Data Analysis, Statistics, Machine Learning

Leland Wilkinson

Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

leland.wilkinson@gmail.com

Distributions

- A probability function is a nonnegative function
  - Its area (or *mass*) is 1
- Distributions are families of probability functions
- Most statistical methods depend on distributions
  - Nonparametric methods are distribution-free
- The Normal (Gaussian) distribution is most popular
  - Other distributions (Binomial, Poisson, …) are often used
  - We use the Normal because of the Central Limit Theorem
    - Variables based on real data are rarely normally distributed
    - But sums or means of random variables tend to be
    - So if we are drawing inferences about means, Normal is usually OK
    - This involves a leap of faith

Distributions

- What is a probability?
  - This is a philosophical question
  - Not needed to prove the mathematical validity of statistical models
    - Kolmogorov did that with his axioms (see below)
  - But needed to do inferences that make sense in the real world
  - Objective (Frequentist) interpretation
    - originated in games of chance
    - relative frequency of a random experiment's outcome, normed to [0, 1]
    - defined over the long run: many replications of the experiment (e.g., toss of coin)
  - Subjective (Bayesian) interpretation
    - degree of belief in an outcome, normed to [0, 1], or [none, …, certain]
    - proportional to how much you are willing to pay when making a bet
  - Frequentists despise the subjective definition
  - Bayesians say they don’t need the frequentist definition

Distributions

What is a probability?

John Hartigan

In the *Foundations of the Theory of Probability*, Kolmogorov did a great disservice to probability because he said "It's just a measure." So everyone thought, "Oh thank God, we've solved the problems of the foundations of probability." Of course we haven't solved anything at all! That's just mere technicalities. Countable additivity and measure theory and whether or not the axiom of choice matters and things like that constitute a total distraction from an understanding of what probability is. I regard Kolmogorov's work as a great step backward, at least in the foundations of probability. In the mathematics of probability, it was a great step forward.

Inference

- Frequentist probability
  - The ratio of the number of times the event occurs in a test series to the total number of trials in the series
    - probabilities are based on frequencies (counting events)
  - Number of trials must be large
  - We must assume independence of trials
    - This is never true in the real world experiments
  - The event must always occur with the same probability
    - This is never true in the real world
    - The definition is circular because it is conditional on defining probability

Inference

Bayesian probability

A measure of the degree of belief that an event will occur

- This measure is close to our intuition
- Not related to the outcome of repeated experiments
- Betworthy events are coherent
- Subjective probability does not mean *arbitrary*
- Objectivity is achieved through *intersubjectivity* (shared culture)

Distributions

Probability Functions

Let S be a countable set of values called a *sample space*

Let an event A be a subset of S

Let $E = 2^S$ be the set of all possible events based on S

The function $P(.)$ on the domain E is a *probability function* if

$$P(A) \geq 0 \text{ for all } A \in E$$
$$P(S) = 1$$
$$P(A \cup B) = P(A) + P(B) \text{ when } A \cap B = \emptyset$$

From these axioms of Kolmogorov, we can deduce

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

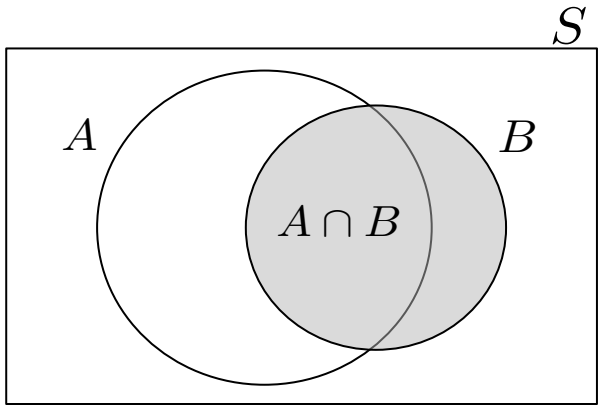
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

if A and B are *independent*, then

$$P(A|B) = P(A)$$
$$P(A \cap B) = P(A)P(B)$$

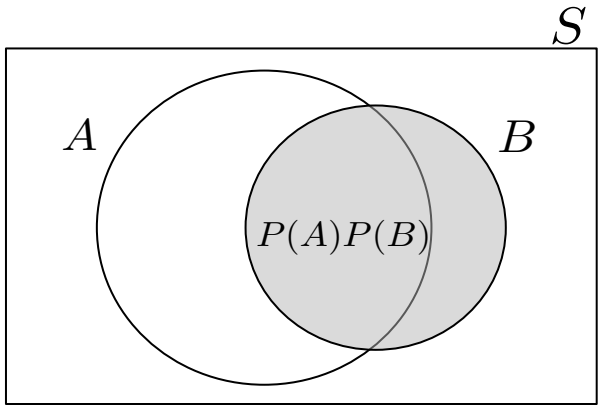
Distributions

Conditional Probabilities



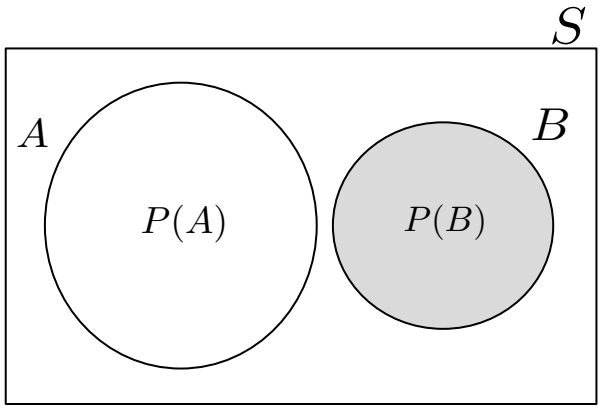
Conditional

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Independent

$$P(A|B) = P(A)$$
$$P(A \cap B) = P(A)P(B)$$



Disjoint

$$P(A|B) = P(B|A) = 0$$
$$P(A \cap B) = 0$$
$$P(A \cup B) = P(A) + P(B)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Distributions

Random Variables

A variable is a mapping from a set of objects to a set of values

$$X \equiv f : O \mapsto V$$

A random variable is a variable whose values are associated with probabilities

- A discrete random variable associates a value with a probability
- A continuous random variable associates an interval with a probability

The mathematical function describing the possible values of a random variable and their associated probabilities is known as a *probability distribution*

Expectations of random variables

The expected value of a random variable is the weighted average of the values of that variable, where the weights are probabilities

We will use this definition later when computing maximum likelihood estimates of the mean

We use this notation for expected value:

$$E[X]$$

Distributions

Random Variables

Algebra of Expectations

$$
\begin{aligned}
E[c] &= c \\
E[cX] &= cE[X] \\
E[c+X] &= c + E[X] \\
E[X+Y] &= E[X] + E[Y] \\
E[XY] &= E[X]E[Y] \text{ if X and Y are independent} \\
VAR(X) &= E[X^2] - E[X]^2 \\
VAR(cX) &= c^2 VAR(X) \\
VAR(c+X) &= VAR(X) \\
VAR(X+Y) &= VAR(X) + VAR(Y) \text{ if X and Y are independent} \\
COV(X,Y) &= E[XY] - E(X)E(Y) \\
VAR(X+Y) &= VAR(X) + VAR(Y) + 2COV(X,Y) \\
VAR(X-Y) &= VAR(X) + VAR(Y) - 2COV(X,Y)
\end{aligned}
$$

Distributions

Random Variables

The law of large numbers

Let $\overline{X}^{(n)} = (X_1 + \cdots + X_n)/n$, where

$$E[X_1] = E[X_2] = \cdots = E[X_n] = \mu$$

Then

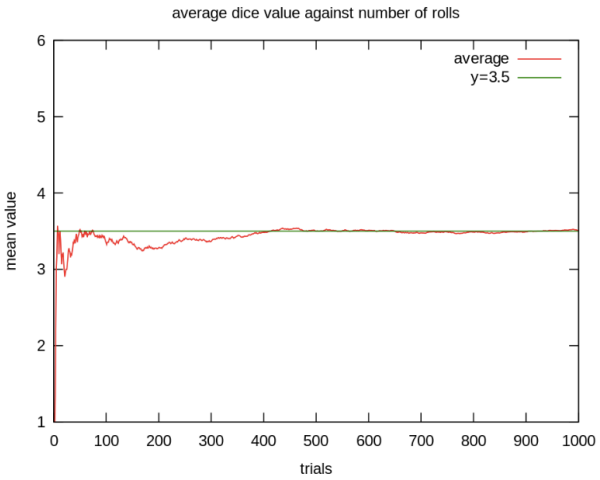
$$\overline{X}^{(n)} \to \mu \text{ as } n \to \infty$$

