# DA503 Applied Statistics

## Lecture 03

## A Primer on Probability

# Agenda

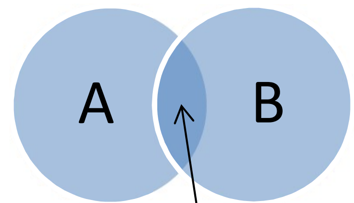
- Basic concepts of probability
- Conditional probability and independence
  - Bayes' theorem
- Random variables (discrete vs continuous)
- Probability distributions
  - Binomial and Poisson distributions
  - Normal (and Standard Normal) distributions
  - Exponential and Gamma distributions
- Joint probability distributions

# Basic rules of probability

Event: A possible outcome of an observation

P(E) is probability of an event E => E is impossible : P(E)=0

E is certain : P(E)=1




For any event E : $0 \le P(E) \le 1$

All events that are not E: $E^c$ => $P(E) + P(E^c) = 1$

Union of two events A and B: $A \cup B$

Probability of all events in A or B or both: $P(A \cup B)$

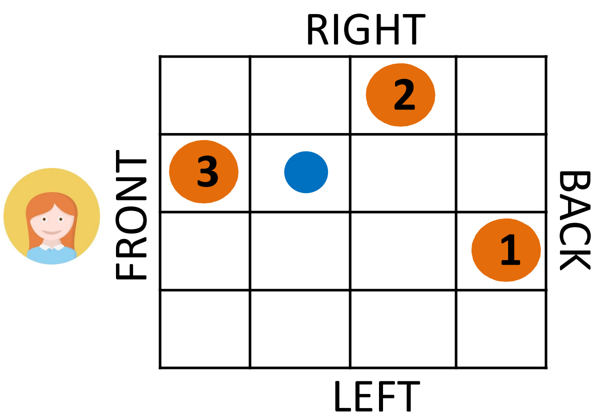
Probability of all events common to both A & B: $P(A \cap B)$

Probability of A or B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Zero if A and B have no elements in common (A and B are mutually exclusive)

# Intuition behind the Bayes' Theorem

- we can never be fully certain of the world, as it is constantly changing. Fundamental principle behind this theorem, is: update and improve our knowledge of reality (prior) as we get more and more data or evidence.
- Suppose we have an invisible blue ball in the field below



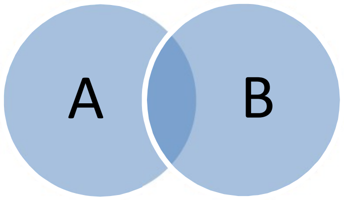

Someone is placing a **visible orange ball** and tells you where the **invisible blue ball** is located with respect to the orange one, and you update your knowledge for the probability of the location of the **blue ball**:

Prior gets updated as new evidence is gathered

0. => Prior : $P(x) = 1/16$
1. "Right" => Posterior: $P(x|1) = 1/8$
2. "Behind" => Posterior: $P(x|1,2) = 1/4$
3. "Front" => Posterior: $P(x|1,2,3) = 1/2$

# Conditional Probability

- Probability of A given that B has occurred:




$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \boxed{1}$$ ( given that P(B) ≠ 0 )

Similarly: $$P(B \mid A) = \frac{P(B \cap A)}{P(A)} \quad \boxed{2}$$ ( given that P(A) ≠ 0 )

Conditional probability (from 1 & 2): $$\boxed{P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}}$$

- <u>Bayes' theorem</u>:

Probability of seeing the evidence/data if the hypothesis is true

(Prior) probability a hypothesis is true (before any evidence/data)

$$P(hypothesis \mid data) = \frac{P(data \mid hypothesis)P(hypothesis)}{P(data)}$$

Posterior probability of hypothesis being true given the evidence/data

Probability of seeing the evidence/data

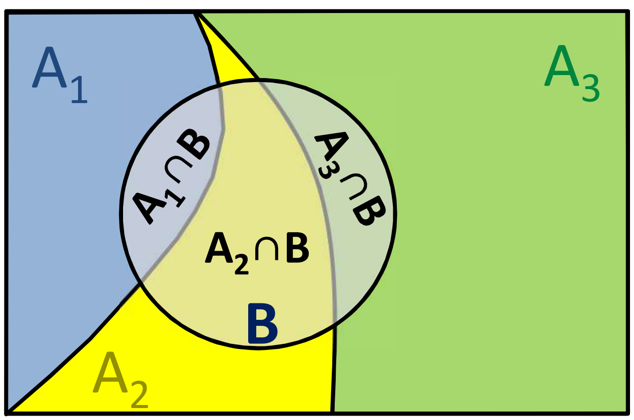
# Conditional Probability – cont'd

- Expanded form:

$$P(B) = P[(B \cap A) \cup (B \cap A^c)] = P(B \cap A) + P(B \cap A^c)$$
$$= P(B / A)P(A) + P(B / A^c)P(A^c)$$

$$\boxed{P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{\underbrace{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}_{\sum_j P(B \mid A_j)P(A_j)}}}$$

- **Total probability theorem:**

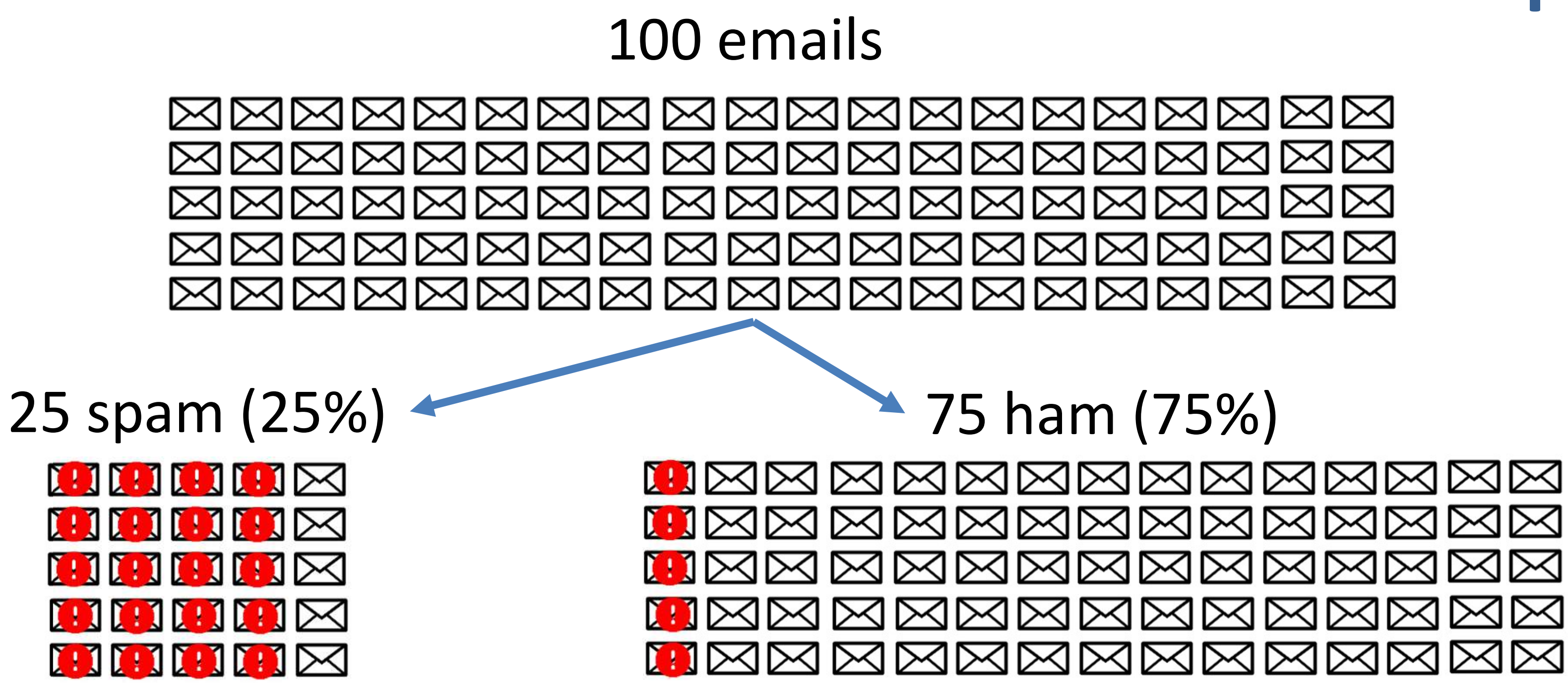



$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

$$P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)$$

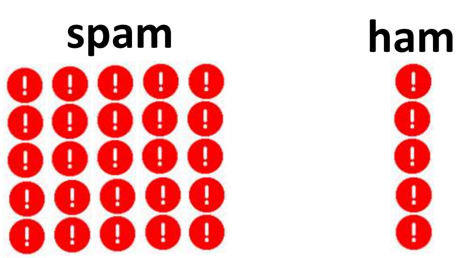
$$\boxed{P(A \mid B) = \frac{P(B \mid A)\,P(A)}{\sum_i P(B \mid A_i)\,P(A_i)}}$$

# Example I




❗ : Messages containing the word "**Buy**"

- Suppose 20 of the messages labeled "**spam**" contains the word "**Buy**"
- And 5 of the messages labeled "**ham**" contains the word "**Buy**"




If a mail contains the word "**Buy**", what is the probability that it is **spam**?
20 spam, 5 ham => p = 20/(20+5) = **80%**

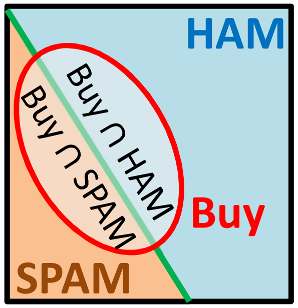
Ref: Naive-Bayes classifier by Luis Serrano

# Example I – cont'd

- Let's solve this by using the Bayes' theorem:

$$P(spam \mid Buy) = ?$$

$$P(spam \mid Buy) = \frac{P(Buy \mid spam)P(spam)}{P(Buy)}$$




$$P(Buy) = P(Buy \cap spam) + P(Buy \cap ham)$$

$$P(Buy) = P(Buy \mid spam)P(spam) + P(Buy \mid ham)P(ham)$$

$$P(Buy \mid spam) = \frac{20}{25} \qquad P(spam) = 0.75$$

$$P(Buy \mid ham) = \frac{5}{75} \qquad P(ham) = 0.25$$

$$P(spam \mid Buy) = \frac{\frac{20}{25} * 0.25}{\frac{20}{25} * 0.25 + \frac{5}{75} * 0.75} = \frac{1/5}{1/5 + 5/100} = \frac{0.2}{0.25} = 0.8$$

# Example II

- You have been diagnosed with **a very rare disease**, which only **affects 0.1% of the population**; that is, 1 in every 1000 persons
- **The test** you have taken to check for the disease **correctly classifies 99% of the people who have the disease**, **misclassifies healthy individuals with a 1% chance**
- If you test positive, what is the probability that you really have the disease?

