AESB2440: Geostatistics & Remote Sensing

Lecture 5: Distributions

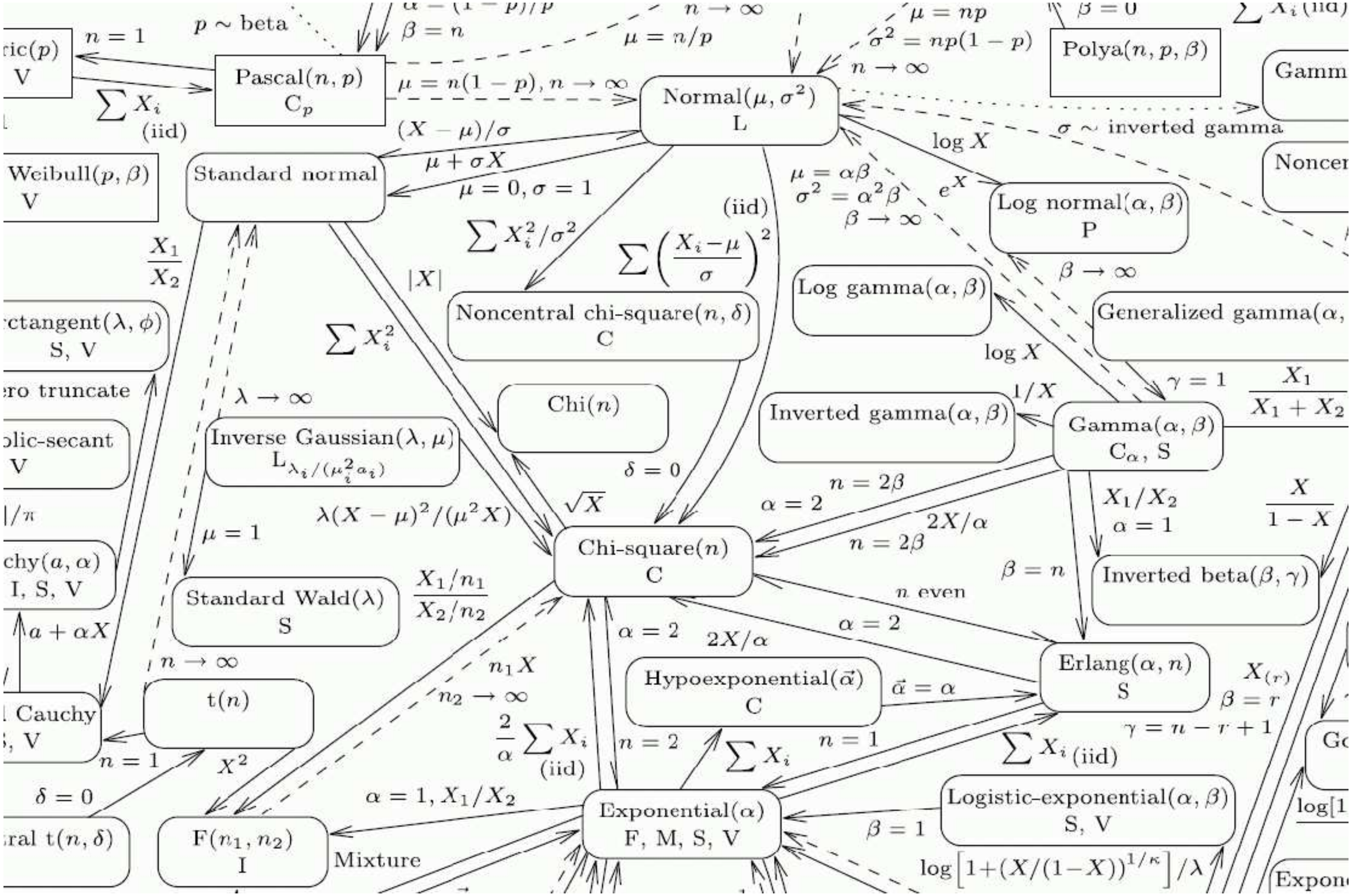
Wednesday, April 29, 2015

Roderik Lindenbergh

Dept. of Geoscience & Remote Sensing

Delft University of Technology

Different distributions



Source http://www.math.wm.edu/~leemis/2008amstat.pdf

Lecture topics

Distribution parameters

Discrete distributions

- Binomial distribution
- Bernoulli distribution

Continous distributions

- Uniform distribution
- Normal distribution
- Mean and Standard deviation
- Exponential distribution

Expectation

- Discrete
- Continuous

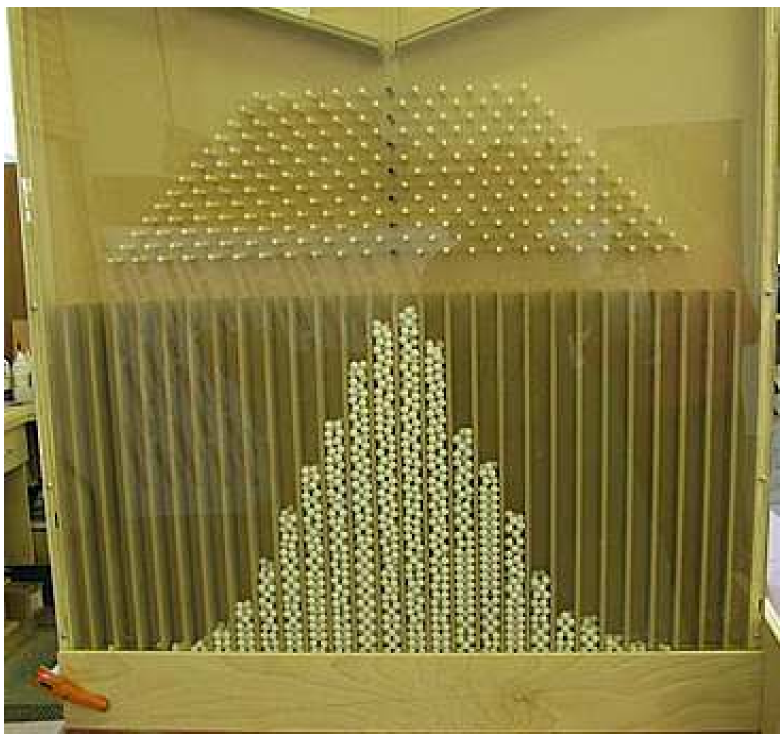
Change of variable

- Linear transformation
- From normal to standard normal

Multivariate statistics

- Joint probability
- Random vector
- Covariance
- Correlation
- Rank statistics

A. Specific Distributions



notation:	$\mathcal{N}(\mu, \sigma^2)$
parameters:	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 \geq 0$ — variance (squared scale)
support:	$x \in \mathbf{R}$ if $\sigma^2 > 0$ $x = \mu$ if $\sigma^2 = 0$
pdf:	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
cdf:	$\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$
mean:	μ
median:	μ
mode:	μ
variance:	σ^2
skewness:	0
ex.kurtosis:	0
entropy:	$\frac{1}{2}\ln(2\pi e\, \sigma^2)$
mgf:	$e^{\mu t + \frac{1}{2}\sigma^2 t^2}$
cf:	$e^{i\mu t - \frac{1}{2}\sigma^2 t^2}$
Fisher information:	$\begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$

Source http://ptrow.com/articles/Galton$_$June$_$07.htm

Recall

Question: what is a distribution function?

Question: what is a probability density function?

Question: what is a probability mass function?

Question: what is the difference between a probability density function and a probability mass function?

Question: what is the relation between a distribution function and its corresponding probability density or mass function?

Random exam

Suppose you totally unprepared attend a multiple choice exam
Each of 10 questions only allows the answer YES or NO.
If your answer is correct you obtain a point.

The random variable X_E equals your total number of points:

$$X_E := \{\text{Number of correct answers}\}.$$

Question: What is $P(X_E = 0)$?
Question: What are generalizations of this problem?
Question: so, what could be parameters describing this problem?

Probability of k correct answers

Question: What is $P(X_E = 1)$?

$P(X_E = 1) = \frac{1}{2} \cdot (\frac{1}{2})^9 \cdot 10,$ or,

(probability that an answer is correct)

$*$

(probability that the other answers are wrong)

$*$

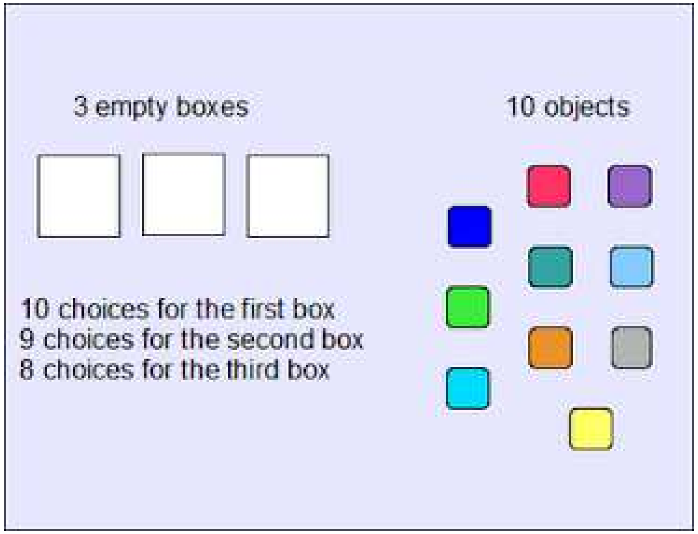
(number of scenarios)

which we generalize to

$$P(X_E = k) = (\frac{1}{2})^k \cdot (\frac{1}{2})^{(10-k)} \cdot C_{10,k}$$

with $C_{10,k}$ the number of scenarios,
i.e. the number of ways to pick k correct answers from a list of 10.

Number of picks



Consider $C_{10,3}$, the number of ways to pick 3 questions from a list of 10.

If order matters, you have $10 \cdot 9 \cdot 8$ possibilities. Otherwise you have to compensate for the double counting of, say, $Q_1 Q_3 Q_2$ and $Q_2 Q_3 Q_1$.

You can order three questions in $3 \cdot 2 \cdot 1$ ways.

$\Rightarrow$ In total there are $\frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdots 1}$ choices of three questions.

Number of possibilities to pick a subset of size k from a set with n elements:

$$\binom{n}{k} := \frac{n(n-1)\dots(n-(k-1))}{k(k-1)\dots 2 \cdot 1} = \frac{n!}{k!(n-k)!}$$

Source http://mymathnotebook.com/

Probability of k good answers

Conclusion: with $n = 10$

$$P(X_E = k) \;=\; \binom{n}{k} \cdot (\frac{1}{2})^k \cdot (\frac{1}{2})^{(n-k)}$$

Question: what values of k make sense?

