Chapter 1

Introduction

1.1 Information about course

This is lecture material for the course ”Data-analysis and Inverse Methods in Astronomy”, DAIM in short. In Finnish, Tähtitieteen data-analyysi ja inversiomenetelmät. Course ID is 53834.

At least at this time, spring 2014, 8 credit points are rewarded from the course. To achieve these points you need to *i)* complete and return weekly exercises, *ii)* pass the final exam, and *iii)* complete a larger project of your own choice. At least 25 % of the weekly exercises need to be done in order to pass, and completing more will earn you a better grade.

Exercises will include both problems that are to be solved analytically, i.e. with pen and paper, and computer tasks that should be completed using some mathematical or statistical software on a computer. We do not specify what kind of software should be used, choose one you are most familiar with or one you would like to learn during the course. Programming or details about specific software are not taught, so you need to have prior knowledge on programming or scientific computing.

Most, if not all of the computer task are possible to do using any general purpose mathematical package such as Matlab, Mathematica, Maple etc. Statistical software packages such as R (free, under GNU GPL) is also an excellent choice for a tool. Lower-level programming tools such as Python can be used, but we do not recommend using very low-level programming such as C or Fortran, since too much effort would probably go to writing code for input/output and for producing graphics. On the other hand, software packages with limited amount of generality and versatile programming capabilities such as Excel or SPSS are not recommended either. The University of Helsinki has a license for SAS software, which is a huge statistical (among others) package that is used quite often in e.g. medical research and business applications, but perhaps because of its vast application areas and history, it is quite complected and a bit cumbersome to use.

Prior knowledge should include mathematical tools that are taught on basic university mathematics courses, e.g. Matemaattiset apuneuvot I and II (53704 and 53705) or Tähtitieteen matemaattiset menetelmät (53966). Especially we will need basic linear algebra and basic multivariate differential calculus.

1.1.1 Spring 2014

Course is held in Physicum, Wednesdays at 10-12 in class D117. The lecture dates are 22.1., 29.1., 5.2., 12.2., 19.2., 26.2., 12.3., 19.3., 26.3., 2.4., 9.4., 16.4., and 30.4. Some changes may happen, but up-to-date version of the dates can be found on the course homepage at `https://wiki.helsinki.fi/display/53834/`. Lecturers are Dr. Antti Penttilä (Antti.I.Penttila (a t) helsinki.fi) and Prof. Heikki Haario.

Course assistant is M.Sc. Olli Wilkman (Olli.Wilkamn (a t) helsinki.fi). If you will participate on the exercise sessions you can return your solutions there. If you cannot participate, you need to return your solutions *before* the session to Olli, either by email or to his mailbox in front of D308 in Physicum building. Exercise times will be announced on the course homepage.

1.1.2 Material

The course material, i.e. this handout and exercises, are based on the following course materials or books:

- A. Ekholm, ″Johdatus todennäköisyyslaskentaan″ and ″Johdatus uskottavuuspäättelyyn″, handouts
- S. Mustonen, Tilastolliset monimuuttujamenetelmät, book, University of Helsinki
- Course material for ″Data-analysis and Inverse Methods in Astronomy, 2012″ by M. Juvela, K. Muinonen, H. Haario and A. Penttilä
- P. Saikkonen, ″Lineaariset mallit″ and ″Epälineaariset mallit″, handouts
- C.P. Robert & G. Casella, Monte Carlo Statistical Methods, book, Springer

1.1.3 Notations

Throughout this material I will try to maintain a uniform and consistent style on symbol notations. If succeeded, the readability of the formulae will probably be better. Normal weight italic symbols are used for scalars: a, b, c, x. Random variables are usually written with capital letters: X, Y. For theoretical variables, i.e. parameters of distributions and/or theoretical and random properties of random variables such as the expected value or variance, Greek letters are usually used: μ, σ^2.

Functions are written with normal weight and non-italic font: $\sin(), \mathrm{P}()$. If possible, named distributions such as normal distribution are marked with calligraphic font, $\mathcal{N}(\mu, \sigma^2)$.

With multidimensional symbols bold weight is used. Vectors are with bold slanted symbols ($\boldsymbol{x}$, $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{\mu}$), and matrices with bold capital non-italics ($\mathbf{X}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{\Sigma}$). Vectors can be constructed from components as $\boldsymbol{x} = (1, 2, 3)$ — using () always refers to column vector, i.e. $n \times 1$ matrix. With [] we always refer to matrices, so $\boldsymbol{x} = [1, 2, 3]^T$ would be also (column)vector.

1.2 Random event, probability and random variable

The concept of random event, probability and random variable is very shortly introduced, since it is probably discussed in previous courses, and we are not going into details behind the philosophical or mathematical measure theory meanings of random variable.

Probability can be interpreted from frequentist viewpoint — if random phenomena or experiment is repeated and its outcome is statistically stable, the ratio of the number of events where result A is observed, n_A, and the number of all events n will estimate the the probability of A. In another words, $\mathrm{P}(A) \approx n_A/n$. Naturally, $0 \leq \mathrm{P}(A) \leq 1$. The actual value of $\mathrm{P}(A)$ may be unknown, but we assume that it is constant.