The probability of the event, given that the hypothesis is true: the probability of being diagnosed positive in the test, given that we have the disease

Prior probability of having the disease before any test has been taken

$$P(D \mid +) = \frac{P(+ \mid D)\,P(D)}{P(+)} = \frac{P(+ \mid D)\,P(D)}{P(+ \mid D)\,P(D) + P(+ \mid ND)\,P(ND)}$$

The probability of the event: the probability of being diagnosed positive for the disease

# Example II – cont'd

- Let's compute the conditional probability:

$$P(D\,|\,+) = \frac{P(+\,|\,D)\,P(D)}{P(+)} = \frac{0.99 \text{ x } 0.001}{0.99 \text{ x } 0.001 + 0.01 \text{ x } 0.999} = 9\%$$

- Think it's small? Well, it was a rare disease, remember?
- But what if we get anxious and decide to take another test. Assuming it turned out to be positive again, what is the probability of having the disease this time?
- we can use exactly the same formula as before, but replacing the initial prior probability (0.1%) with the posterior probability obtained the previous time (9%):

$$P(D\,|\,+) = \frac{P(+\,|\,D)\,P(D)}{P(+)} = \frac{0.99 \text{ x } 0.09}{0.99 \text{ x } 0.09 + 0.01 \text{ x } 0.91} = 91\%$$

- Now we have a much higher chance, 91% of actually having the disease given the new prior.

# A classification example: Naive-Bayes

| Class | Long | Sweet | Yellow | Total |
|---|---|---|---|---|
| **Banana** | **400** | **350** | **450** | 500 |
| **Orange** | **0** | **150** | **300** | 300 |
| **Other** | **100** | **150** | **50** | 200 |
| Total | 500 | 650 | 800 | **1000** |

We have a training set of 1000 fruits

Based on the features of “long”, “sweet” and “yellow”, the fruit can belong to one of the following classes: “Banana, “Orange” or “Other”

- New data: {**long, sweet, yellow**} => What is the likelihood of this fruit being any one of the classes?
- **Step 1**: Let’s calculate the probability that the fruit is a banana

$$P(Banana \mid Long, Sweet, Yellow) = P(B \mid LSY) = \frac{P(LSY \mid B)P(B)}{P(LSY)} = ?$$

- Assume the features are conditionally independent (given B):

$$P(LSY \mid B) = P(L \mid B)P(S \mid B)P(Y \mid B)$$

( Naive assumption of NB )

- **Step 2**: Let’s work out the equations above and plug them in

# A classification example – cont'd

| Class | Long | Sweet | Yellow | Total |
|---|---|---|---|---|
| **Banana** | **400** | **350** | **450** | 500 |
| **Orange** | **0** | **150** | **300** | 300 |
| **Other** | **100** | **150** | **50** | 200 |
| Total | 500 | 650 | 800 | 1000 |

$$P(Long \mid Banana) = 400/500 = 0.8$$
$$P(Sweet \mid Banana) = 350/500 = 0.7$$
$$P(Yellow \mid Banana) = 450/500 = 0.9$$
$$P(Banana) = 500/1000 = 0.5$$

$$P(Banana \mid Long, Sweet, Yellow) = P(B \mid LSY) = \frac{P(LSY \mid B)P(B)}{P(LSY)} = ?$$

$$\boxed{P(LSY \mid B) = 0.8 * 0.7 * 0.9 = 0.504} \qquad \boxed{P(B) = 0.5}$$

$$P(Banana \mid LSY) = \frac{0.504 * 0.5}{P(LSY)} = \frac{0.252}{P(LSY)} = \frac{0.252}{0.27075} = 93\%$$

$$P(LSY) = P(LSY \mid B)P(B) + P(LSY \mid Orange)P(Orange) + P(LSY \mid Other)P(Other)$$

- Carrying out the same computation for the other two classes:

$$\left.\begin{aligned} &P(Orange \mid LSY) = 0 \\ &P(Other \mid LSY) = \frac{0.019}{P(LSY)} = 7\% \end{aligned}\right\}$$

Given the current attributes, Naive-Bayes classifies the fruit as Banana

# Probability distributions

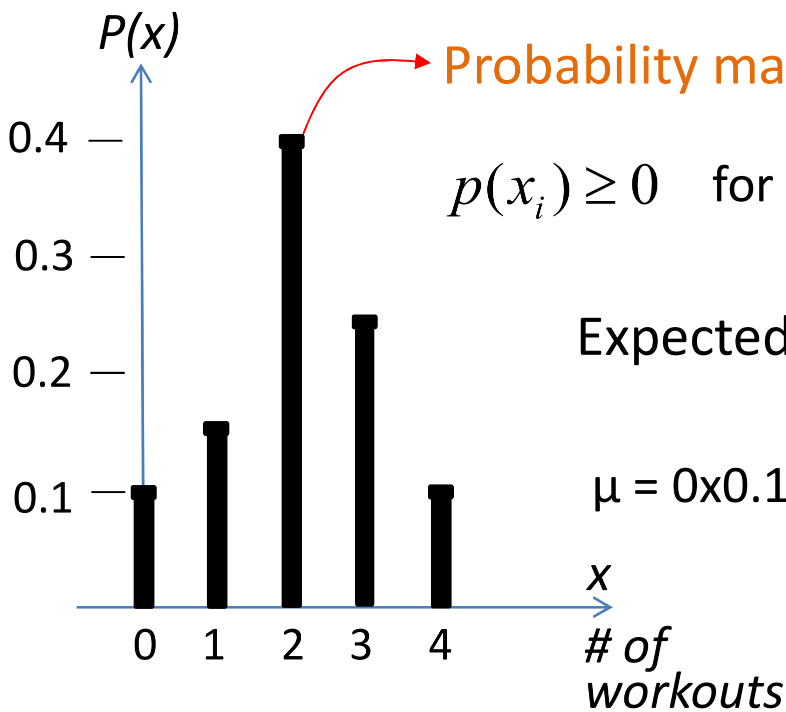
- **Random variable**: A variable whose possible values are numerical outcomes of a random phenomenon (with probabilities specified by its probability distribution). Could be discrete or continuous.
- **Discrete random variable**: RV that can take only a countable (finite) number of distinct values.
- <u>Example</u>: Number of rainy days in Istanbul in May
- **Continuous random variable**: A random variable that can take an infinite number of values between any two given values
- <u>Example</u>: Amount of rainfall in Istanbul in May

# Discrete probability distributions

**Discrete probability function for discrete random variables**: Table/graph that shows all possible values of a discrete RV and the corresponding probabilities.

**Example**: Number of workouts in a week

| X: **# of workouts** | **0** | **1** | **2** | **3** | **4** | (integer values) |
|---|---|---|---|---|---|---|
| P(X) | 0.1 | 0.15 | 0.4 | 0.25 | 0.1 | |




Probability mass function (PMF) : *p(x_i)* such that

$$p(x_i) \geq 0 \quad \text{for all } x_i \text{ and } \sum_{i=1}^{N} p(x_i) = 1$$

Expected value (mean) : $\boxed{E(x) = \mu = \sum_{i=1}^{N} x_i p(x_i)}$

μ = 0x0.1 + 1x0.15 + 2x0.4 + 3x0.25 + 4x0.1 = 2.1

Variance : $\boxed{\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 p(x_i)}$

# Binomial distribution

- It's about **k successes out of n trials**
- A **collection of Bernoulli trials**. A Bernoulli distribution is one of the simplest discrete distributions with 2 outcomes (success and failure): $P(X=\mathrm{x}) = p^{\mathrm{x}}(1-p)^{1-\mathrm{x}}$

  where p is the probability of success.
- If a success occurs, X=1, then: $P(X=1) = p^1(1-p)^0 = p$
- If a failure occurs, X=0, then: $P(X=0) = p^0(1-p)^1 = 1-p$
- Mean and variance of Bernoulli distribution:
- Mean: $E(X) = \Sigma x P(x) = 0.p^0(1-p)^1 + 1.p^1(1-p)^0 = p$
- Variance: $Var(X) = E[(x-\mu)^2] = E(x^2) - [E(x)]^2$

  $= \Sigma x^2 p(x) - p^2 = p - p^2 = p(1-p)$
- Bernoulli distribution is the main building block of many other discrete distributions (Binomial, Geometric, etc.)

# Binomial distribution

- Pre-requisites for Binomial distribution:
  - There are 2 potential outcomes per trial: success/failure
  - The probability of "success" is the same across all trials
  - The number of trials is fixed
  - Each trial is independent
  - We're interested in number of successes
- The number of successes in n independent Bernoulli trials has a Binomial distribution
- **Example**: Let's consider the following problem:
  - We know that only 8% of the population of men is affected by a certain disease
  - If you choose a random sample of 10 men:

# Binomial distribution – cont'd

- **Example** (cont'd)
  - a. What is the probability that all 10 have the disease?

    p = 0.08 (probability of success – disease a success? Weird!)
    n = 10
    $P(x=10) = p^{10} = 0.08^{10} = 1.07\text{x}10^{-11}$

  - b. What is the probability that no men have the disease?

    $P(x=0) = (1-p)^{10} = 0.92^{10} = 0.434$

  - c. What is the probability that 2 men have the disease?