Distributions

Random Variables

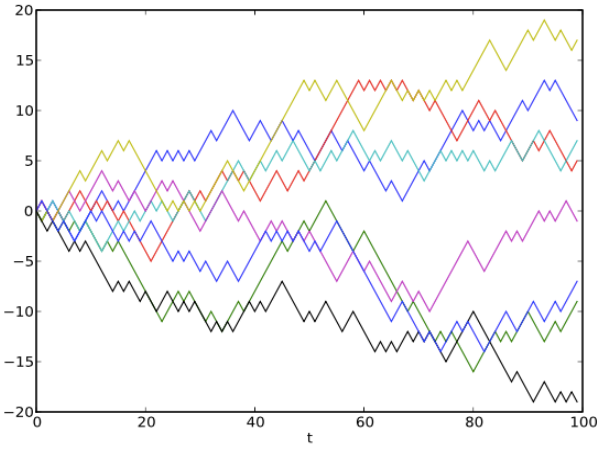
The law of large numbers

YES
the average



Wikipedia

NO
not the sum
(gambler's ruin)



Wikipedia

NO
not eventually
(gambler's fallacy)

someecards.com

Distributions

Bayes' Theorem

Conditional probabilities of twins configurations

$$P(M|B) = \frac{P(M \cap B)}{P(B)} = \frac{\text{light blue}}{\text{light blue} + \text{dark blue}} = \frac{.125}{.125 + .1875} = .4$$

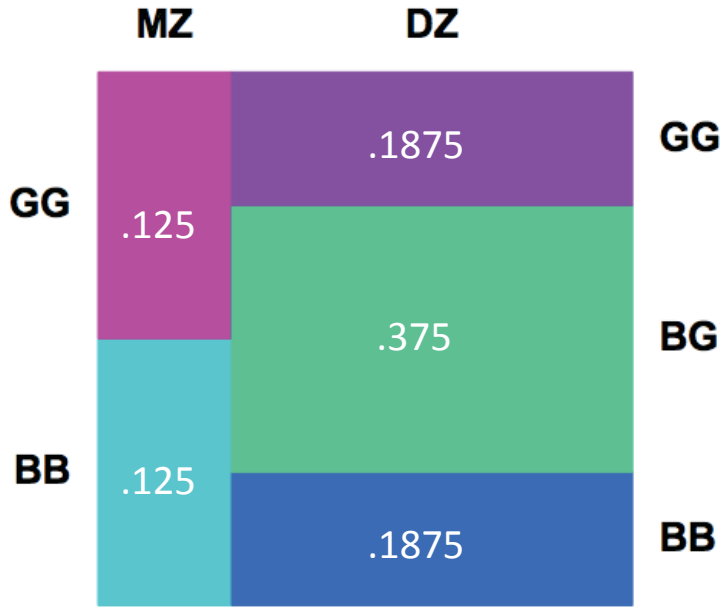
$$P(M|B) = \frac{P(M \cap B)}{P(B)}$$

$$= \frac{P(M \cap B)}{P(M \cap B) + P(\bar{M} \cap B)}$$

$$= \frac{P(B|M)P(M)}{P(B|M)P(M) + P(B|\bar{M})P(\bar{M})}$$

$$P(M|B) = \frac{.5 \times .25}{.5 \times .25 + .25 \times .75} = .4$$

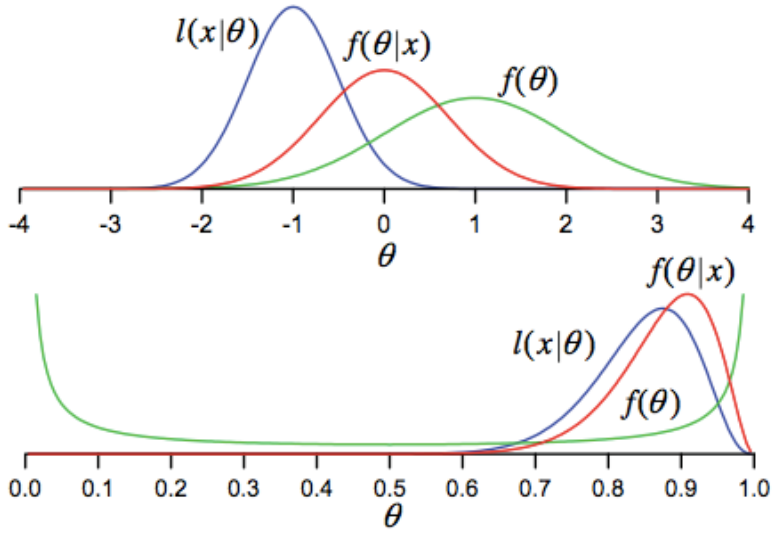
MZ = M = Monozygotic
DZ = D = Dizygotic
GG = G = Girl Girl twins
BB = B = Boy Boy twins



Distributions

Bayes' Theorem for Continuous Random Variables

$$f(\theta|x) = \frac{l(x|\theta)f(\theta)}{\int l(x|\theta)f(\theta)d\theta}$$



Distributions

Common Probability Fallacies

Gambler's fallacy

$$P(H_1 \cap H_2 \cap H_3 \cap H_4, ...) = P(H_1)P(H_2)P(H_3)P(H_4), ...$$

$$P(H_n)|P(H_1 \cap H_2 \cap H_3 \cap H_4, ...) = P(H_n)$$

If events are independent

Distributions

Common Probability Fallacies

Prosecutor's fallacy

$P(E|I) =$ Probability evidence observed when accused is innocent

$P(I|E) =$ Probability accused is innocent given evidence

$P(E|I)$ is usually tiny

So prosecuter concludes $P(I|E)$ is tiny

But $P(I|E) = \dfrac{P(E|I)P(I)}{P(E)}$ where,

$P(E) = P(E|I)P(I) + P(E|\bar{I})[1 - P(I)]$

Distributions

Common Probability Fallacies

Base rate fallacy

$$A \equiv \text{Alarm}$$
$$T \equiv \text{Terrorist}$$
$$I \equiv \text{Innocent}$$

$$P(A|T) = .999$$
$$P(A|I) = .001$$
$$P(T) = .00001$$
$$P(T|A) = .01$$

Chance of false alarming is .99

even though chance of alarming given a terrorist is .999

even though chance of alarming given an innocent person is .001

Because the base rate for terrorists passing through security is only .00001

Distributions

Common Probability Fallacies

Law of averages fallacy

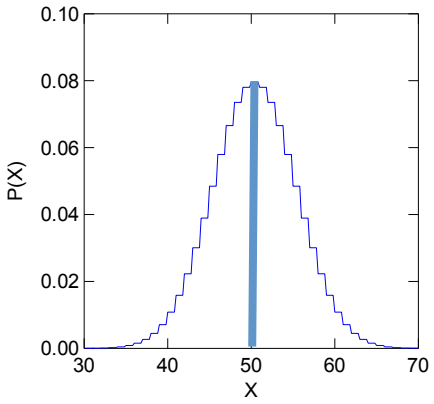
A popular misconception about the law of large numbers

Flip a fair coin 100 times

One might predict that there will be 50 heads and 50 tails

While this is the single most likely outcome,

there is only an 8% chance of it occurring.



Distributions

Common Probability Fallacies

Law of small numbers fallacy

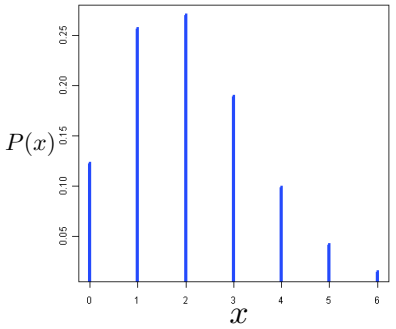
Tversky and Kahneman

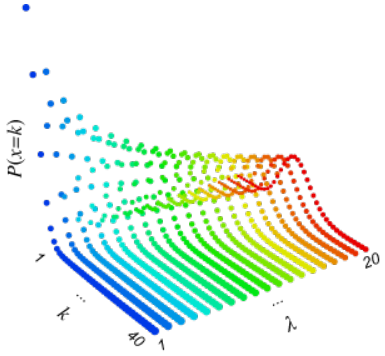
Overconfidence in a small sample

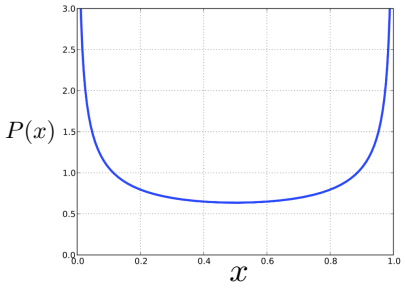
"As I understood clearly only when I taught statistics some years later, the idea that predictions should be less extreme than the information on which they are based is deeply counterintuitive."
(Kahneman Nobel Prize lecture)

