MAATH 3P85 Lecture Notes

Jan Vrbik

Contents

I PROBABILITY **1**

1 Random Experiments **3**
1.1 Important definitions . . . 3
1.1.1 Examples . . . 3
1.1.2 Counting formulas . . . 4
1.2 Set-Theory rules (Boolean Algebra) . . . 5
1.3 Probability of Events . . . 6
1.4 Rules of probability . . . 6
1.4.1 Conditional probability . . . 7
1.4.2 Independence . . . 8

2 Random Variables (discrete type) **9**
2.1 Univariate pmf . . . 9
2.2 Bivariate (joint) pmf . . . 10
2.2.1 Examples . . . 10
2.2.2 Conditional pmf . . . 11
2.2.3 Independence . . . 11
2.3 Multivariate distribution . . . 12

3 Expected Values **13**
3.1 Expected value of a RV . . . 13
3.2 EV of a *function* of a RV . . . 13
3.3 EVs related to bivariate distribution . . . 14
3.4 Moments (univariate) . . . 14
3.5 Joint moments . . . 15
3.6 Probability generating function (PGF) . . . 16
3.7 Conditional expected value . . . 17

4 Common Discrete Distributions **19**
4.1 Binomial . . . 19
4.2 Geometric . . . 20
4.3 Negative Binomial . . . 21
4.4 Hypergeometric . . . 22
4.5 Poisson . . . 24

5 Multivariate Discrete Distributions **25**
5.1 Multinomial 25
5.2 Multivariate Hypergeometric 26

6 Continuous RVs **29**
6.1 Probability density function (pdf) 29
6.2 Distribution function (cdf) 30
6.3 Bivariate (multivariate) pdf 31
6.3.1 Marginal distributions 31
6.3.2 Conditional pdf 32
6.3.3 Independence 33
6.4 Expected value 33
6.4.1 Moment generating function (MGF) 34

7 Common Continuous Distributions **37**
7.1 Uniform 37
7.2 Exponential 38
7.2.1 Median 40
7.3 Gamma 41
7.3.1 Introducing the Γ function 41
7.4 Normal (standardized) 42
7.4.1 Normal (general) 43

8 Central Limit Theorem **45**
8.1 Sampling a distribution 45
8.1.1 Sample mean 45
8.1.2 Stating the CLT 46
8.2 Sampling a bivariate distribution 48
8.2.1 Bivariate CLT 48
8.2.2 Standardized bivariate Normal 49
8.2.3 General bivariate Normal 50

9 Transforming RVs **53**
9.1 Univariate case 53
9.1.1 Distribution-function (or F) technique 53
9.1.2 The pdf (or f) technique 56
9.2 Bivariate Transformations 58
9.2.1 The cdf (or F) technique 58
9.2.2 The pdf (or f) technique 62
9.3 More on Sampling 67
9.3.1 Sample variance 67
9.3.2 Sampling from $N(\mu, \sigma)$ 69
9.3.3 MGF of s^2 69

10 Order Statistics **73**
10.1 Distribution of $X_{(i)}$ 73
10.1.1 Sample median 76
10.2 Bivariate pdf 79
10.2.1 Two *consecutive* order statistics 80
10.2.2 Sample **range** 81
10.2.3 Sample **mid-range** 81

II STATISTICS **85**

11 Estimating Distribution Parameters **87**
11.1 A few definitions 88
11.2 Cramér-Rao inequality 90
11.3 Sufficiency 95
11.4 Method of moments 97
11.4.1 One-parameter estimation 97
11.4.2 Estimating two parameters 99
11.5 Maximum-likelihood technique 100
11.5.1 One-parameter examples 101
11.5.2 Two-parameter examples 104

12 Confidence Intervals **109**
12.1 CI for μ 109
12.1.1 σ unknown 110
12.1.2 Large-sample case 110

Part I

PROBABILITY

Chapter 1

Random Experiments

Typical examples involve rolling a die, dealing cards from a shuffled deck of 52 cards, spinning a wheel with a pointer, and timing customer arrivals. The first two examples are of a DISCRETE type, the last two a CONTINUOUS type (each type requiring a totally different approach).

1.1 Important definitions

SAMPLE SPACE of a specific random experiment is a collection of all possible (complete) outcomes; these are called SIMPLE EVENTS. There is a certain flexibility as to what information to keep.

1.1.1 Examples

Rolling three dice: Here, we may need to keep track of the numbers (of dots) shown, but sometimes it is sufficient to consider only two possibilities - success (getting a six) or failure (anything else). Secondly, we may want to keep track of each particular die (in which case $3, 5, 1$ is considered a different outcome from $1, 5, 3$, for the total of $6^3 = 216$ possibilities), or we may give up on this (the 3 dice look identical and are impossible to tell apart) and consider only how many ones, twos, ... do we get (the sample space will then consist of only $\binom{8}{3} = 56$ outcomes). For a very good reason (it is then easier to assign probabilities), we usually do the *former*.

Dealing five cards If we need to keep track of the order in which the cards are dealt, the sample space will consist of $P_5^{52} = 311,875,200$ simple events, if we only care about the resulting hand (the usual case), then we have only $\binom{52}{5} = 2,598,960$ of those. This time, it is easy to assign probabilities either way.

Spinning wheel The sample space consists of all real numbers from the $[0, 2\pi)$ interval (the angle at which the pointer stops). The number of simple

events is thus a '*continuos*' infinity (as opposed to *countable* infinity).

EVENTS are *subsets* of the sample space; we denote them A, B, C, ... and use a Venn diagram to visualize them. They are usually defined by specifying a *condition* the outcome must meet, e.g. 'the total number of dots is 13' (when rolling three dice), or 'the hand contains exactly 2 spades' (when dealing five cards).

We can utilize SET-THEORY to help us deal with events; only the terminology changes: Universal set Ω is now called the *sample space*, an element of Ω is a *simple event*, a subset of Ω is an *event* and the empty set $\varnothing$ is called NULL EVENT.

We continue to use the words INTERSECTION (notation: $A \cap B$, representing the collection of simple events common to both A *and* B), UNION ($A \cup B$, simple events belonging to either A *or* B or *both*), and COMPLEMENT ($\overline{A}$, simple events *not* belonging to A).

1.1.2 Counting formulas

Suppose we need to select n objects (symbols, numbers, etc.) out of total of K of these; in how many ways can this be done

- when (unrestricted) repetition of symbols is allowed, and the order of selection matters: we have K choices to fill each of our n 'boxes', resulting (applying the MULTIPLICATION PRINCIPLE) in

$$\underbrace{K \cdot K \cdot \ldots \cdot K}_{n \text{ factors}} = K^n$$

  possible ways,

- the n selected objects must be all different (sampling WITHOUT REPLACEMENT); their order still matters: similarly, we have K, $K-1$, $K-2$, $K-n+1$ choices to fill, one by one, our n boxes, resulting in

$$\underbrace{K \cdot (K-1) \cdot \ldots \cdot (K-n+1)}_{n \text{ factors}} = \frac{K!}{(n-K)!} = P_n^K$$

  possible ways (note that when these objects are playing cards, $n!$ of these possibilities yields the *same* card hand),

- select n different objects, but the order in which they come is irrelevant (e.g. we may reach into a bag of marbles and select three of them, all at once - there is no order!): now, we simply remove the $n!$ redundancy of the previous set of possibilities, getting

$$\frac{P_n^K}{n!} = C_n^K$$

  possible ways,

- repetition allowed but the order does not count (we have K choices of fruit, such as apples, bananas, etc. and we need to buy n pieces, any way we like - n apples is fine, if we get lazy): here we use n circles (representing our n choices) and $K-1$ bars (dividing the n circles into K groups of any size - including 0), and permute them to get all possible ways of selection, which is
$$\binom{n+K-1}{n}$$

The last part of this argument is based on yet another important formula which counts the number of different 'words' (codes, etc.) we get by permuting a specific collection of letters (numbers, symbols, etc.), e.g. *aabbbbbccc*. The answer (for this particular example) is

$$\binom{2+5+3}{2,5,3} = \frac{10!}{2!5!3!} = 2520$$

To understand its logic, we start with 10 empty boxes, select 2 of them to place the letter a - this can be done in $\binom{10}{2}$ ways - then, out of the remaining 8 boxes, we select 5 to place the letter b; after that, c is placed in each of the remaining 3 empty boxes. Multiplying the choices, we get

$$\binom{10}{2} \cdot \binom{8}{5} = \frac{10!}{2!8!} \cdot \frac{8!}{5!3!} = \frac{10!}{2!5!3!}$$

thus proving the original formula.

A special case of this: having K *distinct* letters; the formula then reduces to $K!$ permutations.

1.2 Set-Theory rules (Boolean Algebra)

- Both $\cap$ and $\cup$ (individually) are COMMUTATIVE and ASSOCIATIVE
- Intersection is DISTRIBUTIVE over union: $A \cap (B \cup C \cup ...) = (A \cap B) \cup (A \cap C) \cup ...$
- Similarly, union is distributive over intersection: $A \cup (B \cap C \cap ...) = (A \cup B) \cap (A \cup C) \cap ...$
- 'Trivial' rules: $A \cap \Omega = A$, $A \cap \varnothing = \varnothing$, $A \cap A = A$, $A \cup \Omega = \Omega$, $A \cup \varnothing = A$, $A \cup A = A$, $A \cap \overline{A} = \varnothing$, $A \cup \overline{A} = \Omega$, $\overline{\overline{A}} = A$
- Also, when $A \subset B$ (A is a SUBSET of B), we get: $A \cap B = A$ and $A \cup B = B$
- DEMORGAN LAWS: $\overline{A \cap B} = \overline{A} \cup \overline{B}$, and $\overline{A \cup B} = \overline{A} \cap \overline{B}$, or in general
$$\overline{A \cap B \cap C \cap ...} = \overline{A} \cup \overline{B} \cup \overline{C} \cup ...$$
and vice versa (i.e. $\cap \leftrightarrow \cup$)

Definition: A and B are called (mutually) EXCLUSIVE or DISJOINT when

$$A \cap B = \varnothing$$

1.3 Probability of Events

First, we need to assign a probability to each simple event; this is easy when they are all *equally likely* (due to a symmetry of the experiment). In general, this probability is defined as the *relative frequency* of its occurrence in a *long* run (by this, we always mean a *limit* of *infinitely many* independent repetitions of the experiment). To find a probability of an event A, we simply add the probabilities of the simple events A consists of.

Note that

- this will work only when dealing with *discrete* sample spaces (which we are reviewing first, leaving the continuous case for later),
- the general definition of probability (as relative frequency of occurrence) implies that, for most events, we can never know its *exact* value. This gives us two choices: we can ignore this problem and denote the unknown probabilities p, q, etc. (calling them PARAMETERS) - this is how the field of PROBABILITY deals with the issue, or we can concentrate on the actual *estimation* of these parameters, entering the realm of STATISTICS.

This is why this course consists of two parts: we study Probability first, Statistics next.

1.4 Rules of probability

$\Pr(A \cup B) = \Pr(A) + \Pr(B)$ but *only* when $A \cap B = \varnothing$. This implies that $\Pr(\overline{A}) = 1 - \Pr(A)$ as a special case.

This also implies that $\Pr(A \cap \overline{B}) = \Pr(A) - \Pr(A \cap B)$.

For any A and B (possibly overlapping) we have

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

This can be extended to

$$\begin{aligned}\Pr(A \cup B \cup C) &= \Pr(A) + \Pr(B) + \Pr(C) \\ &- \Pr(A \cap B) - \Pr(A \cap C) - \Pr(B \cap C) + \Pr(A \cap B \cap C)\end{aligned}$$

etc.

Using these rules, it is always possible to express the probability of any expression (involving events, their complements, unions and intersections) as a linear combination of probabilities of 'pure' (i.e. no unions nor complements) intersections.

Example:

$$\begin{aligned} & \Pr\left((\bar{A} \cap \bar{B} \cap C) \cup \bar{C}\right) = \Pr(\overline{A \cup B} \cap C) + \Pr(\bar{C}) - \Pr(\varnothing) \\ = \; & \Pr(C) - \Pr\left(C \cap (A \cup B)\right) + 1 - \Pr(C) \\ = \; & 1 - \Pr((A \cap C) \cup (B \cap C)) \\ = \; & 1 - \Pr(A \cap C) - \Pr(B \cap C) + \Pr(A \cap B \cap C) \quad \blacksquare \end{aligned}$$

1.4.1 Conditional probability

of B given A is the long-run relative frequency of an simple event from B when keeping only the outcomes which met condition A. This is the corresponding notation, and the formula for computing it based on ordinary probabilities:

$$\Pr(B|A) \equiv \frac{\Pr(A \cap B)}{\Pr(A)}$$

All basic formulas of probability remain true conditionally, e.g.:

$$\begin{aligned} \Pr(\overline{B}|A) &= 1 - \Pr(B|A) \\ \Pr(B \cup C|A) &= \Pr(B|A) + \Pr(C|A) - \Pr(B \cap C|A), \end{aligned}$$

etc.

Another useful formulas are: the PRODUCT RULE

$$\begin{aligned} \Pr(A \cap B) &= \Pr(A) \cdot \Pr(B|A) \\ \Pr(A \cap B \cap C) &= \Pr(A) \cdot \Pr(B|A) \cdot \Pr(C|A \cap B) \end{aligned}$$

etc.,

and the TOTAL-PROBABILITY FORMULA

$$\Pr(B) = \Pr(B|A_1) \cdot \Pr(A_1) + \Pr(B|A_2) \cdot \Pr(A_2) + ... + \Pr(B|A_k) \cdot \Pr(A_k)$$

where A_1, A_2,...A_k is a PARTITION of the sample space, meaning: the individual A_is must be *mutually exclusive* (no gaps) and their union must *cover* the *whole* sample *space* (no overlaps).

Example: There are two 'black' boxes, one with 2 red and 5 blue marbles, the other one has 3 red and 3 blue marbles. One of the boxes is selected at random an two marbles are drawn from it without replacement.

The probability of selecting the second box and drawing from it two red marbles is

$$\Pr(B_2 \cap 2 \text{ red}) = \Pr(B_2) \cdot \Pr(2 \text{ red} \mid B_2) = \frac{1}{2} \cdot \frac{\binom{3}{2}}{\binom{6}{2}} = \frac{1}{10}$$

(the product rule).

The probability of getting 2 red marbles (from whichever box) is

$$\begin{aligned} \Pr(2 \text{ red}) &= \Pr(2 \text{ red} \mid B_1) \cdot \Pr(B_1) + \Pr(2 \text{ red} \mid B_2) \cdot \Pr(B_2) \\ &= \frac{\binom{2}{2}}{\binom{7}{2}} \cdot \frac{1}{2} + \frac{\binom{3}{2}}{\binom{6}{2}} \cdot \frac{1}{2} = \frac{13}{105} \end{aligned}$$

(total probability formula).

The conditional probability of having selected Box 2 given that both marbles are red:

$$\Pr(B_2 \mid 2 \text{ red}) = \frac{\Pr(B_2 \cap 2 \text{ red})}{\Pr(2 \text{ red})} = \frac{\frac{1}{10}}{\frac{13}{105}} = \frac{21}{26} = 80.77\%$$

(called Bayes' rule - not really a new rule though). ■

1.4.2 Independence

of two or more events is a very natural notion (we should be able to readily tell which events are independent and which are not): when any one of them happens, this does *not* affect the probability of the rest of them. i.e.

$$\Pr(B|A) \equiv P(B)$$

etc.

Independence implies

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

for any two independent events,

$$\Pr(A \cap B \cap C) = \Pr(A) \cdot \Pr(B) \cdot \Pr(C)$$

for any three independent events, etc.

Furthermore, independence of A, B, C, D, ... implies that $A \cap \bar{B}$ and $\bar{C} \cup D$ are also independent (any Boolean combination of one group against a Boolean combination of another, *distinct* group). But we must be careful: $D \cap \bar{B}$ and $\bar{C} \cup D$ are *not* independent (why?).

Having mutually independent events simplifies one of our previous statements: we can now express the probability of any expression involving such events in terms of their individual probabilities.

Example: Assuming that A, B, C and D are mutually independent, find

$$\begin{aligned} & \Pr\left((A \cup \bar{B}) \cap \overline{B \cup C} \cap (C \cup \bar{D})\right) \\ = \; & 1 - \Pr\left((\bar{A} \cap B) \cup B \cup C \cup (\bar{C} \cap D)\right) \end{aligned}$$

Now,

$$\bar{A} \cap B \subset B$$

implies that

$$(\bar{A} \cap B) \cup B = B$$

Similarly,

$$C \cup (\bar{C} \cap D) = (C \cup \bar{C}) \cap (C \cup D) = \Omega \cap (C \cup D) = C \cup D$$

This simplifies the original probability to

$$\begin{aligned} & 1 - \Pr(B \cup C \cup D) = \Pr(\bar{B} \cap \bar{C} \cap \bar{D}) \\ = \; & \big(1 - \Pr(B)\big) \cdot \big(1 - \Pr(C)\big) \cdot \big(1 - \Pr(D)\big) \quad \blacksquare \end{aligned}$$

Chapter 2

Random Variables (discrete type)

A RANDOM VARIABLE (RV for short) returns a *number* for every possible outcome of a random experiment. For the same random experiment, one can define several random variables; we denote them X, Y, Z, ... Technically, a RV is a *mapping* from the sample space into the set of real (or integer) values; thus, representing a fairly abstract notion.

One should appreciate the difference between *random variables* and *events* (the former assigns, to each simple event, a *number*, the latter *yes* or *no*, depending on whether the corresponding condition has been met or not).

2.1 Univariate pmf

We can always compute the probability that a random variable (say X) has a specific value (say i) by adding the probability of all simple events which result in this value of X. Doing this with *all possible* values of X establishes the so-called probability DISTRIBUTION of X. Often, it is possible to express the result of this exercise in terms of a PROBABILITY MASS FUNCTION (pmf for short), defined by

$$f_X(i) \equiv \Pr(X = i)$$

(when this becomes too difficult, a *table* summarizing this information will do just fine). Clearly, the total of these probabilities when summing over all possible values of X must equal to 1.

Example: Defining X as the number of spades in a randomly dealt five-card hand, we have

$$f_X(i) = \frac{\binom{13}{i} \cdot \binom{39}{5-i}}{\binom{52}{5}} \tag{2.1}$$

where $0 \leq i \leq 5$ is the corresponding SUPPORT (the interval of possible values of X); this is important to include whenever specifying a pmf. ■

Once we know the distribution of a RV, we can resolve any related issue without having to go back to the original random experiment (which now, we don't even need to specify); we thus lose the 'real-world' connection, but expediency forces us to do that sooner or later.

A CONDITIONAL pmf of X given that an event A has happened is computed by

$$f_X(i \mid A) \equiv \frac{\Pr(X = i \cap A)}{\Pr(A)}$$

Summing these over all (conditionally) possible values of i must still yield 1; conditional distributions in general have all the usual properties of an ordinary distribution.

2.2 Bivariate (joint) pmf

of *two* random variables is similarly an expression for computing the following probabilities

$$f_{X,Y}(i, j) \equiv \Pr(X = i \cap Y = j)$$

and specifying the 2D support (the set of admissible i and j values).

Based on this, one can always find the corresponding MARGINAL pmf of X by

$$f_X(i) = \sum_{\text{All } j|i} f(i, j)$$

where $j|i$ indicates the CONDITIONAL interval of j values (of the random variable Y) given a specific i (the value of X). Note that this marginal distribution of X is the same as the 'univariate' distribution of X defined originally - the name 'marginal' refers only to the *way* the X distribution was extracted from the joint distribution.

Similarly, one can find the marginal distribution of Y.

2.2.1 Examples

Using a table: Sometimes it is difficult (practically impossible) to express all bivariate probabilities as a simple function of i and j; in that case an explicit table will do just fine, e.g.

$X =$	0	1	2	3	
$Y = 0$	$\frac{7}{50}$	0	0	0	$\frac{7}{50}$
1	$\frac{6}{50}$	$\frac{4}{50}$	0	0	$\frac{10}{50}$
2	$\frac{2}{50}$	$\frac{6}{50}$	$\frac{1}{50}$	0	$\frac{9}{50}$
3	0	$\frac{5}{50}$	$\frac{8}{50}$	$\frac{2}{50}$	$\frac{15}{50}$
4	0	0	$\frac{5}{50}$	$\frac{4}{50}$	$\frac{9}{50}$
	$\frac{15}{50}$	$\frac{15}{50}$	$\frac{14}{50}$	$\frac{6}{50}$	

The table also displays both marginal distributions (thus explaining their name). ■

Dealing 5 cards The bivariate pmf of the number of spades (X) and the number of diamonds (Y) dealt is given by

$$f_{X,Y}(i,j) = \frac{\binom{13}{i}\binom{13}{j}\binom{26}{5-i-j}}{\binom{52}{5}}$$

with the following support

$$0 \le j \le 5-i \qquad \text{while} \qquad 0 \le i \le 5$$

or, equivalently

$$0 \le i \le 5-j \qquad \text{while} \qquad 0 \le j \le 5$$

The marginal distribution of X is of course the same as (2.1). ■

2.2.2 Conditional pmf

of X, given an (observed) value of Y, is defined by

$$f_X(i|Y=\mathbf{j}) \equiv \Pr(X=i \,|\, Y=\mathbf{j}) = \frac{f_{X,Y}(i,\mathbf{j})}{f_Y(\mathbf{j})}$$

where i varies over its *conditional* interval, given $Y=\mathbf{j}$. By boldfacing $\mathbf{j}$ we emphasize the fact that $\mathbf{j}$ has now a specific value and is thus no longer a *variable* of the conditional (*univariate*) distribution (only i is); at best $\mathbf{j}$ can be considered a PARAMETER of this distribution.

This is clearly an (important) special case of a conditional distribution, as introduced two sections ago.

Example: Using our previous example, we get

$i=$	1	2	3
$f_X(i\|Y=3)$	$\frac{5}{15}$	$\frac{8}{15}$	$\frac{2}{15}$

■

2.2.3 Independence

of two random variables X and Y means that the outcome of X cannot influence the outcome of Y (and vice versa) - something we can always gather directly from the nature of the experiment.

This implies that

$$f_{X,Y}(i,j) = f_X(i) \cdot f_Y(j)$$

for every possible combination of i and j, further implying that the joint pmf becomes redundant (we can always build it from the corresponding marginals).

2.3 Multivariate distribution

is an extension of these concepts to *three or more* RVs. The corresponding conditional distributions can themselves be bivariate (or multivariate), e.g.

$$f_{X,Y}(i,j \mid Z=\mathbf{k}) \equiv \Pr(X=i \cap Y=j \mid Z=\mathbf{k}) = \frac{f_{X,Y,Z}(i,j,\mathbf{k})}{f_Z(\mathbf{k})}$$

as opposed to

$$f_X(i \mid Y=\mathbf{j} \cap Z=\mathbf{k}) \equiv \Pr(X=i \mid Y=\mathbf{j} \cap Z=\mathbf{k}) = \frac{f_{X,Y,Z}(i,\mathbf{j},\mathbf{k})}{f_{Y,Z}(\mathbf{j},\mathbf{k})}$$

(the possibilities quickly multiply). We will not go into any more detail.

Chapter 3

Expected Values

3.1 Expected value of a RV

also called its MEAN corresponds (empirically) to its *average* value obtained in a long run of repeated, independent experiments. It is computed by

$$\mu_X \equiv \mathbb{E}(X) \equiv \sum_{\text{All } i} i \cdot f_X(i) \tag{3.1}$$

(a weighted average of the potential values - the weights being the probabilities), where the summation is over all possible values of i. It can be visualized as a center of mass of the corresponding probability histogram.

3.2 EV of a *function* of a RV

First, one has to realize that plugging X into an algebraic expression (such as $\frac{X}{1+X^2}$) creates a new *random variable* (say U); this is called a TRANSFORMATION of X. To get the expected value of U, one does not need to know its pmf (getting it is difficult) - one can find it using the pmf of X, thus:

$$\mathbb{E}[g(X)] = \sum_{\text{All } i} g(i) \cdot f_x(i)$$

where g is any function (usually a simple expression).

Note that, in general, $\mathbb{E}[g(X)] \neq g\left(\mathbb{E}[X]\right)$. For example, $\mathbb{E}[\frac{X}{1+X^2}] \neq \frac{\mu_x}{1+\mu_x^2}$.

An exception is a LINEAR TRANSFORMATION of X; it is true that

$$\mathbb{E}(a \cdot X + c) = a \cdot \mu_X + c$$

where a and c are arbitrary *constants*.

3.3 EVs related to bivariate distribution

When g is any function of *two* arguments, $g(X, Y)$ is again a new (single) random variable (say V). And again, we do not need to know is (univariate) pmf (very difficult to build) to be able to compute its expected value, by

$$\mathbb{E}\left[g(X,Y)\right] = \sum_{\text{All } i,j} g(i,j) \cdot f_{X,Y}(i,j)$$

where the i, j summation is over the 2D support of the X, Y distribution.

The result does *not* equal to $g(\mu_x, \mu_y)$ in general, except when the transformation is *linear*:

$$\mathbb{E}\left(a \cdot X + b \cdot Y + c\right) = a \cdot \mu_X + b \cdot \mu_Y + c$$

Note that this is true regardless whether X and Y are independent or not (a common misconception).

The previous formula easily extends to any number of variables:

$$\mathbb{E}\left(a_1 X_1 + a_2 X_2 + ... + a_k X_k + c\right) = a_1\mu_1 + a_2\mu_2 + ... + a_k\mu_k + c$$

where $\mu_i \equiv \mathbb{E}\left(X_i\right)$. Again, independence is *not* required.

Nevertheless, *independence* of X and Y *does* help when dealing with an expected value of a *product* of RVs, thus:

$$\mathbb{E}(X \cdot Y) = \mu_X \cdot \mu_Y$$

and

$$\mathbb{E}\left[g_1(X) \cdot g_2(Y)\right] = \mathbb{E}\left[g_1(X)\right] \cdot \mathbb{E}\left[g_2(Y)\right]$$

These can be extended to a product of any number of *independent* RVs.

3.4 Moments (univariate)

Simple moments:

$$\mathbb{E}(X^k)$$

Central moments:

$$\mathbb{E}\left[(X - \mu_X)^k\right]$$

There are also factorial moments (introduced later).

This implies that we have yet another name for expected value: it is the first-order ($k = 1$) simple moment.

Of the central moments, the most important is the second one, also called the variance of X

$$\text{Var}(X) = \mathbb{E}\left[(X - \mu_X)^2\right] = \mathbb{E}(X^2) - {\mu_X}^2$$

(prove the last equality). Its square root is the STANDARD DEVIATION of X, notation: $\sigma_x = \sqrt{\text{Var}(X)}$ (or 'sigma' of X). Any variance is clearly non-negative (a sum of non-negative contributions), equal to zero only in the case of a DEGENERATE distribution (X can have only *one* value, i.e. it is a non-random *constant*).

The *mean* and *variance* are two most important CHARACTERISTICS of a univariate distribution.

The interval $\mu - \sigma$ to $\mu + \sigma$ always contains the 'bulk' of the distribution - anywhere from 50 to 90%.

