definition definition Soln definition

Contents

Data Analysis and Interpretation

Milind Sohoni and Ganesh Ramakrishnan
CSE, IIT Bombay

1 Data

The modern world, of course, is dominated by data. Our own common perceptions are governed to a large extend by numbers and figures, e.g., IPL rankings, inflation statistics, state budgets and their comparisons across the years, or some figures and maps, such as naxalite-affected districts, forest cover and so on. The use of, and the belief in data has grown as the world as a whole and we in particular, become more and more industrialized or 'developed'. In fact, most of us even frame our objectives in terms of numeric targets. For example, the Human Development Index (HDI), is a composite of various sets of data and the Millenium Development Goal is for all countries of the world to achieve certain target numbers in the various attributes of the HDI.

That said, there is much argument amongst politicians, journalists, intellectuals, cricket players, students and parents, about whether society is becoming *too much* or *too less* data-driven. This matches calls for more *subjectivity* (e.g., selecting a suitable boy for your sister) or *objectivity* (admitting students into colleges). In fact, these arguments are popular even among national leaders and bureaucrats, where for example, we now have a new area of study called *Evidence-based Policy Design* which aims to put objectives ahead of ideology and studies methods of executing such policies.

Perhaps the first collectors and users of data were the officers of the kings. Much of the kingdom's expenses depended on taxes, in cash and in kind, from artisans and farmers. This called for maintaining records of, say land productivity, over the years, so that the correct tax rate for the region could be evolved. Also, in the past, ownership of the land could be tied to the expertise of the owner in ensuring its productivity. This too needed a careful understanding of data. Note that for data to be put to use, there must be a certain technical sophistication in understanding (i) what needs to be measured and (ii) how is it to be measured, (iii) how is it to be used, and finally (iv) are our conclusions sound. Thus for example, if you have not measured rainfall, or the number of people in the household, then you would make wrong conclusions on the productivity of the farmer.

Another early use of data was in astronomy. The measurement of this data required several sophisticated actions: (i) the universal acceptance of a certain fixed coordinate system, and (ii) a measuring device to measure the various parameters associated with the objects. While agricultural data was much about the past, astronomical data was largely about the future. Using this, astronomers hoped to predict the seasons, eclipses, and so on. Thus, this involved building *models* from the given data with certain predictive capabilities. In fact, even for the simple *panchang* (the almanac, as known in Maharashtra), there are two models, viz., the *Datey panchang* and the more popular *Tilak panchang*.

1.1 The method of science and its use of data

The method of science is of course, intimately connected with data. Perhaps, the astronomy example above is the earliest demonstration of *the method of science*, as it is known today. This method may be described in the following steps:

- **Observe**. To observe is different from *to see*. To observe also assumes a system and a tool for measurement.
- **Document**. This involves a collection of observations arranged systemtically. There may be several attributes by which we organize our observations, e.g., by time of observation, the rainfall that year and so on. The output of this phase is *data*.
- **Model**. This is the part which wishes to explain the data, i.e., to create a *model* which is the first step towards an explanation. This may be *causal*, i.e., a relationship of cause and effect, or *concommitant*, i.e., of coupled variables. It may be *explicit*, i.e., attempt to explain one variable in terms of others, or *implicit*, i.e., a relationship between the variables which may not be easily separated.

  The simplest model will want to explain the observed variable as a simple function of a classifying attributes, e.g., `rainfall`$>$`1000mm` $\Rightarrow$ `yield` $=$ `1000kg`.
- **Theorize**. This is the final step in the method of science. It aims to integrate the given model into an existing set of explanations or laws, which aim to describe a set pf phenomena in terms of certain basic and advanced concepts. Thus, for example, Mechanics would start with the variables *position, velocity, acceleration, coefficient of friction*, etc., and come up with laws relating these variables.

We now see our first piece of data in Fig. 1.1. These are the water levels observed in an *observation bore-well* managed by the Groundwater Survey and Development Agency (GSDA) of the Govt. of Maharashtra. This borewell is located in Ambiste Village of Thane district. On the X-axis are dates on which the observations were taken, and on the Y-axis, the depth of the water from the top of the well.

The science here is of course, *Groundwater Hydro-geology*, the science of explaining the extent and availability of groundwater and the geology which related to it. Since groundwater