Frequentist interpretation has some caveats because we often want to consider probability of events that cannot strictly speaking be repeated. Probability is better interpreted through set theory. The sample space $\mathcal{S}$ includes all the possible events s_i. The sample space can be finite, countably infinite or uncountable infinite. All the probability calculus can be derived from three simple axioms for set A in $\mathcal{S}$:

$$\forall A \text{ holds that } \mathrm{P}(A) \geq 0 \tag{1.1}$$

$$\mathrm{P}(\mathcal{S}) = 1 \tag{1.2}$$

$$\text{If } A_1 \cap A_2 \cap \ldots \cap A_n = \emptyset \text{, then}$$

$$\mathrm{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} \mathrm{P}(A_i) \tag{1.3}$$

The third axiom tells that if events are mutually exclusive, the probability measure is additive. The third axiom also holds for infinite sets. This set theory interpretation of probability can often be graphically studied by means of Venn diagrams, see Fig. 1.1 for an example.

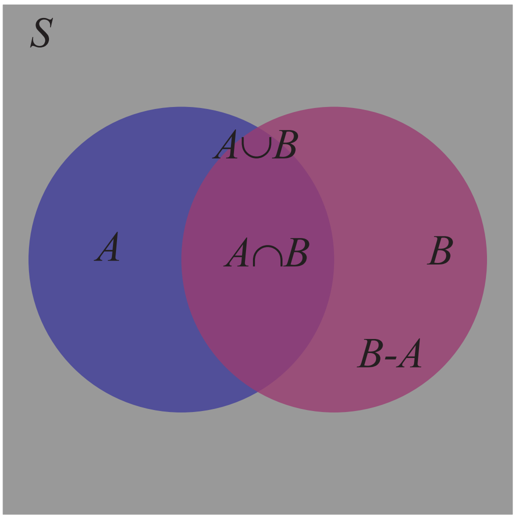

Figure 1.1: Example Venn diagram with some group theory sets.

1.2.1 Some probability laws

Laws of probability can be derived from the three axioms. Some simple and most common definitions are given here. In what follows we will write $A \cap B$ shorter with AB.

Addition:

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(AB) \tag{1.4}$$

that is valid also if $A \cap B \neq \emptyset$.

Conditional probability (*ehdollinen todennäköisyys*): Probability of event A requiring that B has happened, $\mathrm{P}(A|B)$.

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(AB)}{\mathrm{P}(B)} \tag{1.5}$$

Statistical independence (*tilastollinen riippumattomuus*): Events A and B are statistically (or stochastically) independent if and only if $\mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B)$. The usual notation for this is

$$A \perp\!\!\!\perp B \implies \mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B) \tag{1.6}$$

Chain rule:

$$\mathrm{P}(AB) = \mathrm{P}(B)\mathrm{P}(A|B) = \mathrm{P}(A)\mathrm{P}(B|A) \tag{1.7}$$

and theorem of total probability:

$$\mathrm{P}(B) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)\mathrm{P}(B|A_i) \tag{1.8}$$

when the sample space $\mathcal{S}$ has been partitioned into mutually exclusive sets $A_1, \ldots$

Bayes formula:

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(A_i)\mathrm{P}(B|A_i)}{\sum_{i=1}^{\infty}\mathrm{P}(A_i)\mathrm{P}(B|A_i)} \tag{1.9}$$

where $\mathrm{P}(A_i)$ is called prior probability and $\mathrm{P}(A_i|B)$ posterior probability.

We will prove and use some of these formulae in the exercises.

1.2.2 Random variable

Random variable (*satunnaismuuttuja*) is a mapping of the result of a random event into real axis. If Y is a random variable, then every possible outcome $s \in \mathcal{S}$ can be coded into real number y. For example, if there are only two possible outcomes, "A will happen, or A will not happen", it is often coded that $Y(A) = 1, Y(\text{not } A) = 0$.

Probability of certain random event to occur follows from set theory notation, $\mathrm{P}(Y = y)$. This is often written also as $\mathrm{P}_Y(y)$ or even as $\mathrm{P}(y)$ for short, if it is evident what random variable is considered. Evidently, from Eqs. (1.1) and (1.2) it follows that $0 \leq \mathrm{P}(Y = y) \leq 1$.

With discrete random variables are such that the set of possible outcomes is finite or countably infinite. Finite set can be for example three categories where the event will fall, and countable infinite set, for example, the set of natural numbers. It is possible that $\mathrm{P}(Y = y_i) = 0$ for some y_i, but from Eq. (1.2) it follows that there must be at least one y_i for which $\mathrm{P}(Y = y_i) > 0$.

Discrete variables can be divided into different scales according to their properties. The nominal scale is the most simple one. In nominal scale the outcome of the event is in finite set of 'categories' for which there is no natural order. An example would be the party a person is voting for. These categories are coded into numbers, but no arithmetic operations are meaningful with the numbers. One cannot say that category '1' is smaller than category '2'. The only possible probability description of nominal variable is to list the probabilities $\mathrm{P}(Y = y)$. The complete list of outcomes and associated probabilities is the *probability mass function* (*pistetodennäköisyysfunktio*)

$$\mathrm{f}(y) = \mathrm{P}(Y = y). \tag{1.10}$$

With ordinal scale variable the order of the categories is a meaningful concept. For example, many polls may ask if you "agree fully" $(Y = 4)$, "agree partly" $(Y = 3)$, "disagree partly" $(Y = 2)$, or "disagree strongly" $(Y = 1)$. In that case it is meaningful to claim that '4' is more than '3', although operations such as $4 - 3 = 1$ are not meaningful. For ordinal variable, in addition to probability mass function, a *cumulative distribution function* (*kertymäfunktio*) can be defined

$$\mathrm{F}(y) = \mathrm{P}(Y \leq y) = \sum_{u=1}^{y} \mathrm{f}(u). \tag{1.11}$$

