Bioinformatics and Biostatistics BB2440: Biostatistics
Lecture 3: Mean and Variance
Z-Score
More on Normal Distribution
Central Limit Theorem

Timo Koski

TK

09.09.2013

KTH Matematik

Outline of Lecture 3.

- Mean and Variance for Continuous Distributions, Z score

KTH Matematik

Outline of Lecture 3.

- Mean and Variance for Continuous Distributions, Z score
- Normal Distribution (from Lect 2. and new facts)

KTH VETENSKAP OCH KONST
KTH Matematik

Outline of Lecture 3.

- Mean and Variance for Continuous Distributions, Z score
- Normal Distribution (from Lect 2. and new facts)
- Sums of Random Variables, Independence, Means and Variances

KTH Matematik

Outline of Lecture 3.

- Mean and Variance for Continuous Distributions, Z score
- Normal Distribution (from Lect 2. and new facts)
- Sums of Random Variables, Independence, Means and Variances
- Law of Large Numbers

KTH Matematik

Outline of Lecture 3.

- Mean and Variance for Continuous Distributions, Z score
- Normal Distribution (from Lect 2. and new facts)
- Sums of Random Variables, Independence, Means and Variances
- Law of Large Numbers
- Central Limit Theorem

KTH Matematik

Outline of Lecture 3.

- Mean and Variance for Continuous Distributions, Z score
- Normal Distribution (from Lect 2. and new facts)
- Sums of Random Variables, Independence, Means and Variances
- Law of Large Numbers
- Central Limit Theorem
  - Normal Plot

KTH VETENSKAP OCH KONST
KTH Matematik

Outline of Lecture 3.

- Mean and Variance for Continuous Distributions, Z score
- Normal Distribution (from Lect 2. and new facts)
- Sums of Random Variables, Independence, Means and Variances
- Law of Large Numbers
- Central Limit Theorem
  - Normal Plot
  - Examples and Applications

KTH Matematik

Continuous Random Variables

A **continuous random variable** *has infinitely many values, and those values can be associated with measurements on a continuous scale without gaps or interruptions.*

KTH Matematik

A Density Curve

A density curve $f(x)$ is graph of a continuous probability distribution. It must satisfy the following properties

- Total area under the curve must equal 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- Every point on the curve must have a vertical height that is 0 or greater (that is, the curve cannot fall under the x-axis).

KTH Matematik

A Density Curve

Because the total area under a density curve is equal to 1, there is correspondence between area and probability.

KTH Matematik

Probability as Area under Density Curve

X is continuous and represented by a density curve $f(x)$. Then

$$Pr(a < X \leq b) = \int_a^b f(x)dx$$

KTH Matematik

Mean and Variance

Now we introduce something new: the mean and variance of a continuous random variable.

KTH Matematik

Mean and Variance

The **mean** *of a continuous random variable is denoted by* $E(X)$ *and is defined by*

$$E(X) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} xf(x)dx$$

The **variance** *of a continuous random variable is denoted by* $V(X)$ *and is defined by*

$$V(X) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} (x - E(X))^2 f(x)dx$$

The **standard deviation** *of a continuous random variable is denoted by* $D(X)$ *and is equal to* $D(X) = \sqrt{V(X)}$.

KTH Matematik

Mean and Variance

Clearly

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

and

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)dx$$

are not the same as mean and variance in Lect 1., i.e.:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

$$s^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

respectively. We shall point out a link (= statistical inference) between these concepts in the sequel (Lect 4).

KTH VETENSKAP OCH KONST
KTH Matematik

Coefficient of Variation

Definition

The **coefficient of variation** *for a non-negative X is given by*

$$CV = \frac{D(X)}{E(X)}$$

This is free of units of measurement and can be used to compare variation for data taken from different populations. Expressed often as $\frac{D(X)}{E(X)}100\%$.
Sample coefficient of variation is

$$\mathrm{CV} = \frac{s}{\bar{x}}$$

The coefficient of variation is for a set of non-negative data.

KTH VETENSKAP OCH KONST
KTH Matematik

Coefficient of Variation

$$CV = \frac{D(X)}{E(X)} \rightarrow \textit{ Population CV}$$

$$CV = \frac{s}{\bar{x}} \rightarrow \textit{ Sample CV}$$

The statistical attributes population and sample will be explained later (= statistical inference).

KTH Matematik

Z-score or Standard Score

A **standard score** *or a* **Z score** *is*

- *the number of standard deviations that a random variable X is above or below the mean, or*

$$Z = \frac{X - E(X)}{D(X)}$$

- *the number of standard deviations s that a single data x is above or below the mean, or* $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

$$z = \frac{x - \bar{x}}{s}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Z-score or Standard Score: More statistical terminology

$$Z = \frac{X - E(X)}{D(X)} \rightarrow \textit{ Population Z-score}$$

$$z = \frac{x - \bar{x}}{s} \rightarrow \textit{ Sample Z-score}$$

The statistical notions of population and sample will be explained later (= statistical inference).

