18 Appendix: Some matrix algebra

18.1 Definition of matrix

A matrix A is an $n-$rows $m-$columns array of elements the elements are indicated by $a_{i,j}$ where the first index stands for row and the second for column. n and m are called the row and colum dimensions (sometimes shortened in "the dimensions") or sizes of the matrix A. Sometimes we write: A is a nxm matrix.

Sometimes a matrix is indicated as $A \equiv \{a_{ij}\}$.

When $n = m$ we say the matrix is square.

When the matrix is square and $a_{ij} = a_{ji}$ we say the matrix is symmetric.

When a matrix is made of just one row or one colum it is called a row (column) vector.

18.2 Matrix operations

1. Transpose: $A' = \{a_{ji}\}$. $A'' = A$. If A is symmetric then $A' = A$.

2. Matrix sum. The sum of two matrices $C = A + B$ is defined if and only if the dimensions of the two matrices are identical. In this case C has the same dimensions as A and B and $c_{ij} = a_{ij} + b_{ij}$. Clearly $A+B = B+A$ and $(A+B)' = A' + B'$

3. Matrix product. The product $C = AB$ of two matrices nxm and qxk is defined if and only if $m = q$. If this is the case C is a nxk matrix and $c_{ij} = \sum_l a_{il}b_{lj}$. In the matrix case it may well be that AB is defined but BA not. An important property is $C' = B'A'$ or, that is the same, $(AB)' = B'A'$. Provided the products and sums involved in what follows are defined we have $(A + B)C = AC + BC$.

18.3 Rank of a matrix

A row vector x is said to be linearly dependent from the row vectors of a matrix A if it is possible to find a row vector z such that $x = zA$. The same for a column vector. The number of linearly independent vectors

$r(A)$ (or $rank(A)$) the rank or a matrix A is defined as the number of linearly independent rows or (the number is the same) the number of linearly independent columns of A.

A square matrix of size n is called non singular if $r(A) = n$.

If B is any $n \times k$ matrix, then $r(AB) \leq \min(r(A), r(B))$.

If B is an $n \times k$ matrix of rank n, then $r(AB) = r(A)$.

If C is an $l \times m$ matrix of rank m, then $r(CA) = r(A)$.

18.4 Some special matrix

1. A square matrix A with elements $a_{ij} = 0, i \neq j$ is called a diagonal matrix.

2. A diagonal matrix with the diagonal of ones is called identity and indicated with I. $IA = A$ and $AI = A$ (if the product is defined).

3. A matrix which solves the equation $AA = A$ is called idempotent.

18.5 Determinants and Inverse

There are several alternative definitions for the determinant of a square matrix.

The Leibniz formula for the determinant of an $n \times n$ matrix A is $\det(A) = |A| = \sum_{\sigma \in S_n} sgn(\sigma) \prod_{i=1}^{n} A_{i,\sigma_i}$.

Here the sum is computed over all permutations σ of the set $1, 2, ..., n$. $sgn(\sigma)$ denotes the signature of σ; it is $+1$ for even σ and -1 for odd σ. Evenness or oddness can be defined as follows: the permutation is even (odd) if the new sequence can be obtained by an even number (odd, respectively) of switches of numbers.

The inverse of a square matrix A is the solution A^{-1} (or $inv(A)$) to the equations $A^{-1}A = I = AA^{-1}$.

If A is invertible then $(A')^{-1} = (A^{-1})'$

The inverse of a square matrix A exists if and only if the matrix is non singular that is if the size and the rank of A are the same.

A square matrix is non singular if and only if it has non null determinant.

$\det(A^{-1}) = 1/\det(A)$

If the products and inversions in the following formula are defined then $(AB)^{-1} = B^{-1}A^{-1}$.

Inversion has to do with the solution of linear non omogeneous systems.

Problem: find colum vector x such that $Ax = b$ with A and b given.

If A is square and invertible then the unique solution is $x = A^{-1}b$.

If A is $n \times k$ with $n > k$ but $r(A) = k$ then the system $Ax = b$ has no exact solution, however the system $A'Ax = A'b$ has the solution $x = (A'A)^{-1}A'b$.

18.6 Quadratic forms

A quadratic form with coefficient matrix given by the symmetric matrix A and variables vector given by the column vector x (with size of A equal to the number of rows of x) is the scalar given by:

$x'Ax = \sum_i \sum_j a_{ij} x_i x_j$.

A symmetric matrix A is called semi positive definite if and only if

$x'Ax \geq 0$ for all x

It is called positive definite if and only if

$x'Ax > 0$ for all non null x

If a matrix A can be written as $A = C'C$ for any matrix C ten A is surely at least spd. In fact $x'Ax = x'C'Cx$ but this is the product of the row vector $x'C'$ times itself, hence a sum of squares and this cannot be negative. It is also possible to show that any psd matrix can be written as $C'C$ for some C.

18.7 Random Vectors and Matrices (see the following appendix for more details)

A random vector, resp matrix, is simply a vector (matrix) whose elements are random variables.

18.8 Functions of Random Vectors (or Matrices)

- A function of a random vector (matrix) is simply a vector (or scalar) function of the components of the random vector (matrix).
- Simple examples are: the sum of the elements of the vector, the determinant of a random matrix, sums or produts of matrices and vectors and so on.
- We shall be interested in functions of the vector (matrix) X of the kind: $Y = A + BXC$ where A, B and C are non stochastic matrices of dimensions such that the sum and the products in the formula are well defined.
- A quadratic form $x'Ax$ with a non stochastic coefficient matrix A and stochastic vector x is and example of non linear, scalar function of a random vector.

18.9 Expected Values of Random Vectors

- These are simply the vectors (matrices) containing the expected values of each element in the random vector (matrix).
- $E(X') = E(X)'$
- An important result which generalizes the linear property of the scalar version of the operator $E(.)$ for the general linear function defined above, is this $E(A + BXC) = A + BE(X)C$.

18.10 Variance Covariance Matrix

- For random column vectors, and here we mean vectors only, we define the variance covariance matrix of a column vector X as:

$$V(X) = V(X') = E(XX') - E(X)E(X') = E((X - E(X))(X - E(X))')$$

- The Varcov matrix is symmetric, on the diagonal we have the variances ($V(X_i) = \sigma^2_{X_i}$) of each element of the vector while in the upper and lower triangles we have the covariances ($Cov(X_i; X_j)$).

- The most relevant property of this operator is:

$$V(A + BX) = BV(X)B'$$

- From this property we deduce that varcov matrices are always (semi) positive definite as. In fact if $A = V(z)$ x is a column vector of the same size as z, then $V(x'z) = x'Ax$ which cannot be negative for any possible x.

18.11 Correlation Coefficient

- The correlation coefficient between two random variables is defined as:

$$\rho_{X_i;X_j} = \frac{Cov(X_i; X_j)}{\sigma_{X_i}\sigma_{X_j}}$$

  The correlation matrix $\varrho(\mathbf{X})$is simply the matrix of correlation coefficients or, that is the same, the Varcov matrix of the vector of standardized X_i.

- The presence of a zero correlation between two random variables is defined, sometimes, linear independence or orthogonality. The reader should be careful using these terms as they exist also in the setting of linear algebra but their meaning, even if connected, is slightly different. Stochastic independence implies zero correlation, the reverse proposition is not true.

18.12 Derivatives of linear functions and quadratic forms

Often we must compute derivatives of functions of the kind $x'Ax$ (a quadratic form) or $x'q$ (a linear combination of elements in the vector q with weights x) with respect to the vector x.

In both cases we are considering a (column) vector of derivatives of a scalar function w.r.t. a (column) vector of variables (commonly called a 'gradient'). There is a useful matrix notation for such derivatives which, in these two cases, is simply given by:

$$\frac{\partial x'Ax}{\partial x} = 2Ax$$

and

$$\frac{\partial x'q}{\partial x} = q$$

The proof of these two formulas is quite simple. In both cases we give a proof for a generic element k of the derivative column vector.

For the linear combination we have

$$x'q = \sum_j x_j q_j$$

$$\frac{\partial x'q}{\partial x_k} = q_k$$

For the quadrati form

$$\frac{\partial x'Ax}{\partial x'} = 2x'A$$

$$x'Ax = \sum_i \sum_j x_i x_j a_{i,j}$$

$$\frac{\partial \sum_i \sum_j x_i x_j a_{i,j}}{\partial x_k} = \sum_{j \neq k} x_j a_{k,j} + \sum_{i \neq k} x_i a_{i,k} + 2x_k a_{k,k} = \sum_{j \neq k} x_j a_{k,j} + \sum_{j \neq k} x_j a_{k,j} + 2x_k a_{k,k} = 2A_{k,.}x$$

Where $A_{k,.}$means the $k - th$ row of A and we used the fact that A is a symmetric matrix.

An important point to stress is that the derivative of a function with respect to a vector always has the same dimension as the vector w.r.t. the derivative is taken, in this case x, so, for instance

$$\frac{\partial x'Ax}{\partial x} = 2Ax$$

and not

$$\frac{\partial x'Ax}{\partial x} = 2x'A$$

(remember that A is symmetric).

18.13 Minimization of a PD quadratic form, approximate solution of overdetermined linear systems

Now Let us go back to the linear system $Ax = b$ with A an $n \times k$ matrix of rank k. If $n > k$ this system has no solution, however, let's try to solv a similar problem. By solving a system we wish for $Ax - b = 0$ in our case this is not possible so let us try anch change the problem to this $\min_x (Ax - b)'(Ax - b)$. In words try to minimize the sum of squared differences between Ax and b if you cannot make it equal to 0. we have

$$(Ax - b)'(Ax - b) = x'A'Ax + b'b - 2b'Ax$$

Now let us take the derivative of this w.r.t. x

$$\frac{\partial}{\partial x} x'AA'x + b'b - 2b'Ax = 2A'Ax - 2A'b$$

(remember the rule about the size of a derivatives vector). We now create a new linear system equating these derivatives to 0.

$$A'Ax = A'b$$

And the solution is

$$x = (A'A)^{-1}A'b$$

This is the "least squares" approximate solution of a (overdetermined) linear system. (see the Appendix on least squares and Gauss Markov model).

18.14 Minimization of a PD quadratic form under constraints. Simple applications to finance

Suppose we are given a column vector r where $r_i$is the random (linear) return for the stock j.

Suppose we are holding these returns in a portfolio for one time period and that the (known) relative amount of each stock in our portfolio is given by the column vector w such that $1'w = 1$ where 1 indicates a column vector of ones of the same size as w.

Then the random linear return of the portfolio over the same time period is given by $r_\pi = w'r$.