For a *linear transformation* of X, i.e. $U \equiv aX + c$, we get

$$\text{Var}(U) = a^2 \text{Var}(X)$$

which implies

$$\sigma_u = |a| \cdot \sigma_x$$

3.5 Joint moments

SIMPLE moments:

$$\mathbb{E}(X^k \cdot Y^m)$$

CENTRAL moments:

$$\mathbb{E}\left[(X - \mu_X)^k \cdot (Y - \mu_Y)^m\right]$$

The most important of these is the COVARIANCE of X and Y, equal to the $1^{st}, 1^{st}$ central moment:

$$\text{Cov}(X, Y) \equiv \mathbb{E}\left[(X - \mu_x) \cdot (Y - \mu_y)\right] = \mathbb{E}(X \cdot Y) - \mu_x \cdot \mu_y$$

It is equal to *zero* when X and Y are *independent*, but not the reverse: zero covariance does *not* necessarily imply independence. Clearly, $\text{Cov}(X, Y) = \text{Cov}(Y, X)$; also, $\text{Cov}(X, X) = \text{Var}(X)$. Together with the individual means and variances, covariance is the most important *characteristic* of a bivariate distribution.

The DISTRIBUTIVE LAW OF COVARIANCE states that

$$\text{Cov}(a_1 X_1 + a_2 X_2 + c, Y) = a_1 \text{Cov}(X_1, Y) + a_2 \text{Cov}(X_2, Y)$$

This can be extended to any number of RVs, and to the Y part as well.

A related quantity is the CORRELATION COEFFICIENT between X and Y:

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y}$$

(this is the Greek letter 'rho'). The absolute value of this coefficient cannot be greater than 1.

Proof: For any λ

$$\mathrm{Var}(X - \lambda \cdot Y) = \mathrm{Var}(X) + \lambda^2 \mathrm{Var}(Y) - 2\lambda \mathrm{Cov}(X, Y) \geq 0$$

The LHS has its smallest value at λ which makes the corresponding derivative, namely

$$2\lambda \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y)$$

equal to zero; this is clearly

$$\lambda_s = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}$$

Substituting this λ_s for λ in the original inequality yields

$$\begin{aligned} & \mathrm{Var}(X) + \frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Var}(Y)^2} \cdot \mathrm{Var}(Y) - 2\frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)} \cdot \mathrm{Cov}(X, Y) \\ = \quad & \mathrm{Var}(X) - \frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Var}(Y)} \geq 0 \end{aligned}$$

Divide by $\mathrm{Var}(X)$ to get: $1 \geq \rho_{xy}^2$. ■

Note that

$$\rho_{aX+c,Y} = \pm \rho_{xy}$$

depending on the sign of a.

Variance of $aX + bY + c$ is equal to

$$a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y)$$

Independence would make the last term equal to zero (and disappear).

This can be extended to a linear combination of any number of RVs:

$$\begin{aligned} & \mathrm{Var}(a_1 X_1 + a_2 X_2 + ...a_k X_k + c) = a_1^2 \mathrm{Var}(X_1) + a_2^2 \mathrm{Var}(X_2) + ... + a_k^2 \mathrm{Var}(X_k) \\ & +2a_1 a_2 \mathrm{Cov}(X_1, X_2) + 2a_1 a_3 \mathrm{Cov}(X_1, X_3) + ... + 2a_{k-1} a_k \mathrm{Cov}(X_{k-1}, X_k) \end{aligned}$$

Mutual independence of the Xs removes the last line.

3.6 Probability generating function (PGF)

requires the RV have only integer values (negative integers are OK). It is defined by

$$P(s) \equiv \mathbb{E}\left(s^X\right) = \sum_{\mathrm{All}\ i} s^i \cdot f_X(i)$$

where s is a real (auxiliary, i.e. having no direct connection to any probabilities) variable.

When X and Y are *independent*, we get

$$P_{X+Y}(s) = P_x(s) \cdot P_y(s)$$

which represents the easiest way to find the distribution of $X + Y$ (otherwise, one has to construct the so called CONVOLUTION of the two pmf's).

This can be extended to any independent sum. When, furthermore, the independent RVs have the same distribution,

$$P_{X_1+X_2+...+X_k}(s) = P(s)^k$$

It is also obvious that

$$P'(s)|_{s=1} = \mathbb{E}(X)$$

and

$$P''(s)|_{s=1} = \mathbb{E}[X(X-1)] = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)$$

(these are the FACTORIAL moments). This is often the easiest way of deriving the mean and variance of an integer-valued RV.

Example: The PGF of the number of dots obtained when rolling a die is

$$P(s) = \frac{s + s^2 + s^3 + s^4 + s^5 + s^6}{6}$$

This implies that

$$\begin{aligned} \mu &= P'(s)|_{s=1} = \frac{1+2+3+4+5+6}{6} = \frac{7}{2} \\ \sigma^2 &= P'(s)|_{s=1} + \tfrac{7}{2} - \left(\tfrac{7}{2}\right)^2 \\ &= \tfrac{2+3\cdot 2+4\cdot 3+5\cdot 4+6\cdot 5}{6} + \tfrac{7}{2} - \left(\tfrac{7}{2}\right)^2 = \frac{35}{12} \end{aligned}$$

By expanding $P(s)^3$, we get the individual probabilities of the total number of dots when rolling three dice. ■

3.7 Conditional expected value

is simply an expected value computed using a conditional distribution, e.g.

$$\mathbb{E}(X|Y = \mathbf{j}) = \sum_{\text{All } i|\mathbf{j}} i \cdot f_X(i|Y = \mathbf{j})$$

etc. Empirically, it represents the long-run average of the X values, keeping only those outcomes which resulted in $Y = \mathbf{j}$.

This is a special case of a conditional expected value of X given an *event* A, computed by

$$\mathbb{E}(X|A) = \sum_{\text{All } i \mid A} i \cdot f_X(i|A)$$

In this context we get what we call the TOTAL-EXPECTED-VALUE FORMULA (an extension of the total-probability formula), namely

$$\mathbb{E}(X) = \sum_{j=1}^{k} \Pr(A_j) \cdot \mathbb{E}(X|A_j)$$

where A_1, A_2, ...A_k is a *partition* of the sample space (k may be infinite).

Chapter 4

Common Discrete Distributions

First, we review *univariate* distributions. It helps to recall three summation formulas:

$$\begin{aligned} \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i} &= (a+b)^n \qquad \text{binomial Theorem} \\ \sum_{i=0}^{\infty} a^i &= \frac{1}{1-a} \qquad \text{provided} \quad |a| < 1 \\ \sum_{i=0}^{\infty} \frac{a^i}{i!} &= e^a \end{aligned}$$

The next three distributions are based on what can be called a 'roll of a die' type of experiment.

4.1 Binomial

X is the total number of SUCCESSES in a series of n *independent* trials, each TRIAL having only *two* possible outcomes (*success* or *failure*, with the probability of p and $q \equiv 1-p$, respectively). The sample space consists of all n-letter 'words' built based on a two-letter (S and F) alphabet, such as $FFSFSSS...FS$. The corresponding probability of each of these simple events is $p^j q^{n-j}$, where j is the number of letters S (the remaining $n-j$ letters are all F). The probability that $X = i$ (for any specific i, which can be as small as 0 and as large as n) is clearly equal to $p^i q^{n-i}$, multiplied by the *number* of words having exactly i S's and $n-i$ F's (luckily, they all have the same probability). We know that there is exactly $\binom{n}{i}$ such words in our sample space; this implies that the corresponding

pmf is

$$\Pr(X = i) \equiv f(i) = \binom{n}{i} p^i q^{n-i} \qquad \text{where} \qquad 0 \leq i \leq n$$

Based on this, we can find the PGF by

$$P(s) = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} s^i = (q + p\ s)^n$$

(the binomial theorem 'in reverse').

There are several ways of finding the expected value and variance of X (direct summation using (3.1) is the hardest); differentiating the PGF is relatively easy, but we prefer yet another method, namely: X can also be defined as a sum of n *independent* RVs of the so called Bernoulli type, i.e.

$$X = X_1 + X_2 + ... + X_n$$

where X_j is the number of successes in Trial j. Each of these RVs has a very simple distribution, namely

X_j	0	1
Pr	q	p

(4.1)

with the mean of p and the variance of $p - p^2 = pq$. For their independent sum (of n) we thus get

$$\begin{aligned} \mu &= np \\ \text{Var}(X) &= npq \end{aligned}$$

Notation: $X \epsilon\ \mathcal{B}(n, p)$. Note that one needs to know the value of *two* PARAMETERS (n and p) to fully specify this distribution.

Examples: Rolling a die 20 times (or, equivalently, rolling 20 dice) and counting the number of sixes; flipping a coin 14 times and counting the number of heads; playing a series of independent games and counting the number of wins; drawing 6 marbles (WITH REPLACEMENT, i.e. returning the marble after each draw) from a box containing 32 marbles, 12 of which are red, and counting the number of red marbles in our sample of 6; etc.

4.2 Geometric

X is defined as the *number of trials* to get the *first* success in an independent series of trials of the type described in the last section. The sample space now consists of the following (infinitely many) simple events: S, FS, FFS, $FFFS$, ..., implying that

$$\Pr(X = i) \equiv f(i) = pq^{i-1} \qquad \text{where} \qquad i \geq 1$$

The corresponding PGF is therefore

$$\begin{aligned} P(s) &= ps + pqs^2 + pq^2s^3 + pq^3s^4 + ... \\ &= ps(1 + qs + q^2s^2 + q^3s^3 + ...) = \frac{ps}{1-qs} \end{aligned}$$

based on which we get

$$\mu = \left(\frac{ps}{1-qs}\right)'\Big|_{s=1} = \frac{p(1-qs)+qps}{(1-qs)^2}\Big|_{s=1} = \frac{p}{(1-qs)^2}\Big|_{s=1} = \frac{1}{p}$$

since $1 - q = p$, and

$$\begin{aligned} \mathrm{Var}(X) &= \left(\frac{ps}{1-qs}\right)''\Big|_{s=1} + \frac{1}{p} - \frac{1}{p^2} = \frac{2pq}{(1-qs)^3}\Big|_{s=1} + \frac{1}{p} - \frac{1}{p^2} \\ &= \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1}{p^2} - \frac{1}{p} = \frac{1}{p}\left(\frac{1}{p} - 1\right) = \frac{q}{p^2} \end{aligned}$$

Notation: $X \epsilon\ \mathcal{G}(p)$. Clearly, a *one*-parameter distribution.

Example: The number of rolls of a die to get the first six; the number of flips of a coin to get the first head; the number of games to play to get the first win; the number of marbles to draw (with replacement) to get the first red marble; etc.

Note: Sometimes, we want to count only the *failures* before getting the first success (we call this the MODIFIED geometric distribution); for the new, modified X (clearly equal to the old $X - 1$) we have:

$$\begin{aligned} f(i) &= pq^i \qquad \text{where} \qquad i \geq 0 \\ P(s) &= \frac{p}{1-qs} \\ \mu &= \frac{1}{p} - 1 = \frac{q}{p} \end{aligned}$$

The variance, naturally, remains the same.

4.3 Negative Binomial

X is now the *number of trials* until (and including) we get the k^{th} success. It is clearly a sum of k *independent* random variables of the *geometric* type; this yields immediately (using the formulas for dealing with an independent sum of RVs):

$$\begin{aligned} P(s) &= \left(\frac{ps}{1-qs}\right)^k \\ \mu &= \frac{k}{p} \\ \mathrm{Var}(X) &= \frac{kq}{p^2} \end{aligned}$$

Only the pmf is yet to be found (there is no such simple rule here) by using the following argument: for X to have the value of i, one must get $k-1$ successes (anywhere) within the first $i-1$ trials (we know how to answer this using the binomial distribution), followed by a success in the last (i^{th}) trial, whose probability is p. Multiplying the two yields

$$f(i) = \binom{i-1}{k-1} p^k q^{i-k} \equiv \binom{i-1}{i-k} p^k q^{i-k} \qquad \text{where} \qquad i \geq k$$

One can get the same answer (try it) by Taylor-expanding the above PGF (thus double-checking the correctness of both formulas).

Notation: $X \epsilon \mathcal{NB}(k,p)$

Example: Number of times to roll a die till the 3^{rd} six is obtained, etc.

Note: Similarly to the Geometric distribution, we can defined a MODIFIED version of this distribution by counting only the failures - try figure out the new version of the previous formulas.

4.4 Hypergeometric

Suppose there are N objects (such as marbles), K of which have some *special* property (e.g. being red), placed in a 'black box' and properly mixed (we assume that these numbers are known to us). Drawing n of these *randomly* and WITHOUT REPLACEMENT (dealing cards from a well-shuffled deck is another good example) X is the number of the special objects (red marbles, spades, aces, etc.) found in this sample

Finding the corresponding pmf is relatively easy: there is $\binom{K}{i}$ ways of selecting i 'special' object, $\binom{N-K}{n-i}$ ways of selecting $n-i$ of the 'other' objects; multiplying the two yields the number of different samples of this kind. Dividing by the total number of samples of n objects thus results in the following formula:

$$\Pr(X=i) \equiv f(i) = \frac{\binom{K}{i} \cdot \binom{N-K}{n-i}}{\binom{N}{n}} \qquad \text{where} \quad \max(0, n-N+K) \leq i \leq \min(n, K)$$

Note that the range of possible values of i gets kind of tricky, but Maple is forgiving (we may always take i to go from 0 to n - when any of these become impossible, the formula returns 0).

This time, the general expression for the corresponding PGF is too complicated to be of much use (it involves a hypergeometric function; this gives the name to the distribution itself, in case you have wondered). But realize that we can still construct it for any specific case by using Maple (but not to derive *general* formulas for the mean and variance).

To do that, we again express X in the $X_1 + X_2 + ... + X_n$ manner, where X_j is the number of 'special' objects obtained in the j^{th} draw (visualize dealing n cards and placing them, separately from each other, face down; X_j is then the number of diamonds we see by turning the j^{th} card - zero or one, of course). Since each X_j still has the distribution of (4.1), with $p = \frac{K}{N}$ (can you see why?), for the expected value of X we get an analog of the np formula, namely

$$\mu = n \cdot \frac{K}{N}$$

To get the variance of X is more complicated; it is now equal to

$$\begin{aligned}&\text{Var}(X_1) + \text{Var}(X_2) + ... + \text{Var}(X_n) + \\ &2\text{Cov}(X_1, X_2) + 2\text{Cov}(X_1, X_3) + ... + 2\text{Cov}(X_{n-1}, X_n)\end{aligned}$$

Luckily, due to the obvious symmetry of the experiment, all n variances have the same value of $pq = \frac{K}{N} \cdot \frac{N-K}{N}$, and all $\binom{n}{2}$ covariances also have the same value, which can be derived by finding first

$$\mathbb{E}(X_1 \cdot X_2) = \Pr(X_1 = 1 \cap X_2 = 1) = \frac{K}{N} \cdot \frac{K-1}{N-1}$$

(why is it so simple?) which implies

$$\begin{aligned}\text{Cov}(X_1, X_2) &= \mathbb{E}(X_1 \cdot X_2) - \mathbb{E}(X_1) \cdot \mathbb{E}(X_1) = \frac{K}{N} \cdot \frac{K-1}{N-1} - \frac{K}{N} \cdot \frac{K}{N} \\ &= \frac{K}{N} \cdot \frac{N(K-1) - K(N-1)}{N(N-1)} = -\frac{K(N-K)}{N^2(N-1)}\end{aligned}$$

(negative, as expected). Putting it all together:

$$\begin{aligned}\text{Var}(X) &= n\text{Var}(X_1) + n(n-1)\text{Cov}(X_1, X_2) = n \cdot \frac{K}{N} \cdot \frac{N-K}{N} - n(n-1)\frac{K(N-K)}{N^2(N-1)} \\ &= n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \left(1 - \frac{n-1}{N-1}\right) = n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}\end{aligned}$$

The first part of the formula is an analog of the binomial npq (with $p = \frac{K}{N}$); the last factor makes sure that the variance becomes 0 when $n = N$ (as it should - right?).

Notation: $X \epsilon\, \mathcal{HG}(n, N, K)$ - a three-parameter distribution.

Example: Dealing 5 cards from a well shuffled deck and counting the number of spades; drawing 6 marbles (all at once) from a box with 32 marbles, 12 of which are red, and counting the number of red marbles in the sample; etc.

Note that this distribution becomes *binomial* when sampling is done WITH REPLACEMENT.

4.5 Poisson

Assume that customers arrive at a store randomly, at a *constant average* rate of λ per hour, and let X be the number of customers who will arrive during the next T hours.

To find the probability of $X = i$ we first realize that its expected value should equal to $\Lambda \equiv \lambda \cdot T$. We then subdivide T into n tiny subintervals of length $\frac{T}{n}$ and assume that each of these subintervals will receive an arrival with the (very small - that is why we can ignore the possiblity of *more* than one arrival in any such interval) probability of $p = \frac{\Lambda}{n}$ (to make the expected number of arrivals equal to Λ), independently of what will happen at all the other subintervals. We are thus approximating the distribution of X with a binomial distribution with n trials. The corresponding PGF is

$$P_n(s) = \left(1 - \frac{\Lambda}{n} + \frac{\Lambda}{n}s\right)^n$$

To eliminate the possibility that a subinterval may occasionally receive *more* than one arrival, we must take the $n \to \infty$ limit of the previous expression, getting

$$P(s) = \exp\Big(\Lambda(s-1)\Big)$$

yielding the resulting PGF of the Poisson distribution.

Finding the corresponding pmf is easy; since

$$\begin{aligned} P(s) &\simeq \exp(-\Lambda) \cdot \exp\left(1 + \Lambda s + \frac{\Lambda^2 s^2}{2!} + \frac{\Lambda^3 s^3}{3!} + ... + \frac{\Lambda^i s^i}{i!} + ...\right) \\ f(i) &= \frac{\Lambda^i}{i!} \cdot \exp(-\Lambda) \qquad\qquad i = 0, 1, 2, 3, ... \end{aligned}$$

(the coefficient of s^i in the corresponding Taylor expansion). Furthermore,

$$\begin{aligned} \mu &= \left(e^{\Lambda(s-1)}\right)'\Big|_{s=1} = \Lambda e^{\Lambda(s-1)}\Big|_{s=1} = \Lambda \\ \mathrm{Var}(X) &= \left(e^{\Lambda(s-1)}\right)''\Big|_{s=1} + \Lambda - \Lambda^2 = \Lambda^2 e^{\Lambda(s-1)}\Big|_{s=1} + \Lambda - \Lambda^2 = \Lambda \end{aligned}$$

The variance of the Poisson distribution is thus equal to its mean.

Notation: $X \epsilon\ \mathcal{P}(\Lambda)$

Example: Number of customers arriving at a specific store during the next 10 minutes, knowing that the average (long-run) arrival rate is 27.4 customers per hour.

Chapter 5

Multivariate Discrete Distributions

This time, there are only two 'common' cases to consider, by generalizing first the binomial and then the hypergeometric distribution to allow more than just two types of outcome at each trial.

5.1 Multinomial

This is an extension of the binomial distribution, in which each trial can result in 3 (or more) possible outcomes (not just S or F). Below, for the sake of simplicity, we use 3 possibilities; the formulas make it quite obvious how to deal with 4 etc. (6 is of course quite common). We denote the three possible outcomes as W, L and T.

The trials are again repeated, independently, n times; this time we need three RVs, say X, Y and Z, which count the total number of outcomes of the first, second and third type, respectively, assuming that their probabilities in each trial are p_1, p_2 and p_3 (again, respectively). The sample space consists of all n-letter words built out of the three-letter alphabet, e.g. TLLWWTT.....WWL, each having the probability given by $p_1^i p_2^j p_3^k$, where i, j and k is the number of Ws, Ls and Ts found within the word.

To find the $\Pr(X = i \cap Y = j \cap Z = k)$ for any specific selection of the i, j and k integers, we simply need to multiply the last probability (of any one such word) by the number of words with exactly i Ws, j Ls and k Ts; we know that this number is given by $\binom{n}{i,j,k}$. This leads to the following tri-variate pmf:

$$f_{xyz}(i,j,k) = \binom{n}{i,j,k} p_1^i p_2^j p_3^k$$

for any 3 non-negative integers i, j, k which add up to n.

The corresponding joint PGF is given by

$$P_{xyz}(s_1, s_2, s_3) = (s_1 p_1 + s_2 p_2 + s_3 p_3)^n$$

(use the multinomial theorem to prove it), and can be used to compute

$$\mathbb{E}\,(X \cdot Y) = \left.\frac{\partial^2 (s_1 p_1 + s_2 p_2 + s_3 p_3)^n}{\partial s_1 \partial s_2}\right|_{s_1 = s_2 = s_3 = 0} = n(n-1) p_1 p_2$$

which implies

$$\mathrm{Cov}(X, Y) = n(n-1) p_1 p_2 - n p_1 \cdot n p_2 = -n p_1 p_2$$

since the *marginal distributions* are obviously *binomial* (following the corresponding formulas). The last formula clearly applies to *any* pair of multinomial RVs (one must only replace p_1 and p_2 accordingly).

Example: Rolling a die 10 times (or, equivalently, rolling 10 dice) and counting the number of *sixes* (X), *ones* (Y) and any *other* number of dots (Z). With the help of these formulas, we are able to answer questions about $W \equiv X - Y$ (the net win in a game where we get paid \$1 for each six but have to pay \$1 for each one), etc. Note that the corresponding PGF is

$$P_w(s) = P_{xyz}(s, s^{-1}, 1) = \left(\frac{s}{6} + \frac{1}{6\,s} + \frac{4}{6}\right)^{10}$$

and that

$$\mathbb{E}(W) = 10 \cdot \frac{1}{6} - 10 \cdot \frac{1}{6} = 0$$

(a FAIR GAME), and

$$\mathrm{Var}(X) = 10 \cdot \frac{1}{6} \cdot \frac{5}{6} + 10 \cdot \frac{1}{6} \cdot \frac{5}{6} + 2 \cdot 10 \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{10}{3}$$

The probability of breaking even is the constant term in the expansion of $P_w(s)$, namely $\frac{3308\,407}{15\,116\,544} = 21.89\%$. ■

5.2 Multivariate Hypergeometric

is, similarly, an extension of the univariate hypergeometric distribution to the case of having 3 (or more) types of objects (placed and mixed in the proverbial 'black box'). Assuming that the number of objects of Type 1, Type 2 and Type 3 type is K_1, K_2 and K_3 respectively (where $K_1 + K_2 + K_3 = N$), we get

$$f_{xyz}(i, j, k) = \frac{\binom{K_1}{i}\binom{K_2}{j}\binom{K_3}{k}}{\binom{N}{n}}$$

where X, Y and Z count the number of objects of each respective type, found in a random *sample* of n objects, drawn *without* replacement (all at once, if you

wish). Naturally, $i+j+k=n$; otherwise, i, j and k can be any 3 non-negative integers such that $0 \le i \le K_1$ *and* $0 \le j \le K_2$ *and* $0 \le k \le K_3$. Translating this into three consecutive ranges for the values of i, then (conditionally, given i) for the values of j, and finally (given i and j) for the values of k is rather complicated in general, but we can afford to be sloppy, as explained earlier (the formula returns zero when used with an impossible combination of i, j, k values).

The *marginal distribution* of X (and Y, and Z) is clearly *univariate* hypergeometric (introduced a few sections ago), with obvious parameters. This yields the individual expected values and variances of each X, Y and Z (etc., if there are more than three types of objects).

But this time we have yet another important formula, namely

$$\mathrm{Cov}(X,Y) = -n \cdot \frac{K_1}{N} \cdot \frac{K_2}{N} \cdot \frac{N-n}{N-1}$$

and its obvious modification for any other pair of the original RVs.

Proof: Since we do not have a PGF to work with, we must express both X and Y in terms of what happened in each individual draw, namely

$$\begin{aligned} X &= X_1 + X_2 + ... + X_n \\ Y &= Y_1 + Y_2 + ... + Y_n \end{aligned}$$

where X_j (Y_j) is the number of objects of Type 1 (2) obtained in Draw j (each of these RVs can equal to either 0 or 1). Using the distributive law of covariance to compute $\mathrm{Cov}(X,Y)$, we get n contributions of the $\mathrm{Cov}(X_1,Y_1)$ type and $n(n-1)$ contributions of the $\mathrm{Cov}(X_1,Y_2)$ type - due to the obvious symmetry of the experiment (again: visualize dealing cards face down), all the $\mathrm{Cov}(X_j,Y_j)$ values, and all $\mathrm{Cov}(X_j,Y_k)$ where $j \ne k$, are the same. Now,

$$\begin{aligned} \mathbb{E}(X_1 \cdot Y_1) &= \Pr(X_1 = 1 \cap Y_1 = 1) = 0 \\ \mathbb{E}(X_1 \cdot Y_2) &= \Pr(X_1 = 1 \cap Y_2 = 1) = \frac{K_1}{N} \cdot \frac{K_2}{N-1} \end{aligned}$$

imply that

$$\begin{aligned} \mathrm{Cov}(X_1,Y_1) &= 0 - \frac{K_1}{N} \cdot \frac{K_2}{N} = -\frac{K_1 K_2}{N^2} \\ \mathrm{Cov}(X_1,Y_2) &= \frac{K_1}{N} \cdot \frac{K_2}{N-1} - \frac{K_1}{N} \cdot \frac{K_2}{N} = \frac{K_1 K_2}{N} \left(\frac{1}{N-1} - \frac{1}{N} \right) = \frac{K_1 K_2}{N^2 (N-1)} \end{aligned}$$

Putting it together:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= -\frac{n K_1 K_2}{N^2} + \frac{n(n-1) K_1 K_2}{N^2 (N-1)} = -\frac{n K_1 K_2}{N^2} \left(1 - \frac{n-1}{N-1} \right) \\ &= -n \cdot \frac{K_1}{N} \cdot \frac{K_2}{N} \cdot \frac{N-n}{N-1} \end{aligned}$$

(negative, as expected). ■

Note that sometimes we have objects (such as the ace of spades) which do not fit this description (belonging to *both* types). If their number (in the original 'black box') is $K_{1,2}$ we can find the covariance between the number of objects of Type 1 (call the corresponding RV U) and Type 2 (RV V), found in a sample of n, by

$$\begin{aligned} & \mathrm{Cov}(U,V) = \mathrm{Cov}(U_0+T, V_0+T) \\ = \; & -n \cdot \left(\tfrac{K_1-K_{1,2}}{N} \cdot \tfrac{K_2-K_{1,2}}{N} + \tfrac{K_1-K_{1,2}}{N} \cdot \tfrac{K_{1,2}}{N} + \tfrac{K_{1,2}}{N} \cdot \tfrac{K_2-K_{1,2}}{N} - \tfrac{K_{1,2}}{N} \cdot \tfrac{N-K_{1,2}}{N} \right) \cdot \tfrac{N-n}{N-1} \\ = \; & -n \cdot \left(\frac{K_1}{N} \cdot \frac{K_2}{N} - \frac{K_{1,2}}{N} \right) \cdot \frac{N-n}{N-1} \end{aligned}$$

where U_0 (V_0) is the number of objects of Type 1 (2) which are *not* of the mixed type, whereas T is counting only those of the mixed type; the original formulas then apply to each of these three RVs - the distributive law of covariance then does the rest.

If nothing else, one should remember that the covariance between the number of aces and the number of spades found in a random hand of cards (of any size) is always equal to *zero* (as the last formula clearly indicates).

Example: Consider dealing, randomly, five cards from the standard deck of 52, and counting the number of spades (X_1), diamonds (X_2), clubs (X_3) and hearts (X_4) - this time, we have 4 types of objects (that is why we have switched to using a different notation - this way, we never run out of the alphabet).

