Continuous probability distributions; Gaussian distributions; Intro to CLT; Intro to Gauss Mixture Models

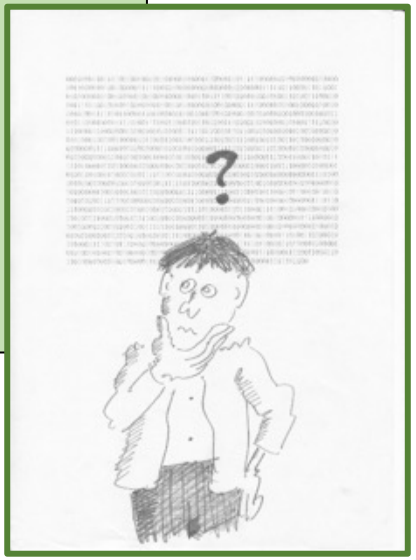
Statistics and data analysis

Ben Galili

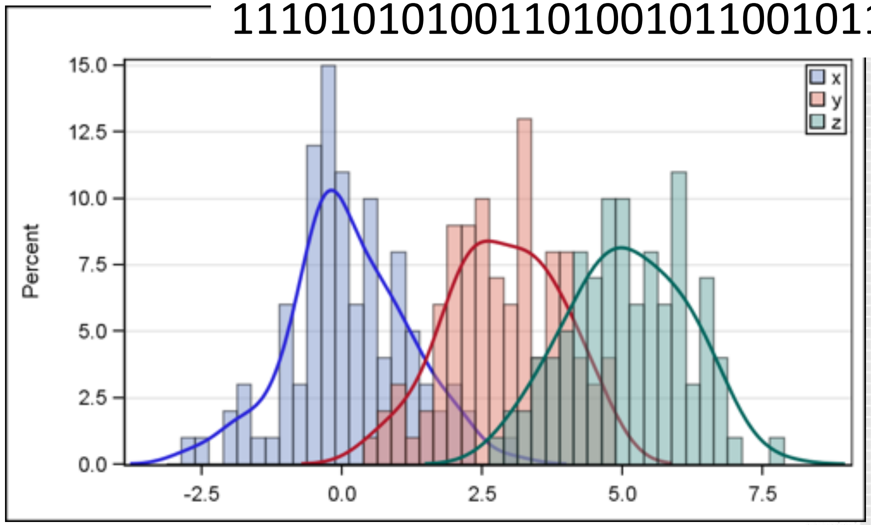
Leon Anavy

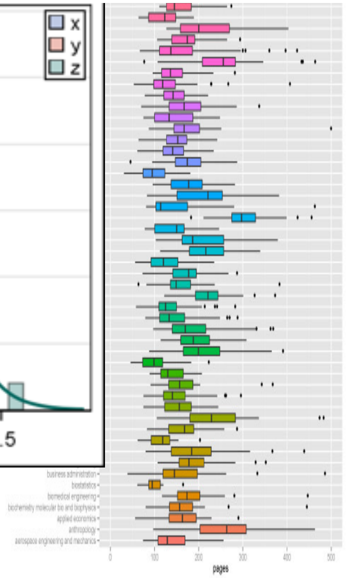
Zohar Yakhini

IDC, Herzeliya



0010011101010100101010100100100010
1010100010101111101011010011001001
1110101010011010010110010110110010





Continuous probability distributions

- A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.
- It is not possible to talk about the probability of the random variable assuming a particular value. It will be 0.
- Instead, we talk about the probability of the random variable assuming a value within a given interval.
- The distribution is defined by a probability density function $p(x)$
- The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the probability density function between x_1 and x_2.

IDC HERZLIYA

CDF of continuous random variables

Consider a density function f : $f \geq 0$, $\int_{-\infty}^{\infty} f(t)dt = 1$
The cumulative distribution function (CDF) of a continuous rv is defined as

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

and for any interval $I = [l, r]$ we can now have two expressions for $P(I)$ = P(X assumes a value in the interval):

$$P(I) = \int_{l}^{r} f(t)dt$$

and

$$P(I) = F(r) - F(l)$$

IDC HERZLIYA

Expected values

$$E(X) = \int_{-\infty}^{\infty} x\, p(x)\, dx = \int_{\Omega} X(\omega)\, d\mu$$

$$E(X) = \sum_{x} x\, p(x) = \sum_{\omega \epsilon \Omega} X(\omega) P(\omega)$$

IDC HERZLIYA

Interesting fact

For a positively valued random variable X we have:

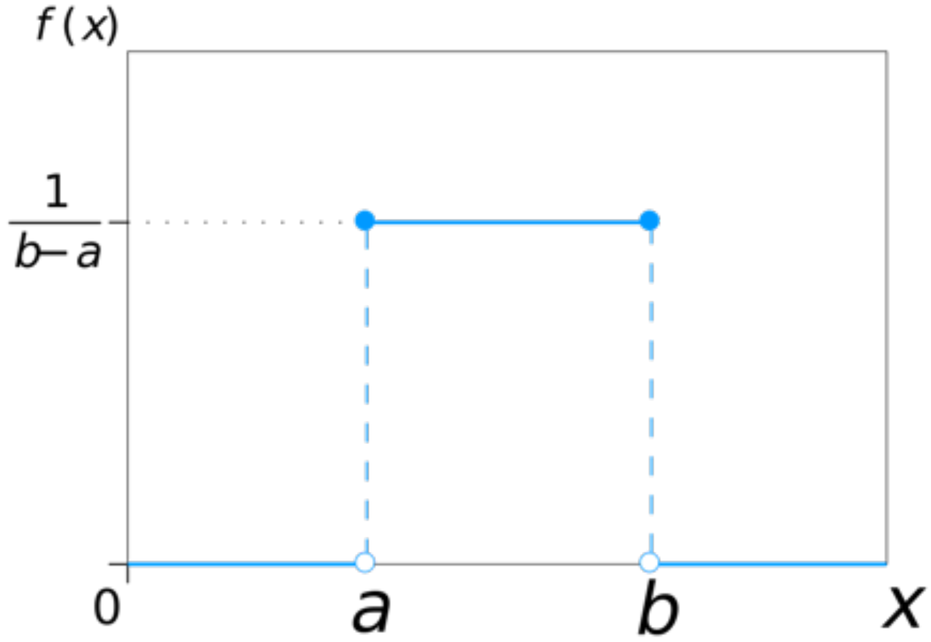
$$E(X) = \int_{-\infty}^{\infty} \left(1 - F(x)\right) dx$$

In the discrete, integer valued, case:

$$E(X) = \sum_{n=1}^{\infty} P(X \geq n)$$

IDC HERZLIYA

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

$$Var(x) = (b - a)^2/12$$

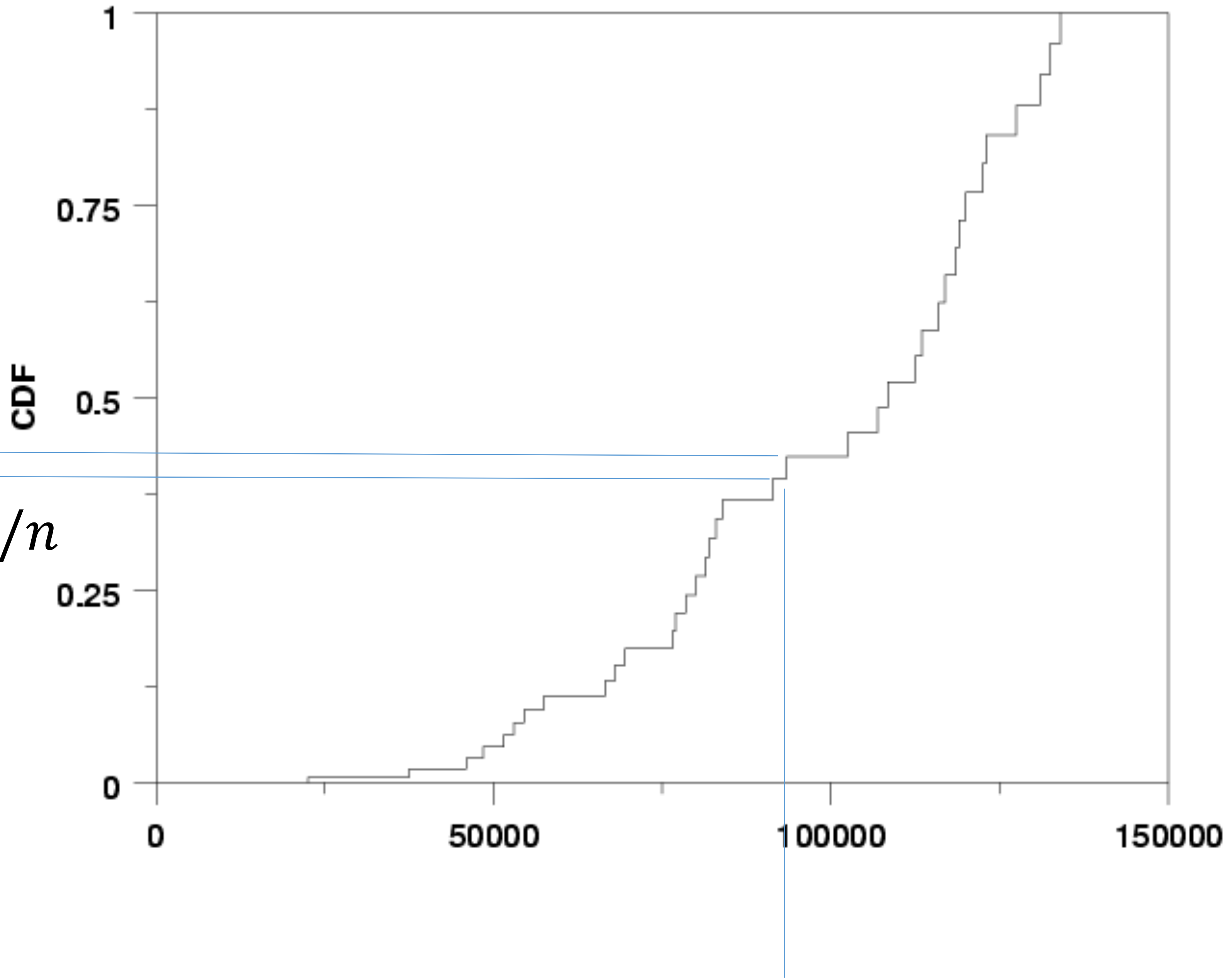
where: a = smallest value the variable can assume

b = largest value the variable can assume

Empirical CDF

$$y_i(U) = i/n$$

$$y_i(L) = (i-1)/n$$



x_i = the value of the i-th smallest value observed in the data

Independent continuous random variables

Two-dimensional density functions can be defined on $\mathbb{R}^2$, leading to probability distributions on $\mathbb{R}^2$.

$$P\big((X,Y)(\omega) \in A\big) = \int_A f(x,y)dxdy$$

For example. Let (X,Y) uniform in $\mathrm{R} = [0,2]\times[5,6]$.
What is the density function?