Since w is known we have $E(w'r) = w'E(r)$ and $V(w'r) = w'V(r)w$.

The fact that, over one period of time, the expected linear return and the variance of the linear return of a portfolio only depend on the expected values and the covariance matrix of the single returns and the weight vector is what allows us to implement a simple optimization theory. For the moment let us suppose that the problem is

$$\min_{w:1'w=1} w'V(r)w$$

In this problem we want to minimize a quadratic form under a linear constraint.

It is to be noticed that, without the constraint, the problem would be solved by $w = 0$ (no investment). The constraint does not allow for this.

Such problems can be solved with the Lagrange multiplier method.

The idea is to artificially express, in a single function, both the need of minimizing the original function and the need to do this with respect to the constraint $1'w = 1$.

In order to do this we define the Lagrangean of the problem given by

$$L(x, \lambda) = w'V(r)w + 2\lambda(1'w - 1)$$

In this function th value of the unconstrained objective function is summed with the value of the constraint multiplied by a dummy parameter 2λ.

We now take the derivatives of the Lagrangean w.r.t. w and λ.

$$\frac{\partial}{\partial w}(w'V(r)w + 2\lambda(1'w - 1)) = 2V(r)w + 2\lambda 1$$
$$\frac{\partial}{\partial \lambda}(w'V(r)w + 2\lambda(1'w - 1)) = 2(1'w - 1)$$

If we set both these to zero we get, supposing $V(r)$ invertible

$$V(r)w = \lambda 1$$
$$1'w = 1$$

Notice the difference between the1-sIn the first equation 1 is a column vector which is required because we cannot equate a vector to a scalar. The same for 1' in the second equation which while the r.h.s. is a scalar one (for dimension compatibility with the l.h.s.). We do not stress this using, e.g., boldface for the vector 1 becouse the meaning follows unanbiglusly from the context.

It is clear that the second equation is satisfied if and only if w satisfies the constraint.

What is the meaning of the first equation (or, better, set of equations)? The unconstrained equation would have been

$$V(r)w = 0$$

whose only solution (due to the fact that $V(r)$is invertible) would be $w = 0$. But this solution does not satisfy the constraint. We still try to get $V(r)w = 0$ but we can't, due to the constraint. What we shall be able to get is $V(r)w = \lambda 1$. For some λ chosen in such a way that the constraint is satisfied.

To find this λ, simply put together the result of the first set of equations: $w = \lambda V(r)^{-1}1$ and the equation expressing the constraint: $1'w = 1$. Both equations are satisfied if and only if

$$\lambda = 1/1'V(r)^{-1}1$$

We now know λ, that is we know of exactly how much we must violate the unconstrained optimization condition (first set of equations) in order to satisfy the constraint (second equation).

In the end, putting this value of λin the solution for the first set of equations, we get

$$w = \frac{V(r)^{-1}1}{1'V(r)^{-1}1}$$

It is to be noticed that these are only necessary conditions but, for our purposes, this is enough.

What we got is the one period "minimum variance portfolio" made of securities whose returns covariance is $V(r)$.

What is the variance of this portfolio?

$$V(w'r) = w'V(r)w = \frac{1'V(r)^{-1}1}{(1'V(r)^{-1}1)^2} = \frac{1}{1'V(r)^{-1}1}$$

The expected value shall be

$$E(w'r) = w'E(r) = \frac{1V(r)^{-1}E(r)}{1'V(r)^{-1}1}$$

18.15 The linear model in matrix notation

Suppose you have a matrix X of dimensions $n \times k$ containing n observations on each of k variables. You also have a $n \times 1$ vector y containing n observations on another variable.

You would like to approximate y with a linear function of X that is: Xb for some $k \times 1$ vector b.

In general, if $n > k$ it shall be possible to exactly fit Xb to y so that the approximation shall imply a vector of errors $\epsilon = y - Xb$.

You would like to minimize ϵ but this is a vector, we must define some scalar function of it we wish to minimize.

A possible solution is $\epsilon'\epsilon$ that is: the sum of sqares of the errors.

We then wish to minimize

$$\epsilon'\epsilon = (y - Xb)'(y - Xb) = y'y + b'X'Xb - 2y'Xb$$

If we take the derivative of this w.r.t b we get

$$\frac{\partial}{\partial b}(y'y + b'X'Xb - 2y'Xb) = 2X'Xb - 2X'y$$

(again remember the size rule and remember that $y'Xb = b'X'y$ each is the transpose of the other but both are scalars).

The solution of this is

$$b = (X'X)^{-1}X'y$$

This simple application of the rule for the approximate solution of an overdetermined system yields tha most famous formula in applied (multivariate) statistics. Whet this problem, for the moment yust a best fit problem, shall be immersed in the appropriate statistical setting, our b shall become the Ordinary Least Squares parameter vector and shall be of paramount relevance in a wide range of applications to economics and finance.

19 Appendix: What you cannot ignore about probability and statistics

Why a Preliminary Course

- The finance master is conceived as a postgraduate course and contains a sizable quantitative section.
- The actual useful development of the program stands on the requirement, for the student, to posses a set (actually not very big) of prerequisite notions which shall be given as granted at the beginning of the master itself.
- This course, while quite introductory, is no exception.
- A student coming from Bocconi undergraduate programs in finance or in economics should posses more than enough knowledge for covering these requirements.
- Students coming from different Universities should, with all probability, have followed similar programs.
- Since, however, at the beginning of a more advanced course there could be some uncertainty about the required initial level of knowledge, we provide the following summary with the only purpose of describing the bare minimal notions, in probability and statistical inference, required for beginning this course (and more in general the two year master).
- The best use of this summary is to read it and to spend some time on the exercises most connected with it in past exams (see previous section).
- If everything is clear and known, no problem, otherwise spending some time with the teacher during office hours in order to agree on some further study could be useful.

Probability

19.1 Probability: a Language

- Probability is a language for building decision models.

- As all languages, it does not offer ready made splendid works of art (that is: right decisions) but simply a grammar and a syntax which point to avoiding inconsistencies. We call this grammar and this syntax "probability calculus".

- On the other hand, any language makes it simple to "say" something, difficult to say something else and there are concepts that cannot be even thought in a ny given language. So, no analysis of what we write in a language is independent on the structure of the language itself, And this is true for probability too.

- The language is useful to deduce probabilities of certain events when other probabilities are given, but the language itself tells us nothing about how to choose such probabilities.

19.2 Interpretations of Probability

- A lot of (often quite cheap) philosophy on the empirical meaning of probability boils down to two very weak suggestions: for results of replicable experiments it may be that probability assessments have to do with long run (meaning what?) frequency;

- For more general uncertainty situations, probability assessments may have something to do with prices paid for bets, provided you are not directly involved in the result of the bet except with regard to a very small sum of money.

- In simple situations, where some symmetry statement is possible, say the standard setting of "games of chance", the probability of relevant events can be reduced to some sum of probabilities of "elementary events" you may accept as "equiprobable".

19.3 Probability and Randomness

- Probability is, at least in its classical applications, introduced when we wish to model a collective "random" phenomenon, that is an instance where we agree that something is happening "under constant conditions" and, this not withstanding, the result is not fully determined by these conditions.

- Traders are interested in returns from securities, actuaries in mortality rates, physicists in describing gases or subatomic particles, gamblers in assessing the outcomes of a given gamble.

- At different degrees of confidence, students in these fields would admit that, in principle, it could be possible to attempt a specific modeling for each instance of the phenomena they observe but that, in practice, such model would require

such a precision in the measurement of initial conditions and parameters to be useless. Moreover computations for solving such models would be unwieldy even in simple cases.

- For these reasons students in these fields are satisfied with a theory that avoids a case by case description but directly models frequency distributions for collectives of observations and uses the probability language for these models.

19.4 Different Fields: Physics

- Quantum physics seems the only field where the "in principle" clause is usually not considered valid.

- In Statistical Physics a similar attitude is held but for a different reason. Statistical physics describes pressure as the result of "random" hits of gas molecules on the surface of a container. In doing this they refrain using standard arguments of mechanics of single particle not because this would be in principle impossible but because the resulting model would be in practice useless (for instance its solution would depend on a precise measurement of position and moment of each gas molecule, something impossible to accomplish in practice).

19.5 Finance

- Finance people would admit that days are by no means the same and that prices are not due to "luck" but to a very complex interplay of news, opinions, sentiments etc. However, they admit that to model this with useful precision is impossible and, at a first level of approximation, days can be seen as similar and that it is interesting to be able to "forecast" the frequency distribution of returns over a sizable set of days.

- The attitude is similar to Statistical Physics where, however, hypotheses of homogeneity of underlying micro behaviours are more easy to sustain. Moreover while we could model in an exact way few particles we cannot do the same even with a single human agent.

19.6 Actuarial Mathematics

- Actuaries do not try to forecast with ad hoc models the lifespan of this or that insured person (while they condition their models to some relevant characteristic the like of age, sex, smoker-no smoker and so) they are satisfied in a (conditional) model ling of the distribution of lifespan in a big population and in matching this with their insured population.

- Gambling
- Gamblers compute probabilities, and sometimes collects frequencies. They would like to be able to forecast each single result but their problem, when the result depends on some physical randomizing device (roulette, die, coin, shuffled deck of cards etc.) is exactly the same as the physicist's problem.

19.7 Wrong Models

- In a sense all probability models are then "wrong". With the exception (perhaps) of Quantum Mechanics, they do not describe the behaviour of each observable instance of a phenomenon but try, with the use of the non empirical concept of probability, to directly and at the same time fuzzily describe aggregate results: collective events.
- For this simple reason they are useful if our payout depends on collectives of events.
- They are not useful for predicting the result of the next coin toss but they are useful for describing coin tossING.

19.8 Meaning of Correct

- When we say that a probability model is "correct" (would be better to call it "satisfactory") we do mean that this model is a full successful description of facts but that its probability statement are well matched by empirical frequencies. Sometimes, probability model are used in cases when the relevant event shall happen only one or few times.
- In this case the model shall be useful for organizing our decision process non for describing its outcome.

19.9 Events and Sets

- Probabilities are assessed for "events" which are propositions concerning facts whose value of Truth can reasonably be assessed at a given future time. However, formally, probabilities are numbers associated with sets of points.
- Sets of points are indicated by capital letters: $A, B, C, ...$ The "universe" set (representing the sure event) is indicated with Ω and the empty set (the impossible event) with $\varnothing$(read: "ou").
- Finite or enumerably infinite collections of sets are usually indicated with $\{A_i\}_{i=1}^{n}$and with $\{A_i\}_{i=1}^{\infty}$.

