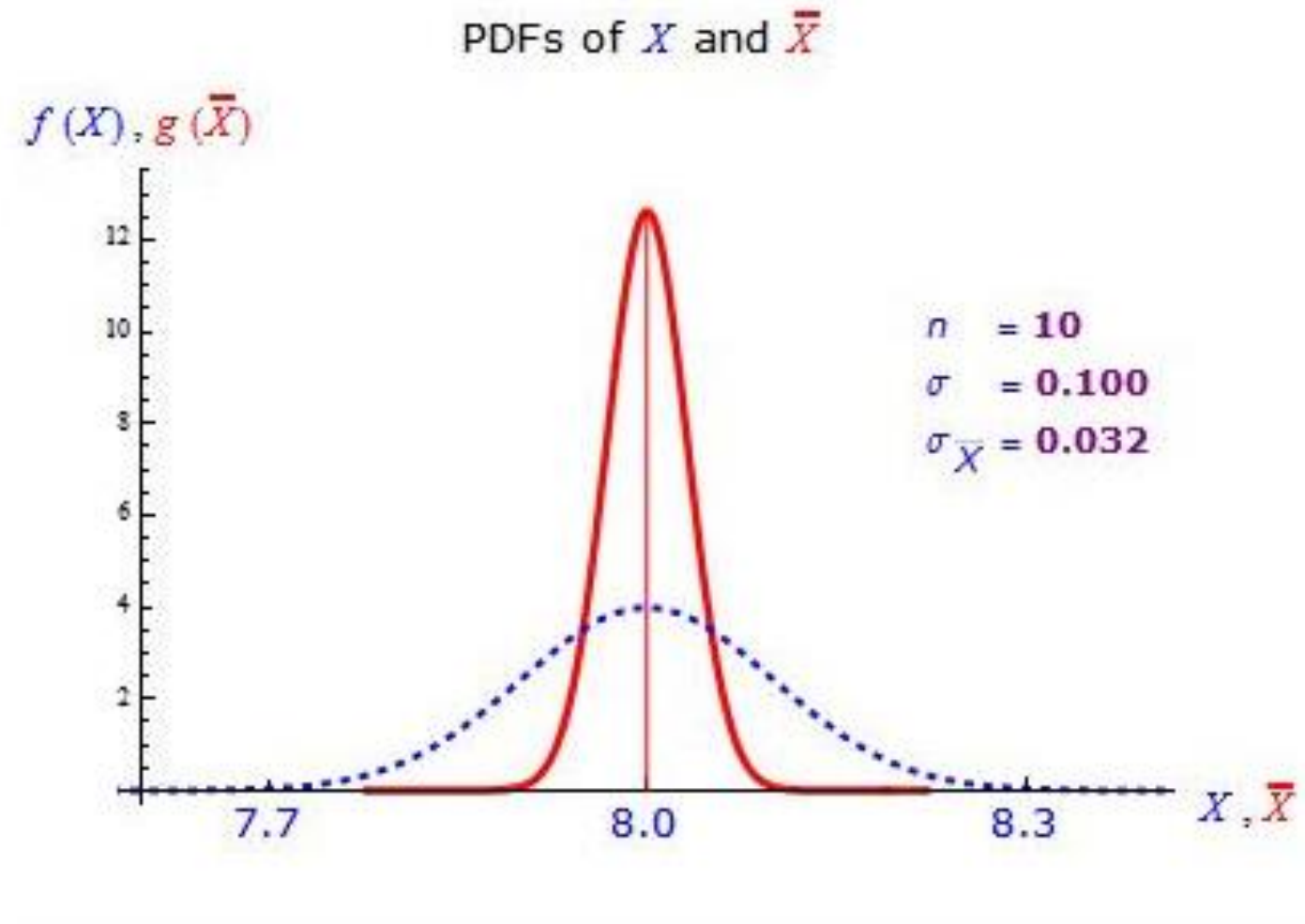
$$\mathrm{var}[M] = E[(M-\mu)^2] = E[M^2 + \mu^2 - 2\mu M] =$$

$$= E[M^2] + \mu^2 - 2\mu^2 = E[M^2] - \mu^2$$

$$E[M^2] = E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mu^2 + \delta_{ij}\sigma^2 =$$

$$= \mu^2 + \frac{1}{n}\sigma^2 \qquad \mathrm{var}[M] = \frac{1}{n}\sigma^2$$

Distribution of the sample means



Distribution of the sample deviations

- If X_i are n independent, normally distributed random variables with mean μ and variance σ^2, then the random variable