    $p^2 \text{ x } (1-p)^8 = 0.08^2 \text{ x } (0.92)^8 = 0.0033$ ?

  – Keep in mind that there is more than one way of selecting 2 men out of 10:

 …

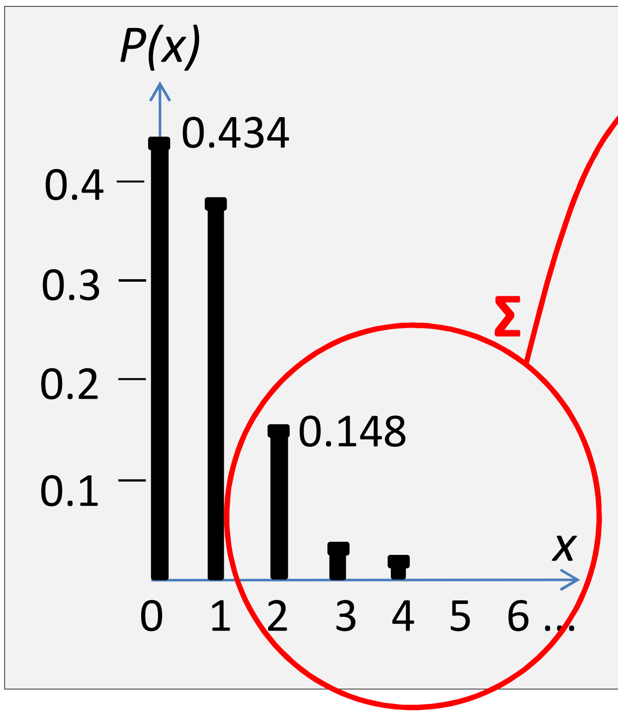
  – We have exactly $C(10,2) = \frac{10!}{2!\,(10-2)!} = \frac{10!}{2!\,8!} = 45$ combinations

# Binomial distribution – cont'd

- where $C(n,k)=\binom{n}{k}=\dfrac{n!}{k!(n-k)!}$ is the combinatorics for possible orderings of a finite number of objects.

P(x=2) = C(10,2) x $p^2$ x $(1\text{-}p)^8$ = 45 x 0.0033 = 0.148

**d**. What is the probability that at least 2 men have the disease?




P(x≥2) = P(x=2) + P(x=3) + … + P(x=9) + P(x=10)

$$P(x \geq 2)=\sum_{k=2}^{10}\binom{10}{k}p^{k}\,(1-p)^{10-k}$$

P(x≥2) = 1 – [ P(x=0) + P(x=1) ]
= 1 – [ 0.378 + 0.434 ]
= 0.188

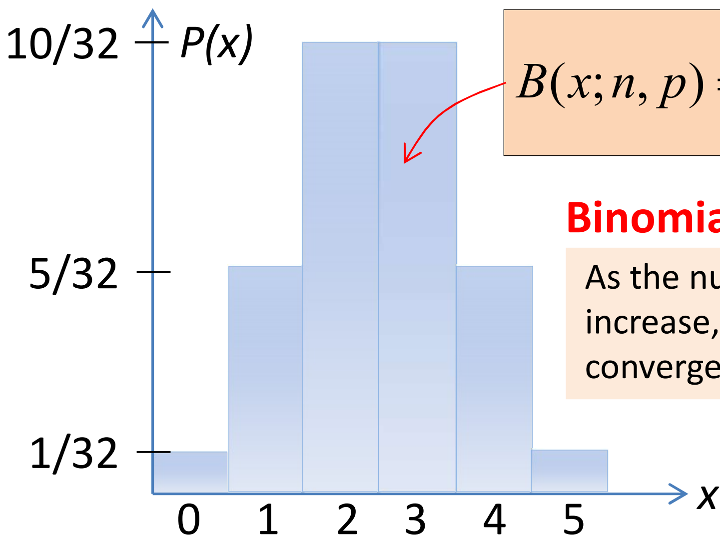
# Binomial distribution – cont'd

**Example**: x is the number of Heads (H) in flipping a coin 5 times (probability of success p=0.5 for a fair coin)

Sample space: $2^5$=32 possible outcomes

If P(x) = P(# of Heads in 5 flips)

$$C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$




$$B(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

**Binomial distribution**

As the number of observations increase, binomial distribution converges to normal distribution

$$P(x=0) = \frac{C(5,0)}{32} = \frac{1}{32}$$

$$P(x=1) = \frac{C(5,1)}{32} = \frac{5}{32}$$

$$P(x=2) = \frac{C(5,2)}{32} = \frac{10}{32}$$

$$P(x=3) = \frac{C(5,3)}{32} = \frac{10}{32}$$

$$P(x=4) = \frac{C(5,4)}{32} = \frac{5}{32}$$

$$P(x=5) = \frac{C(5,5)}{32} = \frac{1}{32}$$

For proof, see: https://www.probabilisticworld.com/binomial-distribution-mean-variance-formulas-proof/

Mean = **μ = np = 5 x (1/2) = 2.5**

Variance = **$\sigma^2$ = np(1-p)** = npq  where q=1-p

# Example I

- You toss a coin 5 times. Given that you have at least 4 heads, what is the probability of getting 5 heads?

$$P(5H \mid at\_least\_4H) = \frac{P(at\_least\_4H \mid 5H)P(5H)}{P(at\_least\_4H)}$$

$$P(5H) = 1/32$$

$$P(at\_least\_4H) = P(4 \le k \le 5) = \sum_{k=4}^{5} \binom{5}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{5-k} = \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) + \binom{5}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0$$

$$P(at\_least\_4H) = \left(\frac{1}{2}\right)^5 (5+1) = \frac{6}{32} = 0.1875$$

```
from scipy.stats import binom
# binom(m,n,p) => probability that you get m or less
# successes out of n when prob of success is p
print(1-binom.cdf(3,5,0.5))
>>> 0.1875
```

**Python code**

$$P(5H \mid at\_least\_4H) = \frac{P(at\_least\_4H \mid 5H)P(5H)}{P(at\_least\_4H)} = \frac{(1)(1/32)}{6/32} = \frac{1}{6} = 0.167$$

# Example II

- You toss a coin 30 times and you see 22 heads. Is this a fair coin?
  - What is the probability of a fair coin showing 22 heads (out of 30 tosses) simply by chance?

$$P(N_H, N_T) = \binom{N}{N_H} \left(\frac{1}{2}\right)^{N_H} \left(1 - \frac{1}{2}\right)^{N_T}$$

number of arrangements (binomial coefficients)

prob of heads

prob of tails

  - The probability of getting 22 heads or more in 30 flips

$$P(k \geq N_H, N) = \sum_{k=N_H}^{N} \binom{N}{k} p^k (1-p)^{N-k}$$

$$P(k \geq 22, 30) = \sum_{k=22}^{30} \binom{30}{k} 0.5^k 0.5^{30-k} = 0.008 = 0.8\%$$

# Poisson distribution

- A probability distribution used to model **count of things for a fixed interval of time or space**
  - Outcome: Success or Failure
  - Poisson constant/rate (**λ**): Average **number of successes** occurring **in a fixed region** (length, area, volume, time) Example: Number of calls a call center receives in an hour
- A **Poisson random variable** is the number of successes resulting from a Poisson experiment
- Poisson distribution: The probability distribution of a Poisson RV (defined for integer values of **k**)

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

E(X) = Var(X) = **λ**

# Poisson distribution

- A Poisson distribution is an approximation for the Binomial distribution when sample size n is large
- Expected value for a Binomial distribution: $E(X) = np$
- Expected value for a Poisson distribution : $E(X) = \lambda$
- A Binomial distribution is given by (where p=λ/n):

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

- What happens to the above equation for large n?

$$\lim_{n\to\infty} \frac{n!}{(n-k)!k!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} = \left(\frac{\lambda^k}{k!}\right) \lim_{n\to\infty} \frac{n!}{(n-k)!}\left(\frac{1}{n^k}\right)\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}$$

$$\lim_{n\to\infty} \frac{n!}{(n-k)!}\left(\frac{1}{n^k}\right) = \lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots\cancel{(n-k)(n-k-1)\ldots(1)}}{\cancel{(n-k)(n-k-1)\ldots(1)}}\left(\frac{1}{n^k}\right) = \lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}$$

# Poisson distribution

$$\lim_{n\to\infty}\frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}=\lim_{n\to\infty}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\ldots\left(\frac{n-k+1}{n}\right)\approx 1$$

- What remains is:

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\approx 1}\approx\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\approx e^{-\lambda}$$

- Finally: $\boxed{P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}=\frac{\lambda^k e^{-\lambda}}{k!}}$

- **Example**: Calls arrive at a call center randomly at an average rate of **2 calls per minute**. What is the probability of observing ≥ 3 calls in a given minute at the call center?