In analogy and extension to the discrete case, the random variables X and Y , with respective pdfs a and b are independent if

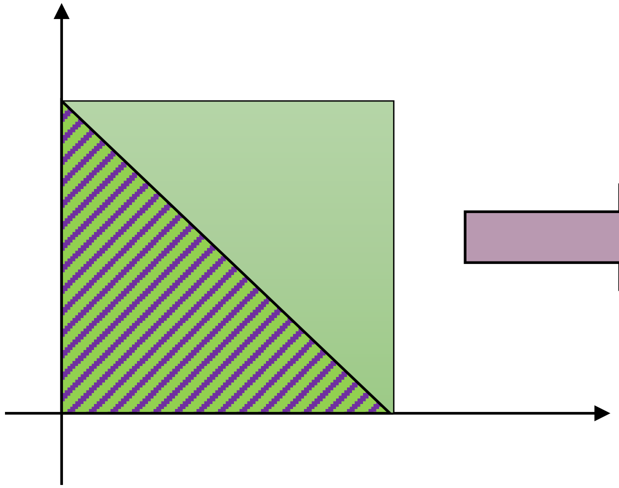
$$\forall(x,y),\ f(x,y) = a(x)b(y)$$

IDC HERZLIYA

Example – bivariate distributions

Consider X and Y independent with pdfs $a(x) = b(x) = 2x$ for $0 \leq x \leq 1$ and 0 otherwise.
What is

$$p = P(X + Y \leq 1)\ ?$$

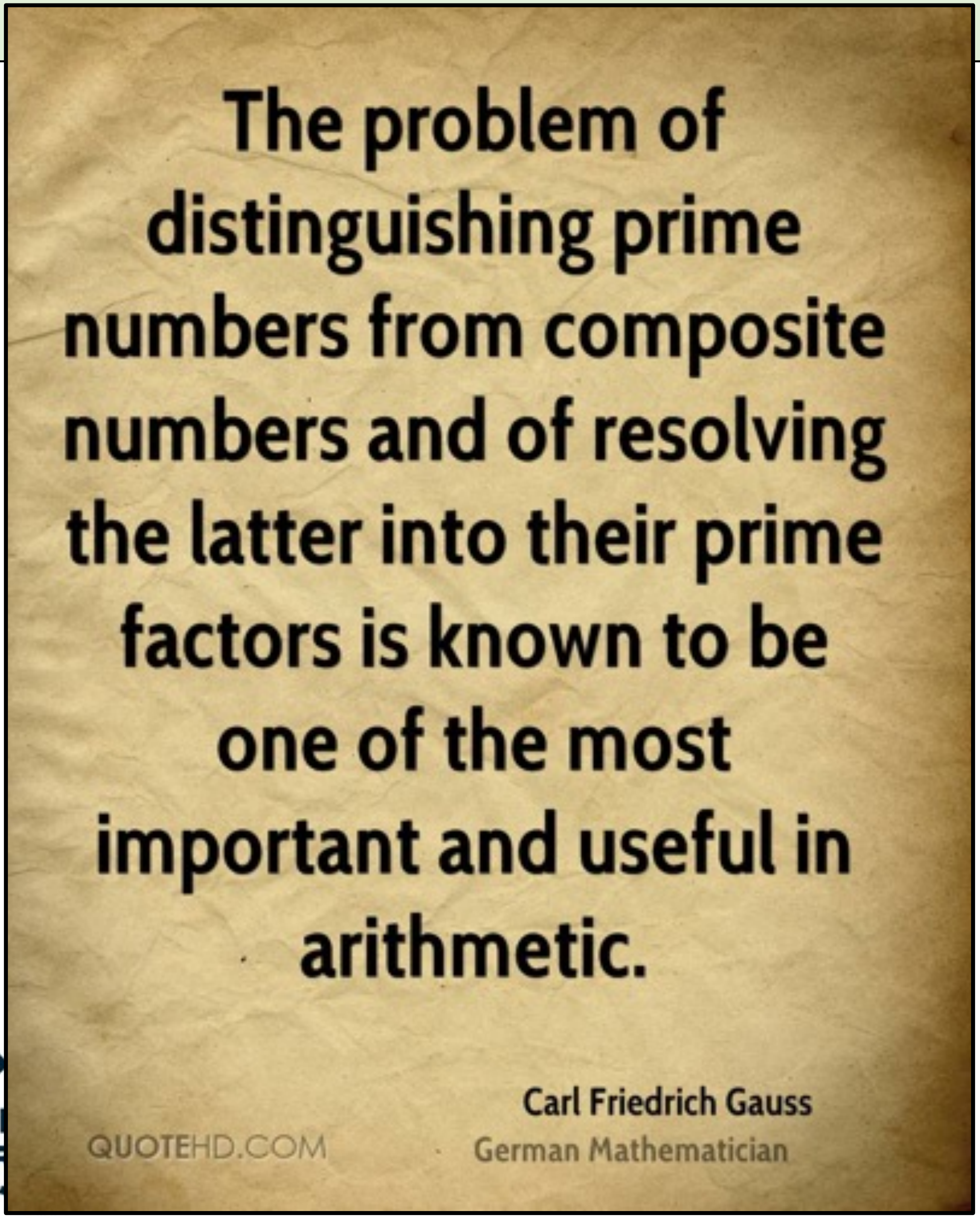


Region T

$$p = \int_T 4xy\ dx\ dy = 4\int_0^1 x\left(\int_0^{1-x} y dy\right) dx = \frac{1}{6}$$

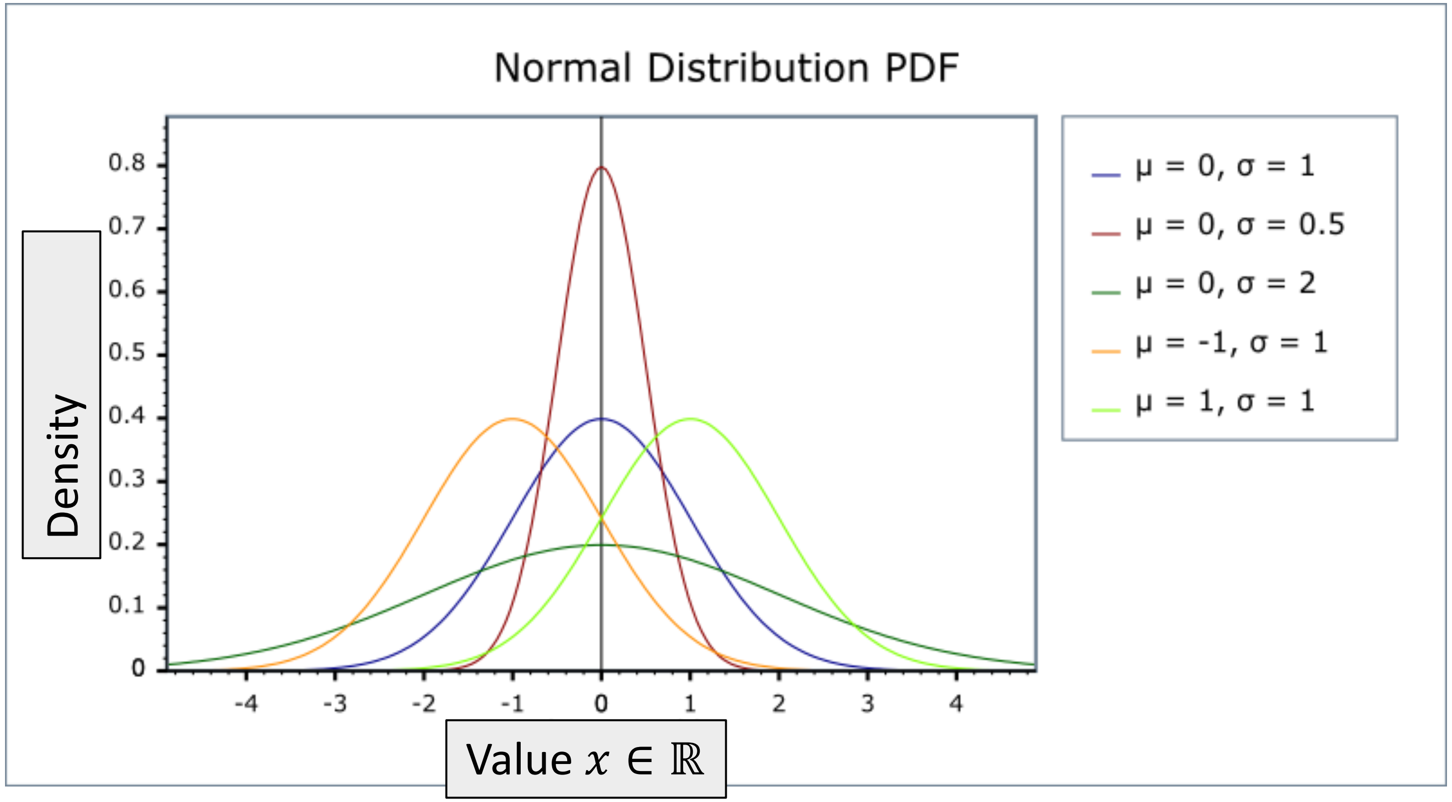
Adapted from De Groot

Gaussian distributions and the Central Limit Theorem



Carl Freidreich Gauss
1777-1855, Germany

Gaussian or Normal Probability Distributions



The shape of the Gaussian, or Laplace-Gauss, or normal, curve is often referred to as a bell-shaped curve.
The highest point on the normal curve is at the mean, which is also the median (and mode) of the distribution.
The normal curve is symmetric.
The standard deviation determines the width of the curve.
The total area under the curve is 1.
Probabilities for the normal random variable are given by areas under the curve.

IDC HERZLIYA

The normal density function

Density functions for Gaussian r.vs:

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

We then say that the r.v X is normally (univariate Gaussian) distributed with mean μ and standard deviation σ.

We write $X \sim N(\mu, \sigma)$

A random variable that has a normal distribution with

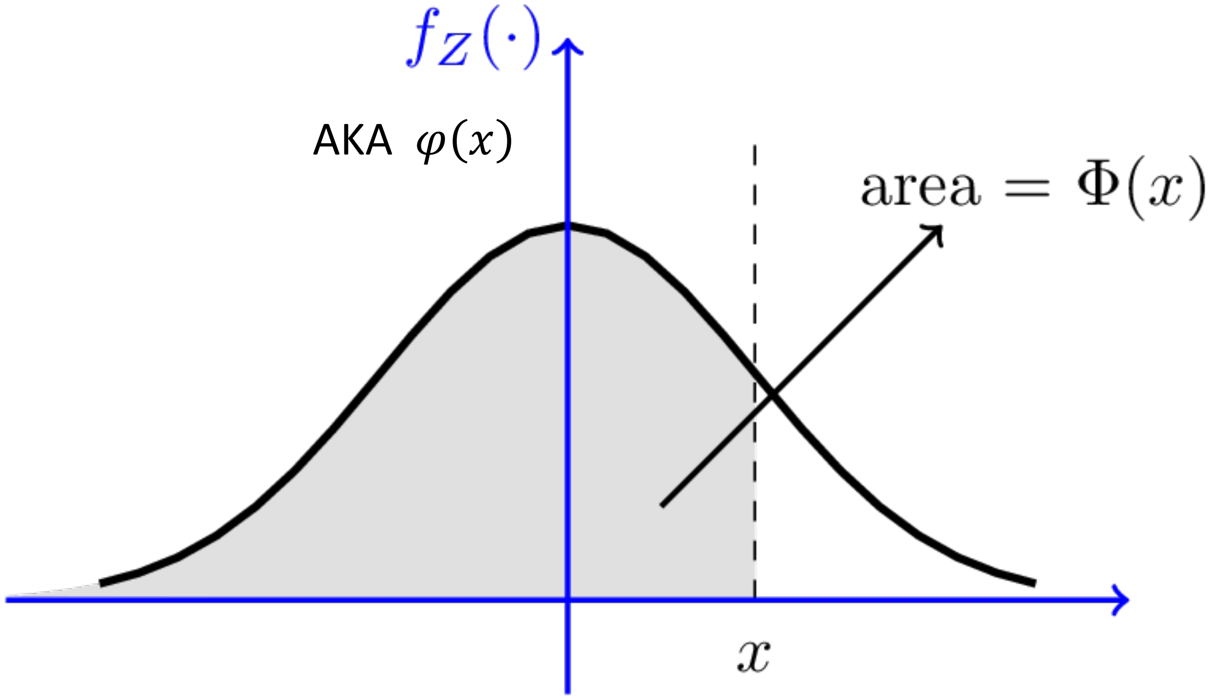
$\mu = 0$ and
$\sigma = 1$

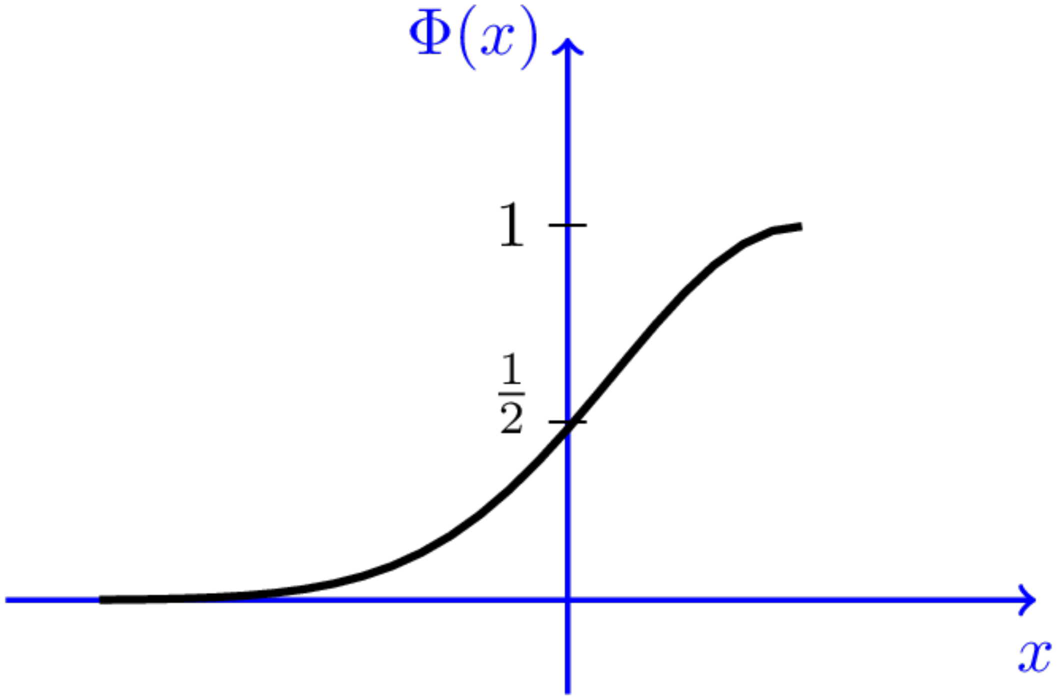
is called Standard Normal.

