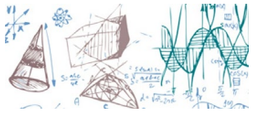
Introduction to probability, statistics and data handling

Agnieszka Obłąkowska-Mucha
Tomasz Szumlak

Faculty of Physics and Applied Computer Science
AGH University of Krakow



The Binomial Distribution

- We already know this distribution – it emerges when we consider an experiment such as tossing a coin, rolling a die or choosing a marble from a box repeatedly.
- So, we considering trials. Each outcome will have constant probability assigned (that should not change in time, and is the parameter of the Bernoulli prob. model family).
  - Sometimes we are also interested in processes where the probability is not constant (out of the scope of our lecture, however)
- We then say that p is a success and q is a failure (in a Bernoulli sense) and can compose the following P.D.F.

$$f(x) = B(n,p) = p(X = x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

- The RV denote the number of successes x in n trials, $x = 0, 1, \dots, n$

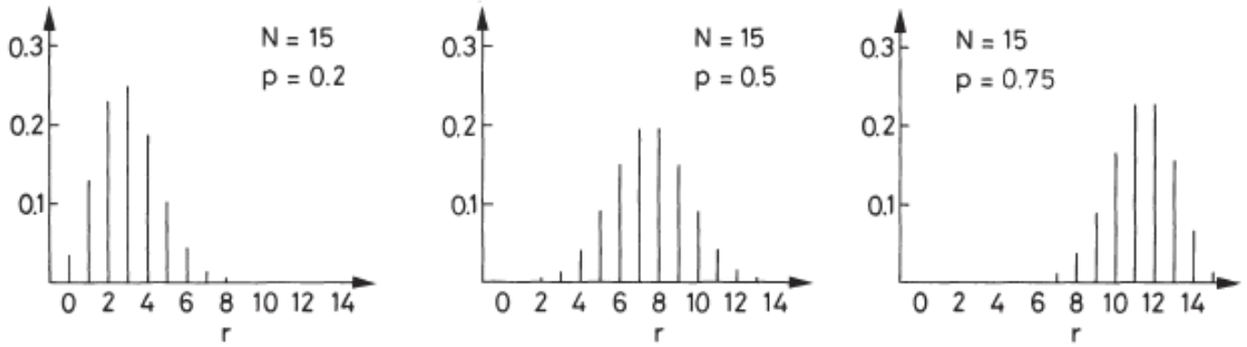
The Binomial Distribution

- The mean and variance can be fairly easy calculated:

$$\mu = \sum_x xP(x) = np$$

$$\sigma^2 = \sum_x (x-\mu)^2 P(x) = np(1-p)$$

- In the limit of „large” n and „no too small” p we can very accurately approximate the Binomial distribution with Gaussian one



The Binomial Distribution

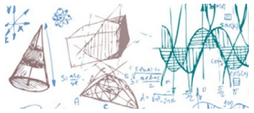
- **Properties of the Binomial P.D.F.**

Mean	$\mu = np$
Variance	$\sigma^2 = npq$
Standard deviation	$\sigma = \sqrt{npq}$
Coefficient of skewness	$\alpha_3 = \frac{q - p}{\sqrt{npq}}$
Coefficient of kurtosis	$\alpha_4 = 3 + \frac{1 - 6pq}{npq}$

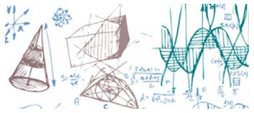
- We can mote something interesting here
- **Theorem 6**. Let X be the RV giving the number of successes in n Bernouli trials, so that $\frac{X}{n}$ is the proportion of successes. Then if p is the probability of success and ϵ is any positive number:

$$\lim_{n\to\infty} prob\left(\left|\frac{X}{n} - p\right| \geq \epsilon\right) = 0$$

Poisson distribution



- The RV of discrete type:
- - the number of outcomes occurring, for instance, during a given time (e.g. number of radioactive decays in a sample of radioactive material) t: $X = X_t = 0, 1, 2, \ldots$
- number of events in a given region of space - e.g. number of typing errors per page
- or:
  - ✓ telephone calls arriving during a (short) period of time;
  - ✓ light quanta (photons) arriving at detecting system;
  - ✓ number of mutations on a strand of DNA (per unit length);
  - ✓ number of customers arriving at a counter;
  - ✓ number of cars arriving at a trac light;
  - ✓ number of Losses/Claims;