Question: what is the probability of outcome six (just passed!)?

Question: how would the formula above change in case the candidate could choose from four answers A, B, C or D for each question?

Question: how is four instead of two answers affecting the probability of outcome six?

Question: what is the probability of outcome six in this case?

Dept. of Geoscience & Remote Sensing

Binomial distribution

A discrete random variable X has a Binomial distribution with parameters p and n, with

$$0 \leq p \leq 1, \quad \text{and} \quad n = 1, 2, \ldots$$

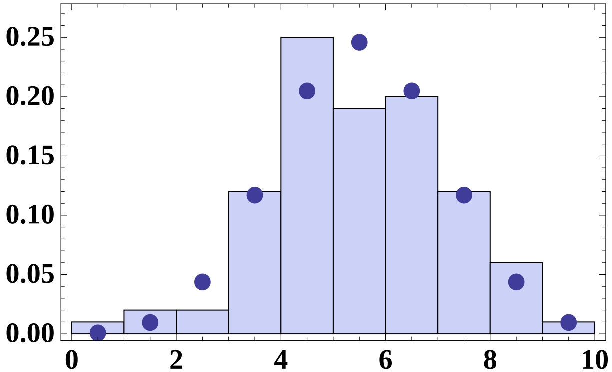
if its probability mass function is given by

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

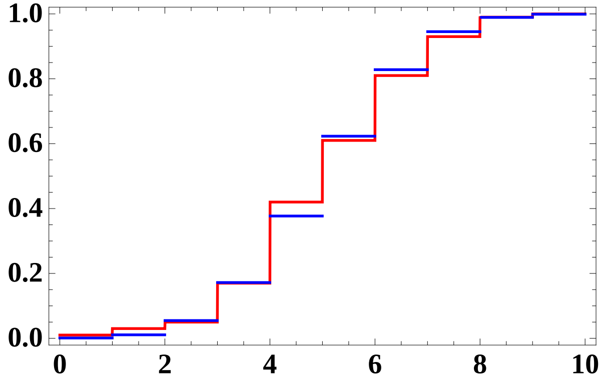
for $k = 0, 1, \ldots, n$.

Notation: $X \sim \text{Bin}(n, k)$

Question: what is in the figures?
Question: what is the chance on passing the exam?

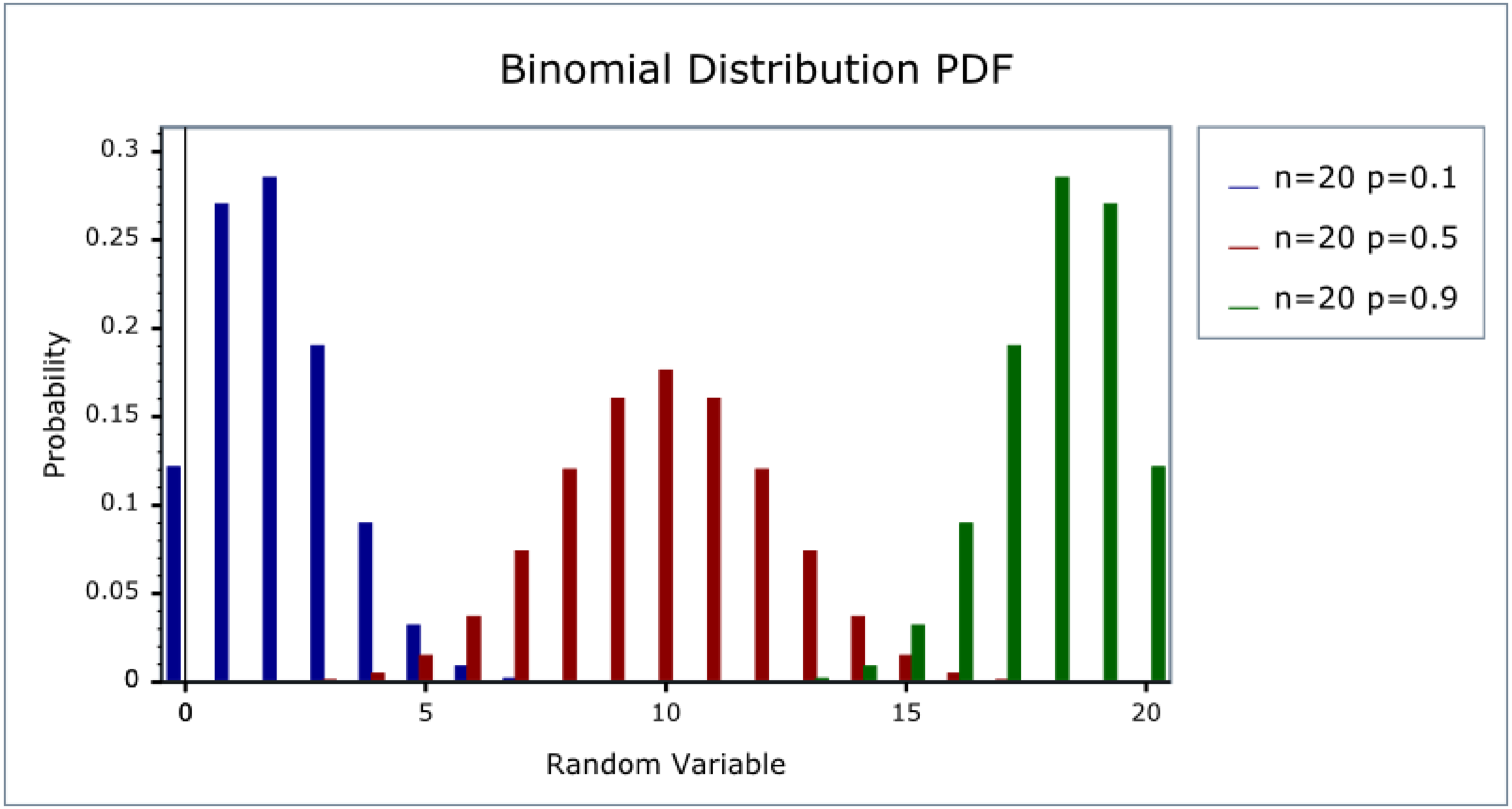


p=.5; n=10, # samples = 100; Dots: theory



p=.5; n=10, Red: # samples = 100; Blue: theory

Examples of binomial distributions



Source http://www.boost.org/doc/libs/1$_$41$_$0/libs/math/doc/sf$_$and$_$dist/html/math$_$toolkit/dist/dist$_$ref/dis

Uniform Distribution

A continuous random variable X has a uniform distribution on the interval $[a, b]$ if its probability density function f is given by

$$f(x) = \begin{cases} 0, & x \text{ not in } [a, b] \\ \frac{1}{b-a}, & \text{for } a \leq x \leq b \end{cases}$$

Notation: $X \sim U(a, b)$

Question
How do

- the probability density function, and the
- cumulative distribution function

of the distribution $U(0, 12)$ look like?

Exponential Distribution

See Exercises

Dept. of Geoscience & Remote Sensing

B. Normal Distributions

THEORIA

MOTVS CORPORVM

COELESTIVM

IN

SECTIONIBVS CONICIS SOLEM AMBIENTIVM

AVCTORE

CAROLO FRIDERICO GAVSS

HAMBVRGI SVMTIBVS FRID. PERTHES ET I. H. BESSER
1809.

Venditur

PARISIIS ap. Treuttel & Würtz. LONDINI ap. R. H. Evans.
STOCKHOLMIAE ap. A. Wiborg. PETROPOLI ap. Klostermann.
MADRITI ap. Sancha. FLORENTIAE ap. Molini, Landi & C°
AMSTELODAMI in libraria: Kunst- und Industrie-Comptoir, dicta.

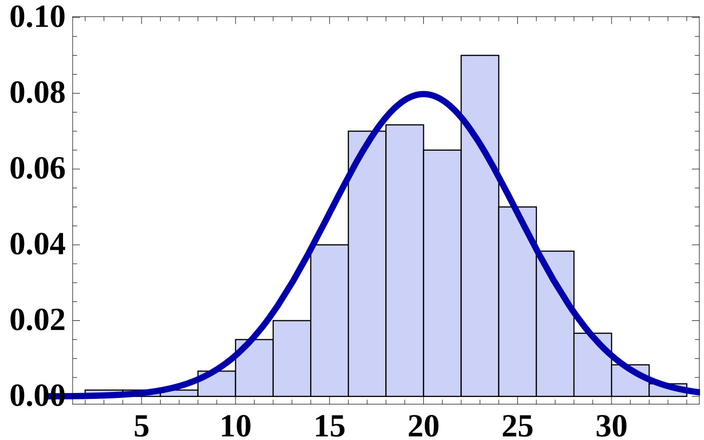
Dept. of Geoscience & Remote Sensing

Normal distribution

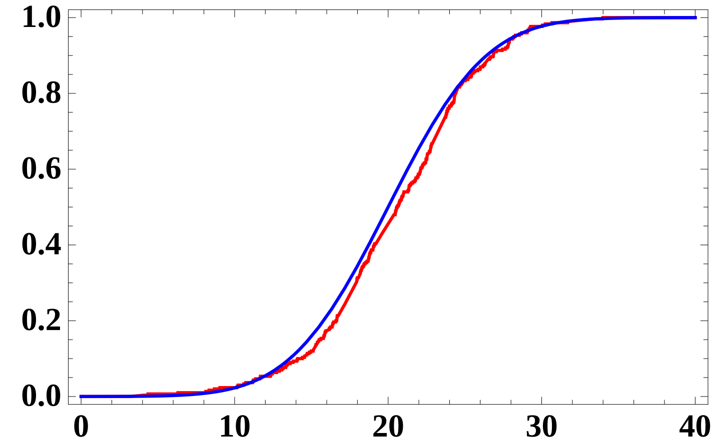
A continuous random variable X has a normal distribution with parameters μ, its mean, and σ^2, its standard deviation, if its probability density function f is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad \text{for} \qquad -\infty \leq x \leq \infty$$

Notation: $X \sim N(\mu, \sigma^2)$



PDF for N(20,5), EPDF with ♯ = 400



CDF for N(20,5); ECDF for ♯=400

Dept. of Geoscience & Remote Sensing

Standard normal distribution

The standard normal distribution is the normal distribution with parameters $\mu = 0$ and $\sigma = 1$.