The density function then becomes:

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

The CDF of a standard normal is often called Φ





$\varphi(x) = \varphi(-x)$

$\Phi(x) = 1 - \Phi(-x)$

Randomistan electric bills

Historical data shows that monthly electric bills for Stochastic Heights households are normally distributed with a mean of 225 RCU and a standard deviation of 55 RCU.
To compensate low income residents for Corona financial damages and to encourage energy efficiency in future years the mayor decided that any household that consumed less than 60 RCU will be exempt from their 2021 payment which is due in January 2022.
SH has 18K households.

- How many monthly bills do we expect to be exempt from payment?
- Can you estimate the cost of this policy to the city budget?

IDC HERZLIYA

A useful fact

If $X \sim N(\mu, \sigma)$

then

$Z = (X - \mu)/\sigma$

is a standard normal random variable

You can think of Z as measuring, for every instance of X drawn, the distance of the obtained value, from the expected value, in units of standard deviations

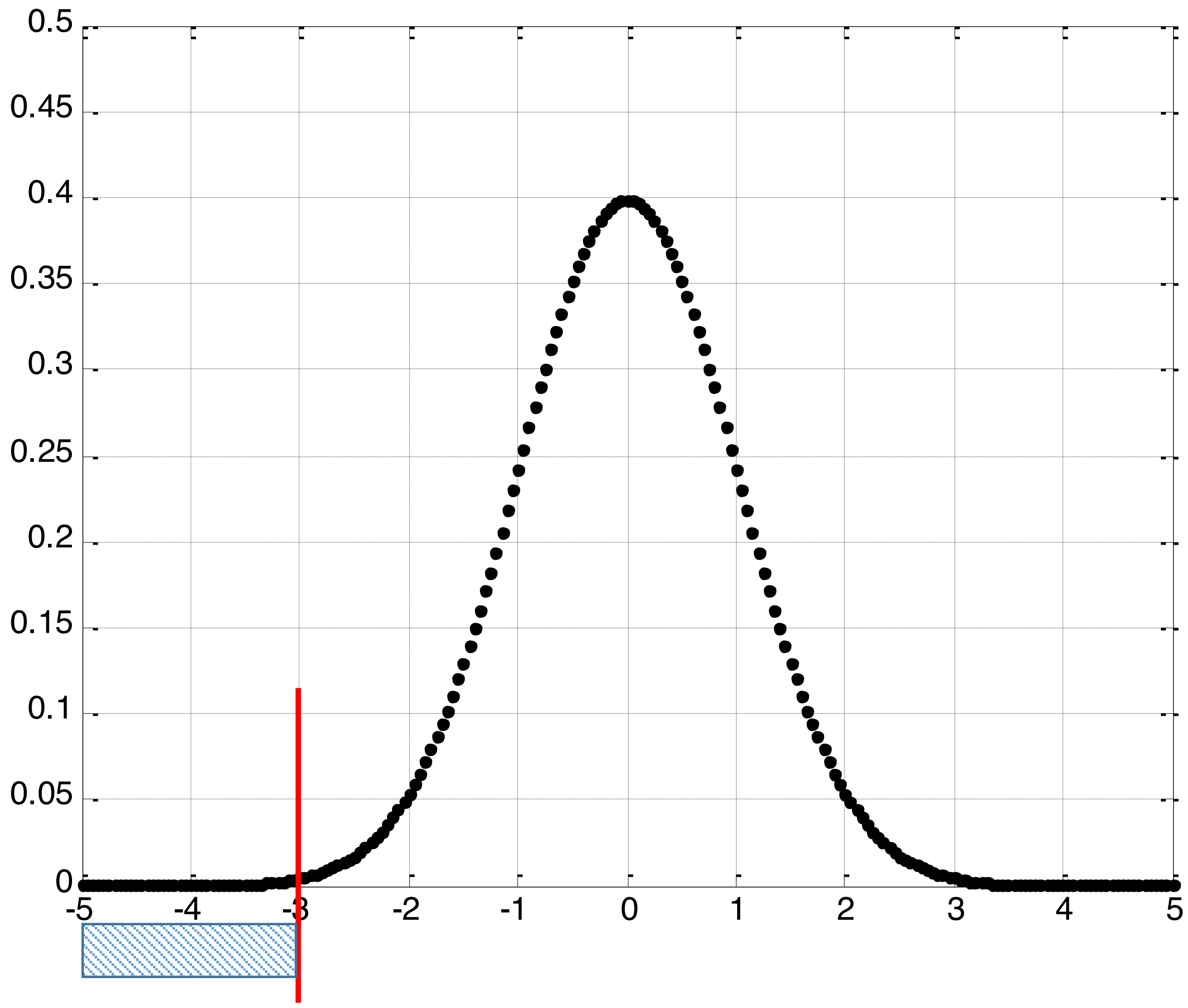
Back to SH…

$$z = (x - \mu)/\sigma$$

$= (60 - 225)/55$

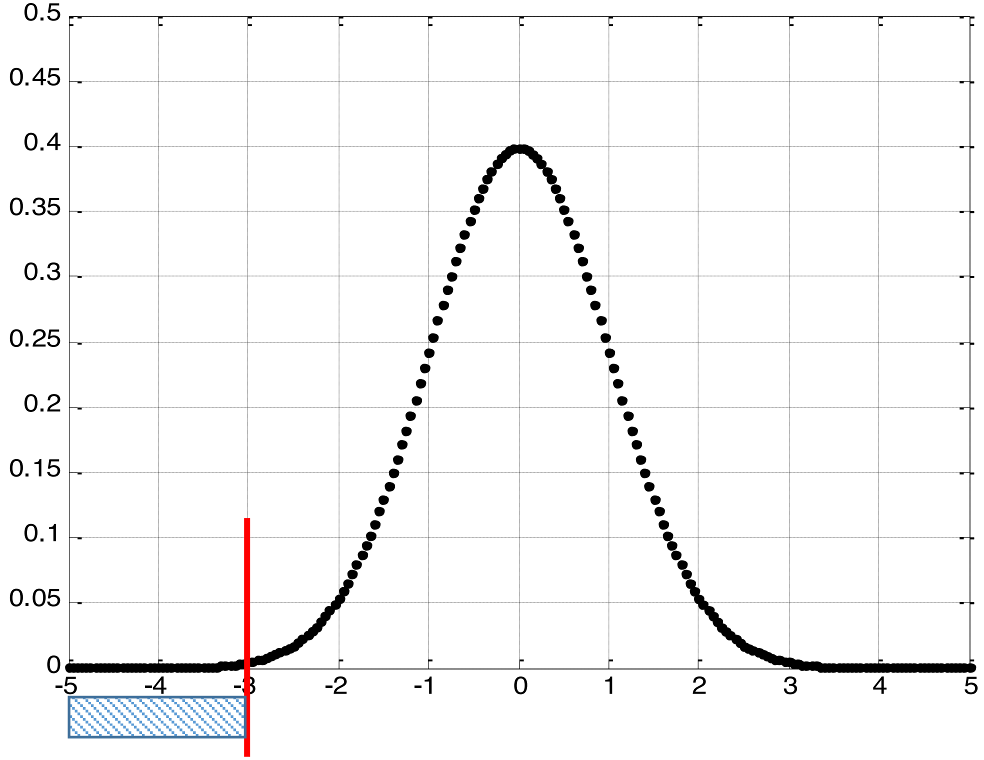
$= -3$

So – we want the area under the graph, to the left of the red line.
We will then multiply it by the total expected number of bills in 2020, namely 12*18K.



IDC HERZLIYA

So…



$$P(\text{Exempt}) = \text{normcdf}(z = (x - \mu)/\sigma)$$
$$= \text{normcdf}(-3) = 0.0013$$

and the mayor should expect ~23.5 exempt bills per month.

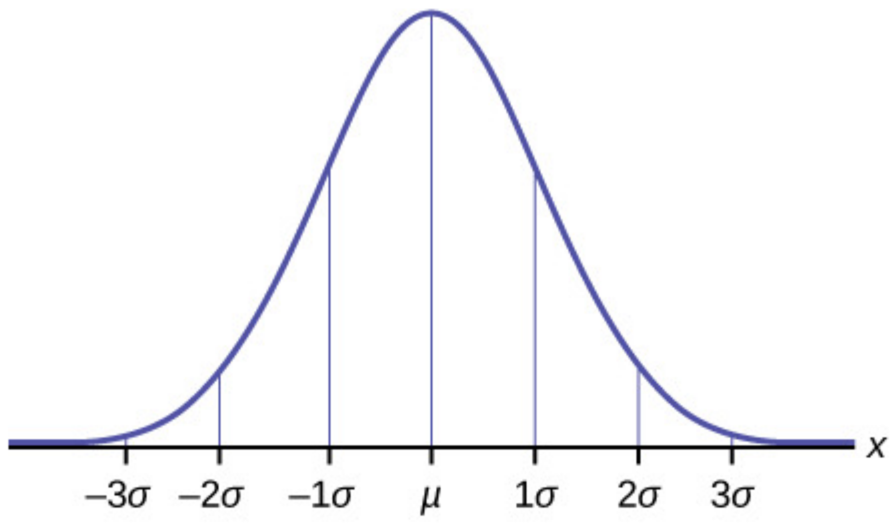
Budget planning for the mayor

- We can therefore bound the expected total cost of the mayor’s policy by 12*24*60 RSU.
- Can we compute a better estimate of this expected cost?

IDC HERZLIYA

Within 1 sigma

What is the probability that an observation from a standard normal distribution will fall in the interval [-1 1]?




```
p = normcdf([-1 1]);
p(2)-p(1)
ans = 0.6827
```

p(2) = normcdf(1)

p(1) = normcdf(-1)

More generally, about 68% of the observations from a normal distribution fall within one standard deviation, σ, of the mean, μ.