Let us now make this into a game in which one is paid \$3 for each spade dealt, but has to pay \$2 for each diamond and \$1 for each club, i.e.

$$W = 3X_1 - 2X_2 - X_3$$

Our formulas imply that

$$\mathbb{E}(W) = 3 \cdot 5 \cdot \frac{13}{52} - 2 \cdot 5 \cdot \frac{13}{52} - 5 \cdot \frac{13}{52} = 0$$

(a fair game again),

$$\begin{aligned} & \mathrm{Var}(X) = 5 \cdot \tfrac{13}{52} \cdot \tfrac{39}{52} \cdot \tfrac{47}{52} \cdot \left(3^2 + (-2)^2 + (-1)^2\right) \\ & -2 \cdot 5 \cdot \tfrac{13}{52} \cdot \tfrac{13}{52} \cdot \tfrac{47}{52} \cdot (3 \cdot (-2) + 3 \cdot (-1) + (-2) \cdot (-1)) = \tfrac{1645}{104} \end{aligned}$$

To answer a probability question about W, we would have to first build, and then expand in powers of s, its PGF

$$P_w(s) = \sum_{i,j,k=0}^{5} \frac{\binom{13}{i}\binom{13}{j}\binom{13}{k}\binom{13}{5-i-j-k} s^{3i-2j-k}}{\binom{52}{5}}$$

Here, we are taking advantage (by being rather sloppy when specifying the summation's limits) of the fact that binomial coefficients with a negative bottom number (also, when the bottom number is bigger than the top number) are equal to zero. ■

Chapter 6

Continuous RVs

These are characterized by having, potentially, any real value (from a specific interval), which implies that the probability of each such number is always equal to zero (ignoring the fact that no measurement can ever be made exactly)! So we are facing a big problem: adding individual probabilities is no longer feasible as a means of finding probability of events. The way out of this conundrum is provided by the introducing the following new concept:

6.1 Probability density function (pdf)

of a RV X is defined by

$$f_X(x) \equiv \lim_{\varepsilon \to 0} \frac{\Pr(x \leq X < x + \varepsilon)}{\varepsilon}$$

Note that we are using the same notation (f_X) for a pdf as we used for a pmf - this should not create any confusion, as we always know what kind of RV we are dealing with. Furthermore (unlike most textbooks), we consistently use $i,\ j,\ k, ...$ to denote *integers* and $x,\ y,\ z...$ to imply *real* values (in the latter case, x indicates a value of X, etc.). Due to this, the X subscript of $f_X(x)$ becomes redundant, and can be removed (writing $f(x)$ instead), which we do from now on whenever we can.

Also note that values of this function do *not* represent a probability of any event; they must be non-negative, but may occasionally exceed 1 (all the way to plus infinity).

Given $f(x)$, we can compute the probability of X resulting in a value from a specific interval by

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

The result is the same regardless whether the boundary values are included or not (a marked difference from the discrete case).

Note that $f(x)$ needs to be often defined in a *piecewise* manner (as we will see shortly).

6.2 Distribution function (cdf)

or more fully the CUMULATIVE DISTRIBUTION FUNCTION is defined by

$$F(x) \equiv \Pr(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

Clearly, it is always a *non-decreasing* function of x.

At this point we must point out another crucial difference between discrete and continuous RVs: in a discrete case, being able to compute the *individual* probabilities is always sufficient (with the help of a computer if necessary) to find any other probability (just by *adding* these); in a continuous case, to compute any probability we always need to *integrate* $f(x)$, and cdf does it for us analytically and 'in advance' (if we cannot find $F(x)$ analytically, we can always integrate $f_X(x)$ *numerically*; this is another big contribution of today's computers).

Univariate example: The (note that it is discontinuous) function

$$f(x) = \begin{cases} \frac{2}{3}(1+x) & \text{when} \quad -1 \leq x < 0 \\ \frac{4}{3}(1-x) & \text{when} \quad 0 \leq x < 1 \\ 0 & \text{otherwise} \end{cases}$$

represents a legitimate pfd of a RV (since it is non-negative and integrates to 1). Realize that it is a *single* function of x (regardless of how many different expressions we need to define it).

The corresponding cdf is

$$F(x) = \begin{cases} 0 & \text{when} \quad x < -1 \\ \frac{1}{3}(1+x)^2 & \text{when} \quad -1 \leq x < 0 \\ 1 - \frac{2}{3}(1-x)^2 & \text{when} \quad 0 \leq x < 1 \\ 1 & \text{when} \quad 1 \leq x \end{cases}$$

Note that, unlike $f(x)$, $F(x)$ is (and must always be) a *continuous* and non-decreasing function of x. Also note that $F'(x) = f(x)$, with the exception of $x = 0$ where $f(x)$ has a discontinuity and $F'(x)$ does not exist. They make a lot of fuss about these things in Calculus; luckily, we do not have to (same as whether to use $\leq$ or $<$). ■

6.3 Bivariate (multivariate) pdf

is defined by

$$f(x,y) = \lim_{\substack{\varepsilon \to 0 \\ \delta \to 0}} \frac{\Pr(x \le X < x+\varepsilon \cap y \le Y < y+\delta)}{\varepsilon \cdot \delta}$$

This formula could actually be made more general by using *any* small 2D region surrounding the (x, y) point in the numerator, the region's area in the denominator, and then squeezing this region (in whichever way) down to the point itself.

To find the probability of the (X, Y) values falling inside a specific 2D region $\mathcal{R}$ (one can always visualize $\mathcal{R}$ as a '*target*' to be hit - or missed) is computed by

$$\iint_{\mathcal{R}} f(x,y)dxdy$$

Recall that computing a double integral is done by two consecutive univariate integrations (students find this quite challenging in terms of specifying the upper and lower limits of the inner and outer integral). But again, Maple comes to the rescue; its latest versions make even this task quite easy (we may like to bypass Maple occasionally and do things 'by hand', not to get too lazy).

When $f(x, y)$ is constant (over its support, 0 otherwise), corresponding to the so called UNIFORM distribution (all points of the 'target' are equally likely to be hit), the value of the above integral equals to the constant value of $f(x, y)$ *multiplied* by the *area* of $\mathcal{R}$; this means that we can normally get the answer 'geometrically', bypassing any integration.

The extension to three or more RVs should be obvious.

6.3.1 Marginal distributions

Given a bivariate pdf $f_{X,Y}(x, y)$, we can eliminate one RV, say Y, and get the marginal *pdf* of X by

$$f_x(x) = \int_{\text{All } y|x} f_{xy}(x,y)\, dy$$

Note that the range of integration is *conditional* (i.e. both the lower and upper limit may depend on x). On the other hand, the resulting interval of possible x values is *marginal* (i.e. y is out of the picture) - making it depend on y is a common mistake!

Also note that, in this case, we have reverted to using subscripts, to emphasize that f_X and $f_{X,Y}$ are two very different functions.

6.3.2 Conditional pdf

of X given that Y has been observed to result in a specific value $\mathbf{y}$ is computed by a simple substitution, thus:

$$f_x(x \mid Y = \mathbf{y}) = \frac{f_{xy}(x, \mathbf{y})}{f_y(\mathbf{y})}$$

valid in the corresponding *conditional* range of x values. Note that the denominator requires us to find the marginal pdf of X first (this can be bypassed by simply 'normalizing' $f_{X,Y}(x, \mathbf{y})$ - dividing it by a constant which adjusts the total conditional probability to 1).

Bivariate example: One can easily verify that

$$f_{xy}(x, y) = \begin{cases} 2x(x-y) & \text{when} \quad -x < y < x \quad \text{while} \quad 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

represents a legitimate *joint* pdf of two RVs (say X and Y), since it is always non-negative and integrates (over all x-y plane) to 1.

The corresponding X marginal (easy) is

$$f_x(x) = \begin{cases} 2x \int_{-x}^{x} (x-y)dy = 4x^3 & \text{when} \quad 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

The Y marginal (more difficult, unless done by Maple) is

$$f_y(y) = \begin{cases} 2\int_{y}^{1} x(x-y)dx = \frac{2}{3} - y + \frac{5}{3}y^3 & \text{when} \quad -1 < y < 0 \\ 2\int_{-y}^{1} x(x-y)dx = \frac{2}{3} - y + \frac{1}{3}y^3 & \text{when} \quad 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

One should realize that (when done 'by hand'), this cannot be done without visualizing the 2D support of the bivariate pdf.

The *conditional* pdf of X given that $Y = -\frac{1}{2}$ is

$$f_x(x \mid Y = -\tfrac{1}{2}) = \begin{cases} \frac{2x(x+\frac{1}{2})}{\frac{2}{3} + \frac{1}{2} - \frac{5}{24}} = \frac{24}{23}x + \frac{48}{23}x^2 & \text{when} \quad \frac{1}{2} < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Finally

$$\Pr(X + Y < 1) = 1 - \Pr(X + Y > 1) = 2\int_{1/2}^{1} x \int_{1-x}^{x} (x-y)\, dy\, dx = \frac{7}{48}$$

6.3.3 Independence

of X and Y implies that $f_{xy}(x,y) = f_x(x) \cdot f_y(y)$, with the usual consequences (same as in the discrete case), most notably $f_x(x \mid Y = \mathbf{y}) = f_x(x)$; each conditional distribution equals to its marginal counterpart ('I don't care about Y' kind of result). This can be extended to the case of three or more RVs.

We should be able to tell independence from the specific form of $f_{xy}(x,y)$: whenever this pdf is *separable* (meaning it is a *product* of a function of x and a function of y) and the *conditional* range of x is (algebraically) independent of y (i.e. neither the lower limit of this range, nor the upper limit depends on y) - or the other way round - (same as saying that the support's boundaries can be only straight lines parallel to one or the other coordinate axis), X and Y are independent.

6.4 Expected value

of a continuous RV is computed by

$$\mathbb{E}(X) = \int\limits_{\text{All } x} x \cdot f(x)dx$$

Similarly:

$$\mathbb{E}[g(X)] = \int\limits_{\text{All } x} g(x) \cdot f(x)dx$$

where $g(..)$ is an arbitrary function.

In the bivariate case this becomes

$$\mathbb{E}[g(X,Y)] = \iint\limits_{\text{All } x,y} g(x,y) \cdot f(x,y)dx\,dy$$

Simple *moments*, central moments, variance, covariance, etc. are defined in the same manner as in the discrete case (except now, instead of summation, we integrate). Also, all previous formulas for dealing with *linear combinations* of RVs still hold, without change.

Example: Using the pdf of the Univariate example, we get

$$\begin{aligned}
\mathbb{E}(X) &= \frac{2}{3}\int_{-1}^{0} x(1+x)dx + \frac{4}{3}\int_{0}^{1} x(1-x)dx = \frac{1}{9} \\
\mathbb{E}\left(\frac{X}{1+X^2}\right) &= \frac{2}{3}\int_{-1}^{0} \frac{x}{1+x^2}(1+x)dx + \frac{4}{3}\int_{0}^{1} \frac{x}{1+x^2}(1-x)dx \\
&= \frac{1}{6}\pi + \frac{1}{3}\ln 2 - \frac{2}{3} = 0.08798
\end{aligned}$$

Similarly, using the Bivariate example

$$\mathbb{E}\left(\frac{X}{1+Y^2}\right) = 2\int_{0}^{1} x^2 \int_{-x}^{x} \frac{(x-y)}{1+y^2}dy\;dx = \frac{2}{3} \quad \blacksquare$$

6.4.1 Moment generating function (MGF)

For continuous RVs we lose the concept of PGF; this is replaced MGF, defined by

$$M(t) \equiv \mathbb{E}\left(e^{t\,X}\right) = \int\limits_{\text{All } x} e^{t\cdot x} \cdot f(x)\,dx$$

where t is a real (auxiliary) variable.

Main results

- $$\mathbb{E}(X^k) = M^{(k)}(t)\Big|_{t=0}$$

  where (k) denotes the k^{th} derivative with respect to t. This implies the following Taylor expansion of a MGF:

  $$M(t) \simeq 1 + \mathbb{E}(X)\cdot t + \mathbb{E}(X^2)\frac{t^2}{2} + \mathbb{E}(X^3)\frac{t^3}{3!} + \mathbb{E}(X^4)\frac{t^4}{4!} + \cdots$$

- For two *independent* RVs, we have

  $$M_{X+Y}(t) = M_x(t)\cdot M_y(t)$$

  This can be extended to any number of mutually independent RVs.

- And, finally

  $$M_{aX+c}(t) = e^{ct}\cdot M_x(at)$$

Unfortunately, converting a MGF back to the corresponding pdf is rather difficult (it requires knowledge of Fourier-transform theory) - we may still attempt to do it in our labs.

Example: For the RV of our Univariate example, we get

$$M(t) = \frac{2}{3}\int_{-1}^{0} e^{x\cdot t}(1+x)dx + \frac{4}{3}\int_{0}^{1} e^{x\cdot t}(1-x)dx = \frac{2}{3}\cdot\frac{2e^t + e^{-t} - t - 3}{t^2}$$

Note that the result has a *removable* singularity at $t = 0$. ■

The concept can be extended to a *bivariate* distribution, namely

$$M_{xy}(t_1, t_2) = \mathbb{E}\left(e^{t_1 X + t_2 Y}\right) = \iint\limits_{\text{All } x,y} e^{t_1 x + t_2 y}\cdot f(x,y)\,dx\,dy$$

We can now find $\mathbb{E}\left(X^k\cdot Y^m\right)$ by

$$\frac{\partial^{k+m} M_{X,Y}(t_1,t_2)}{\partial t_1^k \partial t_2^m}\Big|_{t_1=t_2=0}$$

This implies that $M_{X,Y}(t_1, t_2)$ has the following bivariate Taylor expansion:

$$1 + \mu_X t_1 + \mu_Y t_2 + \mathbb{E}(X^2)\frac{t_1^2}{2} + \mathbb{E}(Y^2)\frac{t_2^2}{2} + \mathbb{E}(X \cdot Y)t_1 t_2 + ... \qquad (6.1)$$

It is also very easy to find the corresponding marginal MGFs:

$$\begin{aligned} M_x(t_1) &= M_{xy}(t_1, t_2 = 0) \\ M_y(t_2) &= M_{xy}(t_1 = 0, t_2) \end{aligned}$$

Chapter 7

Common Continuous Distributions

First we recall a few key integration formulas:

$$\begin{aligned}
\int x^a dx &= \frac{x^{a+1}}{a+1} + C \qquad a \neq -1 \\
\int \frac{dx}{x} &= \ln|x| + C \\
\int e^{b \cdot x} dx &= \frac{e^{b \cdot x}}{b} + C \\
\int x^k e^{-x} dx &= C - k! e^{-x} \left(1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + ... + \frac{x^k}{k!}\right)
\end{aligned}$$

The last formula implies that

$$\int_0^\infty x^k e^{-x} dx = k! \tag{7.1}$$

Now we go over a few basic examples of continuos distributions; more will be introduced in the Transformation Chapter.

7.1 Uniform

Random experiment: Visualize a spinning pointer with a dial labelled by real numbers from a (at 0 degrees) to b (at 180 degrees, thus coinciding with a); X is the value on the dial at which the pointer stops. All real outcomes from a to b are thus equally likely.

This implies that its pdf is constant on the (a, b) interval, equal to zero

otherwise.

$$\begin{aligned} f(x) &= \begin{cases} \dfrac{1}{b-a} & \text{when} \quad a \le x \le b \\ 0 & \text{otherwise} \end{cases} \\ F(x) &= \begin{cases} 0 & \text{when} \quad x < a \\ \dfrac{x-a}{b-a} & \text{when} \quad a \le x \le b \\ 1 & \text{when} \quad x > b \end{cases} \\ \mu &= \frac{a+b}{2} \\ \sigma &= \frac{b-a}{\sqrt{12}} \end{aligned}$$

Proof:

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{b-a}\int_a^b x\, dx = \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2} \\ \mathbb{E}\left(X^2\right) &= \frac{1}{b-a}\int_a^b x^2\, dx = \frac{b^3-a^3}{3(b-a)} = \frac{(b-a)(a^2+ab+b^2)}{3(b-a)} = \frac{a^2+ab+b^3}{3} \end{aligned}$$

implying

$$\begin{aligned} \sigma^2 &= \frac{a^2+ab+b^3}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{4a^2+4ab+4b^3-3a^2-6ab-3b^3}{12} \\ &= \frac{a^2-2ab+b^3}{12} = \frac{(b-a)^2}{12} \end{aligned}$$

■

Notation: $X \epsilon\, \mathcal{U}(a,b)$

Note: Let us introduce HEAVISIDE FUNCTION defined by

$$\mathcal{H}(x) = \begin{cases} 0 & \text{when} \quad x < 0 \\ 1 & \text{when} \quad x \ge 0 \end{cases}$$

With its help, we can rewrite the above pdf in the following manner

$$f(x) = \frac{\mathcal{H}(x-a) - \mathcal{H}(x-b)}{b-a}$$

having the advantage of being correct for any x (convenient when using Maple).

7.2 Exponential

Random experiment: Customers arriving at a store randomly (and independently of each other) at the (long run) average rate of λ (arrivals per unit of time); X is the time till the next arrival (from now).

To find the corresponding cdf at a time x we divide the $(0, x)$ interval into n (many) subintervals of the same length $\frac{x}{n}$, and assume that the probability of an arrival at any of these subintervals is $\frac{\lambda \cdot x}{n}$, and that this is independent of what happens at any of the other subintervals; note that this makes the expected number of arrivals within the $(0, x)$ time interval equal to $\lambda \cdot x$. Using this model, the probability that $X > x$, which is the value of $1 - F(x)$, is given by

$$\left(1 - \frac{\lambda \cdot x}{n}\right)^n$$

(no arrivals in any of these subintervals). The only trouble with this model is that, during any of the subintervals, we may get *more* than one arrival! How do we fix that? Simple: by increasing the value of n indefinitely; in the $n \to \infty$ limit we thus get

$$F(x) = \begin{cases} 0 & \text{when} \quad x < 0 \\ 1 - \exp(-\lambda \cdot x) & \text{when} \quad 0 \le x \end{cases}$$

where λ is the only *parameter* of this distribution. Alternately (which is more common), one can use the average time between consecutive arrivals, namely $\beta \equiv \frac{1}{\lambda}$, instead, getting

$$F(x) = 1 - \exp\left(-\frac{x}{\beta}\right) \qquad \text{when} \qquad 0 \le x$$

('0 otherwise' will be assumed from now on). This implies

$$f(x) = \frac{\exp\left(-\frac{x}{\beta}\right)}{\beta} \qquad \text{when} \quad 0 \le x$$

for the corresponding pdf, which consists of the numerator (a function of x) and the denominator, which is the NORMALIZING CONSTANT (the full area of the numerator), i.e. $\beta = \int_0^\infty \exp\left(-\frac{x}{\beta}\right) dx$.

This yields the following MGF:

$$M(t) = \frac{1}{\beta} \int_0^\infty \exp\left(-\tfrac{x}{\beta} + t \cdot x\right) dx = \frac{1}{\beta\left(t - \frac{1}{\beta}\right)} \exp\left(-\tfrac{x}{\beta} + t \cdot x\right)\Big|_{x=0}^\infty = \frac{1}{1 - \beta \cdot t}$$

assuming that $t < \frac{1}{\beta}$ (the fact that a MGF exists only when t is 'small' is of no consequence to us - we need not worry about these things). Taylor-expanding this MGF yields $M(t) \simeq 1 + \beta \cdot t + 2\beta^2 \frac{t^2}{2} + ...$ which implies that

$$\mu = \mathbb{E}(X) = \beta$$
$$\mathbb{E}\left(X^2\right) = 2\beta^2$$

further implying

$$\sigma^2 = \mathbb{E}\left(X^2\right) - \mu^2 = \beta^2$$

The standard deviation of an exponential distribution is thus equal to its mean value.

Notation: $X\epsilon\ \mathcal{E}(\beta)$

7.2.1 Median

The MEDIAN $\tilde{\mu}$ of a (continuous) distribution is defined by

$$\Pr(X < \tilde{\mu}) = \Pr(X > \tilde{\mu}) = \frac{1}{2}$$

It is usually found by solving

$$F(\tilde{\mu}) = \frac{1}{2}$$

In the case of an exponential distribution, we get

$$\tilde{\mu} = \beta \cdot \ln 2$$

by solving

$$1 - \exp\left(-\tfrac{\tilde{\mu}}{\beta}\right) = \frac{1}{2}$$

Since $\ln 2 = 0.69315$, the median of an exponential distribution is substantially smaller than the corresponding mean (e.g. $\mu = 5$ min but $\tilde{\mu} = 3$ min 28 sec).

Note that when $Y = g(X)$ and g is a non-decreasing (non-increasing) function of X, then

$$\tilde{\mu}_Y = g(\tilde{\mu}_X)$$

Remember that this is true for the two medians, but *not* for the corresponding means (as we already know).

Proof:

$$\Pr\left(Y < \tilde{\mu}_Y\right) = \Pr\left(g(X) < g(\tilde{\mu}_X)\right) = \Pr\left(X < \tilde{\mu}_X\right) = \frac{1}{2}$$

in the non-decreasing case, and

$$\Pr\left(Y < \tilde{\mu}_Y\right) = \Pr\left(g(X) < g(\tilde{\mu}_X)\right) = \Pr\left(X > \tilde{\mu}_X\right) = \frac{1}{2}$$

in the non-increasing case. ■

For example, when X is exponential with the mean of β, the median of $Y = \frac{X}{1+X}$ is simply $\frac{\beta \cdot \ln 2}{1+\beta \cdot \ln 2}$.

7.3 Gamma

This distribution is defined as a sum of k *independent* RVs, all having the $\mathcal{E}(\beta)$ distribution (i.e. the time from now until the k^{th} arrival). This means that we have, immediately ('for free', so to speak):

$$\begin{aligned} \mu &= k\ \beta \\ \sigma &= \sqrt{k}\beta \\ M(t) &= \frac{1}{(1-\beta\cdot t)^k} \end{aligned}$$

The corresponding pdf must therefore equal to

$$f(x) = \frac{x^{k-1}\exp\left(-\frac{x}{\beta}\right)}{(k-1)!\cdot\beta^k} \qquad \text{when} \quad 0 \le x$$

Proof: We show that this $f(x)$ results in the correct MGF:

$$\begin{aligned} M(t) &= \frac{\int_0^\infty x^{k-1}\exp\left(-\frac{x}{\beta}+t\cdot x\right)dx}{(k-1)!\cdot\beta^k} \\ &= \frac{\int_0^\infty (\frac{1}{\beta}-t)^{k-1}x^{k-1}\exp\left(-x\cdot(\frac{1}{\beta}-t)\right)(\frac{1}{\beta}-t)dx}{(\frac{1}{\beta}-t)^k(k-1)!\cdot\beta^k} \\ &= \frac{\int_0^\infty u^{k-1}\exp(-u)\,du}{(\frac{1}{\beta}-t)^k(k-1)!\cdot\beta^k} = \frac{1}{(1-\beta\cdot t)^k} \end{aligned}$$

using the $u = (\frac{1}{\beta}-t)x$ substitution and (7.1). ■

The corresponding cdf is

$$F(x) = 1 - \exp\left(-\frac{x}{\beta}\right)\cdot\sum_{i=0}^{k-1}\left(\frac{x}{\beta}\right)^i \qquad \text{when} \qquad 0 \le x$$

which can be proven by simple differentiation (using the product rule, all terms but one cancel out).

Notation: $X\epsilon$ gamma(k,β)

7.3.1 Introducing the Γ function

The gamma distribution has two parameters, β which must be positive and k which (until now) was a positive *integer*. Surprisingly, all the above formulas (with the obvious exception of cdf) remain valid when k is allowed to have a

positive *real* value (to emphasize that, we will call it α rather than k), as long as we replace $(k-1)!$ in the denominator of $f(x)$ by $\Gamma(\alpha)$, defined as

$$\Gamma(\alpha) \equiv \int_0^\infty x^{\alpha-1} e^{-x} dx$$

(clearly the correct normalizing constant of the $f(x)$ numerator).

Note that we have also lost the original interpretation of X as the time till the k^{th} arrival - there is no such thing as the 3.1784^{th} arrival - but non-integer values of k are useful in other context (as we will discover later).

The main properties of the Γ function are: $\Gamma(k) = (k-1)!$ for an integer argument, and

$$\begin{aligned} \Gamma(\alpha) &= (\alpha-1) \cdot \Gamma(\alpha-1) \\ \Gamma(1) &= 0! = 1 \\ \Gamma(\tfrac{1}{2}) &= \sqrt{\pi} \end{aligned}$$

Proof of the last statement:

$$\Gamma(\tfrac{1}{2}) = \int_0^\infty \frac{\exp(-x)\ dx}{\sqrt{x}} = \sqrt{2} \int_0^\infty \frac{\exp(-\frac{z^2}{2}) z\ dz}{z} = \sqrt{\pi}$$

using the $x = \frac{z^2}{2}$ substitution and () of the next section. ■

7.4 Normal (standardized)

This distribution will be properly introduced in the next chapter; here we just summarize the key formulas.

This distribution is so important that we reserve the letter Z to denote the corresponding RV.

$$\begin{aligned} f(z) &= \frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\pi}} \qquad \text{when} \quad -\infty < z < \infty \\ \mu &= 0 \\ \sigma &= 1 \\ M(t) &= \exp\left(\frac{t^2}{2}\right) \end{aligned}$$

The cdf does not have a simple analytical form (in terms of the 'usual' functions; it is a new, 'special' function); we can always compute probabilities by integrating $f(z)$.

Some proofs: To show that $\sqrt{2\pi}$ is the correct normalizing constant of $f(z)$, we have to use the following trick: instead of evaluating $\int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz$,

we compute

$$\begin{aligned} & \int_{-\infty}^{\infty} \exp\left(-\frac{z_1^2}{2}\right) dz_1 \times \int_{-\infty}^{\infty} \exp\left(-\frac{z_2^2}{2}\right) dz_2 \\ = & \iint\limits_{\text{whole plane}} \exp\left(-\tfrac{z_1^2+z_2^2}{2}\right) dz_1 dz_2 \\ = & \int_0^{2\pi} \int_0^{\infty} \exp\left(-\frac{r^2}{2}\right) r \; dr d\theta = 2\pi \int_0^{\infty} e^{-u} du = 2\pi \end{aligned}$$

with the help of polar coordinates and the $u = \frac{r^2}{2}$ substitution. This implies that

$$\int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz = \sqrt{2\pi} \tag{7.2}$$

To derive the corresponding MGF is now quite easy:

$$\begin{aligned} M(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2} + t \cdot z\right) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{z^2 - 2zt}{2}\right) dz \\ &= \frac{\exp(-\frac{t^2}{2})}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{z^2 - 2zt + t^2}{2}\right) dz \\ &= \frac{\exp(-\frac{t^2}{2})}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(z-t)^2}{2}\right) dz \\ &= \frac{\exp(-\frac{t^2}{2})}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2}\right) du = \exp(-\frac{t^2}{2}) \end{aligned}$$

since the value of the last integral is $\sqrt{2\pi}$. ■

Taylor-expanding

$$\exp(\frac{t^2}{2}) \simeq 1 + \frac{t^2}{2} + ...$$

verifies that the mean of this distribution is 0 and the variance (and the standard deviation) equal to 1.

Notation: $Z \epsilon \mathcal{N}(0,1)$

7.4.1 Normal (general)

A new variable can be introduced by linearly transforming Z, thus:

$$X \equiv \sigma \cdot Z + \mu$$

where $\sigma > 0$.

Note that a linear transformation preserves the shape of the pdf (the new $f_X(x)$ *looks* exactly the same as the old $f_Z(z)$, just the *scale* is different). We now get

$$\begin{aligned} f_X(x) &= \frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\cdot\sigma} \\ \mu_X &= \mu \\ \sigma_X &= \sigma \\ M_X(t) &= \exp\left(\frac{t^2\sigma^2}{2} + \mu\cdot t\right) \end{aligned}$$

The last three lines are based on previous formulas; to get the new pdf we have to learn how to transform RVs (done shortly).

Notation: $Z\epsilon\,\mathcal{N}(\mu,\sigma)$

Chapter 8

Central Limit Theorem

This section links Probability with Statistics, yielding the most important result of these areas of study.