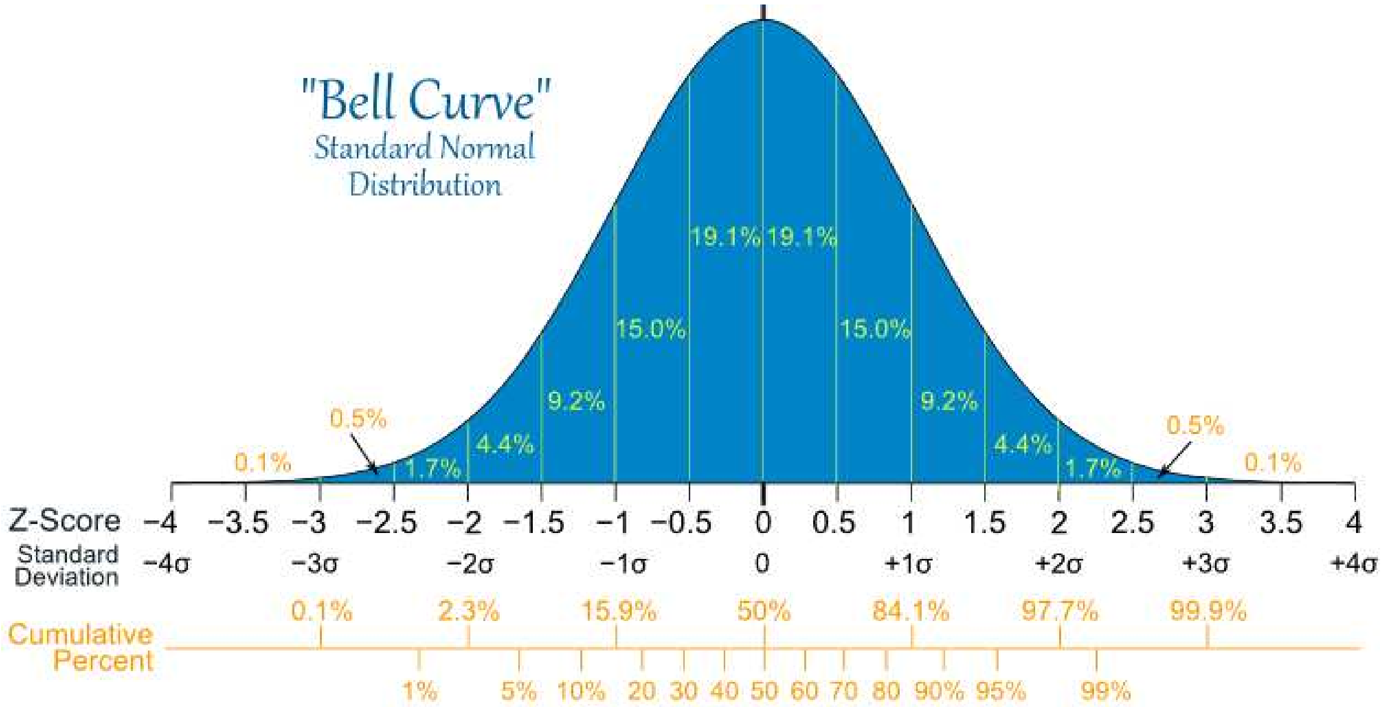
Question. What is $\Phi(x)$, the P(robability) D(ensity) F(unction) of N(0,1)?

Remark. $\Phi(x) = \Phi(-x)$.

Question. What is $\int \Phi(x) dx$?

Dept. of Geoscience & Remote Sensing

Standard normal + standard deviation

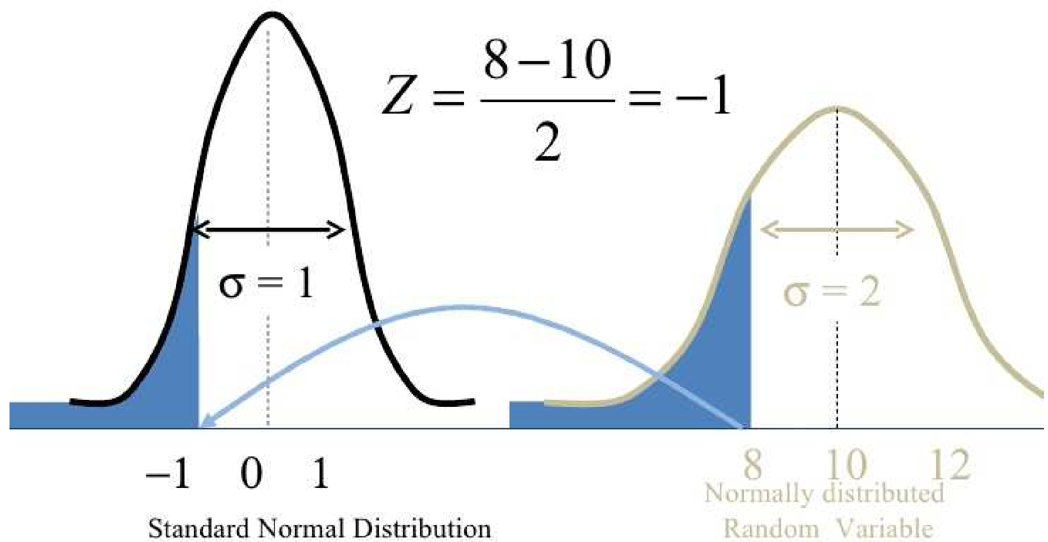


Question: What percentage of points is within 1σ, 2σ and 3σ of the mean?

Source https://www.mathsisfun.com/data/standard-normal-distribution.html

Arbitrary normal → standard normal

This transformation allows us to use the standard normal distribution and the tables of probabilities for the standard normal table to find out the appropriate probability. The Z transformation tells us the 8 on the original distribution is equivalent to -1 on the standard normal distribution. So, the area under the standard normal distribution to the left of -1 represents the same probability as the area under the original distribution to the left of 8.

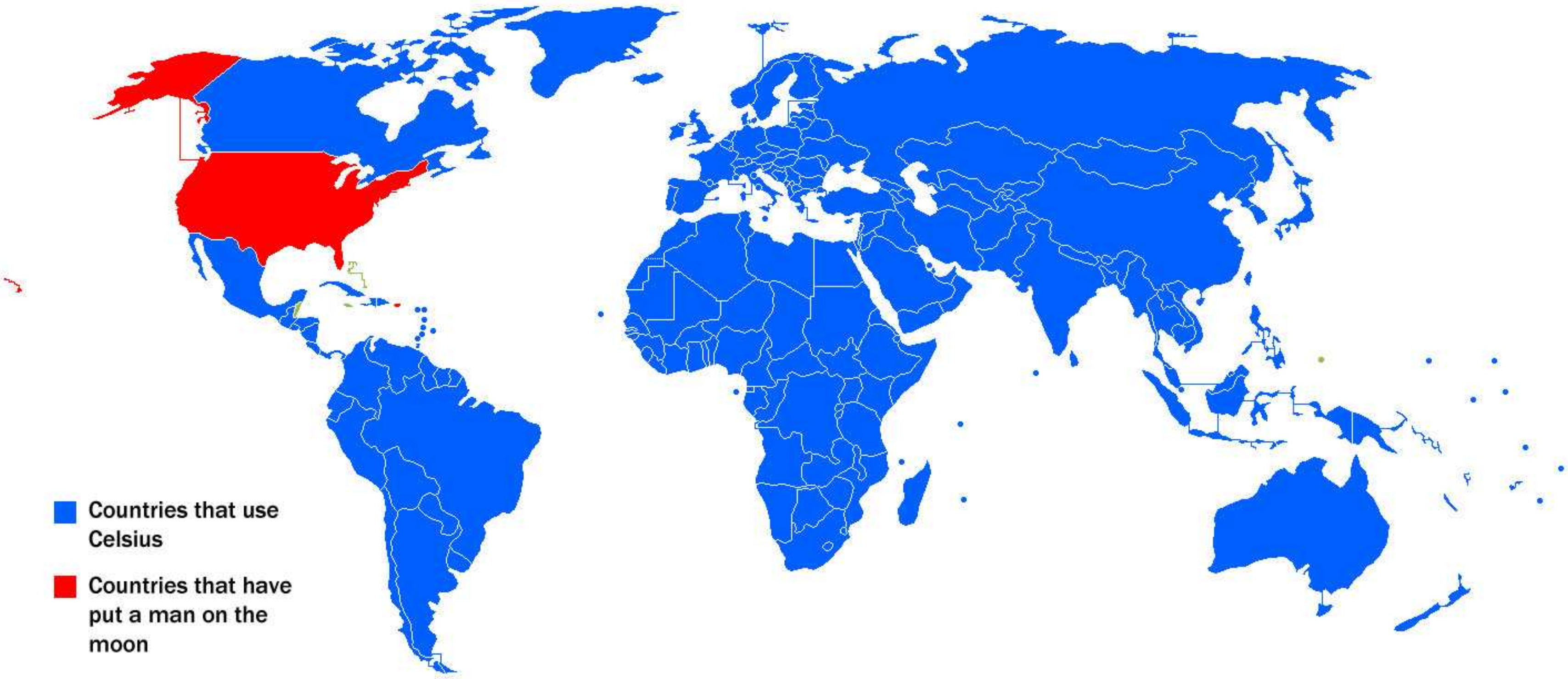


Idea. Compute probabilities for given normal distribution from standard normal distribution:

- Map given mean on standard mean
- Map given standard deviation on standard normal standard deviation

Source http://www.slideshare.net/kkong/demonstration-of-a-z-transformation-of-a-normal-distribution

Celcius and Fahrenheit

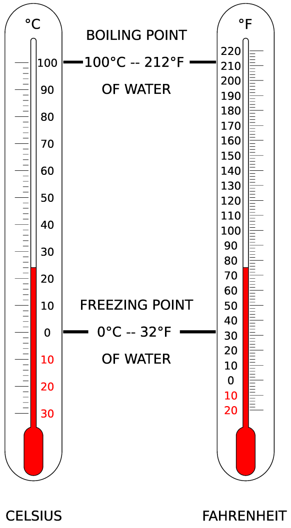


Source http://i.imgur.com/r3xPjAR.jpg

Transformation Fahrenheit-Celcius

$$
\begin{aligned}
X &\quad - \quad \text{Temperature in degrees Celcius} \\
Y &\quad - \quad \text{Temperature in degrees Fahrenheit} \\
Y &\quad = \quad \frac{9}{5}X + 32
\end{aligned}
$$

F_X, F_Y Distribution functions of X and Y



$$
\begin{aligned}
F_Y(a) &= P(Y \leq a) = P(\frac{9}{5}X + 32 \leq a) \\
&= P(X \leq \frac{5}{9}(a - 32)) = F_X(\frac{5}{9}(a - 32))
\end{aligned}
$$

Differentiating to densities:

$$
f_Y(y) = \frac{5}{9} f_X(\frac{5}{9}(y - 32))
$$

Question: What is the type of relation between Celcius and Fahrenheit?