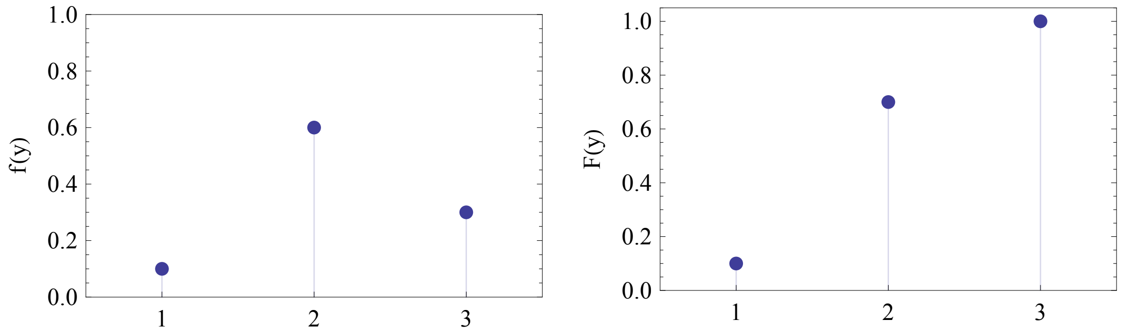

Figure 1.2: Example of probability mass function (on left) and cumulative distribution function (on right) for discrete random variable.

See Fig. 1.2 for examples.

The most advance scale for discrete variables is the interval scale. Variable has countable number of outcomes, and they can be ordered, and their intervals are meaningful and constant, i.e. $1 < 2 < 3$ and $2-1 = 3-2 = 1$. Both probability mass function and cumulative distribution function are defined. Furthermore, one can compute with the outcomes, and especially one can compute descriptive statistics such as mean, median or standard deviation.

Continuous variables are measured in interval or ratio scales. Ratio scale differs from interval scale by having unique and non-arbitrary zero value, but there are no real differences in using continuous interval or ratio scale variables in statistics. Most importantly, continuous variables are uncountable infinite. From that reason the probability of every single outcome is zero. Instead of probability mass function, a non-negative, real valued *probability density function* (pdf, *todennäköisyystiheysfunktio*) is defined so that

$$\mathrm{P}(y_0 < Y \leq y_1) = \int_{y_0}^{y_1} \mathrm{f}(u)du \;\; \text{for } y_0 < y_1. \tag{1.12}$$

The so-called probability density $\mathrm{f}(y)$ can be non-negative although the probability of single event is zero. The *cumulative density function* (cdf) for continuous random variable is defined as

$$\mathrm{F}(y) = \mathrm{P}(Y \leq y) = \int_{-\infty}^{y} \mathrm{f}(u)du. \tag{1.13}$$

See Fig. 1.3 for examples.

1.3 Descriptive statistics

The pdf or cdf of random variable is the complete description of the phenomenon, at least in mathematical sense. However, we often would like to compress that information into some set of numbers that would give us important information on

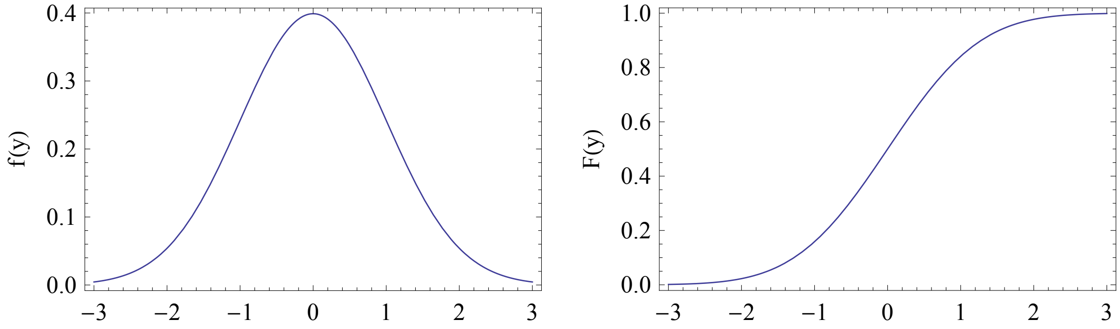

Figure 1.3: Example of probability density function (on left) and cumulative density function (on right) for discrete random variable.

the behavior of the random variable. These numbers are called *statistics* (*tunnusluvut*). In principle everything that is computed from pdf or from random sample is statistics, but there are some common choices on how distributions or samples are described.

We should remember to make clear difference between theoretical statistics and sample statistics. With theoretical statistics we mean quantities that can be derived from the pdf of a random variable, and that pdf might be unknown. The idea is that even though the distribution of random variable is unknown to us, it 'exists' and we can gather knowledge about it by observing the realized outcomes of the random variable. Theoretical statistics are often marked with Greek letters. The most common example of theoretical statistics and its sample counterpart is the expected value (μ) and the sample mean ($\overline{x}$). Actually, sample mean can also be thought to be random variable ($\overline{X}$) and the mean computed from one particular sample ($\overline{x}$) is the realization of that.

1.3.1 Expectation

The expected value (*odotusarvo*) of variable is the 'center of gravity' for a distribution. It is the most common statistics, and many distributions use it as a parameter. Expected value, or the expectation operator $\mathrm{E}(\cdot)$, is defined as

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y\,\mathrm{f}(y)\,dy \tag{1.14}$$

for continuous variable, and

$$\mathrm{E}(Y) = \sum_{y} y\,\mathrm{f}(y) \tag{1.15}$$

for discrete variable. It is said that the expectation does not exists unless the integral

$$\int_{-\infty}^{\infty} |y|\,\mathrm{f}(y)\,dy \tag{1.16}$$

converges, i.e. it has a finite value, and similarly but with sum instead of integral for discrete variable. Famous example of distribution without expected value is the Cauchy distribution.