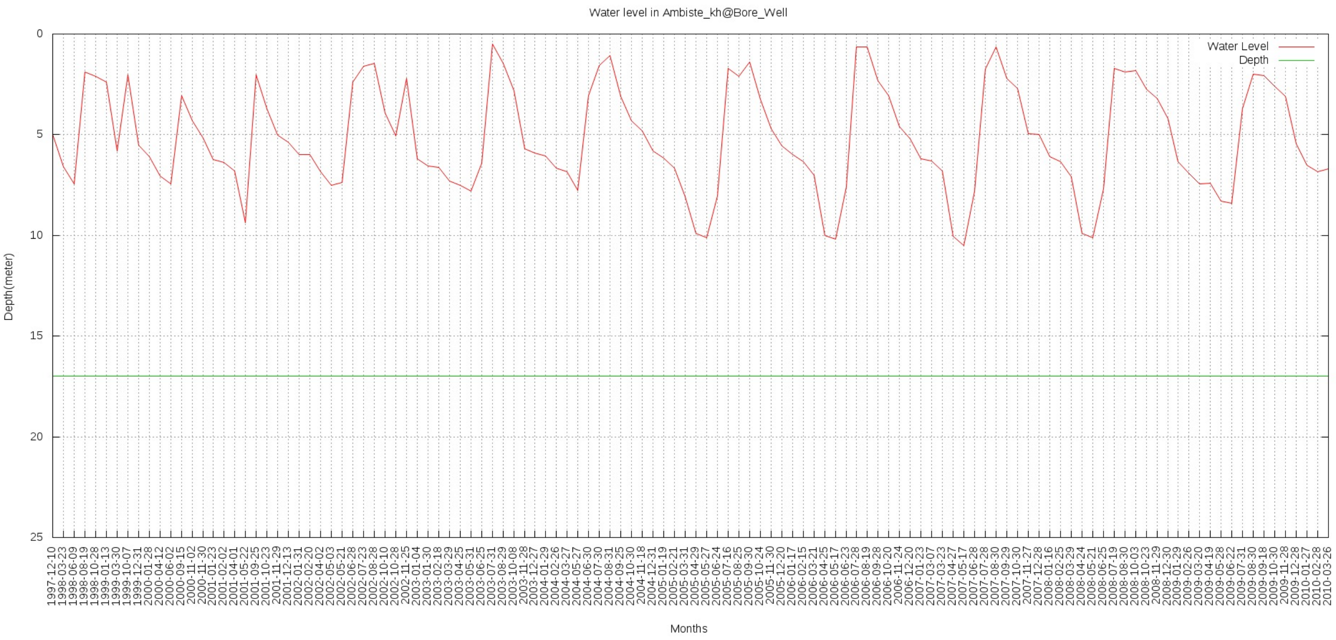

Figure 1: The water levels in a borewell (Courtesy GSDA)

is an important source of drinking water for most indians, almost all states of India have a dedicated agency to supervise the use of groundwater. GSDA does this for Maharashtra. One of the core data-items for GSDA are *observation wells*, i.e., dug-wells and bore-wells which have been set aside purely for observing their levels periodically.

Let us now see how the four steps above apply to this example. Clearly, merely peering down a well or a bore-well (which is harder), does not constitute an observation. We see here that there must have been a device to measure the depth of water and a measuring tape. The next process is documentation. The above graph is one such documentation which wishes to plot the water level with the dates of observations. There is one severe problem with our chosen documentation (*found it?*), and that is that the scale on the X-axis is not uniform on by time, but equi-spaced by observation count. Thus two observations which are 10 days apart and two which are two months apart will appear equally apart in the X-axis. This will need to be rectified. We see here a periodic behaviour, which obviously matches with the monsoons. Thus, groundwater recharges with the rains and then discharges as people withdraw it from the ground through handpumps, wells and borewells. The modelling part could attempt to describe the groundwater levels with time as ideal curves. The science will attempt to explain these curves as arising out of natural laws.

1.2 Data and its attributes

There are two or three important basic concepts that we will associate with data. These are:

- *Variables*: A *variable* is an attribute of any system which may change its value while it is under observation. For example, the number of people in the age group $75 - 79$ in Canada is a variable. There are two basic types of variables, *viz.*, *qualitative* variables and *quantitative* variables.

- *Quantitative*: Qualitative variables take on numeric values. Further, the qualitative variables could be either *discrete* or *continuous* based on whether the variable could take on only whole number values or any number respectively. Typical continuous attributes would be *weight (in kgs.) and* location (in latitude, longitude), *money (in Rupees, USD)*, *height (in inches)*, *age (in days)*, *etc.*. Examples of *discrete* quantitative variables are number of people in New York, number of cars in Germany, the names of *talukas* or anything that can be counted. The discrete set of values is generally regarded as quantitative since its measurement is usually unambiguous.

- *Qualitative*: Qualitative variables take on values that are words – they do not take on numeric values. Examples of qualitative variables include *marital status*, *nationality*, *color of skin*, *gender*, *etc.* Frequently, attributes such as *Satisfaction with Service in a Hotel* are quantfied, in this case, by giving a scale between 1-5. It is obviously unclear if a score of 3 from one customer is better than a 2 from another. Many attributes may be quantitative at first sight but have a hidden quantification rule, e.g., *number of literates in a village*. Here, what should be counted as literacy needs to be defined, and more importantly, the thousands of census workers must be trained to test people by this definition.