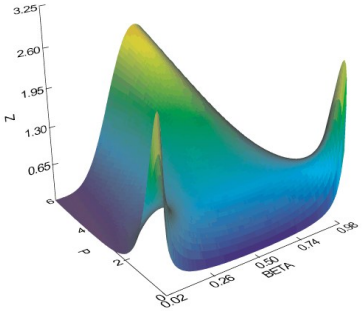
Distributions

Distributions are families of probability functions





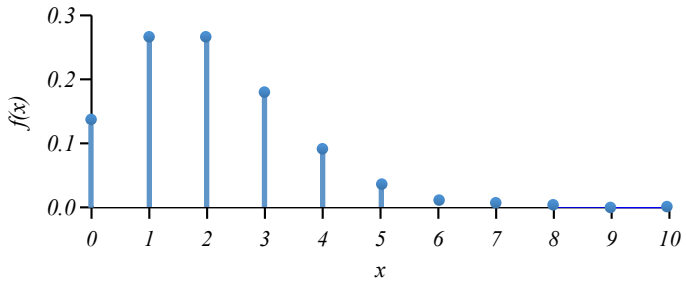




Distributions

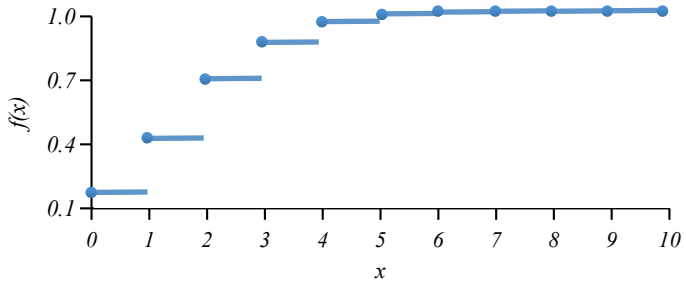
Discrete Density and Distribution Functions

Density function (PDF)



$$f(x) = P(x)$$

Distribution function (CDF)

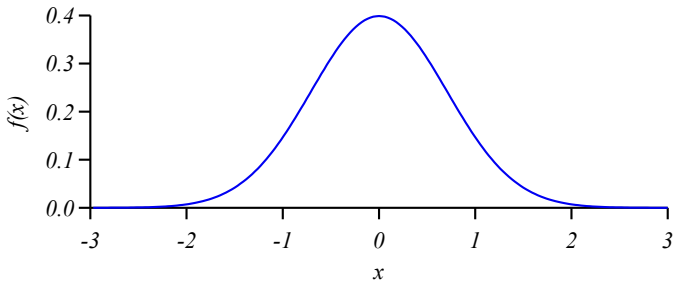


$$F(x) = \sum_{x_i \leq x} f(x_i)$$

Distributions

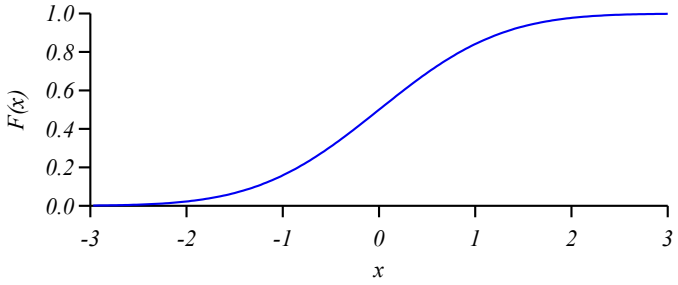
Continuous Density and Distribution Functions

Density function (PDF)



$$f(x) = P(x)$$

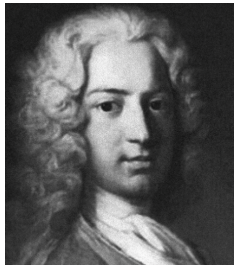
Distribution function (CDF)



$$F(x) = \int_{-\infty}^{x} f(u)du$$

Distributions

The Binomial Distribution

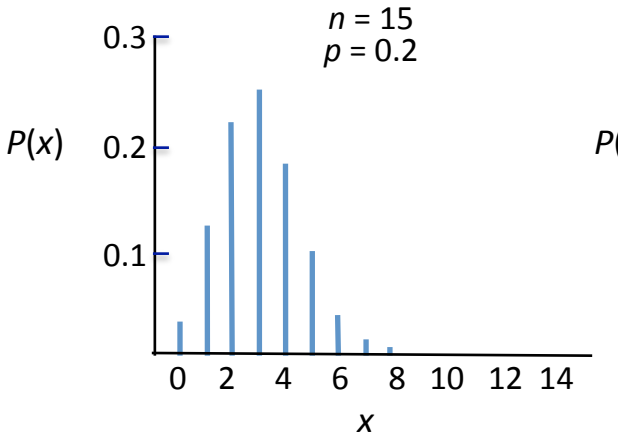


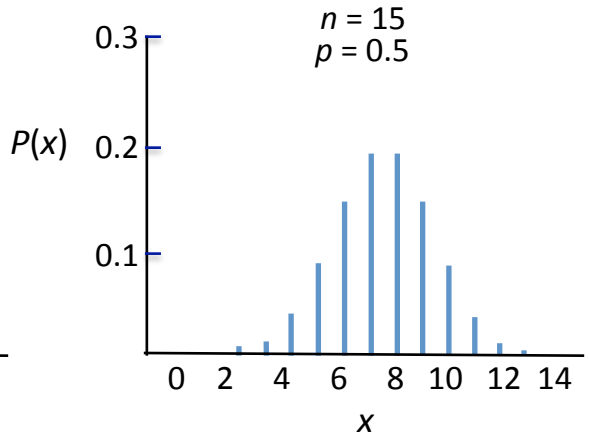
$$P(x; p, n) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

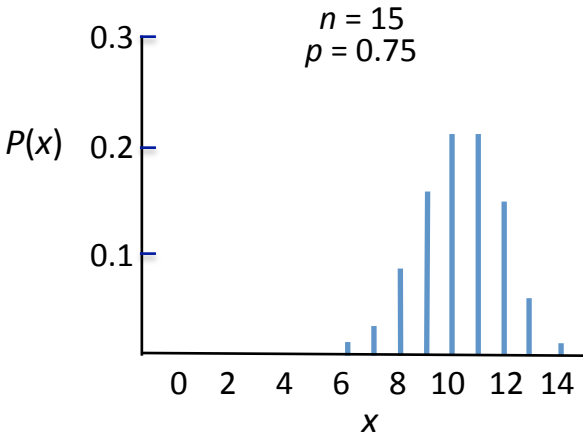
We have *n* independent Bernoulli trials (2 outcomes: *success* or *fail*)

Each trial has probability *p* of success

A fair coin has *p* = .5, so *P*(*x; p, n*) is the probability of exactly *x* heads in *n* tosses







Distributions

The Binomial Distribution

$$P(x; p, n) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

$$(p+q)^n = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

Pascal's triangle

$$\binom{0}{0}$$

$$\binom{1}{0}\ \binom{1}{1}$$

$$\binom{2}{0}\ \binom{2}{1}\ \binom{2}{2}$$

$$\binom{3}{0}\ \binom{3}{1}\ \binom{3}{2}\ \binom{3}{3}$$

$$\binom{4}{0}\ \binom{4}{1}\ \binom{4}{2}\ \binom{4}{3}\ \binom{4}{4}$$

$$\binom{5}{0}\ \binom{5}{1}\ \binom{5}{2}\ \binom{5}{3}\ \binom{5}{4}\ \binom{5}{5}$$

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

Distributions

The Binomial Distribution

The number of successful sales calls

The number of recalls by a car company

The number of females in a software company

The number of days that your cable service fails

Remember the assumptions!

There are *n* trials

- Each trial can result only in a success or a failure
- The probability of success (*p*) is the same for every trial
- The outcomes of different trials are independent
- We are interested in the total number of successes in these *n* trials

Distributions

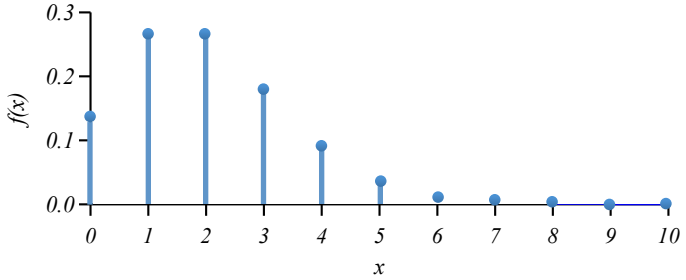
The Poisson Distribution

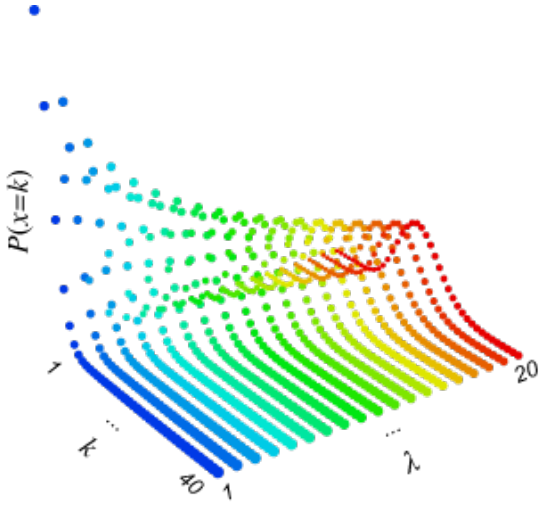
$$P(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Probability of a given number of events occurring in a fixed region (time, length, area, volume, …)

Events must have a known average rate (λ) and be independent across regions

The Poisson distribution is a limiting case of a Binomial distribution when the number of trials (*n*), gets large and the probability of success (*p*), is small