- You are required to know the basic set theoretical operations: $A \cap B$ Intersection, $A \cup B$ Union, $A \setminus B$ Symmetric difference, $\overline{A}$ negation and their basic properties. The same is true for finite and enumerably infinite Unions and intersections: $\underset{i=1...n}{\cup} A_i$, $\underset{i=1...\infty}{\cup} A_i$ and so on.

19.10 Classes of Events

- Probabilities are assigned to classes of events which are usually assumed closed with regard to some set operations.

- The basic class is an Algebra, usually indicated with an uppercase calligraphic letter: $\mathcal{A}$. An algebra is a class of sets which include Ω and is closed to finite intersection and negation of its elements, that is: if two sets are in the class also their intersection and negation is in the class. This implies that also the finite union is in the class and so is the symmetric difference (why?).

- When the class of sets contains more than a finite number of sets, usually also enumerably infinite unions of sets in the class are required to be sets in the class itself (and so enumerable intersections, why?). In this case the class is called a σ-algebra. The name "Event" is from now on used to indicate a set in and algebra or σ-algebra.

19.11 Probability as a Set Function

- A probability is a set function P defined on the elements of an algebra such that: $P(\Omega) = 1$, $P(\overline{A}) = 1 - P(A)$ and for any finite number of disjoint events $\{A_i\}_{i=1}^n$ $(A_i \cap A_j = \emptyset \; \forall i \neq j)$ we have: $P(\underset{i=1...n}{\cup} A_i) = \sum_{i=1}^n P(A_i)$.

- If the probability is defined on a σ-algebra we require the above additivity property to be valid also for enumerable unions of disjoint events.

19.12 Basic Results

- A basic result, implied in the above axioms, is that for any pair of events we have: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- Another basic result is that if we have a collection of disjoint events: $\{A_i\}_{i=1}^n$ $(A_i \cap A_j = \emptyset \; \forall i \neq j)$ and another event B such that $B = \cup_{i=1}^n (A_i \cap B)$ then we can write: $P(B) = \sum_{i=1}^n P(B \cap A_i)$

19.13 Conditional Probability

- For any pair of events we may define the conditional probability of one to the other, say: $P(A|B)$ as a solution to the equation $P(A|B)P(B) = P(A \cap B)$.
- If we require, and we usually do: $P(B) \neq 0$, we have: $P(A|B) = P(A \cap B)/P(B)$.

19.14 Bayes Theorem

Using the definition of conditional probability and the above two results we can prove Bayes Theorem.

Let $\{A_i\}_{i=1}^n$be a partition of Ω in events, that is: $\{A_i\}_{i=1}^n$ $(A_i \cap A_j = \emptyset \, \forall i \neq j)$and $\bigcup_{i=1...n} A_i = \Omega$, we have:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

19.15 Stochastic Independence

- We say that two events are "independent in probability", "stochastically independent" or, simply, when no misunderstandings are possible, "independent" if $P(A \cap B) = P(A)P(B)$.
- If we recall the definition of conditional probability, we see that, in this case, the conditional probability of each one event to the other is again the "marginal" probability of the same event.

19.16 Random Variables

- These are functions $X(.)$ from Ω to the real axis $\mathbb{R}$.
- Not all such functions are considered random variables. For $X(.)$ to be a random variable we require that for any real number t the set B_t given by the points ω in Ω such that $X(\omega) \leq t$ is also an event, that is: an element of the algebra (or σ-algebra).
- The reason for this requirement (whose name is "measurability") is that a basic tool for modeling the probability of values of X is the "probability distribution function" (PDF) (sometimes "cumulative distribution function" CDF) of X defined for all real numbers t as: $F_X(t) = P(\{\omega\} : X(\omega) \leq t) = P(B_t)$ and, obviously, in order for this definition to have a meaning, we need all B_t to be events.

19.17 Properties of the PDF

- From its definition we can deduce some noticeable properties of F_X

1. it is a non decreasing function;
2. its limit for t going to $-\infty$ is 0 and its limit for t going to $+\infty$ is one;
3. we have: $\lim_{h\downarrow 0} F_x(t+h) = F_X(t)$ but this is in general not true for $h \uparrow 0$ so that the function may be discontinuous.

- We may have at most a enumerable set of such discontinuities (they are discontinuities of the first kind).
- Each of these discontinuities is to be understood as a probability mass concentrated on the value t where the discontinuity appears. Elsewhere F is continuous.

19.18 Density and Probability Function

- In order to specify probability models for random variables, usually, we do not directly specify F but other functions more easy to manipulate.
- We usually consider as most relevant two cases (while interesting mix of these may appear):

1. the absolutely continuous case, where F shows no discontinuity and can be differentiated with the possible exception of a set of isolated points
2. the discrete case where F only increases by jumps.

19.19 Density

In the absolutely continuous case we define the probability density function of X as: $f_X(t) = \frac{\partial F_X(s)}{\partial s}|_{s=t}$ where this derivative exists and we complete this function in an arbitrary way where it does not. Any choice of completion shall have the property: $F_X(t) = \int_{-\infty}^{t} f_X(s)ds$.

19.20 Probability Function

In the discrete case we call "support" of X the at most enumerable set of values x_i corresponding to discontinuities of F and we indicate this set wit $Supp(X)$and define the probability function $P_X(x_i) = F_X(x_i) - \lim_{h\uparrow 0} F_X(x_i + h)$ for all $x_i : x_i \in Supp(X)$ with the agreement that such a function is zero on all other real numbers.

19.21 Expected Value

The "expected value" of (in general) a function $G(X)$ is then defined, in the continuous and discrete case as

$$E(G) = \int_{-\infty}^{+\infty} G(s) f_X(s) ds$$

and

$$E(G) = \sum_{x_i \in Supp(X)} G(x_i) P_X(x_i)$$

If G is the identity function $G(t) = t$ the expected value of G is simply called the "expected value", "mathematical expectation", "mean", "average" of X.

19.22 Expected Value

- If G is a non-negative integer power: $G(X) = X^k$, we speak of "the k-th moment of X and usually indicate this with $m_k$or μ_k.
- If $G(X)$ is the function $I(X \in A)$, for a given set A, which is equal to 1 if $X = x \in A$ and 0 otherwise (the indicator function of A) then $E(G(X)) = P(X \in A)$.
- In general $E(G(X)) \neq G(E(X))$ with a noticeable exception: if $G(X) = aX + b$ with a and$b$ constants. In this case we have $E(aX + b) = aE(X) + b$. Sometimes the expected value of X is indicated with μ_X or simply μ.

19.23 Variance

- The "variance" of $G(X)$ is defined as $V(G(X)) = E((G(X) - E(G(X))^2) = E(G(X)^2) - E(G(X))^2$.
- A noticeable property of the variance is that such that $V(aG(X)+b) = a^2 V(G(X))$.
- The square root of the variance is called "standard deviation". For these two quantities the symbols σ^2 and σ are often used (with or without the underscored name of the variable).

19.24 Tchebicev Inequality

- A fundamental inequality which connects probabilities with means and variances is the so called "Tchebicev inequality":

$$P(|X - E(X)| < \lambda\sigma) \geq 1 - \frac{1}{\lambda^2}$$

- As an example: if$\lambda$is set to 2 the inequality gives a probability of at least 75% for X to be between its expected value + and - 2 times its standard deviation.
- Since the inequality is strict, that is: it is possible to find a distribution for which the inequality becomes an equality, this implies that, for instance, 99% probability could require a $\pm$ "10 σ" interval.
- For comparison, 99% of the probability of a Gaussian distribution is contained in the interval $\mu \pm 2.576\sigma$.
- These simple points have a great relevance when tail probabilities are computed in risk management applications.

19.25 Vysochanskij–Petunin Inequality

Tchebicev inequality can be refined by the Vysochanskij–Petunin inequality which, with the added hypothesis that the distribution be unimodal, states that, for any $\lambda > \frac{2}{\sqrt{3}} = 1.632993$

$$P(|X - \mu| < \lambda\sigma) \geq 1 - \frac{4}{9\lambda^2}$$

more than halving the probability outside the given interval given by Tchebicev: the 75% for $\lambda = 2$ becomes now $1 - \frac{1}{9}$ that is 88.(9)%.

19.26 Gauss Inequality

This result is an extension of a result by Gauss who stated that if m is the mode (mind not the expected value: in this is the V-P extension) of a unimodal random variable then

$$P(\mid X - m \mid < \lambda\tau) \geq \begin{cases} 1 - \left(\frac{2}{3\lambda}\right)^2 & \text{if } \lambda \geq \frac{2}{\sqrt{3}} \\ \frac{\lambda}{\sqrt{3}} & \text{if } 0 \leq \lambda \leq \frac{2}{\sqrt{3}}. \end{cases}$$

Where $\tau^2 = E(X - m)^2$.

19.27 Cantelli One Sided Inequality

A less well known but useful inequality is the Cantelli ore one sided Tchebicev inequality, which, phrased in a way useful for left tail sensitive risk managers, becomes :

$$P(X - \mu \geq \lambda\sigma) \geq \frac{\lambda^2}{1 + \lambda^2}$$

and for $\lambda = -2$ this means that at least $\frac{4}{5}$ of the probability (80%) is above the $\mu - 2\sigma$ lower boundary.

19.28 Quantiles

- The "α-quantile" of X is defined as the value $q_\alpha = \inf(t) : F_X(t) \geq \alpha$.
- Notice that in the case of a continuous random variable this equation could be written as $q_\alpha = t : F_X(t) = \alpha$.
- Notice, moreover, that in the discrete case the definition we gave above is just one of the possible definitions.

19.29 Median

- If $\alpha = 0.5$ we call the corresponding quantile the "median" of X and use for it, usually, the symbol M_d.
- It may be interesting to notice that, if G is continuous and increasing, we have $M_d(G(X)) = G(M_d(X))$.

19.30 Univariate Distributions Models

- Models for univariate distributions come in two kinds: non parametric and parametric.
- A parametric model is a family of functions indexed by a finite set of parameters (real numbers) and such that for any value of the parameters in a pre defined parameter space the functions are probability densities (continuous case) or probability functions (discrete case).
- A non parametric model is a model where The family of distributions cannot be indexed by a finite set of real numbers.
- It should be noticed that, in many applications, we are not interested in a full model of the distribution but in modeling only an aspect of it as, for instance, the expected value, the variance, some quantile and so on.