- *Integrity*: This is related to the trustworthiness of the data. There could be many reasons to doubt the veracity–improper measuring instruments or of insufficient tolerance, e.g., temepratures reported only as integers (in degree celsius), instead of with one decimal place. Another frequent problem is the interpretion that different measurers have for the same situation. For example, person A may deem person C as literate while person B may not. Loss of integrity in the data is a severe problem from which recovery is not easy. Thus it is best that integrity planned right at the very beginning. One caution–a reading which does not fit the model does not make it necessarily of less integrity. Most real-life processes are fairly complicated and trying to *correct* a reading which doesnt fit may actually convey a more certain world than it really is. For example, if we had a nice theory relating *inflation* with *stock market* *rates*, with exceptions for a few years, then it would be wise to look into the history of those specific years, rather than suspect the data item. Such *'outliers'* may prove to be important.

- *Coverage and Relevance*: This is whether the data (i) covers the situations that we wish to explain, and (ii) includes observations on variables which may be relevant but which we have missed. For example, groundwater levels may depend on the region and not on the specific location. Thus, the explanation of a groundwater reading may be correlated with levels in nearby wells, which unfortunately, we have not monitored. It may also be that groundwater depends intimately on the rainfall in that specific neighborhood, again, which is not included in the data set.

- *Population vs. Sample*: This is whether the data that we have is the whole collection of data items that there are or is a sampling of the items. This is relevant, e.g., when we wish to understand a village and its socio-economics. Thus, we have visit every individual and make readings for this individual. This data is then called the population data. On the other hand, we may select a *representative* sample and interview these selected persons and obtain their data. This is then called the *sample data*. It is not always easy to cover the whole population, for it may be very large (a city such as Mumbai), or it may inaccesible (all tigers in a reserved forst) and even unknown or irrelevant (e.g., measuring soil quality in an area). In such cases, it is the sample and the method of selecting the sample which is or prime importance.

There are of course, many other factors that we have missed in our discussion. These must be surmised for each situation and must be gathered by interveiwing the people who are engaged in the observations and who are familiar with the terrain or subject matter.

1.3 The purpose and content of this course

This course is meant to give the student the skills of interpreting and analysing data. Data is ubiquitous and is increasingly used to make dramatic conclusions and important decisions. In many such situations, the data which led to these conclusions is publicly available and it is important that as a budding professional, you have the skills to understand how the conclusions arose from the data. Besides this, in your professional life, you will yourself be generating such data and would like to draw conclusions and take decisions. These may be more mundane than national policy, but it may still be important enough for your own work. This may be, e.g., to prove to your customer that your recipe works, or to analyse the work of your junior. It may be an important part of a cost-benefit analysis, or it may simply be a back-of-the-envelope analysis of a situation. Handling data and correctly interpreting what it tells and what it does not, is an important skill.

The course has three main parts.

- **Part I: Statistics and Data Handling**. This will cover the basic notion of data-sets, its attributes and relationships. We will introduce the basic terminology is statistics such as the *sample* and concepts such as the *sample mean* and *sample variance*. We will use the following datasets at different points in the notes for illustrating (a) Thane census 2001 data-set (b) population of Canada by age group for the year 2007. We will also study some elementary methods of representing data such as scatter-plots and histograms. Next, we will study the use of Scilab to manipulate data and to write small programs which will help in representing data and in making our first conclusions. Finally, we will develop the elements of least-square fit and of regressions. This is the first model-building exercise that one does with data. We will uncover some of the mathematics of this and also of errors and their measurement.

- **Part II: Probability**. This is the most mathematical part of the course. It consists of explaining a standard set of models and their properties. These models such as the *exponential, normal* or *binomial* distributions are idealized worlds but may be good approximations to your data sets. This is expecially true of the *normal* distribution. The above will be introduced as example characterizations of a formal object called the *random variable*. We will also study functions of random variable and the important notion of *expectation*, which is a single numeric description of a data set. This includes the *mean* and *variance* as special cases.

- **Part III: Testing and Estimation**. This links statistics and probability. The key notions here are of *parameters*, and their estimation and testing. A parameter is an attribute which we believe, determines the behaviour of the data set. For example, it could be the rate of decline in the water level of the bore-well. We will uncover methods of estimating parmeters and assigning it *confidence*. We will use certain well-known tests such as the Kolmogoroff-Smirnov tests, the χ^2-test (pronounced *chi-squared*) and the *Students t-test*. We will also outline methods of accepting and rejecting certain hypotheses made about he data.

2 Datasets

2.1 Data1: The Hungama data

: This data set[1] is extracted from the corresponding report[2]. We will be primarily using this dataset for assignments. It will be also worth looking at the detailed report survey methodology[3], the household survey tool[4] and the village survey tool[5] that form the basis for data collection using this method.

2.2 Data2: The Thane census dataset

The first important dataset for our discussions in the notes will be the Thane district census 2001 dataset. This is available at `http://www.cse.iitb.ac.in/~sohoni/IC102/thane`. The census is organized by the Govt. of India Census Bureau and is done every 10 years. The data itself is organized in Part I, which deals with the social and employment data, and Part II, which deals with economic data and the amenities data. We will be using **village level** data, which is a listing of all villages in India along with the attributes of Part I and II. A snippet of this data can be seen in the figure below.