$\lambda$ = 2 calls/min

P(x ≥ 3) = P(x=3) + P(x=4) + P(x=5) + . . .
= 1 – P(x < 3) = 1 – [ P(x=0) + P(x=1) + P(x=2) ]
= 1 – [$2^0 e^{-2}$ / 0! + $2^1 e^{-2}$ / 1! + $2^2 e^{-2}$ / 2! ] = 0.323 (32.3%)

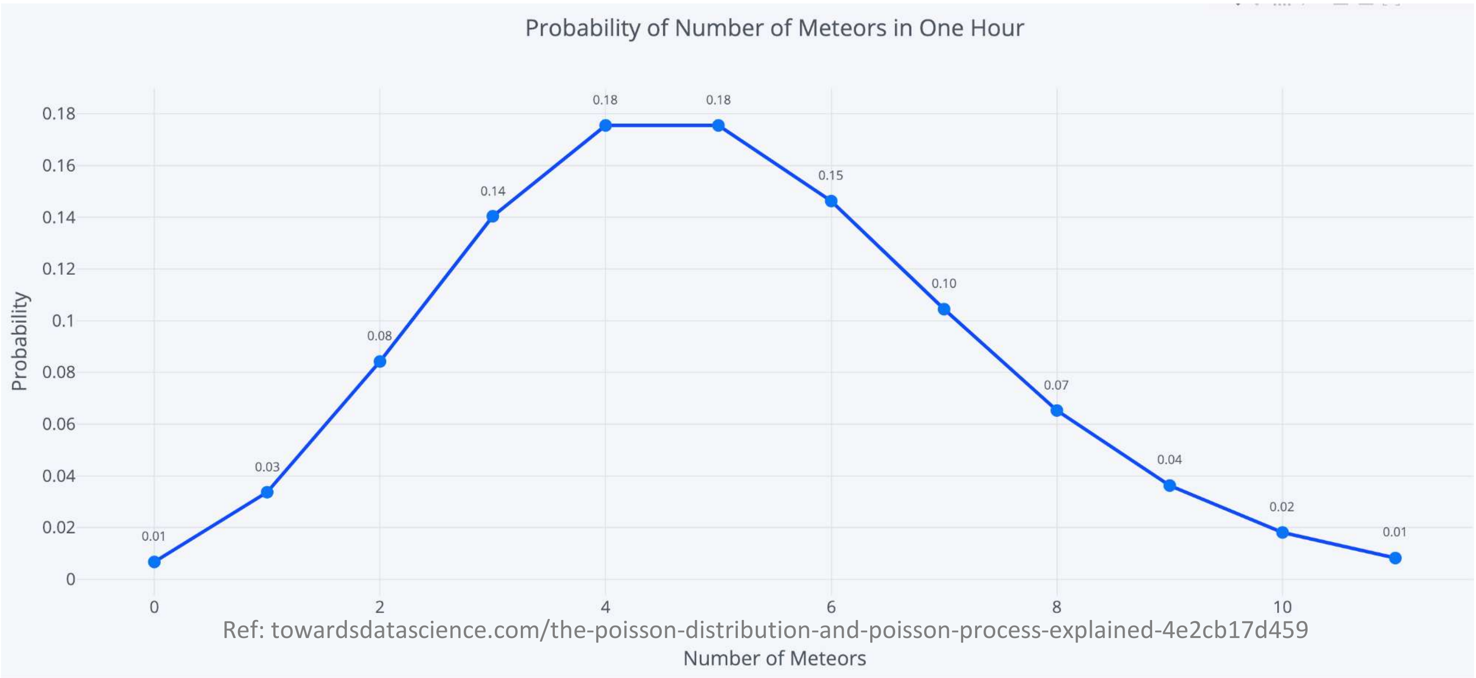
# Example

- Every time you go out at night for sky watching, you get to see 1 meteor in every 12 minutes on average. What is the probability of seeing 3 meteors in an hour?

$$\lambda = \frac{\textbf{events}}{\textbf{interval}} \textbf{ x event lenght} = \frac{1}{12} \textbf{ x } 60 = 5 \textbf{ meteors/hr}$$

$$\boxed{P(X=3) = \frac{\lambda^3}{3!} e^{-\lambda} = \frac{5^3}{3!} e^{-5} = 0.14 \approx 1/7}$$

When you go out every night, you get to see 3 meteors in an hour once in a week.



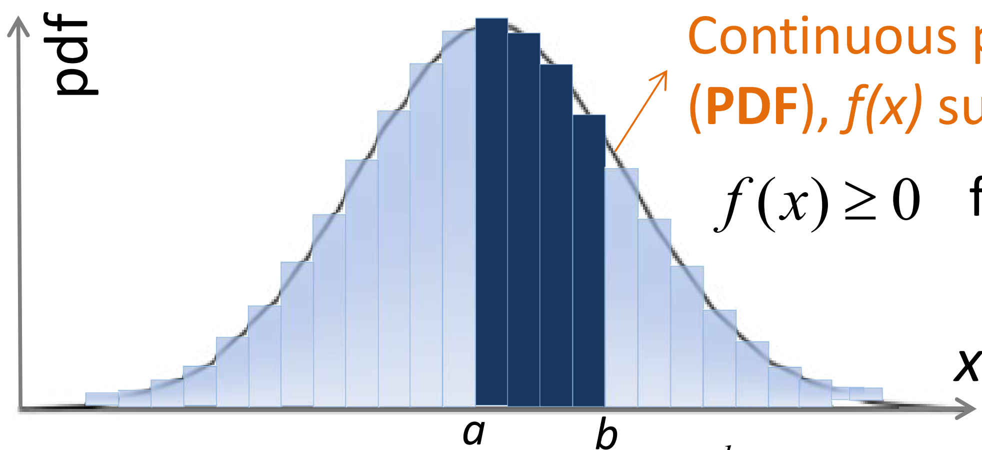

Ref: towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459

# Continuous probability distributions

**Continuous random variable**: A random variable that can take an infinite number of values between any two given values

Example: Exact amount of rainfall in Istanbul on a rainy day

**Continuous probability function**: A smooth curve that closely approximates the relative frequency histogram of many continuous random variables (a.k.a. probability density function)




Continuous probability density function (**PDF**), *f(x)* such that:

$$f(x) \geq 0 \quad \text{for all x and} \quad \int_{-\infty}^{+\infty} f(x)dx = 1$$

Expected value (mean):

$$E(x) = \mu = \int_{-\infty}^{+\infty} xf(x)dx$$

Variance:

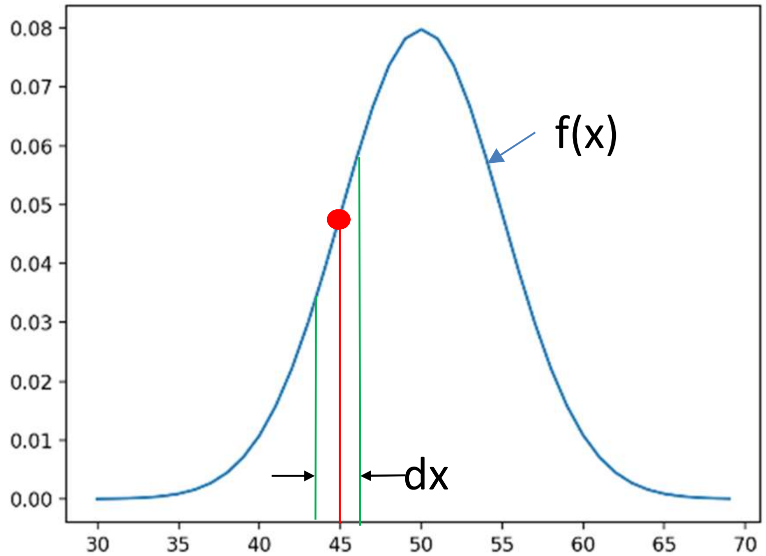
$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$$

Example: Probability(a < x < b) = $\int_a^b f(x)dx$

Probability(x < b) = $\int_{-\infty}^{b} f(x)dx$

# Continuous probability distributions – cont'd

- **Caution**: PDF is not a probability!

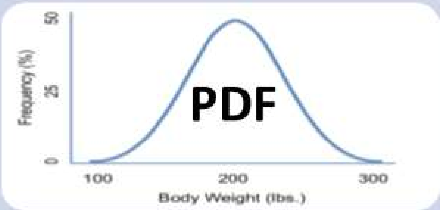


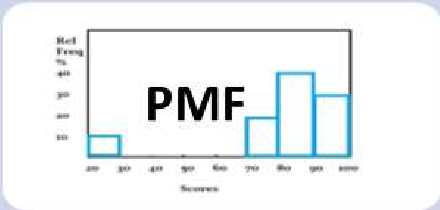

f(x=45) = 0

Area between 45-dx/2 and 45+dx/2 is a probability

- PDF can be greater than 1. Only the total area under a PDF must equal 1.
- Note, on the other hand, that PMF = probability. Because discrete and continuous random variables aren't defined the same way. Remember from Physics that you integrate "density" to get the "mass".

# CDF vs PDF vs PMF

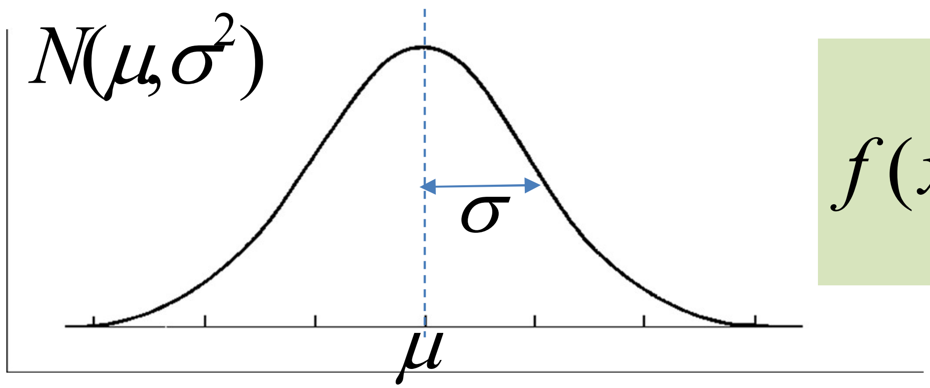




| | **Cumulative Density Function** | **Probability Density Function** | **Probability Mass Function** |
|---|---|---|---|
| Purpose | **Cumulative probability** associated with a function. | Probabilities for continuous **random variables.** | Probabilities for **discrete random variables.** |
| Example | **Cumulative** value from negative infinity up to a random variable X (i.e. x < 10) | Probability of a range of outcomes (e.g. X = 5 to 6) | Probability of a certain outcome (e.g. X = 6) |
| Properties | Integral of the PDF. A CDF has [2]:<br>a/ Left limit = 0, right limit = 1<br>b/ Nondecreasing<br>c/ Right continuous (defined up to a point) [3]. | Derivative of the CDF. A PDF satisfies the following [4]:<br>a/ It is positive everywhere<br>b/ AUC = 1<br>c/ Total probability = integral of f(x) | Satisfies the following[4]:<br>a/ It is positive everywhere<br>b/ AUC = 1<br>c/ Total probability = summations of individual probabilities. |

Ref: https://www.datasciencecentral.com/profiles/blogs/probability-mass-function-vs-probability-density-function

# Normal distribution

- A specific bell-shaped, symmetric curve (aka Gaussian)
- The most important and useful continuous distribution (explains many of the phenomena observed in life)
- The normal curve is represented by the density function:



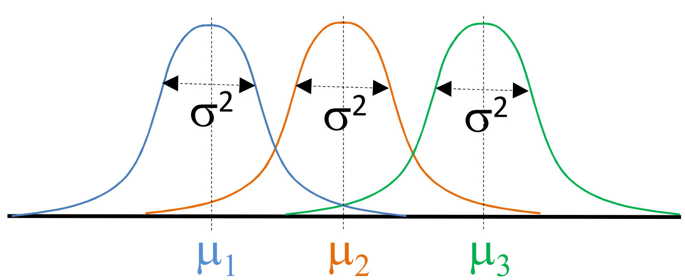

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- The value of $\mu$ is determines where the curve is to be centered, and the value of $\sigma^2$ determines how spread out the curve is.
- The normal curve is completely determined by $\mu$ and $\sigma^2$ thus the notation ***N(*$\mu$ *,*$\sigma^2$*)***.