Expectation is important statistics and is is useful to know some basic properties of $\mathrm{E}(\cdot)$ operator. First, it should be noted that a function of random variable is also a random variable, i.e. if $V = \mathrm{g}(Y)$ then V is a random variable. It can be shown that expectation of V can be derived without knowing the pdf of V by

$$\mathrm{E}(V) = \int_{-\infty}^{\infty} \mathrm{g}(y)\,\mathrm{f}(y)\,dy. \tag{1.17}$$

With discrete variable the same holds but with sum instead of integral. Another property is that expectation is a linear operator, i.e.

$$\mathrm{E}(Y_1 + \cdots + Y_n) = \mathrm{E}(Y_1) + \cdots + \mathrm{E}(Y_n) \tag{1.18}$$

$$\mathrm{E}(cY) = c\,\mathrm{E}(Y)\text{ , where } c \text{ is constant} \tag{1.19}$$

1.3.2 Variance

As expectation is a location measure, variance is a dispersion measure. It describes how much a random variable deviates from its expectation on average. Variance is derived as

$$\mathrm{var}(Y) = \mathrm{E}(Y - \mathrm{E}(Y))^2 = \int_{-\infty}^{\infty} (y - \mathrm{E}(Y))^2\,\mathrm{f}(y)\,dy \tag{1.20}$$

for continuous variable. Variance must be finite to exist. Instead of operators E and var symbols μ and σ^2 are often used.

Some properties of variance are dealt next. First,

$$\mathrm{var}(aY + b) = a^2\mathrm{var}(Y). \tag{1.21}$$

Second, for the variance of sum of *independent* variables $Y_1, \ldots, Y_n \perp\!\!\!\perp$ hold that

$$\mathrm{var}(Y_1 + \cdots + Y_n) = \mathrm{var}(Y_1) + \cdots + \mathrm{var}(Y_n), \tag{1.22}$$

but the same is generally not true if the variables are not independent.

1.3.3 Other statistics

Other commonly used statistics to describe the shape of the distribution include skewness (γ_1, *vinous*) and kurtosis (γ_2, *huipukkuus*). Both are derived from the central moments μ_k of distribution, $\mu_k = \mathrm{E}(Y - \mu)^k$, so that

$$\gamma_1 = \frac{\mu_3}{\sigma^3}\text{ , and } \gamma_2 = \frac{\mu_4}{\sigma^4} - 3. \tag{1.23}$$

Kurtosis is defined so that it is zero for standard normal distribution $\mathcal{N}(0,1)$. Skewness is zero for all symmetric distributions.

One important family of statistics are defined by quantiles. The p'th quantile is the value ξ for which

$$\mathrm{F}(\xi) = p. \tag{1.24}$$

Especially median is the quantile at $1/2$, the middle value of a distribution. Lower or first quartile is at $1/4$ and upper or third quartile at $3/4$. Median and other quartiles are so-called robust statistics, since their values are not heavily effect if the distribution has very wide tails, unlike expectation or variance, for example. An example of some of the abovementioned statistics is given in Fig. 1.4.

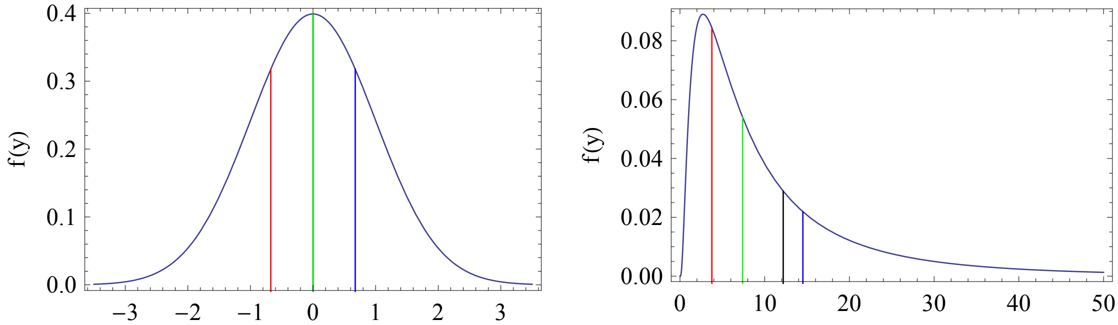

Figure 1.4: Symmetric distribution (normal) on left, and skew distribution (lognormal) on right. For both the place of expected value is marked with black line, median with green, and 1st and 3rd quartiles with red and blue. For symmetric distribution median and μ have the same value.

1.3.4 Covariance

We have not yet introduced multivariate random variables, but still it is best to mention covariance and correlation at this point. As said, covariance deals with two-dimensional random variable (U, V), and it measures the linear dependence between the variables. Definition for covariance is

$$\mathrm{cov}(U,V) = \mathrm{E}[(U - \mathrm{E}(U))(V - \mathrm{E}(V))] = \mathrm{E}(UV) - \mathrm{E}(U)\mathrm{E}(V) \tag{1.25}$$

Without proof we mention that the expectancy of product of two random variables is

$$\mathrm{E}(UV) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uv\,\mathrm{f}(u,v)\,du\,dv \tag{1.26}$$

for continuous variables. The $\mathrm{f}(u, v)$ is the joint distribution (*yhteisjakauma*) of U and V. Correlation is the covariance that is normalized with standard deviations

$$\mathrm{cor}(U,V) = \frac{\mathrm{cov}(U,V)}{\sigma_U \sigma_V} \tag{1.27}$$

Independence is wider concept than only linear independence, so zero covariance does not imply statistical independence, but the opposite is true,

$$U \perp\!\!\!\perp V \implies \mathrm{cov}(U,V) = \mathrm{cor}(U,V) = 0. \tag{1.28}$$

With the concept of covariance we can generalize the Eq. (1.22) about the variance of sum of independent variables to dependent ones,

$$\mathrm{var}(U+V) = \mathrm{var}(U) + \mathrm{var}(V) + 2\mathrm{cov}(U,V), \tag{1.29}$$

even when $U \not\perp\!\!\!\perp V$.