KTH Matematik

Mean of the Exponential distribution Exp(λ)

$X \in \text{Exp}(\lambda)$, $\lambda > 0$

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

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2}$$

KTH Matematik

Normal Distribution (a.k.a. Gaussian Distribution)

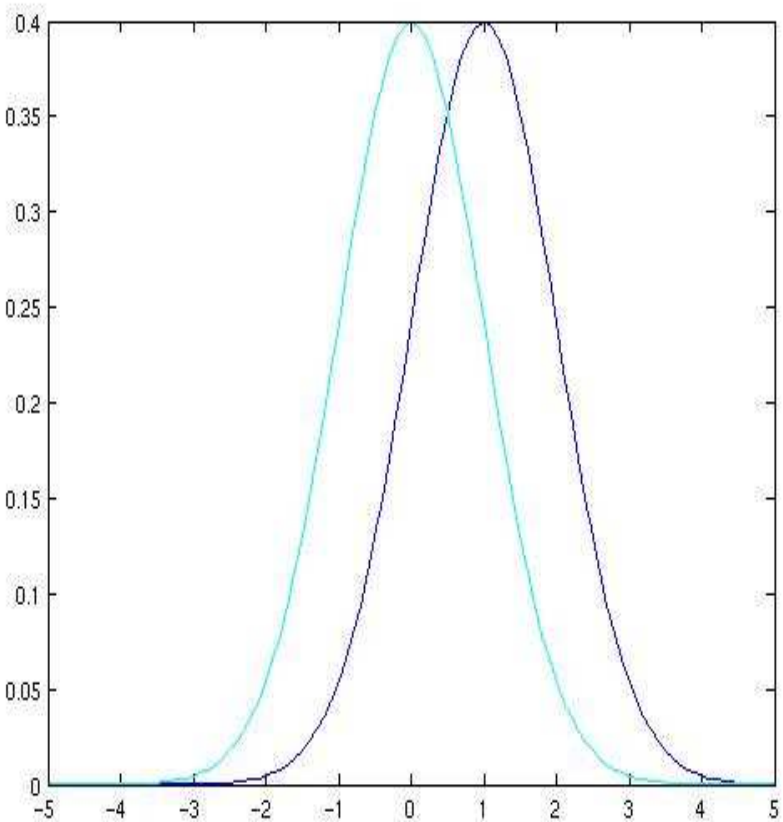
$X \sim N(\mu, \sigma)$

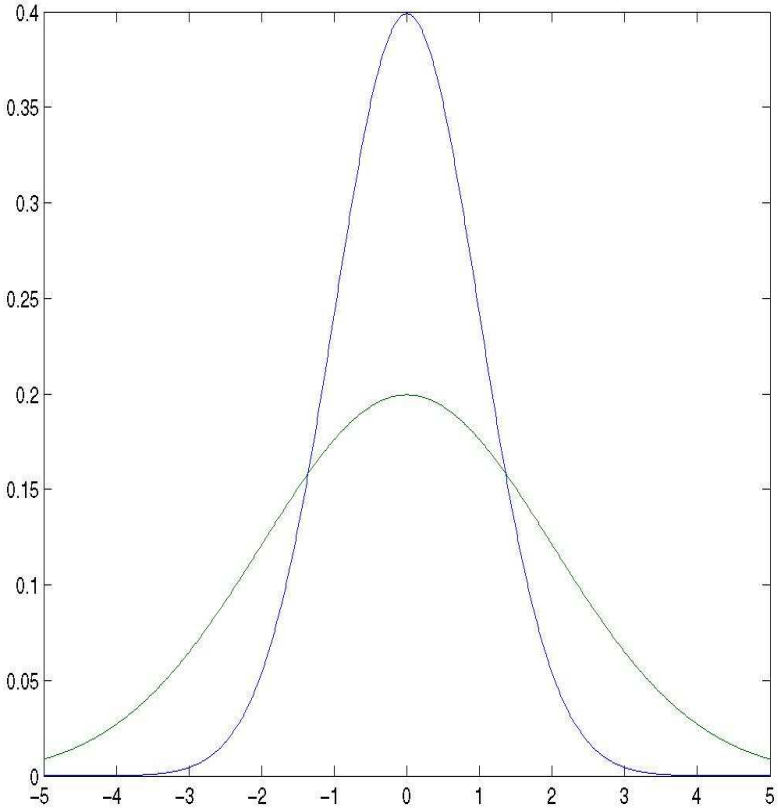
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$E(X) = \mu, \quad V(X) = \sigma^2.$$

KTH Matematik

$N(0,1)$ and $N(1,1)$ and $N(0,1)$ and $N(0,2)$ (from left to the right)





KTH Matematik

Finding the Area Between Two Values Using a Table

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

$$Pr(a < X \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-(x-\mu)^2/2\sigma^2} dx$$

We shall now explain how to find $Pr(a < X \leq b)$. The standard normal distribution is needed for this.

KTH VETENSKAP OCH KONST
KTH Matematik

Standard Normal distribution, $N(0, 1)$

The **standard normal distribution** *is a normal distribution that has a mean of* 0 *and a standard deviation equal to* 1.