19.31 Some Univariate Discrete Distributions

- Bernoulli: $P(x) = \theta$, $x = 1$; $P(x) = 1 - \theta$, $x = 0$; $0 \leq \theta \leq 1$. You should notice the convention: the function is explicitly defined only on the support of the random variable. For the Bernoulli we have: $E(X) = \theta$, $V(X) = \theta(1 - \theta)$.
- Binomial: $P(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$, $x = 0, 1, 2, ..., n$; $0 \leq \theta \leq 1$. We have:$E(X) = n\theta$; $V(X) = n\theta(1 - \theta)$.

- Poisson:$P(x) = \lambda^x e^{-\lambda}/x!,\ x = 0, 1, 2, ..., \infty;\ 0 \leq \lambda$. We have:$E(X) = \lambda;\ V(X) = \lambda$.

- Geometric $P(x) = (1-\theta)^{x-1}\theta,\ x = 1, 2, ..., \infty;\ 0 \leq \theta \leq 1$. We have $E(X) = \frac{1}{\theta}$; $V(X) = \frac{1-\theta}{\theta^2}$

19.32 Some Univariate Continuous Distributions

Negative exponential: $f(x) = \theta e^{-\theta x},\ x > 0,\ \theta > 0$. We have: $E(X) = 1/\theta;\ V(X) = 1/\theta^2$. (Here you should notice that, as it is often the case for distributions with constrained support, the variance and the expected value are functionally related).

19.33 Some Univariate Continuous Distributions

Gaussian: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, x \in \mathbb{R},\ \mu \in \mathbb{R},\ \sigma^2 > 0$. We have $E(X) = \mu$,$V(X) = \sigma^2$. A very important property of this random variable is that, if a and b are constants, then $Y = aX + b$ is a Gaussian if X is a Gaussian.

By the above recalled rules on the E and V operators we have also that $E(Y) = a\mu+b; V(Y) = a^2\sigma^2$. In particular, the transform $Z = \frac{X-\mu}{\sigma}$ is distributed as a "standard" (expected value 0, variance 1) Gaussian.

19.34 Some Univariate Continuous Distributions

The distribution function of this random variable is usually indicated with Φ, so $\Phi(x)$ is the probability of observing values of the random variable X which are smaller then or equal to the number x, in short: $\Phi(x) = P(X \leq x)$. With $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ we indicate the inverse function of Φthat is: the value of the standard Gaussian which leaves on its left a given amount of probability. Obviously $\Phi(\Phi^{-1}(1-\alpha)) = 1-\alpha$.

19.35 Random Vector

- A random vector $\mathbf{X}$ of size n is a n- dimensional vector function from Ω to$\mathbb{R}^n$, that is: a function which assigns to each $\omega \in \Omega$ a vector of n real numbers.

- The name "random vector" is better than the name "vector of random variables" in that, while each element of a random vector is, in fact, a random variable, a simple vector of random variables could file to be a random vector if the arguments ω_i of the different random variables are not constrained to always coincide.

- (If you understand this apparently useless subtlety you are well on your road to understanding random vectors, random sequences and stochastic processes).

19.36 Distribution Function for a Random Vector

- Notions of measurability analogous to the one dimensional case are required to random vectors but we do not mention these here.
- Just as in the case of random variable, we can define probability distribution functions for random vectors as $F_{\mathbf{X}}(t_1, t_2, ..., t_n) = P(\{\omega\} : X_1(\omega) \leq t_1, X_2(\omega) \leq t_2, ..., X_n(\omega) \leq t_n)$ where the commas in this formulas can be read as logical "and" and, please, notice again that the $\omega$for each element of the vector is always the same.

19.37 Density and Probability Function

As well as in the one dimensional case, we usually do not model a random vector by specifying its probability distribution function but its probability function: $P(x_1, ..., x_n)$ or its density: $f(x_1, ..., x_n)$, depending on the case.

19.38 Marginal Distributions

- In the case of random vectors we may be interested in "marginal" distributions, that is: probability or density functions of a subset of the original elements in the vector.
- If we wish to find the distribution of all the elements of the vector minus, say, the i-th element we simply work like this:
- in the discrete case

$$P(x_1, ..., x_{i-1}, x_{i+1}, ...x_n) = \sum_{x_i \in Supp(X_i)} P(x_1, ..., x_{i-1}, , x_i, x_{i+1}...x_n)$$

- and in the continuous case:

$$f(x_1, ..., x_{i-1}, x_{i+1}, ...x_n) = \int_{x_i \in Supp(X_i)} f(x_1, ..., x_{i-1}, x_i, x_{i+1}...x_n) dx_i$$

- We iterate the same procedures for finding other marginal distributions.

19.39 Conditioning

- Conditional probability functions and conditional densities are defined just like conditional probabilities for events.

- Obviously, the definition should be justified in a rigorous way but this is not necessary, for now!

- The conditional probability function of, say, the first i elements in a random vector given, say, the other $n-i$ elements shall be defined as:

$$P(x_1, ..., x_i | x_{i+1}, ...x_n) = \frac{P(x_1, ..., x_n)}{P(x_{i+1}, ...x_n)}$$

- For the conditional density we have:

$$f(x_1, ..., x_i | x_{i+1}, ...x_n) = \frac{f(x_1, ..., x_n)}{f(x_{i+1}, ...x_n)}$$

- In both formulas we suppose denominators to be non zero.

19.40 Stochastic Independence

- Two sub vectors of a random vector, say: the first i and the other $n-1$ random variables, are said to be stochastically independent if the joint distribution is the same as the product of the marginals or, that is the same under our definition, if the conditional and marginal distribution coincide.

- We write this for the density case, for the probability function is the same:

$$f(x_1, ..., x_n) = f(x_1, ..., x_i) f(x_{i+1}, ..., x_n)$$

$$f(x_1, ..., x_i | x_{i+1}, ...x_n) = f(x_1, ..., x_i)$$

- This must be true for all the possible values of the n elements of the vector.

19.41 Mutual Independence

- A relevant particular case is that of a vector of mutually independent (or simply independent) random variables. In this case:

$$f(x_1, ..., x_n) = \prod_{i=1,...,n} f_{X_i}(x_i)$$

- Again, this must be true for all possible $(x_1, ..., x_n)$. (Notice: the added big subscript to the uni dimensional density to distinguish among the variables and the small cap $x_i$which are possible values of the variables).

19.42 Conditional Expectation

- Given a conditional probability function $P(x_1, ..., x_i|x_{i+1}, ...x_n)$ or a conditional density $f(x_1, ..., x_i|x_{i+1}, ...x_n)$ we can define conditional expected values of, in general, vector valued functions of the conditioned random variables.
- Something the like of $E(g(x_1, ..., x_i)|x_{i+1}, ...x_n))$ (the expected value is defined exactly as in the uni dimensional case by a proper sum/series or integral operator).

19.43 Conditional Expectation

- It is to be understood that such expected value is a function of the conditioning variables. If we understand this it should be not a surprise that we can take the expected value of a conditional expected value. In this case the following property is of paramount relevance:

$$E(E(g(x_1, ..., x_i)|x_{i+1}, ...x_n)) = E(g(x_1, ..., x_i))$$

- Where, in order to understand the formula, we must remember that the first expected value in the left hand side of the identity is with respect to (wrt) the marginal distribution of the conditioning variables: $(x_{i+1}, ...x_n)$, while the inner expected value of the same side of the identity is wrt the conditional distribution.

19.44 Conditional Expectation

- On the other hand the expected value on the right end side is to be intended as wrt the marginal distribution of the conditioned variables: $(x_1, ..., x_i)$.
- To be really precise we must say that the notation we use (small printed letters for both the values and the names of the random variables) is approximate: we should use capital letters for variables and small letters for values. However we follow the practice that usually leaves the distinction to the discerning reader.

19.45 Law of Iterated Expectations

In the simplest case of two vectors we have: $E_{\mathbf{Y}}(E_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\mathbf{Y})) = E_{\mathbf{X}}(\mathbf{X})$. For the conditional expectation value, wrt the conditioned vector, all the properties of the marginal expectation hold.

19.46 Regressive Dependence

- Regression function and regressive dependence.
- Usually the conditional expectation $E_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\mathbf{Y})$ is called "regression function" of $\mathbf{X}$ on $\mathbf{Y}$, while $E_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X})$ is the regression function of $\mathbf{Y}$ on $\mathbf{X}$. If $E_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\mathbf{Y})$ is constant wrt $\mathbf{Y}$ we say that $\mathbf{X}$ is regressively independent on $\mathbf{Y}$.
- If $E_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X})$ is independent of $\mathbf{X}$ we say that $\mathbf{Y}$ is regressively independent on $\mathbf{X}$.
- Regressive dependence/independence is not a symmetric concept: it can hold on a side only, however in the case of even one sided regressive independence, zero correlation is implied, the reverse, however, is not true.
- Moreover, stochastic independence implied two sided regressive independence, again, the converse is not true.

19.47 Distribution of the max and the min for independent random variables

- Let $\{X_1, ..., X_n\}$ be independent random variables with distribution functions $F_{X_i}(.)$.
- Let $X_{(1)} = \max\{X_1, ..., X_n\}$ and $X_{(n)} = \min\{X_1, ..., X_n\}$.
- Then $F_{X_{(1)}}(t) = \prod_{i=1}^{n} F_{X_i}(t)$ and $F_{X(n)}(t) = 1 - \prod_{i=1}^{n}(1 - F_{X_i}(t))$.
- If the random variables are also identically distributed we have $F_{X_{(1)}}(t) = \prod_{i=1}^{n} F_{X_i}(t) = F^n(t)$ and $F_{X(n)} = 1 - \prod_{i=1}^{n}(1 - F_{X_i}(t)) = 1 - (1 - F(t))^n$.

19.48 Distribution of the max and the min for independent random variables

- Why? Consider the case of the max. $F_{X_{(1)}}(t)$ is , by definition, the probability that the value of the max among the n random variables is less than or equal to t.
- But the max is less than or equal t if and only if each random variable is less than or equal to t.
- Since they are independent this is given by the product of the F_{X_i} each computed at the same point t, that is $F_{X_{(1)}}(t) = \prod_{i=1}^{n} F_{X_i}(t)$.

- For the min: $1 - F_{X_{(n)}}(t)$ is the probability that the min is greater that t. But this is true if and only if each of the n random variables has a value greater than t and for each random variable this probability is $1 - F_{X_i}(t)$. they are independent so...

