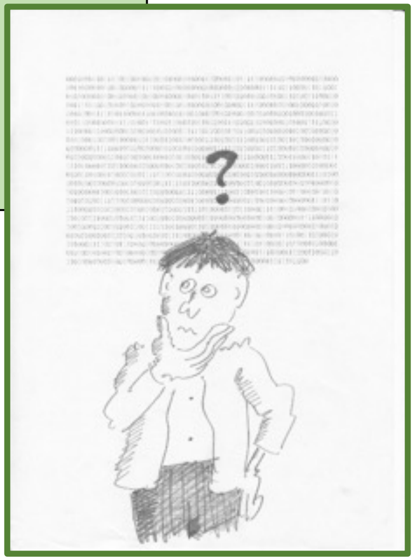
Log-normal distribution, Heavy Tails and the exponential distribution

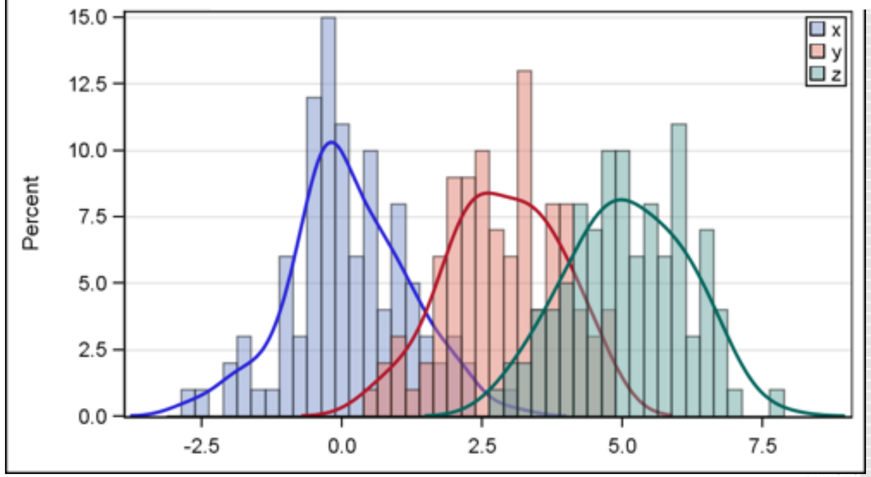
Statistics and data analysis

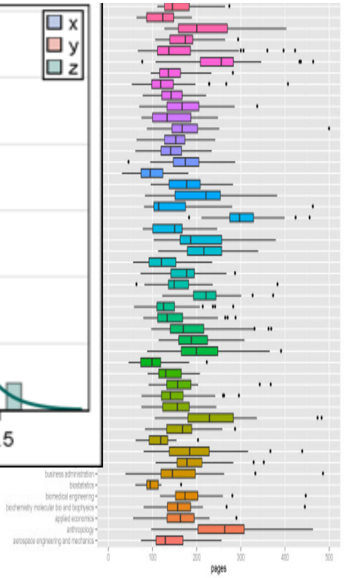
Ben Galili
Leon Anavy
Zohar Yakhini

IDC, Herzeliya

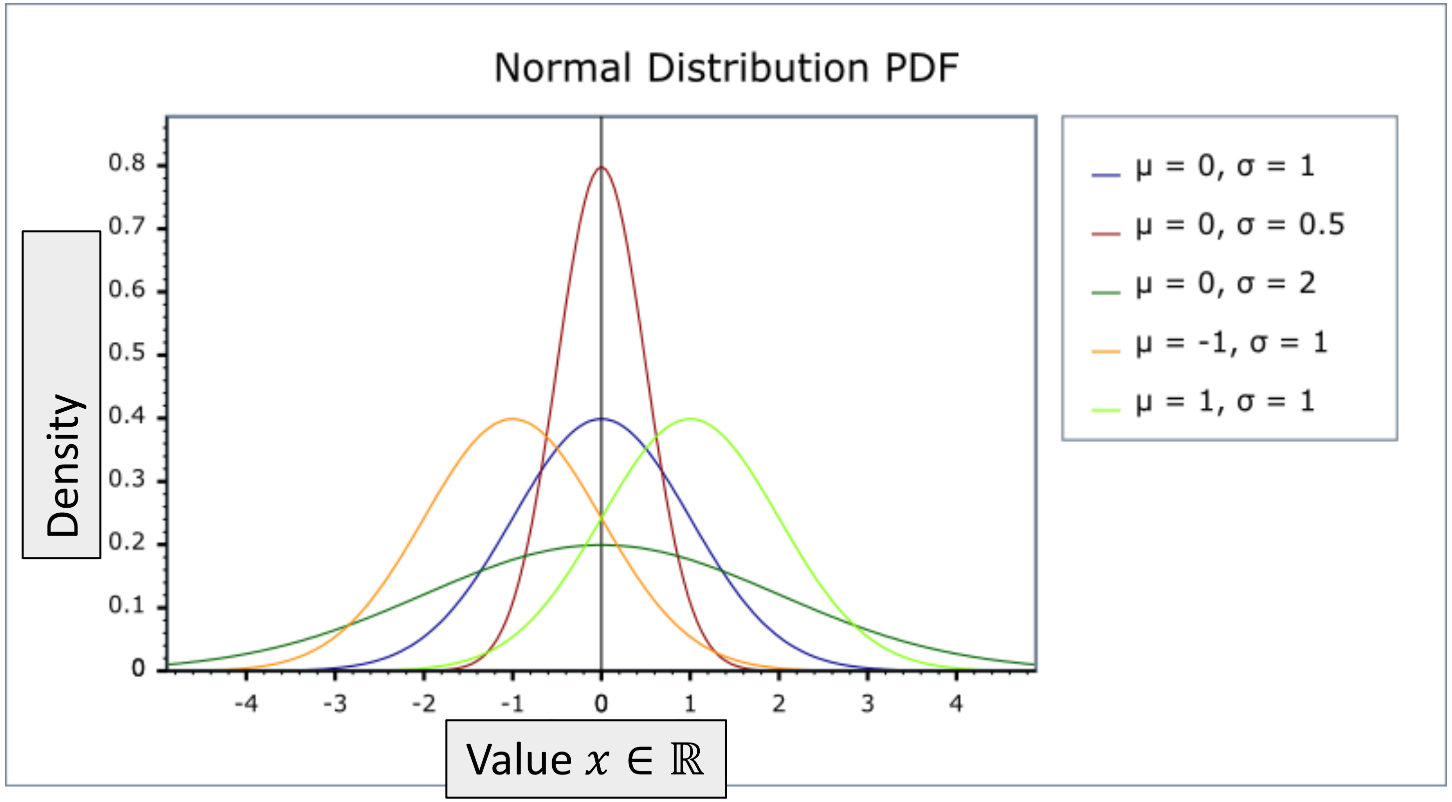


0010011101010100101010100100100010
1010100010101111101011010011001001
1110101010011010010110010110110010





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 distributed with mean μ and standard deviation σ. We write