Let us analyse the structure of Part I data. The data consists of the number of individuals which have a certain set of attributes, e.g., `MARG-HH-M` will list the number of male persons in the village who are marginally employed in household industry. In fact, each attribute is trifurcated as `M,F` and `P`-numbers, which is the male, female and total numbers. We will only list the un-trifurcated attributes:

- `No-HH`: number of houselholds.
- `TOT`: population.
  - `TOT-SC` and `TOT-ST`: SC and ST population.
  - `LIT`: literate population. A person above 7 years of age, who can read or write in any language, with understanding.
  - `06`: population under 6 years of age.
- `TOT-WORK`: total working population. This is classified further under:

[1] `http://www.cse.iitb.ac.in/~IC102/data/hungama_data.xlsx`

[2] `http://www.hungamaforchange.org/HungamaBKDec11LR.pdf`

[3] `http://www.hungamaforchange.org/HUNGaMATrainingManual.pdf`

[4] `http://www.hungamaforchange.org/HUNGaMASurveyTool-Household`

[5] `http://www.hungamaforchange.org/HUNGaMASurveyTool-VillageandAWC`

HH	256	
TOT-P	1287	
P-06	302	
TOT-W	716	
TOT-WORK-MAIN and MARG	374	342
CL	193	171
AL	166	170
HH	0	0
OT	15	1
NON-WORK	571	

Figure 2: Pimpalshet village

  - `MAINWORK`: main working population. This is defined as people who work for more than 6 months in the preceding 1 year.
  - `MARGWORK`: marginal workers, i.e., who have worked less than 6 months in the preceding year.

- `NONWORK`: non-workers, i.e., who have not worked at all in the past year. This typically includes students, elderly and so on.

The attributes `MAINWORK` and `MARGWORK` are further classified under:

- `CL`: cultivator, i.e., a person who works on owned or leased land.
- `AL`: agricultural labourer, i.e., who works for cash or kind on other people's land.
- `HH`: household industry, i.e., where production may well happen in households. Note that household retail is not to be counted here.
- `OT`: other work, including, service, factory labour and so on.

You can find the data for Pimpalshet of Jawhar taluka, Thane in Figure 2.

2.3 Data3: Population of Canada by Age

Table 1 shows a slightly modified estimate of the population of Canada by age group [6] for the year 2007. The first column records the class intervals. *Class intervals* are ranges that

[6] Source: `http://www40.statcan.ca/l01/cst01/demo10a.htm`

the variable is divided into. Each class interval includes the left endpoint but not the right (by convention). The population (second column) is recorded in the thousands. The third column has a record of the percentage of the population that happens to fall in each age group.

Age group	Persons (thousands)	% of total for each group	Height
(class interval)	(count)	(area of histogram)	
0 to 4	1,740.20	5.3	1.06
5 to 9	1,812.40	5.5	1.1
10 to 14	2,060.50	6.2	1.24
15 to 19	2,197.70	6.7	1.34
20 to 24	2,271.60	6.9	1.38
25 to 29	2,273.30	6.9	1.38
30 to 34	2,242.00	6.8	1.36
35 to 39	2,354.60	7.1	1.42
40 to 44	2,640.10	8	1.6
45 to 49	2,711.60	8.2	1.64
50 to 54	2,441.30	7.4	1.48
55 to 59	2,108.80	6.4	1.28
60 to 64	1,698.60	5.2	1.04
65 to 69	1,274.60	3.9	0.78
70 to 74	1,047.90	3.2	0.64
75 to 79	894.7	2.7	0.54
80 to 84	650.8	2	0.4
85 to 89	369.3	1.1	0.22
90 to 95	186.2	0.6	0.12
Total	**32,976.00**	**100**	-

Table 1: A slightly modified estimate of the population of Canada by age group for the year 2007. The population (second column) is recorded in the thousands.

3 Descriptive Statistics: Data representation

Given a large set of data-items, say in hundreds, the mean μ and the variance σ^2 are two attributes of the data (*c.f.* Section 4). A simple representation of the data is the *histogram.* If (y_i) are real numbers, then, we may group the range into a sequence of consecutive intervals and count the *frequencies*, i.e., the number of occurences of data-items for each interval. The histogram will be our first example of a (graphical) descriptive statistic. A histogram provides a picture of the data for single-variable data. We will thereafter discuss the scatter plot (or scatter diagram), which serves as a graphical description of data of two variables.

3.1 Histograms

A histogram is a graphical display of tabulated frequencies. A histogram shows what proportion of cases fall into each of several or many specified categories. In a histogram, it is the area of the bar that denotes the value, not the height. This is a crucial distinction to note, especially when the categories are not of uniform width.

There are three steps to be followed when plotting a histogram for tabulated frequencies as in Table 1.

1. Convert counts to percentages *percent* as shown in the third column of Table 1.

$$percentage = \frac{count}{total\ number\ of\ values}$$

2. Compute height for each class-interval as $height = \frac{percent}{width\ of\ range}$ as shown for the fourth column of Table 1.