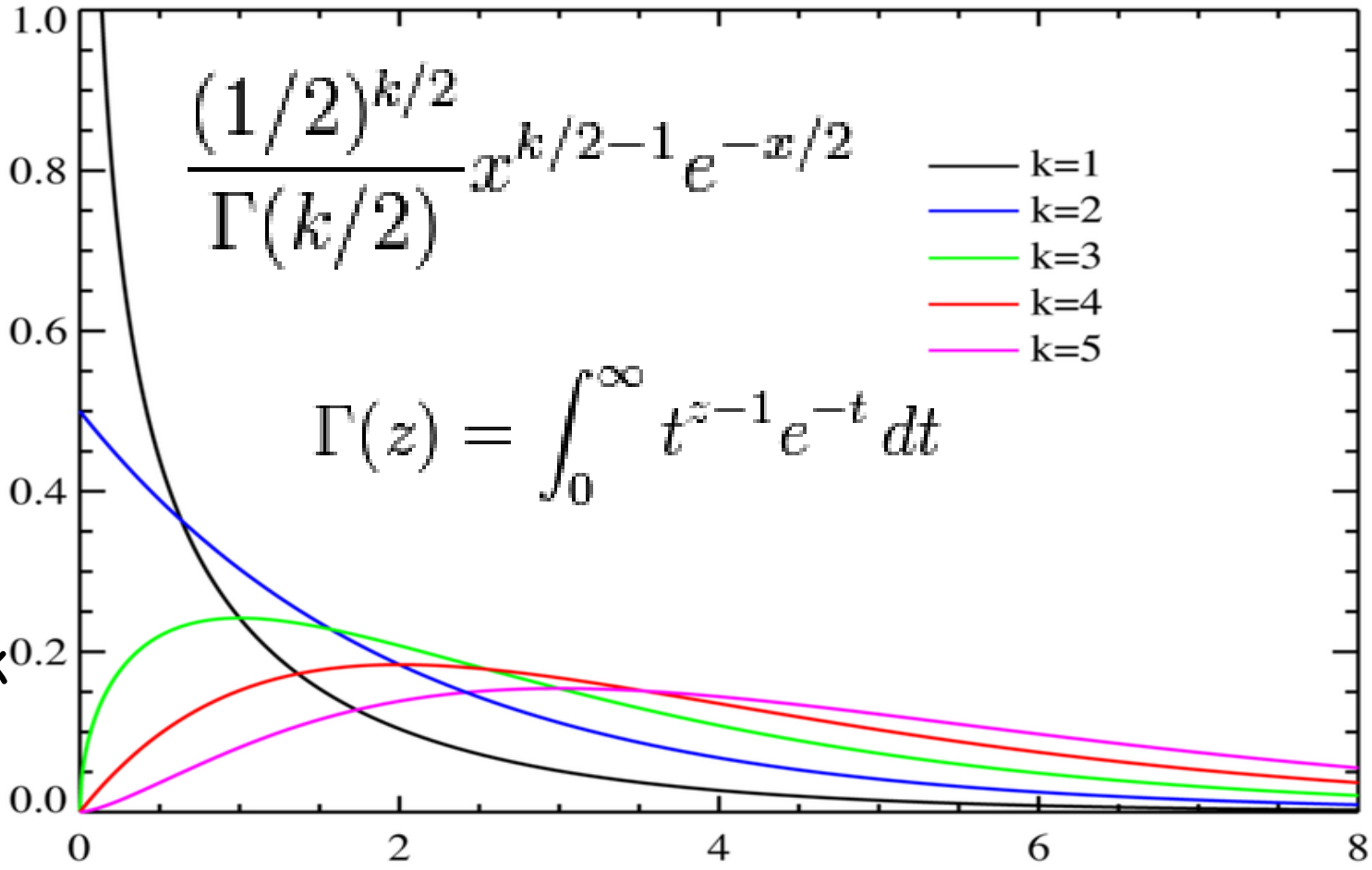
$$Q = \sum_{i=1}^{N} \left(X_i - \mu \right)^2$$