19.49 Distribution of the sum of independent random variables and central limit theorem

- Let $\{X_1, ..., X_n\}$ be independent random variables. Let $S_n = \sum_{i=1}^{n} X_i$ be their sum.
- We know that, if each random variable has expected value μ_i and variance σ_i^2, then $E(S_n) = \sum_{i=1}^{n} \mu_i$ and $V(S_n) = \sum_{i=1}^{n} \sigma_i^2$.
- Can we say something about the distribution of S_n?
- If we knew the distributions of the X_i we could (but could be cumbersome) compute the distribution of the sum.
- However, if we do not know (better: do not make hypotheses on) the distributions of the X_i we still can give proof to a powerful and famous result which, in its simplest form, states:

19.50 Distribution of the sum of independent random variables and central limit theorem

- Let $\{X_1, ..., X_n\}$ be iid random variables with expected value μ and variance σ^2. Then
$$\lim_{n\to\infty} Pr\left(\frac{\frac{S_n}{n} - \mu}{\sigma/\sqrt{n}} \leq t\right) = \Phi(t)$$
Where Φ is the PDF of a standard Gaussian.
- In practice this means that, under the hypotheses of this theorem, if "n is big enough " (a sentence whose meaning is to be, and can be, made precise) we can approximate $F_{S_n}(s)$ with $\Phi(\frac{\frac{s}{n} - \mu}{\sigma/\sqrt{n}})$.

19.51 Distribution of the sum of independent random variables and central limit theorem

- More general versions of this theorem, with non necessarily identically distributed or even non independent X_i exist.

- This result is fundamental in statistical application where confidence levels for confidence intervals of size of errors for test must be computed in non standard settings.

Statistical inference

19.52 Why Statistics

- Probabilities are useful when we can specify their values. As we did see above, sometimes, in finite settings, (coin flipping, dice rolling, card games, roulette, etc.) it is possible to reduce all probability statement to simple statements judged, by symmetry properties, equiprobable.

- In these case we say we "know" probabilities (at least in the sense we agree on its values and, as a first approximation, do not look for some "discovery rule" for probabilities) and use these for making decisions (meaning: betting). In other circumstances we are not so lucky.

- Consider for instance rolling a pyramidal "die": this is a five sided object with four triangular sides a one square side. In this case what is the probability for each single side to be the down side? For some news on dice see `http://en.wikipedia.org/wiki/Dice`

19.53 Unknown Probabilities and Symmetry

- The sides are not identical, so the classical argument for equiprobability does not hold. We may agree that the probability of each triangular face is the same but what is the total value of these four probabilities? Or, that is the same, what is the probability for the square face to be the down one?

- Just by observing different pyramidal dice we could surmise that the relative probability of the square face and of the four triangular faces depend, also, on the effective shape of the triangular faces. We could hypothesize, perhaps, that the greater is the eight of such faces, the bigger the probability for a triangular face to be the down one in comparison to the probability for the square face.

19.54 Unknown Probabilities and Symmetry

- With skillful physical arguments we could come up with some quantitative hypothesis, we understand, however, that this shall not be simple. With much likelihood a direct observation of the results from a series of actual rolls of this dice could be very useful.

- For instance we could observe that the peaker are the triangular sides the smaller the probability for the square side to be down (and conclude that there should be some unknown height such that the probability of each side is 1/5, at least for a given way of trowing the die).

19.55 No Symmetry

- Consider now a different example: horse racing. Here the event whose probability we are interested in is, to be simple, the name of the winner.
- It is "clear" that symmetry arguments here are useless, moreover, while observation of past races results could be relevant, the idea of repeating the same race a number of times in order to derive some numerical evaluation of probability is both unpractical and, perhaps, even irrelevant.

19.56 No Symmetry

- What we may deem useful are data on past races of the contenders but these data regard different track conditions, different tracks and different opponents.
- Moreover they regard different times, hence, a different age of the horse(s), a different period in the years, a different level of training, and so on.
- History, in short.
- This not withstanding, people bet, and bet hard on such events. Where do their probabilities come from?

19.57 Learning Probabilities

- Let us sum up: probability is useful for taking decision (betting) when the only unknown is the result of the game.
- This is the typical case in simple games of chance (not in the, albeit still simple, pyramidal dice case).
- If we want to use probability when numerical values for probability are not easily derived, we are going to be uncertain both on uncertain results and on the probability of such results.
- We can do nothing (legal) about the results of the game but we may do something for building some reasonable way for assessing probabilities. In a nutshell this is the purpose of statistics.

- The basic idea of statistic is that, in some cases, we can "learn" probabilities from repeated observations of the phenomena we are interested in.
- The problem is that for "learning" probabilities we need ... probabilities!

19.58 Pyramidal Die

- Let us work at an intuitive level on a specific problem. Consider this set of basic assumptions concerning the pyramidal die problem.
- We may agree that the probability for each face to be the down one in repeated rollings of the die is constant, unknown but constant.
- Moreover, we may accept that the order with which results are recorded is, for us, irrelevant as "experiments" (rolls of the dice) are made always in the same conditions.
- We, perhaps, shall also agree that the probability of each triangular face is the same.

19.59 Pyramidal Die Model

- Well: we now have a "statistical model". Let us call $\theta_i i = 1, 2, 3, 4$ the probabilities of each triangular face.
- This are going to be non negative numbers (probability theory require this) moreover, if we agree with the statement about their identity, each of these value must be equal to the same θ so the total probability for a triangular face to be the down one shall be 4θ.
- By the rules of probability the probability for the square face is going to be $1-4\theta$ and, since this cannot be negative, we need $\theta \leq .25$ (where we perhaps shall avoid the equal part in the $\leq$sign).

19.60 Pyramidal Die Constraints

- All these statements come from probability theory joint with our assumptions on the phenomenon we are observing.
- In other, more formal, words we specify a probability model for each roll of the die and state this:
- In each roll we can have a result in the range 1,2,3,4,5;

- The probability of each of the first four values is θ and this must be a number not greater than .25.
- With just these words we have hypothesized that the probability distribution of each result in a single toss is an element of a simple but infinite and very specific set of probability distributions completely characterized by the numerical value of the "parameter" θ which could be any number in the "parameter space" given by the real numbers between 0 and .25 (extrema included if you like).

19.61 Many Rolls

- This is a model for a single rolling. But, exploiting our hypotheses, we can easily go on to a model for any set of rollings of the dice.
- In fact, if we suppose, as we did, that each sequence of results of given length has a probability which only depends on the number of triangular and square faces observed in the series (in technical terms we say that the observation process produces an "exchangeable" sequence of results, that is: sequences of results containing the same number of 5 and non 5 have the same probability).
- Just for simplicity in computation let us move on a step: we shall strengthen our hypothesis and actually state that the results of different rollings are stochastically independent (this is a particular case of exchangeability that is: implies but is not implied by exchangeability).

19.62 Probability of Observing a Sample

- Under this hypothesis and the previously stated probability model for each single roll, the joint probability of a sample of size n, were we only record 5s and not 5s, is just the product of the probabilities for each observation.
- In our example: suppose we roll the dice 100 times and observe 40 times 5 (square face down) and 60 times either 1 or 2 or 3 or 4, since each of these faces is incompatible with the other and each has probability θ, the probability of "either 1 or 2 or 3 or 4" is 4θ.
- The joint probability of the observed sample is thus $(4\theta)^{60}(1-4\theta)^{40}$.

19.63 Pre or Post Observation?

But here there is a catch, and we must understand this well: are we computing the probability of a possible sample before observation, or the probability of the observed sample? In the first case no problems, the answer is correct, but, in the second, we

must realize that the probability of observing the observed sample is actually one, after all we DID observe it!

- Let us forget, for the moment, this subtlety which is going to be relevant in what follows. We have the probability of the observed sample, since the sample is given, the only thing in the formula which can change value is the parameter θ.
- The probability of observing the given sample shall, in general, be a function of this parameter.

19.64 Maximize the Probability of the Observed Sample

- The value which maximizes the probability of the observed sample among the possible values of θ is (check it!) $\widehat{\theta}$ =60/400=3/20=.15
- Notice that this value maximizes $(4\theta)^{60}(1-4\theta)^{40}$: the probability of observing the given sample (or any given specific sample containing 40 5s and 60 non 5s) but also maximizes $\binom{100}{40}(4\theta)^{60}(1-4\theta)^{40}$ that is: the probability of observing A sample in the set of samples containing 40 5s and 60 non 5s. (Be careful in understanding the difference between "the given sample " and "A sample in the set", moreover notice that $\binom{100}{40} = \binom{100}{60}$).

19.65 Maximum Likelihood

- Stop for a moment and fix some points. What did we do, after all? Our problem was to find a probability for each face of the pyramidal dice. The only thing we could say a priori was that the probability of each triangular face was the same. From this and simple probability rules we derived a probability model for the random variable X whose values are $1, 2, 3, 4$ when the down face is triangular, and 5 when it is square.
- We then added an assumption on the sampling process: observations are iid (independent and identically distributed as X). The two assumptions constitute a "statistical model" for X and are enough for deriving a strategy for "estimating" θ (the probability of any given triangular face).
- The suggested estimate is the value $\widehat{\theta}$ which maximizes the joint probability of observing the sample actually observed. In other words we estimated the unknown parameter according to the maximum likelihood method.

19.66 Sampling Variability

- At this point we have an estimate of θ and the first important point is to understand that this actually is just an estimate, it is not to be taken as the "true" value of θ.
- In fact, if we roll the dice another 100 times and compute the estimate with the same procedure, in all probability a different estimate shall come up and for another sample, another one and so on and on.
- Statisticians do not only find estimates, most importantly they study the worst enemy of someone which must decide under uncertainty and unknown probabilities: sampling variability.

19.67 Possibly Different Samples

- The point is simple: consider all possible different samples of size 100. Since, as we assumed before, the specific value of a non 5 is irrelevant, let us suppose, for simplicity, that all that is recorded in a sample is a sequence of 5s and non 5s.
- Since in each roll we either get a 5 or a non 5 the total number of these possible samples is 2^{100}.
- On each of these samples our estimate could take a different value, consider, however, that the value of the estimate only depends of how many 5 and non 5 were observed in the specific sample (the estimate is the number of non 5 divided by 4 times 100).
- So the probability of observing a given value of the estimate is the same as the probability of the set of samples with the corresponding number of 5s.

19.68 The Probability of Our Sample

- But it is easy to compute this probability: since by our assumptions on the statistical model, every sample containing the same number of 5s (and so of non 5s) has the same probability, in order to find this probability we can simply compute the probability of a generic sample of this kind and multiply it times the number of possible samples with the same number of 5s.
- If the number of 5s is, say, k we find that the probability of the generic sample with k 5s and 100-k non 5s is (see above): $(4\theta)^{100-k}(1-4\theta)^k$.