Poisson distribution

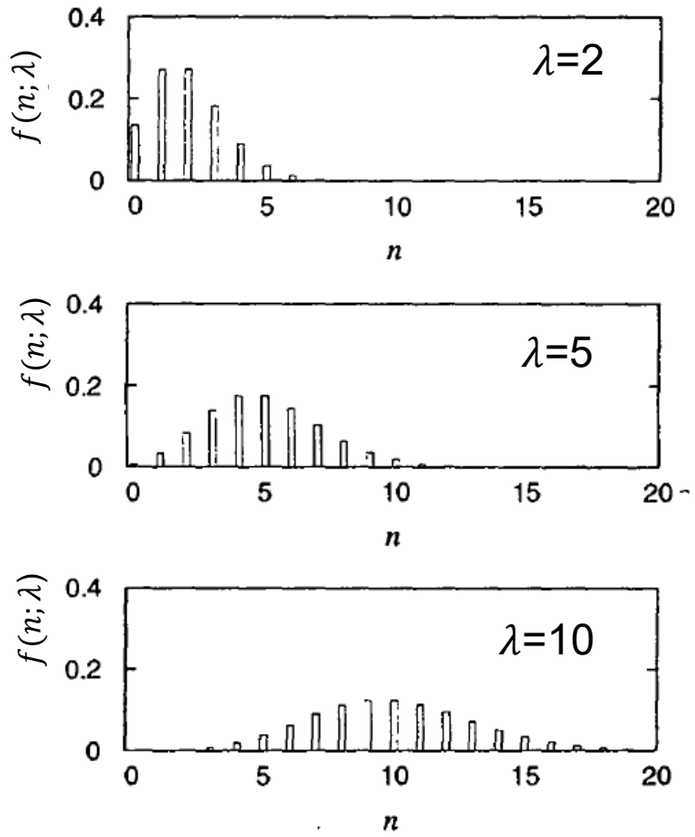
- Numbers of outcomes occurring in one time interval t are independent of each other, i.e. the number occurring in one time interval is independent of the number that occurs in any other disjoint time interval (Poisson process has no memory)
- The probability that a single outcome will occur during a very short time interval t is PROPORTIONAL to the length of interval:

$$P(X_{\Delta t} = 1) \sim \Delta t$$

- Let me to introduce the Poisson distribution:

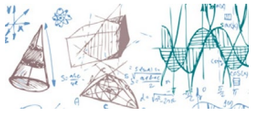
$$P(X_t = k, \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Poisson distribution



- The number of decays of radioactive material in a fixed time period follows the Poisson distribution (given that the decay probability is constant over the time period)

Poisson distribution



- Let's take the binomial distribution, it can be shown that in the limit of large n and very small p (given that np is finite) we get a new distribution:

$$f(n; n \cdot p = \lambda) = \frac{\lambda^n}{n!} e^{-\lambda}$$

- This Poisson distribution is valid for integer variable n ($n = 0, 1, \ldots$) and has a single parameter $\boldsymbol{n \cdot p = \nu}$

- Its mean value and variance

$$E[n] = \sum_{n/0}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} = \lambda$$

$$V[n] = \sum_{n/0}^{\infty} (n - \lambda)^2 \frac{\lambda^n}{n!} e^{-\lambda} = \lambda$$

- In many cases the n can be treated as a continuous variable. Also, if ν is large the Poisson random variable can be treated as a continuous variable similar to the **normal distribution**