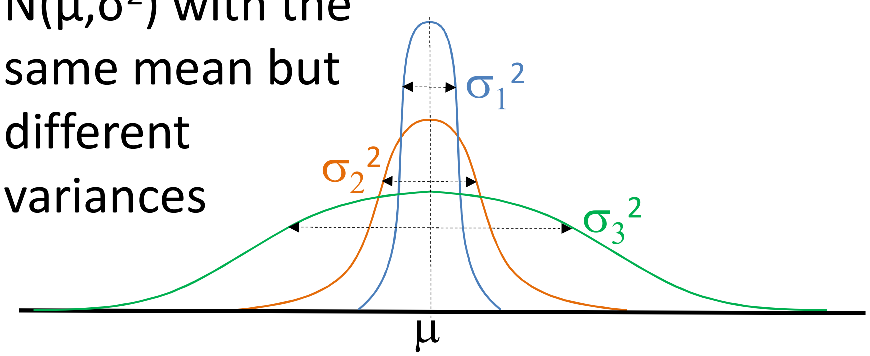
## Properties of N($\mu$ ,$\sigma^2$)

- Characteristics of a Normal distribution
  - A symmetric (bell-shaped) curve around $x = \mu$
  - Extends from $-\infty$ to $+\infty$
  - Total area under the curve is 1 (as in all pdf's)
  - Curve is always above the horizontal axis, $f(x) \geq 0$
  - The mean, the median and the mode are all equal
  - No closed form solution exists for $\int_a^b f(x)dx$

N($\mu$,$\sigma^2$) with different means but equal variances

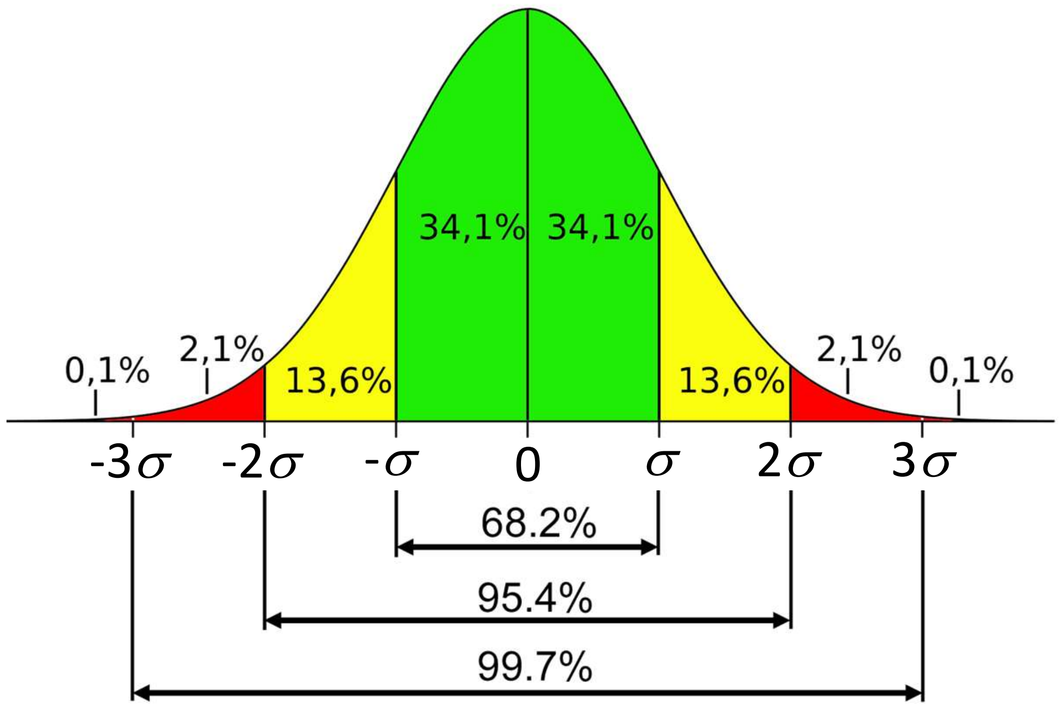



N($\mu$,$\sigma^2$) with the same mean but different variances



# Properties of N($\mu$ ,$\sigma^2$)

- Characteristics of a Normal distribution




  – 68% of all observation fall within 1 standard deviation,
  – 95% of all observation fall within 2 standard deviations, 99.7% of all observations fall within 3 standard deviations away from the mean to both directions

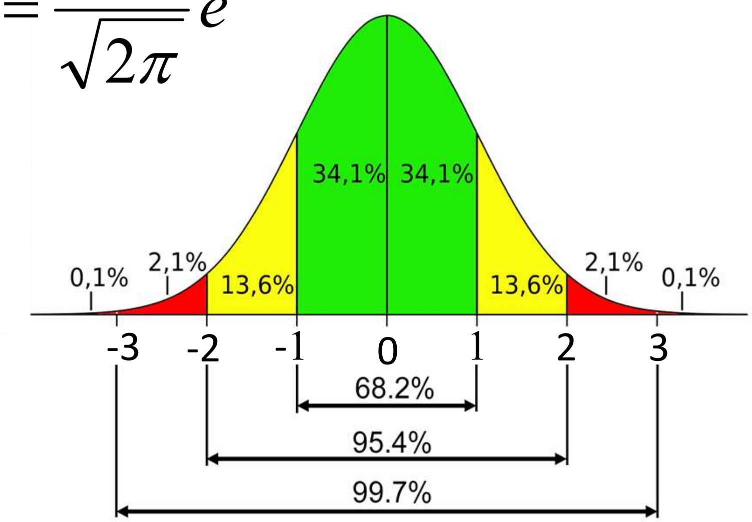
# Standard normal distribution

- How can we compare distributions of different scales?
- We cannot! Therefore we need to put them on a same scale for proper comparison: Standard normal distribution
- A standard normal distribution is a normal distribution with 0 mean and unit variance: *N(0,1)* where $\mu=0$ and $\sigma^2=1$
- **Standardization transformation**:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$f(Z)=\frac{1}{\sqrt{2\pi}}e^{-Z^2/2}$$

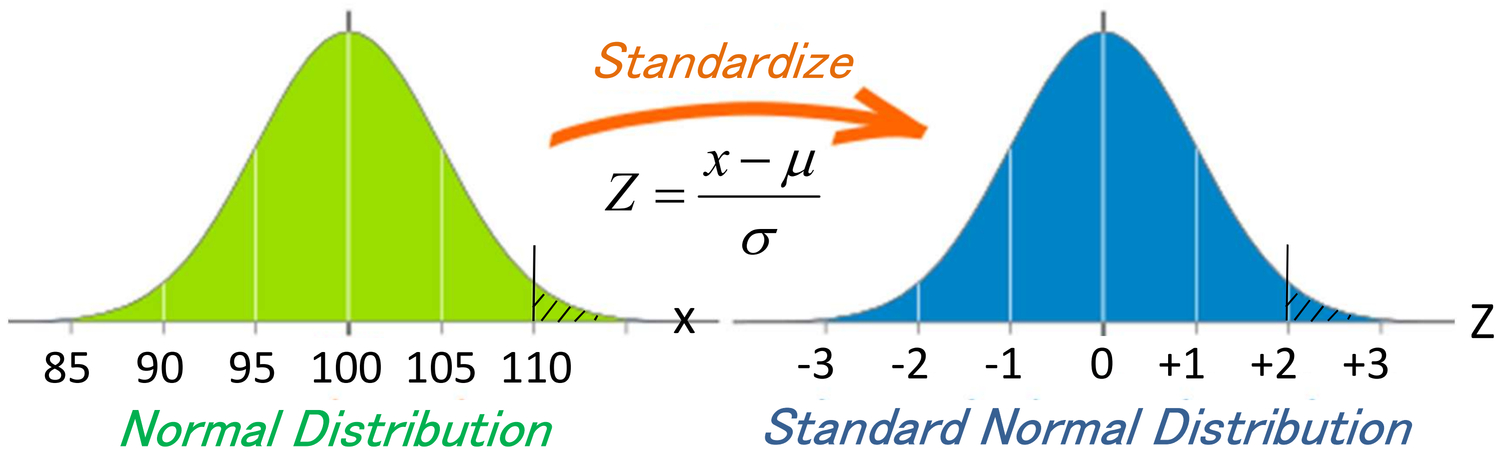
$$Z=\frac{x-\mu}{\sigma}$$




where Z (Z-score) represents the standard score of *x*

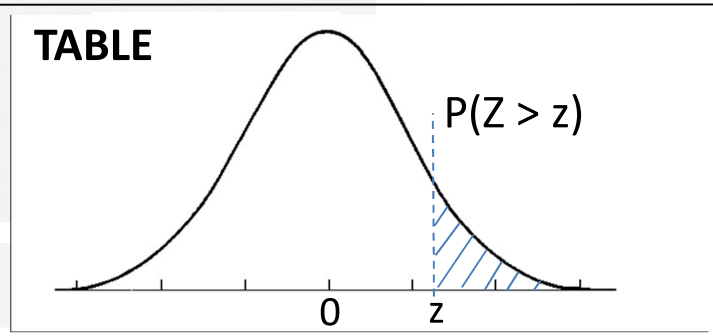
# Standard normal distribution – An example