8.1 Sampling a distribution

A RANDOM INDEPENDENT SAMPLE (RIS) of SIZE n from a specific distribution is a collection of n *independent* RVs X_1, X_2, ...X_n, each of them having this distribution (this is achieved by performing the corresponding experiment, independently, that many times). It is important to realize that, being RVs, they have *not* been observed to attain any specific value yet (the n repetitions of the experiment are yet to be done). The Xs are sometimes referred to as being IID (independent, identically distributed).

A SAMPLE STATISTIC is any function (expression) of these n RVs, thus defining a new (single) RV. A prime example is provided by the so called

8.1.1 Sample mean

defined as the simple average of the X_i's, thus

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

The sample mean (unlike all the other 'means' we have seen before), is clearly a RV, having its own expected value (we could also call it the 'mean of the sample mean'), variance, and distribution (also called, in this context, its SAMPLING DISTRIBUTION). The obvious task is to relate these to the distribution from which the sample is drawn (the *sampled distribution* - the names get confusing!).

This is easy to do for the expected value and variance:

$$\mathbb{E}\left(\bar{X}\right) = \frac{\sum_{i=1}^{n} \mathbb{E}\left(X_i\right)}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \frac{n \cdot \mu}{n} = \mu$$

and

$$\text{Var}\left(\bar{X}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(X_i\right) = \frac{n\cdot\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Note that this implies

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

(one of the most important formulas of Statistics).

But how about the distribution itself, namely: how does the shape of the distribution of $\bar{X}$ (in terms of its pdf) relate to pdf of the distribution from which we sample (of the individual, single Xs)?

When $n = 1$, the answer is simple: the two distributions are identical.

But, as soon as we reach $n = 2$, the two shapes will already visibly differ (as an example, compare the $\mathcal{E}(1)$ and $\gamma(2,1)$ distributions).

And, as n increases (e.g. the gamma$(n,1)$ distribution), something very surprising happens: the resulting shape has *nothing* to do with the original distribution; it looks the *same* in all cases (regardless which distribution we sample)! What exactly is this common ASYMPTOTIC (implying $n \to \infty$) distribution?

8.1.2 Stating the CLT

We now prove that the pdf of

$$Z_n \equiv \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \tag{8.1}$$

approaches, as $n \to \infty$, the pdf of the standardized Normal distribution (as defined earlier).

Note that Z_n is standardized (i.e. it has the mean of 0 and standard deviation equal to 1) for *each* value of n (only its *shape* changes).

Proof: Instead of dealing with pdf, we find the limit of the corresponding MGFs (a lot easier).

First we note that (8.1) can be rewritten as follows:

$$Z_n = \sum_{i=1}^{n}\frac{X_i-\mu}{\sigma\sqrt{n}} \equiv \sum_{i=1}^{n}\frac{U_i}{\sqrt{n}}$$

which is a sum of independent, identically distributed RVs (note that the U_i are all also standardized). The MGF of Z_n is therefore the MGF of a single $\frac{U_i}{\sqrt{n}}$, raised to the power of n.

We know that the MGF of $\frac{U_i}{\sqrt{n}}$ can be expanded in the following manner:

$$\begin{aligned} & 1+\frac{\mathbb{E}(U_i)}{\sqrt{n}}\cdot t+\frac{\mathbb{E}(U_i^2)}{n}\cdot\frac{t^2}{2}+\frac{\mathbb{E}(U_i^3)}{n^{3/2}}\cdot\frac{t^3}{6}+\frac{\mathbb{E}(U_i^4)}{n^2}\cdot\frac{t^4}{24}+... \\ = \quad & 1+\frac{1}{n}\cdot\frac{t^2}{2}+\frac{\alpha_3}{n^{3/2}}\cdot\frac{t^3}{6}+\frac{\alpha_4}{n^2}\cdot\frac{t^4}{24}+... \end{aligned}$$

where α_3 and α_4 is the skewness and kurtosis (respectively) of the original X_i distribution. This means that the MGF of Z_n is

$$\left(1+\frac{1}{n}\cdot\frac{t^2}{2}+\frac{\alpha_3}{n^{3/2}}\cdot\frac{t^3}{6}+\frac{\alpha_4}{n^2}\cdot\frac{t^4}{24}+...\right)^n \tag{8.2}$$

and all we have to do is finding the $n\to\infty$ limit.

One has to recall (from Calculus) that

$$\lim_{n\to\infty}\left(1+\frac{a}{n}+\frac{b}{n^{3/2}}+...\right)^n=e^a$$

(terms with higher-than-one power of n in the denominator do not affect it) to see that the limit of (8.2) is

$$\lim_{n\to\infty}M_{Z_n}(t)=\exp\left(\frac{t^2}{2}\right)$$

clearly identifiable as the MGF of $\mathcal{N}(0,1)$. ■

We may then use the CLT to approximate the distribution of $\bar{X}$ by the Normal distribution with the mean of μ and the standard deviation of $\frac{\sigma}{\sqrt{n}}$. The accuracy of this approximation increases with n and deteriorates when n becomes too small. How small is 'small' depends on the distribution from which we sample; for some such distributions $n\geq 10$ may be more than sufficient, for others, averaging even thousands of observations may not be enough! The usual requirement that n be at least 30 should thus be seen only as a rule of thumb which can be applied to most, but definitely not to all situations.

Example: A standard deck of cards is shuffled and 10 cards are dealt from it; let X represent the number of spades obtained. This is repeated, independently, 30 times. Use the Normal approximation to find the $\Pr\left(\bar{X}>3\right)$.

Solution:

$$\Pr\left(\bar{X}>3\right)\simeq\Pr\left(\frac{\bar{X}-10\cdot\frac{1}{4}}{\sqrt{\frac{10\cdot\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{42}{51}}{30}}}>\frac{\frac{90.5}{30}-2.5}{0.22687}\right)$$

$$=\Pr\left(Z>2.2774\right)=\frac{1}{\sqrt{2\pi}}\int_{2.2774}^{\infty}\exp\left(-\tfrac{z^2}{2}\right)dz=1.138\%$$

Here one should realize that, when the total number of spades one gets in 30 rounds of this experiment is 90, the $\bar{X}>3$ condition is still not met (90 is the largest value still *not* contributing), whereas 91 is the first value which meets this condition; note that in our computation we took the average of the two values (namely 90.5) - this is called the CONTINUITY CORRECTION and it helps to improve the approximation's accuracy.

In this case it is possible to find (with the help of PGF) the exact answer, which turns out to be 1.226%. The error of the approximation is thus only 0.088%.

8.2 Sampling a bivariate distribution

This time, the RIS consists of n independent pairs of (X, Y) observations; it is important to realize that the Xs are mutually independent and so are the Ys; furthermore X_i is also independent of Y_j when $i \neq j$, but X_i and Y_i are correlated (for each i from 1 to n). Their (common) covariance (let us denote it C) can be computed based on the bivariate distribution being sampled.

For the two resulting sample means $\bar{X}$ and $\bar{Y}$ the univariate formulas still hold, namely

$$\begin{aligned} E\left(\bar{X}\right) &= \mu_x \\ \mathrm{Var}(\bar{X}) &= \frac{\sigma_x^2}{n} \\ E\left(\bar{Y}\right) &= \mu_y \\ \mathrm{Var}(\bar{Y}) &= \frac{\sigma_y^2}{n} \end{aligned}$$

but how about their covariance? Well, it is not to difficult to see that

$$\mathrm{Cov}(\bar{X}, \bar{Y}) = \frac{\sum_{i,j=1}^{n} \mathrm{Cov}(X_i, Y_j)}{n^2} = \frac{\sum_{i=1}^{n} \mathrm{Cov}(X_i, Y_i)}{n^2} = \frac{C}{n}$$

due to the distributive law of covariance. This implies that

$$\rho_{\bar{X},\bar{Y}} = \frac{\frac{C}{n}}{\frac{\sigma_X}{\sqrt{n}} \cdot \frac{\sigma_X}{\sqrt{n}}} = \frac{C}{\sigma_X \cdot \sigma_X} = \rho$$

i.e. the correlation coefficient between the two sample means is the same as the correlation coefficient between a single pair of X, Y values (unlike the variances, which both decrease with n).

8.2.1 Bivariate CLT

Taking the same approach as in the univariate case (skipping some of the details) and defining

$$\begin{aligned} U_i &\equiv \frac{X_i - \mu_x}{\sigma_x} \\ V_i &\equiv \frac{Y_i - \mu_y}{\sigma_y} \end{aligned}$$

we can expand the joint MGF of $(\frac{U_i}{\sqrt{n}}, \frac{V_i}{\sqrt{n}})$ using (6.1):

$$1 + \frac{t_1^2}{2n} + \frac{t_2^2}{2n} + \frac{\rho}{n} t_1 t_2 + ...$$

Raising this to the power of n and taking the $n \to \infty$ limit yields the joint MGF of the two *standardized* sample means, getting

$$\exp\left(\frac{t_1^2 + t_2^2 + 2\rho\ t_1 t_2}{2}\right) \tag{8.3}$$

This corresponds to the so-called

8.2.2 Standardized bivariate Normal

distribution, which has the following joint pdf:

$$f(z_1, z_2) = \frac{\exp\left(-\frac{z_1^2 + z_2^2 - 2\rho\ z_1 z_2}{2(1-\rho^2)}\right)}{2\pi\sqrt{1-\rho^2}} \qquad \text{everywhere}$$

where we have removed the subscripts from ρ (it is helpful to plot a few examples - use Maple). Let us verify (by a relatively simple double integration) that this distribution has a joint MGF given by (8.3).

$$\begin{aligned}
& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{z_1^2 + z_2^2 - 2\rho\ z_1 z_2}{2(1-\rho^2)} + t_1 z_1 + t_2 z_2\right) dz_1 dz_2 \\
= & \int_{-\infty}^{\infty} \exp\left(-\frac{z_2^2 - 2t_2 z_2(1-\rho^2)}{2(1-\rho^2)}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{z_1^2 - 2\rho\ z_1 z_2 - 2t_1 z_1(1-\rho^2)}{2(1-\rho^2)}\right) dz_1 dz_2 \\
= & \exp\left(\frac{t_1^2(1-\rho^2)}{2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{z_2^2 - 2t_2 z_2(1-\rho^2) - \rho^2 z_2^2 - 2\rho\ z_2\ t_1(1-\rho^2)}{2(1-\rho^2)}\right) \times \\
& \int_{-\infty}^{\infty} \exp\left(-\frac{\left(z_1 - \rho\ z_2 - t_1(1-\rho^2)\right)^2}{2(1-\rho^2)}\right) dz_1 dz_2 \\
= & \exp\left(\frac{t_1^2(1-\rho^2)}{2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{z_2^2 - 2t_2 z_2 - 2\rho\ z_2\ t_1}{2}\right) dz_2 \times \sqrt{2\pi(1-\rho^2)} \\
= & \exp\left(\frac{t_1^2(1-\rho^2) + t_2^2 + t_1^2\rho^2 + 2\rho t_1 t_2}{2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{(z_2 - t_2 - \rho\ t_1)^2}{2}\right) dz_2 \times \sqrt{2\pi(1-\rho^2)} \\
= & \exp\left(\frac{t_1^2 + t_2^2 + 2\rho t_1 t_2}{2}\right) \times 2\pi\sqrt{(1-\rho^2)}
\end{aligned}$$

yet to be divided by $2\pi\sqrt{(1-\rho^2)}$; this yields (8.3). ■

It is obvious that both marginals of this distribution are $\mathcal{N}(0,1)$, i.e. standardized Normal; let us find the conditional distribution of Z_1 given $Z_2 = \mathbf{z}$.

This is a routine exercise:

$$f(z_1|Z_2 = \mathbf{z}) = \frac{f(z_1, \mathbf{z})}{\frac{\exp(-\frac{\mathbf{z}^2}{2})}{\sqrt{2\pi}}} = \frac{\exp\left(\frac{z_1^2 + \mathbf{z}^2 - 2\rho\ z_1\mathbf{z} - \mathbf{z}^2 + \mathbf{z}^2\rho^2}{2(1-\rho^2)}\right)}{\sqrt{2\pi} \cdot \sqrt{1-\rho^2}}$$

$$= \frac{\exp\left(\frac{(z_1 - \rho\ \mathbf{z})^2}{2(1-\rho^2)}\right)}{\sqrt{2\pi} \cdot \sqrt{1-\rho^2}}$$

which can be easily identified as $\mathcal{N}(\rho\,\mathbf{z}, \sqrt{1-\rho^2})$; one should be able to visualize this.

Note that $\rho = 0$ *does* imply that Z_1 and Z_2 are *independent* (true *only* for the Normal distribution).

8.2.3 General bivariate Normal

distribution results from re-scaling Z_1 and Z_2, thus:

$$X = \sigma_x Z_1 + \mu_x$$
$$Y = \sigma_y Z_2 + \mu_y$$

(Please note that these two RVs are no longer the X and Y of the distribution which we sampled two sections ago.) Also note that such a linear transformation does not change the value of a correlation coefficient; X and Y thus have the same ρ as the original Z_1 and Z_2 pair did.

This also implies that any further linear transformation of either X or Y (or, individually, each X and Y, e.g. $U = 3X - 4$ and $V = -2Y + 7$) will result in the new pair still having a bivariate Normal distribution - only the corresponding means and sigmas need to be recomputed. The new correlation coefficient ρ_{uv} will have the same absolute value as that of ρ_{xy}, but it will *change sign* whenever the *linear coefficients* (3 and -2 of our example) have opposite signs.

In a more complicated transformation of the $U = 5X - 3Y + 2$ and $V = 4X + 2Y - 4$ type, the joint distribution of U and V also remains bivariate Normal, but this time we have to recompute $\rho_{U,V}$ as well as the two means and sigmas.

Notation: $(X, Y) \epsilon\, \mathcal{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$; we need five parameters to specify this distribution.

The corresponding *marginal* distributions of (individually) X and Y are $\mathcal{N}(\mu_x, \sigma_x)$ and $\mathcal{N}(\mu_y, \sigma_y)$ respectively. This follows from their joint MGF (the next formula) by substituting $t_2 = 0$ and $t_1 = 0$ respectively.

The *joint* pdf of the new pair gets kind of messy, but it is a routine exercise

to spell it out as well (we will not do it at this point). The joint MGF becomes

$$M_{xy}(t_1, t_2) = \mathbb{E}\left(e^{X\cdot t_1 + Y\cdot t_2}\right) = \mathbb{E}\left(e^{\sigma_x Z_1\cdot t_1 + \sigma_y Z_2\cdot t_2}\right) e^{\mu_x t_1 + \mu_{yY} t_2}$$

$$= \exp\left(\frac{\sigma_x^2\cdot t_1^2 + \sigma_y^2\cdot t_2^2 + 2\mathrm{Cov}(X,Y)\cdot t_1 t_2}{2} + \mu_x t_1 + \mu_y t_2\right)$$

The *conditional* distribution of X given $Y = \mathbf{y}$ equals the conditional distribution of $\sigma_x Z_1 + \mu_x$ given $Z_2 = \frac{\mathbf{y}-\mu_y}{\sigma_y}$, namely

$$\mathcal{N}\left(\mu_x + \sigma_x\rho\frac{\mathbf{y}-\mu_y}{\sigma_y}\ ,\ \sigma_x\sqrt{1-\rho^2}\right)$$

To answer a probability question about such X and Y, we can always convert it into a question about Z_1 and Z_2 and (double) integrate the corresponding pdf.

Chapter 9

Transforming RVs

In this chapter we investigate the following issue: Given the distribution of X, how do we find the distribution of $Y \equiv g(X)$ for any specific function g. We will deal exclusively with continuous distributions (the discrete case is less interesting and more messy).

We refer to X as the *old* RV, Y is the *new* RV (X is TRANSFORMED into Y).

Eventually, we also learn how to transform *two* RVs into *one* or *two* new RVs, etc.

9.1 Univariate case

by which we mean transforming *one* old RV into *one* new RV.

There are two techniques for doing this: the cdf technique (more general but also computationally more clumsy) and the pdf technique (it requires the transformation to be one-to-one, but it is then more elegant and faster).

9.1.1 Distribution-function (or F) technique

works as follows: we find the cdf of the new RV Y by computing $\Pr(Y \leq y) = \Pr\left(g(X) \leq y\right)$. This amounts to solving the $g(X) \leq y$ inequality for X (usually resulting in an interval of values), and then integrating $f(x)$ over this interval.

Cauchy example: Consider $X \epsilon \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$; this corresponds to a spinning wheel with a *two-directional* pointer, say a laser beam attached to it, where X is the pointer's angle from a fixed direction when the wheel stops spinning. We want to know the distribution of $Y = \tilde{\sigma}\tan(X) + \tilde{\mu}$; this represents the location of a dot our laser beam would leave on a screen placed $\tilde{\sigma}$ units from the wheel's center, with a scale whose origin is $\tilde{\mu}$ units off center.

Solution: We start by writing down

$$F_x(x) = \frac{x + \frac{\pi}{2}}{\pi} = \frac{x}{\pi} + \frac{1}{2} \qquad \text{when} \qquad -\frac{\pi}{2} < x < \frac{\pi}{2}$$

To get $F_y(y)$ we proceed as follows:

$$\begin{aligned} \Pr\left(\tilde{\sigma}\tan(X)+\tilde{\mu}\le y\right) &= \Pr\left(X\le \arctan(\frac{y-\tilde{\mu}}{\tilde{\sigma}})\right) = \\ F_x\left(\arctan(\frac{y-\tilde{\mu}}{\tilde{\sigma}})\right) &= \frac{1}{\pi}\arctan(\frac{y-\tilde{\mu}}{\tilde{\sigma}})+\frac{1}{2} \end{aligned}$$

where y can have any real value. Usually, we can relate better to the corresponding pdf, namely

$$f_y(y)=\frac{1}{\pi}\cdot\frac{\tilde{\sigma}}{\tilde{\sigma}^2+(y-\tilde{\mu})^2}$$

This function looks superficially similar to the Normal pdf (it is bell shaped), but its properties are very different. ■

The distribution we have just discovered is so important that we dedicate the next section to it; we call it the

Cauchy distribution

Since the $\int_{-\infty}^{\infty} y\cdot f_Y(y)\,dy$ integral leads to $\infty-\infty$, its mean does not exist (its INDEFINITE), similarly, its variance has an infinite value. Yet it possesses a clear *center* at $y=\tilde{\mu}$ and a well-defined *width* equal to $\tilde{\sigma}$. But if these are not the mean and standard deviation, what are they?

The answer: $\tilde{\mu}$ is the so called MEDIAN of the distribution, defined (for *any* distribution, not just Cauchy) as the unique solution to

$$F(\tilde{\mu})=\frac{1}{2}$$

(the corresponding RV has a 50% chance to be smaller than $\tilde{\mu}$), and $\tilde{\sigma}$ is its *semi-inter-quartile range* (QUARTILE DEVIATION for short), defined by

$$\tilde{\sigma}\equiv\frac{Q_U-Q_L}{2}$$

where Q_U and Q_L are the UPPER and LOWER QUARTILES, i.e. solutions to $F(Q_U)=\frac{3}{4}$ and $F(Q_L)=\frac{1}{4}$ respectively. One can easily verify that, in the case of Cauchy distribution, $Q_L=\tilde{\mu}-\tilde{\sigma}$ and $Q_U=\tilde{\mu}+\tilde{\sigma}$.

Notation: $Y\epsilon\ C(\tilde{\mu},\tilde{\sigma})$

More examples

of a univariate transformation.

Example: Let X have the following pdf:

$$f(x)=6x(1-x)\qquad\text{when}\qquad 0<x<1$$

with the understanding (from now on) of: *zero otherwise*. Find the pdf of $Y=X^3$.

Solution: First we realize that $0 < Y < 1$. Secondly, we find

$$F_x(x) = 6 \int_0^x (x - x^2)\, dx = 6(\frac{x^2}{2} - \frac{x^3}{3}) = 3x^2 - 2x^3$$

And finally:

$$\begin{aligned} F_y(y) &\equiv \Pr(Y \le y) = \Pr(X^3 \le y) = \\ \Pr(X &\le y^{\frac{1}{3}}) = F_x(y^{\frac{1}{3}}) = 3y^{\frac{2}{3}} - 2y \end{aligned}$$

This easily converts to

$$f_y(y) = 2y^{-\frac{1}{3}} - 2 \qquad \text{when} \qquad 0 < y < 1$$

(zero otherwise - one last time). Note that when $y \to 0$ this *pdf* becomes infinite. ■

Example (transforming Uniform to Exponential): Let $X \epsilon\, \mathcal{U}(0,1)$. Find and identify the distribution of $Y = -\ln X$ (its support is obviously $0 < y < \infty$).

Solution: First we need $F_x(x) = x$ when $0 < x < 1$. Then

$$\begin{aligned} F_y(y) &= \Pr(-\ln X \le y) = \Pr(X \ge e^{-y}) \\ &= 1 - F_x(e^{-y}) = 1 - e^{-y} \qquad \text{when} \quad y > 0 \end{aligned}$$

This implies

$$f_y(y) = e^{-y} \qquad \text{when} \quad y > 0$$

which can be identified a $\mathcal{E}(1)$. Note that $Y = -\beta \cdot \ln X$ (where $\beta > 0$) would be $\mathcal{E}(\beta)$.

Example (introducing χ^2_1): Assuming that $Z \epsilon\, \mathcal{N}(0,1)$, find the distribution of $Y = Z^2$.

Solution:

$$F_y(y) = \Pr(Z^2 \le y) = \Pr(-\sqrt{y} \le Z \le \sqrt{y}) = F_Z(\sqrt{y}) - F_Z(\sqrt{y}).$$

Since we do not have an explicit expression for $F_Z(z)$ it would appear that we are stuck at this point, but we can still get the corresponding $f_Y(y)$ by simple differentiation:

$$\begin{aligned} f_y(y) &= \frac{dF_y(y)}{dy} = \frac{dF_z(\sqrt{y})}{dy} - \frac{dF_z(-\sqrt{y})}{dy} \\ &= \frac{1}{2} y^{-\frac{1}{2}} f_z(\sqrt{y}) + \frac{1}{2} y^{-\frac{1}{2}} f_z(-\sqrt{y}) = \frac{y^{-\frac{1}{2}} e^{-\frac{y}{2}}}{\sqrt{2\pi}} \end{aligned}$$

when $y > 0$. This can be identified as the $\mathsf{gamma}(\frac{1}{2}, 2)$ distribution, but due to its importance, we also call it a CHI-SQUARE distribution with *one* degree of freedom. Note that its MGF is thus equal to $(1 - 2t)^{-1/2}$. ■

General chi-square

is defined as the distribution of a sum of k *independent* RVs of the $\mathcal{N}(0,1)$ type. Its MGF is thus given by

$$M(t) = \left((1-2t)^{-1/2}\right)^k = \frac{1}{(1-2t)^{k/2}}$$

which can be identified as MGF of the $\mathsf{gamma}(\frac{k}{2}, 2)$ distribution (we know all its properties already). It is also called the chi-square distribution with k (integer) degrees of freedom and denoted χ^2_k.

Linear transformation

Note that *any* RV X can be transformed into $Y = \sigma X + \mu$ where $\sigma > 0$ becomes the so-called SCALE (scaling) parameter, and μ is the LOCATION parameter of the new distribution. Often Y represents the original X expressed in new units (ie. converting temperature from Celsius to Fahrenheit, weight from kilograms to pounds, distance from km to miles, etc.), thus preserving the *shape* of the distribution This implies that

$$\begin{aligned} F_y(y) &= \Pr(\sigma X + \mu \le y) = \Pr\left(X \le \frac{y-\mu}{\sigma}\right) = F_x\left(\frac{y-\mu}{\sigma}\right) \\ f_y(y) &= \frac{1}{\sigma} \cdot f_x\left(\frac{y-\mu}{\sigma}\right) \end{aligned}$$

In each of these cases, *any* formula relating to X can then be easilty converted into a corresponding formula for Y. This means that, by understanding the $\mathcal{C}(0,1)$, $\mathcal{N}(0,1)$, $\mathcal{E}(1)$ and $\mathcal{U}(0,1)$ distributions, we can then easily deal with $\mathcal{C}(\tilde{\mu}, \tilde{\sigma})$, $\mathcal{N}(\mu, \sigma)$, $\mathcal{E}(\beta)$ and $\mathcal{U}(a,b)$, since $\tilde{\mu}$, μ and a are the respective location parameters and $\tilde{\sigma}$, σ and $b-a$ are similarly the scaling parameters.

In the case of $\mathsf{gamma}(\alpha, \beta)$ distribution, while β remains a scale parameter (inherited from exponential), α is the so-called SHAPE parameter; distinct values of α lead to distributions of different shapes and properties.

9.1.2 The pdf (or f) technique

is a bit faster and usually somehow easier (technically) to carry out, but it works for *one-to-one* (typically either increasing or decreasing - such functions are called MONOTONE) transformations *only* (e.g. it would not work in our last $Y = Z^2$ example). This can be easily established by plotting the transformation function over the *support* of the old RV (i.e. ignoring values which cannot happen; one can also ignore individual, ie. single-point exceptions to the one-to-one rule).

Incidentally, note that for *monotone* transformations

$$\tilde{\mu}_Y = g(\tilde{\mu}_X)$$

i.e., unlike the mean, the *median* does transfrom according to g. In any such (monotone) case, we get

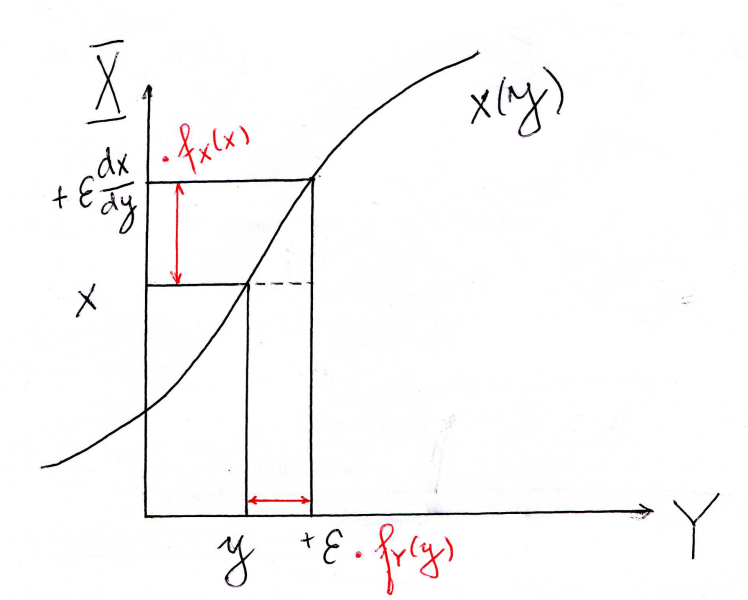
$$F_y(y) = \Pr(g(X) < y) = \begin{cases} \Pr(X < g^{-1}(y)) = F_x\left(g^{-1}(y)\right) & \text{increasing} \\ \Pr(X > g^{-1}(y)) = 1 - F_x\left(g^{-1}(y)\right) & \text{decreasing} \end{cases}$$

To build the pdf of $Y = g(X)$, all we have to do is

$$f_y(y) = \pm f_x\left(g^{-1}(y)\right) \cdot \frac{dg^{-1}(y)}{dy} = f_x\left(g^{-1}(y)\right) \cdot \left|\frac{dg^{-1}(y)}{dy}\right|$$

(similar to what we do when replacing $\int f_x(x)\, dx$ by $\int \ldots\, dy$ integration).