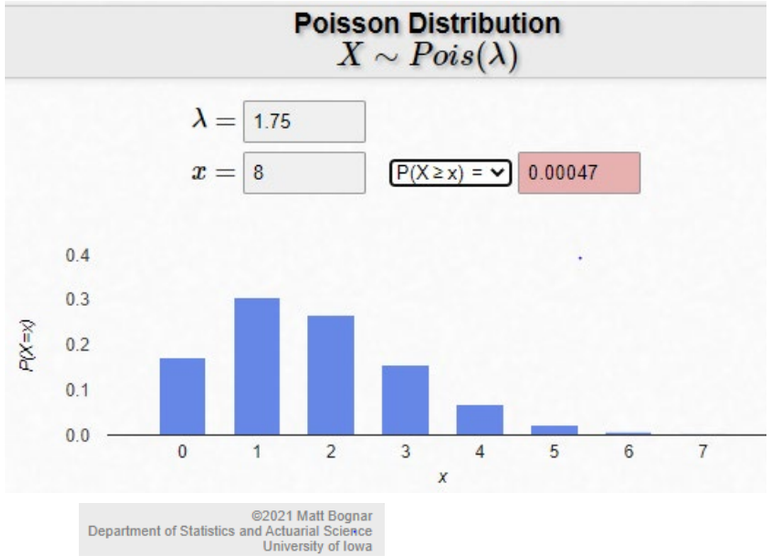
Poisson distribution

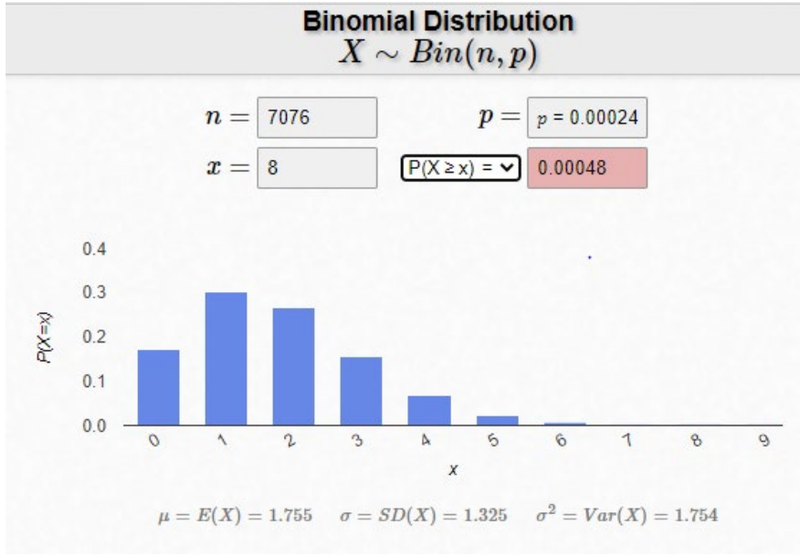
- The probability of getting leukemia is $p = 0.000248$. Using the approximation Bernoulli-to-Poisson find the P of eight or more leukemia cases in a population of size $n = 7076$.

$$np = 7076 \times 0.000248 = 1.75 = \lambda.$$

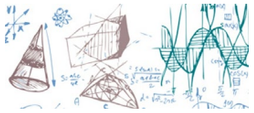
$$\mathcal{P}(X \leq 7) = \sum_{0 \leq x \leq 7} e^{-1.75} \frac{1.75^x}{x!} = 0.999518.$$

hence $\mathcal{P}(X \geq 8) \approx 1 - 0.999518 = 0.000482.$





Uniform distribution



- Uniform P.D.F. is defined for the C.R.V. and given by

$$f(x;a,b)=\begin{cases}\dfrac{1}{b-a} & a\le x\le b\\ 0 & \text{otherwise}\end{cases}$$

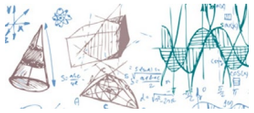
- So, we say that x is equally likely to be found inside our interval of interest (a,b). The mean and variance:

$$E[x]=\int_a^b \frac{x}{b-a}dx=\frac{1}{2}(a+b)$$

$$V[x]=\int_a^b \left(x-\frac{1}{2}(a+b)\right)^2\frac{1}{b-a}dx=\frac{1}{12}(b-a)^2$$

- There is a very important application of the uniform model related to the fact that for any R.V. x with P.D.F. $f(x)$ we can easily find transformation to a new variable that is uniform.

Uniform distribution



- If we call the new transformed variable y, the transformation rule is simply related with calculating the C.D.F. Cool!

$$x \to y: y = F(x)$$

- Remember, x – any P.D.F., y uniform P.D.F. Next, for any C.D.F. the following is true

$$\frac{dy}{dx} = \frac{d}{dx}\int_{-\infty}^{x} f(x')dx' = f(x)$$

- And using the rule for change of variables

$$g(y) = f(x)\left|\frac{dx}{dy}\right| = f(x)\left|\frac{dy}{dx}\right|^{-1} = 1, \qquad (0 \le y \le 1)$$

- This is the fundamental rule of random number generator programs that can give us a list of numbers with any distribution. We are going to look at this in more detail.

Exponential distribution

- Exp. distribution describes the amount of time between some (relatively rare) events
- This model is used for C.R.V. x: $0 \leq x \leq \infty$ and is defined by

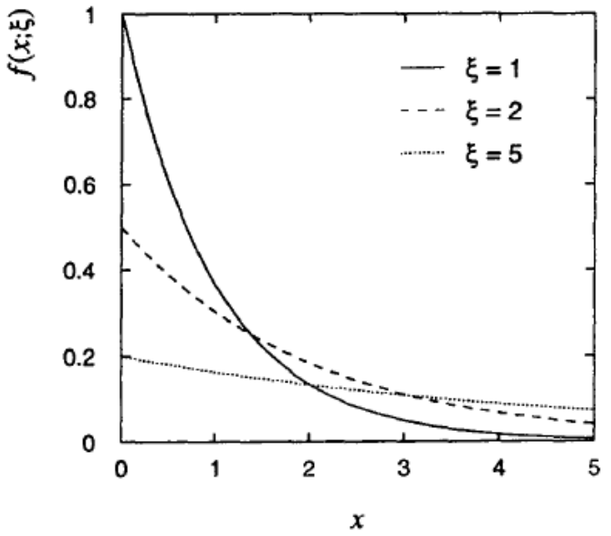
$$f(x;\xi) = \frac{1}{\xi} e^{-\frac{x}{\xi}}$$