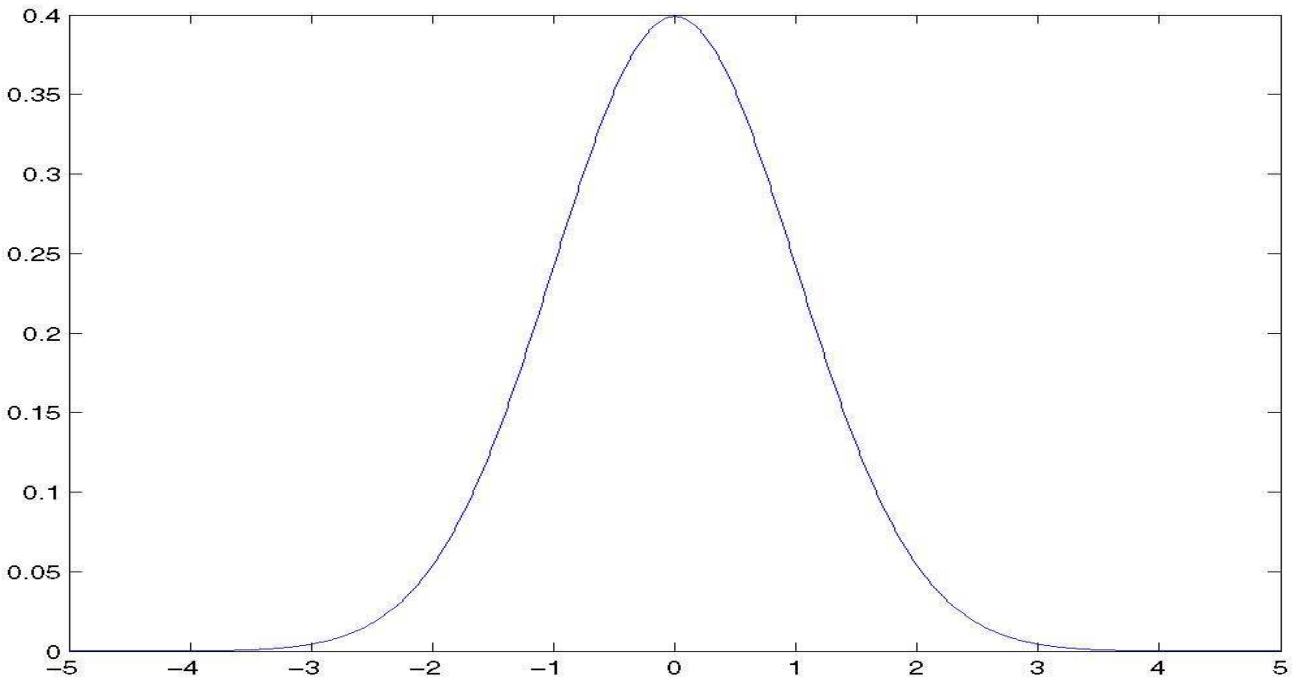
KTH VETENSKAP OCH KONST
KTH Matematik

Standard Normal distribution, $N(0,1)$

The distribution $N(0,1)$ density curve

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

is called standard normal density curve



KTH VETENSKAP OCH KONST
KTH Matematik

Standard Normal distribution, $N(0,1)$

Definition

A random variable Z is said to have standardized normal distribution if $Z \sim N(0,1)$, i.e., if it has the density/density curve

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Standard Normal distribution, $N(0,1)$

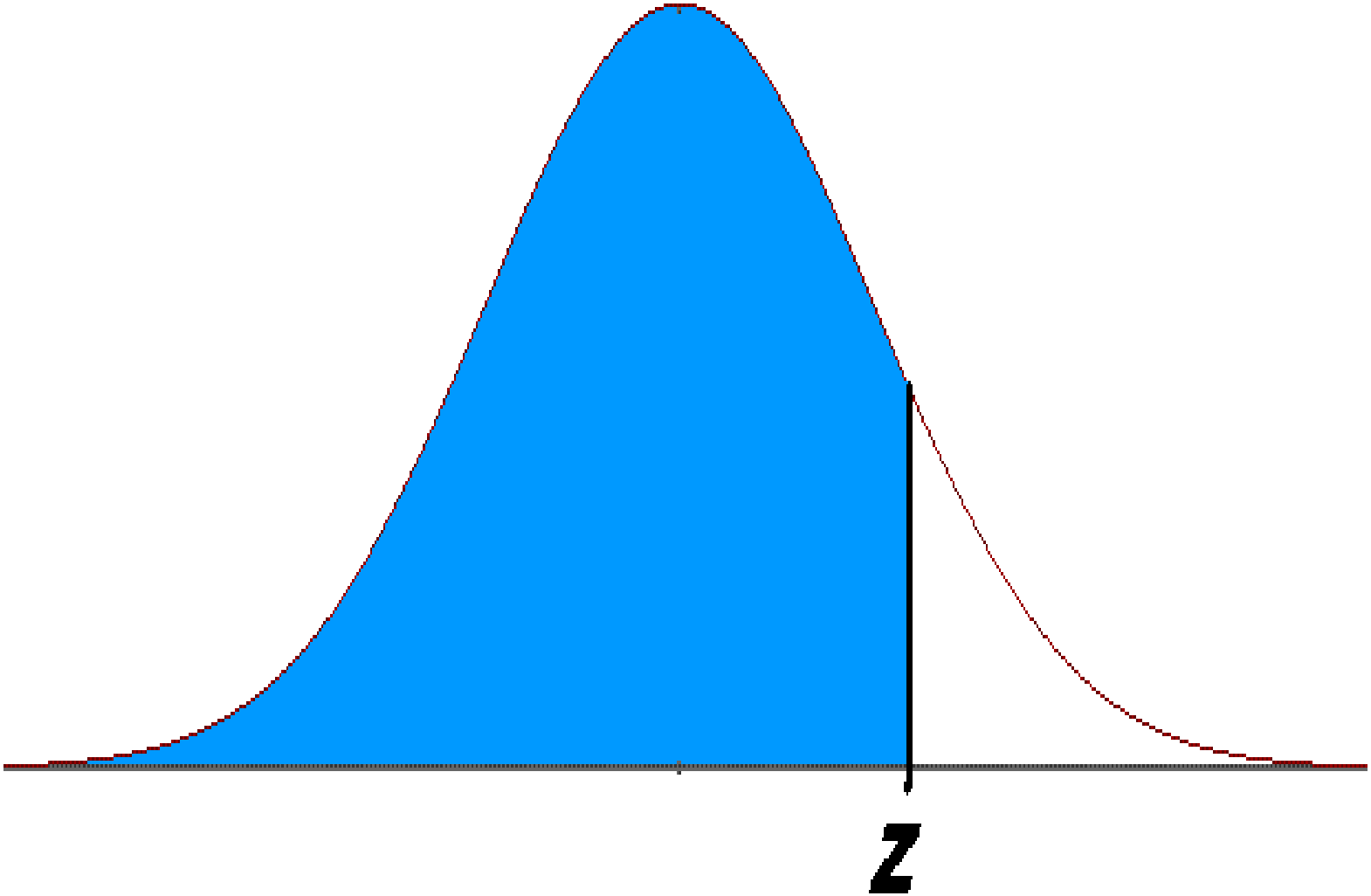
The area under the curve from $-\infty$ to z is denoted by $\Phi(z)$. The function $\Phi(z)$ is called the *cumulative density function* for standard normal distribution.