The proof rests on the following picture:



The procedure then consists of three simple steps (after verifying that the transformation is one-to-one and establishing the interval of possible Y values as a by-product), namely:

(i) Solve the $Y = g(X)$ equation for x (in terms of y), switching to small letters; we denote this solution by $x(y)$ - this is always a *specific* function of y.

(ii) Substitute $x(y)$ for the argument of $f_X(x)$, getting a function of y.

(iii) Multiply this by $\left|\frac{dx(y)}{dy}\right|$.

This results in $f_y(y)$ of the new RV.
Let us quickly re-do the examples used to demonstrate the cdf technique.

- $X \epsilon\, \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$ and $Y = \tilde{\sigma}\tan(X) + \tilde{\mu}$

Solution:

$$\begin{aligned} x(y) &= \arctan(\frac{y-\tilde{\mu}}{\tilde{\sigma}}) \\ f_x(x(y)) &= \frac{1}{\pi} \\ f_y(y) &= \frac{1}{\pi}\cdot\frac{dx(y)}{dy} = \frac{1}{\pi}\cdot\frac{\tilde{\sigma}}{\tilde{\sigma}^2+(y-\tilde{\mu})^2} \end{aligned}$$

where y can have any real value. ■

- $f(x) = 6x(1-x)$ when $0 < x < 1$, and $Y = X^3$

Solution:

$$\begin{aligned} x(y) &= y^{1/3} \\ f_x(x(y)) &= 6y^{1/3}(1-y^{1/3}) \\ f_y(y) &= 6y^{1/3}(1-y^{1/3})\cdot\frac{1}{3}y^{-2/3} = 2(y^{-1/3}-1) \end{aligned}$$

when $0 < y < 1$. ■

- $X \epsilon\, \mathcal{U}(0,1)$ and $Y = -\ln X$

Solution:

$$\begin{aligned} x(y) &= e^{-y} \\ f_x(x(y)) &= 1 \\ f_y(y) &= 1\cdot e^{-y} = e^{-y} \quad \text{when } y > 0 \end{aligned}$$ ■

Let us now move on to

9.2 Bivariate Transformations

First we discuss the case of transforming *two* RVs into a *single* one.

9.2.1 The cdf (or F) technique

follows essentially the same pattern as the univariate case:

The new random variable Y is now defined by $Y \equiv g(X_1, X_2)$, where we know the bivariate distribution of X_1 and X_2 (they are not necessarily independent, even though in most of our examples they are).

We find

$$F_y(y) = \Pr(Y \le y) = \Pr\left(g(X_1, X_2) \le y\right)$$

by realizing that the $g(X_1, X_2) \le y$ inequality corresponds to (for each y) a *2D region* in the (x_1, x_2) plane. Once we establish how this region looks like, all we need to compute is the double integral of $f(x_1, x_2)$ over this region. The technique is thus simple in principle, but *not* in technical details.

Example (ratio of two exponentials): Suppose that X_1 and X_2 *are* independent RVs, both from $\mathcal{E}(1)$, and $Y = \dfrac{X_2}{X_1}$. Note that we would be getting the same answer were both X_1 and X_2 from $\mathcal{E}(\beta)$, as long as it is the *same* (positive) β.

Solution:

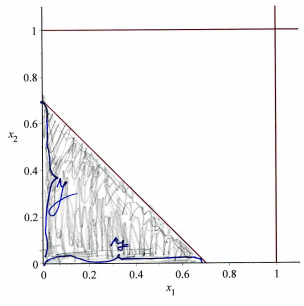
$$\begin{aligned} F_y(y) &= \Pr\left(\frac{X_2}{X_1} \le y\right) = \Pr(X_2 \le y\, X_1) = \iint\limits_{0 \le x_2 \le y x_1} e^{-x_1 - x_2}\, dx_1\, dx_2 \\ &= \int_0^\infty e^{-x_1} \int_0^{y\, x_1} e^{-x_2}\, dx_2 dx_1 = \int_0^\infty e^{-x_1}(1 - e^{-y\, x_1}) dx_1 \\ &= \int_0^\infty \left(e^{-x_1} - e^{-x_1(1+y)}\right) dx_1 = 1 - \frac{1}{1+y} \qquad \text{when} \quad y > 0 \end{aligned}$$

This implies that

$$f_y(y) = \frac{1}{(1+y)^2} \quad \text{when} \quad y > 0$$

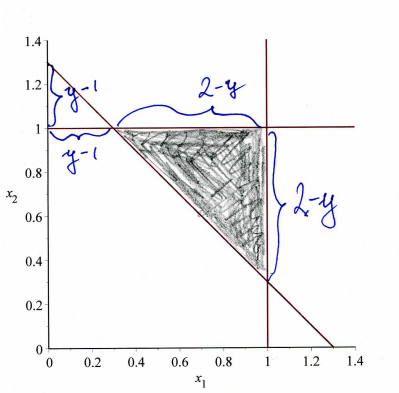
which is an example of a RV having an infinite mean and variance (sometimes we say that they 'do not exist', meaning that they do not have finite values). It is also a special case of a so-called *Fisher* distribution (introduced later). ■

Example - sum of two $\mathcal{U}(0,1)$: To find $F_Y(y)$ of $Y = X_1 + X_2$, where the two Xs are independent, each having the $\mathcal{U}(0,1)$ distribution, we must integrate their joint pdf, identically equal to 1, over the following region



when $0 < y < 1$. This answer is simply the area of the corresponding triangle, equal to $\frac{y^2}{2}$.

Similarly, when $1 < y < 2$, $\Pr(X_1 + X_2 < y)$ is equal to 1 minus the area of the following triangle



which yields $1 - \frac{(2-y)^2}{2}$. Differentiating, we get the corresponding pdf:

$$f_y(y) = \begin{cases} y & 0 < y < 1 \\ 2 - y & 1 < y < 2 \end{cases} \qquad \blacksquare$$

χ^2_2 **example:** This time Z_1 and Z_2 are independent RVs from $\mathcal{N}(0,1)$ and $Y = Z_1^2 + Z_2^2$ (we have done this already and know the answer, but let us proceed anyhow).

Solution:

$$F_y(y) = \Pr(Z_1^2 + Z_2^2 \le y) = \frac{1}{2\pi} \iint\limits_{z_1^2+z_2^2 \le y} e^{-\frac{z_1^2+z_2^2}{2}} dz_1 dz_2$$

$$\overset{\text{go polar}}{=} \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\sqrt{y}} e^{-\frac{r^2}{2}} \cdot r \, dr d\theta \overset{w=\frac{r^2}{2}}{=} \int_0^{\frac{y}{2}} e^{-w} dw = 1 - e^{-\frac{y}{2}}$$

when $y > 0$. This is the cdf of $\mathcal{E}(2)$ - same as χ^2_2. $\blacksquare$

Convolution example: Assume that X_1 and X_2 are independent RVs whose pdf is $f_1(x_1)$ and $f_2(x_2)$ respectively. Find the distribution of $Y = X_1 + X_2$.

Solution:

$$F_y(y) = \Pr(X_1 + X_2 \le y) = \iint\limits_{x_1+x_2 \le y} f_1(x_1) \cdot f_2(x_2) dx_1 dx_2$$

$$= \int_{-\infty}^{\infty} f_1(x_1) \cdot \left(\int_{-\infty}^{y-x_1} f_2(x_2) dx_2 \right) dx_1$$

Differentiating it with respect to y (remember how it is done?) results in

$$\begin{aligned} f_y(y) &= \int_{-\infty}^{\infty} f_1(x_1) \cdot f_2(y - x_1) dx_1 \\ &\equiv \int_{-\infty}^{\infty} f_1(x) \cdot f_2(y - x) dx \end{aligned}$$

Doing it the other way around must yield the same answer:

$$f_y(y) = \int_{-\infty}^{\infty} f_2(x) \cdot f_1(y - x) dx$$

Combining two pdfs in this manner is called their CONVOLUTION. ■

Special cases of convolution

- When X_1 and X_2 are independent and each has the $\mathcal{U}(0, 1)$ distribution, the pdf of $Y \equiv X_1 + X_2$ is

$$f_y(y) = \int_{\max(0,y-1)}^{\min(1,y)} 1 \, dx = \begin{cases} y & \text{when} \quad 0 < y < 1 \\ 2 - y & \text{when} \quad 1 < y < 2 \end{cases}$$

  We will call this (rather informally) the *triangular* distribution. ■

- Replacing $\mathcal{U}(0, 1)$ by $\mathcal{C}(0, 1)$, whose pdf is $f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$, we get

$$f_y(y) = \frac{1}{\pi^2} \int_{-\infty}^{\infty} \frac{1}{1 + x^2} \cdot \frac{1}{1 + (y - x)^2} dx = \frac{2}{\pi} \cdot \frac{1}{4 + y^2}$$

  (for any real y).

  The last result can be easily converted to the pdf of $U = \frac{Y}{2} = \frac{X_1+X_2}{2}$ (the *sample mean* of two observations), yielding

$$f_u(u) = 2 \cdot f_y(2u) = 2 \cdot \frac{2}{\pi} \cdot \frac{1}{4 + (2u)^2} = \frac{1}{\pi} \cdot \frac{1}{1 + u^2}$$

  Thus, the sample mean has the *same* Cauchy distribution as each of the two individual observations (furthermore, this can be extended to a sample of *any* size). This is a shocking result: it implies that the sample mean of even millions of values from a Cauchy distribution cannot estimate the location of its center (of the laser gun hidden behind a screen) any better than a *single* observation!!? We knew that the CLT (which requires μ and σ to be finite) would break down in this case, but few of us expected this. But, if the sample mean fails so spectacularly to locate the center, is there something else (a different *sample statistic*) we could use for this purpose? Something which would *improve* (in terms of its standard deviation getting smaller) when the sample size increases. We address this issue later.

- Replacing $\mathcal{C}(0,1)$ by $\mathcal{E}(\beta)$, the pdf of $Y = X_1 + X_2$ is

$$f_y(y) = \frac{1}{\beta^2} \int_0^y \exp(-\beta\ x) \exp(-\beta(y-x))\ dx = \frac{y\ \exp(-\beta\ y)}{\beta^2} \quad y > 0$$

  which can be identified as the pdf of gamma(2,β) - we had expected that, did not we? ■

9.2.2 The pdf (or f) technique

works somehow faster, even though it may *appear* to be more complicated. It is based on the same principles as when changing variables in a double integral.

Example: Using polar coordinates, this is how we would deal with the following double integral

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{z_1^2+z_2^2}{2}\right) dz_1 dz_2 &= \int_0^{2\pi} \frac{1}{2\pi} \int_0^{\infty} \exp\left(-\frac{r^2}{2}\right) r\ dr\ d\theta \\ &= \int_0^{2\pi} \frac{1}{2\pi}\ d\theta \times \int_0^{\infty} \exp\left(-\frac{r^2}{2}\right) r\ dr \end{aligned}$$

implying that, when Z_1 and Z_2 and independent standardized Normal, $R = \sqrt{Z_1^2 + Z_2^2}$ is a RV with a pdf equal to $r \exp\left(-\frac{r^2}{2}\right)$ when $0 < r < \infty$, and $\Theta = \arctan(Z_2, Z_1)$ is $\mathcal{U}$niform$(0, 2\pi)$.

The technique thus requires the following steps:

- It can work only for *one-to-one* (i.e. 'invertible') transformations. The necessary (but *not* sufficient) condition to make this possible requires transforming two old RVs into *two* new RVs. This implies that the new RV $Y \equiv g(X_1, X_2)$ must be accompanied by yet another *arbitrarily* chosen function of X_1 and/or X_2 (taken to be the second *new* RV). We usually choose this auxiliary RV in the simplest possible manner, i.e. we make it equal to X_2 (or X_1). We then re-label the original Y as Y_1 and call the auxiliary RV Y_2. We then:

- **Invert** the transformation, i.e. solve the two equations $y_1 = g(x_1, x_2)$ and $y_2 = x_2$ for x_1 and x_2 (ie. express each x_1 and x_2 in terms of y_1 and y_2). Getting a *unique* solution guarantees that the transformation is one-to-one.

- **Substitute** this solution $x_1(y_1, y_2)$ and $x_2(y_2)$ into the joint pdf of the old X_1, X_2 pair (yielding a function of y_1 and y_2).

- **Multiply** the resulting function by the transformation's JACOBIAN, defined as $\begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$. This yields the joint pdf of Y_1 and Y_2. At the same

time, we must establish the 2D region of possible (y_1, y_2) values (this is often the most difficult part of the procedure - this should be done by spelling out the *conditional* range of Y_2 and the *marginal* range of Y_1).

- **Eliminate** Y_2 (the 'phoney' RV) by dy_2 integration of the joint pdf of Y_1 and Y_2 over the conditional range of Y_2.

We will go over a few examples.

'Catching fish' example: X_1, $X_2 \epsilon\ \mathcal{E}(1)$, independent; $Y \equiv \frac{X_1}{X_1+X_2}$ (the time of the first 'catch', relative to the time needed to catch two fish). Again, changing $\mathcal{E}(1)$ to $\mathcal{E}(\beta)$, as long as it is the same β for both Xs, does no affect the answer, since $\frac{X}{\beta} \epsilon\, \mathcal{E}(1)$ when $X \epsilon\, \mathcal{E}(\beta)$.

Solution: Re-labelling Y as Y_1 and adding $Y_2 \equiv X_2$, we get

$$\begin{aligned} x_1 y_1 + x_2 y_1 &= x_1 \\ y_2 &= x_2 \end{aligned}$$

which implies

$$\begin{aligned} x_2(y_2) &= y_2 \\ x_1(y_1, y_2) &= \frac{y_1 \cdot y_2}{1 - y_1} \end{aligned}$$

Substitute into $f(x_1, x_2) = e^{-x_1 - x_2}$ getting

$$e^{-y_2\left(1+\frac{y_1}{1-y_1}\right)} = e^{-\frac{y_2}{1-y_1}}$$

Multiply it by

$$\left| \begin{array}{cc} y_2 \frac{1-y_1+y_1}{(1-y_1)^2} & \frac{y_1}{1-y_1} \\ 0 & 1 \end{array} \right| = \frac{y_2}{(1-y_1)^2}$$

getting

$$f(y_1, y_2) = \frac{y_2 \cdot e^{-\frac{y_2}{1-y_1}}}{(1-y_1)^2} \qquad \text{when} \quad 0 < y_1 < 1 \quad \text{and} \quad y_2 > 0$$

Eliminate Y_2 by

$$\int_0^\infty \frac{y_2 \cdot e^{-\frac{y_2}{1-y_1}}}{(1-y_1)^2} dy_2 = \frac{1}{(1-y_1)^2} \cdot (1-y_1)^2 = 1 \quad \text{when} \quad 0 < y_1 < 1$$

using

$$\int_0^\infty x^k e^{-\frac{x}{a}}\, dx = k! \cdot a^{k+1} \tag{9.1}$$

The distribution of Y is thus $\mathcal{U}(0, 1)$. ■

Another example: Same X_1 and X_2 as in the previous example, $Y = \frac{X_2}{X_1}$ (solved earlier, using the F technique).

Solution: This time, we reverse the labels: $Y_1 \equiv X_1$ and $Y_2 = \frac{X_2}{X_1} \Rightarrow x_1 = y_1$ and $x_2 = y_1 \cdot y_2$ (to simplify our notation, x_1 and x_2 are functions of y_1 and y_2 only *implicitly*). Substitute into $e^{-x_1-x_2}$ to get $e^{-y_1(1+y_2)}$, times

$$\left| \begin{array}{cc} 1 & 0 \\ y_2 & y_1 \end{array} \right| = y_1$$

yields the joint pdf when $y_1 > 0$ and $y_2 > 0$. Eliminate y_1 by

$$\int_0^\infty y_1 e^{-y_1(1+y_2)} dy_1 = \frac{1}{(1+y_2)^2}$$

when $y_2 > 0$. Thus

$$f_y(y) = \frac{1}{(1+y)^2} \qquad \text{when} \quad y > 0$$

(eliminating the now-redundant subscript). ■

Beta distribution

Let X_1 and X_2 be independent RVs from the gamma distribution with parameters (k, β) and (m, β) respectively, and let $Y_1 \equiv \frac{X_1}{X_1+X_2}$. This is the relative time needed to catch the first k fish out of $k+m$ ('relative' means: divided by the total time).

Solution Using a previous argument, one can show that β 'cancels out', and we can assume that $\beta = 1$ without affecting the answer. Since $x_1(y_1, y_2)$, $x_2(y_2)$ and the Jacobian is the same as in the 'fishing example', all we need to do is to substitute into

$$f(x_1, x_2) = \frac{x_1^{k-1} x_2^{m-1} e^{-x_1-x_2}}{\Gamma(k) \cdot \Gamma(m)}$$

and multiply by the Jacobian, getting

$$f(y_1, y_2) = \frac{y_1^{k-1} y_2^{k-1} y_2^{m-1} e^{-\frac{y_2}{1-y_1}}}{\Gamma(k)\Gamma(m)(1-y_1)^{k-1}} \cdot \frac{y_2}{(1-y_1)^2}$$

when $0 < y_1 < 1$ and $y_2 > 0$. Integrating over y_2 with the help of (9.1) results in:

$$\begin{aligned} & \frac{y_1^{k-1}}{\Gamma(k)\Gamma(m)(1-y_1)^{k+1}} \int_0^\infty y_2^{k+m-1} e^{-\frac{y_2}{1-y_1}} dy_2 \\ = \quad & \frac{\Gamma(k+m)}{\Gamma(k) \cdot \Gamma(m)} \cdot y_1^{k-1} (1-y_1)^{m-1} \end{aligned}$$

when $0 < y_1 < 1$. This is a pdf of a new two-parameter distribution denoted beta(k, m). As a by-product of this exercise, we have effectively proved the following formula:

$$\int_0^1 y^{k-1}(1-y)^{m-1}dy = \frac{\Gamma(k)\cdot\Gamma(m)}{\Gamma(k+m)}$$

for any $k, m > 0$. This enables us (skipping the details) to find the distribution's *mean*

$$\frac{k}{k+m}$$

and *variance*

$$\frac{k\,m}{(k+m+1)\,(k+m)^2}$$

It is easy to see that the distribution of $1 - Y = \frac{X_2}{X_1+X_2}$ is also beta, what are its two parameters? Note that beta$(1, 1)$ is the same distribution as $\mathcal{U}(0, 1)$. Also note that k and m do *not* need to be integers (as long as they are both positive).

Student t distribution

We start with two independent RVs $X_1 \epsilon\, \mathcal{N}(0, 1)$ and $X_2 \epsilon\, \chi^2_m$, and introduce a new RV by $Y_1 \equiv \dfrac{X_1}{\sqrt{\frac{X_2}{m}}}$. The resulting distribution is called Student's t distribution with m degrees of freedom, notation: t_m.

Solution: Introducing $Y_2 \equiv X_2$, we get

$$\begin{aligned} x_2 &= y_2 \\ x_1 &= y_1 \cdot \sqrt{\frac{y_2}{m}} \end{aligned}$$

Substituting into

$$f(x_1, x_2) = \frac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} \cdot \frac{x_2^{\frac{m}{2}-1} e^{-\frac{x_2}{2}}}{\Gamma(\frac{m}{2})\cdot 2^{\frac{m}{2}}}$$

and multiplying by

$$\left| \begin{array}{cc} \sqrt{\frac{y_2}{m}} & \frac{1}{2}\cdot\frac{y_1}{\sqrt{m\ y_2}} \\ 0 & 1 \end{array} \right| = \sqrt{\frac{y_2}{m}}$$

results in

$$f(y_1, y_2) = \frac{e^{-\frac{y_1^2 y_2}{2m}}}{\sqrt{2\pi}} \cdot \frac{y_2^{\frac{m}{2}-1} e^{-\frac{y_2}{2}}}{\Gamma(\frac{m}{2})\cdot 2^{\frac{m}{2}}} \cdot \sqrt{\frac{y_2}{m}}$$

when $y_2 > 0$ (and any real y_1). To eliminate y_2, we do

$$\begin{aligned} & \frac{1}{\sqrt{2\pi}\Gamma(\frac{m}{2})\,2^{\frac{m}{2}}\sqrt{m}} \int_0^\infty y_2^{\frac{m-1}{2}} e^{-\frac{y_2}{2}(1+\frac{y_1^2}{m})} dy_2 \\ = & \frac{\Gamma(\frac{m+1}{2}) \cdot 2^{\frac{m+1}{2}}}{\sqrt{2\pi}\Gamma(\frac{m}{2}) \cdot 2^{\frac{m}{2}}\sqrt{m}\left(1+\frac{y_1^2}{m}\right)^{\frac{m+1}{2}}} \\ = & \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})\sqrt{m\pi}} \cdot \frac{1}{\left(1+\frac{y_1^2}{m}\right)^{\frac{m+1}{2}}} \end{aligned}$$

Note that when $m = 1$, this results in

$$\frac{1}{\pi} \cdot \frac{1}{1+y_1^2}$$

which is the $\mathcal{C}(0,1)$ distribution. When $m \to \infty$ ($m > 30$ is sufficient to be reasonably close to this limit) we get

$$\frac{\exp\left(-\frac{y_1^2}{2}\right)}{\sqrt{2\pi}}$$

clearly the $\mathcal{N}(0,1)$ distribution. Due to the $f(y_1) = f(-y_1)$ symmetry of the pdf, the corresponding mean is zero (when is exists, i.e. when $m \geq 2$). The variance can be computed (skipping the details) to have the following value:

$$\frac{m}{m-2}$$

when $m \geq 3$ (for $m = 1$ and 2, the variance is infinite).

Fisher F distribution

Let X_1 and X_2 be independent, having the χ^2_k and χ^2_m distribution, respectively, and let $Y_1 \equiv \frac{\frac{X_1}{k}}{\frac{X_2}{m}}$. The resulting distribution is called Fisher's F distribution with k and m degrees of freedom (numerator, followed by denominator). Notation: $\mathsf{F}_{k,m}$.

Solution: Introducing $Y_2 \equiv X_2$, we get

$$\begin{aligned} x_2 &= y_2 \\ x_1 &= \frac{k}{m} y_1 y_2 \end{aligned}$$

The Jacobian the equals to $\frac{k}{m} y_2$. Substituting into

$$\frac{x_1^{\frac{k}{2}-1} e^{-\frac{x_1}{2}}}{\Gamma(\frac{k}{2}) \cdot 2^{\frac{k}{2}}} \cdot \frac{x_2^{\frac{m}{2}-1} e^{-\frac{x_2}{2}}}{\Gamma(\frac{m}{2}) \cdot 2^{\frac{m}{2}}}$$

and multiplying by the Jacobian yields

$$\frac{(\frac{k}{m})^{\frac{k}{2}}}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})\cdot 2^{\frac{k+m}{2}}} y_1^{\frac{k}{2}-1}\cdot y_2^{\frac{k+m}{2}-1} e^{-\frac{y_2(1+\frac{k}{m}y_1)}{2}}$$

when $y_1 > 0$ and $y_2 > 0$. Integrating over y_2 (from 0 to ∞) yields the following final formula

$$f(y_1) = \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}(\frac{k}{m})^{\frac{k}{2}}\cdot \frac{y_1^{\frac{k}{2}-1}}{(1+\frac{k}{m}y_1)^{\frac{k+m}{2}}}$$

when $y_1 > 0$, The corresponding expected value is

$$\frac{m}{m-2}$$

when $m \geq 3$ (the mean is infinite for $m = 1$ and 2); the variance equals

$$\frac{2m^2(k+m-2)}{(m-2)^2(m-4)k}$$

when $m \geq 5$ (infinite for $m = 1, 2, 3$ and 4). It is obvious that $\frac{1}{Y}$ also has the Fisher's distribution (what are its degrees of freedom?).

9.3 More on Sampling

At this point, we need to return (for a while) to our discussion of **random independent sampling**.

9.3.1 Sample variance

Recall the definition of RIS of size n from a specific distribution as a collection of n RVs yet to be observed, independently, from this distribution. We have already dealt with the *sample mean* $\bar{X}$, and we know that, for large enough n, it has, approximately, the $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ distribution. The question of what is 'large enough' is usually answered by saying that n must be bigger than 30, but in reality it *does* depend on the distribution from which we sample (for some of them we get a good fit even when $n = 10$, but there are others where even millions would not suffice - such as playing a lottery).

In addition to $\bar{X}$ we now define the so called SAMPLE VARIANCE by

$$s^2 \equiv \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

where s, the corresponding square root, is the SAMPLE STANDARD DEVIATION (the sample variance does not have its own symbol).

To find its expected value, we first simplify its numerator thus

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}\left((X_i-\mu)-(\bar{X}-\mu)\right)^2\\ &= \sum_{i=1}^{n}(X_i-\mu)^2-2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+n\cdot(\bar{X}-\mu)^2\\ &= \sum_{i=1}^{n}(X_i-\mu)^2-n\cdot(\bar{X}-\mu)^2\end{aligned}$$

implying that

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right) &= \sum_{i=1}^{n}\mathrm{Var}(X_i)-n\cdot\mathrm{Var}(\bar{X})\\ &= n\,\sigma^2-n\cdot\frac{\sigma^2}{n}=\sigma^2(n-1)\end{aligned}\qquad(9.2)$$

Later on, we also need

$$\begin{aligned}\mathrm{Cov}(\bar{X},X_1) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}(X_i,X_1)=\frac{1}{n}\mathrm{Cov}(X_1,X_1)+0\\ &= \frac{1}{n}\mathrm{Var}(X_1)=\frac{\sigma^2}{n}\end{aligned}$$

Note that $\mathrm{Cov}(\bar{X},X_2)$, $\mathrm{Cov}(\bar{X},X_3)$, ... must clearly have the same value.

Dividing (9.2) by $n-1$ yields

$$\mathbb{E}(s^2)=\sigma^2$$

Does this imply that $s\equiv\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}$ have the expected value of σ? The answer is 'no' in general, s is usually (slightly) BIASED, meaning that

$$\mathbb{E}(s)\neq\sigma$$

('bias' is the difference between the two).

Note: The *bar notation* can be used for the *sample average* of any function of X_i; for example

$$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$$

can be written as

$$\overline{(X-\bar{X})^2}$$

or (less cryptically) expanded to read

$$\overline{X^2}-\bar{X}^2$$

The last simplification is possible because

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} {X_i}^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2$$
$$= \sum_{i=1}^{n} {X_i}^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2$$

9.3.2 Sampling from $N(\mu, \sigma)$

To be able to say anything more about s^2, we need to know the distribution from which we are sampling. We will thus assume that the distribution is *Normal*, with the mean μ and variance σ^2. This immediately simplifies the distribution of $\bar{X}$, which must also be Normal (with the mean μ and standard deviation of $\frac{\sigma}{\sqrt{n}}$, as we already know) for *any* sample size n (not just 'large').

Proof: Since the MGF of the individiual X_is is

$$M(t) = \exp\left(\frac{\sigma^2 t^2}{2} + \mu \cdot t\right)$$

the MGF of $\bar{X}$ is therefore

$$M\left(\frac{t}{n}\right)^n = \exp\left(\frac{\sigma^2 t^2}{2n^2} + \mu \cdot \frac{t}{n}\right)^n = \exp\left(\frac{\sigma^2 t^2}{2n} + \mu \cdot t\right)$$

which implies that the corresponding distribution is $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$. ■

9.3.3 MGF of s^2

The $(n+1)$-dimensional distribution of $X_i - \bar{X}$ and $\bar{X}$ is jointly Normal, described by the usual parameters (means, variances and covariances); what is important here is that all n covariances between $\bar{X}$ and each of the $X_i - \bar{X}$ are *zero*; this implies that $\bar{X}$ will be independent of any RV built out of the $X_i - \bar{X}$'s such as

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$$

At the same time, we know that this last RV is equal to

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} - \frac{n}{\sigma^2} \cdot (\bar{X} - \mu)^2$$

which implies that

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \frac{n}{\sigma^2} \cdot (\bar{X} - \mu)^2$$

Since the RHS RVs are independent, we can correspoindingly relate the three MGFs (of which we know the first and last) to get:

$$\frac{1}{(1-2t)^{n/2}} = \frac{1}{(1-2t)^{(n-1)/2}} \cdot \frac{1}{(1-2t)^{1/2}}$$

This proves that $\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$ has the χ^2_{n-1} distribution (and is independent of $\bar{X}$).