- The model depends on a single parameter θ. The mean and variance are as follow:

$$E[x] = \frac{1}{\xi}\int_0^\infty x e^{-\frac{x}{\xi}} dx = \xi$$

$$V[x] = \frac{1}{\xi}\int_0^\infty (x-\xi)^2 e^{-\frac{x}{\xi}} dx = \xi^2$$

- The decay time of an unstable particle measured in its rest frame follows the exponential distribution. The parameter of the distribution is then interpreted as the mean lifetime.



Poisson and Exponential distribution

- Exp distribution: time between two events $X \rightarrow Exp\ (\mu)\ E[x] = \mu$.
- If the time between events is not affected by the times between previous events (times are independent) then number of events per unit time has Poisson distribution with $\lambda = 1/\mu$:

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

- Conversely, if the number of events per unit time follows a Poisson distribution, then the amount of time between events follows the exponential distribution.

Poisson and Exponential distribution

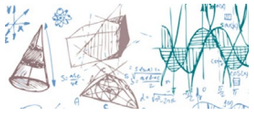
- There is a direct connection between the exponential distribution and the Poisson distribution (process).
- We may remember that the unique parameter of **Poisson** distribution can be interpreted as a **mean number of events per unit time**.
- So, if we consider a time interval of the length t the number of events should be $\boldsymbol{N = \lambda \cdot t}$
- Consider now a RV described by the time required for the first event to occur, X.
- The probability that the length of time until the event will exceed x is equal to the probability that **no** $(k = 0)$ Poisson events will occur in x.

$$P(k = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} \qquad \text{therefore:} \qquad P(X \geq x) = e^{-\lambda t}$$

- CDF for X is: $F(x) = P(0 < X < x) = 1 - e^{-\lambda t}$

  if we differentiate CDF: $f(x) = \lambda e^{-\lambda t}$ exponential distribution with $\xi = 1/\lambda$

The Normal Distribution



- This is definitely one of the most fundamental PDF with great significance in statistics

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < \infty$$

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(v-\mu)^2/2\sigma^2}\, dv$$

- We can also introduce the standardised variable corresponding to X

$$Z = \frac{X-\mu}{\sigma}, \mu_Z = 0, \sigma_Z = 1$$

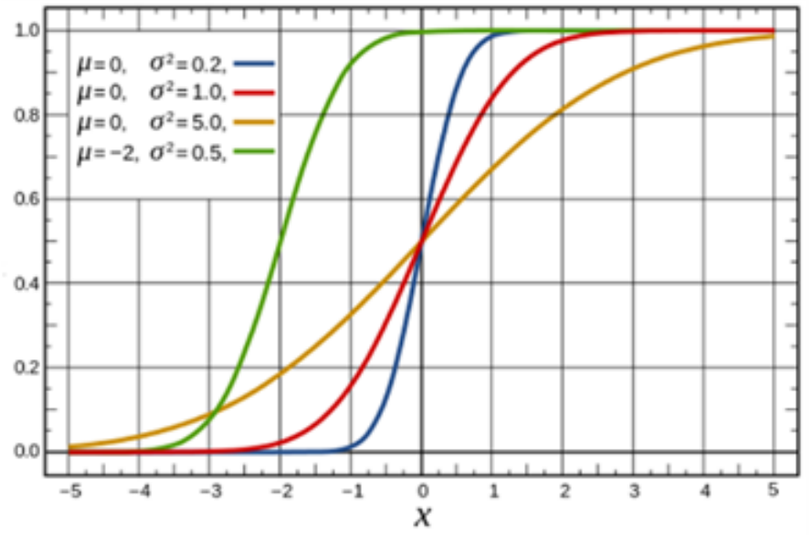
$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2}, -\infty < z < \infty$$

$$F(z) = P(Z \leq z) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{z} e^{-v^2/2}\, dv = \frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int_{0}^{z} e^{-v^2/2}\, dv$$