Chebyshev inequality

For any random variable X that has a finite variance we have:

$$\forall b \quad P(|X - E(X)| \geq b\sigma(X)) \leq \frac{1}{b^2}$$

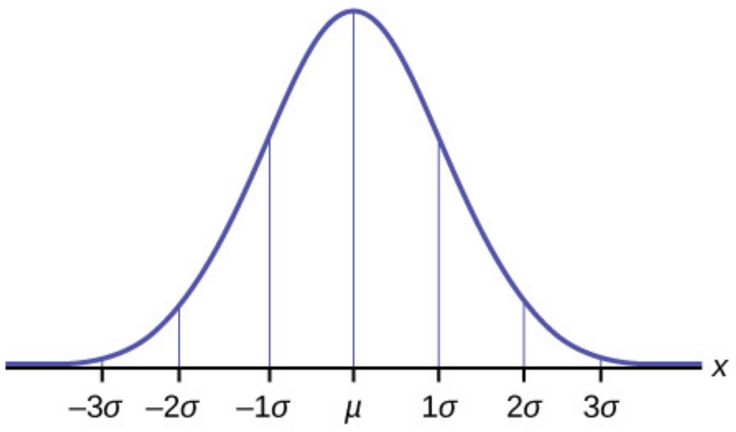
IDC HERZLIYA

Chebichev is not tight for Gaussian distributions …

- Chebyshev's Theorem.
  Let Y be a random variable with mean μ and std σ. Then:
  $P(\mu-b\sigma \leq Y \leq \mu+b\sigma) \geq 1-(1/b^2)$ for $b > 0$

  - b=1: $P(\mu-1\sigma \leq Y \leq \mu+1\sigma) \geq 1-(1/1^2) = 0$ (trivial result – compare to 0.68 above)
  - b=2: $P(\mu-2\sigma \leq Y \leq \mu+2\sigma) \geq 1-(1/2^2) = ¾$
  - b=3: $P(\mu-3\sigma \leq Y \leq \mu+3\sigma) \geq 1-(1/3^2) = 8/9$

- For Gaussian distributions:
  - $P(\mu-1\sigma \leq Y \leq \mu+1\sigma) \approx 0.68$
  - $P(\mu-2\sigma \leq Y \leq \mu+2\sigma) \approx 0.95$
  - $P(\mu-3\sigma \leq Y \leq \mu+3\sigma) \approx 0.997$



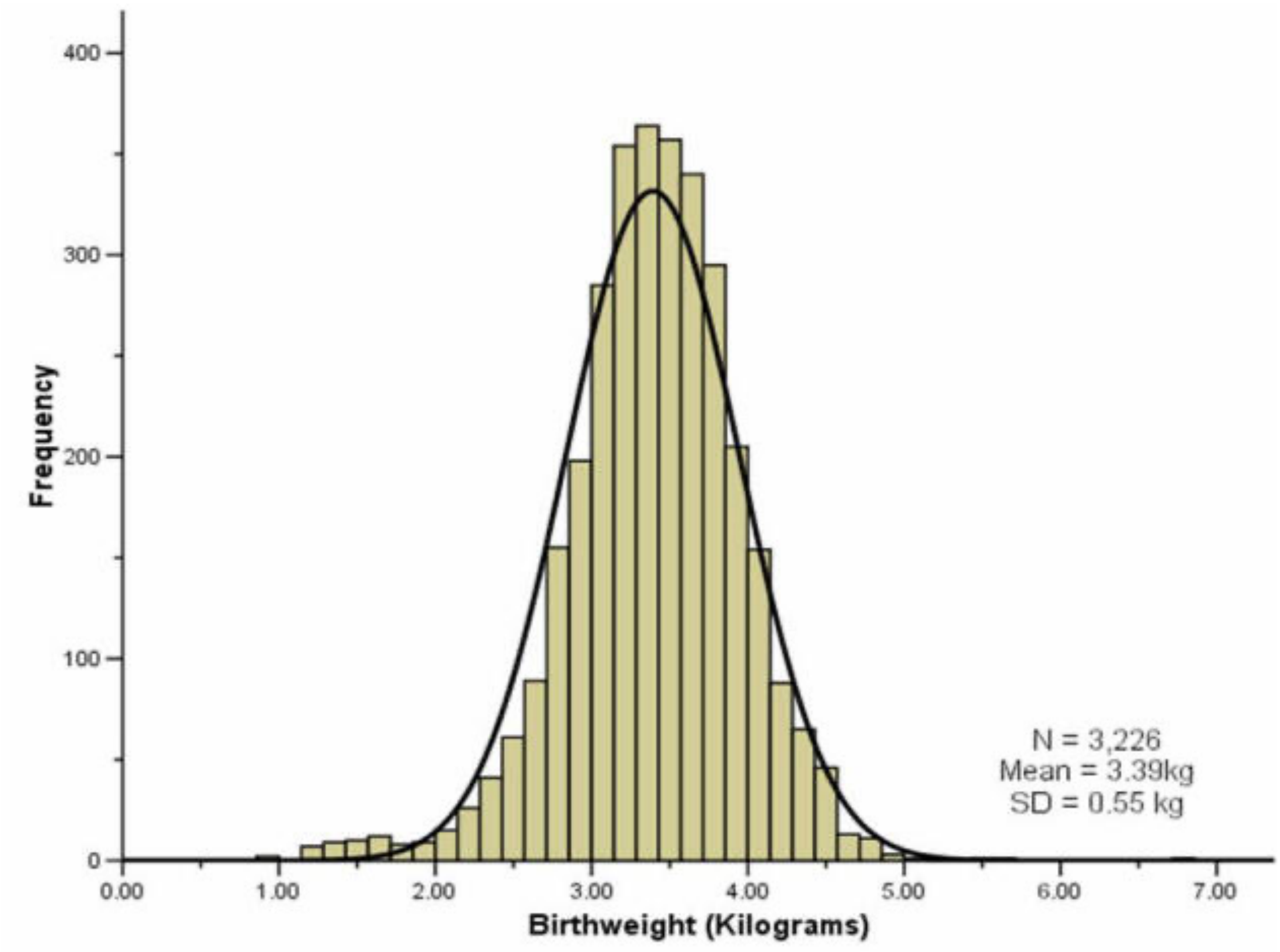
IDC HERZLIYA

Normal distribution – a strong assumption??

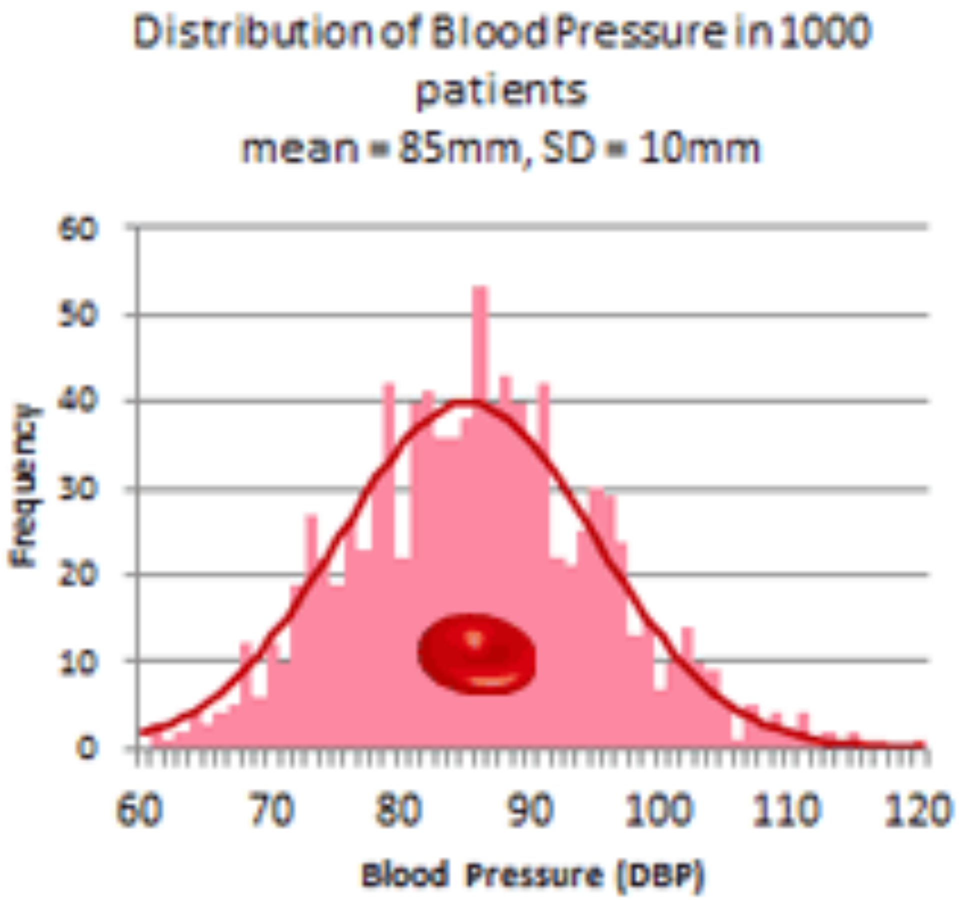
- A foundational assumption in the electricity bills story was that they determined/assumed a normal distribution for the monthly bills.
- It enabled us to use the normal distribution characteristics.
- When can such an assumption be made?
  How ubiquitous is the normal distribution?
- What can be done in more general cases?

IDC HERZLIYA

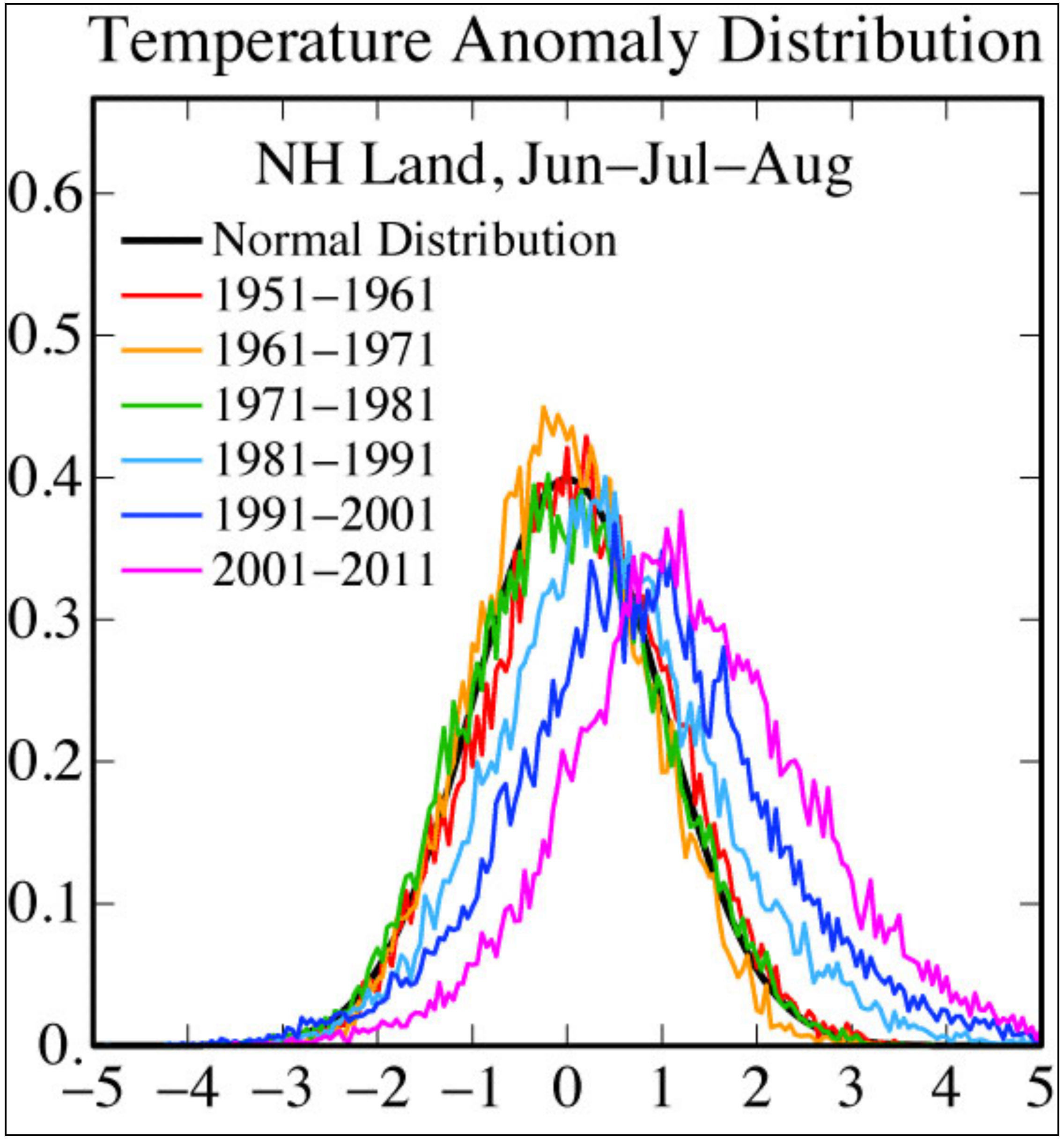
Distribution of birth weight in 3,226 newborn babies (data from O' Cathain et al 2002)



IDC HERZLIYA



Distribution of Blood Pressure in 1000 patients
mean = 85mm, SD = 10mm
Frequency
60
50
40
30
20
10
0
60
70
80
90
100
110
120
Blood Pressure (DBP)



Deviation of daily summer temp in NH, from the 1951-1980 mean.