Dept. of Geoscience & Remote Sensing

Transformation to standard normal

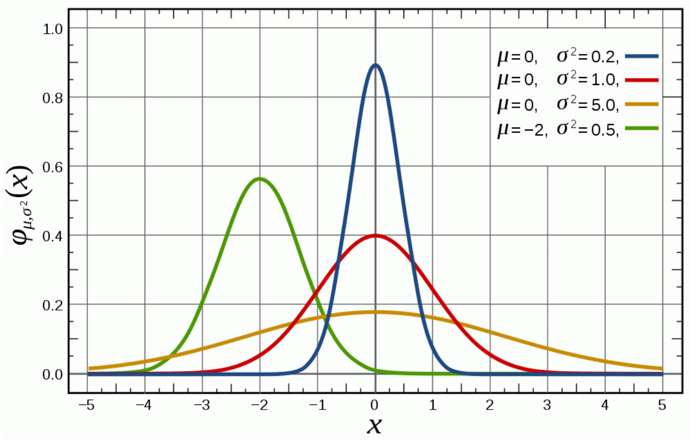
A probability $P(x_1 < X < x_2)$ for a normal distribution $X \sim N(\mu, \sigma)$ can be expressed in terms of the standard normal distribution $Z = N(0,1)$.

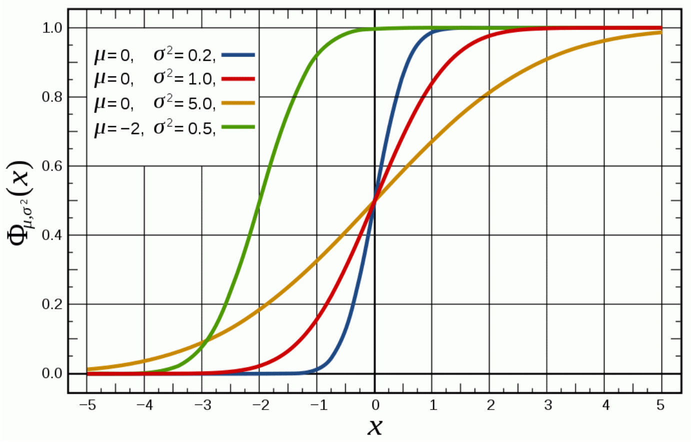
Let

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

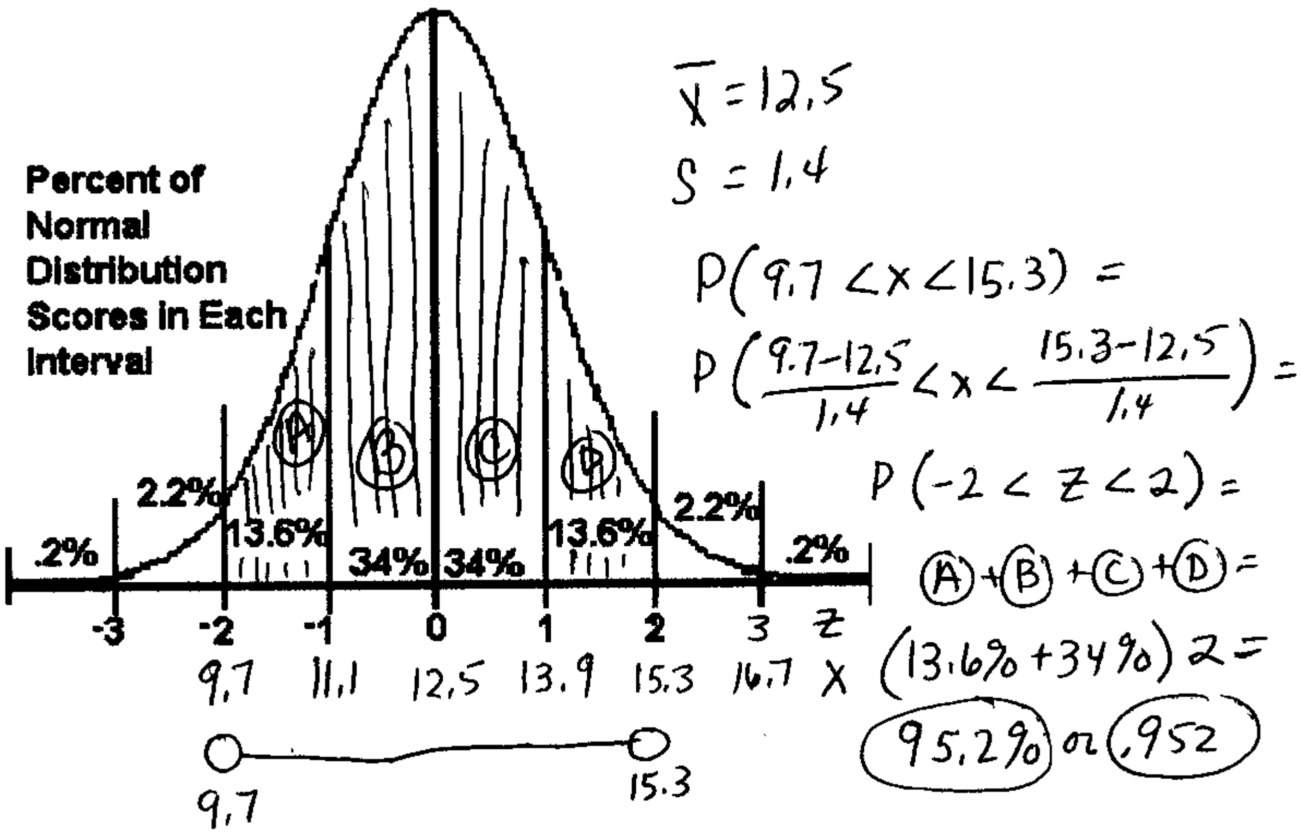
Then, with $z_{1,2} = \frac{x_{1,2} - \mu}{\sigma}$,

$$P(x_1 < X < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz = P(z_1 < Z < z_2)$$





Example: probs for arbitrary normal



Source http://www.mathnstuff.com/math/spoken/here/2class/90/standrd.htm

C. Expectation

Source http://thepetitegeek.blogspot.nl/2010/06/reading-groups.html

Expectation

The expectation of a continuous random variable X with prob. density function f is the number

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

Question. What is the discrete equivalent?

Question. What is $E\{X\}$ when f is an even function?

Question. What is an example of an even probability density function?

Well-known expectations

Question. What is the expectation of the

1. Uniform distribution?

2. Normal distribution?

3. The discrete Bernoulli distribution:

$R(p)$ has a Bernoulli distribution with parameter p, if

$$\begin{aligned} R &= 1, \quad && \text{with probability } p \\ R &= 0, \quad && \text{with probability } 1-p. \end{aligned}$$

4. (Binomial distribution?)

Change Of Variable Formula

Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a function and let X be a random variable.

If X is continous, with probability density function f_X, then

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

Question. What is $E\{g(X)\}$, if g is linear?

Write $g(X) = rX + s$ and apply integral formula above...

This is the linearity of expectation.

Binomial Expectation

Recall. The discrete equivalent, i.e. the expectation of $E\{g(X)\}$, in case X is a discrete random variable, taking values $a_1, a_2, \ldots, a_n$ is given by: $E\{g(X)\} = \sum_{i=1}^{n} g(a_i)P(X = a_i)$

What is the expectation of the Binomial distribution?

Any Bin(n, p) distribution can be written as

$$X = R_1 + R_2 + \ldots R_n$$

where the R_i are independent ‘Coin‘(p) distributions, that is: $R_i = 1$ with probability p, and 0 with probability $1 - p$. (Official name: *Bernoulli distribution*). As

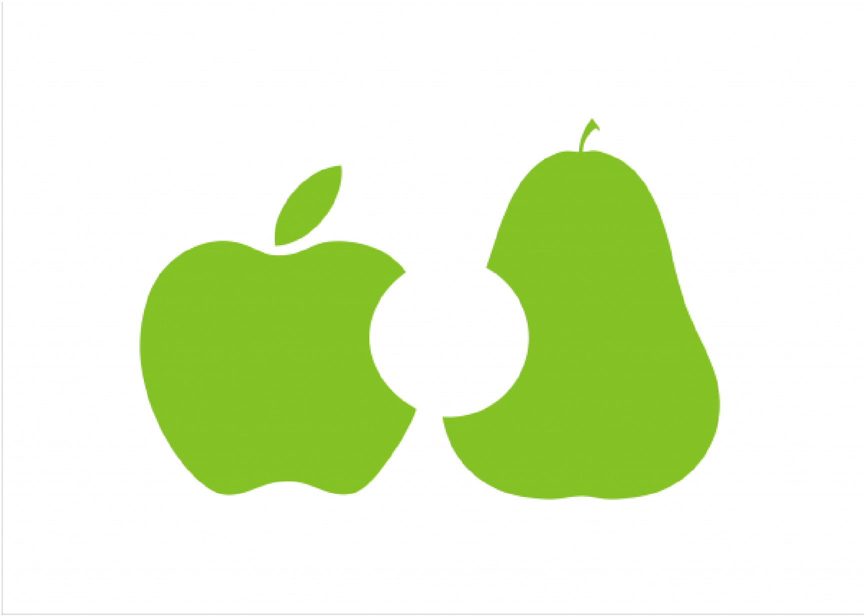
$$E(R_i) = 0 \cdot (1 - p) + 1 \cdot p = p,$$

the linearity of expectation gives: $E(X) = n \cdot p$.

Question: What is the expected mark of the guessing student?

- In case of two answers, A and B?
- In case of four answers, A, B, C or D?