$$\Phi(z) = Pr(Z \leq z) = \int_{-\infty}^{z} \varphi(x)\, dx = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

KTH Matematik

Cumulative density function, $N(0,1)$

$$\Phi(z) = Pr(Z \leq z) = \int_{-\infty}^{z} \varphi(x)\, dx$$



Finding the Area Between Two Values

$Z \sim N(0,1)$. The function $\Phi(z)$ is given in tables for $z \geq 0$. If $z < 0$ we use the fact that $\Phi(-z) = 1 - \Phi(z)$. The probability of an outcome between a and b, $a < b$ is then $\Phi(b) - \Phi(a)$. (Why?) (Draw a figure to see this).

KTH Matematik

$Pr(a < Z \leq b) = \Phi(b) - \Phi(a)$

$$Pr(Z \leq b) = Pr(a < Z \leq b \quad \text{or} \quad Z \leq a)$$

The events $a < Z \leq b$ and $Z \leq a$ are mutually exclusive, the rule of addition gives

$$Pr(Z \leq b) = Pr(a < Z \leq b) + Pr(Z \leq a)$$

and by defintion of the cumulative density

$$\underbrace{Pr(Z \leq b)}_{=\Phi(b)} = Pr(a < Z \leq b) + \underbrace{Pr(Z \leq a)}_{=\Phi(a)}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Finding the Area Between Two Values: An Example

Example

$Z \sim N(0,1)$. Probability of $-0.5 < Z \leq 1.1$ equals
$\Phi(1.1) - \Phi(-0.5) = \Phi(1.1) - (1 - \Phi(0.5)) = \Phi(1.1) + \Phi(0.5) - 1 = 0.864 + 0.691 - 1 = 0.555$ by the table.

KTH Matematik

A Table for $Pr(Z \leq z)$

Standard normal, cumulative density, $P(Z<z)$

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.52
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.56
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.60
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.64
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.68
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.71
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.74
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.77
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.80
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.83
1	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.85
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.87

KTH VETENSKAP OCH KONST

KTH Matematik

Z-Score & Finding the Area Between Two Values

$X \sim N(\mu, \sigma)$. We look now at the Z-score

$$Z = \frac{X - \mu}{\sigma}.$$

It can be checked that $Z \sim N(0, 1)$. Furthermore

$$a < X \leq b \Leftrightarrow \frac{a - \mu}{\sigma} < Z \leq \frac{b - \mu}{\sigma}$$

are two events that are equivalent. Thus

$$Pr(a < X \leq b) = Pr(\frac{a - \mu}{\sigma} < Z \leq \frac{b - \mu}{\sigma})$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-x^2/2} dx$$

KTH Matematik

Z-Score & Finding the Area Between Two Values

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

$$Pr(a < X \leq b) = \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-x^2/2} dx$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

i.e., the final result is

$$Pr(a < X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

KTH Matematik

Finding the Area Between Two Values: Example

$X \sim N(28.0, 0.25)$.

$$Pr(27.5 < X \leq 28.5) = \Phi\left(\frac{28.5 - 28.0}{0.25}\right) - \Phi\left(\frac{27.5 - 28.0}{0.25}\right)$$

$$= \Phi(2) - \Phi(-2) = \Phi(2) - (1 - \Phi(2)) = 0.977 - (1 - 0.977) = 0.954$$

We found $\Phi(2)$ in the table (Biostatistics, Fall 2012).