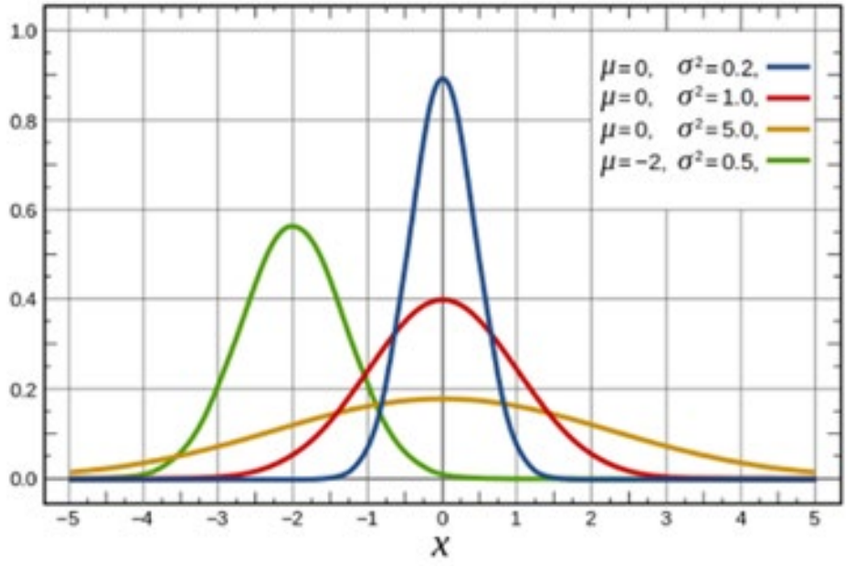
The Normal Distribution -CDF

- In statistics we deal with two major players: probability density function $\boldsymbol{PDF}$ and cumulative density function $\boldsymbol{CDF}$

Cumulative Density Function



Probability Density Function



The Normal Distribution

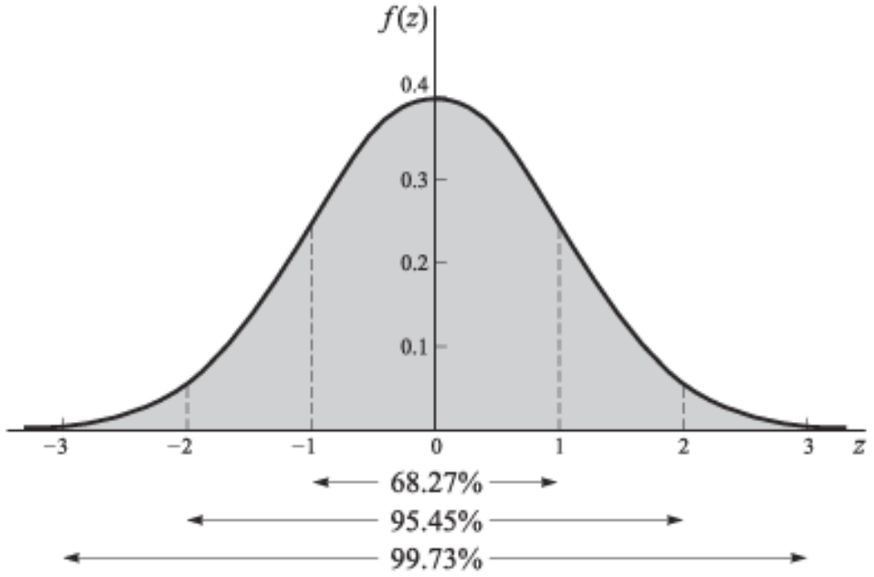
- We then call the Z the standard score and the distribution function $F(Z)$ can be related to error function (tabulated) $\mathrm{erf}(z)$

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{z} e^{-u^2} dv \qquad F(z) = \frac{1}{2}\left[1 + erf\left(\frac{z}{\sqrt{2}}\right)\right]$$