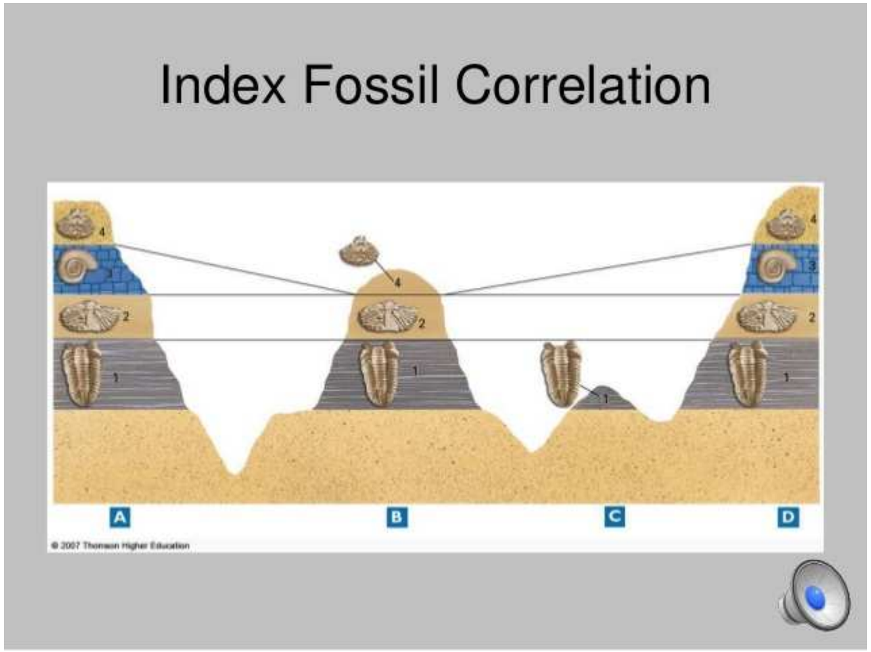
D. Relating attributes



TU Delft

Attributes

Attribute example: amount or precense of guide fossils in geological layers

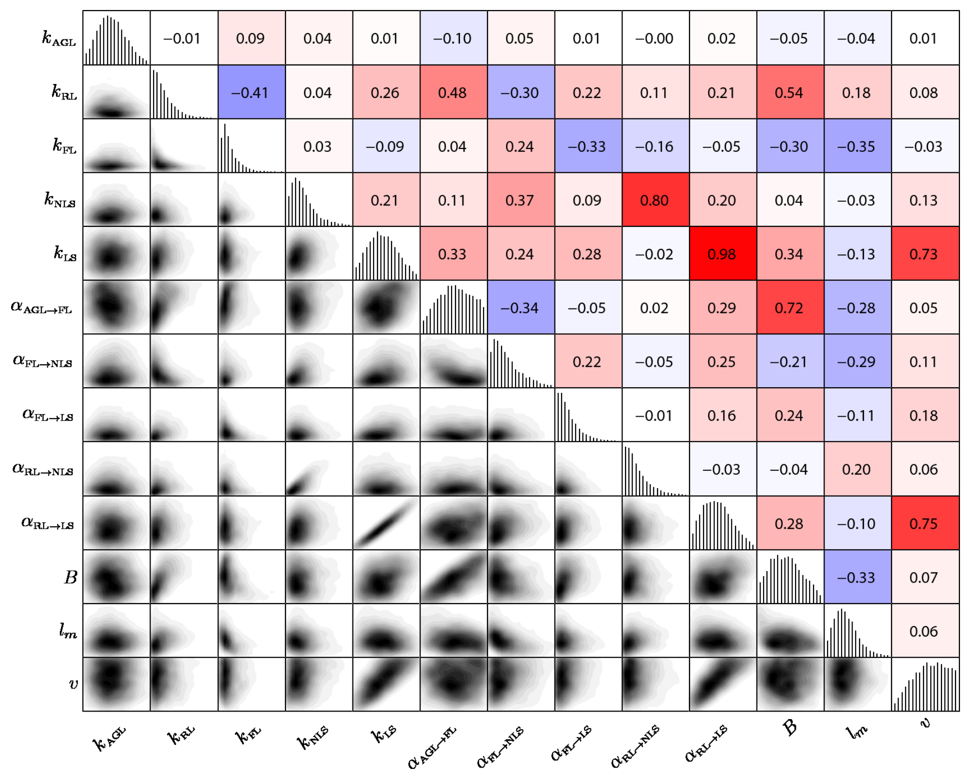


Idea: link layers having similar fossile contents

Source http://www.slideshare.net/cooperk2/guide-to-rock-dating-chap-4

Correlation Matrix

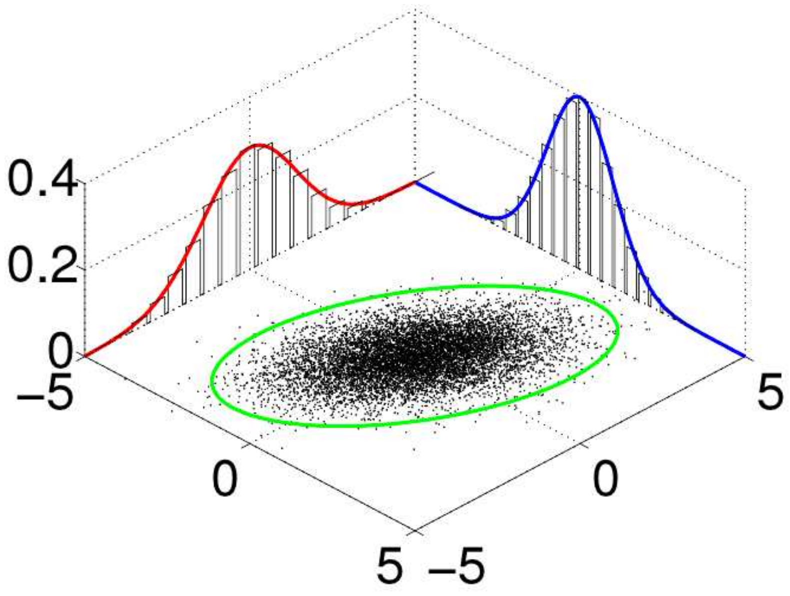
Goal: quantify information sharing between different attributes.



Idea: very red or blue entries correspond to pairs of attributes that carry very similar information.

Source https://www.bgc-jena.mpg.de/bgi/index.php/People/MaartenBraakhekke

Joint distributions



Source http://en.wikipedia.org/wiki/Joint$_$probability$_$distribution

Random vector + Joint distribution

A random vector $\mathbf{X} : \Omega \to I\!R^n$ is a number of random variables:

$$\mathbf{X} = \{X_1, X_2, \ldots, X_n\}^T$$

GPS Example. The random vector $\mathbf{X} = \{N, E, H\}$ consists of the random variables N, North, E, East, and, H, height.

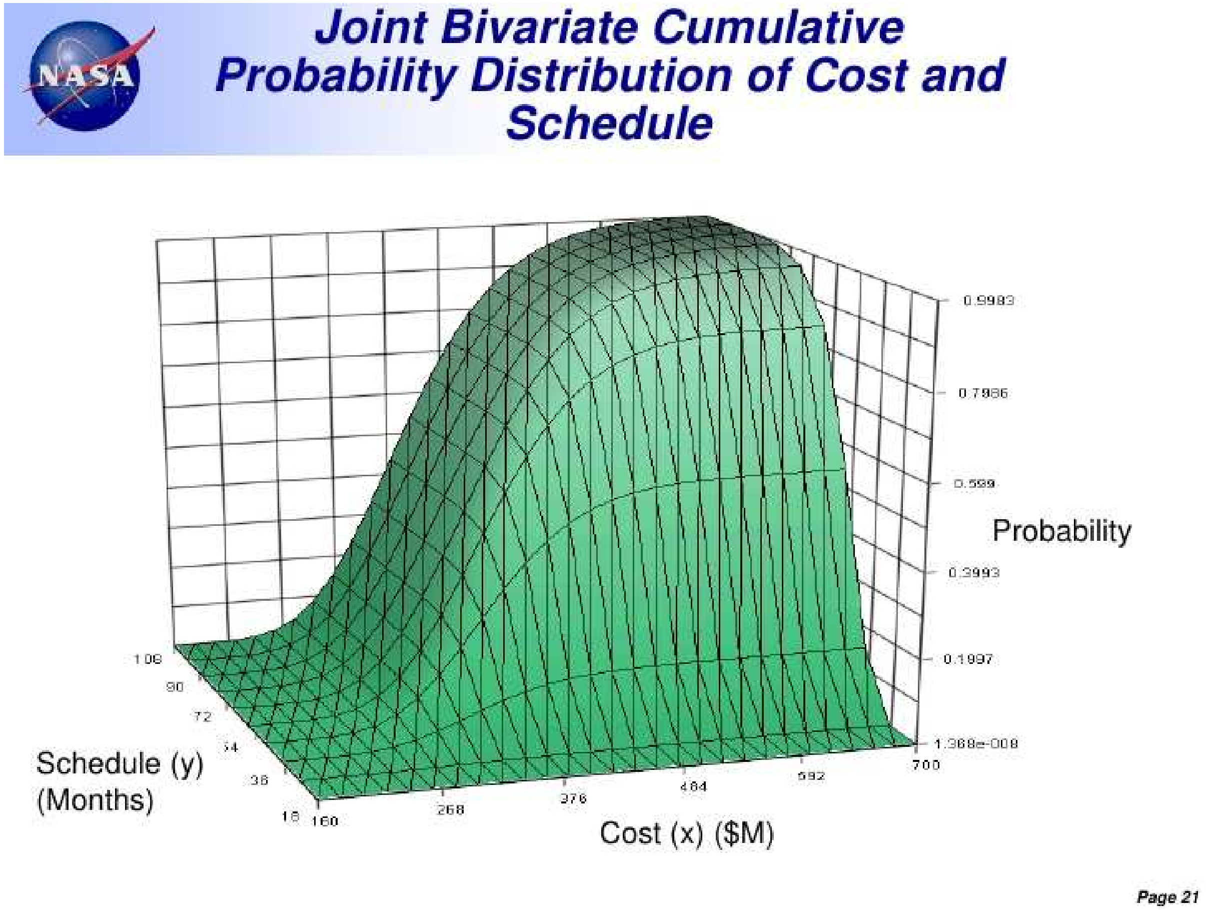
The Joint distribution function of the random vector $\mathbf{X}$ is:

$$F_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

Properties:

1. $F_{\mathbf{X}}(-\infty, \ldots, -\infty) = 0,$
2. $F_{\mathbf{X}}$ is increasing with increasing x_i
3. $F_{\mathbf{X}}(\infty, \ldots, \infty) = 1.$

Joint cumulative distribution function



Source http://www.slideshare.net/NASAPMC/sandra-smalley

Marginal distribution

Within a joint distribution $F_{\mathbf{X}}$, the marginal distribution of each of the X_i is given by:

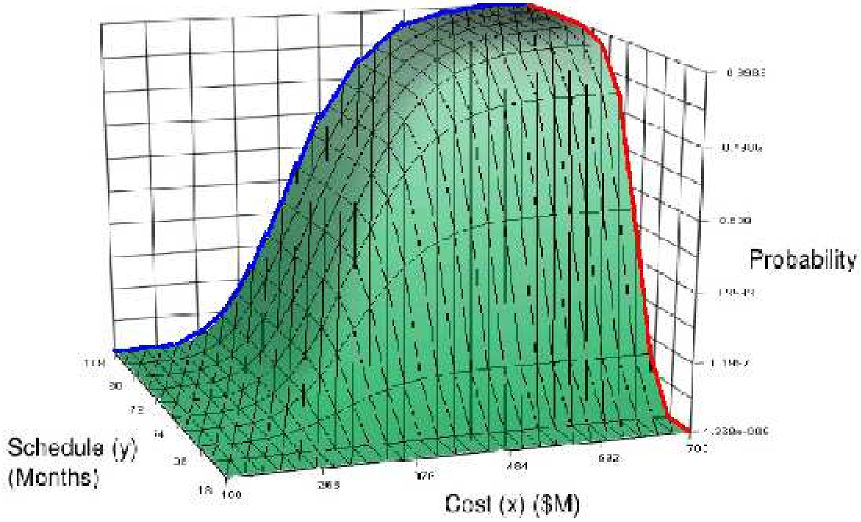
$$F_{X_i}(x_i) \quad = \quad F_{\mathbf{X}}(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty)$$

That is, only the i-th component of the joint distribution function is considered.