KTH Matematik

Sums of Random Variables

A New Topic: Sums Random of Random Variables

Sums of Random Variables: Introduction

Illumina BeadArray platform is a microarray technology offering highly replicable measurements of gene expression in a biological sample. Each probe is measured on average of thirty to sixty beads randomly distributed on the surface of the array, avoiding spatial artifacts and the reported probe intensity is the robust mean of the bead measurements. Fluorescence intensity measured on each bead is subject to several sources of noise (non-specific binding, optical noise,).

KTH Matematik

Sums of Random Variables: Introduction

Thus the intensities produced by the microarray require a background correction in order to account measurement error. For that purpose, Illumina microarray design includes a set of non specific negative control probes which provides an estimate of the background noise distribution. In genome-wide microarrays, the observed intensity of a probe is usually modeled as the sum of a signal and a background noise. Namely, let X be the **observed intensity** of a given probe, we assume that

$$X = S + B$$

where S is the **true signal** which counts for the abundance of the probe complementary sequence in the target sample and is independent of the **background noise** B. Only X is observed but the quantity of interest is the signal S.

KTH Matematik

Sums of Random Variables: Introduction

$$X = S + B$$

It is often taken that

$$S \sim \text{Exp}(\lambda), B \sim N(\mu, \sigma)$$

and S and B are **independent**. E.g.,

Plancade, Sandra and Rozenholc, Yves and Lund, Eiliv: Generalization of the normal-exponential model: exploration of a more accurate parametrisation for the signal distribution on Illumina BeadArrays BMC bioinformatics, *13, p. 329 -, 2012*

KTH VETENSKAP OCH KONST
KTH Matematik

Sums of Random Variables. Q1: Independence ?

What do we mean by independence of S and B? The logical answer is that

$$Pr\left(a < B \leq b \quad \text{and} \quad c < S \leq d\right) = Pr\left(a < B \leq b\right) \cdot Pr\left(c < S \leq d\right)$$

for all numbers a, b, c and d.

KTH Matematik

Sums of Random Variables. Q2: Mean ?

$$X = S + B$$

The mean of X is found as

$$E(X) = E(S + B) = E(S) + E(B)$$

If now $S \sim \text{Exp}(\lambda)$ and $B \sim N(\mu, \sigma)$, then

$$E(X) = \frac{1}{\lambda} + \mu$$

KTH Matematik

Sums of Random Variables. Q3: Variance ?

$$X = S + B$$

The variance of X is found as (this is where we need independence for the formula to hold)

$$V(X) = V(S + B) = V(S) + V(B)$$

If now $S \sim \text{Exp}(\lambda)$ and $B \sim N(\mu, \sigma)$, then

$$V(X) = \frac{1}{\lambda^2} + \sigma^2$$

KTH Matematik

General Expressions for Mean of a Sum

$$E(X+Y) = E(X) + E(Y)$$
$$E(X-Y) = E(X) - E(Y)$$

KTH Matematik

General Expressions for Variance of a Sum

If X and Y are independent:

$$V(X+Y) = V(X) + V(Y)$$
$$V(X-Y) = V(X) + V(Y).$$

KTH Matematik

General Expressions for Mean of a Sum

$X_1, \ldots, X_n$

$$Y = c_1 X_1 + \ldots + c_n X_n + a.$$

then

$$E(Y) = c_1 E(X_1) + \ldots + c_n E(X_n) + a$$

KTH Matematik

General Expressions for Mean of a Sum

$X_1, \ldots, X_n$ are independent and

$$Y = c_1 X_1 + \ldots + c_n X_n + a.$$

then

$$V(Y) = c_1^2 V(X_1) + \ldots + c_n^2 V(X_n)$$

Observe the disappearance of a. Why is this ?

KTH Matematik

Application: Arithmetic Mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$X_1, X_2, \ldots, X_n$ are independent have the same distribution (=equally distributed) (e.g., $X_1 \sim N(\mu, \sigma), X_2 \sim N(\mu, \sigma), \ldots, X_n \sim N(\mu, \sigma)$) with mean μ and standard deviation σ. Then

$$E(\bar{X}) = \mu, \quad V(\bar{X}) = \frac{\sigma^2}{n} \quad \textit{and} \quad D(\bar{X}) = \frac{\sigma}{\sqrt{n}}\,.$$

Found from the preceding with $c_1 = \ldots = c_n = \frac{1}{n}$ and $a = 0$.

KTH Matematik

Limits

A New Topic: Law of Large Numbers and the Central Limit Theorem. Although we shall discuss a theorem, we do not include proofs.

Law of Large Numbers

$X_1, X_2, \ldots, X_n$ are independent have the same distribution (=equally distributed) with mean μ and standard deviation σ.
If n is very large we have

The Law of Large Numbers

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mu$$

where μ is the common mean of $X_1, X_2, \ldots, X_n, \ldots$
This is because

$$V(\bar{X}) = \frac{\sigma^2}{n} \to 0$$

as $n \to \infty$

KTH VETENSKAP OCH KONST
KTH Matematik

Central Limit Theorem

We have invested quite a lot of time and effort in the normal distribution. That would have been questionable if it were not so that the normal distribution is frequently present.