1.3.5 Sample statistics

All the abovementioned theoretical statistics all have their sample counterparts, or sample estimates (*otosestimaatti*), to be exact. The concept and derivation of estimate is introduced only in the next chapter, but for now we list formulae for these common statistics without proving their estimate properties.

Sample mean $\overline{x}$ is computed as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1.30}$$

(sample) standard error as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \tag{1.31}$$

and sample covariance as

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}). \tag{1.32}$$

The denominator $n-1$ is needed instead of n for the estimator to be unbiased, but this is again a topic of estimation theory and not dealt with here. Estimates for different quantiles are self-evident and can be made by sorting the sample with n observations and searching for k'th value so that $k/n = p$.

Mean and variance are not robust statistics. If the underlying distribution has heavy tails, i.e. the probability for extreme values is not 'small', the sample estimate may vary a lot from one sample to another. With astronomical observations, for example, sampling more and more is often not an option, so it is difficult to know whether observations come from heavy-tail distribution or not, or if some of the observations are simply wrong or affected by another process. Therefor, it is quite difficult to objectively say if some observations are outliers and should be left out from the analysis or not. However, duo to the large effect that 'unusual' observations can have in mean or variance estimates, they are sometimes left out, i.e. data is censored or trimmed. Common practices include e.g. trimming out observations with distance to mean larger than three standard deviations and then computing mean and variance again.

1.4 Distributions

The distribution, either probability mass function for discrete variable or probability density function for continuous, is the complete description of the random variable. Alternatively, cumulative functions can be used. One should note that all random variables have distribution, but that there are infinitive number of distributions and only few of them are 'known' in the sense that they are named and their formula is given. In this chapter we will list some univariate distributions and their statistics.

1.4.1 Discrete distributions

Bernoulli

Most simple discrete distribution is the Bernoulli distribution for binary random variable, i.e. with two possible outcomes, 0 and 1. If the probability of having 1 is π, then

$$Y \sim \mathcal{B}(\pi) \implies \mathrm{f}(y) = \pi^y (1-\pi)^{1-y}, \tag{1.33}$$

$$\mathrm{E}(Y) = \pi, \ \mathrm{var}(Y) = \pi(1-\pi), \ y \in \{0, 1\}. \tag{1.34}$$

Notice the notation, $Y \sim \mathcal{B}(\pi)$ should be read as Y has/obeys Bernoulli distribution with parameter π.

Binomial distribution

When more than one identical and independent Bernoulli trials are sampled, the total number of successes (outcome 1) is given by binomial distribution

$$Y \sim \mathrm{Bin}(n, \pi) \implies \mathrm{f}(y) = \frac{n!}{y!(n-y)!}\pi^y (1-\pi)^{n-y}, \tag{1.35}$$

$$\mathrm{E}(Y) = n\pi, \ \mathrm{var}(Y) = n\pi(1-\pi), \ y = 0, \ldots, n. \tag{1.36}$$

Poisson distribution

Poisson distribution can be used to model counts, i.e. how many times some (rare) event has occurred in one time unit. Good example could be the number of photons that hit the CCD sensor per time unit. When the intensity parameter, i.e. expected number of events per unit time, is λ, the distribution is

$$Y \sim \mathcal{P}(\lambda) \implies \mathrm{f}(y) = \exp(-\lambda)\frac{\lambda^y}{y!}, \tag{1.37}$$

$$\mathrm{E}(Y) = \lambda, \ \mathrm{var}(Y) = \lambda, \ y = 0, \ldots. \tag{1.38}$$

Examples of Poisson and binomial pdf's are shown in Fig. 1.5.

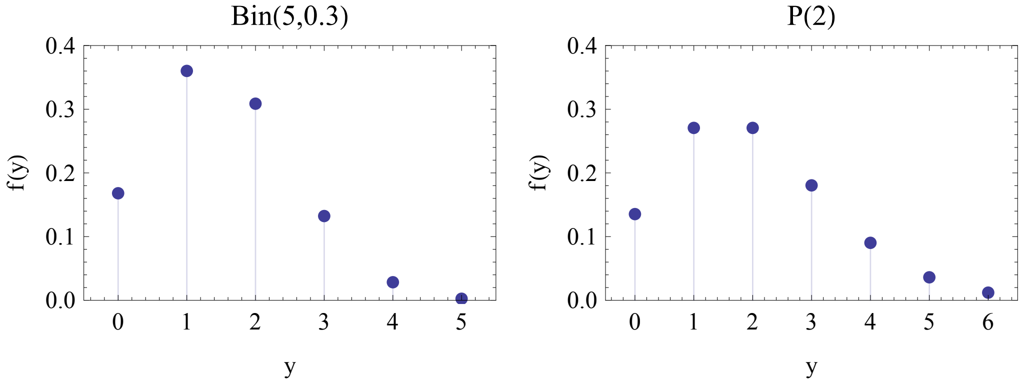

Figure 1.5: Pdf's of binomial (on left) and Poisson (on right) distributions.