Source: NASA/GISS

IDC HERZLIYA

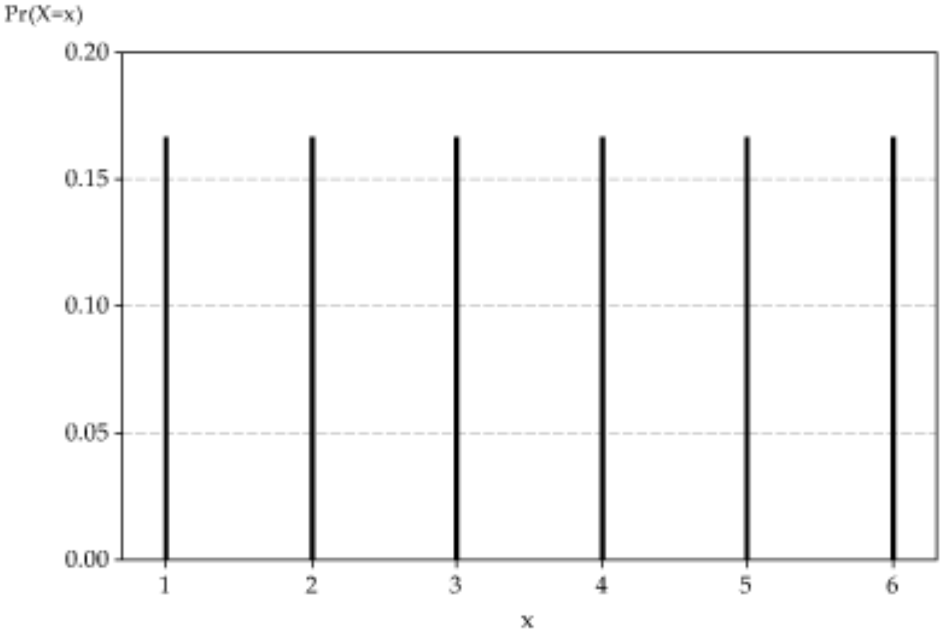
Bad news - not all is really normal …

IDC HERZLIYA

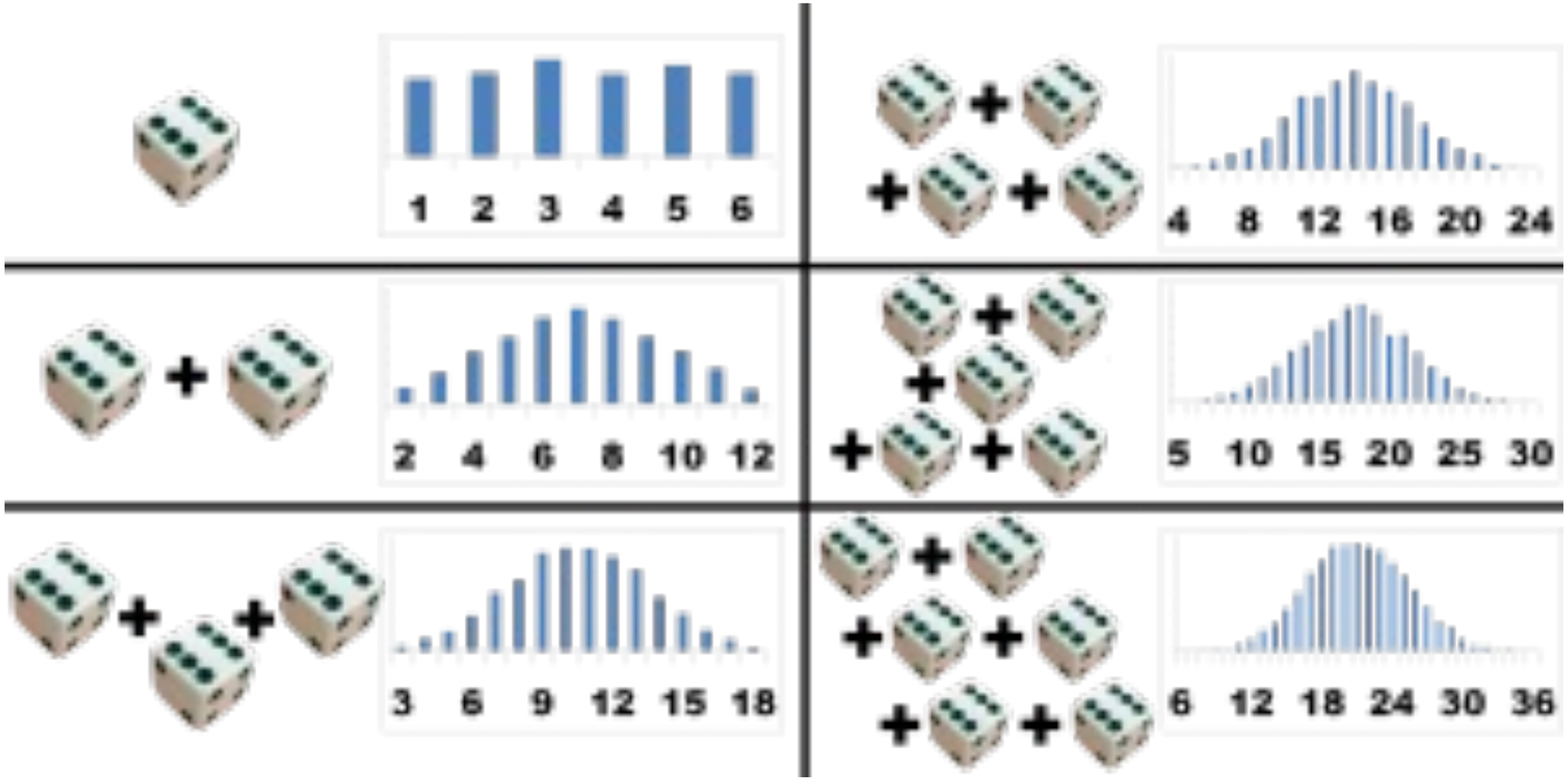
The good news:

… but most distributions ***average*** to be normal:
the Central Limit Theorem

IDC HERZLIYA

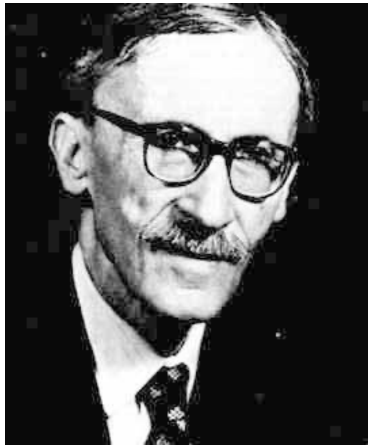


Pr(X=x)
0.20
0.15
0.10
0.05
0.00
1
2
3
4
5
6
x



IDC HERZLIYA

The Central Limit Theorem
(Lindeberg and Levy 1920)



Let $X_1, X_2, X_3, \ldots, X_n$ be random variables all sampled **independently** from **the same distribution** with mean μ and (finite non 0) variance σ^2.
Let $\overline{X_n}$ be the average of $X_1, X_2, X_3, \ldots, X_n$.(*)
Then for any fixed number x we have

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}}{\sigma}\left(\overline{X_n} - \mu\right) \leq x\right) = \Phi(x)$$

where $\Phi(x)$ is the standard normal density function.

$$(*) \qquad \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Paul Levy, France, 1886-1971
JW Lindeberg, Finland, 1876-1932

Earlier version (for binomial) proven by De Moivre (London, 18th Century). Generalized versions proven later and are active research

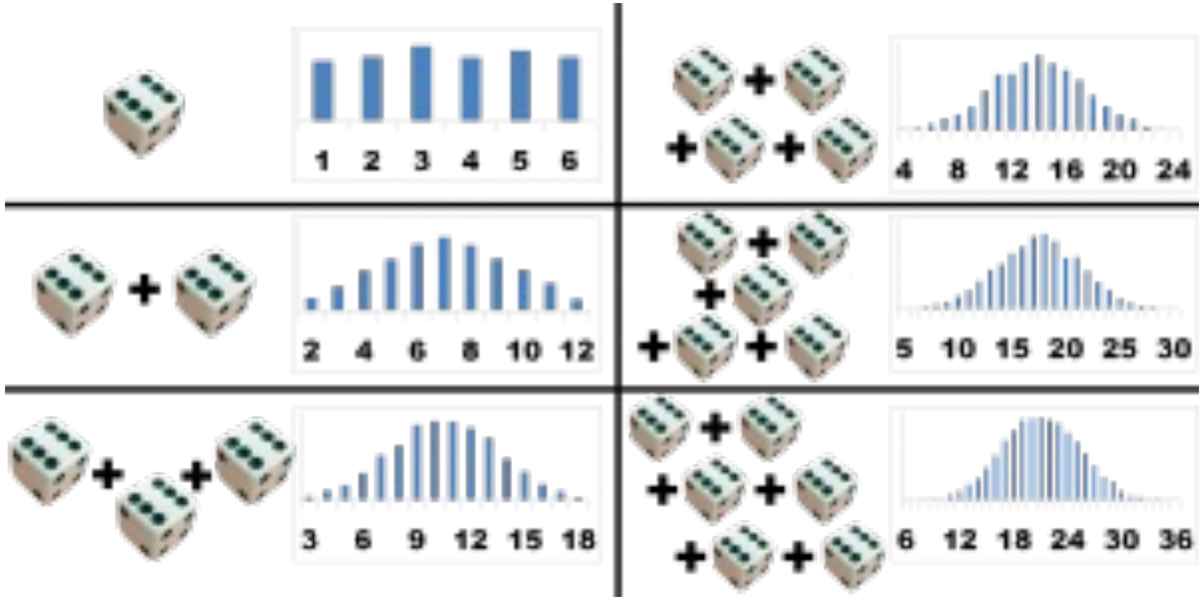
IDC HERZLIYA

In other words ….

For (**almost**) any distribution, if we sample it (**sufficiently**) many times, and then average, then our distance from the mean, properly scaled, is (**very close to**) standard normally distributed.

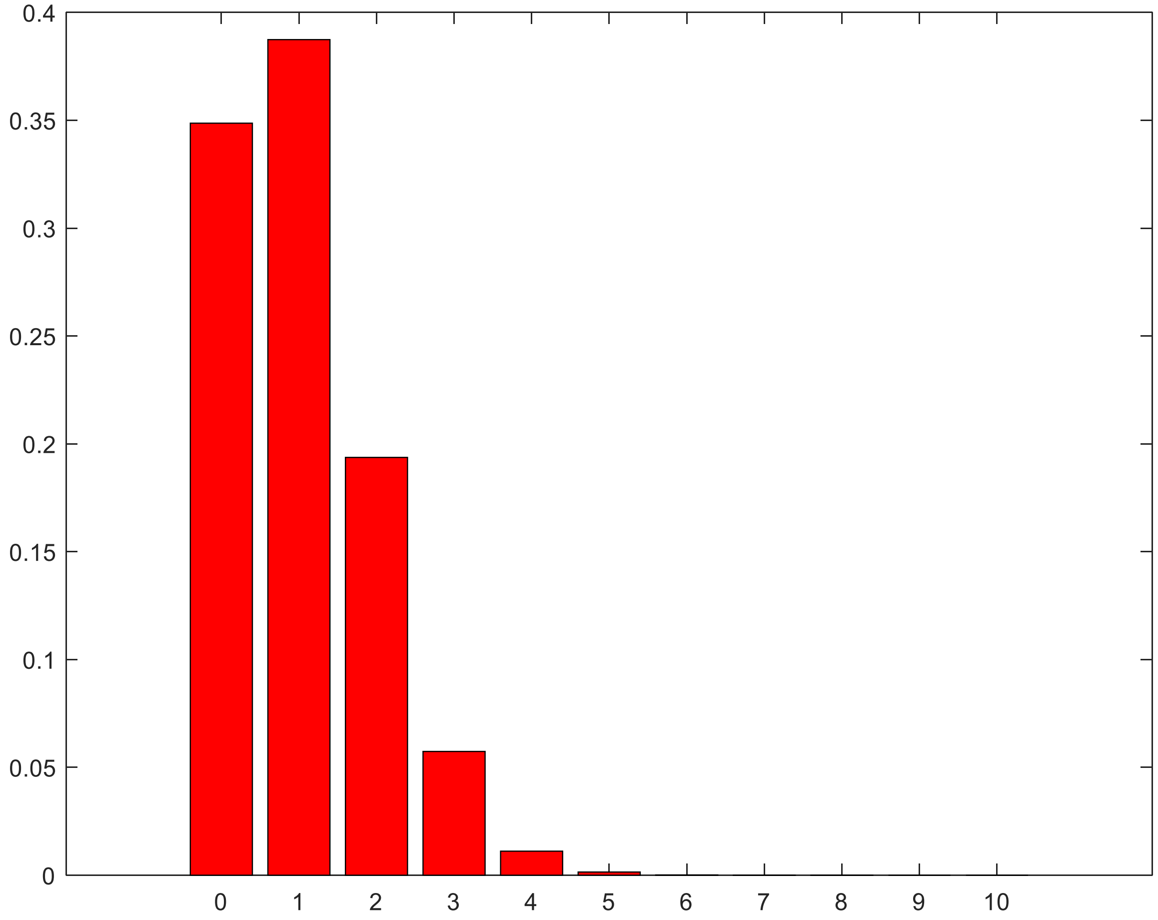
Alternatively - $\overline{X_n}$ is approximately normally distributed with mean μ and variance σ^2/n.