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

A random variable that has a normal distribution with

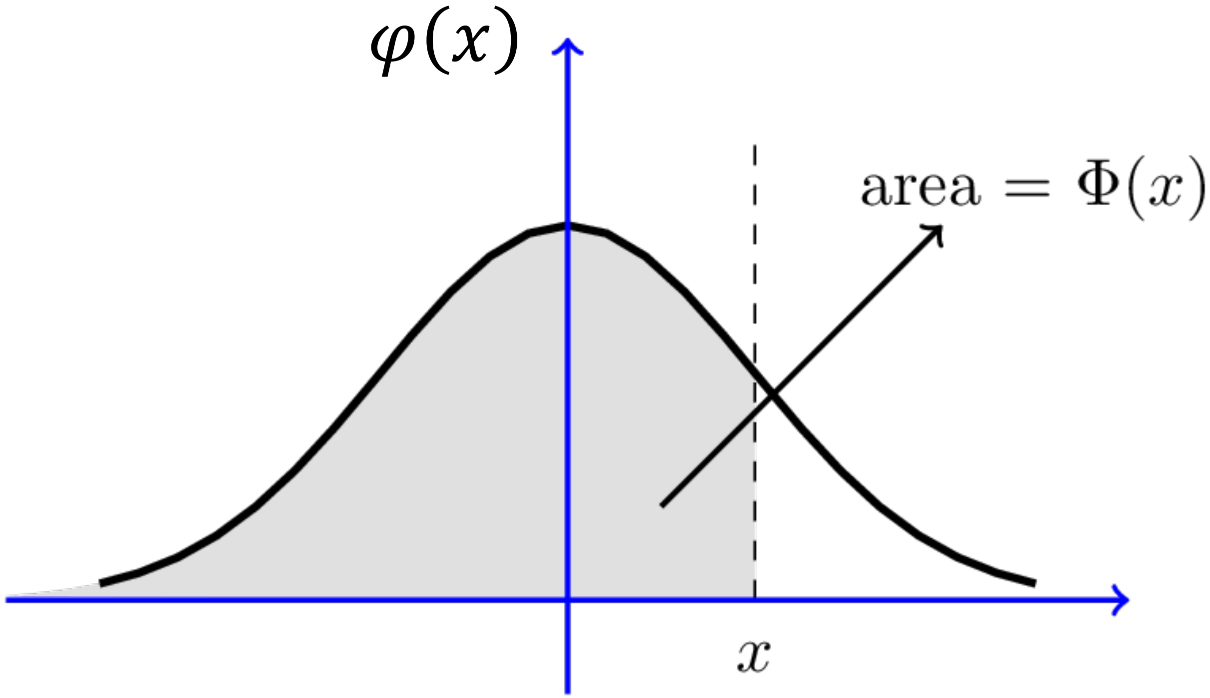
$\mu = 0$ and $\sigma = 1$

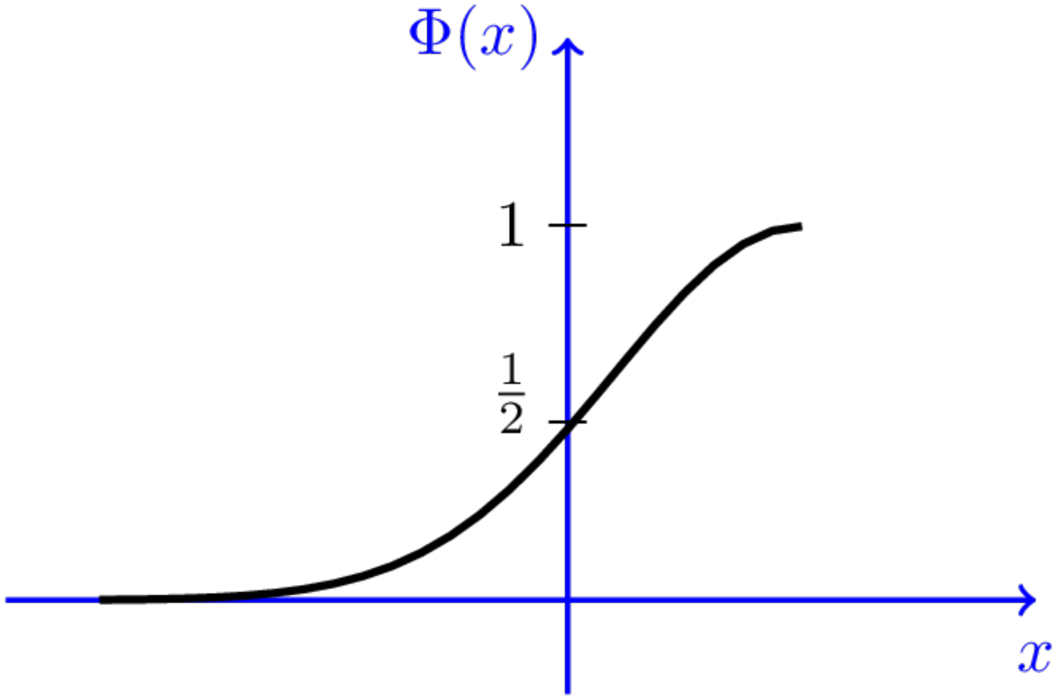
is called Standard Normal: $\mathrm{Z} \sim N(0, 1)$

The density function then becomes: $f_Z(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

IDC HERZLIYA

The CDF of a standard normal is often called Φ

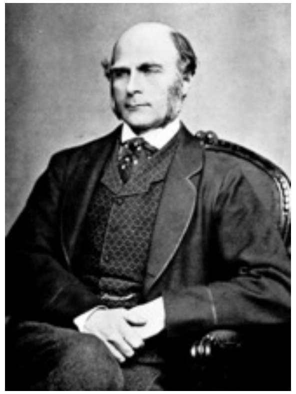




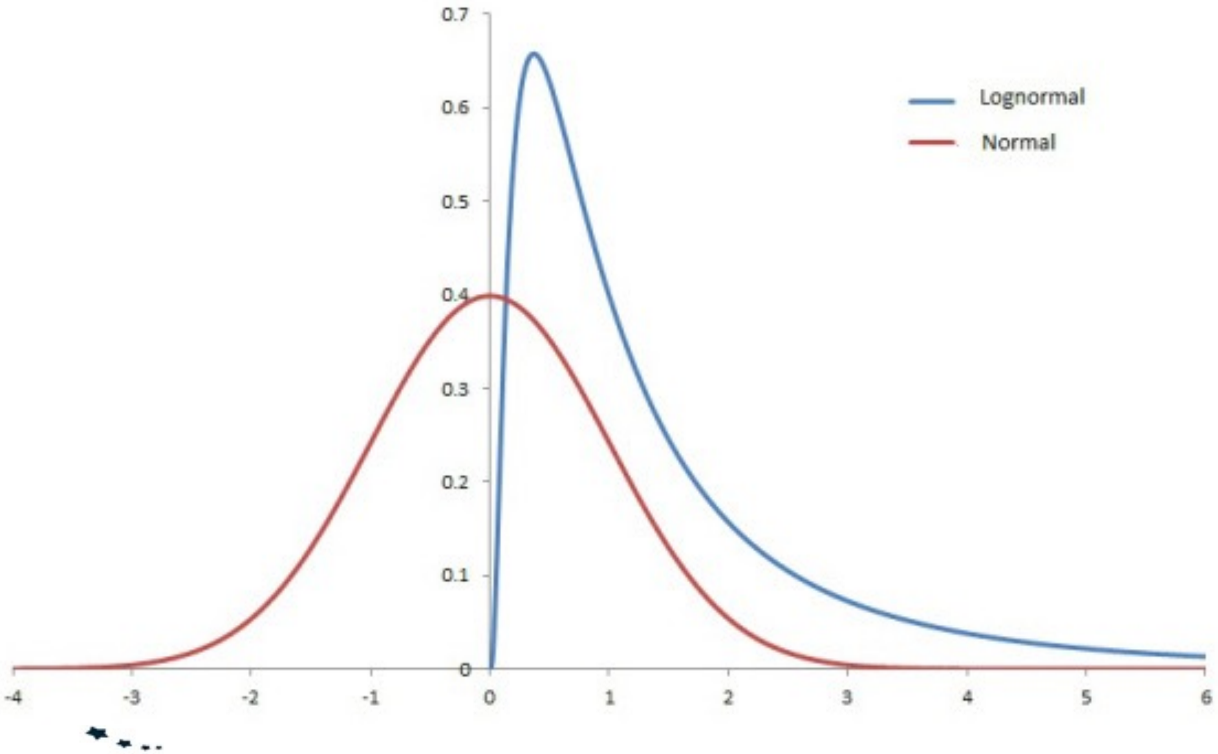
IDC HERZLIYA

Log-normal (Galton) distributions

A random variable Y is said to have a log-normal distribution if its log, log(Y), has a normal (Gaussian) distribution.
In other words – Y is log-normal if $Y = e^X$ for some Gaussian X.
That is: $Y = e^{\mu+\sigma Z}$, where Z is standard normal.