1.4.2 Continuous distributions

Normal distribution

Normal distribution is by far the most common distribution due to the fact that it is the limiting distribution of many derived random variables by the central limit theorem, and thus can be used as approximative distribution to many otherwise too complicated or non-traceable distributions. Gauss derived the distribution to describe errors observed in the movements of planets and planetoids. With parameters μ and σ^2 the distribution is

$$Y \sim \mathcal{N}(\mu, \sigma^2) \implies \mathrm{f}(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \tag{1.39}$$

$$\mathrm{E}(Y) = \mu,\ \mathrm{var}(Y) = \sigma^2,\ y \in \mathbb{R}. \tag{1.40}$$

The term standardization (*standardointi*) means that the expected value (or mean) is subtracted from the original value, and the result is scaled (divided) with the standard deviation. This operation is not limited to normal distribution in any way, but if general normal variable $Y \sim \mathcal{N}(\mu, \sigma^2)$ is standardized, the results has $\mathcal{N}(0, 1)$ distribution, a.k.a. standard normal distribution.

The probability mass in normal distribution between $\mu - k\sigma$ and $\mu + k\sigma$ is approximately 68% with $k = 1$, 95% with $k = 2$, and 99% with $k = 3$. These are the famous one, two and three-sigma intervals that are commonly used in statistical tests and error limits.

The central limit theorem states that, under quite common conditions, pdf of the scaled sum Z of independent and identically distributed (i.i.d.) random variables approaches to normal distribution when the number of summed variables increases

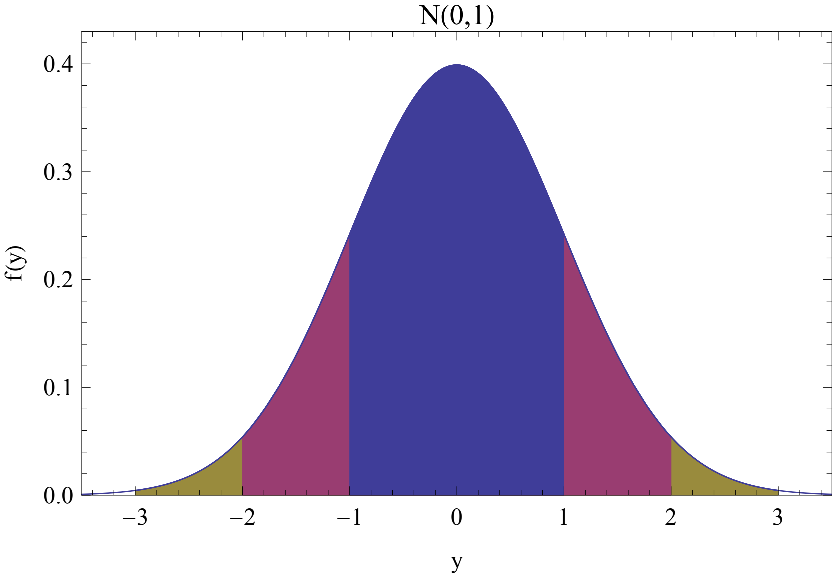

Figure 1.6: One (blue), two (red) and three-sigma (yellow) areas in normal distribution.

without limit. Precisely

$$\text{For i.i.d } Y_1, \dots, Y_n \text{ with } \mathrm{E}(Y_i) = 0 \text{ and } \mathrm{var}(Y_i) = \sigma^2, \tag{1.41}$$
$$Z = \frac{1}{\sqrt{n}} \sum_i^n Y_i \overset{approx}{\sim} \mathcal{N}(0, \sigma^2), \text{ as } n \to \infty.$$

This has evident implication to the sample mean $\overline{X}$ as a random variable, for large samples the sample mean should have normal distribution around the true, unknown mean, and the variance of sample mean around the true value is σ^2/n.

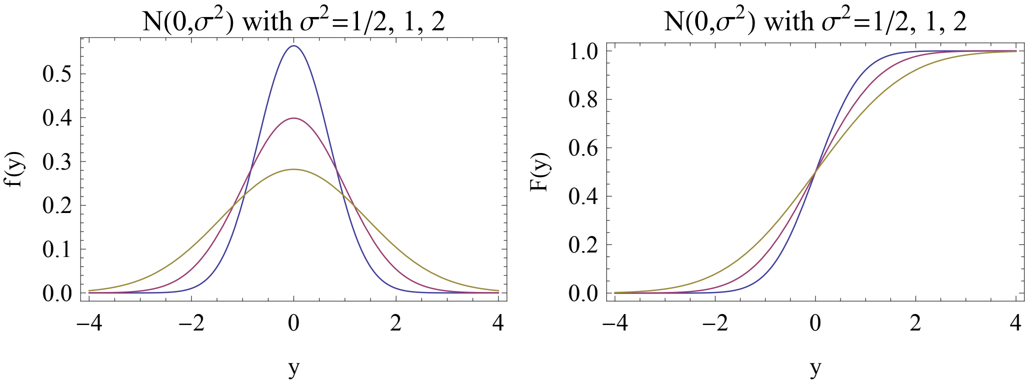

Figure 1.7: Pdf and cdf of normal distributions with different σ.