19.69 The Probability of a Similar Estimate

- This is the same for any sample with k 5 and 100-k non 5. There are many samples of this kind, depending on the order of results. The number of possible samples of this kind can be computed in this simple way: we must put k 5s in a sequence of 100 possible places.
- We can insert the first 5 in any of 100 places, the second in any of 99 and so on.
- We get $100 * 99 * ... * (100 - k) = \frac{100!}{(100-k)!}$ however there are k ways to choose the first 5 $k - 1$ for the second and so on up to 1 for the k^{th} and for all these $k!$ ways (they are called "combinations" the sample is always the same, so the number of different samples is $\frac{100!}{k!(100-k)!} = \binom{100}{k}$.

19.70 The Probability of a Similar Estimate

- This is the number of different sequences of "strings" of 100 elements each containing k 5s and 100-k non 5s.
- Summing up: the probability of observing k 5s on 100 rolls, hence of computing and estimate of θ equal to $k/400$, is precisely: $\binom{100}{k}(4\theta)^{100-k}(1-4\theta)^k$ (which is a trivial modification of the binomial).

19.71 The Probability of a Similar Estimate

- So, before sampling, for any possible "true" value of θ we have a different probability for each of the (100 in this case) possible values of the estimate.
- The reader shall realize that, for each given value of θ the a priori (of sampling) most probable value of the estimate is the one corresponding to the integer number of 5s nearest to $100(1 - 4\theta)$ (which in general shall not be integer).

19.72 The Estimate in Other Possible Samples

- Obviously, since this is just the most probable value of the estimate if the probability is computed with this θ, it is quite possible, it is in fact very likely, that a different sample is observed.
- Since our procedure is to estimate θ with $\frac{100-k}{400}$ this immediately implies that, in the case the observed sample is not the most probable for that given θ, the value of the estimate shall NOT be equal to θ, in other words it shall be "wrong" and the reason of this is the possibility of observing many different samples for each given "true" θ, that is: sampling variability.

- In general, using the results above, for any given θ, the probability of observing a sample of size n which gives as an estimate $\frac{n-k}{4n}$ is (as above) $\binom{n}{k}(4\theta)^{n-k}(1-4\theta)^k$

19.73 The Estimate in Other Possible Samples

- So, for instance, the probability, given this value of θ, of observing a sample such that, for instance, the estimate $\frac{n-k}{4n}$ is equal to the parameter value, is, if we suppose that the value for θ which we use in computing this probability can be written as $\frac{n-k}{4n}$ (otherwise the probability is 0 and we must use intervals of values)

$$\binom{n}{k}(4\frac{n-k}{4n})^{n-k}(1-4\frac{n-k}{4n})^k = \binom{n}{k}(\frac{n-k}{n})^{n-k}(1-\frac{n-k}{n})^k$$

- Due to what we did see above the value $\frac{n-k}{4n}$ is the most probable value of the estimate when $\theta = \frac{n-k}{4n}$ but many other values may have sizable probability so that, eve if the "true value" is $\theta = \frac{n-k}{4n}$ it is possible to observe estimates different than $\frac{n-k}{4n}$ with non negligible probability.

19.74 Sampling Variability

- The study of the distribution of the estimate given θ is called the study of the "sampling variability" of the estimate: the attitude of the estimate to change in different samples and can be done in several different ways.

- For instance, using again our example, we see clearly that there does not exist a single "sampling distribution" of the estimate as there is one for each value of the parameter.

- On one hand this is good, because otherwise the estimate would give us quite poor information about θ: the information we get from the estimate comes exactly from the fact that for different values of θ different values of the estimate are more likely to be observed.

- On the other it does not allow us to say which is the "sampling distribution" of the estimate but only gives us a family of such distribution.

19.75 Sampling Variability

- However, even if we do not know the value of the parameter we may study several aspects of the sampling distribution.

- For instance, for each θ we can compute, given that θ the expected value of the estimate for the distribution of the estimate with that particular value of θ. In other words we could compute

$$\sum_{k=0}^{n} \frac{n-k}{4n} \binom{n}{k} (4\theta)^{n-k}(1-4\theta)^k$$

and by doing this computation we would see that the result is θ itself, no matter which value has θ. So that we say that the estimate is unbiased.

19.76 Sampling Variability

- Again, for each θ we can compute the variance of the of the estimate for the distribution of the estimate with that particular value of θ. That is, we could compute

$$\sum_{k=0}^{n} (\frac{n-k}{4n})^2 \binom{n}{k} (4\theta)^{n-k}(1-4\theta)^k - \theta^2 = \frac{4\theta(1-4\theta)}{4n}$$

the "sampling variance" of the estimate, and see that, while this is a function of θ (whose value is unknown to us) for any value of θ it goes to 0 when n goes to infinity. This, joint with the above unbiasedness result, implies (Tchebicev inequality) that the probability of having

$$\frac{n-k}{4n} \in [\theta \pm c]$$

that is: of observing a value of the estimate different than θ at most of c, goes to 1 for ANY $c > 0$ no matter the value of θ. This is called "mean square consistency".

19.77 Sampling Variability

- A curiosity. In typical applications the sampling variance depends on the unknown parameter(s).

- While any reasonable estimate must have a sampling distribution depending on the unknown parameter(s) there are cases where tha sampling variance could be independent on unknown parameter(s).

- For instance, in iid sampling from an unknown distribution with unknown expected value μ and known standard deviation σ the usual estimate of μ, the arithmetic mean of the data, has a sampling variance equalt to $\frac{\sigma^2}{n}$ which does not depend on unknown parameters (repeat: we assumed σ known).

19.78 Estimated Sampling Variability

- In the end, if, say we wish for some "number" for the sampling variance when, as in our case, it depends on the unknown parameter and not the simple formula $\frac{4\theta(1-4\theta)}{4n}$, or some specific distribution in the place of the family of distributions $\binom{n}{k}(4\theta)^{n-k}(1-4\theta)^k$ we could "estimate" these substituting in the formula the estimate of θ to the unknown value $\hat{\theta} = \frac{n-k}{4n}$ and get

- $\hat{V}(\hat{\theta}) = \frac{4\hat{\theta}(1-4\hat{\theta})}{4n}$ and $\hat{P}(\hat{\theta} = \frac{n-k}{4n}) = \binom{n}{k}(4\hat{\theta})^{n-k}(1-4\hat{\theta})^k$ and always remember to notice the "hats" on V and P.

19.79 Quantifying Sampling Variability

- Whatever method we use for dealing with sampling variability the point is to face it

- We could find different procedures for computing our estimate, however, for the same reason (for each given true value of θ many different samples are possible) any reasonable estimate always a sampling distribution (in reasonable cases depending on θ), so we would in any case face the same problem:sampling variability.

- The point is not to avoid sampling variability but to live with it. In order to do this it is better to follow some simple principles.

- Simple, yes, but so often forgotten, even by professionals, as to create most problems encountered in practical applications of statistics.

19.80 Principle 1

- The first obvious principle to follow in order to be able to do this is: "do not forget it".

- An estimate is an estimate is an estimate, it is not the "true" θ.

- This seems obvious but errors of this kind are quite common: it seems human brain does not like uncertainty and, if not properly conditioned, it shall try in any possible way, to wrongly believe that we are sure about something on which we only posses some clue.

19.81 Principle 2

- The second principle is "measure it".
- An estimate (point estimate) by itself is almost completely useless, it should always be supplemented with information about sampling variability.
- At the very least information about sampling standard deviation should be added. Reporting in the form of confidence intervals could be quite useful.
- This and not point estimation is the most important contribution statistics may give to your decisions under uncertainty.

19.82 Principle 3

- The third principle is "do not be upset by it".
- Results of decision may upset you even under certainty. This is obviously much more likely when chance is present even if probabilities are known.
- We are at the third level: no certainty, chance is present, probability are unknown!
- The best statistics can only guarantee an efficient and logically coherent use of available information.
- It does not guarantee Luck in "getting the right estimates" and obviously it cannot guarantee that, even if probabilities are estimated well something very unlikely does not happen! (And no matter what, People shall always expect, forgive the joke, that what is most probable is much more likely than it is probable).

19.83 The Questions of Statistics

- This long discussion should be useful as an introduction to the statistical problem:
- why we need to do inference and do not simply use probability?
- what can we expect from inference?
- Now let us be a little more precise.

19.84 Statistical Model

- This is made of two ingredients.
- The first is a probability model for a random variable (or more generally a random vector, but here we shall consider only the one dimensional case).
- This is simply a set of distributions (probability functions or densities) for the random variable of interest. The set can be indexed by a finite set of numbers (parameters) and in this case we speak of a parametric model. Otherwise we speak of a non parametric model.
- The second ingredient is a sampling model that is: a probabilistic assessment about the joint distribution of repeated observation on the variable of interest.
- The simplest example of this is the case of independent and identically distributed observations (simple random sampling).

19.85 Specification of a Parametric Model

- Typically a parametric mode is specified by choosing some functional form for the probability or density function (here we use the symbol P for both) of the random variable X say: $X \rightsquigarrow P(X;\theta)$ and a set of possible values for $\theta : \theta \in \Theta$(in the case of a parametric model).
- Sometimes we do not fully specify P but simply ask, for instance, for $X$to have a certain expected value or a certain variance.

19.86 Statistic

- A fundamental concept is that of "estimate" or "statistic". Given a sample: $\underline{X}$ and estimate is simply a function of the sample and nothing else: $T(\underline{X})$.
- In other words it cannot depend on unknowns the like of parameters in the model. Once the sample is observed the estimate becomes a number.

19.87 Parametric Inference

- When we have a parametric model we typically speak about "parametric inference", and we are going to do so here.
- This may give the false impression that statistician are interested in parameter values. S

- sometimes this may be so but, really, statisticians are interested in assessing probabilities for (future) values of X, parameters are just "middlemen" in this endeavor.

19.88 Different Inferential Tools

- Traditionally parametric inference is divided in three (interconnected) sections:
- Point estimation;
- Interval estimation;
- Hypothesis testing.

19.89 Point Estimation

- In point estimation we try to find an estimate $T(\underline{X})$ for the unknown parameter θ (the case of a multidimensional parameter is completely analogous).
- In principle, any statistic could be an estimate, so we discriminate between good and bad estimates by studying the sampling properties of these estimates.
- In other words we try to asses whether a given estimate sampling distribution (that is, as we did see before, the probability distribution of the possible values of the statistic as induced by the probabilities of the different possible samples) enjoys or not a set of properties we believe useful for a good estimate.

19.90 Unbiasedness

- An estimate $T(\underline{X})$ is unbiased for θ iff $E_\theta(T(\underline{X})) = \theta$, $\forall\theta \in \Theta$. In order to understand the definition (and the concept of sampling distribution) is important to realize that, in general, the statistic T has a potentially different expected value for each different value of θ (hence each different distribution of the sample).
- What the definition ask for is that this expected value always corresponds to the θ which indexes the distribution used for computing the expected value itself.