Francis Galton, 1822-1911, British statistician



Log-normals are always positive.
Can be useful in modelling intrinsically positive quantities.
Mean, mode and median are different from each other.
The log-normal distribution has a heavy right side tail.
μ and σ are called the *location* and *scale* of Y.
They are NOT the mean and std of Y.
They are the mean and std of $X = \ln(Y)$

The density of the log-normal distribution

Let Y be a **standard** log-normal random variable. Let f(y) and F(y) be the PDF and CDF of a standard log-normal.
Let Z be standard normal and $\varphi(z)$ and $\Phi(z)$ denote the PDF and CDF of the standard normal distribution.

$$F(y) =$$

The density of the log-normal distribution

Let Y be a **standard** log-normal random variable. Let f(y) and F(y) be the CDF and PDF of a standard log-normal.
Let Z be standard normal and $\Phi(z)$ and $\varphi(z)$ denote the CDF and PDF of the standard normal distribution.

$$F(y) = P(Y \le y) = P(e^Z \le y) = P(Z \le \ln y) = \Phi(\ln y)$$

Therefore, since the PDF is the derivative of the CDF, we get:

$$f(y) = F'(y) = \frac{\Phi'(\ln y)}{y} = \frac{\varphi(\ln y)}{y}$$

IDC HERZLIYA

The density of the log-normal distribution

Let Y be a log-normal random variable with location μ and scale σ.
What are the CDF and the PDF of Y.
Let Z be standard normal and $\Phi(x)$ and $\varphi(x)$ denote the CDF and PDF of the standard normal distribution.

$$F(y) = P(Y \ \le y) = P(\, e^{X} \ \le y) = P(\, e^{\mu+\sigma Z} \ \le y) =$$

$$P(\mu + \sigma Z \le \ \ln y) = P\left(Z \le \frac{\ln y - \mu}{\sigma}\right) = \Phi\left(\frac{\ln y - \mu}{\sigma}\right)$$

Therefore, since the PDF is the derivative of the CDF, we get:

$$f(y) = F'(y) = \frac{\Phi'\left(\frac{\ln y - \mu}{\sigma}\right)}{\sigma y} = \frac{\varphi\left(\frac{\ln y - \mu}{\sigma}\right)}{\sigma y}$$

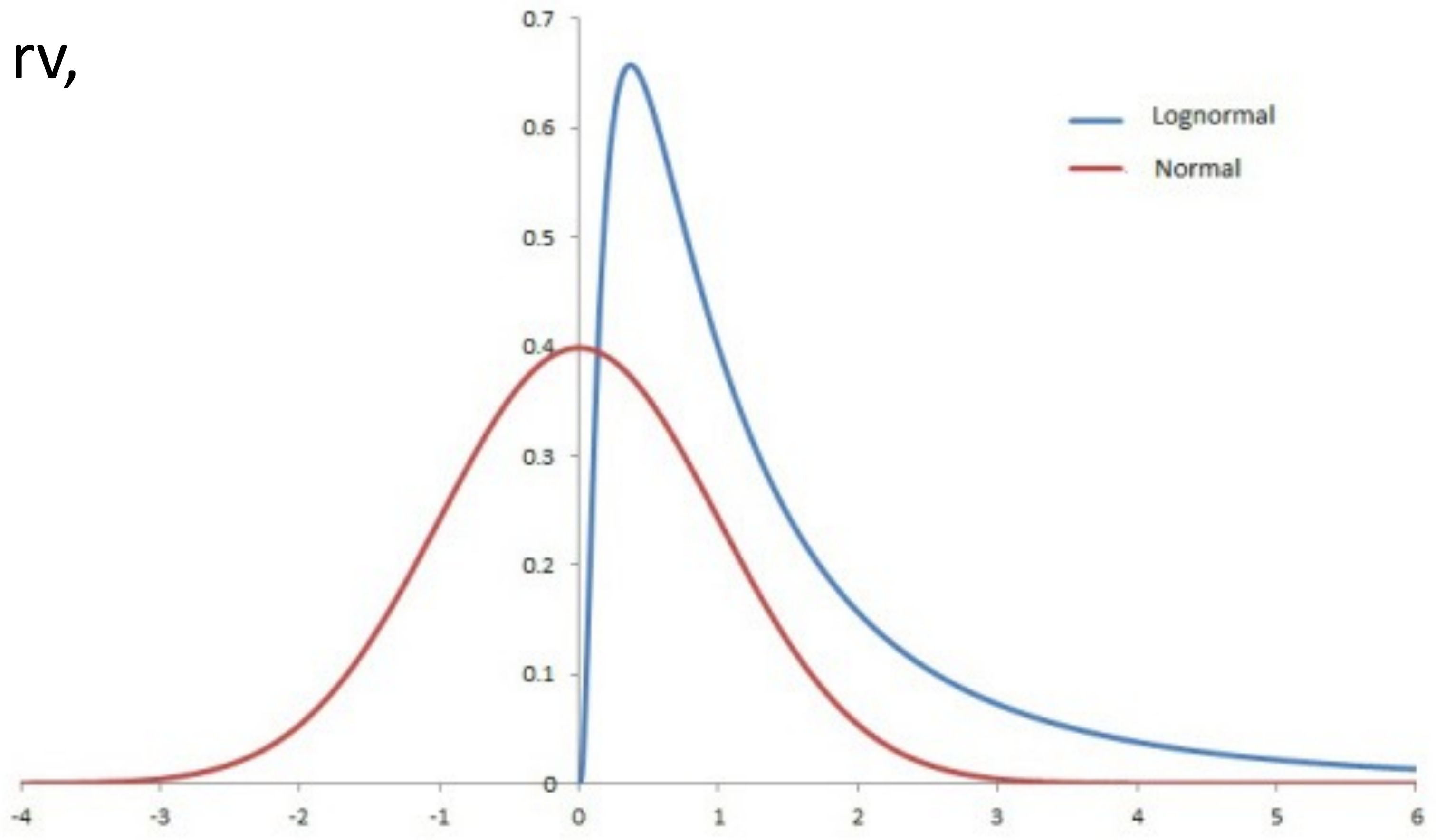
IDC HERZLIYA

Median, mode and expected value of a log-normal

Let $Y = e^{\mu+\sigma z}$ be a log-normal rv,

Calculate:

- Median(Y)
- Mode(Y)
- E(Y)



IDC HERZLIYA

Median, mode and expected value of a log-normal

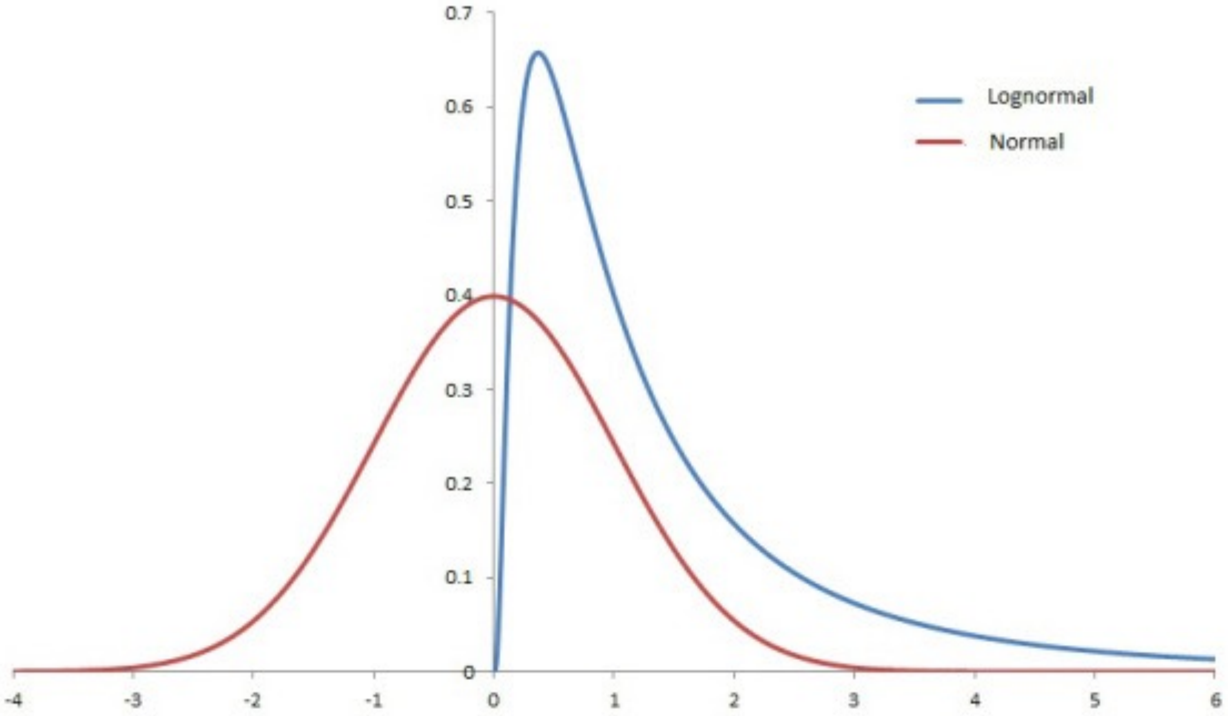
Let $Y = e^{\mu+\sigma z}$ be a log-normal rv, then:

- Median(Y) = e^{μ}

$$P(Y \leq m) = 0.5$$

$$P\left(Z \leq \frac{\ln m - \mu}{\sigma}\right) = 0.5$$

$$\ln m = \mu \rightarrow m = e^{\mu}$$



IDC HERZLIYA

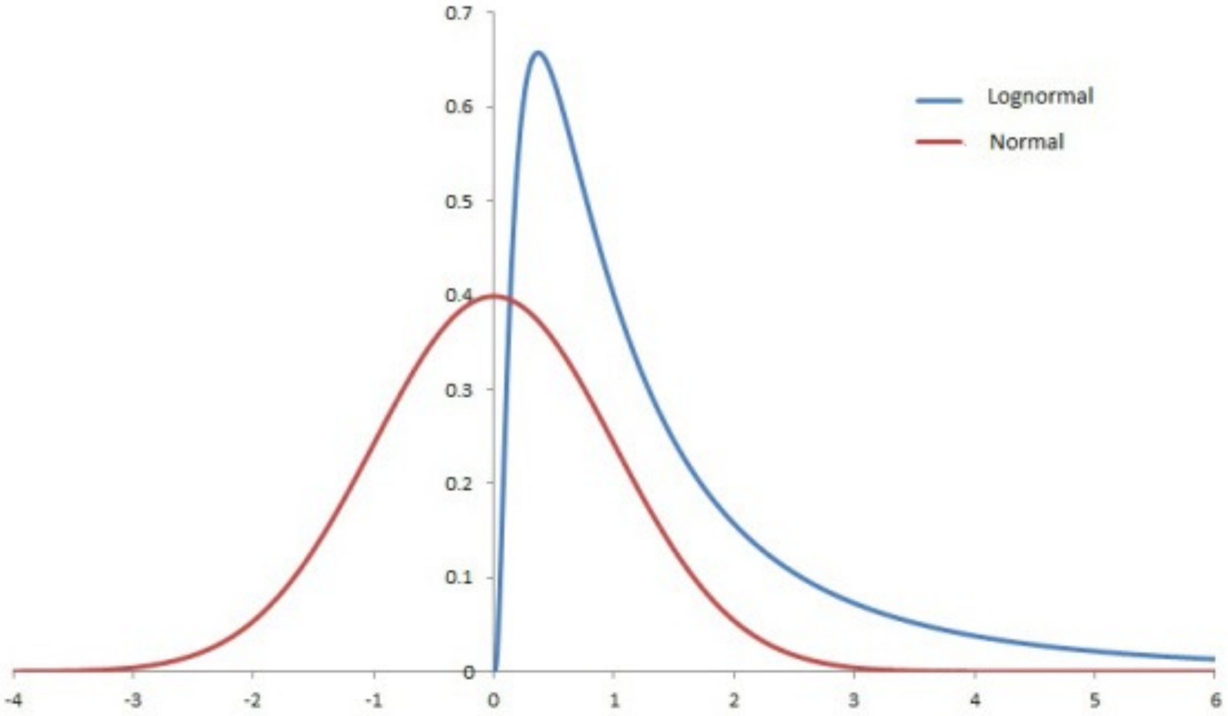
Median, mode and expected value of a log-normal

Let $Y = e^{\mu+\sigma Z}$ be a log-normal rv, then:

- Mode(Y)=$e^{\mu-\sigma^2}$

Use the derivative of f(y)

$$f'(mode) = 0$$



IDC HERZLIYA

Median, mode and expected value of a log-normal

Let $\mathrm{Y} = e^{\mu+\sigma z}$ be a log-normal rv, $\mathrm{E(Y)} =?$

For the **standard** log-normal case ($\mu = 0, \sigma = 1$):

$$E(Y) = \int_0^\infty y f(y) dy = \int_0^\infty \frac{y\varphi(\ln y)}{y} dy = \int_0^\infty \varphi(\ln y) dy =^{(*)}$$

IDC HERZLIYA

Median, mode and expected value of a log-normal

Let $\mathrm{Y} = e^{\mu+\sigma z}$ be a log-normal rv, $\mathrm{E(Y)} = ?$
For the **standard** log-normal case ($\mu = 0, \sigma = 1$):

$(*)\ t = \ln y \rightarrow y = e^t, \frac{dy}{dt} = e^t$
$(**)\ (t-1)^2 = t^2 - 2t + 1$

$$\int_0^\infty \varphi(\ln y) dy =^{(*)} \int_{-\infty}^\infty \varphi(t) e^t dt = \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} e^t dt$$

$$=^{(**)} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{(t-1)^2}{2}+\frac{1}{2}} dt = e^{\frac{1}{2}} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{(t-1)^2}{2}} dt = e^{\frac{1}{2}}$$

IDC HERZLIYA

What will the general case be?

Log-normal examples

- Comment length in internet discussions
- Company size
- City size
- Particle size

Multiplicative processes

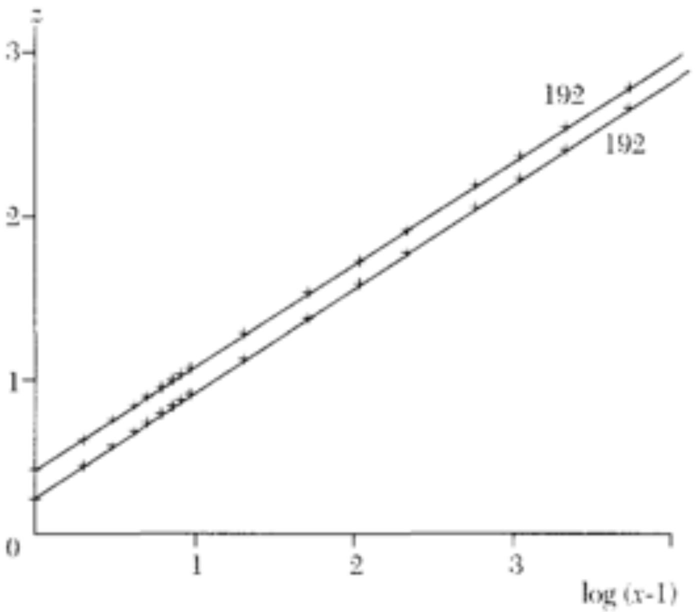
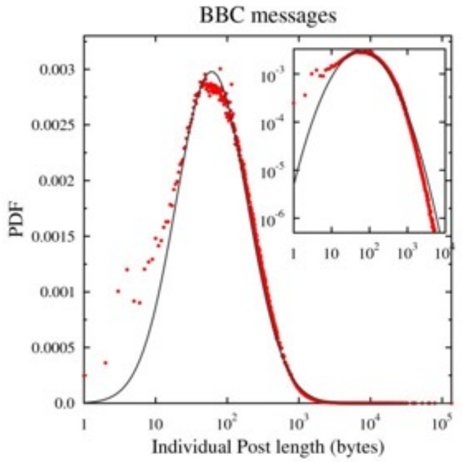
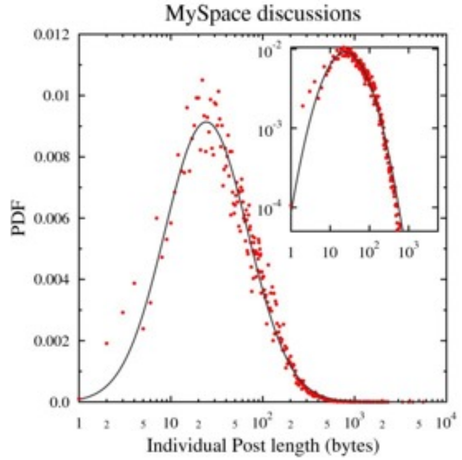