Distributions

The Poisson Distribution

The number of cars that pass through a mile marker on an expressway

The number of car accidents in a given time interval

The number of phone calls at a call center each hour

The number of packets a web node receives each minute

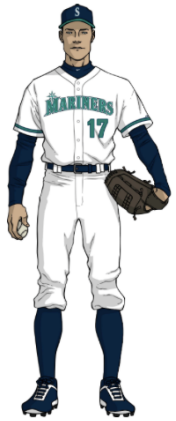
The number of cactuses per acre

The number of typing errors on a page

The number of light bulbs that burn out in a year

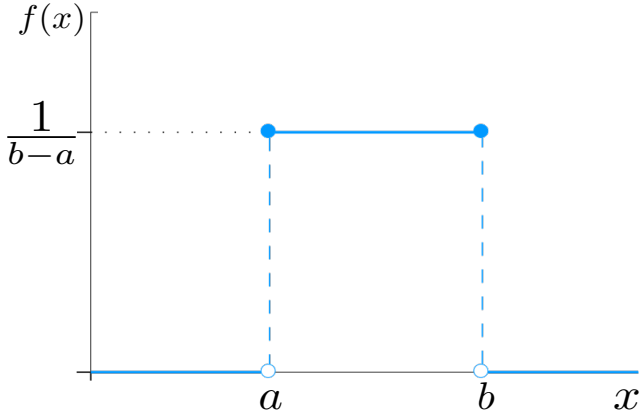
Distributions

The Uniform Distribution



Sorry, bad pun

$$P(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$



Distributions

The Uniform Distribution

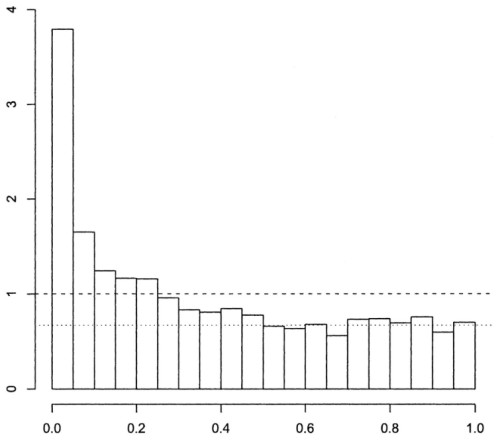
A set of numbers from a pseudorandom number generator

The position of an oxygen molecule in a room

Trailing digits in a collection of numerical data (Benford)

The *p*-values for a set of independent experiments when the null is true

- This example shows how to leverage this idea
- Values above the dotted line are likely significant
- Because null distribution would have been level



Story & Tibshirani (2003), *PNAS, 100*, 9440–9445

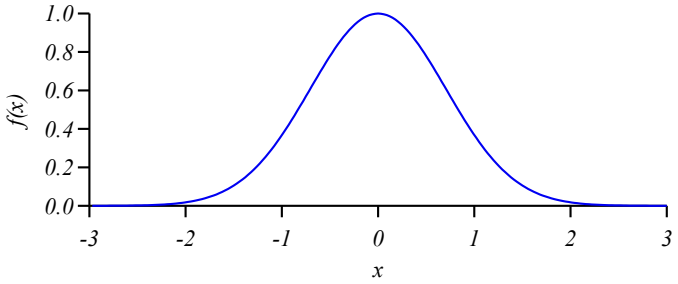
Distributions

The Normal (Gaussian) Distribution

First devised (by Laplace and Gauss) for aggregating measurements

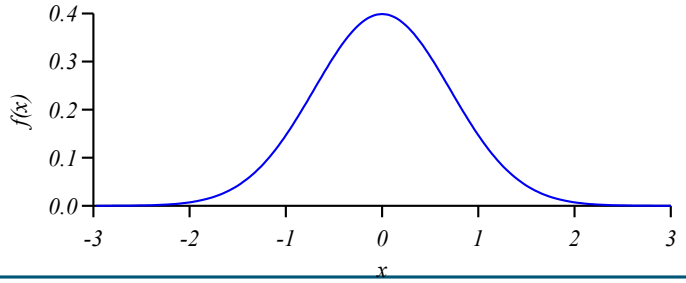
The word Normal is probably due to Karl Pearson

$$f(x) = e^{-x^2}$$



Error function

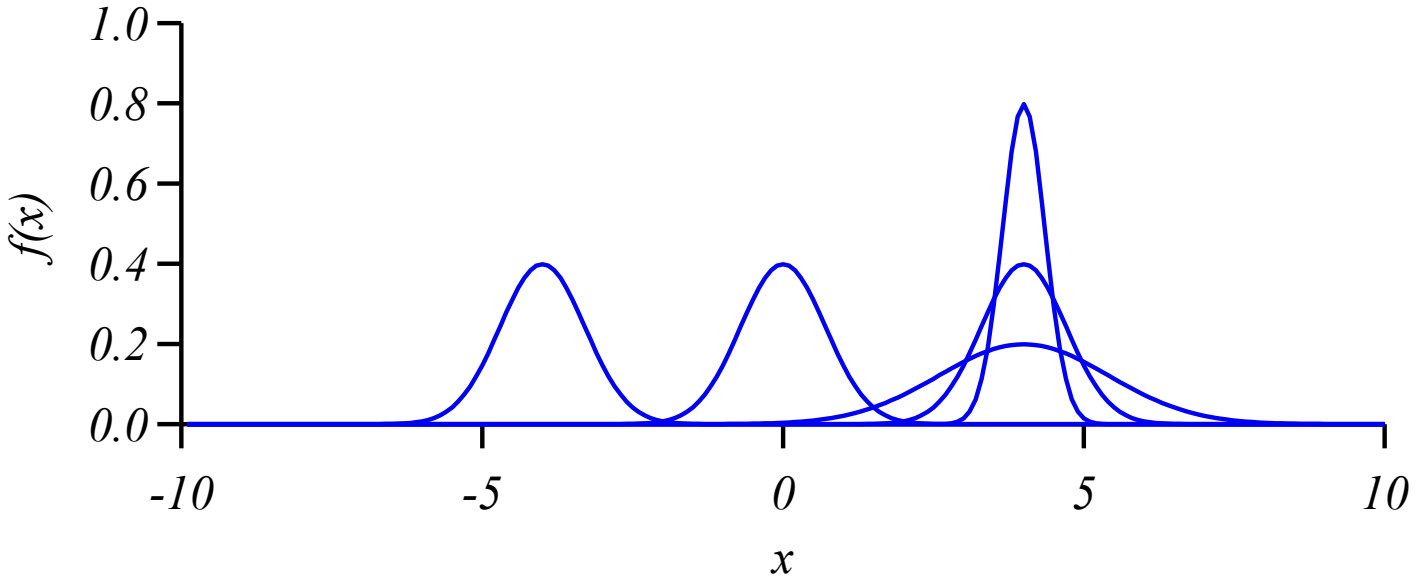
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Gaussian function

Distributions

The Normal (Gaussian) Distribution



Distributions

The Central Limit Theorem (Laplace)

- ✓ We have *n* independent random variables
- ✓ They are identically distributed
- ✓ They have finite mean μ and variance σ^2

As *n* increases, the distribution of their mean → Normal

with mean μ

and variance σ^2/n

There are several different versions, but this is the basic one

Distributions

The Central Limit Theorem

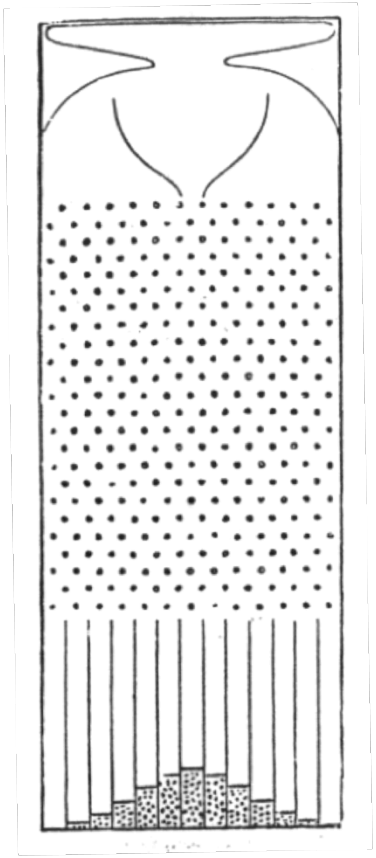
Implications

The expected error of a poll rests on the size of the sample (*n*)

NOT the size of the population (or size relative to population)

Distributions

Galton's Quincunx



Each pin randomly bounces a ball left or right (a Bernoulli trial)

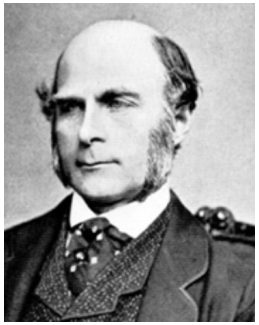
The outcome of the device represents a Binomial distribution