Remark

1. The random variable X_i is completely determined by its marginal distribution, but,
2. The joint distribution is in general not yet known if all marginal distributions are known.

Question: Why not?



Dept. of Geoscience & Remote Sensing

TU Delft

Independence

Consider the random variables

$$\begin{aligned} X_1 : \Omega &\rightarrow \mathbb{R} \\ X_2 : \Omega &\rightarrow \mathbb{R}, \\ &\vdots \\ X_n : \Omega &\rightarrow \mathbb{R} \end{aligned}$$

These random variables are mutually independent iff the events

$$\{X_1 < x_1\}, \{X_2 < x_2\} \ldots, \{X_n < x_n\}$$

are independent, that is:

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

Question: How are the marginal and joint distribution functions related in this special case?

Covariance

Consider the random variables

$$X_1 : \Omega :\longrightarrow \mathbb{R}, \qquad \text{and} \qquad X_2 : \Omega :\longrightarrow \mathbb{R}.$$

Assume the expectations $E\{X_1\} = \bar{x}_1$, $E\{X_2\} = \bar{x}_2$ and $E\{X_1 X_2\}$ are all finite.

The Covariance of X_1 and X_2 is defined as:

$$\begin{aligned}
\text{cov}(X_1, X_2) &= E\{(X_1 - \bar{x}_1)(X_2 - \bar{x}_2)\} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) f_{X_1,X_2}(x_1, x_2) dx_1 dx_2 \\
&= \ldots \big(\text{Expand and recollect}\big) \\
&= E\{X_1 X_2\} - \bar{x}_1 \bar{x}_2
\end{aligned}$$

Covariance and Independence

Let X_1 and X_2 be two independent random variables with finite means $E\{X_1\} = \bar{x}_1$ and $E\{X_2\} = \bar{x}_2$. Then

$$E\{X_1 X_2\} \;=\; E(X_1)E(X_2) \;=\; \bar{x}_1 \bar{x}_2$$

X_1 and X_2 are independent iff

$$f_{X_1,X_2}(x_1, x_2) \quad = \quad f_{X_1}(x_1) f_{X_2}(x_2)$$

Then

$$\begin{aligned}
E\{X_1 X_2\} &= \int\int x_1 x_2 f_{X_1,X_2}(x_1, x_2) dx_1 dx_2 \\
&= \int\int x_1 f_{X_1}(x_1) x_2 f_{X_2}(x_2) dx_1 dx_2 \\
&= \int x_1 f_{X_1} dx_1 \int x_2 f_{X_2} dx_2 \\
&= E\{X_1\} E\{X_2\}
\end{aligned}$$

Variance and Covariance

Let X be a random variable. The variance of X is the number

$$\text{var}(X) = E\{(X - E(X))(X - E(X))\}$$

The covariance between X and an additional random variable Y is given by

$$\text{cov}(X, Y) = E\{(X - E(X))(Y - E(Y))\}$$

Example. Let the random variable $Z = X + Y$ be the sum of the two random variables X and Y.

Question: What is the mean of Z?

Claim: $\text{var}(Z) = \text{var}(X) + \text{var}(Y) + 2 \cdot \text{cov}(X, Y)$

$$\begin{aligned}\text{var}(Z) &= E[(Z - \bar{Z})^2] = E[(X - \bar{X} + Y - \bar{Y})^2] \\ &= E[(X - \bar{X})^2 + E(Y - \bar{Y})^2 + 2 \cdot E(X - \bar{X})(Y - \bar{Y})] \\ &= \sigma_X^2 + \sigma_Y^2 + 2 \cdot \text{cov}(X, Y)\end{aligned}$$

Variance-Covariance matrix

Consider the random vector

$$\mathbf{X} = \{X_1, X_2, \ldots, X_n\}^T,$$

with expectation

$$E(\mathbf{X}) \quad = \quad \bar{\mathbf{X}} \;=\; \{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n\}^T$$

The variance-covariance matrix of $\mathbf{X}$, denoted Q_{xx}, is given by

$$\begin{aligned} Q_{xx} &= E((\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T) \\ &= \begin{pmatrix} \sigma^2_{X_1} & \text{cov}(X_1, X_2) & \ldots & \text{cov}(X_1, X_n) \\ \text{cov}(X_1, X_2) & \sigma^2_{X_2} & \ldots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_1, X_n) & \text{cov}(X_2, X_n) & \ldots & \sigma^2_{X_n} \end{pmatrix} \end{aligned}$$

Question: Why is Q_{xx} symmetric?

Experimental Covariance

Let Z_i and Z_j denote two random functions with standard deviations σ_i and σ_j resp.

Theoretical and Experimental covariance.

$$\begin{aligned}
\text{cov}(Z_i, Z_j) &= E\{(Z_i - E\{Z_i\})\} \cdot E\{(Z_j - E\{Z_j\})\} \\
&= E\{(Z_i - \mu_i)(Z_j - \mu_j)\} \; = \; \sigma_{ij} \\
&\updownarrow \\
\sigma_{ij} &= \frac{1}{n} \sum_{k=1}^{n} (z_{k,1} - \mu_1)(z_{k,2} - \mu_2)
\end{aligned}$$

Correlation

Disadvantage of covariance: arbitrary number

Let X_1 and X_2 be two random variables with finite variances $\sigma^2_{X_1}$ and $\sigma^2_{X_2}$.
The correlation coefficient ρ of X_1 and X_2 is given by:

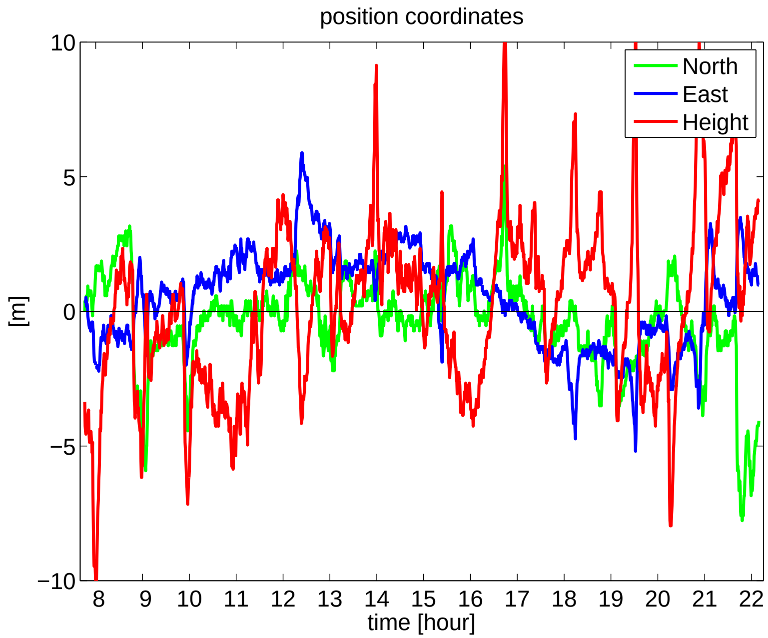
$$\rho(X_1, X_2) \quad = \quad \frac{\text{cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}$$

Corrollary. $\text{cov}(X_1, X_2) = 0$ iff $\rho(X_1, X_2) = 0$.

Short formula. (Pearson's) coefficient.

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \cdot \sigma_j} \in [-1, 1]$$

Are the GPS offsets correlated?



Matrix of covariances:

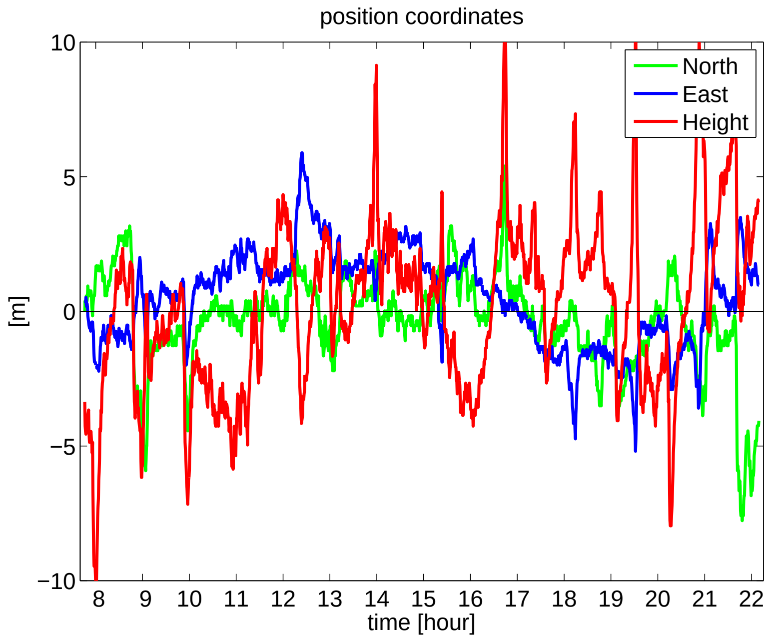
$$C(i,j) = \begin{pmatrix} 3.05 & 0.36 & -0.36 \\ 0.36 & 3.07 & -0.49 \\ -0.36 & -0.49 & 9.67 \end{pmatrix}$$

Question. What could be reason for correlation between the offsets (in N, E and H)?