$$p(-1 < z < 1) = 0.6827$$

$$p(-2 < z < 2) = 0.9545$$

$$p(-3 < z < 3) = 0.9973$$



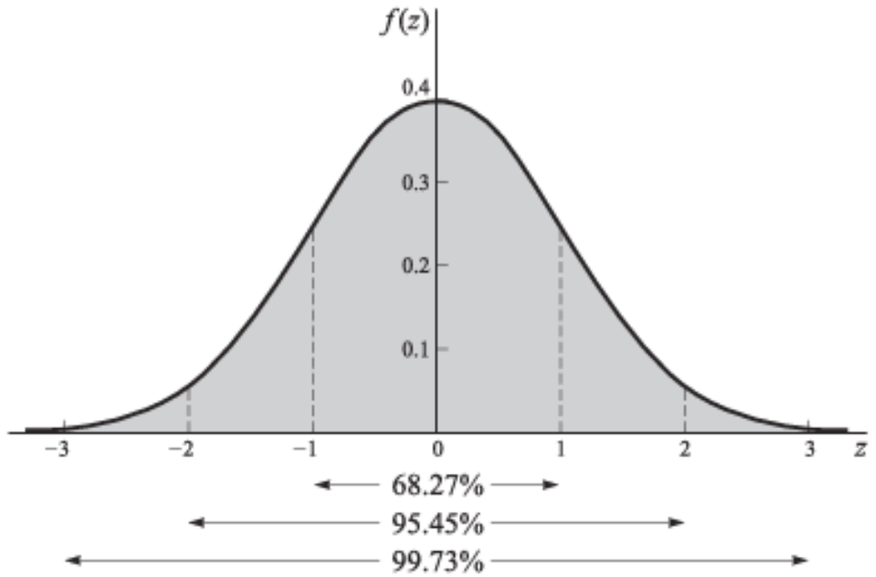
The Normal Distribution

- The Z score tells how many standard deviations the value *x* is above (to the right of) or below (to the left of) the mean, μ.

$$p(-1 < z < 1) = 0.6827$$

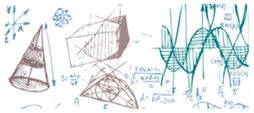
$$p(-2 < z < 2) = 0.9545$$

$$p(-3 < z < 3) = 0.9973$$



The empirical rule known as the **68-95-99.7 rule**.

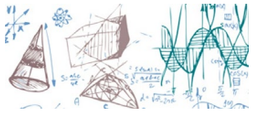
The Normal Distribution



Exercise (use statistical tables and calculator only):
A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

a) Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.
b) The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.
c) Find the 90^{th} percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.
d) The middle 40% of mandarin oranges from this farm are between _____ and _____.
e) Find the 16^{th} percentile and interpret it in a complete sentence.

The Normal Distribution



- The properties of the Gaussian distribution

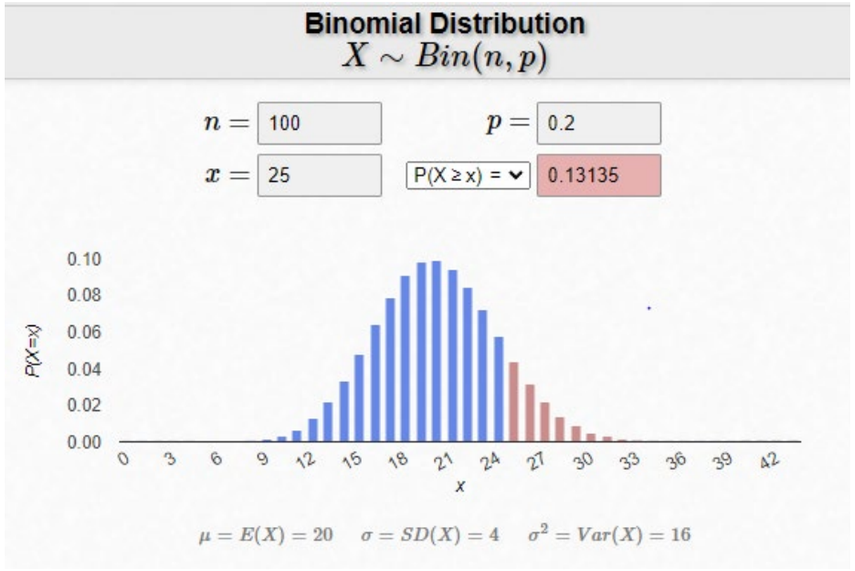
Mean	μ
Variance	σ^2
Standard deviation	σ
Coefficient of skewness	$\alpha_3 = 0$
Coefficient of kurtosis	$\alpha_4 = 3$

- Relation between binomial and normal distribution

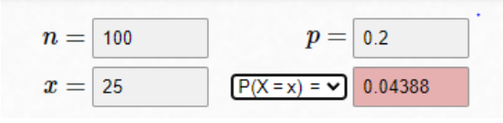
$$Z = \frac{X - np}{\sqrt{npq}}$$

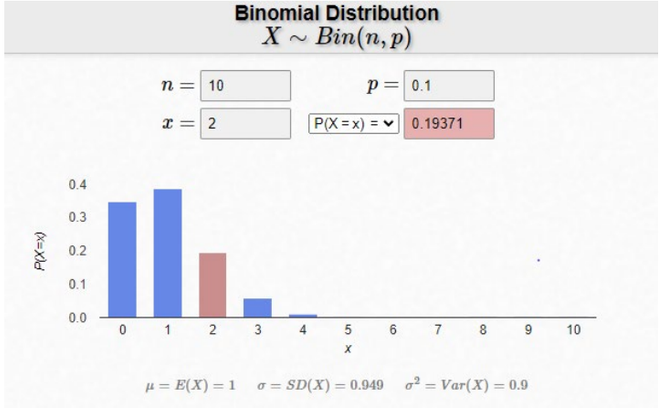
$$\lim_{n\to\infty} p\left(a \le \frac{X - np}{\sqrt{npq}} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-v^2/2}\, dv$$