The probability that the ball ends up in the *k*th bin is

$$P(k;p,n) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

As *n* becomes large,

the CLT says the Binomial approaches Normal



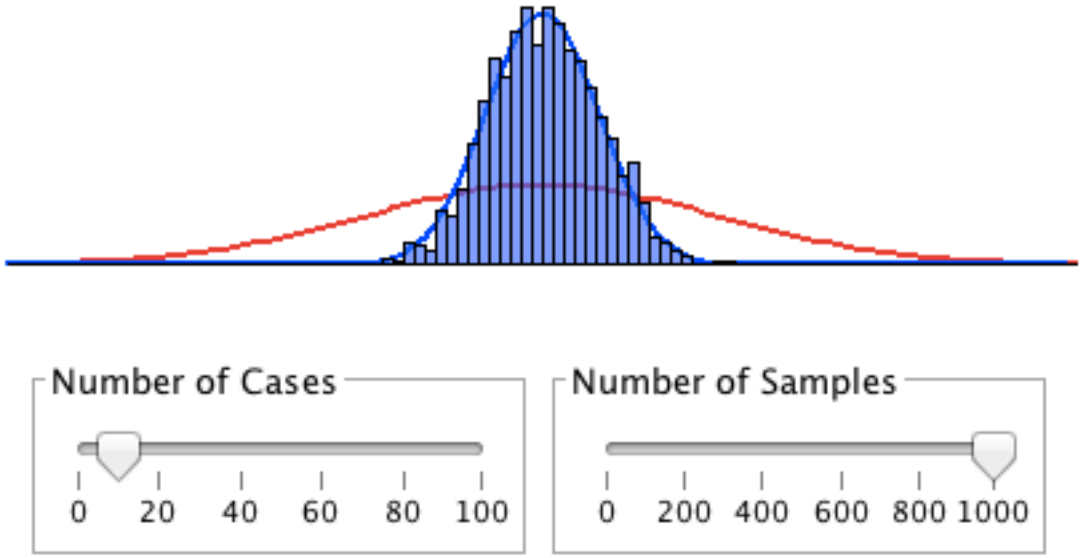
Distributions

A demonstration

CLT Applet

(http://www.cs.uic.edu/~wilkinson/Applets/clt.html)

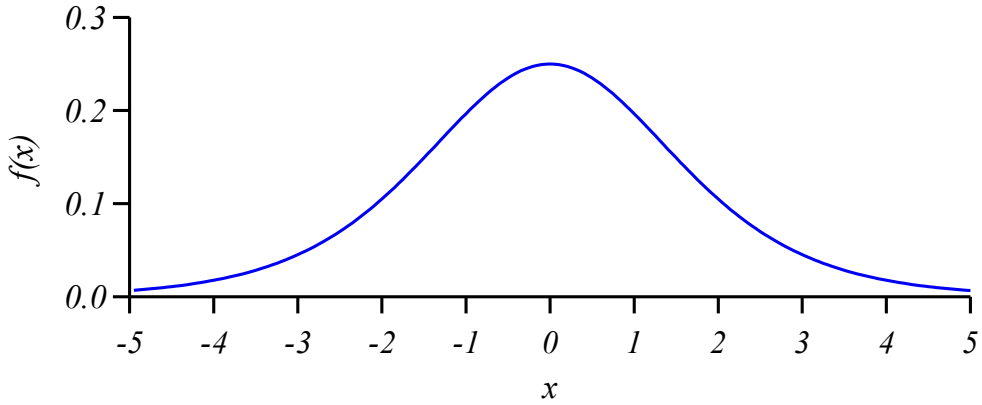
(you may need to set permissions to allow Java applets to execute)



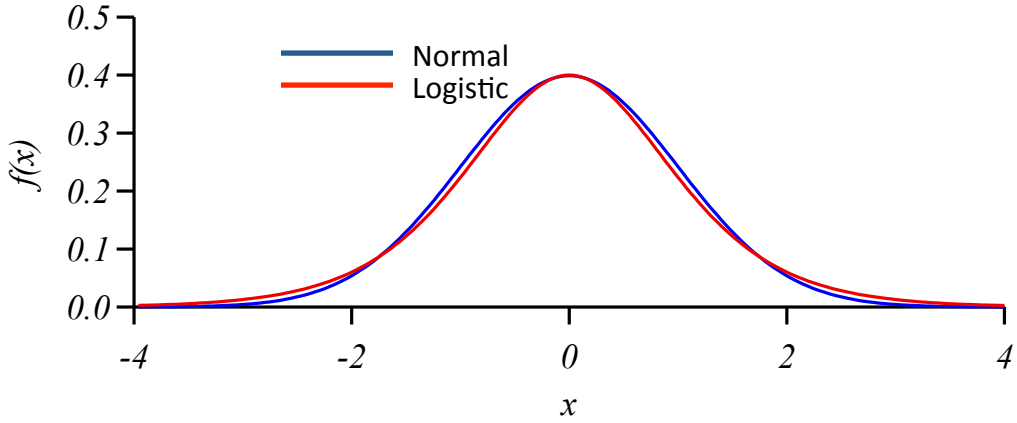
Distributions

The Logistic Distribution

$$f(x) = \frac{e^x}{(1+e^x)^2}$$



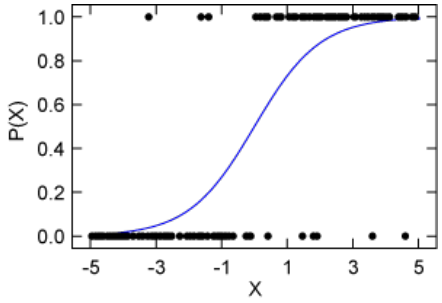
$$P(x;\alpha,\beta) = \frac{1}{\beta}\frac{e^{-(x-\alpha)/\beta}}{(1+e^{-(x-\alpha)/\beta})^2}$$



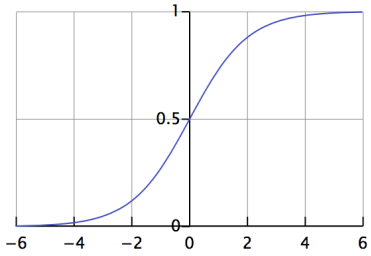
Distributions

The Logistic Distribution

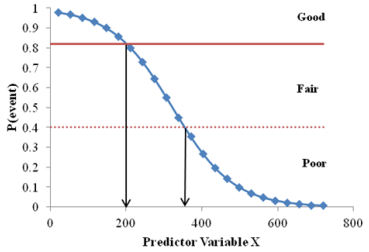
$$F(x) = \frac{e^x}{1 + e^x}$$



Logistic regression



Growth curves

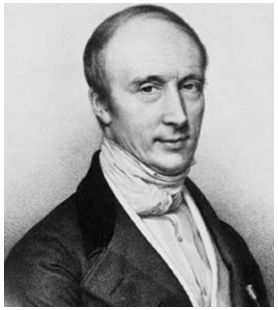


Reliability

Distributions

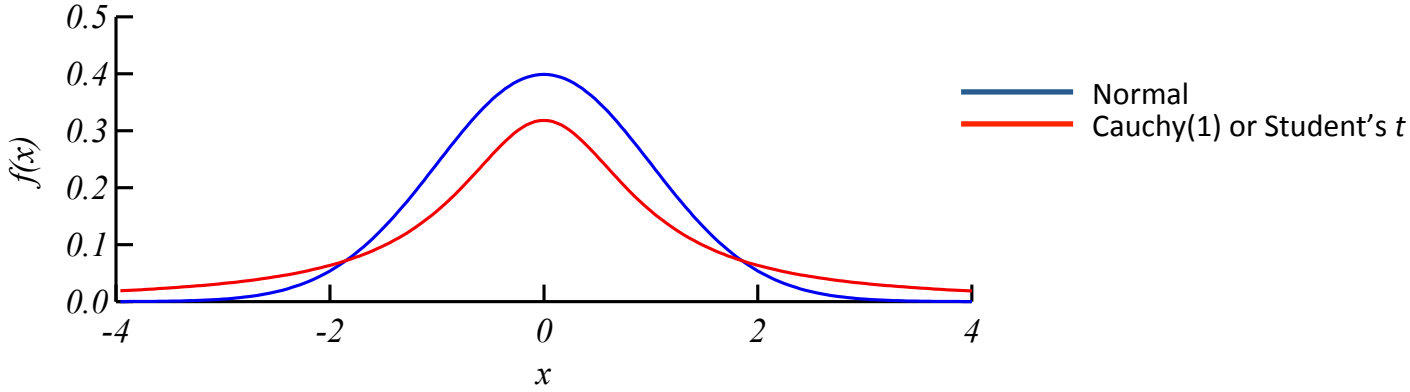
The Cauchy Distribution

$$P(x;\alpha,\lambda) = \frac{1}{\pi\lambda\left\{1+\left(\frac{x-\alpha}{\lambda}\right)^2\right\}}$$