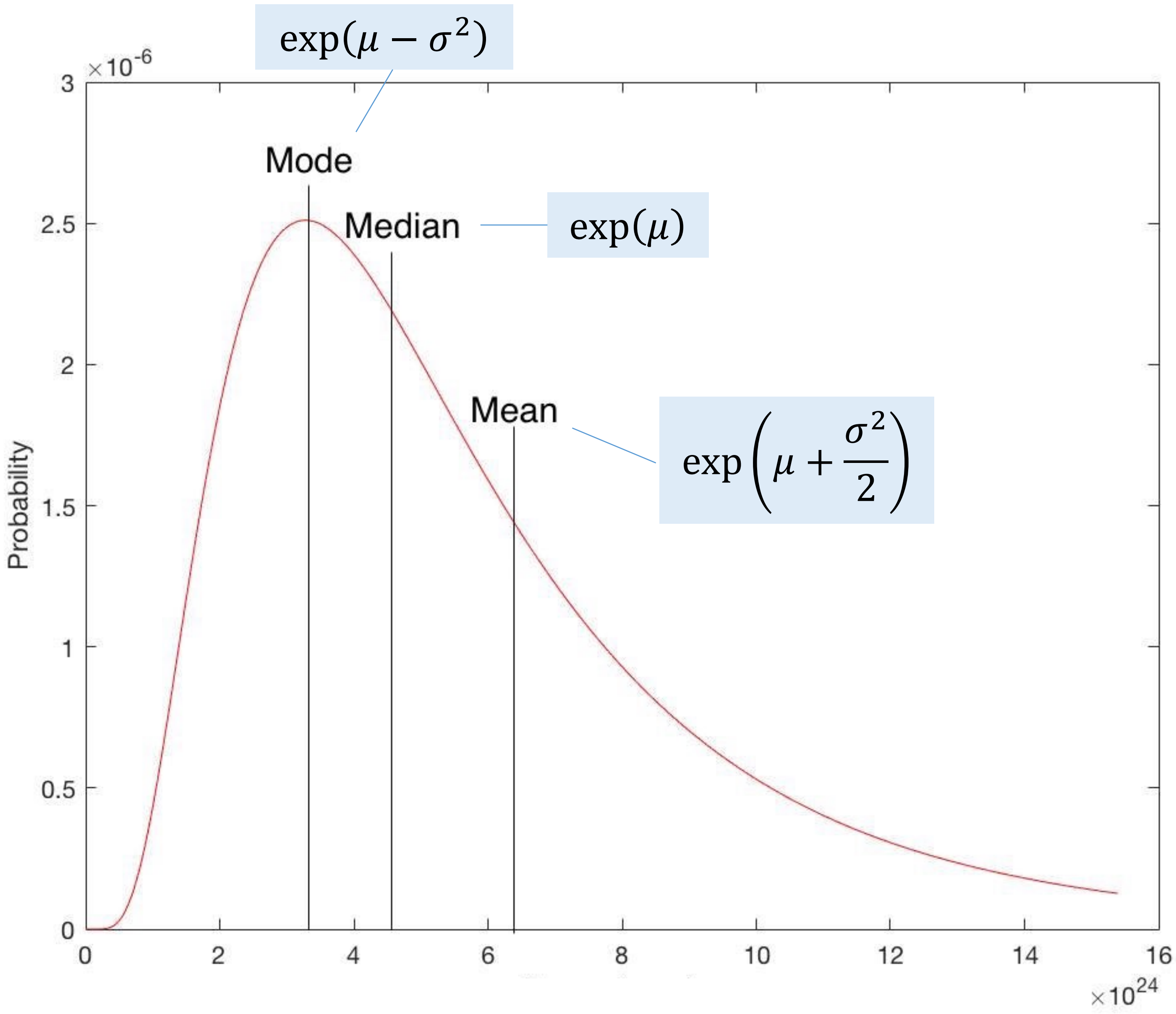
Figure 1. Gibrat's Data for French Manufacturing Establishments in 1920 and 1921