The **important consequence** of this is that

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

also has the $\mathtt{t}_{n-1}$ distribution, since

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{s^2(n-1)}{\sigma^2}}{n-1}}} \equiv \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

Furthermore, the ratio of two sample *variances* based on two *independent* RISs (of sizes n_1 and n_2 respectively) from the *same* Normal distribution is equal to

$$\frac{s_1^2}{s_2^2} = \frac{\frac{(n_1-1)s_1^2}{(n_1-1)\sigma^2}}{\frac{(n_2-1)s_2^2}{(n_2-1)\sigma^2}} \equiv \frac{\frac{\chi^2_{n_1-1}}{n_1-1}}{\frac{\chi^2_{n_2-1}}{n_2-1}}$$

and therefore has the Fisher F_{n_1-1,n_2-1} distribution.

Example (Student): For a RIS of size 15 from $\mathcal{N}(\mu, \sigma)$, compute

$$\Pr\left(\left|\bar{X} - \mu\right| \le \frac{s}{3}\right)$$

Solution: This equals

$$\Pr\left(\left|\frac{\bar{X}-\mu}{s} \cdot \sqrt{15}\right| \le \sqrt{\tfrac{5}{3}}\right) = \Pr\left(|\mathrm{t}_{14}| \le \sqrt{\tfrac{5}{3}}\right)$$

$$= \frac{\Gamma\left(\frac{15}{2}\right)}{\Gamma\left(\frac{14}{2}\right)\sqrt{14\pi}} \int_{-\sqrt{\frac{5}{3}}}^{\sqrt{\frac{5}{3}}} \left(1 + \tfrac{y^2}{14}\right)^{-\frac{15}{2}} dy = 78.24\%$$

This implies (you learned it in MATH 2P82) that $\bar{X} \pm \frac{s}{3}$ constitutes the so called 78.24% CONFIDENCE INTERVAL for the value of μ (assumed unknown, together with σ). Note that, for an actual sample, both $\bar{X}$ and s would be easily computable *numbers*.

Continuation (chi-square): Similarly, compute

$$\Pr\left(0.9\sigma \leq s \leq 1.1\sigma\right)$$

Solution: This equals

$$\Pr\left(14 \cdot 0.9^2 \leq \tfrac{14s^2}{\sigma^2} \leq 14 \cdot 1.1^2\right) = \Pr\left(\tfrac{1134}{100} \leq \chi^2_{14} \leq \tfrac{1694}{100}\right)$$

$$= \frac{1}{6!2^7} \int_{\frac{1134}{100}}^{\frac{1694}{100}} y^6 \exp(-\tfrac{y}{2}) dy = 39.98\%$$

implying that

$$\frac{s}{1.1} \leq \sigma \leq \frac{s}{0.9}$$

is the corresponding 39.98% confidence interval for the (unknonwn) value of σ.

Continuation (Fisher): Suppose another person takes his own RIS of size 12 from the same distribution; find the probability that the two sample variances will not differ from each other by more than a factor of 1.3 .

Solution:

$$\Pr\left(\frac{1}{1.3} < \frac{s_1^2}{s_2^2} < 1.3\right) = \frac{\Gamma(\frac{14+11}{2})(\frac{14}{11})^7}{\Gamma(\frac{14}{2})\Gamma(\frac{11}{2})} \int_{1/1.3}^{1.3} y^6 (1+\tfrac{14}{11}y)^{-\frac{14+11}{2}} dy = 34.77\%$$

Chapter 10

Order Statistics

In this section we consider a RIS of size n from a *continuous* distribution (not necessarily Normal), calling the individual, independent, would-be observations X_1, X_2, ..., X_n. Based on these, we define a new set of RVs denoted $X_{(1)}$, $X_{(2)}, ...X_{(n)}$ (some textbooks use Y_1, $Y_2, ...Y_n$) to be the *smallest* sample value ($X_{(1)}$), the *second smallest* value ($X_{(2)}$), ..., the *largest* value ($X_{(n)}$). Even though the original X_i's were independent, $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$ are clearly *strongly correlated*. They are called the first, the second, ..., and the last ORDER STATISTIC, respectively. When n is odd, $X_{(\frac{n+1}{2})}$ is also called the sample MEDIAN; notation: $\tilde{X}$.

10.1 Distribution of $X_{(i)}$

It is obvious that $\Pr(X_k \leq x) = F(x)$ for each of the original (not yet arranged from smallest to largest) observations, where $F(x)$ is the distribution function of the *sampled* distribution. Since these events are independent of each other, the probability of exactly j of them happening is equal to (using Binomial pmf)

$$\binom{n}{j} F(x)^j \left(1 - F(x)\right)^{n-j}$$

This implies that the distribution function of $X_{(i)}$ is given by

$$F_{(i)}(x) = \sum_{j=i}^{n} \binom{n}{j} F(x)^j \left(1 - F(x)\right)^{n-j}$$

(to have $X_{(i)} \leq x$ requires *at least* i of the *original*, yet unsorted observations be $\leq x$, right?). This formula then enables us to answer any probability question about $X_{(i)}$. In principle, it should easily yield the corresponding pdf, by simple differentiation. Unfortunately, this results in a rather messy expression which requires extensive simplification (all terms but one cancel out, but in fairly non-trivial manner) before revealing its final form. To find it, it is easier to start

from scratch, utilizing the original, formal definition of a pdf. This leads to

$$\begin{aligned} & f_{(i)}(x) \equiv \lim_{\triangle \to 0} \frac{\Pr(x \le X_{(i)} < x + \triangle)}{\triangle} \\ = & \lim_{\triangle \to 0} \binom{n}{i-1, 1, n-i} F(x)^{i-1} \frac{F(x+\triangle) - F(x)}{\triangle} \left(1 - F(x+\triangle)\right)^{n-i} \\ = & \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1} \left(1 - F(x)\right)^{n-i} f(x) \end{aligned} \tag{10.1}$$

To understand how we computed $\Pr(x \le X_{(i)} < x + \triangle)$, realize that $i-1$ of the original X_j observations must be smaller than x, one must be between x and $x + \triangle$, and the rest must be bigger than x. The resulting pdf has the same support as the original distribution.

We will go over a few examples.

Exponential example Consider a RIS of size 7 from $\mathcal{E}(\beta = 23\,\text{min}.)$; this can be interpreted as seven equally skilled fishermen independently catching *one* fish each, assuming the that long-run average time to catch a fish is 23 minutes.

- Find $\Pr(X_{(3)} < 15\text{ min}.)$, i.e. the probability that the third catch of the group will not take longer than 15 min.

Solution: First find the probability that *any* one of the original 7 *independent* observations is < 15 min.: $\Pr(X_i < 15\text{ min}.) = 1 - e^{-\frac{15}{23}} = 0.479088 \equiv p$. We interpret the same sampling as a *binomial* experiment, where a value smaller than 15 min. defines a *success*, and a value bigger than 15. min. represents a *failure*. The question is: what is the probability of getting *at least* 3 successes (the complement of 'no more than 2')? Using binomial probabilities, we get

$$1 - \left(q^7 + 7pq^6 + \binom{7}{2} p^2 q^5\right) = 73.77\% \quad \blacksquare$$

- Now, find the mean and standard deviation of $X_{(3)}$.

Solution: First we have to construct the corresponding pdf. By (10.1) this equals:

$$\frac{7!}{2!4!}(1 - e^{-\frac{x}{\beta}})^{3-1}(e^{-\frac{x}{\beta}})^{7-3} \cdot \frac{1}{\beta} e^{-\frac{x}{\beta}} = \frac{105}{\beta}(1 - e^{-\frac{x}{\beta}})^2 e^{-\frac{5x}{\beta}}$$

where $\beta = 23$ min. This yields the following mean of $X_{(3)}$:

$$\begin{aligned} & 105 \int_0^\infty x \cdot (1 - e^{-\frac{x}{\beta}})^2 e^{-\frac{5x}{\beta}} \frac{dx}{\beta} \overset{(\frac{x}{\beta} \to u)}{=} 105\beta \int_0^\infty u \cdot (e^{-5u} - 2e^{-6u} + e^{-7u}) du \\ = & \; 105 \cdot 23 \left(\tfrac{1}{5^2} - \tfrac{2}{6^2} + \tfrac{1}{7^2}\right) = 11.72\,\text{min}. \end{aligned}$$

The second simple moment $\mathbb{E}(X_{(3)}^2)$ is similarly

$$105\beta^2 \int_0^\infty u^2 \cdot (e^{-5u} - 2e^{-6u} + e^{-7u}) du = 105 \cdot 23^2 \left(\tfrac{2}{5^3} - \tfrac{4}{6^3} + \tfrac{2}{7^3}\right) = 184.0 \,\text{min}.^2$$

implying that

$$\sigma_{X_{(3)}} = \sqrt{184 - 11.72^2} = 6.830 \,\text{min}.$$

Note that if each fisherman continued fishing (after getting his first, second, ... catch), the distribution of the time of the third catch of the group would be $\mathsf{gamma}(3, \frac{23}{7})$, with the mean of 9.86 min. and $\sigma = 5.69$ min.; naturally, somehow shorter than our previous answer. ■

Note: By a different approach, one can derive the following general formulas (applicable only for sampling from *Exponential* distribution):

$$\begin{aligned} \mathbb{E}(X_{(i)}) &= \beta \sum_{j=0}^{i-1} \frac{1}{n-j} \\ \text{Var}(X_{(i)}) &= \beta^2 \sum_{j=0}^{i-1} \frac{1}{(n-j)^2} \end{aligned}$$

Verify that these give the same answers as our lengthy computation above.

Uniform example: Consider a RIS of size 5 form $\mathcal{U}(0,1)$. Find the mean and standard deviation of $X_{(2)}$.

Solution: The corresponding pdf is

$$\frac{5!}{1!3!} x(1-x)^3 \qquad \text{when} \qquad 0 < x < 1$$

which can be readily identified as $\mathsf{beta}(2,4)$; we have: $X_{(i)} \epsilon\, \mathsf{beta}(i, n+1-i)$ in general. By previous formulas

$$\mathbb{E}(X_{(2)}) = \frac{2}{2+4} = \frac{1}{3}$$

and

$$\text{Var}(X_{(2)}) = \frac{2 \times 4}{(2+4)^2(2+4+1)} = \frac{2}{63}$$

implying that $\sigma_{X_{(2)}} = 0.1782$. ■

Note: These results can be easily extended to sampling from $\mathcal{U}(a,b)$ by utilizing the

$$Y = (b-a)X + a$$

transformation.

10.1.1 Sample median

is one of the most important order statistics, denoted $\tilde{X}$ and equal to (when n is odd, i.e. $n = 2k+1$) $X_{(k+1)}$. This means that k observations are smaller than $\tilde{X}$ and k are bigger than $\tilde{X}$. When n is even (i.e. $n = 2k$) we make

$$\tilde{X} = \frac{X_{(k)} + X_{(k+1)}}{2}$$

To simplify things, we will assume that n is odd from now on.

Let us see what happens to $\tilde{X}$ when n is *large*. One can show that its pdf then (in the $n \to \infty$ limit) becomes *approximately Normal*, with the mean of $\tilde{\mu}$ (the sampled distribution's median) and the standard deviation of

$$\frac{1}{2f(\tilde{\mu})\sqrt{n}}$$

Proof: The sample median $\tilde{X} \equiv X_{(k+1)}$ has the following pdf:

$$\frac{n!}{k! \cdot k!} F(x)^k \left(1 - F(x)\right)^k f(x)$$

where $k \equiv \frac{n-1}{2}$. To explore what happens when $n \to \infty$, we introduce a new RV by

$$Y \equiv 2(\tilde{X} - \tilde{\mu}) f(\tilde{\mu}) \sqrt{n}$$

We build its pdf in the usual three steps:

$$x = \tilde{\mu} + \frac{y}{2f(\tilde{\mu})\sqrt{n}}$$

$$\frac{n!}{k!\cdot k!} \cdot F(\tilde{\mu} + \frac{y}{2f(\tilde{\mu})\sqrt{n}})^k \left(1 - F(\tilde{\mu} + \frac{y}{2f(\tilde{\mu})\sqrt{n}})\right)^k f(\tilde{\mu} + \frac{y}{2f(\tilde{\mu})\sqrt{n}}) \quad (10.2)$$

and multiply this by $\frac{1}{2f(\tilde{\mu})\sqrt{n}}$

To take the limit of the resulting pdf, we utilize the following Taylor expansion:

$$\begin{aligned} F(\tilde{\mu} + \frac{y}{2f(\tilde{\mu})\sqrt{n}}) &\simeq F(\tilde{\mu}) + F'(\tilde{\mu}) \frac{y}{2f(\tilde{\mu})\sqrt{n}} + \frac{F''(\tilde{\mu})}{2} \frac{y^2}{4f(\tilde{\mu})^2 n} + ... = \\ &\frac{1}{2} + \frac{y}{2\sqrt{n}} + \frac{f'(\tilde{\mu})}{2} \frac{y^2}{4f(\tilde{\mu})^2 n} + ... \end{aligned}$$

which implies that

$$1 - F(\tilde{\mu} + \frac{y}{2f(\tilde{\mu})\sqrt{n}}) \simeq \frac{1}{2} - \frac{y}{2\sqrt{n}} - \frac{f'(\tilde{\mu})}{2} \frac{y^2}{4f(\tilde{\mu})^2 n} + ...$$

Multiplying the two results yields

$$F(\tilde{\mu} + \frac{y}{2f(\tilde{\mu})\sqrt{n}}) \left(1 - F(\tilde{\mu} + \frac{y}{2f(\tilde{\mu})\sqrt{n}})\right) \simeq \frac{1}{4} - \frac{y^2}{4n} + ...$$

the dots implying terms proportional to

$$\frac{1}{n^{3/2}}, \frac{1}{n^2}, ...$$

which do not affect the subsequent limit. Substituting into (10.2) and multiplying by the Jacobian results in

$$\frac{n!}{2\cdot 4^k\cdot k!\cdot k!\cdot\sqrt{n}}\cdot\left(1-\frac{y^2}{n}+...\right)^{\frac{n-1}{2}}\cdot\frac{f\left(\tilde{\mu}+\frac{y}{2f(\tilde{\mu})\sqrt{n}}\right)}{f(\tilde{\mu})}$$

Taking the $n\to\infty$ limit leads to (note that the limit of the first term - hard to figure out - must match the limit of the second term):

$$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)$$

This is the pdf of $\mathcal{N}(0,1)$. As a by-product (of no further consequence to us), we have just derived the so called Wallis formula

$$\frac{(2k+1)!}{2^{2k+1}\cdot k!\cdot k!\cdot\sqrt{2k+1}}\underset{k\to\infty}{\longrightarrow}\frac{1}{\sqrt{2\pi}}$$

And, since

$$\tilde{X}=\tilde{\mu}+\frac{Y}{2f(\tilde{\mu})\sqrt{n}}$$

the distribution of the sample median (for a large, fixed n) must be, approximately, $\mathcal{N}(\tilde{\mu},\frac{1}{2f(\tilde{\mu})\sqrt{n}})$. ■

We will now go over a few examples.

Cauchy example: Consider a RIS of size 1001 from $\mathcal{C}(0,1)$. Find $\Pr(-0.1<\tilde{X}<0.1)$, both exactly and using the Normal approximation. Compare with $\Pr(-0.1<\bar{X}<0.1)$.

Solution: Let us recall that, for this Cauchy distribution, we have

$$\begin{aligned} f(x) &= \frac{1}{\pi}\cdot\frac{1}{1+x^2} \\ F(x) &= \frac{1}{2}+\frac{1}{\pi}\arctan(x) \end{aligned}$$

By direct integration of the exact pdf we get

$$\Pr(-0.1<\tilde{X}<0.1)=\frac{1001!}{(500!)^2\pi}\int_{-0.1}^{0.1}\left(\frac{1}{4}-\frac{\arctan(x)^2}{\pi^2}\right)^{500}\frac{dx}{1+x^2}=95.56\%$$

Using the $\tilde{X}\ \tilde{\epsilon}\ \mathcal{N}(0,\frac{\pi}{2\sqrt{1001}})$ approximation (somehow easier - especially without a computer), we get

$$\frac{1}{\sqrt{2\pi}}\int_{-\frac{0.2\sqrt{1001}}{\pi}}^{\frac{0.2\sqrt{1001}}{\pi}}\exp(-\frac{z^2}{2})dz=95.60\%$$

(a decent agreement). For the sample *mean* we get

$$\Pr(-0.1<\bar{X}<0.1)=\frac{1}{\pi}\arctan(x)\Big|_{x=-0.1}^{0.1}=6.35\%$$

only (furthermore, it does not improve with n). Clearly, for a Cauchy distribution, the sample *median* is a lot better way of estimating the central location. Could there be even a *better* way of doing it - we will address this issue later. ■

Normal example: When sampling from $\mathcal{N}(\mu, \sigma)$, is it better to estimate μ by the sample *mean* or by the sample *median*?

Solution: Since

$$\bar{X} \; \epsilon \; \mathcal{N}(\mu, \tfrac{\sigma}{\sqrt{n}})$$

and

$$\tilde{X} \; \tilde{\epsilon} \; \mathcal{N}(\mu, \sqrt{\tfrac{\pi}{2}} \cdot \tfrac{\sigma}{\sqrt{n}})$$

it is obvious that $\tilde{X}$'s standard deviation (in the context of estimation called STANDARD ERROR) is $\sqrt{\frac{\pi}{2}} = 1.253$ times bigger than that of $\bar{X}$. Thus, this time, we are better off using $\bar{X}$. To estimate μ to the same accuracy as $\bar{X}$ does, $\tilde{X}$ would have to use $\frac{\pi}{2} = 1.571$ times bigger sample; the sample mean is, in this case, 57.1% more EFFICIENT than the sample median. ■

Example: Consider a RIS of size 349 from a distribution with $f(x) = 2x$ (when $0 < x < 1$). Find $\Pr(\tilde{X} < 0.75)$, both exactly and using the Normal approximation. Also, approximate $\Pr(\bar{X} < 0.70)$.

Solution: By direct integration of the exact pdf we get

$$\tfrac{2 \cdot 349!}{(174!)^2} \int_0^{0.75} x^{349}(1-x^2)^{174} dx = 99.05\%$$

Based on $F(x) = x^2$, the distribution's median $\tilde{\mu}$ equals to $\frac{1}{\sqrt{2}}$. The Normal approximation thus yields $\Pr(Z < 2.26645) = 98.83\%$ (0.22% off). To deal with the last part of the question we first need

$$\begin{aligned} \mu &= \tfrac{2}{3} \\ \sigma^2 &= \tfrac{1}{18} \end{aligned}$$

Since

$$\bar{X} \; \tilde{\epsilon} \; \mathcal{N}(\tfrac{2}{3}, 0.0126168)$$

$$\Pr(\bar{X} < 0.7) = 99.59$$

This time, evaluating the exact answer would be short of impossible. ■

We mention in passing that $X_{(pn+p)}$, where $0 < p < 1$, tends to the Normal distribution whose mean is the solution to $F(x) = p$, denoted x_p and called the $(100 \cdot p)^{\text{th}}$ PERCENTILE of the sampled distribution, and whose standard deviation is $\frac{\sqrt{p(1-p)}}{f(x_p) \cdot \sqrt{n}}$.

10.2 Bivariate pdf

We now construct the *joint* distribution of *two* order statistics $X_{(i)}$ and $X_{(j)}$ $(i < j)$. By our former definition,

$$f(x,y) = \lim_{\Delta \to 0} \frac{\Pr(x \le X_{(i)} < x + \Delta \cap y \le X_{(j)} < y + \Delta)}{\Delta^2}$$

where x is the value of $X_{(i)}$ and y is the value of $X_{(j)}$. To make the event in parentheses happen, exactly $i-1$ observations must have a value less than x, 1 observation must fall in the $[x, x+\Delta)$ interval, $j-i-1$ observations must be between $x+\Delta$ and y, 1 observation must fall in $[y, y+\Delta)$ interval, and $n-j$ observations must be bigger than $y+\varepsilon$. By our *multinomial* formula, this probability equals to

$$\binom{n}{i-1,1,j-i-1,1,n-j} F(x)^{i-1} \left(F(x+\Delta) - F(x)\right) \left(F(y) - F(x+\Delta)\right)^{j-i-1} \cdot \left(F(y+\Delta) - F(y)\right) \left(1 - F(y+\Delta)\right)^{n-j}$$

Dividing by Δ^2 and taking the $\Delta \to 0$ limit yields

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!} F(x)^{i-1} \left(F(y) - F(x)\right)^{j-i-1} \left(1 - F(y)\right)^{n-j} f(x) f(y) \quad (10.3)$$

with $L < x < y < H$, where L and H is the lower and upper limit (respectively) of the original support.

Example: Consider a RIS of size 8 from a $\mathsf{gamma}(2,3)$ distribution. Find $\mathrm{Cov}(X_{(3)}, X_{(5)})$ and $\Pr(X_{(5)} - X_{(3)} > 2)$.

Solution: For the $\mathsf{gamma}(2,3)$ distribution

$$\begin{aligned} f(x) &= \frac{x \cdot \exp(-\frac{x}{3})}{9} \\ F(x) &= 1 - \left(1 + \tfrac{x}{3}\right) \exp(-\tfrac{x}{3}) \end{aligned}$$

The joint pdf of $X_{(3)}$ and $X_{(5)}$ is therefore given by

$$f_{3,5}(x,y) = \frac{8!}{2 \cdot 3!} F(x)^2 \left(F(y) - F(x)\right) \left(1 - F(y)\right)^3 f(x) f(y) \quad \text{when} \quad 0 < x < y$$

implying

$$\begin{aligned} \mathbb{E}\left(X_{(3)}\right) &\equiv \mu_{(3)} = \int_0^\infty \int_0^y x \cdot f_{3,5}(x,y) dx dy = 3.6163 \\ \mathbb{E}\left(X_{(5)}\right) &\equiv \mu_{(5)} = \int_0^\infty y \int_0^y f_{3,5}(x,y) dx dy = 5.8298 \\ \mathrm{Cov}(X_{(3)}, X_{(5)}) &= \mathbb{E}\left((X_{(3)} - \mu_{(3)}) \cdot (X_{(5)} - \mu_{(5)})\right) \equiv \\ &= \int_0^\infty (y - \mu_{(5)}) \int_0^y (x - \mu_{(3)}) f_{3,5}(x,y) dx dy = 1.532 \\ \Pr(X_{(5)} - X_{(3)} &> 2) = \int_2^\infty \int_0^{y-2} f_{3,5}(x,y) dx dy = 47.86\% \end{aligned}$$

Let us now discuss two important special cases of (10.3).

10.2.1 Two *consecutive* order statistics

namely $X_{(i)}$ and $X_{(i+1)}$ (with values x and y respectively) have therefore the following joint pdf:

$$f(x,y) = \tfrac{n!}{(i-1)!(n-i-1)!} F(x)^{i-1} \left(1-F(y)\right)^{n-i-1} f(x)\, f(y)$$

when $L < x < y < H$.

This reduces to

$$\tfrac{n!}{(i-1)!(n-i-1)!} x^{i-1}(1-y)^{n-i-1} \qquad 0 < x < y < 1 \qquad (10.4)$$

when the *sampled* distribution is $\mathcal{U}(0,1)$.

Uniform example: Based on this, find the distribution of $U = X_{(i+1)} - X_{(i)}$.

Solution: We introduce a second RV $V \equiv X_{(i)}$. Then

$$\begin{aligned} x &= v \\ y &= u + v \end{aligned}$$

Substituting into (10.4) and multiplying by $\left|\frac{dy}{du}\right|$ (equal to 1) yields

$$f(u,v) = \tfrac{n!}{(i-1)!(n-i-1)!} v^{i-1} (1-u-v)^{n-i-1}$$

when $0 < u < 1$ and $0 < v < 1-u$. The marginal pdf of u is thus

$$f_U(u) = \tfrac{n!}{(i-1)!(n-i-1)!} \int_0^{1-u} v^{i-1}(1-u-v)^{n-i-1} dv$$

$$\overset{v=(1-u)z}{=} \tfrac{n!}{(i-1)!(n-i-1)!} (1-u)^{n-1} \int_0^1 z^{i-1}(1-z)^{n-i-1} dz$$

$$= \tfrac{n!}{(i-1)!(n-i-1)!} (1-u)^{n-1} \tfrac{\Gamma(i)\Gamma(n-i)}{\Gamma(n)} = n(1-u)^{n-1} \quad \text{when } 0 < u < 1$$

which is the same for all i values! The value of the last integral follows when recalling the beta distribution.

To see what happens to this distribution in the $n \to \infty$ limit, we must first introduce

$$W \equiv U \cdot n$$

(why?). Then, clearly,

$$f_W(w) = n(1-\tfrac{w}{n})^{n-1} \left|\tfrac{du}{dw}\right| = (1-\tfrac{w}{n})^{n-1} \qquad \text{when } 0 < w < n$$

In the $n \to \infty$ limit this pdf tends to

$$e^{-w} \qquad \text{when} \quad 0 < w$$

which we can identify as $\mathcal{E}(1)$. This is what we have always used for the time between two consecutive arrivals (and now we understand why). ■

10.2.2 Sample range

We start by spelling out the joint pdf of the *first* and *last* order statistics, $X_{(1)}$ and $X_{(n)}$ (associated with x and y respectively); another special case of (10.3):

$$f(x,y) = n(n-1)\left(F(y)-F(x)\right)^{n-2} f(x)f(y) \tag{10.5}$$

when $L < x < y < H$.

Let us now investigate the distribution of the SAMPLE RANGE $U \equiv X_{(n)} - X_{(1)}$. Taking $V \equiv X_{(1)}$, we get

$$\begin{aligned} x &= v \\ y &= u+v \end{aligned}$$

Substituting into (10.5) yields

$$f(u,v) = n(n-1)\left(F(u+v)-F(v)\right)^{n-2} f(v)f(u+v)$$

when $L < v < H$ and $u < H - v$ or, equivalently, $0 < u < H - L$ and $L < v < H - u$. This implies that

$$f(u) = n(n-1)\int_L^{H-u} \left(F(u+v)-F(v)\right)^{n-2} f(v)f(u+v)dv$$

when $0 < u < H - L$. Note that, when $H = \infty$, $H - u$ reduces to ∞.

Uniform example: When we sample $\mathcal{U}(0,1)$, this becomes:

$$f(u) = n(n-1)u^{n-2}\int_0^{1-u} dv = n(n-1)u^{n-2}(1-u)$$

when $0 < u < 1$, which is $\mathsf{beta}(n-1,2)$ with

$$\begin{aligned} \mu_u &= \tfrac{n-1}{n+1} \\ \sigma_u^2 &= \tfrac{2(n-1)}{(n+2)(n+1)^2} \end{aligned}$$

These results can be easily extended to sampling from $\mathcal{U}(a,b)$ - just multiply μ_u by $b-a$ and σ_u^2 by $(b-a)^2$. Note that, for large n, $\sigma_u \approx \frac{\sqrt{2}\cdot(b-a)}{n}$; the standard error of $X_{(n)} - X_{(1)}$ as an *estimator* of the 'population' range $b-a$ thus goes down to zero with $\frac{1}{n}$ (not the usual $\frac{1}{\sqrt{n}}$) - a very efficient way of estimating it! ■

10.2.3 Sample mid-range

When the sampled distribution is $\mathcal{U}(0,1)$, the joint pod of $X_{(1)}$ and $X_{(n)}$ simplifies to

$$f(x,y) = n(n-1)(y-x)^{n-2} \qquad \text{when} \quad 0 < x < y < 1$$

In this case, let us also find the distribution of $U \equiv \frac{X_{(1)}+X_{(n)}}{2}$ (sample MID-RANGE).