is distributed according to the chi-square distribution with *n-1* degrees of freedom. This is usually written as:

$$Q \approx \sigma^2 \chi^2_{n-1}$$

The chi-square distribution has one parameter: *(n-1)* that is a positive integer that specifies the number of degrees of freedom (i.e. the number of independent X_i -μ)

Chi-square distribution



Mean = k

Variance = 2k

Distribution of the sample deviations

- So, being the unbiased sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - M\right)^2$$

is distributed according to the chi-square distribution with *k* degrees of freedom.

$$S^2 \approx \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

It follows that

$$E\left[S^2\right] \approx \frac{\sigma^2}{n-1}E\left[\chi^2_{n-1}\right] = \frac{\sigma^2}{n-1}\left(n-1\right) = \sigma^2$$

Unbiased

Distribution of the sample deviations: Proof for 1 degree of freedom

Let random variable Y be defined as $Y = X^2$ where X has normal distribution with mean 0 and variance 1

We can compute the cumulative function

$$y < 0,\ P(Y < y) = 0$$

$$y \geq 0,\ P(Y < y) = P(X^2 < y) =$$

$$= P(|X| < \sqrt{y}) = F_x(\sqrt{y}) - F_x(-\sqrt{y})$$

Distribution of the sample deviations: Proof for 1 degree of freedom

y>0

$$f_y(y) = f_x(\sqrt{y})\frac{\partial(\sqrt{y})}{\partial y} - f_x(-\sqrt{y})\frac{\partial(-\sqrt{y})}{\partial y}$$