19.91 Mean Square Error

- We define the mean square error of an estimate T as: $MSE_\theta(T) = E_\theta((T-\theta)^2)$.
- Notice how, in this definition, we stress the point that the MSE is a function of θ (just like the expected value of T).

- We recall the simple result:

$$E_\theta((T-\theta)^2) = E_\theta((T - E_\theta(T) + E_\theta(T) - \theta)^2) =$$

$$= E_\theta((T - E_\theta(T))^2 + (E_\theta(T) - \theta)^2$$

  where the first term in the sum is the sampling variance of the estimate and the second is the "bias".

- Obviously, for an unbiased estimate, MSE and sampling variance are the same.

19.92 Mean Square Efficiency

- Suppose we are comparing two estimates for θ, say: T_1 and T_2.
- We state that T_1 is not less efficient than T_2 if and only if $MSE_\theta(T_1) \leq MSE_\theta(T_2)$ $\forall\theta \in \Theta$.
- As is the case of unbiasedness the most important point is to notice the "for all" quantifier ($\forall$).
- This implies, for instance, that we cannot be sure, given two estimates, whether one is not worse than the other under this definition.
- In fact it may well happen that mean square errors, as functions of the parameter "cross", so that one estimate is "better" for some set of parameter values while the other for a different set.
- In other words, the order induced on estimates by this definition is only partial.

19.93 Meaning of Efficiency

If an estimate is T_1 satisfies this definition wrt another estimate T_2, this means (use Tchebicev inequality and the above decomposition of the mean square error) that it shall have a bigger (better: not smaller) probability of being "near" θ for any value of this parameter, than T_2.

19.94 Mean Square Consistency

- Here we introduce a variation. Up to now properties consider only fixed sample sizes. here, on the contrary, we consider the sample size n as a variable.
- Obviously, since an estimate is defined on a given sample, this new setting requires the definition of a sequence of estimates and the property we are about to state is not a property of an estimate but of a sequence of estimates.

19.95 Mean Square Consistency

- A sequence $\{T_n\}$ of estimates is termed "mean square consistent if and only if $\lim_{n\to\infty} MSE_\theta(T_n) = 0$, $\forall\theta \in \Theta$.
- You should notice again the quantifier on the values of the parameter.
- Given the above decomposition of the MSE the property is equivalent to the joint request: $\lim_{n\to\infty} E_\theta(T_n) = \theta, \forall\theta \in \Theta$ and $\lim_{n\to\infty} V_\theta(T_n) = 0$, $\forall\theta \in \Theta$.
- Again, using Tchebicev, we understand that the requirement implies that, for any given value of the parameter, the probability of observing a value of the estimate in any given interval containing θ goes to 1 if the size of the sample goes to infinity.

19.96 Methods for Building Estimates

We could proceed by trial and error: this would be quite time consuming. better to devise some "machinery" for creating estimates which can reasonably expect to be "good" in at least some of the above defined senses.

19.97 Method of Moments

- Suppose we have a iid (to be simple) sample $\underline{X}$ from a random variable X distributed according to some (probability or density) $P(X;\theta)$ $\theta \in \Theta$where the parameter is, in general, a vector of k components.
- Suppose, moreover, X has got, say, n moments $E(X^m)$ with $m = 1, ..., n$.
- In general we shall have $E(X^m) = g_m(\theta)$ that is: the moments are functions of the unknown parameters.

19.98 Estimation of Moments

- Now, under iid sampling, it is very easy to estimate moments in a way that is, at least, unbiased and mean square consistent (and also, under proper hypotheses, efficient).
- In fact the estimate: $\widehat{E}(X^m) = \sum_{i=1,...,n} X^m/n$ that is: the $m-$th empirical moment is immediately seen to be unbiased, while its MSE (the variance, in this case) is $\frac{V(X^m)}{n}$ which (if it exists) obviously goes to 0 if the size n of the sample goes to infinity.

19.99 Inverting the Moment Equation

- The idea of the method of moment is simple. Suppose for the moment that θ is one dimensional.
- Choose any g_m and suppose it is invertible (if the model is sensible, this should be true. Why?).
- Estimate the correspondent moment of order m with the empirical moment of the same order and take as an estimate of θ the function $\widehat{\theta}_m = g_m^{-1}(\sum_{i=1,\ldots,n} X^m/n)$.
- In the case of k parameter just solve with respect to the unknown parameter a system on k equation connecting the parameter vector with k moments estimated with the corresponding empirical moments.

19.100 Problems

- This procedure is intuitively alluring. However we have at least two problem. The first is that any different choice of moments is going to give us, in general, a different estimate (consider for instance the negative exponential model and estimate its parameter using different moments).
- The Generalized Method of Moments tries to solve this problem(do not worry! this is something you may ignore, for the moment).
- The second is that, while empirical moments under iid sampling are, for instance, unbiased estimates of corresponding theoretical moments, this is usually not true for method of moments estimates. This is due to the fact that the $g_m$we use are typically not linear.
- Under suitable hypotheses we can show that method of moments estimates are means square consistent but this is usually all we can say.

19.101 Maximum Likelihood

- Maximum likelihood method (one of the many inventions of Sir R. A. Fisher: the creator of modern mathematical statistics and modern mathematical genetics).
- Here the idea is clear if we are in a discrete setting (i.e. if we consider a model of a probability function).
- The first step in the maximum likelihood method is to build the joint distribution of the sample.
- In the context described above (iid sample) we have $P(\underline{X};\theta) = \prod_i P(X_i;\theta)$.

- Now, observe the sample and change the random variables in this formulas (X_i) into the corresponding observations (x_i).
- The resulting $P(\underline{x};\theta)$ cannot be seen as a probability of the sample (the probability of the observed sample is, obviously, 1), but can be seen as a function of θ given the observed sample: $L_{\underline{x}}(\theta) = P(\underline{x};\theta)$.

19.102 Maximum Likelihood

- We call this function the "likelihood".
- It is by no means a probability, either of the sample or of θ, hence the new name.
- The maximum likelihood method suggests the choice, as an estimate of θ, of the value that maximizes the likelihood function given the observed sample,formally: $\widehat{\theta}_{ml} = \arg\max_{\theta\in\Theta} L_{\underline{x}}(\theta)$.

19.103 Interpretation

- If P is a probability (discrete case) the idea of the maximum likelihood method is that of finding the value of the parameter which maximizes the probability of observing the actually a posteriori observed sample.
- The reasoning is exactly as in the example at the beginning of this section.
- While for each given value of the parameter we may observe, in general, many different samples, a set of these (not necessarily just one single sample: many different samples may have the same probability) has the maximum probability of being observed given the value of the parameter.

19.104 Interpretation

- We observe the sample and do not know the parameter value so, as an estimate, we choose that value for which the specific sample we observe is among the most probable samples.
- Obviously, if , given the parameter value, the sample we observe is not among the most probable, we are going to make a mistake, but we hope this is not the most common case and we can show, under proper hypotheses, that the probability of such a case goes to zero if the sample size increases to infinity.

19.105 Interpretation

- A more satisfactory interpretation of maximum likelihood in a particular case.
- Suppose the parameter θ has a finite set (say m) of possible values and suppose that, a priori of knowing the sample, the statistician considers the probability of each of this values to be the same (that is $1/m$).
- Using Bayes theorem, the posterior probability of a given value of the parameter given the observed sample shall be:$P(\theta_j|\underline{x}) = \frac{P(\underline{x}|\theta_j)\frac{1}{m}}{\sum_j P(\underline{x}|\theta_j)\frac{1}{m}} = h(\underline{x})L_{\underline{x}}(\theta_j)$.

19.106 Interpretation

- In words: if we consider the different values of the parameter a priori (of sample observation) as equiprobable, then the likelihood function is proportional to the posterior (given the sample) probability of the values of the parameter.
- So that, in this case, the maximum likelihood estimate is the same as the maximum posterior probability estimate.
- In this case, then, while the likelihood is not the probability of a parameter value (it is proportional to it) to maximize the likelihood means to choose the parameter value which has the maximum probability given the sample.

19.107 Maximum Likelihood for Densities

- In the continuous case the interpretation is less straightforward. Here the likelihood function is the joint density of the observed sample as a function of the unknown parameter and the estimate is computed by maximizing it.
- However, given that we are maximizing a joint density and not a joint probability the simple interpretation just summarized is not directly available.

19.108 Example (Discrete Case)

Example of the two methods. Let X be distributed according to the Poisson distribution, that is: $P(x;\theta) = \frac{\theta^x e^{-\theta}}{x!}$ $x = 0, 1, 2, ...$ Suppose we have a simple random sample of size n.

19.109 Example Method of Moments

- For this distribution all moment exist and, for instance $E(X) = \theta$, $E(X^2) = \theta^2 + \theta$.

- If we use the first moment for the estimation of θ we have $\widehat{\theta}_1 = \bar{x}$ but, if we choose the second moment, we have: $\widehat{\theta}_2 = (-1 + \sqrt{1 + 4\overline{x_2}})/2$ where $\overline{x_2}$ here indicates the empirical second moment (the average of the squares).

19.110 Example Maximum likelihood

- The joint probability of a given Poisson sample is: $L_{\underline{x}}(\theta) = \prod_i \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \frac{\theta^{\sum_i x_i} e^{-n\theta}}{\prod_i x_i!}$.
- For a given θ this probability does not depend on the specific values of each single observation but only on the sum of the observations and the product of the factorials of the observations.
- The value of θ which maximizes the likelihood is $\widehat{\theta}_{ml} = \overline{x}$ which coincides with the method of moments estimate if we use the first moment as the function to invert.

19.111 More Advanced Topics

- Sampling standard deviation, confidence intervals, tests, a preliminary comment.
- The following topics are almost not touched in standard USA like undergraduate economics curricula, and scantly so in other systems.
- They are, actually, very important but only vague notions of these can be asked to a student as a prerequisite.
- In the following paragraphs such vague notions are shortly described.

19.112 Sampling Standard Deviation and Confidence Intervals

- As stated above, a point estimate is useless if it is not provided with some measure of sampling error.
- A common procedure is to report the point estimate joint with some measure related to sampling standard deviation.
- We say "related" because, in the vast majority of cases, the sampling standard deviation depends on unknown parameters, hence it can only be reported in an "estimated" version.