The **Central Limit Theorem** *is the main argument for the normal distribution. This is one of the most important results in probability and statistics.*

KTH Matematik

Central Limit Theorem

Let $X_1, X_2, \ldots$ be independent and have the same distribution with mean μ and standard deviation σ. Then

$$\sum_{i=1}^{n} X_i \text{ approximatively} \sim N\left(n\mu, \sigma\sqrt{n}\right)$$

KTH Matematik

Central Limit Theorem: Arithmetic Mean

A usual application of the central limit theorem in biostatistics is on the arithmetic mean.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We get

$$\bar{X} \text{ approximatively} \sim N(\mu, \sigma/\sqrt{n})$$

Or, with with the cumulative density $\Phi(x)$

$$P(a < \bar{X} \leq b) \approx \Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right)$$

if n is big enough.

KTH Matematik

Central Limit Theorem

It is regrettably not possible to give general rules for how large n must be in order that the normal approximation will be applicable. This depends on how like normal the individual variables X_k are. A rule of thumb is that if X_ks have roughly symmetric density curves, then relatively small values of n will do, say some ten or twenty. If X_ks have very skew density curves, then n must be hunreds.

KTH Matematik

Central Limit Theorem

In particular, there should not be some X_k which is very dominating. There may, e.g., be a dominating source of error in some measurement process, which sees to it that normal approximation will not be applicable.

KTH Matematik

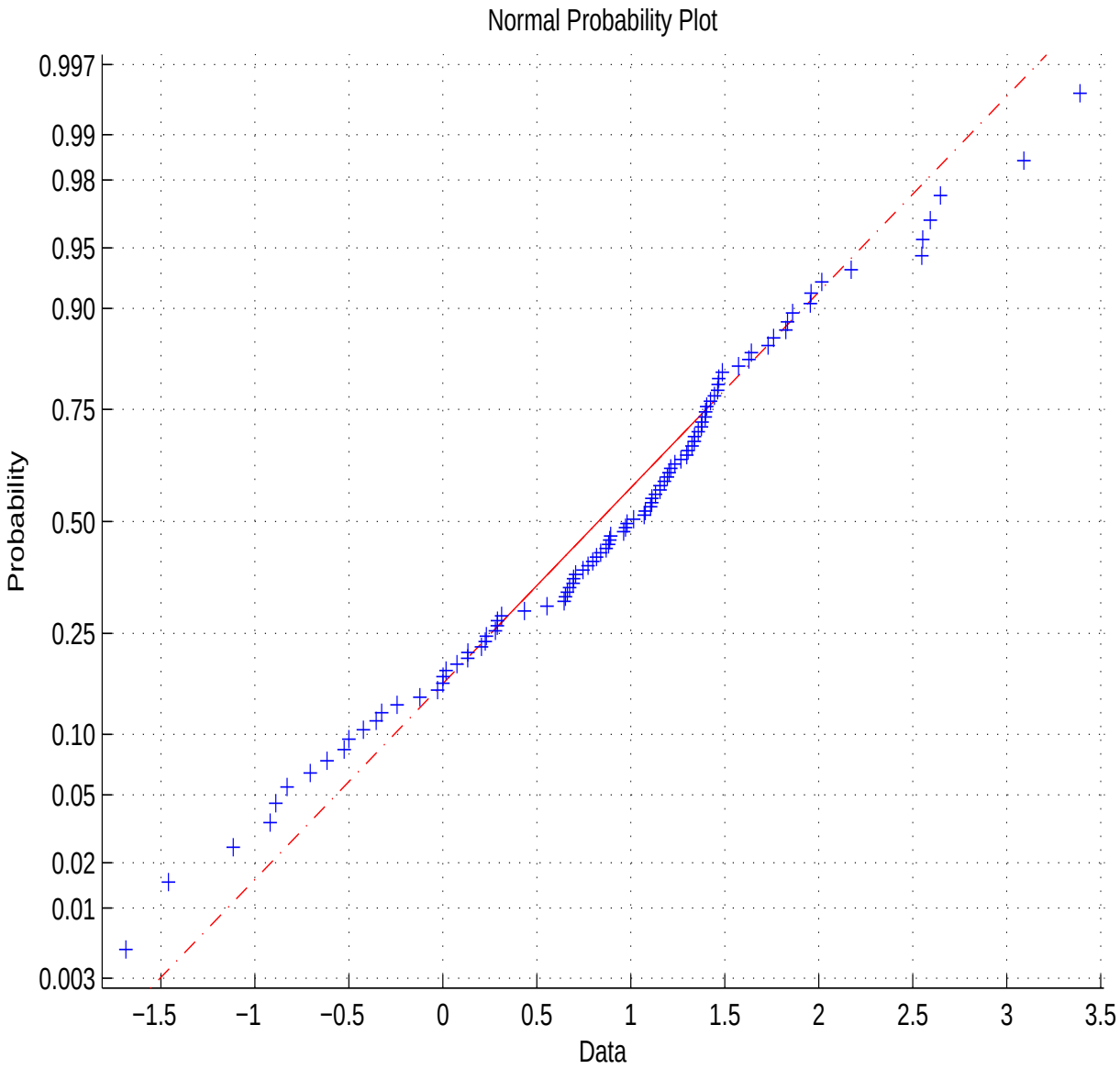
Assessing Normality

A **normal quantile plot** *or a* **normal probability plot** *The data are plotted against a theoretical normal distribution (expected z-scores $*$) in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality.*

It is a very messy thing to explain $*$, I trust a software demonstration.