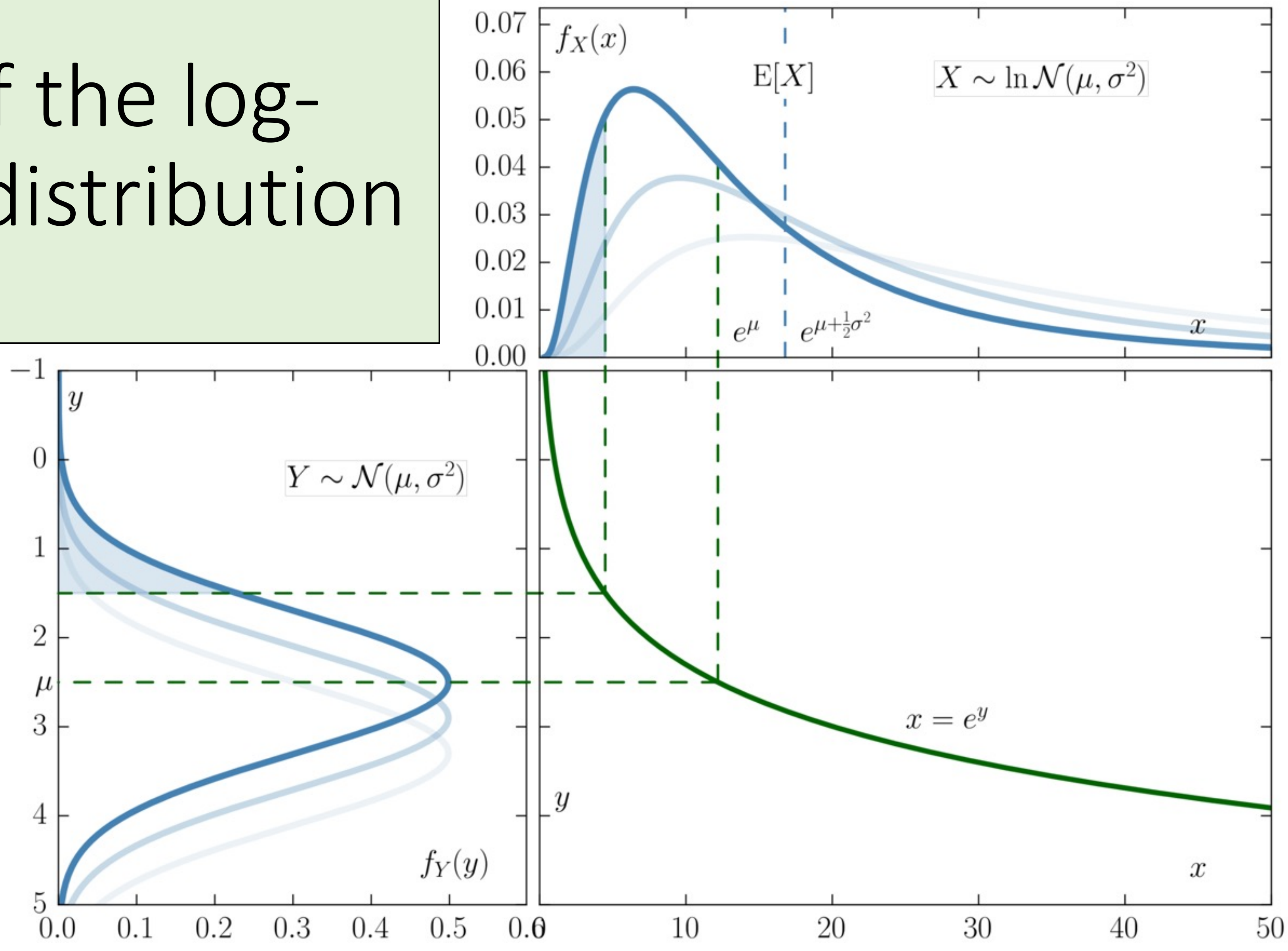


IDC HERZLIYA

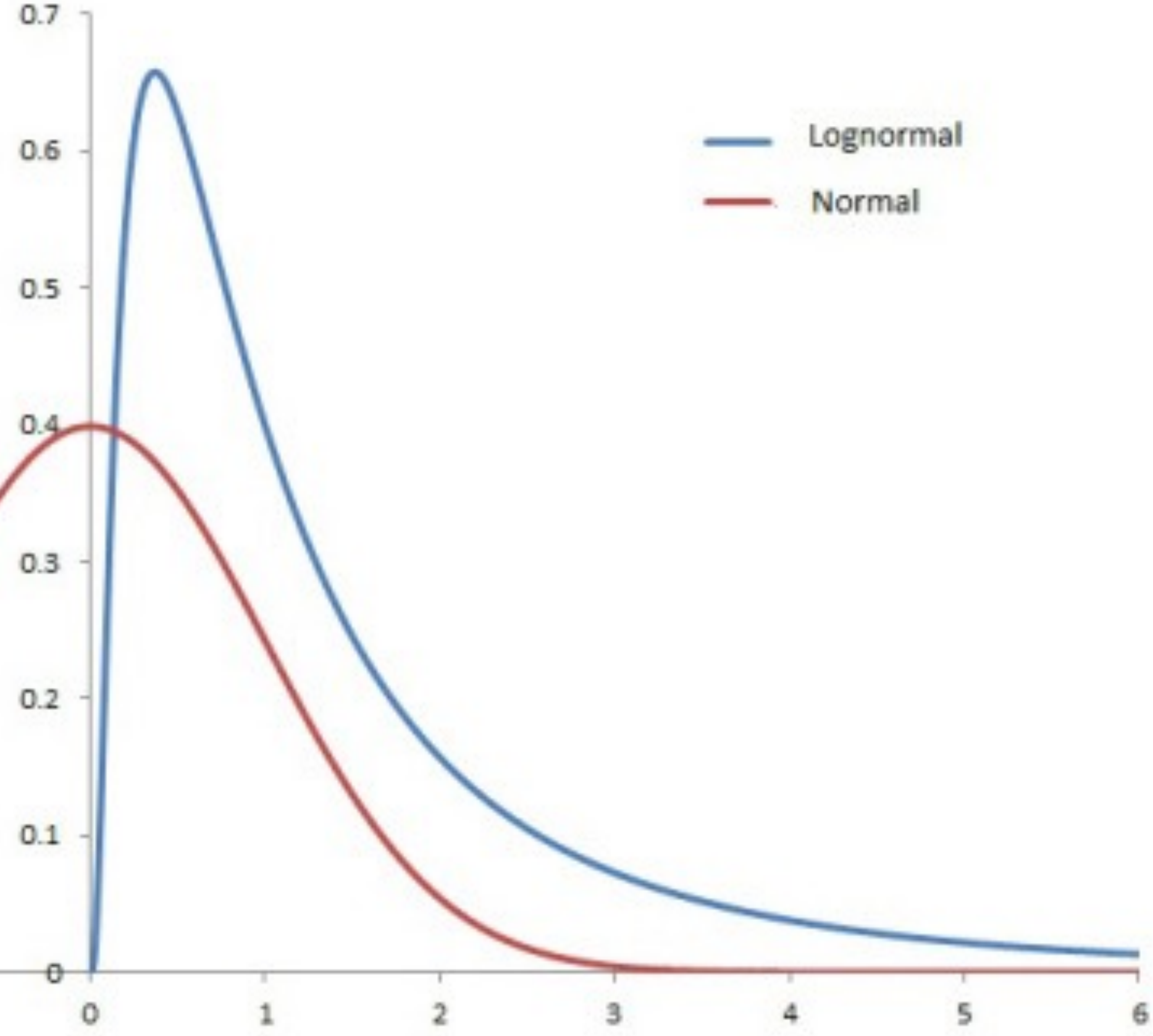
Shape of the log-normal distribution



Shape of the log-normal distribution

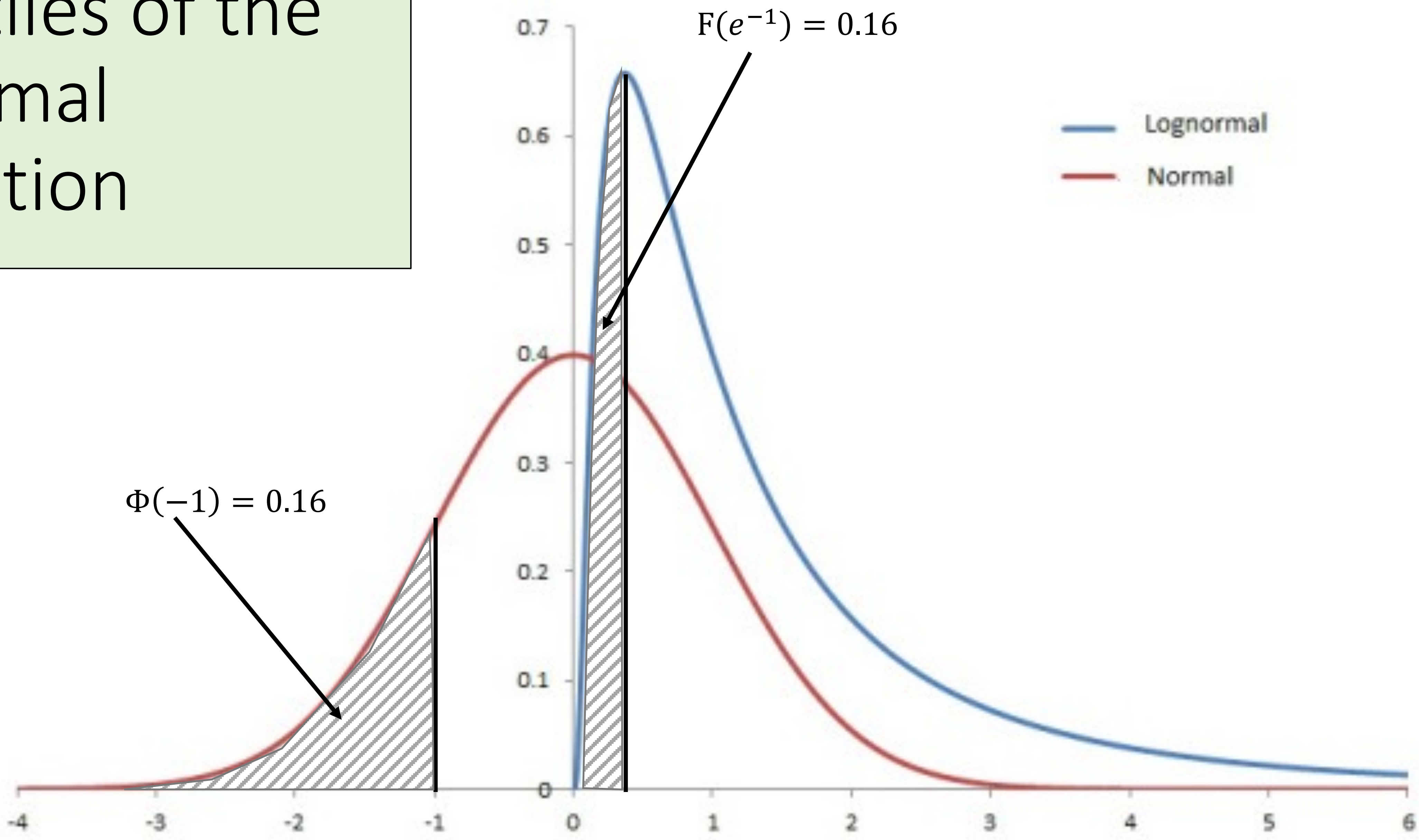


Percentiles of the log-normal distribution



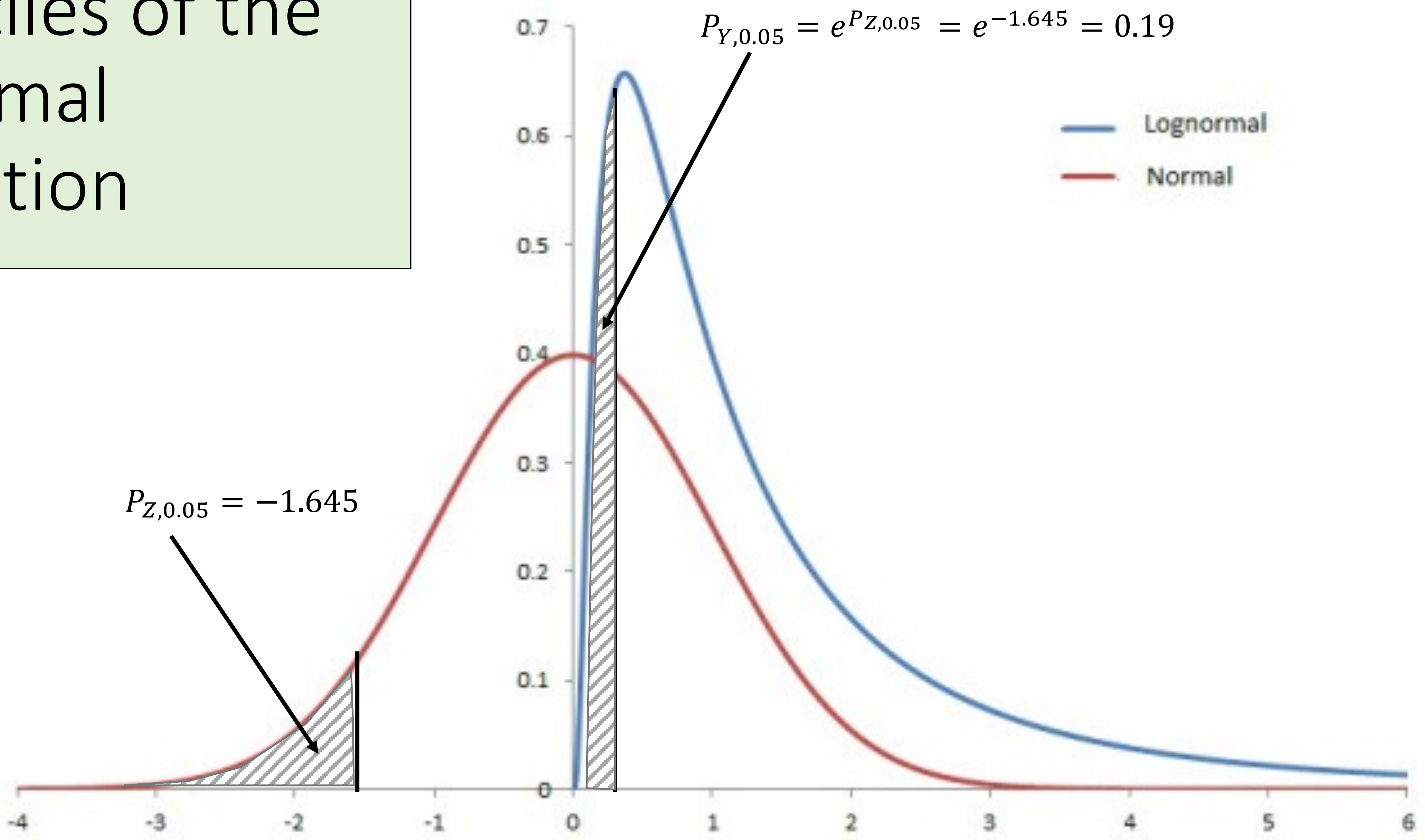
IDC HERZLIYA

Percentiles of the log-normal distribution



IDC HERZLIYA

Percentiles of the log-normal distribution



Heavy Right Tail

A distribution is said to have a heavy right tail if its tail probabilities vanish slower than any exponential:

$$\forall t > 0 \quad \lim_{x \to \infty} e^{tx} P(X > x) = \infty$$

IDC HERZLIYA

Normal distribution is not heavy tailed

$$\forall t > 0 \quad \lim_{x\to\infty} e^{tx} P(X > x) = \infty$$

$$X \sim N(0,1)$$

$$\lim_{x\to\infty} e^{tx}\left(1 - \int_{-\infty}^{x} \varphi(u)du\right) = \lim_{x\to\infty} \frac{\left(1 - \int_{-\infty}^{x} \varphi(u)du\right)}{e^{-tx}} =$$

$$\lim_{x\to\infty} \frac{-\varphi(x)}{-te^{-tx}} = \lim_{x\to\infty} \frac{e^{\frac{-x^2}{2}}}{te^{-tx}} = \lim_{x\to\infty} \frac{1}{t} e^{\left(\frac{-x^2}{2} + tx\right)} = 0, \forall t$$

IDC HERZLIYA

Log-normal distribution is heavy tailed

In the HW

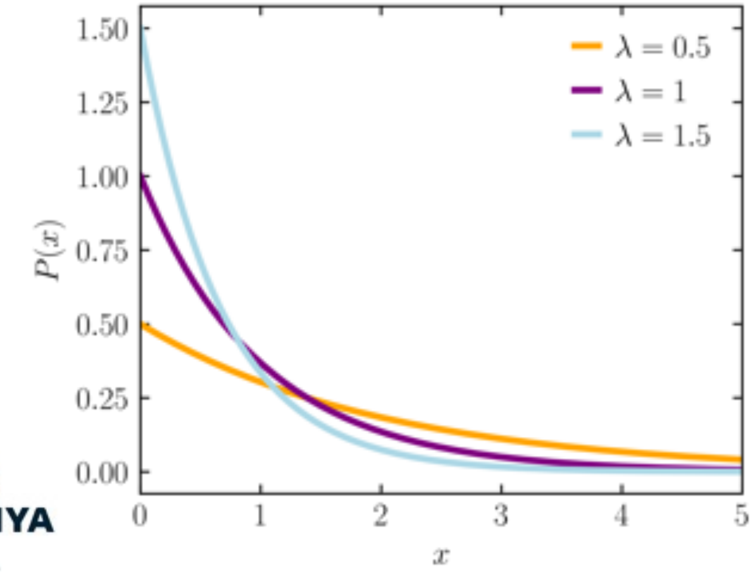
Exponential distribution

A random variable X is said to have an exponential distribution with **rate** $\lambda > 0$ if its PDF is

$$X \sim \exp(\lambda)$$

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0 \end{cases}$$

Siméon Poisson,
1781-1840,
French mathematician



- Exponential distribution describes the wait time in a Poisson process.
- It is the continuous analogue of the Geometric distribution.
- It is memoryless.

Exponential distribution

A random variable X is said to have an exponential distribution with **rate** $\lambda > 0$
if its PDF is

$$X \sim \exp(\lambda)$$

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0 \end{cases}, \qquad F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0 \end{cases}$$

$$F(x) = \int_0^x f(u)du = \int_0^x \lambda e^{-\lambda u} du = \left[-\frac{1}{\lambda}\lambda e^{-\lambda u}\right]_0^x = -e^{-\lambda x} + 1 = 1 - e^{-\lambda x}$$

Mean, Variance and Median

$$(*)\ \int u dv = uv - \int v du$$
$$u = x, v = -e^{-\lambda x}, dv = \lambda e^{-\lambda x}, du = 1$$

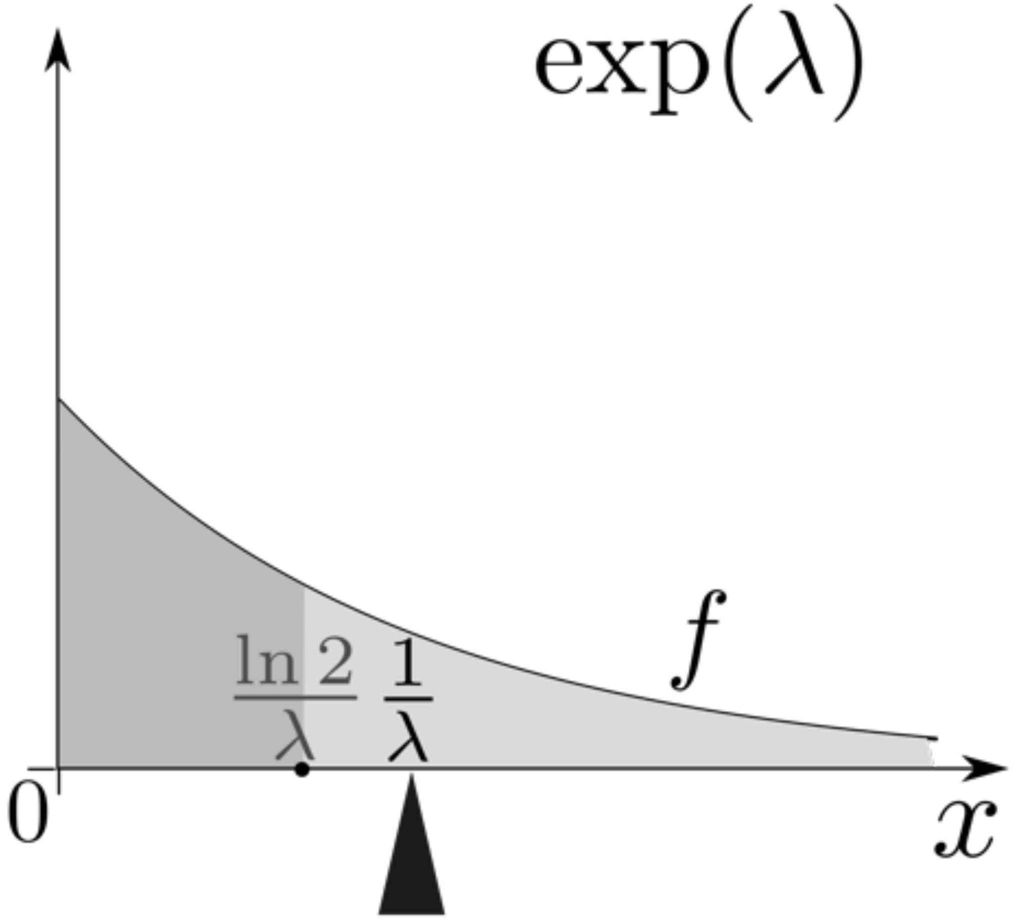
- $E(X) = \int_0^\infty x\lambda e^{-\lambda x} dx =^{(*)} -xe^{-\lambda x}\Big|_0^\infty - \int_0^\infty -e^{-\lambda x} dx = 0 - \frac{1}{\lambda} e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$
- Median