Exponential distribution

Exponential distribution can be used to model waiting times between two successive events from Poisson distributed variable. When intensity parameter (same interpretation as with Poisson) is λ, the distribution is

$$Y \sim \mathrm{Exp}(\lambda) \implies \mathrm{f}(y) = \lambda \exp(-\lambda y), \tag{1.42}$$

$$\mathrm{E}(Y) = 1/\lambda,\ \mathrm{var}(Y) = 1/\lambda^2,\ y \geq 0. \tag{1.43}$$

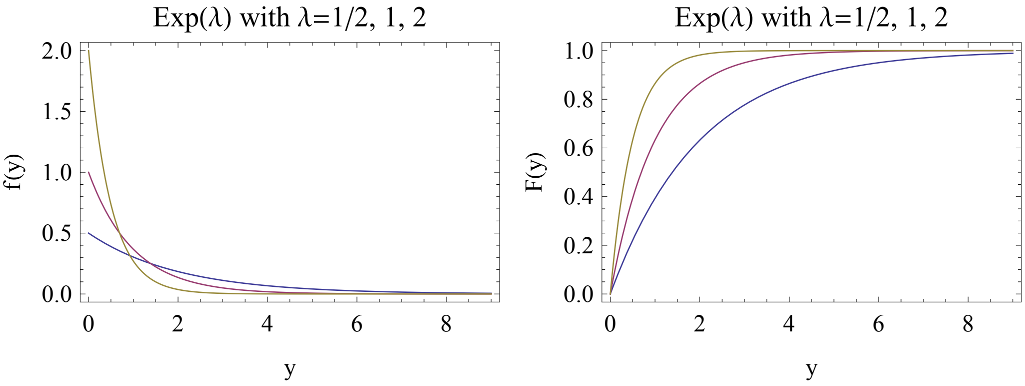

Figure 1.8: Pdf and cdf of exponential distributions with different λ.

Gamma distribution

Gamma distribution is a general case of exponential distribution, exponential is gamma with index $\kappa = 1$. Gamma is flexible distribution and is used to model lifetimes and other distances before event. With index κ and scale λ the distribution is

$$Y \sim \mathrm{Gamma}(\kappa, \lambda) \implies \mathrm{f}(y) = \frac{\lambda^\kappa y^{\kappa-1} \exp(-\lambda y)}{\Gamma(\kappa)}, \tag{1.44}$$

$$\mathrm{E}(Y) = \kappa/\lambda,\ \mathrm{var}(Y) = \kappa/\lambda^2,\ y \geq 0 \tag{1.45}$$

where $\Gamma()$ is the gamma function.

Log-normal distribution

Log-normal distribution is yet another distribution for positive-valued variable, and as its name suggest, it is the result of logarithm of normal-distributed variable. As the normal distribution can be justified through central limit theorem and sum of i.i.d. variables, log-normal is the limiting distribution for the product of

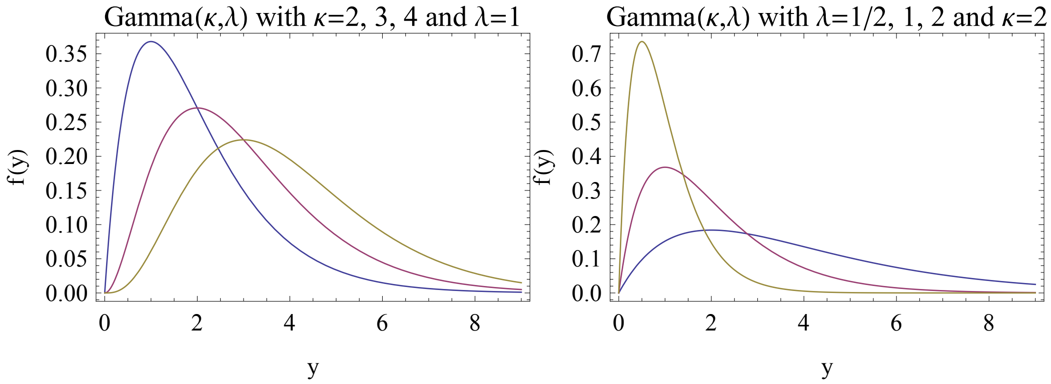

Figure 1.9: Pdf's of gamma distributions with different κ and λ.

i.i.d. variables. With parameters μ and σ^2, which refer to the underlying normal distribution, the log-normal distribution is

$$Y \sim \mathcal{LN}(\mu, \sigma^2) \implies \mathrm{f}(y) = \frac{1}{y\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln(y) - \mu)^2}{2\sigma^2}\right), \tag{1.46}$$

$$\mathrm{E}(Y) = \exp\left(\mu + \frac{1}{2}\sigma^2\right), \ \mathrm{var}(Y) = \exp(\sigma^2 - 1)\exp(2\mu + \sigma^2), \ y \geq 0. \tag{1.47}$$

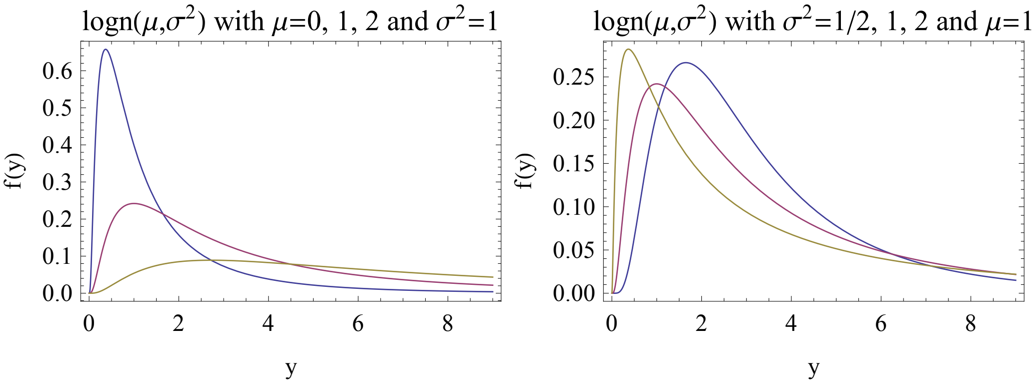

Figure 1.10: Pdf's of log-normal distributions with different μ and σ^2.