KTH Matematik

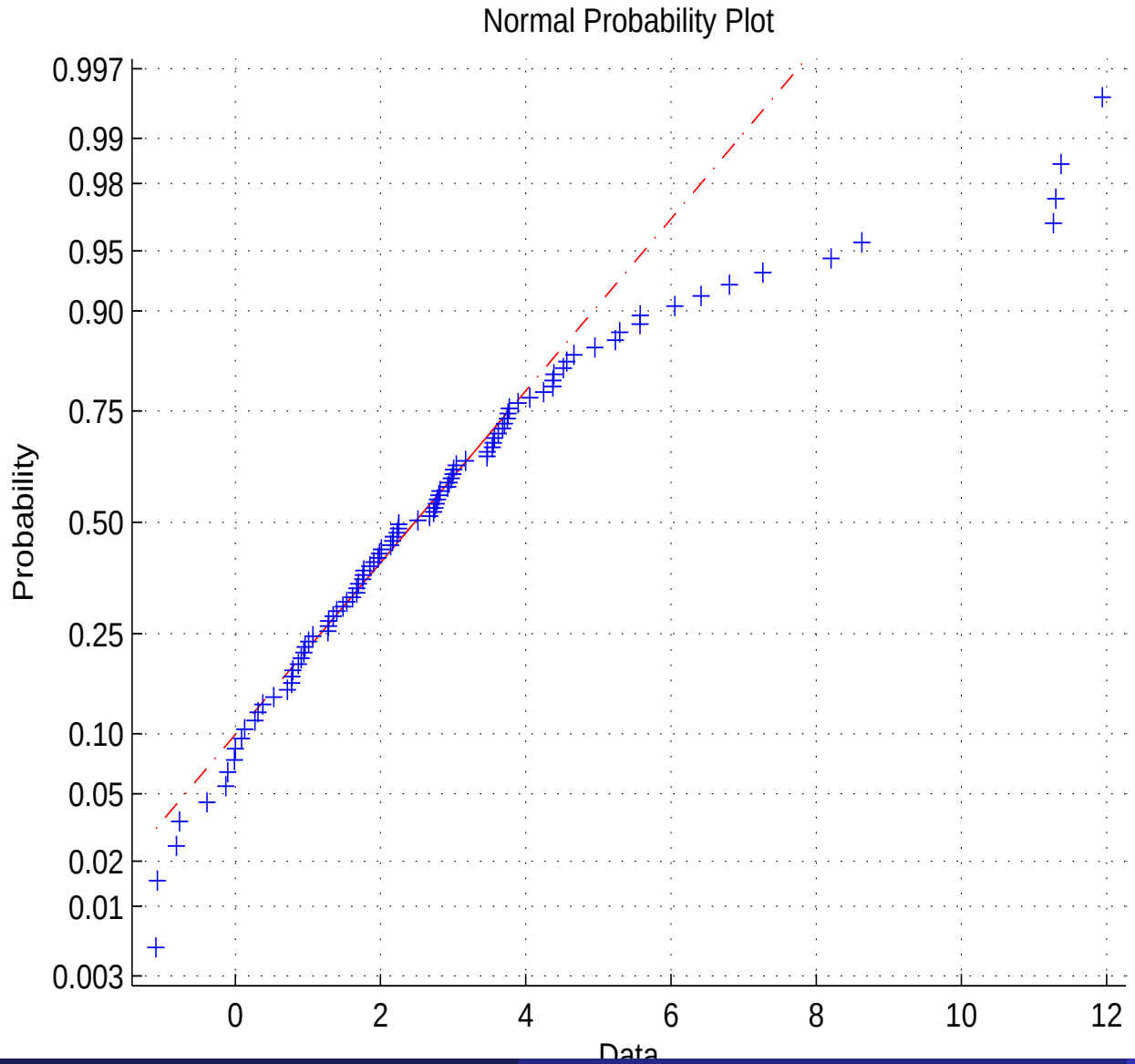
Normal Probability plot for $N(1,1)$



KTH VETENSKAP OCH KONST
KTH Matematik

Normal Probability Plot for $X = S + B$

$S \sim \text{Exp}(2)$, $B \sim N(1, 1)$



Bin(n, p) approximatively $N(np, \sqrt{npq})$

If $X \sim \text{Bin}(n, p)$ *and* $np(1-p) \geq 10$ *then* X *is approximatively* $N(np, \sqrt{npq})$.

This means that

$$\left.\begin{array}{l} P(X \leq k) \\ P(X < k) \end{array}\right\} \approx \Phi\left(\frac{k - np}{\sqrt{npq}}\right).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Multiplication & Lognormal Distribution

$X_1, X_2, \ldots, X_n$ are independent and have the same distribution and are positive. Suppose

$$Y = X_1 \cdot X_2 \cdots X_n$$

Then central limit theorem suggests that the logarithm

$$\ln Y = \ln X_1 + \ln X_2 + \ldots + \ln X_n$$

is approximatively normal. We say that Y is approximately **lognormal** distributed, since its logarithm is normally distributed.