- If x is a normal RV with μ=100 and σ=5, what is the probability that X is larger than 110? P(X > 110) = ?




$$Z = \frac{x-\mu}{\sigma}$$

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 |
|---|---|---|---|---|---|---|---|
| 0.0 | 0.5000 | 0.4960 | 0.4920 | 0.4880 | 0.4841 | 0.4801 | 0.4761 |
| 0.1 | 0.4602 | 0.4562 | 0.4522 | 0.4483 | 0.4443 | 0.4404 | 0.4364 |
| 0.2 | 0.4207 | 0.4168 | 0.4129 | 0.4090 | 0.4052 | 0.4013 | 0.3974 |
| . . . | | | | | | | |
| 2.0 | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 |
| 2.1 | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 |



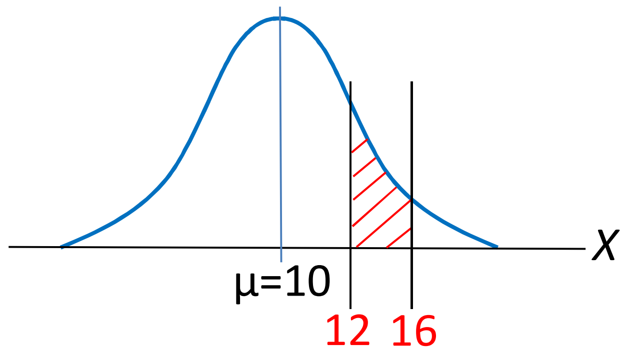

$$P(x > 110) = P\left(\frac{x-\mu}{\sigma} > \frac{110-100}{5}\right) = P(Z > 2.0) = 0.0228 = 2.28\%$$

# Example

- Calculating areas under a Normal distribution
  - If a random variable X has a normal distribution with mean 10 and variance 25. What is the area under the curve between 12 & 16?
  - Standardizing transformation:

$$Z_1 = \frac{x-\mu}{\sigma} = \frac{12-10}{5} = 0.4$$

$$Z_2 = \frac{x-\mu}{\sigma} = \frac{16-10}{5} = 1.2$$




P(data|distribution)

P(12 ≤ X ≤ 16) = P(0.4 ≤ Z ≤ 1.2) = P(0 ≤ Z ≤ 1.2) - P(0 ≤ Z ≤ 0.4)

= 0.3849 – 0.1554 = 0.2295

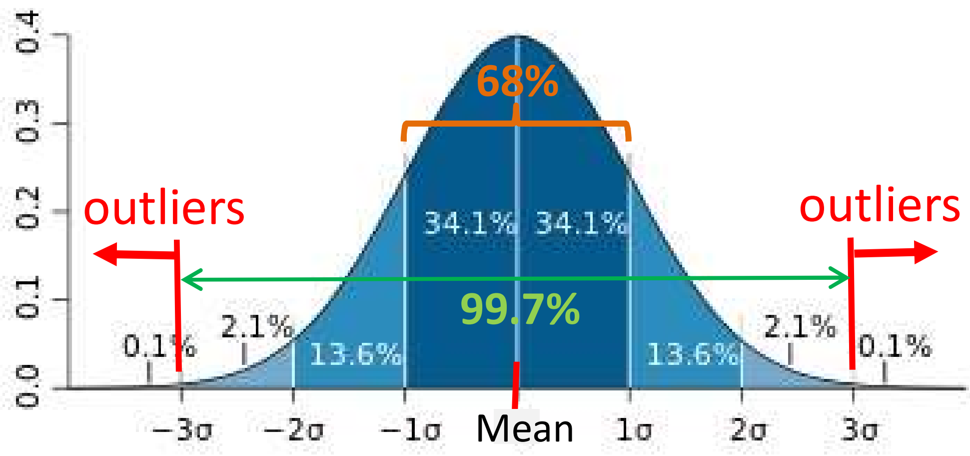
P(12 ≤ X ≤ 16 | μ=12, σ=4) : Probability of RV's falling in between 12 & 16 given a population distribution with μ=12 and σ=4

```
import scipy.stats as stats
print(stats.norm.cdf(1.2)-stats.norm.cdf(0.4))
# computes left tail probability by default
>>> 0.229508588168
```

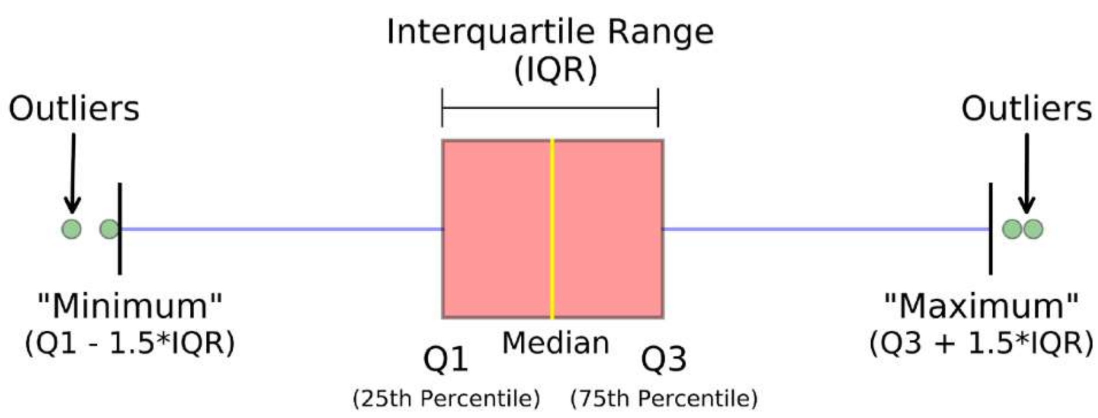
**Python code**

# Detecting outliers

- Capping: Values beyond ±**3σ** from the mean in a distribution

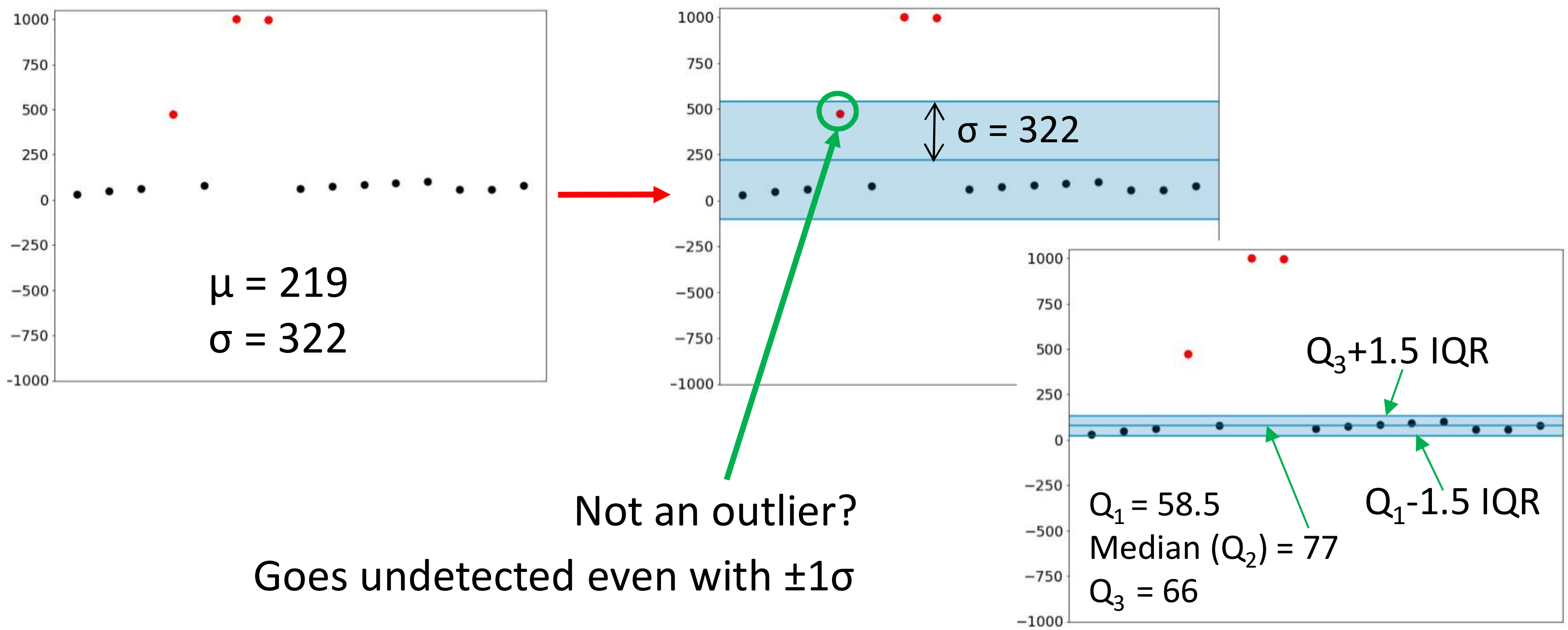



- Remember in a boxplot we had:



# Detecting outliers – cont'd

- IQR vs Capping based on standard deviation (σ)
- Using σ for detecting outliers might be problematic in the presence of extreme outliers. They inflate the σ so much that even a moderate outlier can go undetected.
- Suppose we have a data set:
- D=[ 30, 50, 63, 474, 78, 999, 997, 61, 74, 83, 92, 100, 55, 56, 77 ]




Ref: medium.com/@davidnh8/outlier-detection-101-median-and-interquartile-range-cc9dde94c0ac