Questions

1. What does the -0.49 represent?
2. And what the 3.07?
3. Why is C symmetric?
4. Why is $C(3,3)$ the largest entry?
5. What is the total variance of the data set?

Correlations between GPS offsets



Matrix of correlations:

$$P(i,j) = \begin{pmatrix} 1 & 0.12 & -0.07 \\ 0.12 & 1 & -0.09 \\ -0.07 & 0.09 & 1 \end{pmatrix}$$

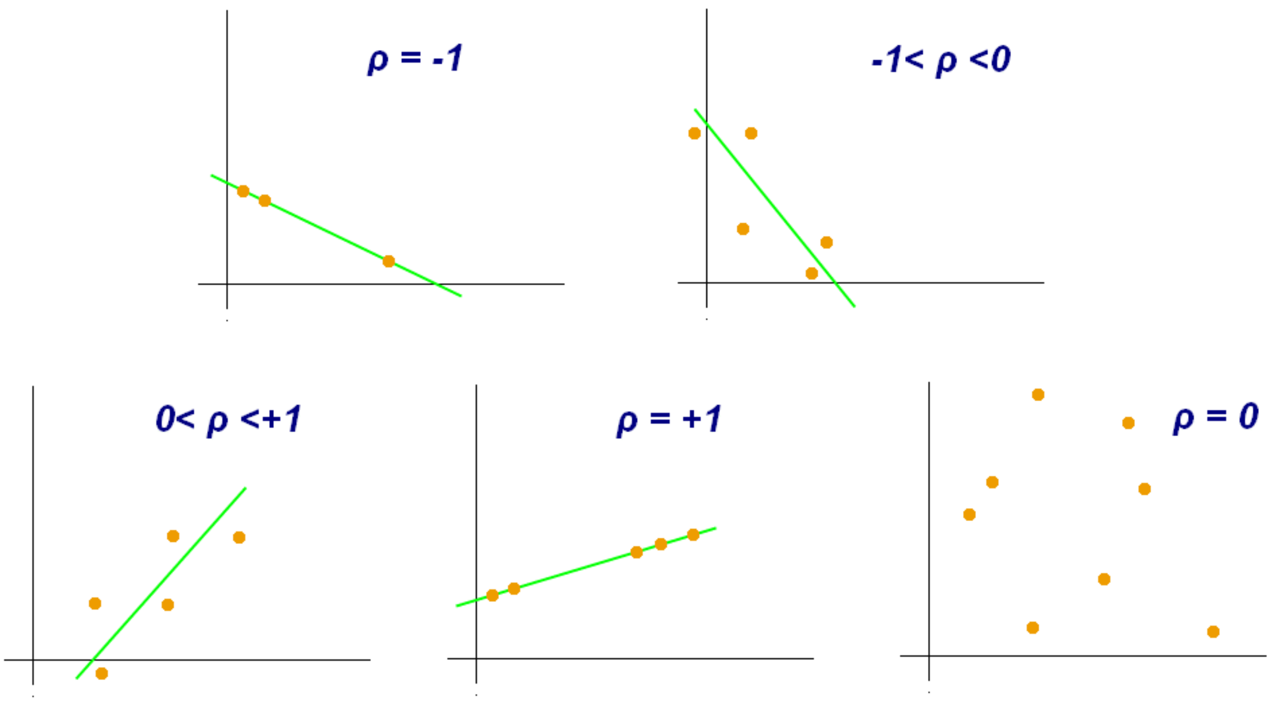
Questions

1. Why are all diagonal elements equal to 1?
2. Which two attributes are most correlated?
3. And which two least?
4. Why are some correlations positive and some negative?
5. What is wrong with this matrix?

Exercise.
Determine the correlation matrix, starting from the covariance matrix.

Dept. of Geoscience & Remote Sensing

Types of Correlation



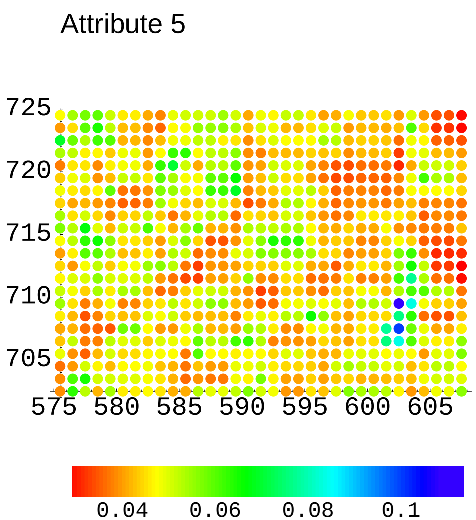
Source http://en.wikipedia.org/wiki/Pearson$_$product-moment$_$correlation$_$coefficient

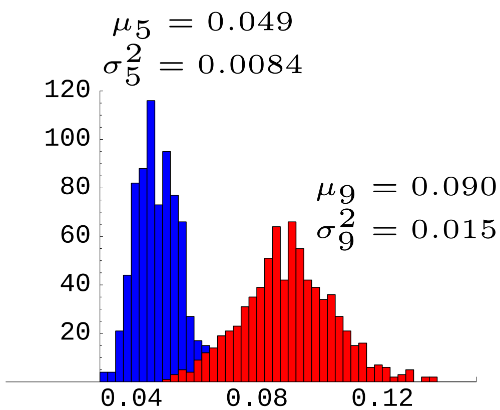
Question. What kind of relation do the covariance and correlation coefficient reveal?

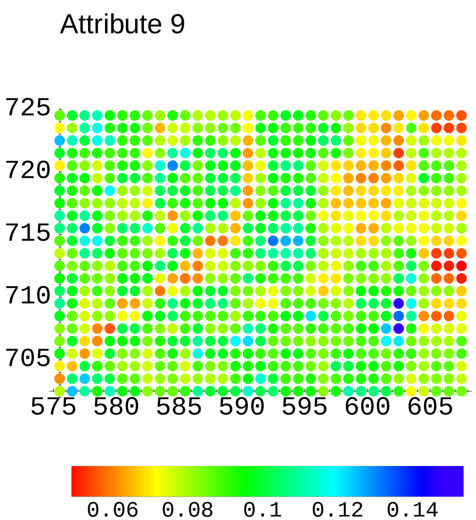
Dept. of Geoscience & Remote Sensing

Example: correlation between attributes.

Are attributes 5 and 9 related?







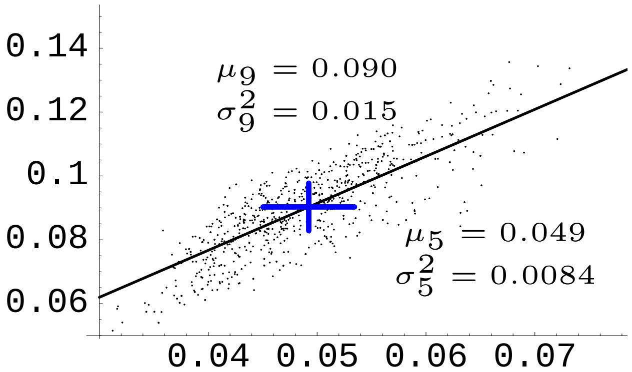
Covariance: $\sigma_{59} = \frac{\sum_{i=1}^{n}(z_{5,i}-\mu_5)(z_{9,i}-\mu_9)}{n} = 0.0001037.$

Correlation coefficient: $\rho_{5,9} = \frac{\sigma_{59}}{\sigma_5 \cdot \sigma_9} = 0.83.$

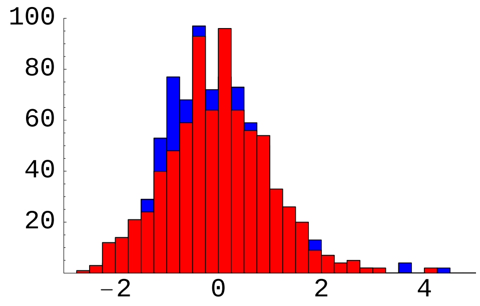
Example, scaling of point clouds.

Normalize all points: $\tilde{z}_i = \frac{z_i - \mu}{\sigma}$

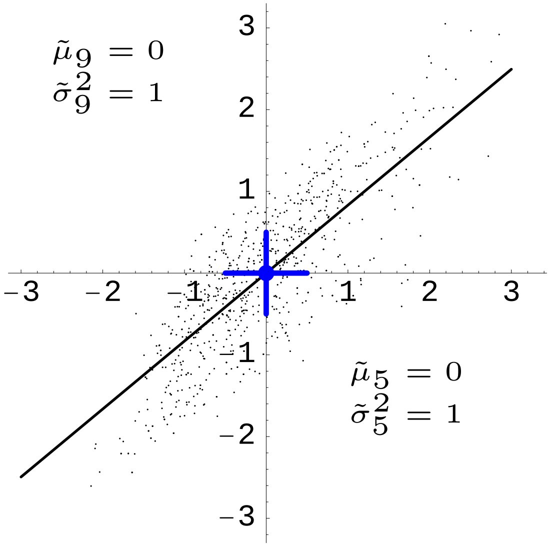
Original point cloud



Normalized histogram



Normalized point cloud



Normalized covariance: $\tilde{\sigma}_{5,9} = 0.83 = \frac{\tilde{\sigma}_{5,9}}{\tilde{\sigma}_5 \cdot \tilde{\sigma}_9} = \tilde{\rho}_{5,9} = \rho_{5,9}$.

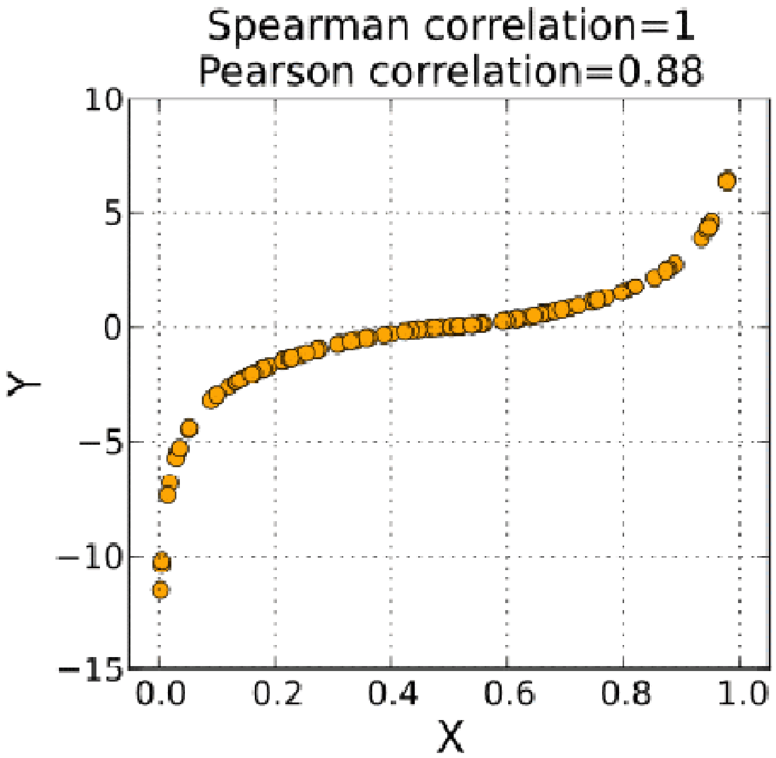
Remark. Minimal sum of least squares for linear fit equals $\sigma_9^2(1 - (\rho_{5,9})^2)$.