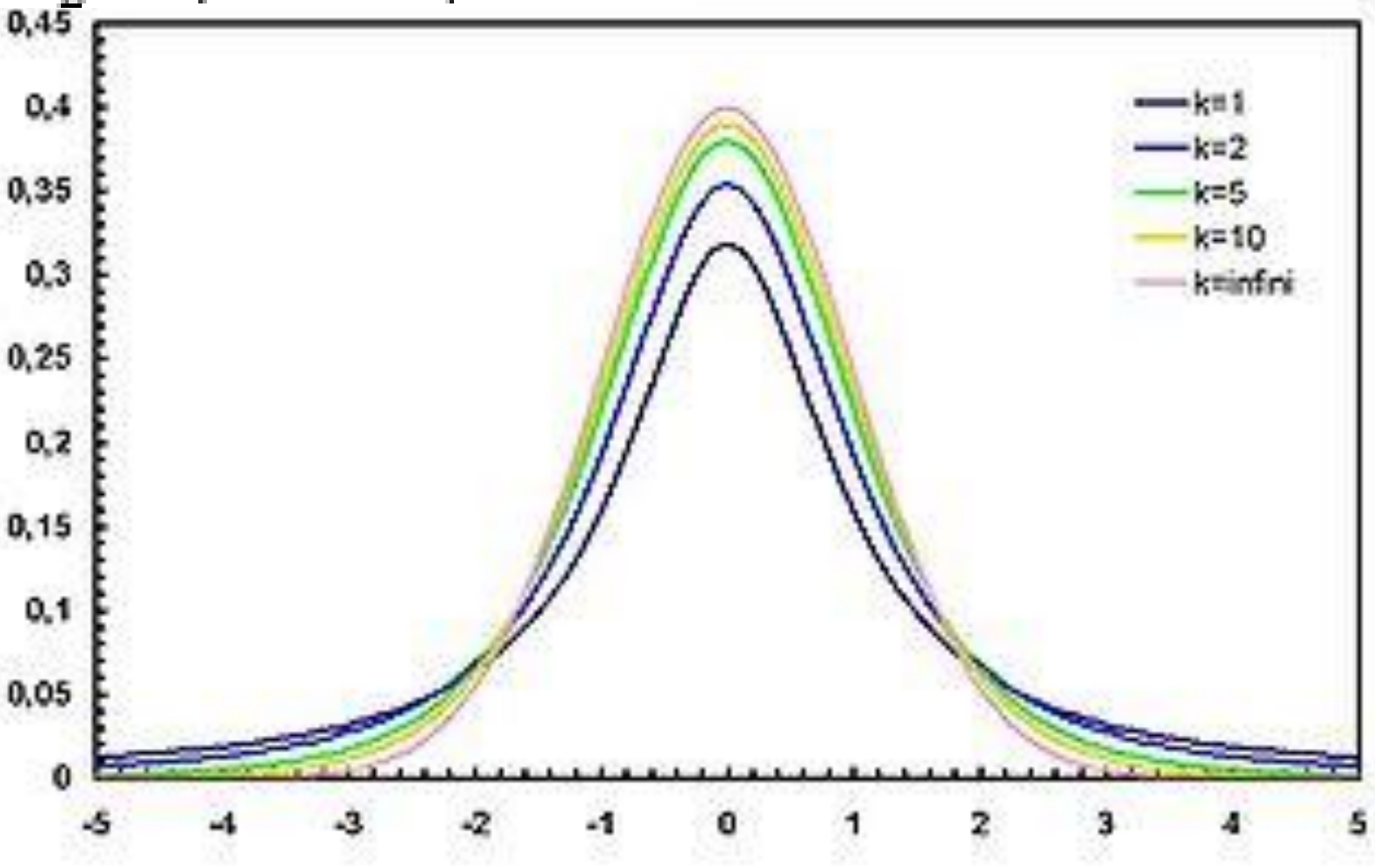
$$= \frac{1}{\sqrt{2\pi}}e^{\frac{-y}{2}}\frac{1}{2y^{1/2}} + \frac{1}{\sqrt{2\pi}}e^{\frac{-y}{2}}\frac{1}{2y^{1/2}}$$

$$= \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})}y^{\frac{1}{2}-1}e^{\frac{-y}{2}} \qquad Y = X^2 \sim \chi_1^2$$