3. Draw axes and label them. Label the class intervals (age groups in this case) along the $x-$axis and the heights along the $y-$axis.

4. Along each class interval on the $x-$axis, draw a rectange of corresponding height and width as shown in Figure 3. This is precisely the histogram for the tabulated data in Table 1.

Figure 3 shows the histogram corresponding to Table 1. Note that the sum total area of all the bars is 100 (percent).

Histograms for discrete and continuous variables look slightly different. For histograms of continuous variables, class intervals (such as age ranges) are marked along the $x-$axis of the histogram and the width of the bars could be positive real number. On the other hand, histograms of discrete variables have generally a default width of 1 for every bar and different values assumed by the discrete variable are marked along the $x-$axis. Each bar must be centered on the corresponding value of the discrete variable and the height of every bar is the percentage value.

As another example, consider the taluka of Vasai and the item (y_i) of the number of house-holds in village i. This is a data-set of size 100. The mean is 597, the variance 34100

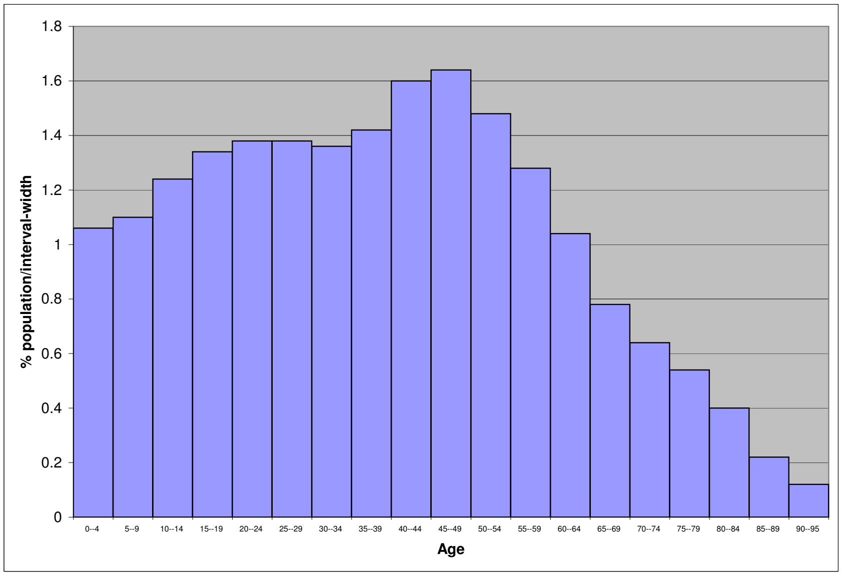

Figure 3: Histogram for Table 1.

and the standard deviation 583 (*c.f.* Section 4), and the maximum size of a village is 3152 households. We may construct intervals $[0, 99]$, $[100, 199]$, $[200, 299]$ and count the number of villages with the number of households in each interval. This aggregated data may be shown in a table:

0-100	100-200	200-300	...
4	15	38	...

This table may be conveniently represented as a *histogram* as in, Fig 3.1. Locate the mean 597 in the diagram and the points $\mu \pm 3\sigma$, viz., roughly 0 and 2200. We notice that there are very few points outside this range. In fact, this is a routine occurence and σ actually is a measure of the dispersion in the data so that most of the data is within $\mu \pm 3\sigma$.

While plotting histograms, there is usually ample room for innovation for selecting the actual variable and the intervals. Here is an example. Consider for example, the data set composed of the tuple (s_i, c_i, n_i, a_i) of drinking water schemes for villages in Thane district sanctioned in the years 2005-2011. Here, n_i is the village name, a_i is the sanctioned amount, s_i is the sanction year and and c_i is the completion year. There are about 2000 entries in this data-set. Here would be a table to illustrate a fragment of this data:

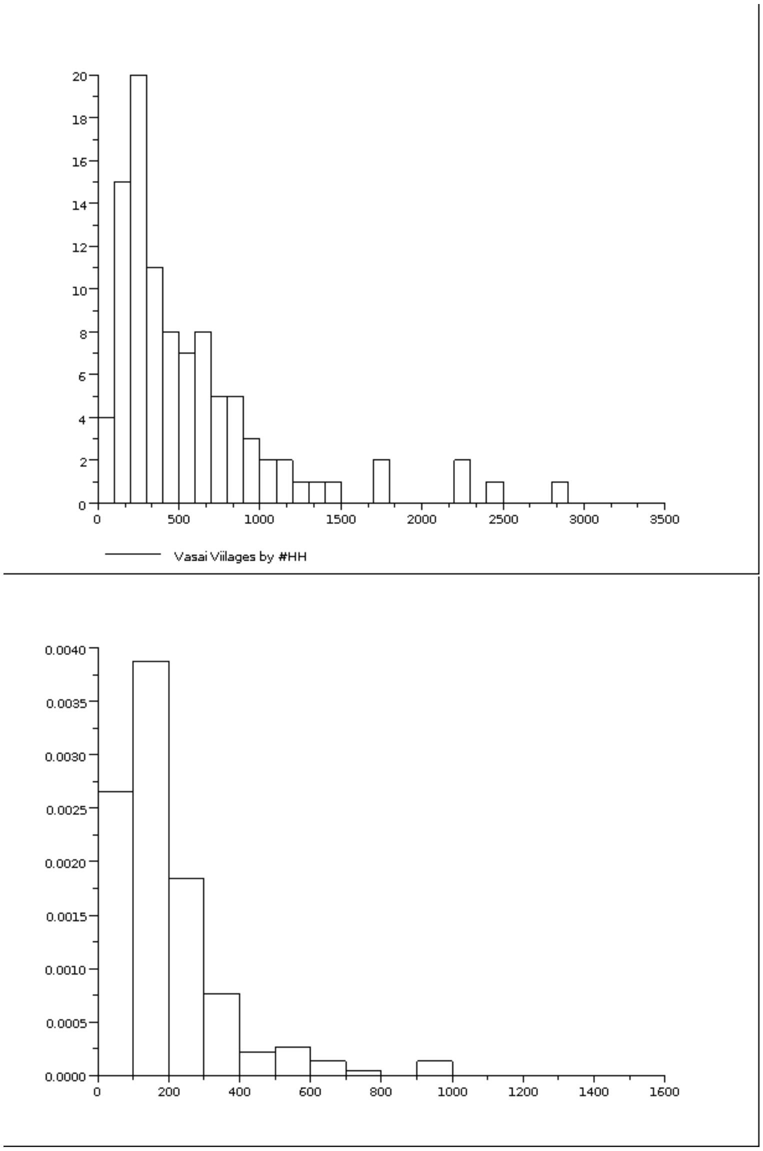

Figure 4: Number of households in villages in Vasai and Shahpur

	Completion Year							
Sanction Year	2005	2006	2007	2008	2009	2010	Incomplete	Total
2005	0	0	3	15	10	13	15	56
2006		0	6	18	33	63	72	182
2007			1	11	12	15	36	75
2008				0	34	55	160	249
2009					1	13	83	97

Reading across a row tells us the fate of the schemes sanctioned in a given year, which reading a column gives us an idea of the number of schemes completed in a particular year. We see that there are considerable variations in the data with 2007 being a lean year and 2008 being an active year in sanctioning and 2009 in completing. In fact, both these years did mark some event in the national drinking water policy.

3.1.1 Density Scale

The height plotted along the y-axis of a histogram is often referred to as the *density scale.* It measures the 'crowdedness' of the histogram in units of '% per x unit'; taller the histogram bar, more is the density. In the last example, the unit was census. Using Table 1 and the corresponding density estimates in Figure 3, one can estimate that the percentage of population aged between 75 and 77 years of age is around $\frac{2.7}{5} \times 3 = 1.62\%$. This is assuming that the density of population in the age group $75 - 79$ is evenly distributed throughout the interval (that is the bar is really flat). But a close look at the bars surrounding that for $75-79$ will suggest that the density in the interval $75 - 59$ is probably not quite evenly distributed. While it would accurate and lossless to have population counts corresponding to every age (instead of intervals), such data may not be as easy to digest as the population estimates based on intervals. There is a tradeoff between summarization and elaborate accounting or equivalently between wider bars and lots of bars.

3.2 Scatter Diagram

Suppose we are prodived data comparing the marks (out of 100) obtained by some 500 students in the mathematics subjects in the year 1 and year 2 of a certain college. Consider a plot with 'year 1 marks' plotted along the x-axis and 'year 2 marks' plotted around the y-axis. The scatter diagram (or plot) for this marks data will consist of a point marked per

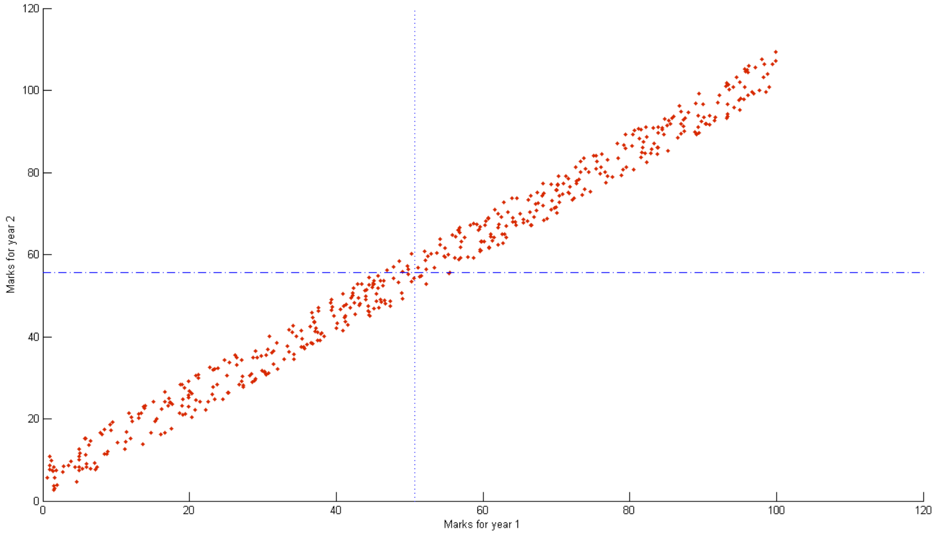

Figure 5: A sample scatter plot.

student with its coordinates give by '(marks in year 1, marks in year 2)'. Figure 5 shows the scatter plot for some such hypothetical data. The dotted vertical and horizontal lines mark the average marks for year 1 and year 2 respectively. It can be seen from the plot that most students either performed well in both years or performed poorly in both years.