Alternative: Spearman correlation

Alternative for (Pearson) correlation:

Compare order statistics, not the real values.

1. **Input**: Two vectors $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$.
2. Replace each entry x_i by its (increasing) rank. This gives a vector R_X
3. Make the vector R_Y in the same way.
4. **Output**: (ordinary) correlation between vectors R_X and R_Y.



Source: Wikipedia

Remark. In case of ex aequo ranks, take the average.

Example: Spearman correlation

A group of AES students scored the following marks:

Soil Mechanics	5.6	7.5	5.5	7.1	6.1	6.4	5.8	10	9.1	6.1
Extraction of Resources	6.6	7.0	1	6.0	6.5	1.2	5.8	7.1	6.7	6.3

Which results in these ranks:

Soil Mechanics	9	3	10	4	6.5	5	8	1	2	6.5
Extraction of Resources	4	2	10	7	5	9	8	1	3	6

The Spearman correlation between the vectors of marks equals the ordinary correlation between the two rank vectors, which is 0.67

Exercise.
What is the correlation between the marks for Soil Mechanics and Extraction of Resources?

Covariance/correlation properties

Covariance

- Depends on the measurement scale.
- Positive covariance $\Leftrightarrow$ Residuals have the same sign $\Leftrightarrow$ Data values are on the same side of the mean
- High absolute covariance $\Leftrightarrow$ Both residuals are far way from mean.

Correlation

- Scale free.
- Uncorrelated variables $\Leftrightarrow$ Residuals are arbitrary $\Leftrightarrow$ $\rho_{ij} = \sigma_{ij} = 0$.
- $-1 \leq \rho_{ij} \leq 1$, while equality holds if and only if a linear relation exists between Z_i and Z_j with probability one.
- Only measures linear relations.
- Sensitive to outliers.

Alternatives???

Conclusions

Distributions with a name are

- characterized by a small number of parameters
- often linked to very particular experiments

Distributions are often related:

- Arbitrary normal to standard normal
- Binomial as multiple Bernoulli

Notably linear relationships are easy to work with

From univariate to multivariate:

- Consider several attributes simultaneosuly
- E.g. by their covariance or correlation

Exercises

Exercise 5.1 The random variable X has an Exponential distribution if its probability density function equals

$$f_X(x) = \begin{cases} \lambda E^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

a). Sketch the PDF and the CDF for $\lambda = 2$.

b). Determine mean and variance for general λ

Exercise 5.2 Assume the random variable X has a normal distribution. Let $\bar{x}$ denote its expectation. Use tabulated values of $1 - \Phi(X)$ or Matlab to show that

a). $P(|X - \bar{x}| \leq \sigma_X) = 0.683$

b). $P(|X - \bar{x}| \leq 2\sigma_X) = 0.954$

c). $P(|X - \bar{x}| \leq 1.96\sigma_X) = 0.95$

d). $P(|X - \bar{x}| \leq 2.58\sigma_X) = 0.99$

Exercises

Exercise 5.3 Assume that the duration of horse pregnancies varies according to a normal distribution with mean 336 days and standard deviation 3 days. Find the percentage of horse pregnancies that are longer than 339 days.

Exercise 5.4 Let $X_1, X_2, \ldots, X_n$ be n independent random variables, all with variance σ^2. Show that the variance of $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ is equal to σ^2/n.

Exercise 5.5 Determine the matrix of GPS correlations from the matrix of GPS covariances. (Slides 42 and 43).

Exercise 5.6 The random variable X has a uniform distribution of the interval (-1,3), i.e. $X \sim U(-1, 3)$. What is the mean of the random variable $Y = X^3 + 4$?

Dept. of Geoscience & Remote Sensing

Answers, Exercise 5.1

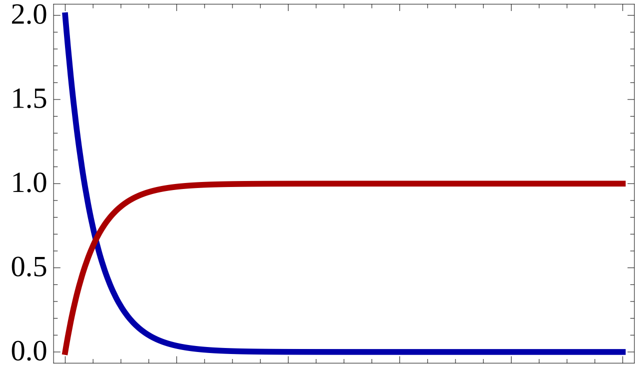
The random variable X has an Exponential distribution if its probability density function equals

$$f_X(x) = \begin{cases} \lambda E^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

1. Sketch the PDF and the CDF for $\lambda = 2$.
2. Determine mean and variance for general λ

Mean: $\bar{x} = \lambda \int_0^\infty x e^{-\lambda x} dx = \frac{1}{\lambda}$
Variance: $\sigma^2 = \lambda \int_0^\infty (x - \frac{1}{\lambda})^2 e^{-\lambda x} dx = \frac{1}{\lambda^2}$
Solve both integrals using Integrating by parts:

$$\int_a^b uv' = [uv]_a^b - \int_a^b vu'$$



Exponential Distr. λ=2; Blue: PDF; Red: CDF

Dept. of Geoscience & Remote Sensing

Answers, Exercise 5.2

Let $\bar{x}$ denote the expectation. Use tabulated values of $1 - \Phi(z)$ or Matlab to show that

1. $P(|X - \bar{x}| \leq \sigma_X) \;=\; 0.683$
2. $P(|X - \bar{x}| \leq 2\sigma_X) \;=\; 0.954$
3. $P(|X - \bar{x}| \leq 1.96\sigma_X) \;=\; 0.95$
4. $P(|X - \bar{x}| \leq 2.58\sigma_X) \;=\; 0.99$

1. $P(\frac{X-\bar{x}}{\sigma_X} < 1)$ = $P(|z| \leq 1)$ = $1 - 2P(z < -1)$ = $1 - 2 \cdot \Phi(-1)$ = $1 - 2 \cdot 0.1587$ = $.6826$
   (Evaluate the CDF of $N(0,1)$ at z = -1)
2. $P(|z| \leq 2)$ = $1 - 2 \cdot \Phi(-2)$ = $1 - 2 \cdot 0.0228$ = 0.9544
   (Evaluate the CDF of $N(0,1)$ at z = -2)
3. $P(|z| \leq 1.96)$ = $1 - 2 \cdot \Phi(-1.96)$ = $1 - 2 \cdot 0.025$ = 0.95
   (Evaluate the CDF of $N(0,1)$ at z = -1.96)
4. $P(|z| \leq 2.58)$ = $1 - 2 \cdot 0.0048$ = 0.9902
   (Evaluate the CDF of $N(0,1)$ at z = -2.58)

Answers, Exercise 5.3

Assume that the duration of horse pregnancies varies according to a normal distribution with mean 336 days and standard deviation 3 days. Find the percentage of horse pregnancies that are longer than 339 days.

$X \approx N(336, 3)$. Therefore,

$$
\begin{aligned}
P(X \geq 339) &= P(\frac{X - \bar{X}}{\sigma_x} \geq \frac{339 - 336}{3}) \\
&= P(Z \geq 1)
\end{aligned}
$$

Matlab: $P(Z \geq 1) = .1587 \approx 16\%$, with $Z \approx N(0, 1)$.

(Compare previous exercise)

Answers, Exercise 5.4

Let $X_1, X_2, \ldots, X_n$ be n independent random variables, all with variance σ^2. Show that the variance of $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ is equal to σ^2/n.

In the lecture it has been shown that

$$\text{var}(X_1 + X_2) \quad = \quad \text{var}(X_1) + \text{var}(X_2) + 2\text{cov}(X_1, X_2)$$

X_1 is independent from X_2, so the covariances vanish. Therefore

$$\text{var}(X_1 + X_2 + \cdots + X_n) \quad = \quad n\text{var}(X_i) \; = \; n\sigma^2$$

Moreover, $\text{var}(\frac{1}{n}Z) = \frac{1}{n}^2\text{var}(Z)$. So, the result follows with $Z = X_1 + \ldots X_n$.

Dept. of Geoscience & Remote Sensing

Answers, Exercise 5.5

Determine the matrix of GPS correlations from the matrix of GPS covariances. (Slides 42 and 43).

To obtain entry $P(i,j)$ for $i = 1,2,3$, and $j = 1,2,3$ in the matrix on Slide 43, apply the following formula:

$$P(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)}\sqrt{C(j,j)}}$$

For example,

$$1 \;=\; P(1,1) \;=\; \frac{3.05}{\sqrt{3.05}\sqrt{3.05}}$$

and

$$-0.09 \;=\; P(2,3) \;=\; \frac{-0.49}{\sqrt{3.07}\sqrt{9.67}}$$

Answers, Exercise 5.6

The random variable X has a uniform distribution of the interval (-1,3), i.e. $X \approx U(-1,3)$. What is the mean of the random variable $Y = X^3 + 4$?

Probability density function: $f_X(x) = \frac{1}{4}$ (compare before)

Determine first the expectation of X:

$$E\{X\} = \int x f_X \, dx = \int_{-1}^{3} x \frac{1}{4} dx = 1$$

Let $Y \approx X^3 + 4$. Then

$$E\{Y\} = \int y f_Y \, dy = \int (x^3 + 4) f_X(x) dx = \int_{-1}^{3} (x^3 + 4) \frac{1}{4} dx = \frac{1}{4} [\frac{x}{4}^4 + 4x]_{-1}^{3} = 9$$