Student's *t*

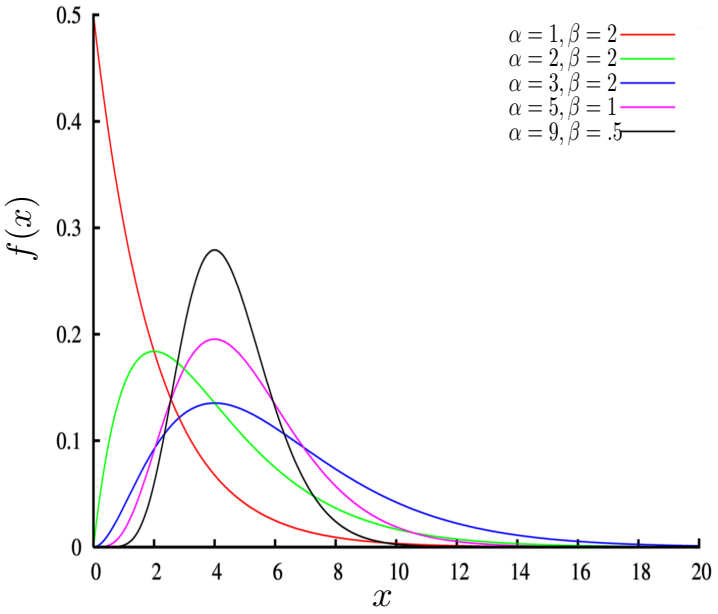
$$\alpha = 0, \lambda = 1$$

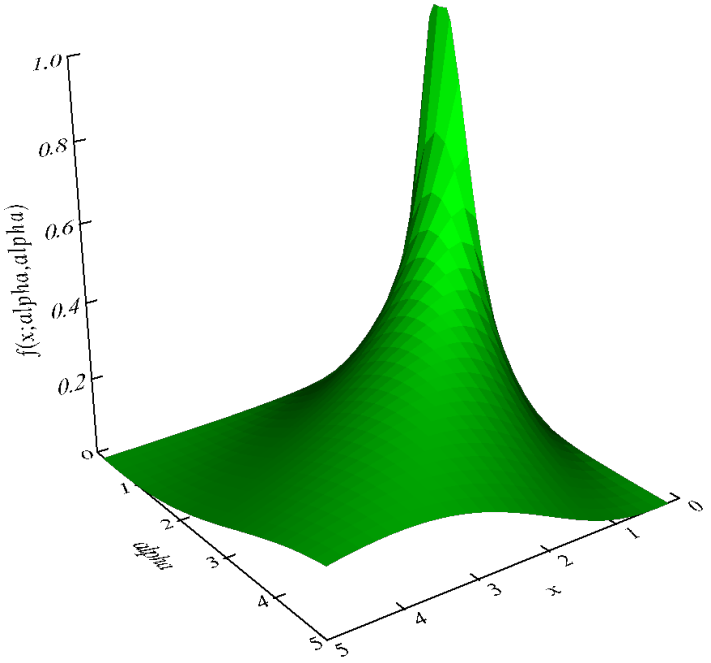


Distributions

The Gamma Distribution

$$P(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$





Distributions

The Gamma Distribution

The chi-square distribution with v degrees of freedom is Gamma(v/2, 2)

The amount of rainfall accumulated in a reservoir

The size of loan defaults

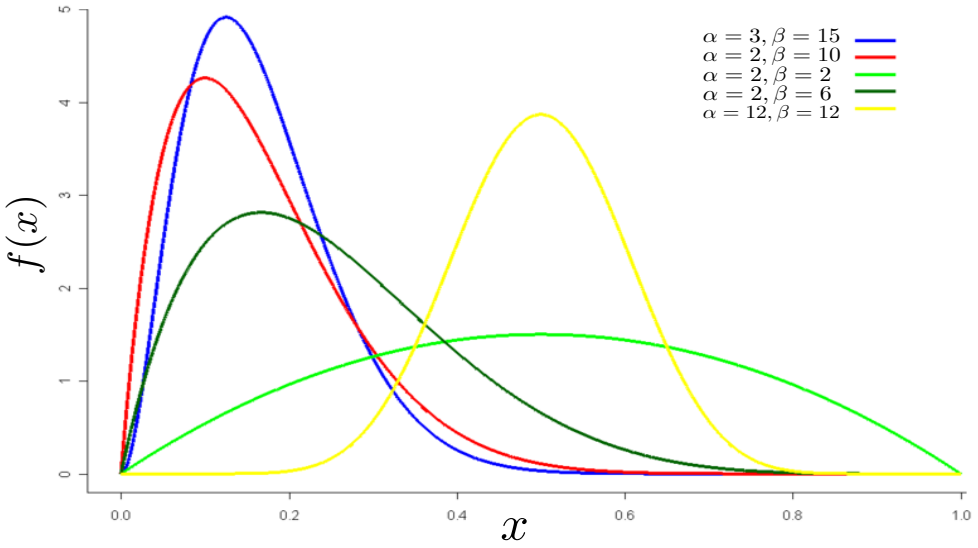
The flow of items through manufacturing or distribution processes

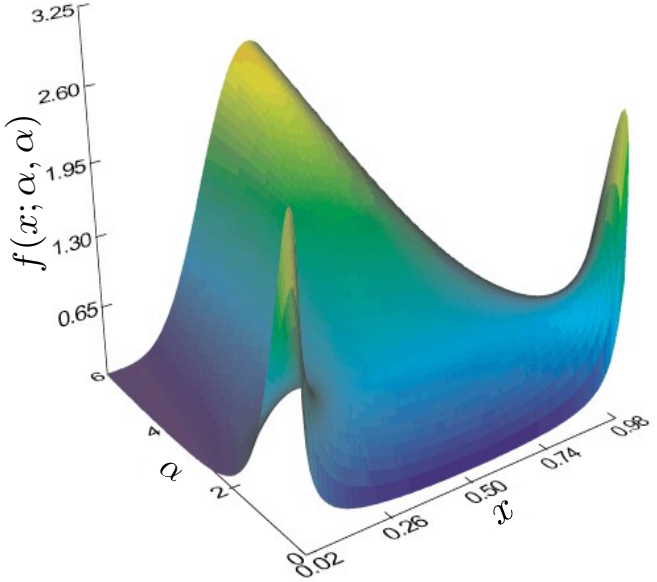
The load on web servers

Distributions

The Beta Distribution

$$P(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$





Distributions

The Beta Distribution

For an F distribution with v_1 and v_2 degrees of freedom,

$v_1F / (v_2 + v_1F)$ is Beta ($v_1/ 2, v_2/ 2$).

The distribution of R^2 is Beta($(p - 1)/2, (n - p)/2$) when $\rho = 0$

The time it takes to complete a task

The proportion of defective items in a shipment

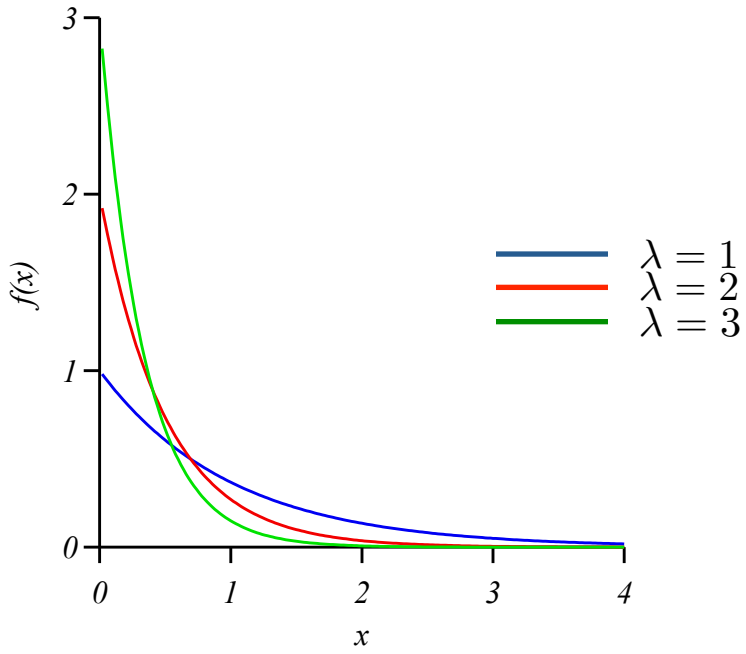
A distribution of prior probabilities

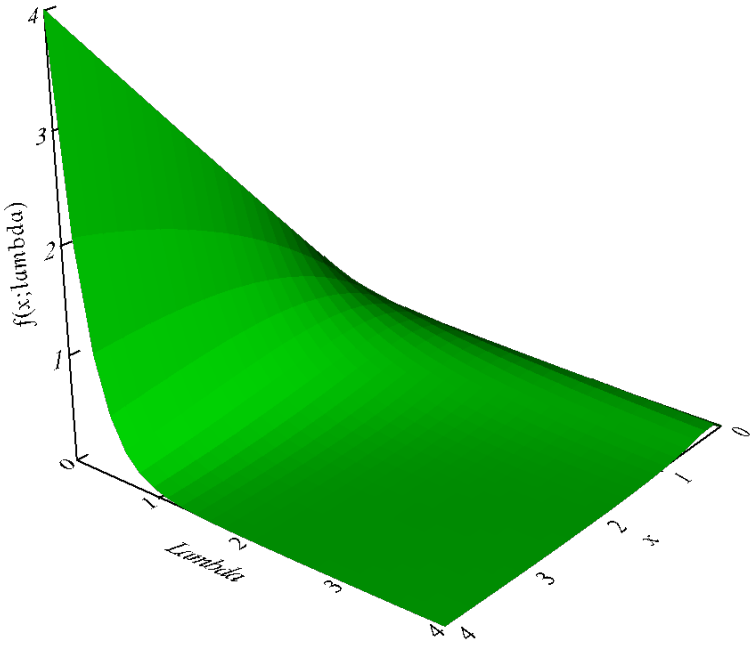
Approximates just about any distribution