# Summary of distributions

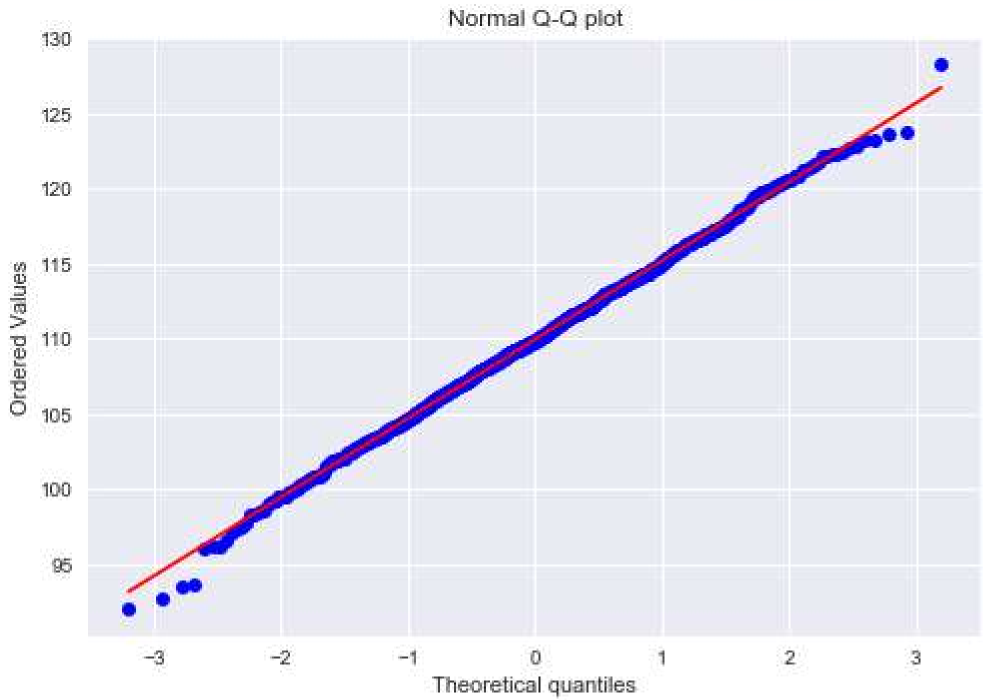
- **Discrete Random Variables**
  - Bernoulli Distribution/Trial
    - Apps: coin toss, success/failure experiments
  - **Binomial Distribution** (collection of iid Bernoulli trials)
    - Apps: Obtaining k heads in n tossings of a coin; Receiving k bits correctly in n transmitted bits; Batch arrivals of k packets from n inputs at an ATM switch
  - Geometric Distribution
    - Apps: To represent the number of dice rolls needed until you roll a six; Queueing theory and discrete Markov chains
  - Poisson Distribution
    - Apps: The number of phone calls at a call center per minute; The number of times a web page is accessed per minute; The number of spelling mistakes one makes while typing a single page
- **Continuous Random Variables**
  - Exponential Distribution
    - Apps (similar to Poisson): The time it takes before your next telephone call
  - Uniform Distribution
    - Apps: To test a statistic for the simple Null Hyp
  - **Normal Distribution**
    - Apps: Many physical phenomena like noise, measurement errors, financial variables etc.
  - Log-normal Distribution
    - Apps: The **long-term** return rate on a stock investment
  - Gamma (Erlang) Distribution
    - Apps: Waiting time for n calls made to a switching center

# Quantile-Quantile plot

- Q-Q plot is a visual tool to assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

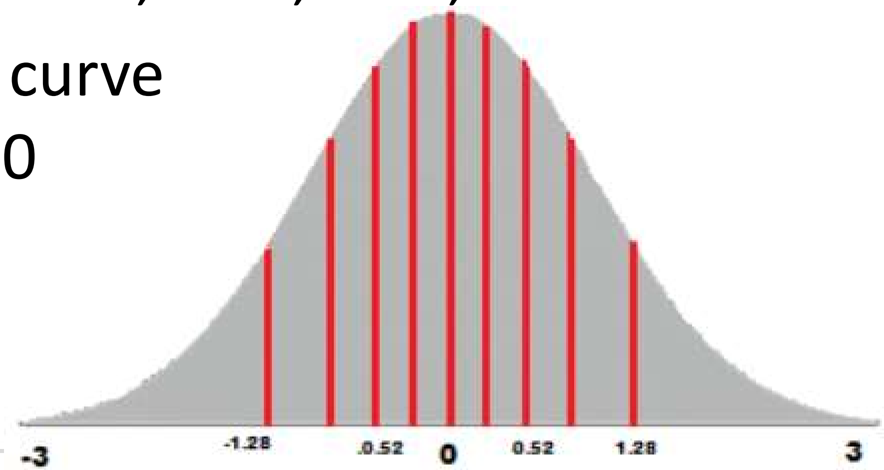
```
import scipy.stats as stats
stats.probplot(IQ, dist="norm", plot=plt)
plt.title("Normal Q-Q plot")
plt.show()
```

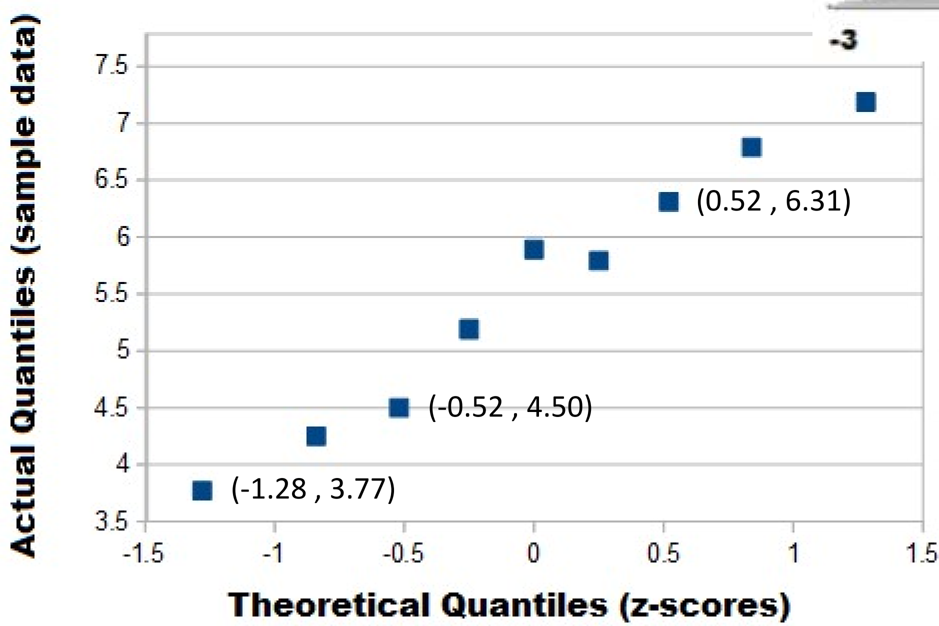
**Python code**



# Q-Q plot – example

- Do the followings come from a normal distribution?
- 3.77, 4.25, 4.50, 5.19, 5.79, 5.89, 6.31, 6.79, 7.19
- Draw a normal distribution curve (divide it into n+1 = 9+1 = 10 equally sized areas):

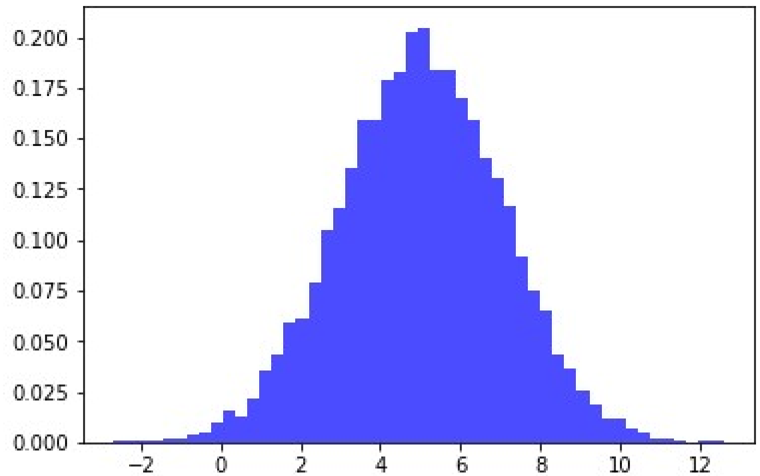


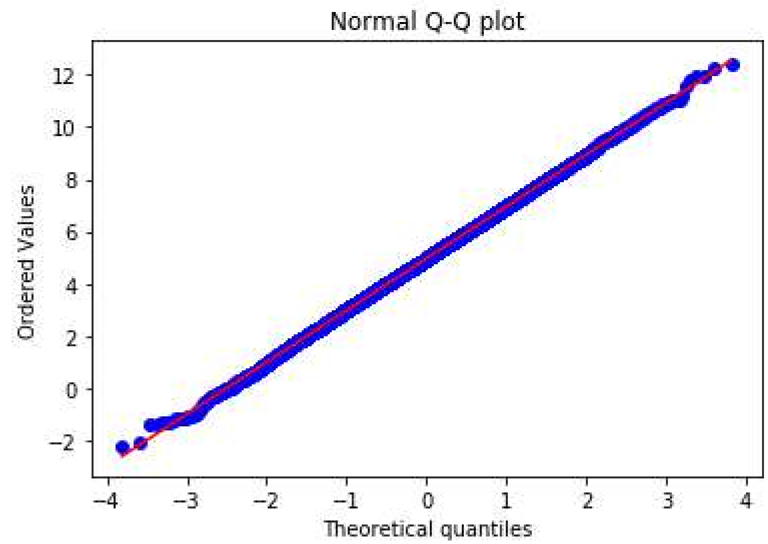




- 10% = -1.28
- 20% = -0.84
- 30% = -0.52
- 40% = -0.25
- 50% = 0
- 60% = 0.25
- 70% = 0.52
- 80% = 0.84
- 90% = 1.28
- 100% = 3.0