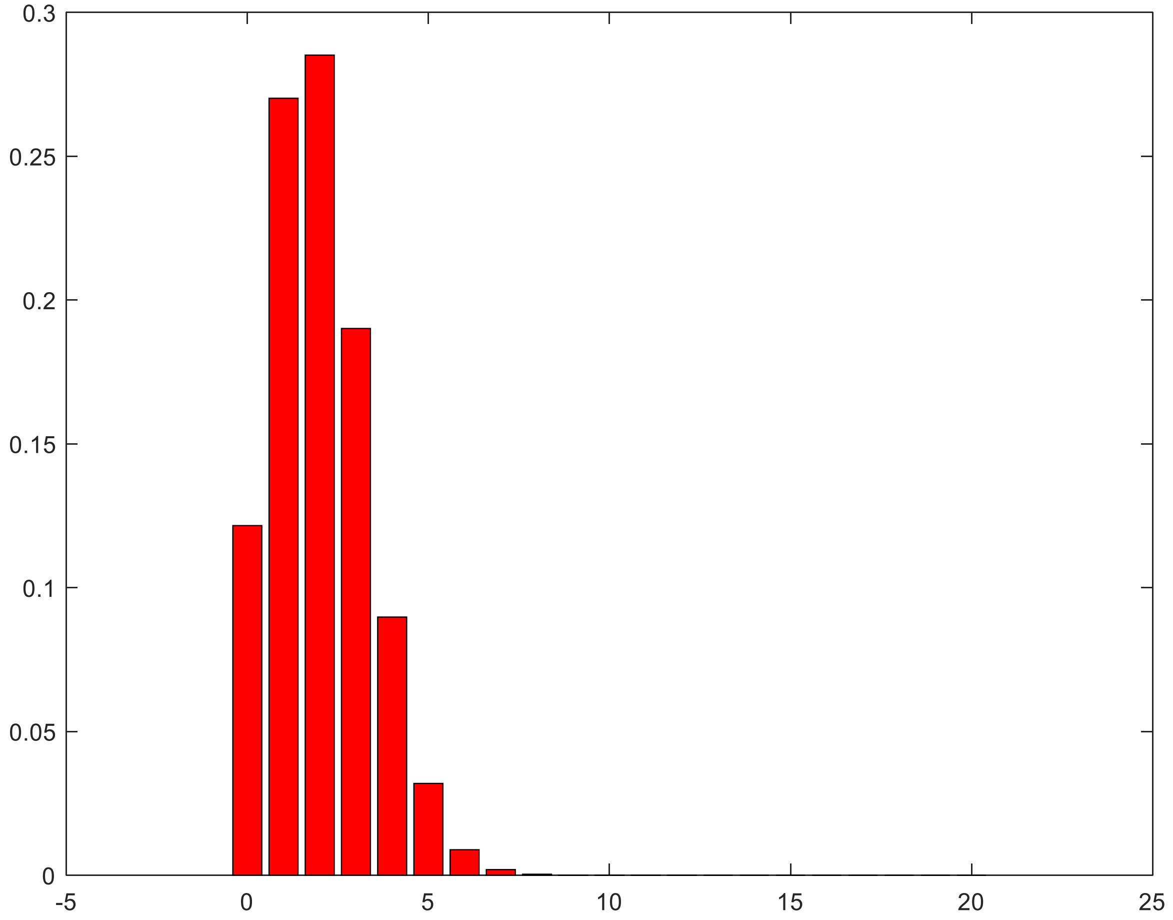
Or: S_n is approximately normally distributed with mean $n\mu$ and variance $n\sigma^2$.