Distributions

The Exponential Distribution

$$P(x;\lambda) = \lambda e^{-\lambda x}$$





Distributions

The Exponential Distribution

distance or time between random events

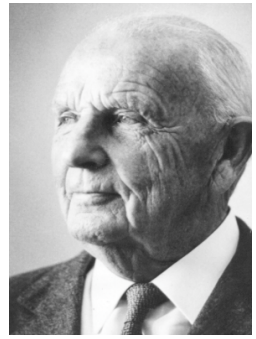
- mutations on a strand of DNA
- time between cars on a highway
- time before next telephone call
- distance between road kill

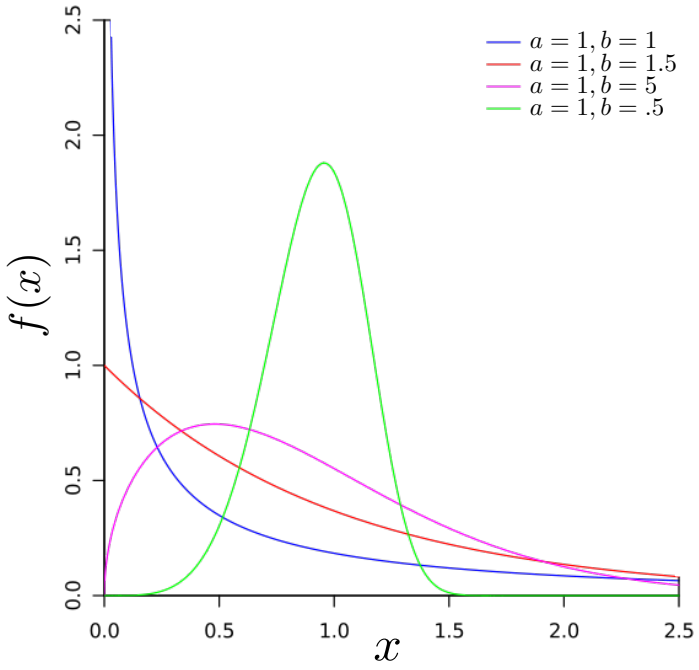
reliability of a machine

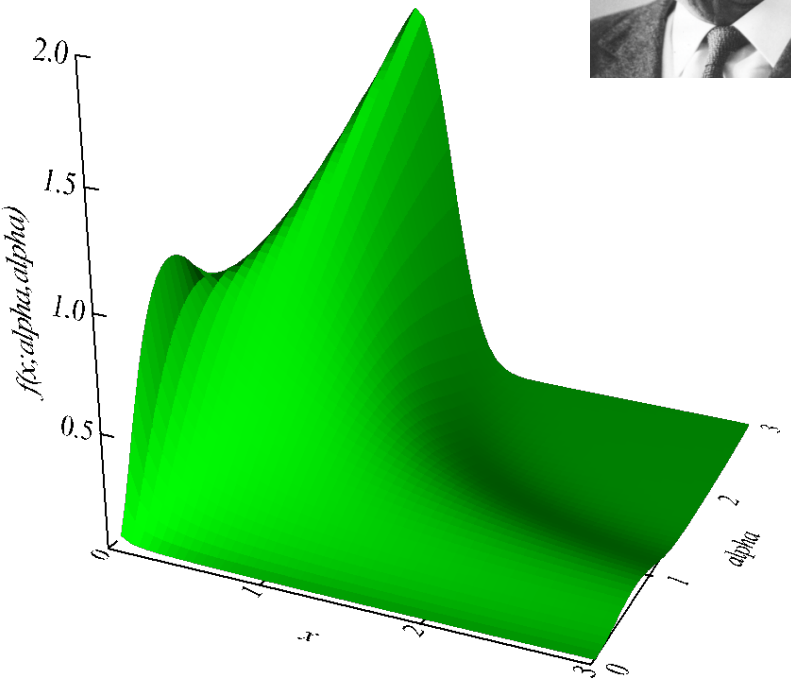
Distributions

The Weibull Distribution

$$P(x;a,b) = \begin{cases} abx^{b-1}e^{-ax^b} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$