A point corresponding to an observation that is numerically far away from the rest of the points in a scatter plot is called an *outlier*. Statistics derived from data sets that include outliers can be misleading. Figure 6 shows the scatter plot of Figure 5 with an outlier introduced (in the form of a black point). The outlier results in a relatively drastic change in mean values of marks for years 1 and 2. While the mean value along the x-axis drops from 50.72 to 50.68, the mean value along the y-axis increases from 55.69 to 55.74.

The scatter plot is used for a data-set consisting of tuples (x_i, y_i) where both are numeric quantities. For example, we could take Shahpur taluka and let x_i be the fraction of literate people in the i-th village. Thus, $x_i =$`P-LIT/TOT-P`. Let y_i be the fraction of people under 6 years of agei, i.e., $y_i =$`P-06/TOT-P`. Thus, we for any village i, we have the tuple (x_i, y_i) of numbers in $[0, 1]$. Now the scatter plot below merely puts a cross at the point (x_i, y_i). Note that we see that as literacy increases, the fraction of people under 6 years of age decreases. However, one must be very careful to assume causality! In other words, it is not clear that one caused the other. It could well be that few children induced people to study.

Warning 1 *The reader should be aware that each village is our individual data item. For example, while calculating the mean literacy of the village, we should add up* `P-LIT` *for all*

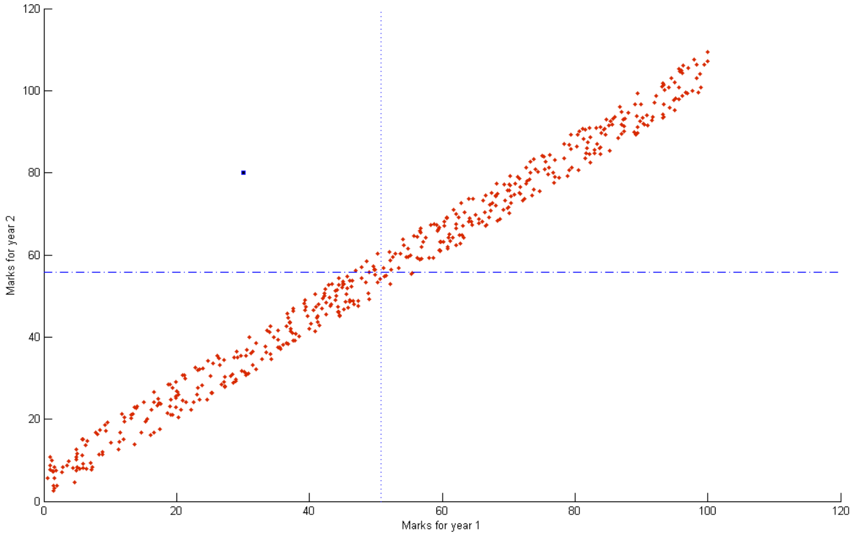

Figure 6: The sample scatter plot of Figure 5 with an outlier $(30, 80)$.

villages and divide it with the sum of `TOT-P`*. However, we have chosen not to do this. One reason is that it tends to drop the identity of the village as site for many correlations which cannot be understood at the individual level. For example, suppose that* `P-LIT`*=450 and* `P-ST`*=300 for a village with* `TOT-P`*=600. At the individual level, it would be impossible from this data to come up with a correlation on ST and literacy. Thus, for correlation purposes, it is only the aggregate which makes sense. There is another reason and that is the lack of independence. For example, if the overall literacy in Murbad is 0.7, then for a village of size 300, if an individual's literacy is independent of others, then the number of literates in the village should be very close to 210. But thats simply not true. Many large villages will show substantial deviation from the mean. The reason of course is that the literacy of an individual in a village is* not independent *of other individuals in the village.*

Not all scatter-plots actually lead to insights. Here is another example where we plot the P-06 fraction vs. the size of the village (measured as the number of households). In this example, we dont quite see anything useful going on.