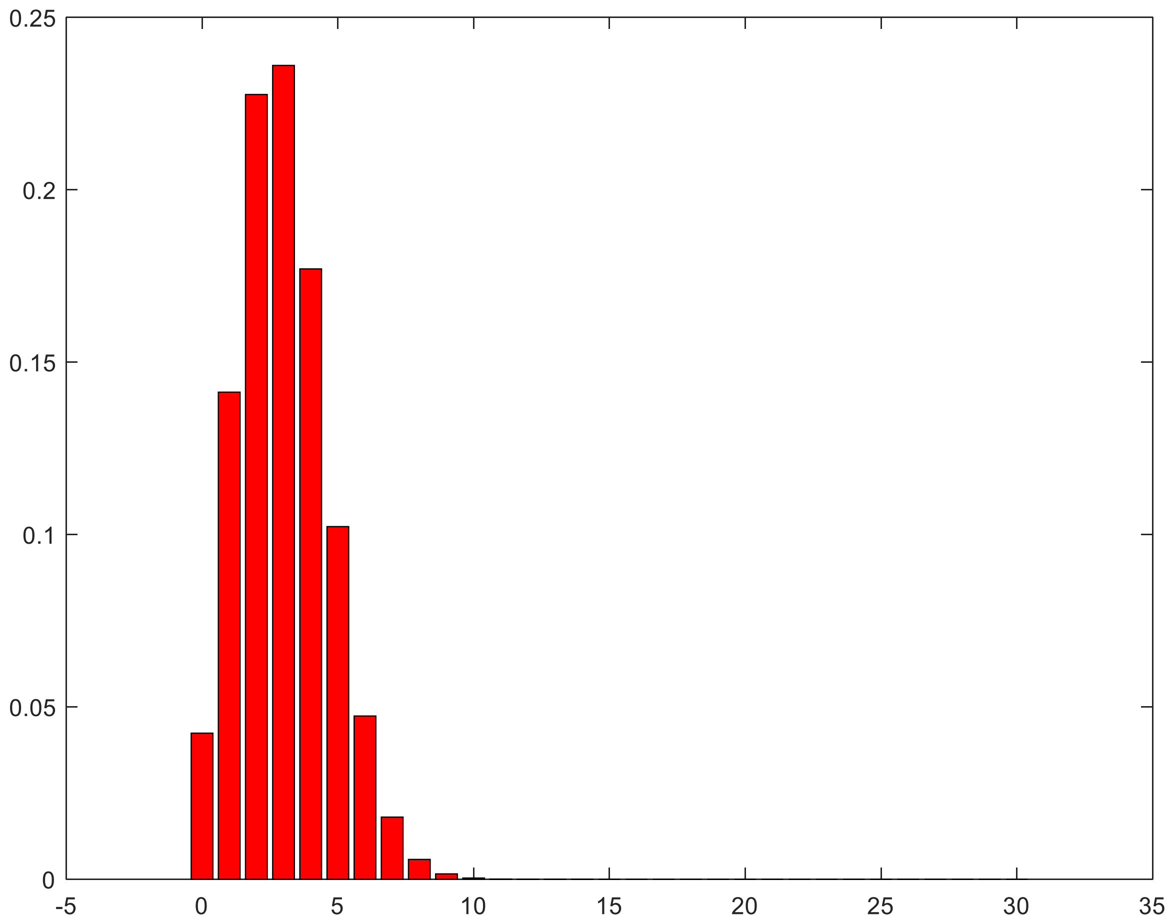
IDC HERZLIYA

Binomials w p=0.1

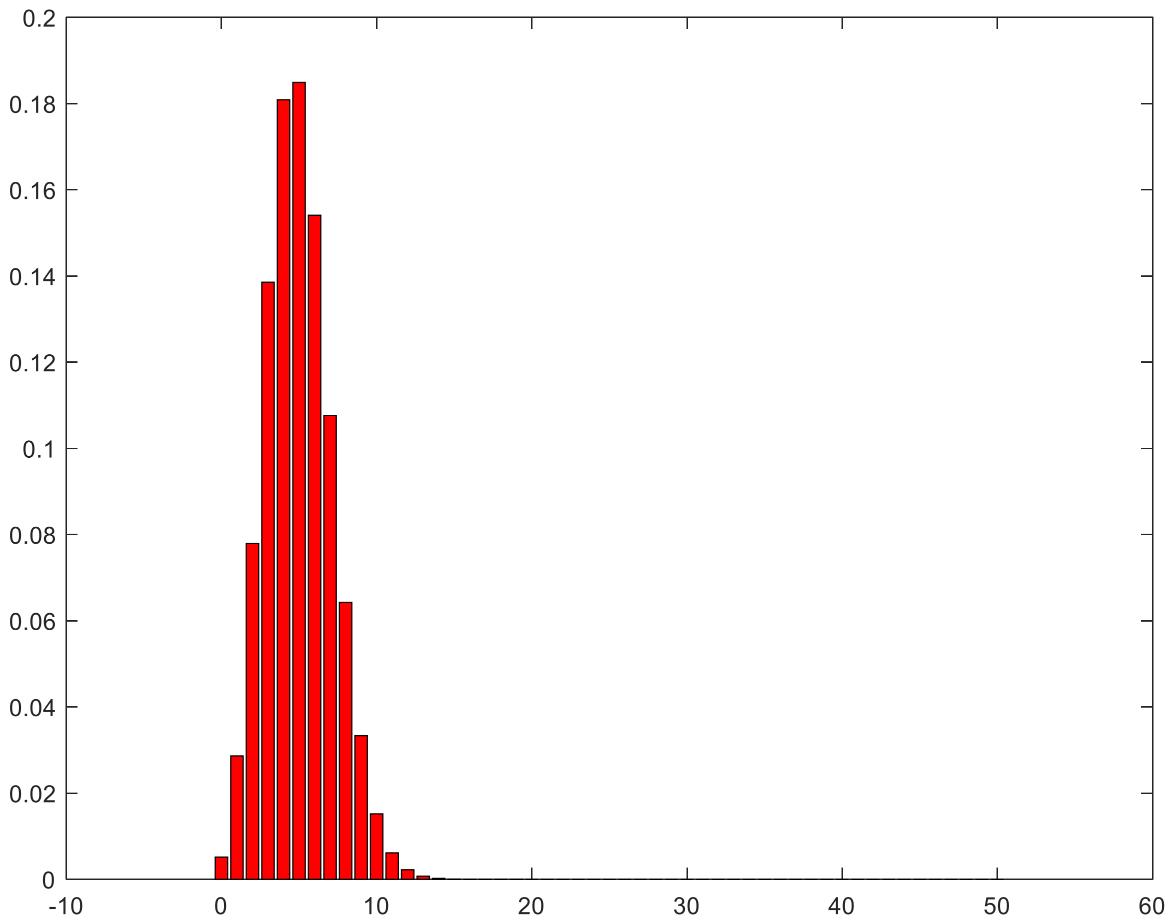


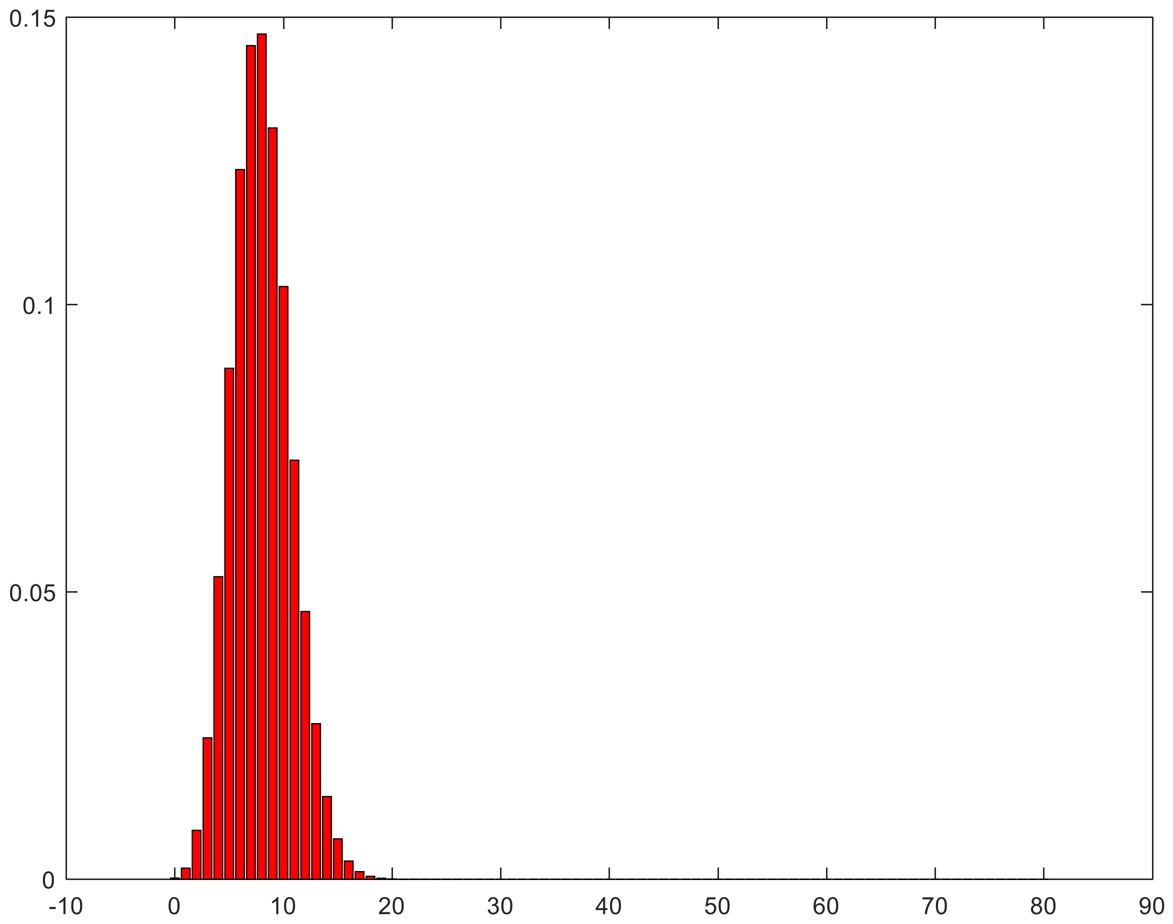


0.3
0.25
0.2
0.15
0.1
0.05
0
-5
0
5
10
15
20
25



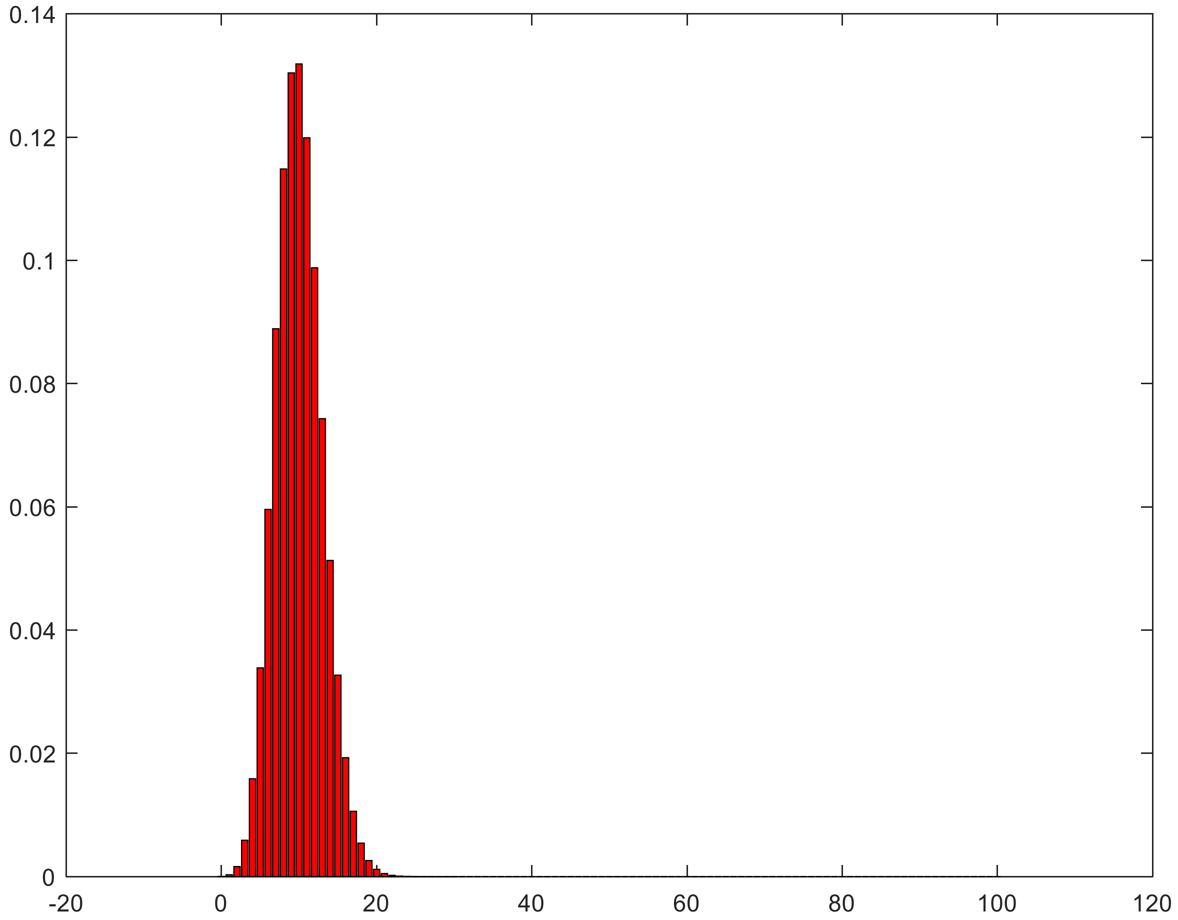
0.25
0.2
0.15
0.1
0.05
0
-5
0
5
10
15
20
25
30
35


0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-10
0
10
20
30
40
50
60



IDC HERZLIYA

$$P\left(\frac{\sqrt{n}}{\sigma}\ (\overline{X_n} - p)\ \leq x\right) \sim \Phi(x)$$



IDC HERZLIYA

Proof(s) of the CLT

Uses moment generating functions.

For an r.v X the mgf is $M(t) = E(\exp(tX))$

<u>The MGF Continuity Lemma (P Levy 1920)</u>:
Consider cdfs F_n with corresponding mgfs M_n.
Also consider a continuous cdf F and a corresponding mgf M.
If $M_n(t) \to M(t)$ in some open interval around 0 then $F_n(x) \to F(x)$ for all $x \in \mathbb{R}$.

This lemma and other properties of the mgf can be used to prove the general Levy-Lindeberg CLT

IDC HERZLIYA

Proofs(s) – cont

$$P\left(\frac{(S_n - np)}{\sqrt{npq}} = \mathrm{round}(x)\right) \sim \varphi(x)$$

The de Moivre version (for Binomilas, 18th century) can be proven directly using Stirling approximation and a lot of algebra.

Stirling: $n! \sim n^n e^{-n}\sqrt{2\pi n}$ as $n \to \infty$

$$\text{binom.pdf}(n, p, np) = \; p^{np}\; q^{nq}\,\frac{n!}{(np)!(nq)!} \;\approx$$

$$\frac{p^{np}}{(np)^{np}\sqrt{2\pi np}} \cdot \frac{q^{nq}}{(nq)^{nq}\sqrt{2\pi nq}} \cdot \frac{n^n\sqrt{2\pi n}}{1} = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{npq}}$$

IDC HERZLIYA

Gaussian mixtures

X is a Gaussian Mixture RV if the density function for X's distribution is:

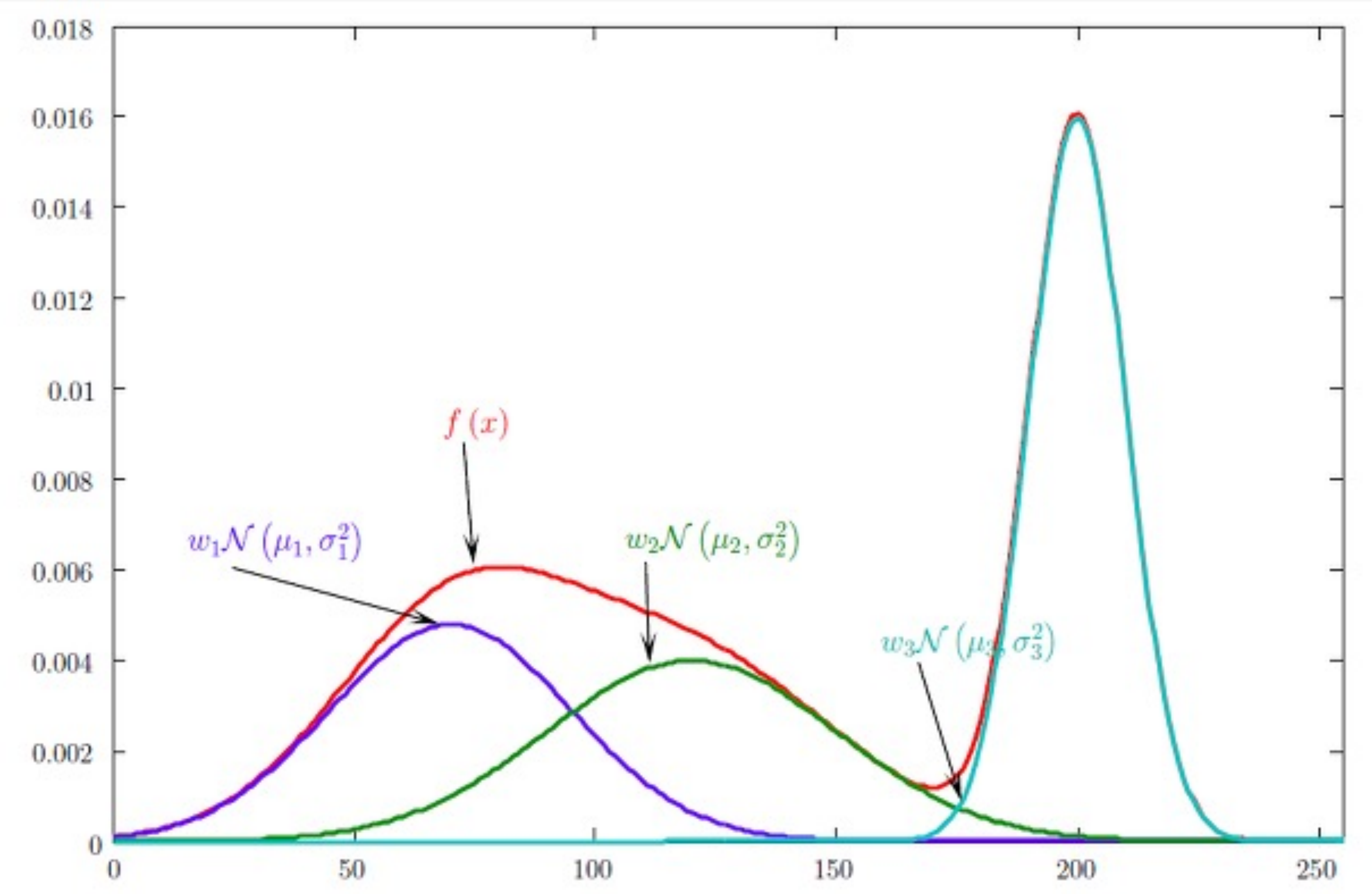
$$f(x) = \sum_{i=1}^{k} w_i f_i(x)$$

Where each one of the density functions, f_i , is a Gaussian density:

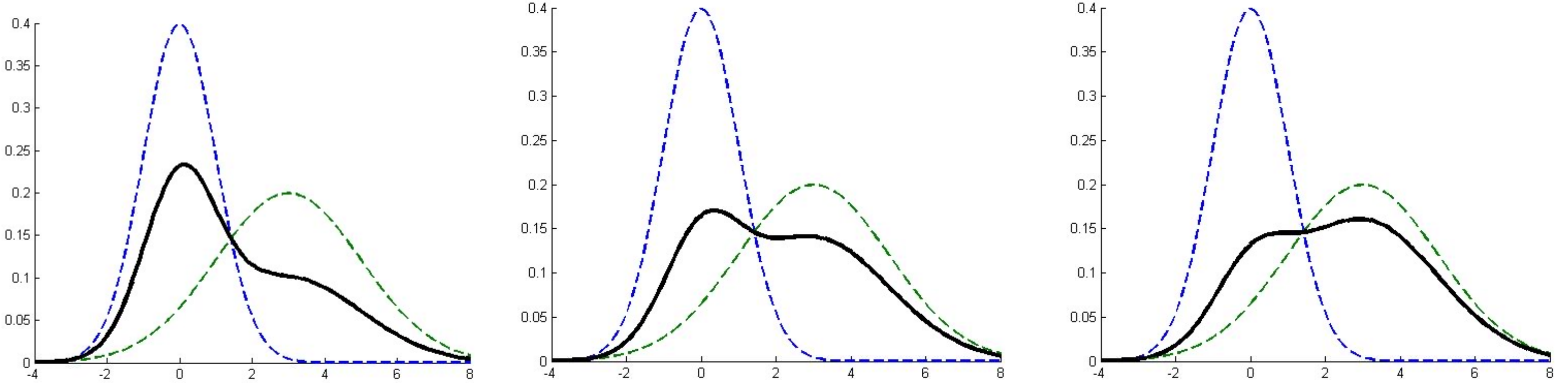
$$f_i(x) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-(x-\mu_i)^2/2\sigma_i^2}$$

IDC HERZLIYA

What does f look like?