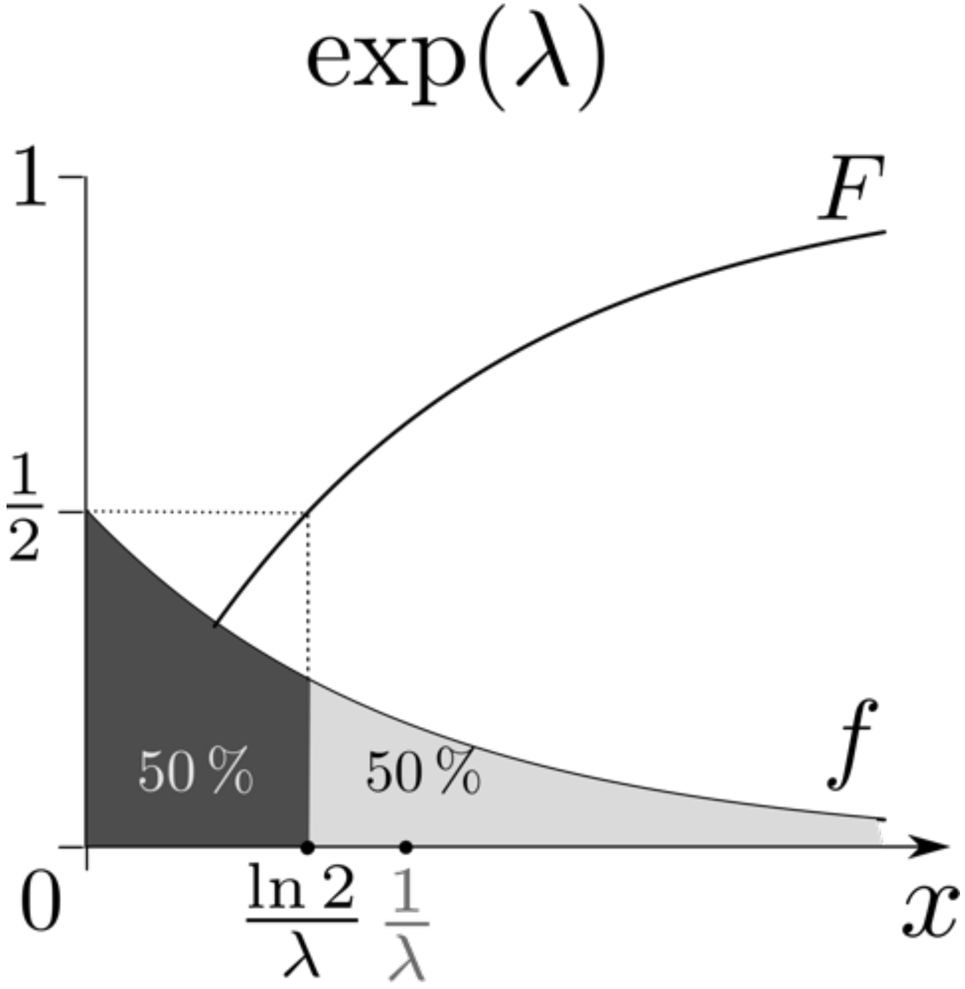
$$P(X \le m) = F(m) = 1 - e^{-\lambda m} = 0.5 \ \rightarrow e^{-\lambda m} = 0.5$$

$$\rightarrow m = -\frac{\ln(0.5)}{\lambda} = \frac{\ln 2}{\lambda}$$

What would be the q'th percentile?

Mean, Variance and Median





The exponential distribution is memoryless

$X \sim \exp(\lambda)$

$P(X > x + a | X > a) = P(X > x)$

$$P(X > x + a | X > a) = \frac{P(X > x + a, X > a)}{P(X > a)} = \frac{P(X > x + a)}{P(X > a)}$$

$$= \frac{1 - F(x + a)}{1 - F(a)} = \frac{e^{-\lambda(x+a)}}{e^{-\lambda a}} = e^{-\lambda x} = P(X > x)$$

IDC HERZLIYA

Special properties of exponential distribution

$X_1, \ldots X_n$ independent exponential r.v. $X_i \sim \exp(\lambda_i)$

$\mathrm{Y} = \min(X_1, \ldots, X_n) \sim \exp(\sum \lambda_i)$

$$P(Y > y) = P(X_1 > y, X_2 > y, \ldots, X_n > y) = \prod_{i=1}^{n} P(X_i > y)$$

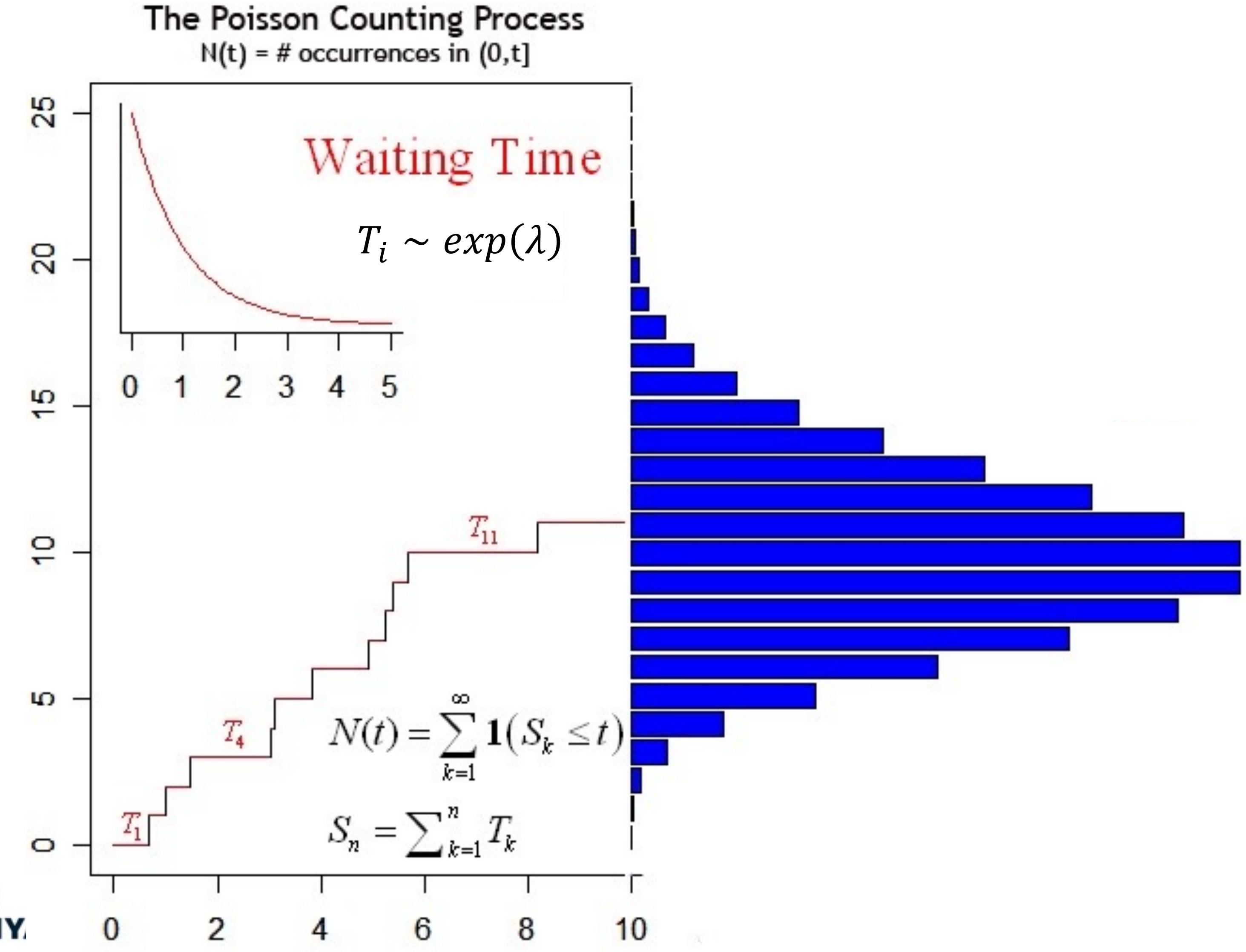
$$= \prod_{i=1}^{n} 1 - F_i(y) = \prod_{i=1}^{n} e^{-\lambda_i y} = e^{-\sum \lambda_i y}$$

Special properties of exponential distribution

$X_1 \sim \exp(\lambda_1), X_2 \sim \exp(\lambda_2)$ independent exponential r.v.

$P(X_1 < X_2) =$ In the HW

Exponential distribution and Poisson processes



$$N(t) \sim Poisson(\lambda t)$$

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

$$N(1) \sim Poisson(\lambda)$$

$$P(N(1) = 0) = e^{-\lambda}$$

$$P(T_1 > 1) = 1 - F(1) = e^{-\lambda}$$

$$N(2) \sim Poisson(2\lambda)$$

$$P(N(2) = 0) = e^{-2\lambda}$$

$$P(T_1 > 2) = 1 - F(2) = e^{-2\lambda}$$

Exponential distribution is not heavy tailed

$$\forall t > 0 \quad \lim_{x\to\infty} e^{tx} P(X > x) = \infty$$

$$\lim_{x\to\infty} e^{tx} P(X > x) = \lim_{x\to\infty} e^{tx} e^{-\lambda x} = \lim_{x\to\infty} e^{(t-\lambda)x} = 0 \text{ for } t < \lambda$$

Summary

- Log Normal distribution
- Heavy tailed distributions
- Exponential distribution