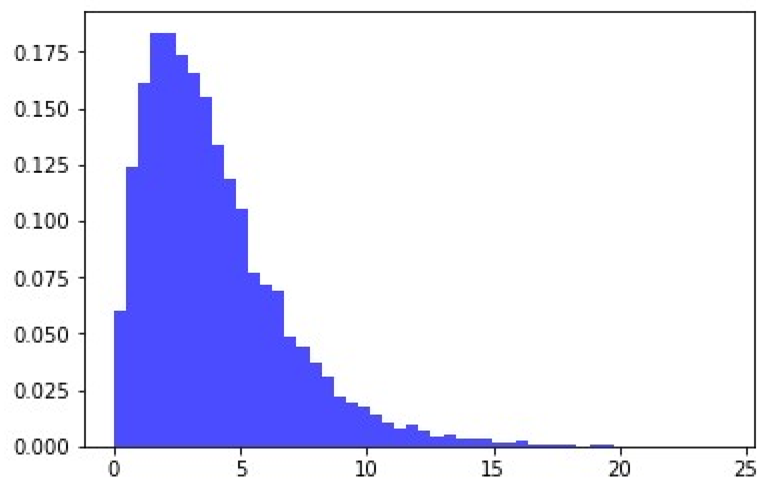
# Q-Q plot – cont'd

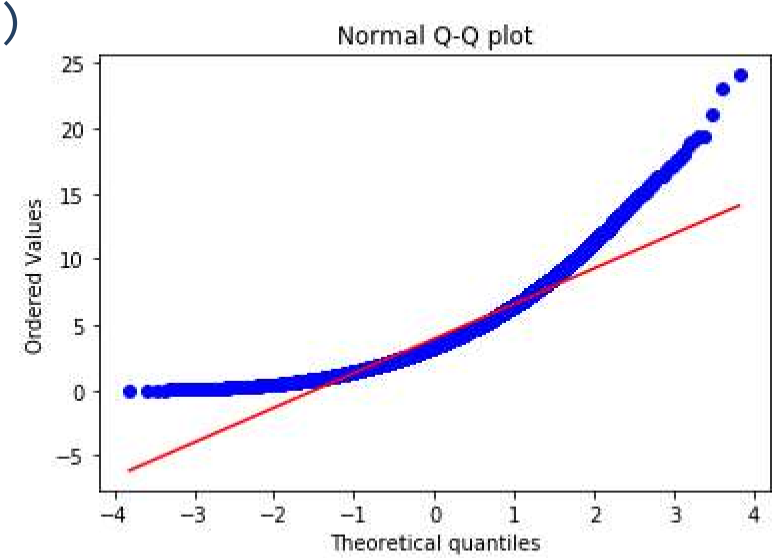
```
y = np.random.normal(5.0, 2.0, size=10000) #mean=5.0 and sd=2.0
plt.hist(y, bins=50 ,normed=True)
```





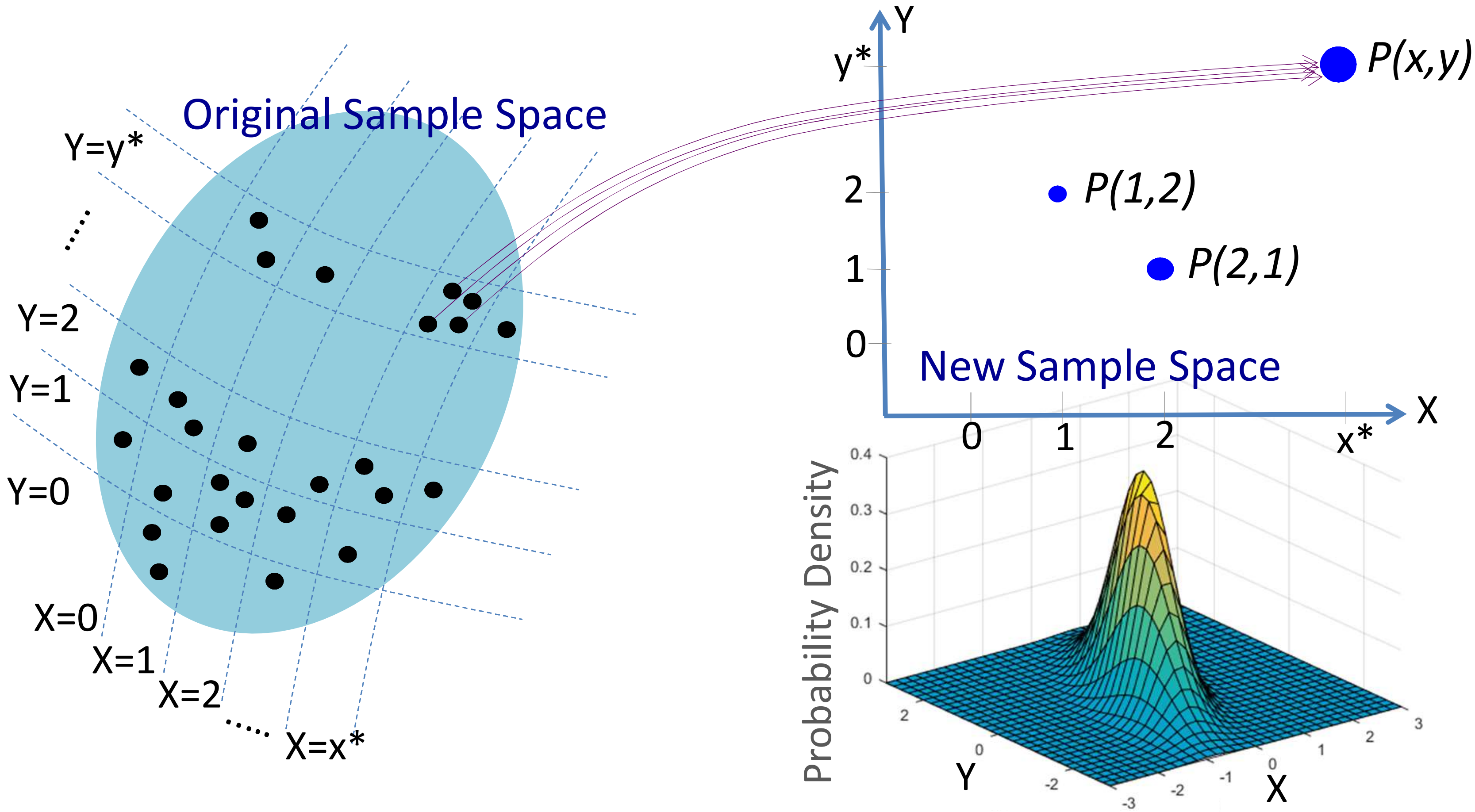

```
y = np.random.gamma(shape=2, scale=2, size=10000)
plt.hist(y, bins=50 ,normed=True)
```





# Joint probability distributions

- In the presence of 2 or more RV's, the resulting probability distribution is a joint probability density function
- In the special case of 2 RV's: Bivariate probability function



# Joint probability distributions – cont'd

- Formal definition of the joint probability density:

  $f_{XY}(x,y) = P(X=x, Y=y) = P(X=x \cap Y=y)$

- **Marginal probability** functions: $P(X=x)$ and $P(Y=y)$
  - Individual probability distribution for X or for Y given the joint probability distribution for X and Y
  - Distribution for X becomes:

    $P(X=x) = \sum_{y} P(x,y)$ and $f_X(X=x) = \int_{y} f_{XY}(x,y)\, dy$

- **Conditional probability**:

$$f_{Y|X}(y \mid x) = P(Y = y \mid X = x) = \frac{\overbrace{P(X = x \cap Y = y)}^{\textit{Joint density of X and Y}}}{\underbrace{P(X = x)}_{\textit{Marginal density of X}}} = \frac{f_{XY}(x, y)}{f_X(x)}$$

- Independence of discrete RV's (x and y are independent):

  $P(X = x \mid Y = y) = P(X = x)P(Y = y)$ for all possible values of x and y

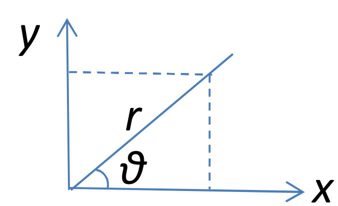
# Addendum

- Normal distribution N(0,1) has a probability density function f(z) in the form of: $f(Z) = ce^{-Z^2/2}$
- We've seen that $c = 1/\sqrt{2\pi}$ Where does it come from?
  - "c" is a normalizing constant to make the area "1"
  - And $e^{-Z^2/2}$ decays (goes to zero) fast as $z \to \infty$
- We know that: $\int_{-\infty}^{+\infty} ce^{-Z^2/2} dz = 1$ (an indefinite integral and is impossible to solve in closed form)
- Let's define $I$ as $I = \int_{-\infty}^{+\infty} e^{-Z^2/2} dz$ and compute $I^2$ instead:

$$I^2 = \int e^{-Z^2/2} dz \int e^{-Z^2/2} dz = \int e^{-x^2/2} dx \int e^{-y^2/2} dy = \int e^{-(x^2+y^2)/2} dx dy$$

- By transformation of coordinates:

$$x^2 + y^2 = r^2 \qquad x = r\cos\theta \qquad y = r\sin\theta$$



## Addendum (cont'd)

- Jacobian: $\left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| dr d\theta = r dr d\theta$

$I^2 = \int_{\theta=0}^{2\pi}\int_{r=0}^{\infty} e^{-r^2/2} r dr d\theta$ using $u=r^2/2$ and $du=rdr$

$$I^2 = \int_0^{2\pi} d\theta \int_0^{\infty} e^{-r^2/2} r dr = \int_0^{2\pi}\left[\int_0^{\infty} e^{-u} du\right] d\theta = 2\pi \Rightarrow I = \sqrt{2\pi}$$

$$\int_{-\infty}^{+\infty} c e^{-Z^2/2} dz = cI = 1 \Rightarrow c = 1/\sqrt{2\pi}$$

- So the standard normal distribution is: $f(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}$
- Mean (expected value): $E(Z) = 1/\sqrt{2\pi}\int_{-\infty}^{+\infty} z e^{-Z^2/2} dz = 0$

  (by symmetry)
- Variance: *Var(Z) = E(Z²)- [E(Z)]² = E(Z²)*

$$Var(Z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} z^2 e^{-Z^2/2} dz = \frac{2}{\sqrt{2\pi}}\int_0^{\infty} z^2 e^{-Z^2/2} dz = 1$$

Prove (hint: integration by parts)