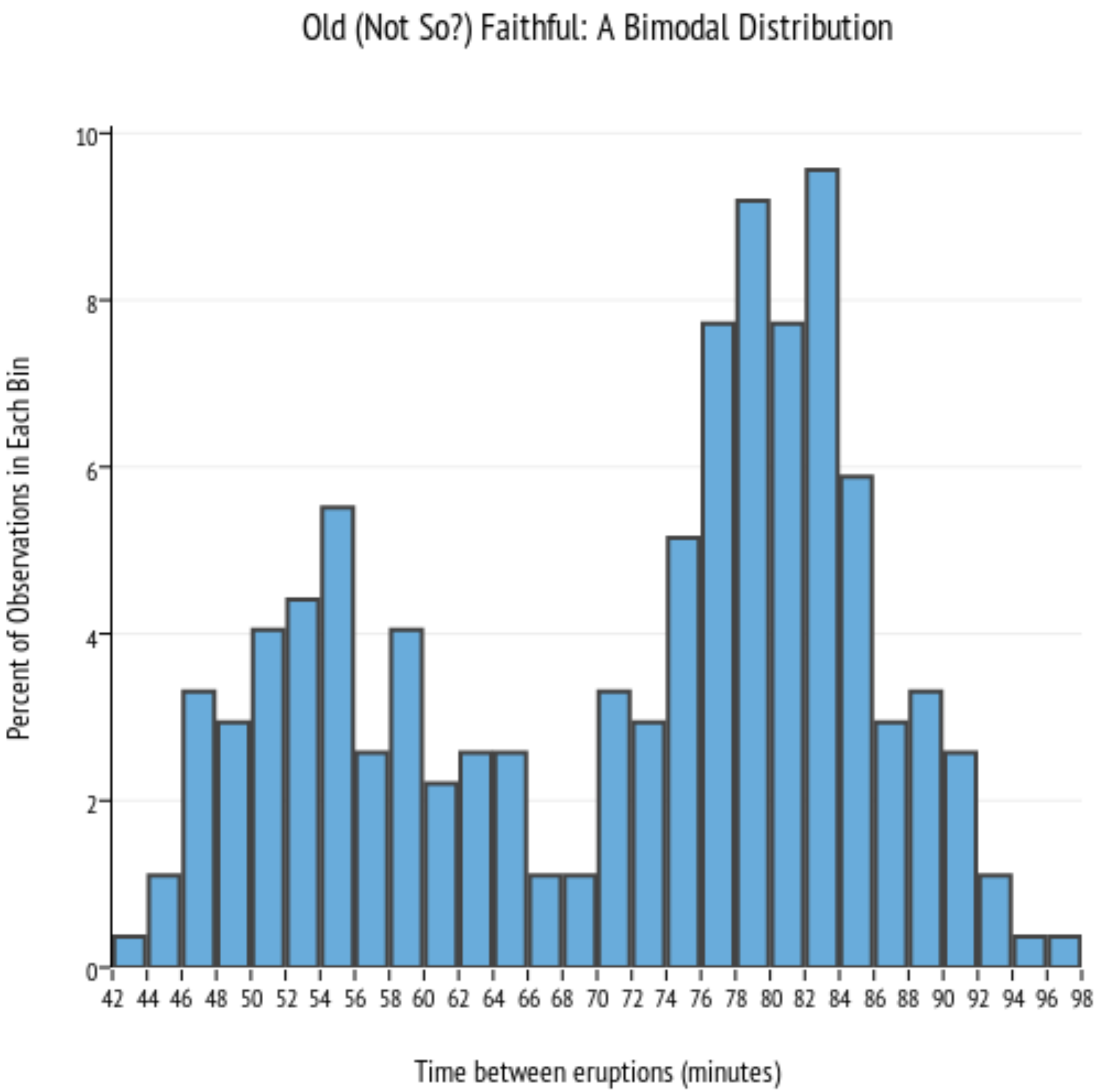


GMD dependence on weights



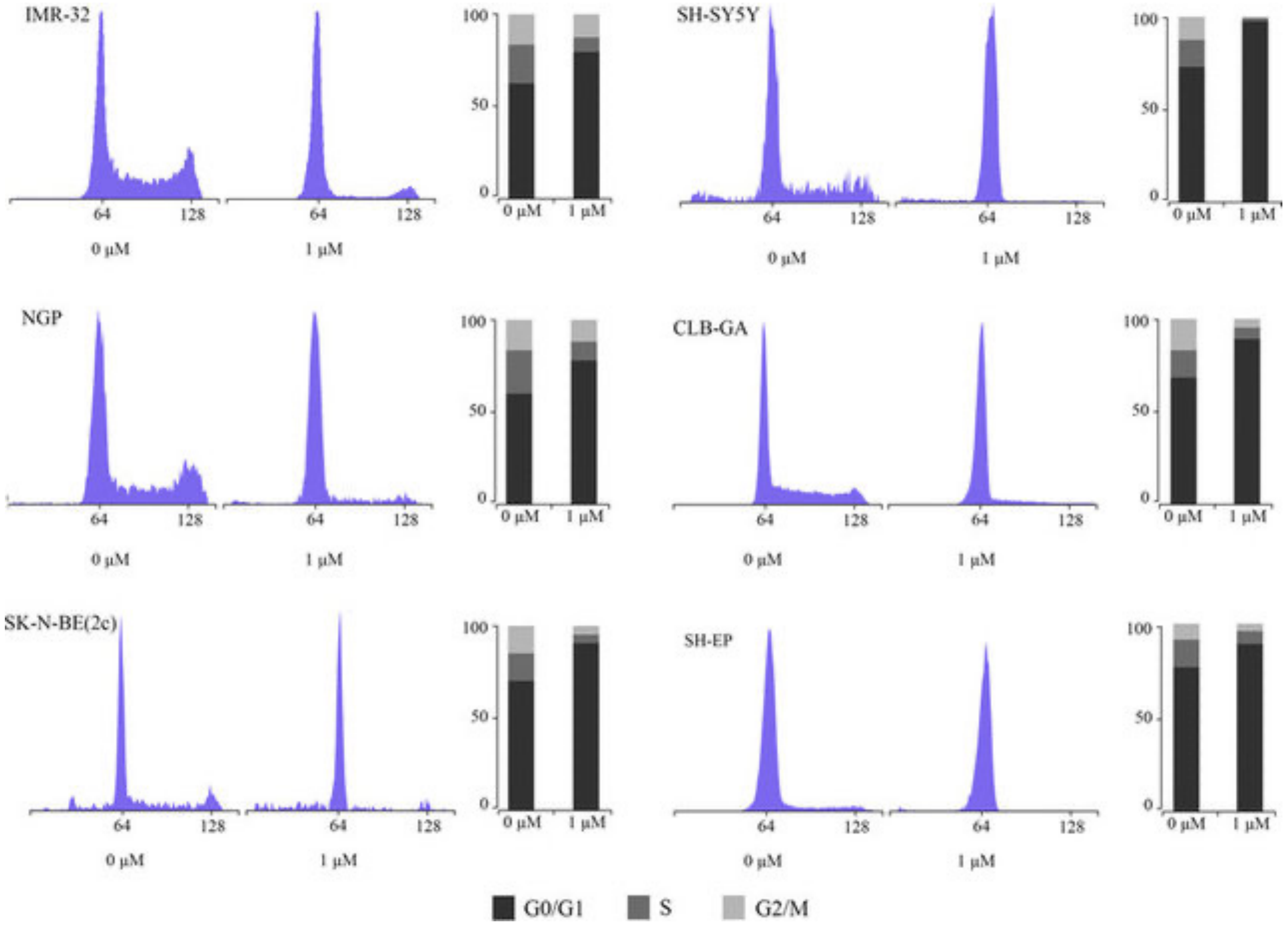
IDC HERZLIYA

Old faithful eruptions



IDC HERZLIYA

Dividing cells



Rihani et al, 2015

IDC HERZLIYA

Expectation and variance of Gaussian mixtures

$$X_i \sim f_i(x)$$
$$X \sim f(x) = \sum_i w_i f_i(x)$$

What is $F(x)$?

$$F(x) = \sum_i w_i F_i(x)$$

Consider some function $g: \mathbb{R} \to \mathbb{R}$
What is $E\big(g(X)\big)$?

$$E\big(g(X)\big) = \int_{-\infty}^{\infty} g(x) f(x) dx = \int_{-\infty}^{\infty} g(x) \left(\sum_i w_i f_i(x) \right) dx$$
$$= \sum_i w_i \left(\int_{-\infty}^{\infty} g(x) f_i(x) dx \right) = \sum_i w_i E\big(g(X_i)\big)$$

Now let $g(x) = x$ and we get $E(X) = \sum_i w_i E(X_i)$

IDC HERZLIYA

Variance of Gaussian mixtures

$$Var(X) = E\left((X-\mu)^2\right)$$
$$= \sum_i w_i E\left((X_i-\mu)^2\right)$$

$$= \sum_i w_i E\left(\left((X_i-\mu_i)+(\mu_i-\mu)\right)^2\right)$$

$$= \sum_i w_i E\left((X_i-\mu_i)^2 + 2(X_i-\mu_i)(\mu_i-\mu) + (\mu_i-\mu)^2\right)$$

$$= \sum_i w_i E\left((X_i-\mu_i)^2\right) + \sum_i 2w_i(\mu_i-\mu)E(X_i-\mu_i) + \sum_i w_i(\mu_i-\mu)^2$$

$$= \sum_i w_i E\left((X_i-\mu_i)^2\right) + \sum_i w_i(\mu_i-\mu)^2$$

IDC HERZLIYA

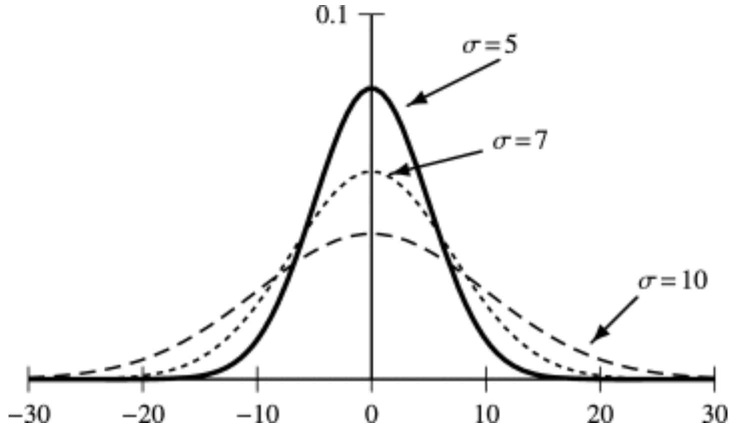
What the CLT does NOT imply

Some interesting CLT variants and extensions

- Different distributions – relaxing the identically distributed assumption
- Markov – relaxing the independence assumption
- MultiD – limits will be multiD Gaussians
- Berry-Essen – a rate on the convergence

IDC HERZLIYA

Summary



- The normal distribution is a convenient distribution to work with and it also represents many naturally occurring and real-life distributions.
- It also represents the limit behavior of repeated sample averages of (almost) any distribution.
- It is easy to make inferences, test hypotheses, calculate probabilities, when using normal distributions.
- Gaussian mixture Models – a useful extension of Gaussians
- Multi-var Gaussians introduced. More next time

IDC HERZLIYA