Solution: Including $V \equiv X_{(1)}$ implies

$$\begin{aligned} x &= v \\ y &= 2u - v \end{aligned}$$

Substituting and multiplying by 2 (Jacobian) yields

$$f(u,v) = 2n(n-1)\,(2u-2v)^{n-2}$$

when $0 < v < 1$ and $v < u < \frac{v+1}{2}$. To find the U marginal, we have to consider two cases (not uncommon):

$$f(u) = 2^{n-1}n(n-1)\int_0^u (u-v)^{n-2}dv = 2^{n-1}n\,u^{n-1} \quad \text{when } 0 < u < \tfrac{1}{2}$$

and

$$f(u) = 2^{n-1}n(n-1)\int_{2u-1}^u (u-v)^{n-2}dv = 2^{n-1}n\,(1-u)^{n-1} \quad \text{when } \tfrac{1}{2} < u < 1$$

Note that this defines one, not two ordinary (not conditional) univariate pdf, with the following properties:

$$\begin{aligned} \mu_u &= \tfrac{1}{2} \\ \sigma_u^2 &= \tfrac{1}{2(n+2)(n+1)} \end{aligned}$$

(make sure you understand how it is done, based on a two-piece pdf). Also note that the (approximate) variance of $\bar{X}$ is $\frac{1}{12n}$ and that of $\tilde{X}$ equals $\frac{1}{2n}$; for large n, the sample mid-range is thus a much more efficient estimator of the distribution's mid-range than either $\bar{X}$ or $\tilde{X}$ - this can be easily generalized to sampling from $\mathcal{U}(a,b)$, where $\frac{a+b}{2}$ is to be estimated. ■

Example: Consider a RIS of size 1001 from $\mathcal{U}(0,1)$. Compute and compare the following probabilities:

$$\begin{aligned} \Pr(0.499 &< \tfrac{X_{(1)}+X_{(1001)}}{2} < 0.501) = 86.52\% \\ \Pr(0.499 &< \bar{X} < 0.501) \simeq 8.73\% \\ \Pr(0.499 &< \tilde{X} < 0.501) \simeq 5.05\% \end{aligned}$$

This again demonstrates that, for a *uniform* distribution, the sample mid-range is a lot more likely to be close to the true center than either the sample mean or the sample median. ■

In a similar manner we could find the joint *pdf* of three, four, etc. order statistics. We mention in passing that the joint pdf of *all* n order statistics is given by

$$n! \prod_{i=1}^{n} f(x_{(i)}) \qquad \text{when} \qquad L < x_{(1)} < x_{(2)} < ... < x_{(n)} < H$$

Note that this (due to their support) makes them highly correlated.

Also, in passing: one can show that the asymptotic (i.e. in the $n \to \infty$ limit) correlation coefficient between two *sample* percentiles $X_{(p_1 n+p_1)}$ and $X_{(p_2 n+p_2)}$ is equal to

$$\sqrt{\frac{p_1(1-p_2)}{p_2(1-p_1)}}$$

Part II

STATISTICS

Chapter 11

Estimating Distribution Parameters

Until now we have studied PROBABILITY, proceeding as follows: we assumed *parameters* of all distributions to be *known* and, based on this, *computed probabilities* of various outcomes. We now make the essential transition to STATISTICS, which is concerned with the exact opposite: a random experiment is performed (usually many times) and the individual outcomes recorded; based on these, we want to *estimate* values of the distribution parameters (one or more). Until the last two sections, we restrict our attention to the (easier and most common) case of estimating only *one* parameter of a distribution.

Example: Propose to estimate the mean μ of a Normal distribution $\mathcal{N}(\mu, \sigma)$, based on a RIS of size n?

Solution: The sample mean $\bar{X}$ seems to be a 'reasonable' ESTIMATOR of μ. Note that this name applies to the RV $\bar{X}$, before the sampling is done; as soon as the experiment is completed and a particular value of $\bar{X}$ computed, this specific *number* is called an ESTIMATE of μ.

Immediately, a few related issues comes to mind (and needs to be resolved, one by one):

- How do we judge the quality of an estimator - is there a criterion or a set of criteria to help us decide?
- Is it possible to find the *best* estimator of a parameter, at least in some restricted sense?
- Would not it be better to use, instead of a single number (the so called POINT ESTIMATOR, which can never precisely agree with the exact value of the unknown parameter), an *interval* of values, which can give us a good idea about the accuracy of the estimate?

The rest of this section tackles the first two issues. We must start with

11.1 A few definitions

First we allow an ESTIMATOR of a parameter θ to be *any* sample statistic, say $\hat{\Theta}(X_1, X_2, ..., X_n)$. Note that, when specifying $\hat{\Theta}$, we may include (in the actual expression) values of the other parameters, if they are known (also, the sample size n). We concentrate on estimating parameters with real (rather than integer) values.

To narrow down our choices, we will first insist that our estimators be UNBIASED, meaning

$$\mathbb{E}(\hat{\Theta}) = \theta$$

or at least ASYMPTOTICALLY UNBIASED , i.e.

$$\mathbb{E}(\hat{\Theta}) \underset{n\to\infty}{\longrightarrow} \theta$$

The $\mathbb{E}(\hat{\Theta}) - \theta$ difference is called the BIAS of an estimator, and can be easily removed (constructing unbiased estimators is thus not a major challenge).

Example: Propose an estimator for the variance σ^2 (which thus becomes our θ) of a $\mathcal{N}(\mu, \sigma)$ distribution, assuming that the value of μ is also unknown.

Solution: Starting with

$$\hat{\Theta} \equiv \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

we recall that

$$\mathbb{E}(\hat{\Theta}) = \frac{n-1}{n}\sigma^2$$

Our estimator is thus only *asymptotically* unbiased. The bias can be easily removed by defining a new estimator

$$s^2 \equiv \frac{n}{n-1}\hat{\Theta}$$

(the sample variance) which is fully unbiased. Since

$$\frac{n-1}{\sigma^2}s^2 \epsilon\ \chi^2_{n-1}$$

we can also find the variance of s^2 to be

$$\mathrm{Var}(s^2) = \left(\frac{\sigma^2}{n-1}\right)^2 \cdot 2(n-1) = \frac{2\sigma^4}{n-1}$$

Supplementary: Does this imply that s is an unbiased estimator of σ? The answer is NO, as we can see from

$$\mathbb{E}\left(\sqrt{\chi^2_{n-1}}\right) = \frac{1}{\Gamma(\frac{n-1}{2})\,2^{\frac{n-1}{2}}}\int_0^\infty \sqrt{x}\cdot x^{\frac{n-3}{2}}e^{-\frac{x}{2}}\,dx = \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}$$

implying

$$\mathbb{E}(s) = \frac{\sigma}{\sqrt{n-1}} \cdot \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \approx \sigma(1 - \frac{1}{4n} - \frac{7}{32n^2} + ...)$$

But, we already know how to fix this: use

$$\sqrt{\frac{n-1}{2}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} s$$

instead; this is a fully unbiased estimator of σ. ■

Yet, making an estimator unbiased (or at least asymptotically so) is *not enough* to make it even acceptable (let alone 'good'). Consider estimating μ of a distribution by taking $\hat{\Theta} = X_1$ (the first observation only), throwing away the rest of the sample! We get a fully unbiased estimator which is evidently unacceptable, since we are wasting nearly all of the available information. Thus, being unbiased is only *one* essential ingredient of a good estimator, the other one is its *variance*, which we would like to keep as small as possible.

This leads to two new definitions:

Definition: CONSISTENT ESTIMATOR must have two properties:

$$\mathbb{E}(\hat{\Theta}) \underset{n\to\infty}{\longrightarrow} \theta$$

i.e. be asymptotically unbiased, and

$$\mathrm{Var}(\hat{\Theta}) \underset{n\to\infty}{\longrightarrow} 0$$

meaning that its variance must tend to zero with increasing sample size. ■

This implies that we can converge on the *exact* value of θ by indefinitely increasing the sample size. Nice as it sounds, this represents only the *minimal standard* (or even less) to be expected of an estimator - some of them may still be so wasteful to make them inacceptable. For example, discarding every second observation to estimate μ by averaging the remaining observations (i.e. wasting *half* of our sample) still yields a consistent (but rather silly) estimator.

Definition of MVUE: MINIMUM VARIANCE UNBIASED ESTIMATOR is an *unbiased estimator* whose *variance* is smaller or equal to the variance of any other *unbiased* estimator for all potential values of θ (the 'unbiased' requirement is essential: an arbitrary *constant* may be totally nonsensical as an estimator in all but 'lucky-guess' situations, yet no estimator can compete with its variance). ■

Having such an estimator would of course be ideal, but we run into several difficulties:

- The variance of an estimator is, in general, a function of θ, which means that we are now comparing *functions*, not values. It may easily happen that two unbiased estimators have variances such that one is smaller in some range of θ values and bigger in another. Neither estimator is then (uniformly) better than the other, and the MVUE estimator may therefore not exist.
- Even when the MVUE estimator exists, how do we know that it does and, finally,
- how do we find it?

To partially answer the second point: luckily, there is a *theoretical lower bound* on the variance of all unbiased estimators; when an estimator achieves this bound, it is automatically MVUE. The relevant details are contained in the following theorem:

11.2 Cramér-Rao inequality

Consider a parameter θ which does *not* affect the boundaries of the distribution's support (the so-caled REGULAR CASE); as an example of two parameters which are *not* regular, consider $\mathcal{U}(a, b)$.

The variance of any *unbiased* estimator $\hat{\Theta}$ of such a parameter must meet the following inequality:

$$\mathrm{Var}(\hat{\Theta}) \geq \frac{1}{n\mathbb{E}\left[\left(\frac{\partial \ln f(x|\theta)}{\partial\theta}\right)^2\right]} = \frac{1}{-n\mathbb{E}\left(\frac{\partial^2 \ln f(x|\theta)}{\partial\theta^2}\right)} \tag{11.1}$$

where $f(x|\theta)$ stands for the old $f(x)$ - we are now emphasizing its functional dependence on the parameter θ (it does *not* imply a conditional pdf - θ is not a RV).

Proof: Consider a RV X having a general PDF which we θ denote by

$$f(x|\theta)$$

For example, when this distribution is Exponential, we have

$$f(x|\theta) = \frac{\exp(-\frac{x}{\theta})}{\theta}$$

We now transform X into a new RV U, defined by

$$U = \frac{\partial \ln f(X|\theta)}{\partial\theta} = \frac{\frac{\partial f(X|\theta)}{\partial\theta}}{f(X|\theta)}$$

e.g. (Exponential)

$$U = \frac{\partial}{\partial\theta}\left(-\frac{X}{\theta} - \ln\theta\right) = \frac{X}{\theta^2} - \frac{1}{\theta}$$

It is simple to show that

$$\mathbb{E}(U) = \int_L^H \frac{\frac{\partial f(x|\theta)}{\partial \theta}}{f(x|\theta)} \cdot f(x|\theta)\ dx = \int_L^H \frac{\partial f(x|\theta)}{\partial \theta}\ dx = \frac{\partial}{\partial \theta} \int_L^H f(x|\theta)\ dx = 0$$

since

$$\int_L^H f(x|\theta)\ dx = 1$$

for each θ, and interchanging integration and θ-derivative is, in this case, legitimate (the fact that both L and H are *not* allowed to depend on θ is crucial). Similarly

$$\mathrm{Var}(U) = \mathbb{E}\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right] = \int_L^H \left(\frac{\partial \ln f(x|\theta)}{\partial \theta}\right)^2 \cdot f(x|\theta)\ dx$$

By differentiating

$$\int_L^H \frac{\partial \ln f(x|\theta)}{\partial \theta} \cdot f(x|\theta)\ dx = 0$$

(proven three lines ago) with respect to θ we get

$$\begin{aligned} & \int_L^H \frac{\partial^2 \ln f(x|\theta)}{\partial \theta^2} \cdot f(x|\theta)\ dx + \int_L^H \frac{\partial \ln f(x|\theta)}{\partial \theta} \cdot \frac{\partial f(x|\theta)}{\partial \theta}\ dx \\ = & \int_L^H \frac{\partial^2 \ln f(x|\theta)}{\partial \theta^2} \cdot f(x|\theta)\ dx + \int_L^H \left(\frac{\partial \ln f(x|\theta)}{\partial \theta}\right)^2 \cdot f(x|\theta)\ dx = 0 \end{aligned}$$

shows that there is a alternate way of computing the variance of U, namely

$$\mathrm{Var}(U) = -\mathbb{E}\left(\frac{\partial^2 \ln f(X|\theta)}{\partial \theta^2}\right)$$

Out of the two formulas, we can always choose the more convenient one.

And, true enough, for our Exponential example

$$\mathbb{E}\,(U) = \frac{\mathbb{E}(X)}{\theta^2} - \frac{1}{\theta} = \frac{\theta}{\theta^2} - \frac{1}{\theta} = 0$$

while the variance of U equals

$$\mathbb{E}\left[\left(\frac{X}{\theta^2} - \frac{1}{\theta}\right)^2\right] = \frac{\mathbb{E}\left(X^2\right)}{\theta^4} - \frac{2\mathbb{E}\,(X)}{\theta^3} + \frac{1}{\theta^2} = \frac{2\theta^2}{\theta^4} - \frac{2\theta}{\theta^3} + \frac{1}{\theta^2} = \frac{1}{\theta^2}$$

or, using the second formula

$$-\mathbb{E}\left(-\frac{2X}{\theta^3} + \frac{1}{\theta^2}\right) = \frac{2\mathbb{E}\,(X)}{\theta^3} - \frac{1}{\theta^2} = \frac{1}{\theta^2} \qquad \text{(check)}$$

With these preliminaries under our belt, we are now ready for the main part of the proof: When taking a RIS of size n from this distribution, the joint PDF of the individual X_is is simply

$$\prod_{i=1}^{n} f(x_i|\theta)$$

Assuming that $\hat{\Theta}$ is an *unbiased* estimator of θ, which means that it is a function of X_1, $X_2, ...X_n$, but *not* of θ, and also

$$\mathbf{E}\left(\hat{\Theta}\right) = \int_{L}^{H} \cdots \int \hat{\Theta} \cdot \prod_{i=1}^{n} f(x_i|\theta)\ dx_1 dx_2 ... dx_n = \theta$$

(**E** denoting a *multivariate* expected value).

Differentiating the last equation with respect to θ yields

$$\begin{aligned}
& \int_{L}^{H} \cdots \int \hat{\Theta} \cdot \left(\prod_{i=1}^{n} f_i\right)' dx_1 dx_2 ... dx_n = \int_{L}^{H} \cdots \int \hat{\Theta} \cdot \frac{\left(\prod_{i=1}^{n} f_i\right)'}{\prod_{i=1}^{n} f_i} \cdot \prod_{i=1}^{n} f_i\ dx_1 dx_2 ... dx_n \\
= & \int_{L}^{H} \cdots \int \hat{\Theta} \cdot \left(\ln \prod_{i=1}^{n} f_i\right)' \cdot \prod_{i=1}^{n} f_i\ dx_1 dx_2 ... dx_n = \int_{L}^{H} \cdots \int \hat{\Theta} \cdot \left(\sum_{i=1}^{n} \ln f_i\right)' \cdot \prod_{i=1}^{n} f_i\ dx_1 dx_2 ... \\
= & \int_{L}^{H} \cdots \int \hat{\Theta} \cdot \sum_{i=1}^{n} (\ln f_i)' \cdot \prod_{i=1}^{n} f_i\ dx_1 dx_2 ... dx_n = \int_{L}^{H} \cdots \int \hat{\Theta} \cdot \sum_{i=1}^{n} U_i \cdot \prod_{i=1}^{n} f_i\ dx_1 dx_2 ... dx_n \\
= & \mathbf{E}\left(\hat{\Theta} \cdot \sum_{i=1}^{n} U_i\right) = \mathrm{Cov}\left(\hat{\Theta}, \sum_{i=1}^{n} U_i\right) = 1
\end{aligned}$$

where, to simplify the notation, we have replaced $f(x_i|\theta)$ by f_i and have indicated θ-differentiation by a simple prime, i.e.

$$f_i' \equiv \frac{\partial f(x_i|\theta)}{\partial \theta}$$

We know (proved earlier) that in general (for *any* two RVs)

$$\mathrm{Cov}\left(\hat{\Theta}, \sum_{i=1}^{n} U_i\right)^2 \leq \mathrm{Var}(\hat{\Theta}) \cdot \mathrm{Var}\left(\sum_{i=1}^{n} U_i\right) = \mathrm{Var}(\hat{\Theta}) \cdot n\mathrm{Var}\,(U)$$

which, in this particular case yields (11.1). The RHS of this inequality is called the Cramer-Rao variance (CRV) - either version will do, as the results are the same.

Note that this proof holds in the case of a *discrete* distribution as well (each *integration* then has to be replaced by the corresponding *summation*).

For the Exponential distribution the CRV equals to $\frac{\theta^2}{n}$ (no unbiased estimator of θ can have a better variance than this, whatever the value of θ is). Since the expected value of $\bar{X}$ is θ and its variance equals to $\frac{\theta^2}{n}$, the sample mean is the MVUE of θ (in the particular case of Exponential distribution - not necessarily true for any other distribution). ■

Based on this C-R bound we define the so called EFFICIENCY of an *unbiased* estimator $\hat{\Theta}$ as the ratio of the theoretically (perhaps) achievable CRV to the *actual* variance of $\hat{\Theta}$, thus:

$$\frac{\mathsf{CRV}}{\mathrm{Var}(\hat{\Theta})}$$

usually expressed in percent. An estimator whose variance is as small as CRV is called EFFICIENT (note that, from what we know already, this makes it automatically the MVUE estimator of θ). An estimator which reaches 100% efficiency only in the $n \to \infty$ limit is called ASYMPTOTICALLY EFFICIENT (we will take it - it is usually the best estimator one can find; besides, our samples are usually large).

One can also define RELATIVE EFFICIENCY of one estimator, say $\hat{\Theta}_1$ with respect to another, say $\hat{\Theta}_2$, as

$$\frac{\mathrm{Var}(\hat{\Theta}_2)}{\mathrm{Var}(\hat{\Theta}_1)}$$

Normal example: Is $\bar{X}$ the best way to estimate μ of the Normal distribution $\mathcal{N}(\mu, \sigma)$.

Solution: We know that its variance is $\frac{\sigma^2}{n}$. To compute C-R bound, we do

$$\frac{\partial^2}{\partial \mu^2}\left[-\ln(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2}\right] = -\frac{1}{\sigma^2}$$

Thus, CRV equals $\frac{1}{\frac{n}{\sigma^2}} = \frac{\sigma^2}{n}$ implying that $\bar{X}$ is MVUE of μ. ■

Bernoulli example: Suppose we want to estimate p of a Bernoulli distribution by the proportion of successes we get in n independent trials (effectively, the sample mean of the 0 or 1 observations), thus:

$$\hat{\Theta} = \frac{\sum_{i=1}^{n} X_i}{n}$$

The mean of our estimator is $\frac{np}{n} = p$ (unbiased), its variance equals $\frac{npq}{n^2} = \frac{pq}{n}$, since $\sum_{i=1}^{n} X_i$ has the $\mathcal{B}(n,p)$ distribution. Is this the best we can do?

Solution: Let us compute the corresponding CRV (in two steps):

$$\frac{\partial^2}{\partial p^2}\left[x\ln p+(1-x)\ln p\right]=-\frac{x}{p^2}-\frac{1-x}{(1-p)^2}$$

followed by

$$\mathbb{E}\left(\frac{X}{p^2}+\frac{1-X}{(1-p)^2}\right)=\frac{1}{p}+\frac{1}{1-p}=\frac{1}{pq}$$

which implies that CRV$=\frac{pq}{n}$. So, the answer is yes, our estimator is MVUE. ■

Poisson example: Similarly, is $\bar{X}$ MVUE of Λ of the Poisson distribution?

Solution:

$$\begin{aligned}\frac{\partial^2}{\partial\Lambda^2}\left[x\ln\Lambda-\ln(x!)-\Lambda\right] &= -\frac{x}{\Lambda^2}\\ \mathbb{E}\left[\frac{X}{\Lambda^2}\right] &= \frac{1}{\Lambda}\end{aligned}$$

which implies that CRV$=\frac{\Lambda}{n}$. Since $\mathbb{E}(\bar{X})=\frac{n\Lambda}{n}=\Lambda$ and $\mathrm{Var}(\bar{X})=\frac{n\Lambda}{n^2}=\frac{\Lambda}{n}$, the answer is YES. ■

Uniform example: Let us try estimating θ of the uniform distribution $\mathcal{U}(0,\theta)$. This is *not* a regular case, so we don't have CRV and the concept of (absolute) efficiency. All we can do is to compute *relative efficiency* of two unbiased estimators, say $2\bar{X}$ and $\frac{n+1}{n}X_{(n)}$.

Solution: For the former, we get $\mathbb{E}(2\bar{X})=\theta$ (check) and $\mathrm{Var}(2\bar{X})=\frac{\theta^2}{3n}$. As for the latter, we realize that

$$\frac{X_{(n)}}{\theta}\epsilon\ \mathsf{beta}(n,1)$$

which implies that

$$\mathbb{E}(\frac{n+1}{n}X_{(n)})=\frac{n+1}{n}\cdot\frac{n}{n+1}\cdot\theta=\theta\qquad\text{(check)}$$

and

$$\mathrm{Var}(\frac{n+1}{n}X_{(n)})=\frac{\theta^2}{(n+2)n}$$

Its relative efficiency with respect to $2\bar{X}$ is therefore $\frac{n+2}{3}$ i.e., in the large-sample limit, $\frac{n+1}{n}X_{(n)}$ is 'infinitely' more efficient than $2\bar{X}$. But how can we establish whether $\frac{n+1}{n}X_{(n)}$ is the 'best' unbiased estimator, lacking the C-R bound? Obviously, something else is needed in a case like this. ■

What will help us deal with non-regular cases (such as the previous example) is the concept of

11.3 Sufficiency

which not only guarantees that a sufficient estimator is MVUE, but (unlike the C-R bound) actually helps us *find* it. Furthemore, the concept of sufficiency can be used even in cases which are not regular (e.g. it applies to both parameters of the uniform distribution). The problem is that, for some distributions (such as Cauchy), sufficient estimators do not exist (this is something we discover only when we try to find them - it is hard to tell in advance). But, on the plus side, when they do exist, finding them is quite easy.

Definition: A sample statistic $\hat{\Phi}(X_1, X_2, ...X_n)$ is called a SUFFICIENT STATISTIC (not an estimator yet) for estimating θ when the joint pdf of the sample can be written as a product of a function of θ and $\hat{\Phi}$ only (no other x_is), and of another function of the x_is (but no θ), i.e.

$$\prod_{i=1}^{n} f(x_i|\theta) = g(\theta, \hat{\Phi}) \cdot h(x_1, x_2, ...x_n)$$

where $g(\theta, \hat{\Phi})$ thus takes care of the joint pdf's θ dependence, *including* support boundaries.

Equivalently, and in many cases more easily, we can similarly 'split' the *logarithm* of the joint pdf, into a *sum* of two such functions, namely

$$\sum_{i=1}^{n} \ln f(x_i|\theta) = \tilde{g}(\theta, \hat{\Phi}) + \tilde{h}(x_1, x_2, ...x_n)$$

Such an $\hat{\Phi}$ (if it exists) contains all the information relevant for estimating θ. All we have to do to convert $\hat{\Phi}$ into the best possible *estimator* of θ is to make it *unbiased* (by some transformation, which is usually easy to design). One can then show that the resulting estimator is MVUE even if it does not reach the C-R limit (it will reach it *asymptotically*, i.e. in the $n \to \infty$ limit).

Let us go over a few examples.

Bernoulli example: Since the sample's joint pdf is

$$p^{x_1+x_2+...+x_n} (1-p)^{n-x_1-x_2-...-x_n}$$

is itself a function of p and of a *single combination* of the x_is, namely $\sum_{i=1}^{n} x_i$, this last sum (after replacing x_i by X_i) is a sufficient statistic for estimating p (dividing it by n makes it into an unbiased estimator). ■

Normal example: Sampling from $N(\mu, \sigma)$, ln of the joint *pdf* is

$$-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - n \ln\left(\sqrt{2\pi}\sigma\right)$$

$$\frac{2\mu \sum_{i=1}^{n} x_i - n\ \mu^2}{2\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} {x_i}^2 - n \ln\left(\sqrt{2\pi}\sigma\right)$$

where the first term is a function of only a single combination of the x_is, namely their sum, and the μ parameter (the second parameter σ is to be treated as a known constant); the rest of the expression has no μ in it. Thus, the sum of the X_is is a sufficient statistics for estimating μ, and $\bar{X}$ is the corresponding unbiased estimator. ■

Exponential example: Since

$$\prod_{i=1}^{n} f(x_i|\beta) = \frac{1}{\beta^n} \exp\left(-\frac{1}{\beta}\sum_{i=1}^{n} x_i\right)$$

we reach the same conclusion as in the previous two cases (when β is the parameter to be estimated. To make this example a bit more challenging, suppose that instead of β, we want to estimate $\theta \equiv \frac{1}{\beta}$ (the arrival *rate*); $\sum_{i=1}^{n} X_i \epsilon$ $\mathsf{gamma}(n, \frac{1}{\theta})$ is still a sufficient statistics, but making it unbiased is now a bit more challenging. Since

$$\mathbb{E}\left(\frac{1}{\sum_{i=1}^{n} X_i}\right) = \frac{\theta^n}{(n-1)!}\int_0^\infty \frac{1}{u}\cdot u^{n-1}e^{-\theta\, u}du = \frac{\theta}{n-1}$$

$$\hat{\Theta} = \frac{n-1}{\sum_{i=1}^{n} X_i}$$

makes the sum into an unbiased estimator of θ. Its variance can be found (by a similar integration) to be $\frac{\theta^2}{n-2}$, whereas the corresponding CRV is $\frac{\theta^2}{n}$. The efficiency of this MVUE is thus $\frac{n-2}{n}$ (less than 100%) - the estimator is efficient only asymptotically. ■

Gamma example: Since

$$\prod_{i=1}^{n} f(x_i|\beta) = \frac{\exp\left(-\frac{1}{\beta}\sum_{i=1}^{n} x_i\right)}{\beta^{\alpha\ n}} \cdot \frac{\left(\prod_{i=1}^{n} x_i\right)^{\alpha-1}}{\Gamma(\alpha)^n}$$

$\sum_{i=1}^{n} X_i$ a sufficient statistics for estimating β. Similarly, $\prod_{i=1}^{n} X_i$ would be a sufficient statistics for estimating α (the two are then *jointly sufficient* for estimating α and β, but we will not go into that). Since

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = n\ \alpha\ \beta$$

$\frac{\sum_{i=1}^{n} X_i}{n\ \alpha}$ is the corresponding unbiased estimator (to estimate β in this manner can thus be done only when the value of α is known). Its variance equals to $\frac{\beta^2}{n\ \alpha}$, which agrees with the C-R bound, making it 100% efficient. ■

Uniform example: Show that $X_{(n)}$ is a sufficient statistic for estimating θ of the uniform $\mathcal{U}(0, \theta)$ distribution.