KTH VETENSKAP OCH KONST
KTH Matematik

Lognormal distribution and Microarrays

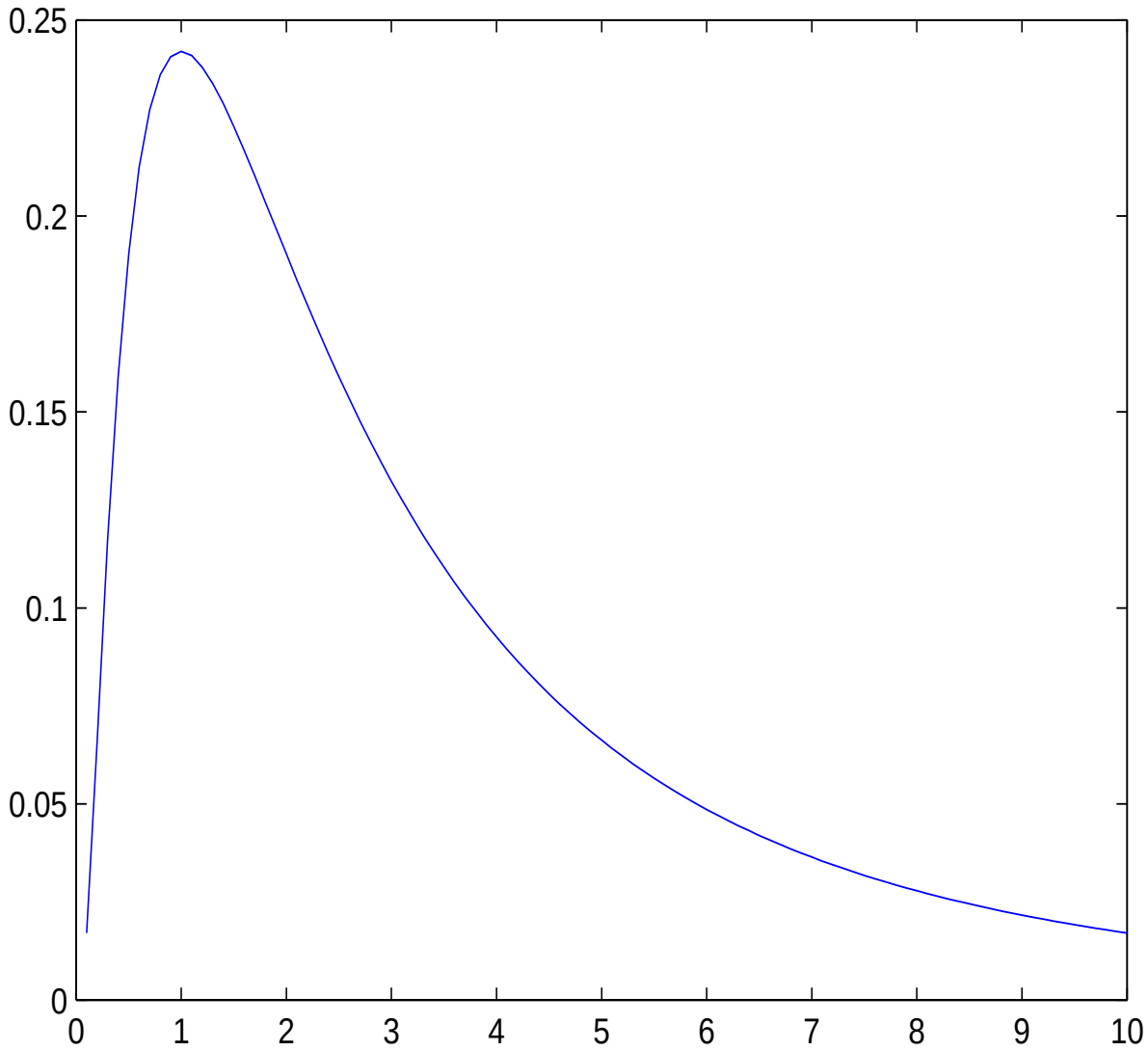
Hoyle, D.C. and Rattray, M. and Jupp, R. and Brass, A: Making sense of microarray data distributions, Bioinformatics, *18, 4, pp. 576−584, 2002*

Establishes that useful biological findings may result from analyzing microarray data at the level of entire intensity distributions. The distribution of the bulk of microarray spot intensities is well approximated by a log-normal.

KTH VETENSKAP OCH KONST
KTH Matematik

Lognormal density curve: a skew density

The associated normal distribution is $N(1, 1)$.



Uniform Random Variables

Then $X \sim \mathrm{U}(a, b)$, uniformly distributed if its density is with $a < b$

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{elsewhere.} \end{cases}$$

$$E(X) = \frac{a+b}{2}, V(X) = \frac{(b-a)^2}{12}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Uniform Random Variables

Simulate hundred uniform $U(0,1)$ variables and multiply them

$$Y = U_1 \cdot U_2 \cdot \ldots \cdot U_{100}$$

Take $\log(Y)$. Repeat thousand times. Then we have thousand values $\log(Y)$ and we expect this to be normally distributed.

KTH Matematik

Normal Probability plot for $\log(Y)$