Distribution of function of random variable

Functions of random variable introduce new random variables which have their own distributions. The new distribution can be found by replacing original variable by inverse transform function and scaling by the derivative of the transform. More formally, let us have original random variable U with known distribution $\mathrm{f}_U(u)$, and a transform function from U to V: $V = \mathrm{g}(U)$. With the following method function g must be differentiable. With inverse transform $\mathrm{h}(V) = \mathrm{g}^{-1}(V) = U$ we can define that

$$\mathrm{f}_V(v) = \mathrm{f}_U(\mathrm{h}(v))\,|\frac{d\mathrm{h}(v)}{dv}| \tag{1.48}$$

Please note that if inverse transform $u = \mathrm{h}(v)$ is multiple-valued function, for example $u = \pm\sqrt{v}$, then all the possible pdf's must be summed together for $\mathrm{f}_V(v)$, e.g. $\mathrm{f}_V(v) = \mathrm{f}_U(-\sqrt{v})\,|d| + \mathrm{f}_U(\sqrt{v})\,|d|$.

1.5 Statistical plots

A large part of data analysis is to describe the data with methods that compress the important information with numbers (statistics) or with figures. We show here a few typical plots for one-dimensional data, and scatterplots for multi-dimensional data. Previous pages have already shown examples of probability distribution plots for both discrete and continuous variables. The corresponding plot for sample data is histogram.

Histogram collects data into bins, and plots the bins so that their height (discrete variable) or area (continuous variable) corresponds to the frequency of the observations in bins. If the purpose of histogram is to compare against theoretical distribution, the frequencies must be scaled so that their heights (discrete) or areas (continuous) sum up to one. The number of bins can be chosen freely, but one 'rule-of-thumb' suggests to use number of bins between $\sqrt{n}$ and $2\sqrt[3]{n}$ for data with n observations.

For data with outliers or otherwise long tails, the widths of the bins may differ in the histogram. Especially then one must remember that the area of the 'bar' in the histogram is what counts, not the height. An example is shown in Fig. 1.11.

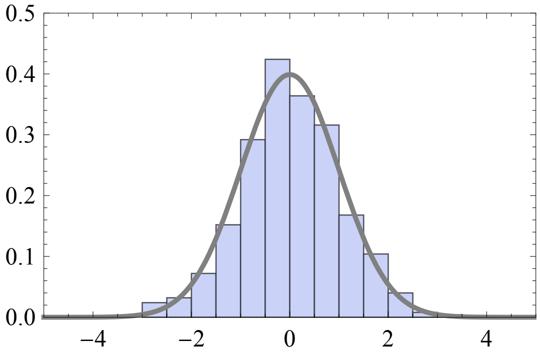

Figure 1.11: Pdf of normal distribution and histogram of 500 normal-distributed random numbers.

If distributions of several variables need to be compared in one figure, a box-and-whiskers plot is quite handy choice. Box-and-whiskers plot shows the range where data is, and its quartiles. In that way one gets a rough idea on how the data is spread, and about the symmetric / non-symmetric properties of the distribution and tails. The plot is drawn using smallest and largest values of data as 'whiskers', and a box from first to third quartile. Median or mean value is drawn in the middle of the box. Example in Fig. 1.12 will enlighten the principle. If there seems to be outliers in the data, the 'whiskers' might use, e.g., 1% and 99% quantiles as the endpoints instead of smallest and largest value.

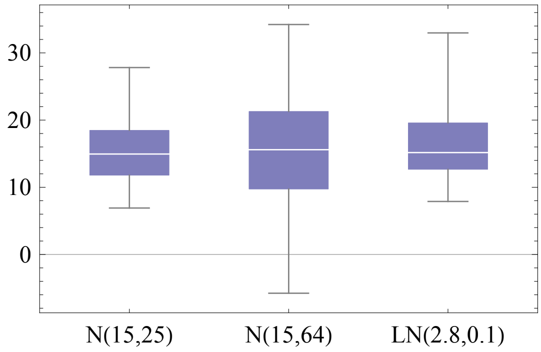

Figure 1.12: Box-and-whiskers plot of three samples of 100 observations from different distributions. First two are from normal distribution, and third from lognormal.

Scatterplots (*sirontakuviot*) are used to show dependence between two or more variables. With many variables the individual i vs. j plots can be organized into matrix of scatterplots.

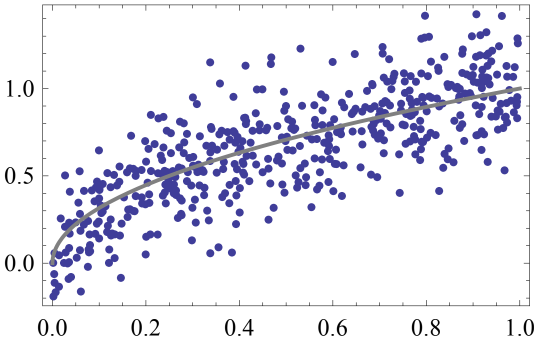

Figure 1.13: Scatterplot of data with $\sqrt{x}$-dependence and normally distributed errors.