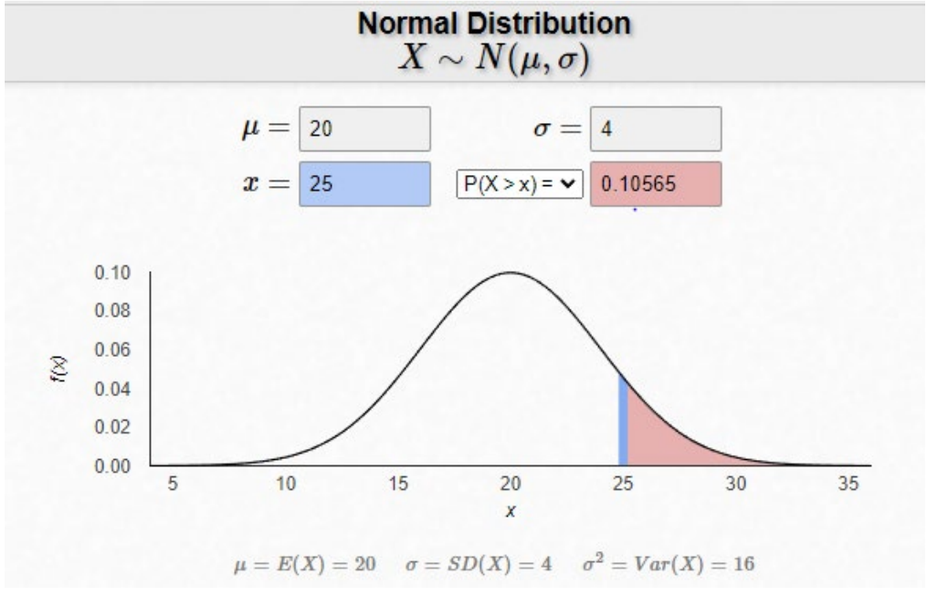
- We say, that the variable $\frac{X-np}{\sqrt{npq}}$ is **asymptotically normal**!

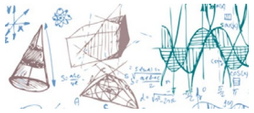

Binomial Distribution
$X \sim Bin(n, p)$
$n =$ 100
$p =$ 0.2
$x =$ 25
P(X ≥ x) =
0.13135
P(X=x)
0.10
0.08
0.06
0.04
0.02
0.00
0 3 6 9 12 15 18 21 24 27 30 33 36 39 42
x
$\mu = E(X) = 20$ $\sigma = SD(X) = 4$ $\sigma^2 = Var(X) = 16$


PROBLEMS
AHEAD


$n =$ 100
$p =$ 0.2
$x =$ 25
P(X = x) =
0.04388


Binomial Distribution
$X \sim Bin(n, p)$
$n =$ 10
$p =$ 0.1
$x =$ 2
P(X = x) =
0.19371
P(X=x)
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10
x
$\mu = E(X) = 1$ $\sigma = SD(X) = 0.949$ $\sigma^2 = Var(X) = 0.9$


Normal Distribution
$X \sim N(\mu, \sigma)$
$\mu =$ 20
$\sigma =$ 4
$x =$ 25
P(X > x) =
0.10565
f(x)
0.10
0.08
0.06
0.04
0.02
0.00
5 10 15 20 25 30 35
x
$\mu = E(X) = 20$ $\sigma = SD(X) = 4$ $\sigma^2 = Var(X) = 16$



More variables

- Especially important for the inference is operating with samples of measurements (data), and usually we have $\boldsymbol{n}$ of them
- Define the CDF for this case:

$$F(x_1, x_2, \dots, x_n) = P(X_1 < x_1, X_2 < x_2, \dots, X_n < x_n)$$

- And the PDF in this case:

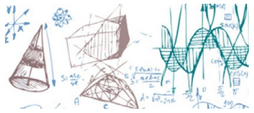
$$f(x_1, x_2, \dots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F(x_1, x_2, \dots, x_n)$$

- Any marginal PDF of RV x_k

$$g_k(x_k) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_{k-1} dx_{k+1} \cdots dx_n$$

- … and the mean value for x_k

$$E[x_k] = \mu_k = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_k f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n$$



More variables

- And the same using the marginal PDF of x_k

$$E[x_k] = \mu_k = \int_{-\infty}^{+\infty} x_k g_k(x_k) dx_k$$

NICE!