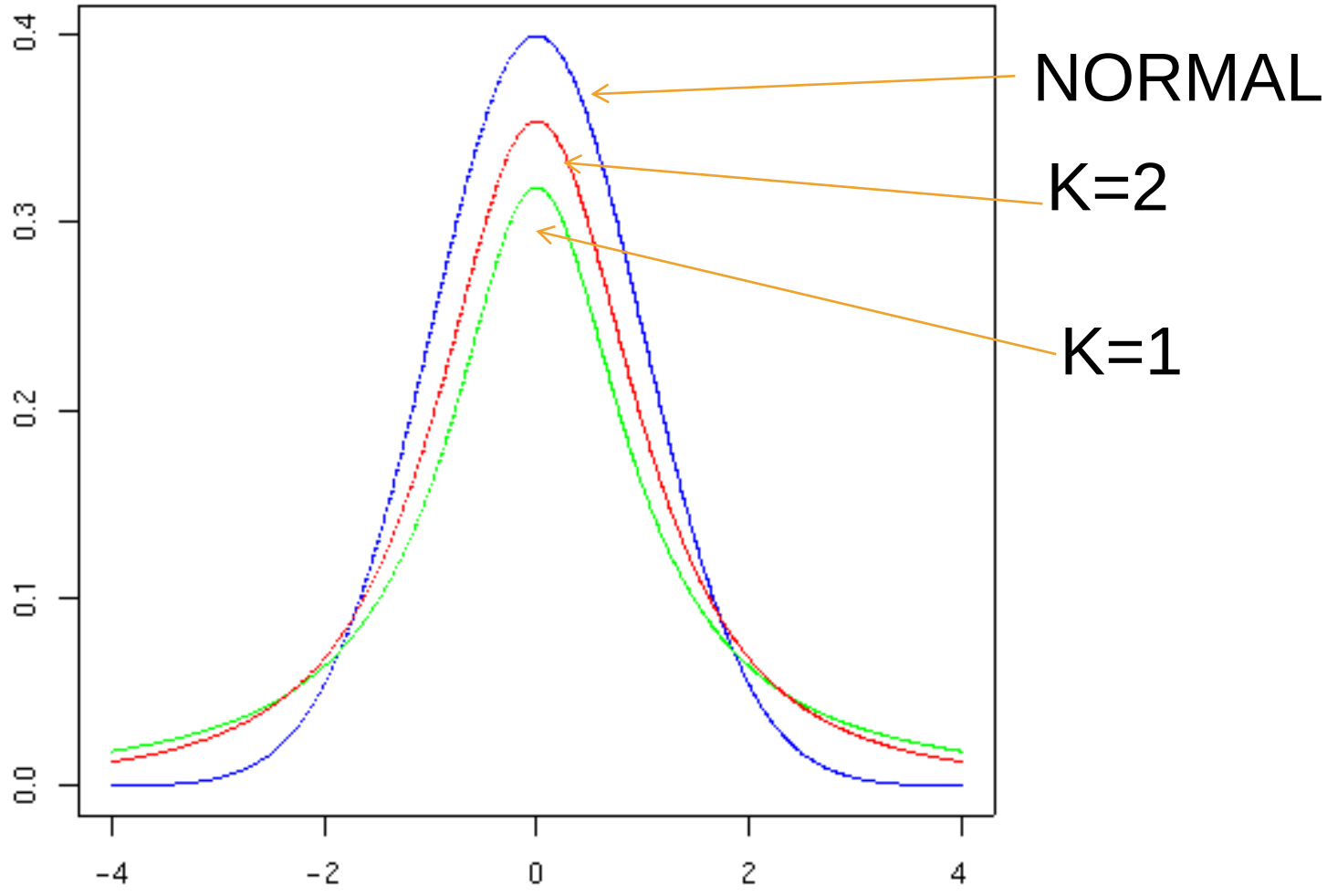
Student's t-distribution

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

v : degrees of freedom



Normal Vs t- distributions

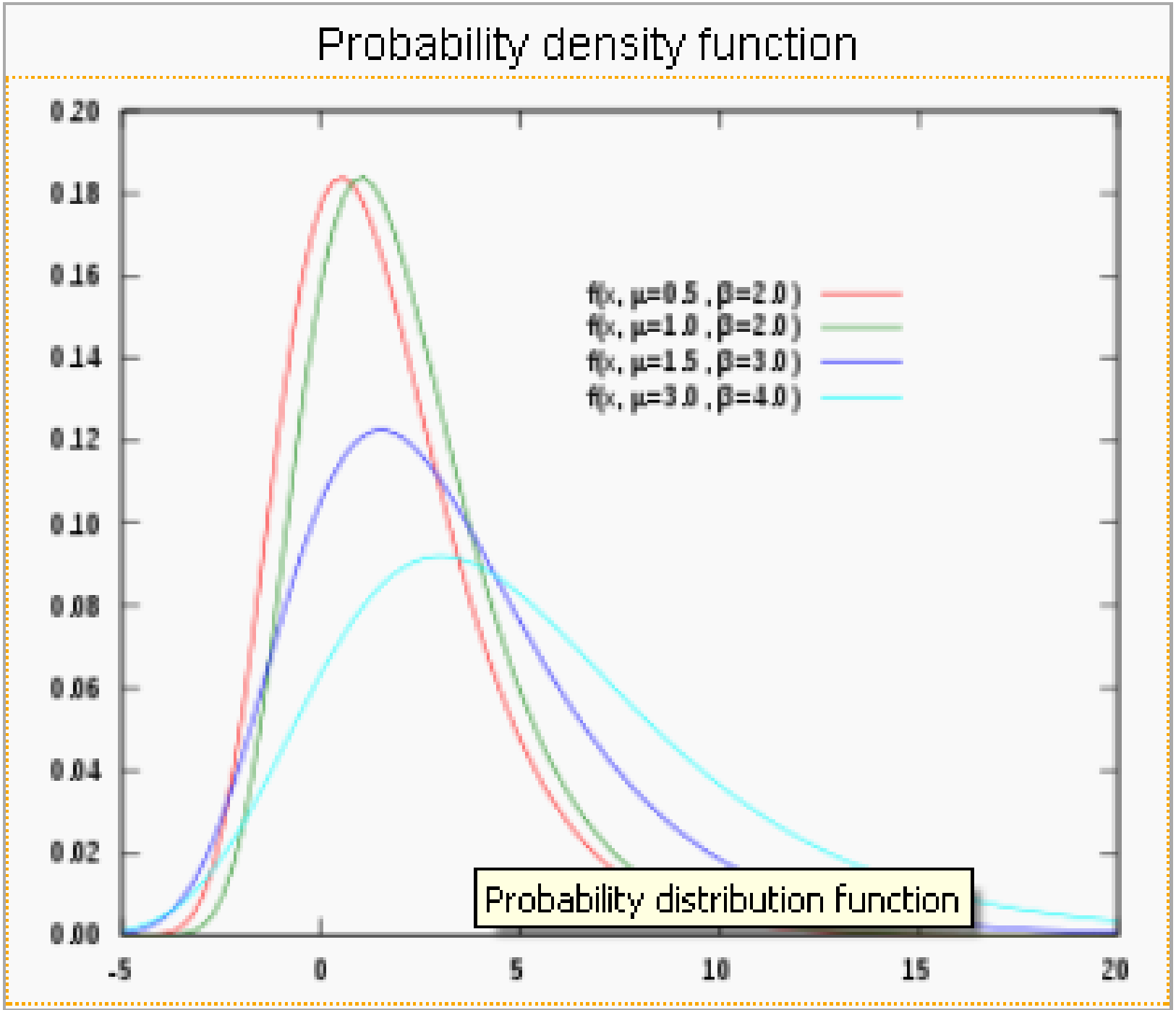


Extreme value distribution

- the **Gumbel distribution** is used to model the distribution of the maximum (or the minimum) of a number of samples of various distributions.
- For example we would use it to represent the distribution of the maximum level of a river in a particular year if we had the list of maximum values for the past ten years. It is useful in predicting the chance that an extreme earthquake, flood or other natural disaster will occur.

Extreme value distribution

support:	$x \in (-\infty; +\infty)$
pdf:	$\frac{z\, e^{-z}}{\beta}$ where $z = e^{-\frac{x-\mu}{\beta}}$
cdf:	$\exp(-e^{-(x-\mu)/\beta})$
mean:	$\mu + \beta\,\gamma$
median:	$\mu - \beta \ln(\ln(2))$
mode:	μ
variance:	$\frac{\pi^2}{6}\beta^2$


Probability density function
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
-5
0
5
10
15
20
f(x, μ=0.5 , β=2.0)
f(x, μ=1.0 , β=2.0)
f(x, μ=1.5 , β=3.0)
f(x, μ=3.0 , β=4.0)
Probability distribution function

BLAST

- In accordance with the Gumbel EVD, the probability p of observing a score S equal to or greater than x is given by the equation

$$p(S \geq x) = 1 - \exp\left(-e^{-\lambda(x-\mu)}\right)$$

$$\mu = {[\log(Km'n')]}/{\lambda}$$

BLAST: E-value

- The expect score E of a database match is the number of times that an unrelated database sequence would obtain a score S higher than x by chance. The expectation E obtained in a search for a database of D sequences is given by

$$E \approx pD$$