Distributions

The Weibull Distribution

component failures

returns during warranty period

automobile recalls

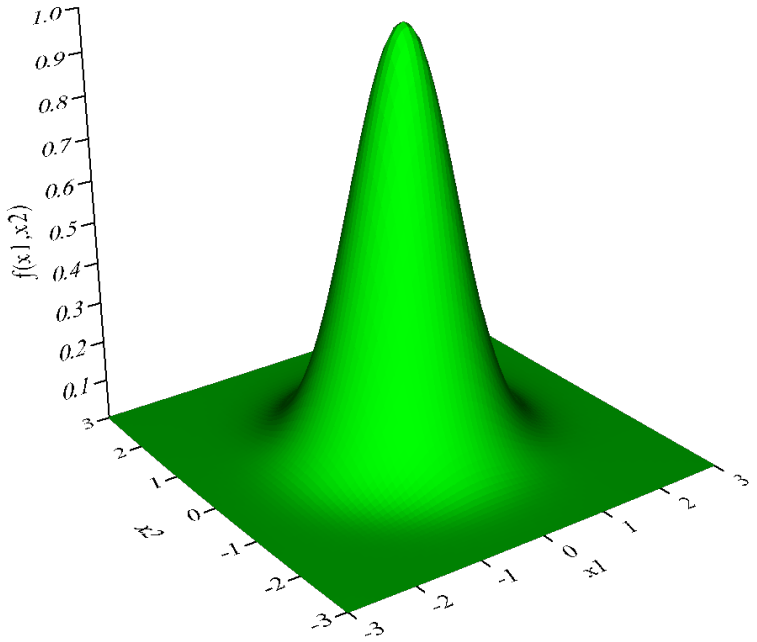
optimal replacement intervals for parts

survival analysis

Distributions

The Bivariate Error Function

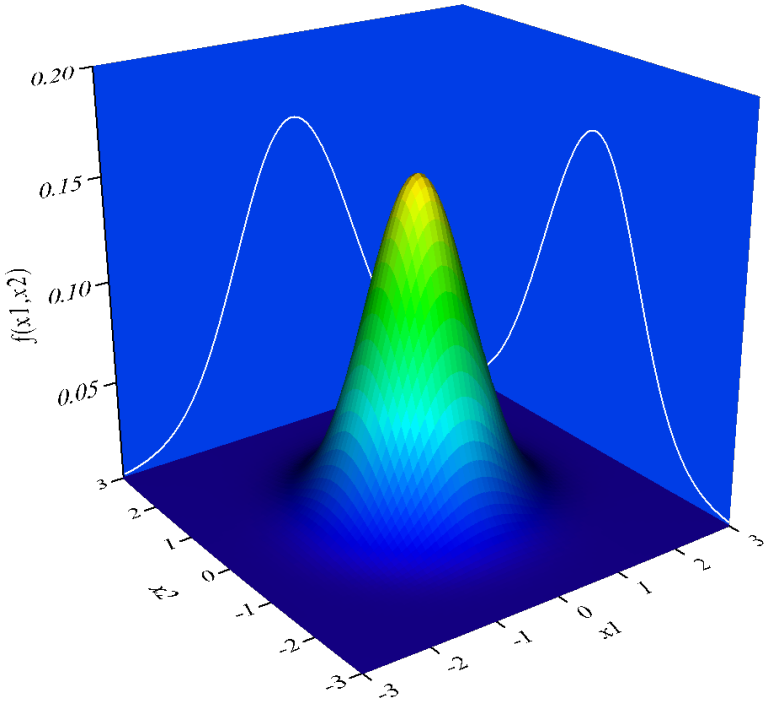
$$f(x_1, x_2) = e^{-(x_1^2 + x_2^2)}$$



Distributions

The Bivariate Standard Normal Distribution

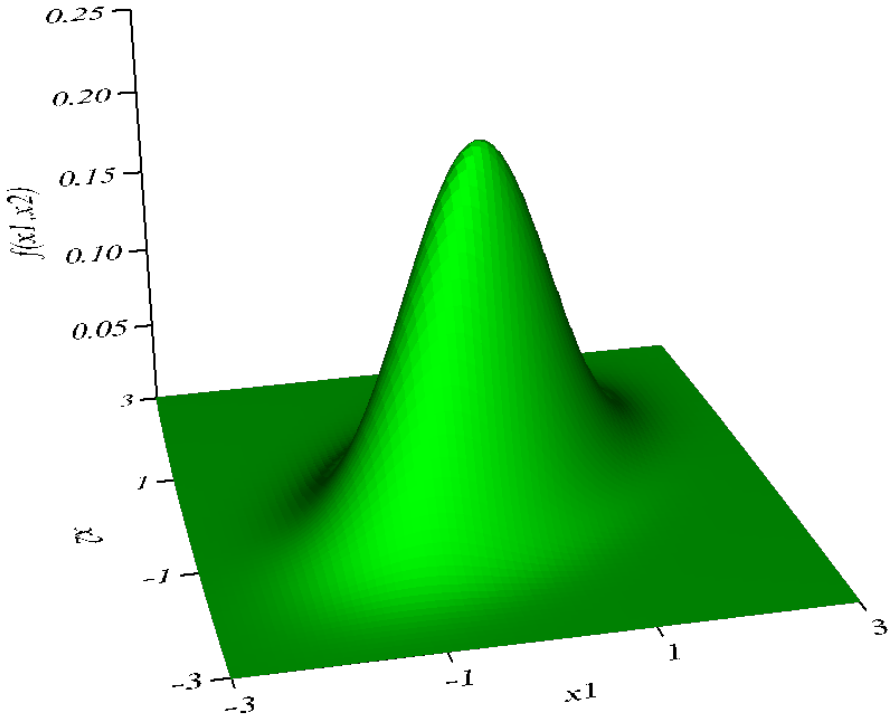
$$f(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)}$$



Distributions

The Bivariate Normal Distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

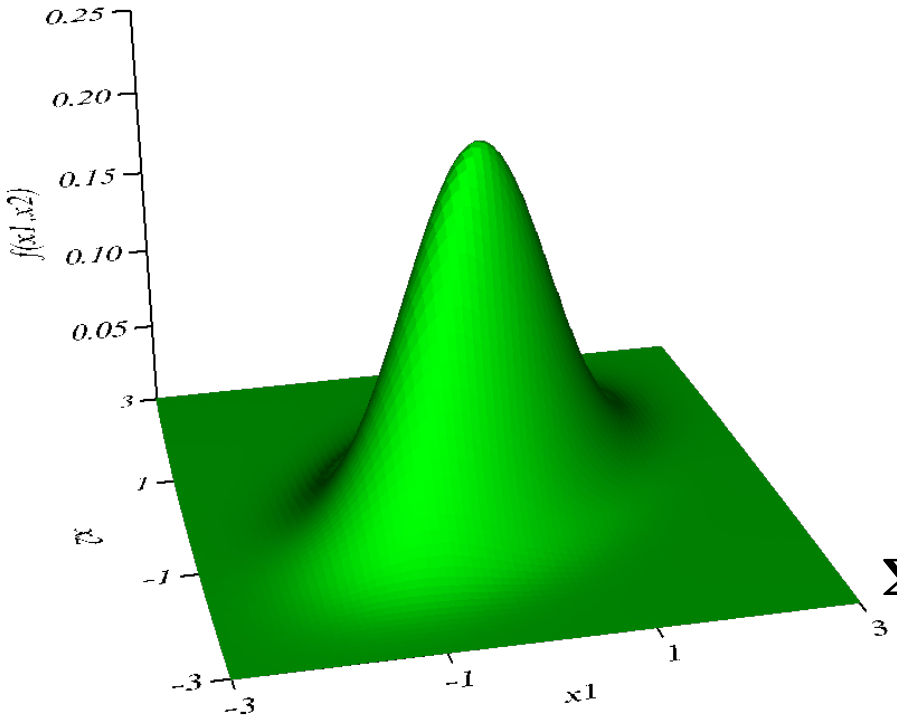
$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right\}}$$



Distributions

The Multivariate Normal Distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$



$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} {\sigma_1}^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & {\sigma_2}^2 \end{pmatrix}$$

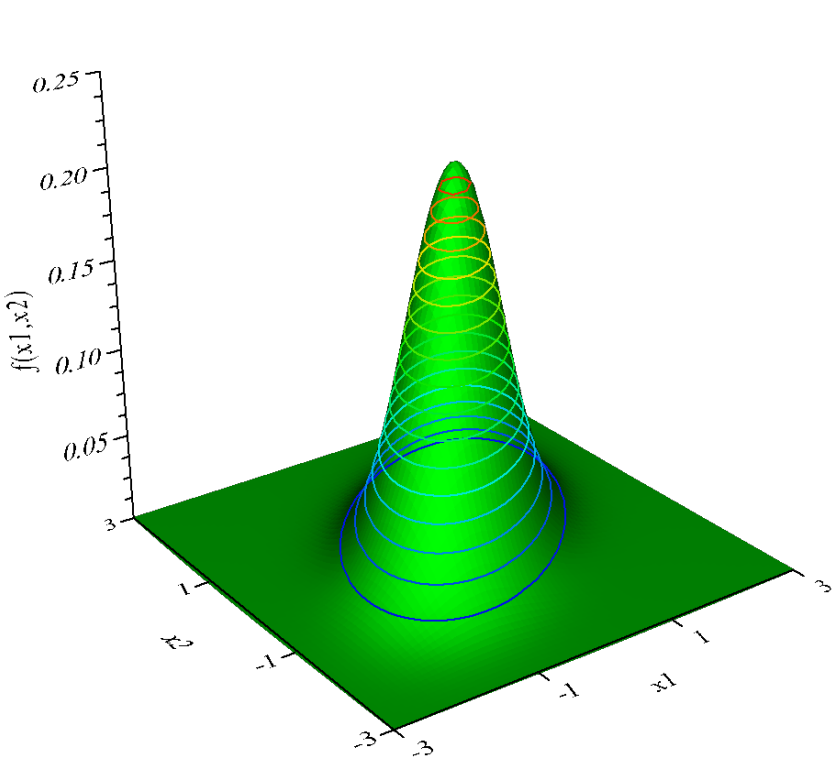
$$\begin{aligned} |\boldsymbol{\Sigma}| &= {\sigma_1}^2{\sigma_2}^2 - \rho\sigma_1\sigma_2\rho\sigma_2\sigma_1 \\ &= {\sigma_1}^2{\sigma_2}^2(1-\rho^2) \end{aligned}$$

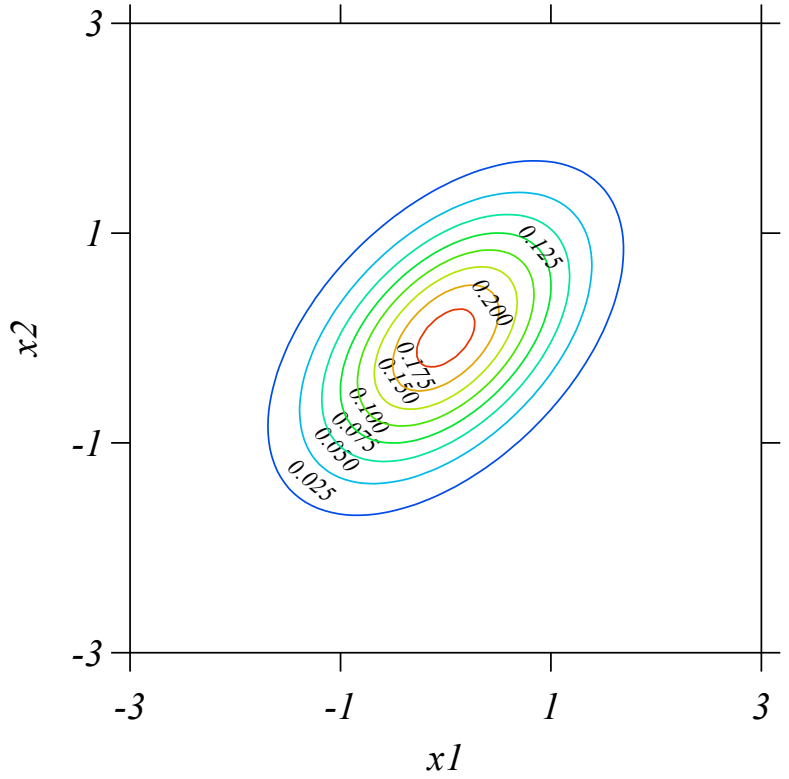
$$\boldsymbol{\Sigma}^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_2^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_2\sigma_1} & \frac{1}{{\sigma_1}^2} \end{pmatrix}$$

Distributions

Bivariate Normal instance of Elliptical Distribution

The slices are ellipses





Distributions

Mahalanobis Distance

Formula for unit circle $1 = x_1^2 + x_2^2$

Formula for ellipse $1 = ax_1^2 + bx_2^2$

Formula for rotated ellipse $1 = ax_1^2 + bx_2^2 - cx_1x_2$

Mahalanobis distance $\chi^2 = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$