- Now, in this convention let's write out the mean, variance and covariance

$$E[x_i] = \boldsymbol{\mu_i}$$

$$E[(x_i - E[x_i])^2] = E[(x_i - \mu_i)^2] = \boldsymbol{\sigma_i^2}$$

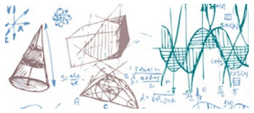
$$Cov(x_i, x_j) = E[(x_i - E[x_i])(x_j - E[x_j])] = E[(x_i - \mu_i)(x_j - \mu_j)] = \boldsymbol{c_{ij}}$$

- We can also introduce a pseudo-vector notation

$$\vec{x} = \{x_1, x_2, \dots, x_n\}, \vec{X} = \{X_1, X_2, \dots, X_n\}$$

$$f(\vec{x}) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F(\vec{x})$$

Nice and compact!



More variables

- We can also put all our variances and covariances in one structure that we call **covariance matrix**

$$\mathcal{C} = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix} \quad c_{ii} = \sigma_i^2, c_{ij} = c_{ji}$$

- Also, we can do similar thing („vectorisation”) for the means

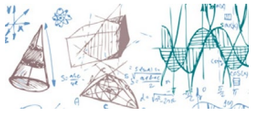
$$E[\vec{x}] = \vec{\mu} \qquad c_{ij} = E\left[(x_i - \mu_i)\left(x_j - \mu_j\right)\right]$$

$$\boldsymbol{\mathcal{C} = E[(\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^T]}$$

- The respective elements can be written explicitly (take 2 RV)

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \vec{x}^T = (x_1, x_2) \qquad \vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \vec{\mu}^T = (\mu_1, \mu_2)$$

More variables



- Now make the complete calculations

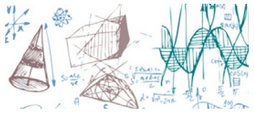
$$(\vec{x}-\vec{\mu})^T = (x_1-\mu_1, x_2-\mu_2), \vec{x}-\vec{\mu} = \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}$$

$$E[(\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^T] = \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}(x_1-\mu_1, x_2-\mu_2) =$$

$$= \begin{pmatrix} (x_1-\mu_1)(x_1-\mu_1) & (x_1-\mu_1)(x_2-\mu_2) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)(x_2-\mu_2) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & c_{12} \\ c_{21} & \sigma_2^2 \end{pmatrix}$$

- We can use our new and compact notation to derive one super important rule in statistics: **error propagation formula**
- It combines variable change and multivariate functions of RV
- Interested already? Go to the next page!

N-dim Gaussian



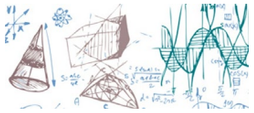
- Soon, we discuss the central limit theorem that states that the sum on n independent C.R.V. with finite means μ_i and variances σ_i becomes a Gaussian R.V. with mean $\mu = \sum_i \mu_i$ and variance $\sigma = \sum_i \sigma_i$ for n being large.
- Note, this is the justification that the random measurement errors should follow the Gaussian distribution!
- The N-dim generalisation of the Gaussian formula is given by:

$$f(\vec{x};\vec{\mu},\mathcal{C}) = \frac{1}{(2\pi)^{N/2}|\mathcal{C}|^{1/2}} exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T\mathcal{C}^{-1}(\vec{x}-\vec{\mu})\right]$$

$|\mathcal{C}|$ - determinant of covariance matrix

$$E[x_i] = \mu_i, V[x_i] = \sigma_i = \mathcal{C}_{ii}, \mathrm{cov}[x_i, x_j] = \mathcal{C}_{ij}$$

N-dim Gaussian



- For the pedagogical reasons let's write explicitly the 2-dim case of Gaussian distribution with

$$\rho = \left. \mathrm{cov}[x_i, x_j] \right/ \sigma_1 \sigma_2$$

$$f(x_1, x_2, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho}} \times$$

$$\times exp\left[-\frac{1}{2(1-\rho^2)} \left(\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) \right) \right]$$

- Which for the independent variables reduces to

$$f(x_1, x_2, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho = 0) = \frac{1}{2\pi\sigma_1\sigma_2} \times exp\left[-\frac{1}{2} \left(\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 \right) \right]$$