19.113 Sampling Variance of the Mean

- The simplest example is this.
- Suppose we have n iid observations from an unknown distribution about which we only know that it possesses expected value μ and variance σ^2 (by the way, are we considering here a parametric or a non parametric model?)
- In this setting we know that the arithmetic mean is an unbiased estimate of μ.
- By recourse to the usual properties of the variance operator we find that the variance of the arithmetic mean is σ^2/n.
- If (as it is very frequently the case) σ^2 is unknown, even after observing the sample we cannot give the value of the sampling standard deviation.

19.114 Estimation of the Sampling Variance

- We may estimate the numerator of the sampling variance: σ^2(typically using the sample variance, with n or better $n-1$ as a denominator) and we usually report the square root of the estimated sampling variance.
- Remember: this is an estimate of the sampling standard error, hence, it too is affected by sampling error (in widely used statistical softwares, invariably, we see the definition "standard deviation of the estimate" in the place of "estimated standard deviation of the estimate": this is not due to ignorance of the software authors, just to the need for brevity, but could be misleading for less knowledgeable software users).

19.115 $n\sigma$ Rules

- In order to give a direct joint picture of estimate an its (estimated) standard deviation, $n\sigma$ "rules" are often followed by practitioners.
- They typically report "intervals" of the form Point Estimate $\pm n$ Estimated Standard Deviation. A popular value of n outside Finance is 2, in finance we see value of up to 6.
- A way of understanding this use is as follows: if we accept the two false premises that the estimate is equal to its expected value and this is equal to the unknown parameter and that the sampling variance is the true variance of the estimate, then Tchebicev inequality assign a probability of at least .75 to observations of the estimate in other similar samples which are inside the " $\pm 2\sigma$" interval (more than .97 for the " $\pm 6\sigma$" interval).

19.116 Confidence Intervals

- A slightly more refined but much more theoretically requiring behavior is that of computing "confidence intervals" for parameter estimates.
- The theory of confidence intervals typically developed in undergraduate courses of statistics is quite scant.
- The proper definition is usually not even given and only one or two simple examples are reported but with no precise statement of the required hypotheses.

19.117 Confidence Intervals

- These examples are usually derived in the context f simple random sampling (iid observations) from a Gaussian distribution and confidence intervals for the unknown expected value are provided which are valid in the two cases of known and unknown variance.
- In the first case the formula is $\left[\overline{x} \pm z_{1-\alpha/2}\sigma/\sqrt{n}\right]$ and in the second $\left[\overline{x} \pm t_{n-1,1-\alpha/2}\widehat{\sigma}/\sqrt{n}\right]$ where $z_{1-\alpha/2}$ is the quantile in the standard Gaussian distribution which leaves on its left a probability of $1-\alpha/2$ and $t_{n-1,1-\alpha/2}$is the analogous quantile for the T distribution with $n-1$ degrees of freedom.

19.118 Confidence Intervals

- With the exception of the more specific choice for the "sigma multiplier" these two intervals are very similar to the "rule of thumb" intervals we introduced above.
- In fact it turns out that, if α is equal to .05, the z in the first interval is equal to 1.96, and, for n greater than, say, 30, the t in the second formula is roughly 2.

19.119 Hypothesis testing

- The need of choosing actions when the consequences of these are only partly known, is pervasive in any human endeavor. However few fields display this need in such a simple and clear way as the field of finance.
- Consequently almost the full set of normative tools of statistical decision theory have been applied to financial problems and with considerable success, when used as normative tools (much less success, if any, was encountered by attempts to use such tools in the description of actual empirical human behavior. But this has to be expected).

19.120 Parametric Hypothesis

- Statistical hypothesis testing is a very specific an simple decision procedure. It is appropriate in some context and the most important thing to learn, apart from technicalities, is the kind of context it is appropriate for
- Statistical hypothesis. here we consider only parametric hypotheses. Given a parametric model, a parametric hypothesis is simply the assumption that the parameter of interest, θ lies in some subset $\Theta_0 \in \Theta$.

19.121 Two Hypotheses

- In a standard hypothesis testing, we confront two hypotheses of this kind with the requirement that, wrt the parameter space, they should be exclusive (they cannot be both true at the same time) and exhaustive (they cover the full parameter space.
- So, for instance, if you are considering a Gaussian model and your two hypotheses are that the expected value is either 1 or 2, this means, implicitly, that no other values are allowed.

19.122 Simple and Composite

- A statistical hypothesis is called "simple" if it completely specifies the distribution of the observables, it is called "composite" if it specifies a set of possible distributions. the two hypotheses are termed "null" (H_0) hypothesis and "alternative" hypothesis (H_1).
- The reason of the names lies in the fact that, in the traditional setting where testing theory was developed, the "null" hypothesis corresponds to some conservative statement whose acceptance would not imply a change of behavior in the researcher while the "alternative" hypothesis would have implied, if accepted, a change of behavior.

19.123 Example

- The simplest example is that of testing a new medicine or medical treatment.
- In a very stylized setting, let us suppose we are considering substituting and already established and reasonably working treatment for some illness with a new one.
- This is to be made on the basis of the observation of some clinical parameter in a population.

- We know enough as to be able to state that the observed characteristic is distributed in a given way if the new treatment is not better than the old one and in a different way if this is not the case.
- In this example the distribution under the hypothesis that the new treatment is not better than the old shall be taken as the null hypothesis and the other as the alternative.

19.124 Critical Region, Acceptance Region

- The solution to a testing problem Is the partition of the set of possible samples into two subsets. If the actually observed sample falls in the acceptance region $\underline{x} \in A$ we are going to accept the null, if it falls in the rejection or critical region $\underline{x} \in C$ we reject it.
- We assume that the union of the two hypotheses cover the full set of possible samples (the sample space) while the intersection is empty (they are exclusive). this is similar to what is asked to the hypotheses wrt the parameter space but has nothing to do with it.
- The critical region stands to testing theory in the same relation as the estimate is to estimation theory.

19.125 Errors of First and Second Kind

- Two errors are possible:

1. $\underline{x} \in C$ but the true hypothesis is H_0, this is called error of the first kind;
2. $\underline{x} \in A$ but the true hypothesis is H_1, this is called error of the second kind.

- We should like to avoid these errors, however, obviously, we do not even know (except in toy situations) whether we are committing them, just like we do not know how much wrong our point estimates are.
- Proceeding in a similar way as we did in estimation theory we define some measure of error.

19.126 Power Function and Size of the Errors

- Power function and size of the two errors. Given a critical region C, for each $\theta \in \Theta_0 \cup \Theta_1$ (which sometimes but not always corresponds to the full parameter space Θ) we compute $\Pi_C(\theta) = P(\underline{x} \in C; \theta)$ that is the probability, as a function of θ, of observing a sample in the critical region, so that we reject H_0.

- We would like, ideally, this function to be near 1 for $\theta \in \Theta_1$ while we would like this to be near 0 for $\theta \in \Theta_0$.
- We define $\alpha = \sup_{\theta \in \Theta_0} \Pi_C(\theta)$ the (maximum) size of the error of the first kind and $\beta = \sup_{\theta \in \Theta_1} (1 - \Pi_C(\theta))$ the (maximum) size of the error of the second kind.

19.127 Testing Strategy

- There are many reasonable possible requirements for the size of the two errors we would the critical region to satisfy.
- The choice made in standard testing theory is somewhat strange: we set α to an arbitrary (typically small) value and we try to find the critical region that, given that (or a smaller) size of the error of the first kind, minimize (among the possible critical regions) the error of the second kind.
- The reason of this choice is to be found in the traditional setting described above. If accepting the null means to continue in some standard and reasonably successful therapy, it could be sensible to require a small probability of rejecting this hypothesis when it is true and it could be considered as acceptable a possibly big error of the second kind.

19.128 Asymmetry

The reader should consider the fact that this very asymmetric setting is not the most common in applications.

19.129 Some Tests

- One sided hypotheses for the expected value in the Gaussian setting. Suppose we have an iid sample from a Gaussian random variable with expected value $\mu$and standard deviation σ.
- We want to test $H_0 : \mu \leq a$ against $H_1 : \mu \geq b$ where $a \leq b$ are two given real numbers. It is reasonable to expect that a critical region of the shape: $C : \{\underline{x} : \overline{x} > k\}$ should be a good one.
- The problem is to find k.

19.130 Some Tests

- Suppose first σ is known. The power function of this critical region is (we use the properties of the Gaussian under standardization):

$$\Pi_C(\theta) = P(\underline{x} \in C; \theta) = P(\overline{x} > k; \mu, \sigma) = 1 - P(\frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \leq \frac{k - \mu}{\sigma/\sqrt{n}}) =$$

$$= 1 - \Phi(\frac{k - \mu}{\sigma/\sqrt{n}})$$

- Where Φ is the usual cumulative distribution of the standard Gaussian distribution.

19.131 Some Tests

- Since this is decreasing in μ the power function is increasing in μ, hence, its maximum value in the null hypothesis region is for $\mu = a$.
- We want to set this maximum size of the error of the first kind to a given value α so we want: $1 - \Phi(\frac{k-a}{\sigma/\sqrt{n}}) = \alpha$ so that $\frac{k-a}{\sigma/\sqrt{n}} = z_{1-\alpha}$ so that $k = a + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$.
- When the variance is unknown the critical region is of the same shape but $k = a + \frac{\widehat{\sigma}}{\sqrt{n}} t_{n-1,1-\alpha}$ where $\widehat{\sigma}$ and$t$ are as defined above.

19.132 Some Tests

The reader should solve the same problem when the hypotheses are reversed and compare the solutions.

19.133 Some Tests

- Two sided hypotheses for the expected value in the Gaussian setting and confidence intervals.
- By construction the confidence interval for μ (with known variance): $\left[\overline{x} \pm z_{1-\alpha/2}\sigma/\sqrt{n}\right]$ contains μ with probability (independent on μ) equal to $1 - \alpha$.
- Suppose we have $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ for some given μ_0. The above recalled property of the confidence interval implies that the probability with which $\left[\overline{x} \pm z_{1-\alpha/2}\sigma/\sqrt{n}\right]$ contains μ_0, when H_0 is true, is $1 - \alpha$.

19.134 Some Tests

- The critical region: $C : \left\{\underline{x} : \mu_0 \notin \left[\overline{x} \pm z_{1-\alpha/2}\sigma/\sqrt{n}\right]\right\}$ or, that is the same: $C : \left\{\underline{x} : \overline{x} \notin \left[\mu_0 \pm z_{1-\alpha/2}\sigma/\sqrt{n}\right]\right\}$ has only α probability of rejecting $H_0$when $H_0$is true.

- Build the analogous region in the case of unknown variance and consider the setting where you swap the hypotheses.