Solution: First introduce

$$G_{a,b}(x) \equiv \begin{cases} 1 & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

(note that it needs to be seen as a function of three arguments). The joint pdf of X_1, X_2, ..., X_n can then be written as

$$\frac{1}{\theta^n} \prod_{i=1}^{n} G_{0,\theta}(x_i) = \frac{G_{0,\theta}(x_{(n)})}{\theta^n} \cdot G_{0,x_{(n)}}(x_{(1)})$$

where the first factor is a function of θ and $x_{(n)}$, and the second factor is θ-free. Knowing that

$$\mathbb{E}(X_{(n)}) = \frac{n}{n+1}\theta$$

we can easily see that $\frac{n+1}{n} X_{(n)}$ is an *unbiased* and therefore MVUE of θ (since there is no CRV, the concept of efficiency does not apply here). ■

The only difficulty with the approach of this section arises when a sufficient statistic does *not* exist (e.g. the case of Cauchy distribution). One can then resort to using one of the following two techniques for finding an estimator of a parameter (or joint estimators of two or more parameters). The second of these is guaranteed to provide MVUE, at least asymptotically; this remains true even when having to estimate two or more parameters.

11.4 Method of moments

is the simpler of the two; it can find adequate (often MVUE) estimators in many cases, but when it fails, it fails rather badly. It should be considered old-fashioned, if not obsolete. That is why we will be as brief as possible.

11.4.1 One-parameter estimation

In this case, the method provides a very simple prescription: since $\mathbb{E}(X)$ must be a (usually simple) function, say $g(\theta)$, of the parameter (let us call it θ), we make $g(\theta)$ equal to $\bar{X}$ and solve for the estimator $\hat{\Theta}$, thus:

$$\hat{\Theta} \equiv g^{-1}(\bar{X})$$

The method clearly fails when $\mathbb{E}(X)$ does not exist (such as Cauchy).

Example: When sampling from a distribution with

$$f(x) = \frac{2x}{a} e^{-\frac{x^2}{a}} \qquad \text{when} \quad x > 0$$

use this technique to estimate $a > 0$.

Solution: Since

$$\mathbb{E}(X) = \int_0^\infty \frac{2x^2}{a} e^{-\frac{x^2}{a}} dx \stackrel{u=\frac{x^2}{a}}{=} \int_0^\infty \sqrt{au} e^{-u} du = \sqrt{a}\Gamma(\tfrac{3}{2}) = \frac{\sqrt{a\pi}}{2}$$

we get

$$\hat{a} = \frac{4\bar{X}^2}{\pi}$$

Similarly,

$$\mathbb{E}(X^2) = \int_0^\infty \frac{2x^3}{a} e^{-\frac{x^2}{a}} dx \stackrel{u=\frac{x^2}{a}}{=} a \int_0^\infty u\ e^{-u} du = a$$

which enables us to compute

$$\mathbb{E}\left(\bar{X}^2\right) = \frac{n\mathbb{E}\left(X_1^2\right) + n(n-1)\mathbb{E}\left(X_1 \cdot X_2\right)}{n^2} = \frac{a}{n} + \frac{n-1}{n} \cdot \frac{a\pi}{4}$$

implying that

$$\mathbb{E}\left(\hat{a}\right) = a + \frac{a}{n}(\frac{4}{\pi} - 1)$$

This shows that our estimator is unbiased only asymptotically (making it fully unbiased would be easy). Its variance can be computed with the help of (to investigate *asymptotic* efficiency only, these would not even be needed, as we will see shortly):

$$\begin{aligned}
\mathbb{E}(X^3) &= \int_0^\infty \frac{2x^4}{a} e^{-\frac{x^2}{a}} dx \stackrel{u=\frac{x^2}{a}}{=} a^{\frac{3}{2}} \int_0^\infty u^{\frac{3}{2}} e^{-u} du = \frac{3a^{\frac{3}{2}}\sqrt{\pi}}{4} \\
\mathbb{E}(X^4) &= \int_0^\infty \frac{2x^5}{a} e^{-\frac{x^2}{a}} dx \stackrel{u=\frac{x^2}{a}}{=} a^2 \int_0^\infty u^2 e^{-u} du = 2a^2
\end{aligned}$$

which imply

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^4\right] = \binom{4}{4} n\mathbb{E}(X_1^4) + \binom{4}{3,1} n(n-1)\mathbb{E}(X_1^3 X_2) + \binom{4}{2}\binom{n}{2}\mathbb{E}(X_1^2 X_2^2) \\
&+ \binom{4}{2,1,1} n\binom{n-1}{2}\mathbb{E}(X_1^2 X_2 X_3) + \binom{4}{1,1,1,1}\binom{n}{4}\mathbb{E}(X_1 X_2 X_3 X_4) \\
= \ & a^2 \frac{\pi^2 n^4 + 6\pi(4-\pi)n^3 + (48 - 48\pi + 11\pi^2)n^2 - 2(8 - 12\pi + 3\pi^2)n}{16} \\
\simeq \ & a^2 \frac{\pi^2 n^4 + 6\pi(4-\pi)n^3 + ...}{16}
\end{aligned}$$

Note that we have discarded terms which do not affect asymptotic efficiency; to get the approximate answer, only the $\mathbb{E}(X_1^2 X_2 X_3)$ and $\mathbb{E}(X_1 X_2 X_3 X_4)$ part of the original expansion would have been required.

This results in

$$\begin{aligned}
\text{Var}\,(\hat{a}) &= \left(\frac{4}{\pi}\right)^2 \left(\frac{\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^4\right]}{n^4} - \mathbb{E}\left(\bar{X}^2\right)^2 \right) \\
&= \left(\frac{4}{\pi}\right)^2 \left(a^2 \frac{\pi^2 n^4 + 6\pi(4-\pi)n^3}{16n^4} - a^2 \frac{(\pi n + 4 - \pi)^2}{16n^2} \right) \\
&\simeq \frac{4a^2(4-\pi)}{\pi n} + O(\frac{1}{n^2})
\end{aligned}$$

Since

$$\frac{\partial^2 \ln f(x)}{\partial a^2} = \frac{a - 2x^2}{a^3}$$

the corresponding CRV is

$$\frac{a^2}{n}$$

The asymptotic efficiency of our estimator is therefore equal to

$$\frac{\pi}{4(4-\pi)} = 91.49\% \qquad \blacksquare$$

11.4.2 Estimating two parameters

When having to estimate two parameters (say θ_1 and θ_2), we do something similar (this time using the sample mean *and* variance), namely:

$$\begin{aligned}
\mathbb{E}(X) &= g(\theta_1, \theta_2) \equiv \bar{X} \\
\text{Var}(X) &= h(\theta_1, \theta_2) \equiv s^2
\end{aligned}$$

and solving for θ_1 and θ_2 (expressing each in terms of $\bar{X}$ and s^2) - these are then the corresponding estimators.

Uniform example: Based on a RIS of size n from $\mathcal{U}(a, b)$, estimate both a and b.

Solution: Since

$$\begin{aligned}
\mathbb{E}(X) &= \tfrac{a+b}{2} \equiv \bar{X} \\
\text{Var}(X) &= \tfrac{(b-a)}{12}^2 \equiv s^2
\end{aligned}$$

($\equiv$ is to be read: 'make it equal to'), solving for a and b yields

$$\begin{aligned}
\hat{a} &= \bar{X} + \sqrt{3s^2} \\
\hat{b} &= \bar{X} - \sqrt{3s^2}
\end{aligned}$$

This proves to be a very **in**efficient way of estimating a and b, since the standard error of both estimators is proportional to $\frac{1}{\sqrt{n}}$ (we will not bother to compute them), compared to ML estimators (see the next section) whose standard error decreases with $\frac{1}{n}$, as the sample size increases. ■

Beta example: Similarly, sampling from a $\mathsf{beta}(k, m)$ distribution, estimate both n and m.

Solution:

$$\begin{aligned} \mathbb{E}(X) &= \tfrac{k}{n+m} \equiv \bar{X} \\ \mathrm{Var}(X) &= \tfrac{k\ m}{(k+m)^2(k+m+1)} \equiv s^2 \end{aligned}$$

imply

$$\begin{aligned} \hat{k} &= \bar{X} \cdot \left(\tfrac{\bar{X}(1-\bar{X})}{s^2} - 1 \right) \\ \hat{m} &= (1-\bar{X}) \cdot \left(\tfrac{\bar{X}(1-\bar{X})}{s^2} - 1 \right) \end{aligned}$$

It would be rather difficult to investigate properties (such as bias, etc.) of these estimators, but they are certainly not expected to be very efficient. ■

Estimating three or more parameters, one would have to move to higher moments of the sampled distribution.

11.5 Maximum-likelihood technique

always performs very well; in general, ML estimators are as good or better than those found by other techniques, even though they may be only *asymptotically* unbiased (but, as we already know, removing bias is usually not too difficult). The only problem is that, in some cases, they can be found only by solving, *numerically*, one or more potentially non-linear equations (i.e. they may not have the form of a neat analytical expression); in the age of computers, this is not a major issue.

The way to find MLEs of parameters of a distribution is rather simple (at least in principle):

In the joint pdf of X_1, X_2, ..., X_n, i.e. in

$$\prod_{i=1}^{n} f(x_i|\theta_1, \theta_2, ...) \tag{11.2}$$

replace each x_i by the actually *observed* value of X_i, thus obtaining the so called LIKELIHOOD FUNCTION. Note that (unlike the original pdf), this is a function of the *parameters* only (the x_is have become fixed numbers).

Then, one has to *maximize* this likelihood function with respect to θ_1, θ_2, ...; the θ-values which achieve the maximum value of (11.2) are the respective ML estimates. Note that it is frequently easier (but equivalent) to maximize the ln of the likelihood function instead.

11.5.1 One-parameter examples

Exponential: Sampling from $\mathcal{E}(\beta)$, find the MLE of β.

Solution: We have to maximize

$$-n \ln \beta - \frac{\sum_{i=1}^{n} X_i}{\beta}$$

with respect to β. Making the corresponding first derivative equal to zero yields:

$$-\frac{n}{\beta} + \frac{\sum_{i=1}^{n} X_i}{\beta^2} = 0$$

implying that $\hat{\beta} = \bar{X}$. We know that the exact distribution of this estimator is $\mathsf{gamma}(n, \frac{\beta}{n})$. ■

Uniform: Sampling from $\mathcal{U}(0, \theta)$, find the MLE of θ.

Solution: We have to maximize

$$\frac{1}{\theta^n} G_{0,\theta}(X_{(n)}) \cdot G_{0,X_{(n)}}(X_{(1)})$$

with respect to θ; this can be achieved by choosing the smallest possible value for θ while keeping $G_{0,\theta}(X_{(n)}) = 1$, leading to $\hat{\theta} = X_{(n)}$. We know that the exact distribution of $\frac{\hat{\theta}}{\theta}$ is $\mathsf{beta}(n, 1)$. ■

Geometric: Sampling from $\mathcal{G}(p)$, find the MLE of p.

Solution: Maximize

$$n \ln p + \left(\sum_{i=1}^{n} X_i - n \right) \ln(1 - p)$$

by solving

$$\frac{n}{p} - \frac{\sum_{i=1}^{n} X_i - n}{1 - p} = 0$$

which yields

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} X_i} = (\bar{X})^{-1}$$

We know that the exact distribution of $\sum_{i=1}^{n} X_i$ is $\mathcal{NB}(n, p)$, which would enable us to investigate basic properties of this estimator. ■

Reyleigh: Sampling from a distribution with

$$f(x) = \frac{2x}{a} e^{-\frac{x^2}{a}} \qquad \text{when} \quad x > 0$$

find the MLE of $a > 0$.

Solution: Maximize

$$n \ln 2 - n \ln a + \ln \prod_{i=1}^{n} X_i - \frac{\sum_{i=1}^{n} X_i^2}{a}$$

by solving

$$-\frac{n}{a} + \frac{\sum_{i=1}^{n} X_i^2}{a^2} 0$$

which implies that

$$\hat{a} = \frac{\sum_{i=1}^{n} X_i^2}{n} = \overline{X^2}$$

(unbiased estimator - done earlier). Based on

$$\frac{\partial^2}{\partial a^2} \left(\ln(2X) - \ln a - \frac{X^2}{a} \right) = \frac{1}{a^2} - 2\frac{X^2}{a^3}$$

whose expected value is $-\frac{1}{a^2}$, the C-R bound is $\frac{a^2}{n}$. Since

$$\mathrm{Var}(\overline{X^2}) = \frac{\mathrm{Var}(X^2)}{n} = \frac{\mathbb{E}(X^4) - a^2}{n} = \frac{a^2}{n}$$

the MLE is MVUE. ■

Normal: Sampling from $\mathcal{N}(\mu, \sigma)$, find the MLE of σ^2 assuming that μ is known.

Solution: Maximize

$$\frac{n}{2} \ln(2\pi) - n \ln \sigma - \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2\sigma^2}$$

by solving (note that we can differentiate with respect to σ rather than σ^2)

$$-\frac{n}{\sigma} + \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^3} = 0$$

which implies that

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}$$

Based on

$$\frac{\partial^2}{(\partial \sigma^2)^2} \left(-\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \sigma^2 - \frac{(x-\mu)^2}{2\sigma^2} \right) = \frac{1}{2\sigma^2} - \frac{(x-\mu)^2}{\sigma^6}$$

(this time we do have to differentiate with respect to σ^2), whose expected value is $-\frac{1}{2\sigma^4}$, CRV equals to $\frac{2\sigma^4}{n}$. Since

$$\begin{aligned} \mathrm{Var}(\hat{\sigma}^2) &= \mathbb{E}\left[\left(\frac{\sum_{i=1}^{n} [(X_i - \mu)^2 - \sigma^2]^2}{n} \right) \right] = \frac{\mathbb{E}[(X-\mu)^4] - 2\mathbb{E}[(X-\mu)^2]\sigma^2 + \sigma^4}{n} \\ &= \frac{3\sigma^4 - 2\sigma^4 + \sigma^4}{n} = \frac{2\sigma^4}{n} \end{aligned}$$

the MLE is MVUE. ■

Gamma (estimating β): Sampling from gamma(α, β), find the MLE of β (assuming that α is known).

Solution: Maximize

$$(\alpha - 1) \ln \prod_{i=1}^{n} X_i - \frac{\sum_{i=1}^{n} X_i}{\beta} - n \ln \Gamma(\alpha) - n \ \alpha \ln \beta \tag{11.3}$$

by solving

$$\frac{\sum_{i=1}^{n} X_i}{\beta^2} - \frac{n \ \alpha}{\beta} = 0$$

for β, getting $\widehat{\beta} = \dfrac{\bar{X}}{\alpha}$. We know that the exact distribution of this estimator is gamma$(n\alpha, \frac{\beta}{n\alpha})$. ■

Gamma (estimating α): Sampling the gamma(α, β) distribution, find the MLE of α (assuming that β is known).

Solution: Maximize (11.3) by solving

$$\sum_{i=1}^{n} \ln X_i - n \, \psi(\alpha) - n \ln \beta = 0$$

where $\psi(\alpha)$ is the so called Euler's PSI function (also called digamma function), defined as the α derivative of $\ln \Gamma(\alpha)$. This implies that

$$\psi(\hat{\alpha}) = \overline{\ln \frac{X}{\beta}}$$

which can be easily and uniquely solved for $\hat{\alpha} > 0$ (graphically or numerically). Thus, for example, based on the following RIS of size fifteen: $3.82, 4.26, 13.98, 8.06, 4.47, 1.53, 7.70, 1.62, 7.93, 4.93, 3.91, 5.75, 10.99, 6.04, 5.91$ and knowing that $\beta = 1.5$, we get

$$\overline{\ln \frac{X}{\beta}} = 1.242$$

and, consequently, $\hat{\alpha} = 3.95$. The asymptotic variance (equal to CRV) of this estimator is $\frac{1}{n \cdot \psi'(\alpha)}$; this enables us to estimate the standard error of our estimate as $\sqrt{\frac{1}{15 \cdot \psi'(3.95)}} = 0.48$. ■

Cauchy (estimating median): Sampling $\mathcal{C}(a, 1)$, find the MLE of a.

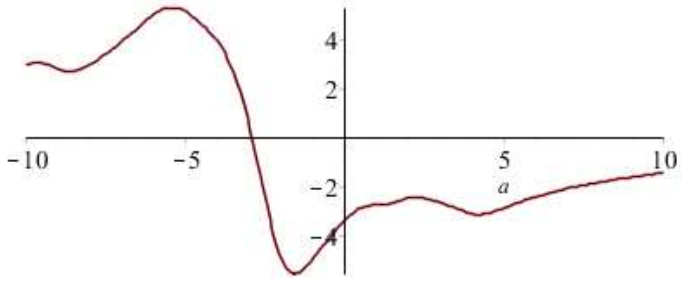
Solution: Maximize

$$-n \ln \pi - \sum_{i=1}^{n} \ln \left(1 + (X_i - a)^2\right)$$

by solving

$$\sum_{i=1}^{n} \frac{X_i - a}{1 + (X_i - a)^2} = 0$$

Unfortunately, in this case there is no *analytic* solution for a. But, as soon as a RIS is taken, the individual X_is become simple numbers, and the equation can be easily solved *numerically*. Thus, for example, when $n = 20$ and the corresponding RIS results in 3.67, 0.69, −9.11, −2.07, −3.89, −3.25, −2.85, 2.60, −3.27, 17.09, −3.72, −2.27, −2.42, −2.88, −2.28, −2.42, −4.60, −5.37, −4.68, −2.80, one can easily plot the LHS of the last equation (as a function of a) to see that it intersects the a axis at −2.91 (to get it accurately enough, one may have to do a bit of zooming in) which thus becomes the corresponding ML ESTIMATE:



To find the standard error of the answer, we utilize the fact that any MLE must be (at least asymptotically) 100% efficient. Based on

$$\mathbb{E}\left[\left(\frac{\partial \ln f}{\partial a}\right)^2\right] = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{4(x-a)^2\,dx}{\left(1+(x-a)^2\right)^3} = \frac{1}{2}$$

the CRV is equal to $\frac{2}{n}$; the standard error of our estimate is thus $\sqrt{0.1} = 0.32$.

In this context, we recall that the asymptotic variance of the *sample median* $\tilde{X}$ (which in this case has the value of −2.825) equals to $\frac{\pi^2}{4n}$, its efficiency is thus only $\frac{8}{\pi^2} = 81.06\%$ (the corresponding standard error is 0.35). But $\tilde{X}$ has one big advantage: unlike the ML estimator, it does *not* need to know the value of the second parameter! Furthermore, $\tilde{X}$ (unlike the MLE) remains a sensible estimator of distribution's median even when the distribution is not quite Cauchy (after all, we may be uncertain not only about the exact values of various parameters, but about the shape and nature of the distribution itself). Estimators of this kind (insensitive to a breakdown of some of our assumptions) are called ROBUST. ■

11.5.2 Two-parameter examples

Normal: Sampling from $\mathcal{N}(\mu, \sigma)$, find the MLEs of both μ and σ.

Solution: Maximize

$$\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}$$

by setting both derivatives equal to zero, i.e.

$$\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma^2}=0$$

and

$$-\frac{n}{\sigma}+\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^3}=0$$

The first one of these clearly yields $\hat{\mu}=\bar{X}$, the second one (after substituting $\bar{X}$ for μ) yields

$$\hat{\sigma}=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}}$$

The properties of the two estimators follow from what we know about the joint distribution of $\bar{X}$ and s^2.

Uniform: Sampling $\mathcal{U}(a,b)$, find the MLEs of a and b.

Solution: Maximize

$$\frac{1}{(b-a)^n}\prod_{i=1}^{n}G_{a,b}(X_i)=\frac{G_{a,b}(X_{(n)})G_{a,X_{(n)}}(X_{(1)})}{(b-a)^n}$$

by choosing a and b as close to each other as the G-functions allow (before dropping to zero). Obviously, a cannot be any bigger that $X_{(1)}$ and b cannot be any smaller than $X_{(n)}$, so these are the corresponding MLEs. Based on the 'Order Statistics' chapter, we know how to find their exact joint distribution. ■

Gamma: Sampling from $\mathsf{gamma}(\alpha,\beta)$, find the MLEs of both parameters.

Solution: Maximize (11.3). The two derivatives result in the following two equations:

$$\sum_{i=1}^{n}\ln X_i-n\,\psi(\alpha)-n\ln\beta=0$$

and

$$\frac{\sum_{i=1}^{n}X_i}{\beta^2}-\frac{n\;\alpha}{\beta}=0$$

Based on the second one, we get

$$\widehat{\beta}=\frac{\bar{X}}{\hat{\alpha}}$$

The first equation can be then re-written as follows

$$\ln\hat{\alpha}-\psi(\hat{\alpha})=\ln\bar{X}-\overline{\ln X}$$

and needs to be solved numerically for $\hat{\alpha}$ (there is always a unique solution). Using the RIS of size 15 from the previous section (where we were estimating the value of α only), we now get

$$\ln\bar{X}-\overline{\ln X}=0.1543$$

which leads to $\hat{\alpha} = 3.40$ and $\widehat{\beta} = \frac{6.06}{3.40} = 1.78$. One can show that their *approximate* joint distribution is bivariate Normal, with the two means given by α and β respectively (i.e. the exact, albeit still unknown values of the two parameters), and the variance-covariance matrix equal to

$$\begin{aligned} & \frac{1}{-n} \cdot \begin{bmatrix} \mathbb{E}\left(\frac{\partial^2 \ln f}{\partial a^2}\right) & \mathbb{E}\left(\frac{\partial^2 \ln f}{\partial \alpha \ \partial \beta}\right) \\ \mathbb{E}\left(\frac{\partial^2 \ln f}{\partial \alpha \ \partial \beta}\right) & \mathbb{E}\left(\frac{\partial^2 \ln f}{\partial \beta^2}\right) \end{bmatrix}^{-1} \\ = & \frac{1}{n} \cdot \begin{bmatrix} \psi'(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{bmatrix}^{-1} \\ = & \frac{1}{n \cdot (\alpha \ \psi'(\alpha) - 1)} \cdot \begin{bmatrix} \alpha & -\beta \\ -\beta & \beta^2 \psi'(\alpha) \end{bmatrix} \end{aligned}$$

This yields (in this particular case - note that we need to use our MLEs to evaluate the previous matrix) an approximate standard error of $\hat{\alpha}$ to be 1.18 and that of $\widehat{\beta}$ to be 0.67; furthermore, the correlation correlation between the two is roughly equal to -0.93 (meaning that, when $\hat{\alpha}$ is *overestimating* the value of α, $\widehat{\beta}$ is most likely *underestimating* the value of β).

Cauchy: Sampling $\mathcal{C}(a, b)$, we have to maximize

$$-n \ln \pi + 2n \ln b - \sum_{i=1}^{n} \ln \left(b^2 + (X_i - a)^2\right)$$

by solving (numerically)

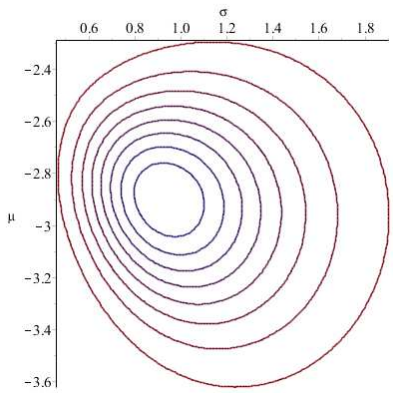
$$\sum_{i=1}^{n} \frac{X_i - a}{b^2 + (X_i - a)^2} = 0$$

and

$$\frac{n}{b} - \sum_{i=1}^{n} \frac{b}{b^2 + (X_i - a)^2} = 0$$

Alternately, one can find the same solution graphically, based on the contour plot of the likelihood function itself. Using the RIS of size 20 from

the previous section, we get



which, after a bit of zooming in, yields $\hat{a} = -2.90$ and $\hat{b} = 0.94$. The corresponding approximate variance-covariance matrix is now

$$\frac{2b^2}{n} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

which makes both standard errors equal to 0.30; this time, the two estimators are practically uncorrelated.

Chapter 12

Confidence Intervals

The previous section considered the issue of so called POINT ESTIMATION, which uses a single (originally random) number as an approximation to the parameter's true, exact value (we should also realize that we will never know this value). Doing this gives us no information about how accurate the estimate is; it is therefore desirable to also have some idea about the size of its error. To provide that, we build a so called CONFIDENCE INTERVAL (CI) around the estimate, which tells us (with a specific LEVEL OF CONFIDENCE) that the true value of the parameter should be inside this interval. It is easy to see that saying: 'we are 90% confident that the true value of the parameter is inside the 8.3 ± 0.1 interval' is quite different from a similar statement which places it inside the 8.3 ± 1.0 limits.

The *level of confidence* (denoted $1 - \alpha$ in general) is defined as the original (i.e. *before* the sample is taken) probability that the resulting CI will contain the true value of the parameter (depending on how lucky we are when doing the sampling). Please realize that, after we have collected a sample and converted it into a CI, there is no randomness left: when we say the true value of the parameter is inside the CI, we are either 100% right or 100% wrong - that is why we speak of 'confidence' rather than 'probability'.

From now on, we assume that our samples are taken from $\mathcal{N}(\mu, \sigma)$.

12.1 CI for μ

We first assume that, even though the (unknown) value of μ is to be estimated, we *do know* the exact value of σ (based on past experience; the implicit assumption being that μ may change in time, but σ does not).

We know that

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

has the $\mathcal{N}(0, 1)$ distribution. This means (a statement about something yet to

happen) that

$$\Pr\left(\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right) = \Pr\left(\left|\bar{X}-\mu\right| < z_{\alpha/2}\cdot\sigma/\sqrt{n}\right) = 1-\alpha \qquad (12.1)$$

where $z_{\alpha/2}$ is the so called CRITICAL VALUE found from

$$\Pr(Z > z_{\alpha/2}) = \tfrac{\alpha}{2}$$

Once we have taken the sample and evaluated $\bar{X}$, the same line (12.1) can be re-written and re-interpreted as follows:

$$\bar{X} - z_{\alpha/2}\cdot\sigma/\sqrt{n} < \mu < \bar{X} + z_{\alpha/2}\cdot\sigma/\sqrt{n}$$

claimed with the $1-\alpha$ level of confidence. Note that the most (by far) commonly used value of α is 5% (leading to the confidence level of 95%).

12.1.1 σ unknown

In this case, we have to replace σ by the next best thing, which is of course sample standard deviation s. We know that the distribution of

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \qquad (12.2)$$

changes from $\mathcal{N}(0,1)$ to $\mathtt{t}_{n-1}$. This means that we also have to change $z_{\alpha/2}$ to $\mathtt{t}_{\alpha/2,n-1}$, the rest remains the same. A $100\cdot(1-\alpha)\%$ confidence interval for μ is then constructed by

$$\bar{X} - \mathtt{t}_{\alpha/2,n-1}\cdot s/\sqrt{n} < \mu < \bar{X} + \mathtt{t}_{\alpha/2,n-1}\cdot s/\sqrt{n}$$

12.1.2 Large-sample case

When n is 'large' ($n \geq 30$), there is practically no difference between $z_{\alpha/2}$ and $\mathtt{t}_{\alpha/2,n-1}$, and we can use $z_{\alpha/2}$ even when σ needs to be replaced by s. Furthermore, the distribution from which we sample does not even have to be Normal! The *approximate* CI for μ is then

$$\bar{X} - z_{\alpha/2}\cdot s/\sqrt{n} < \mu < \bar{X} + z_{\alpha/2}\cdot s/\sqrt{n}$$

claimed with $1-\alpha$ level of confidence.$\frac{n}{\sigma^2}\cdot(\bar{X}-\mu)^2$