4 Summary Statistics: Elementary properties of data

The simplest example of data is of course, the table, e.g.,

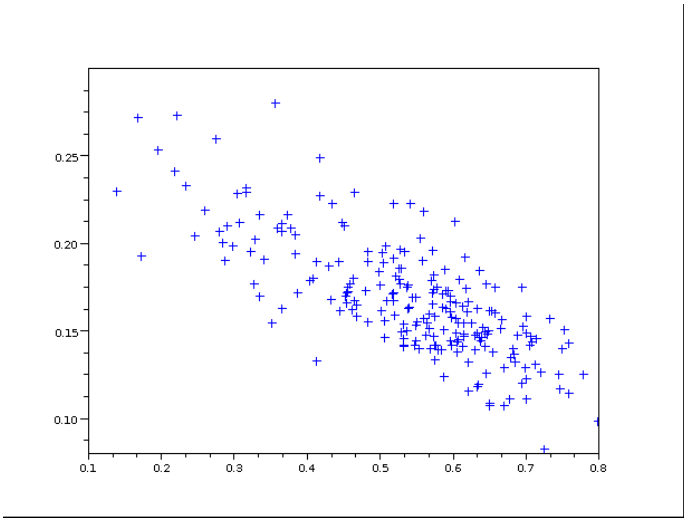

Figure 7: Population under 6 vs. literacy fractions for Shahpur

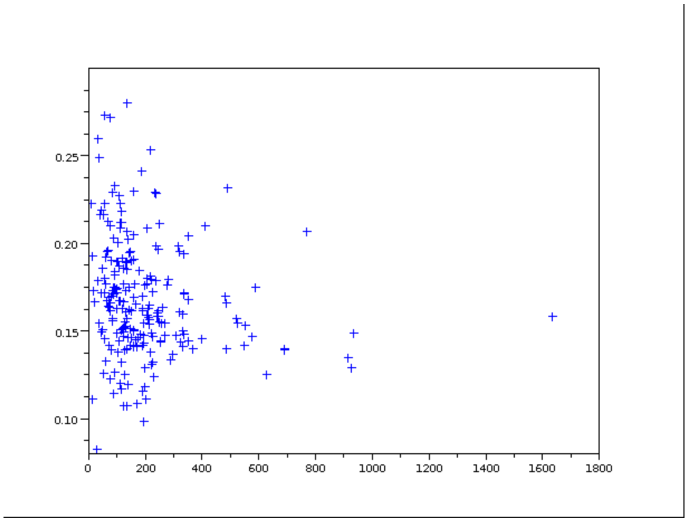

Figure 8: Population under 6 fraction vs. number of HH for Shahpur

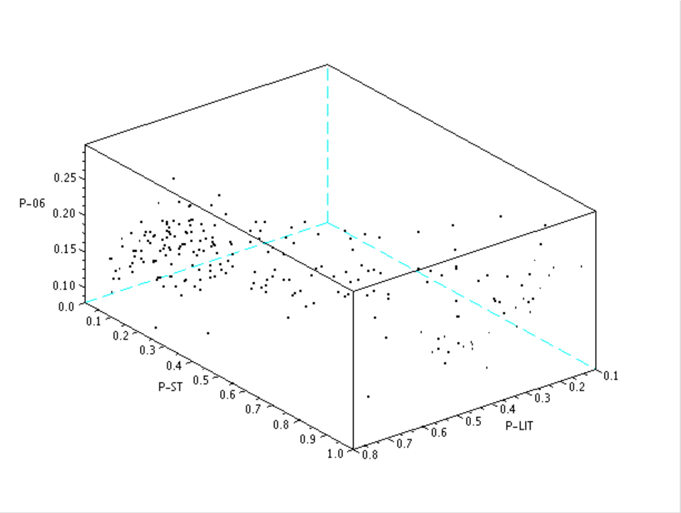

Figure 9: A 3-way plot for Shahpur

Name	Weight (kgs)
Vishal	63
Amit	73
Vinita	58
⋮	
Pinky	48

This may be abstracted as a sequence $\{(x_i, y_i) | i = 1, \ldots n\}$ where each x_i is a name, in this case, and $y_i \in \mathbb{R}$, is a real number in kilos. Summary statistics are numbers that summarize different features of a dataset. There are summary statistics such as mean, median, standard deviation for data of single variables and measures such as correlation coefficient for two variables.

4.1 Standard measures of location

The *arithmatic mean* μ (or simply the *mean*) is one of the most popular summary statistics and it is the average value assumed by the variable. If a variable V assumes values given by the set of data $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, then the mean μ of V is computed as

$$\mu = \frac{\sum_{i=1}^{N} v_i}{N}$$

The mean is also the balancing point on a histogram (*c.f.* Section 3); that is, if you think of the $x-$axis of the histogram as the beam of weight balance and weights to be proportional to the areas of the bars, then the fulcrum placed at the mean point will ensure that the beam stays horizontal.

Another measure of 'center' for a list of numbers is the *median*. The median is the number ν such that at least half the numbers in $\mathcal{V}$ are less than or equal to ν and at least half the numbers are greater than or equal to ν. In order to determine the median, you need to sort the numbers (in either the ascending or descending order) and just pick the center. If more than one value qualifies as the middle value, their average (arithmatic mean) is taken to be the median. The median is the point on the x-axis of a histogram such that half of the total area of the bars lies to its left and half to its right.

As an example, the average of the set $\mathcal{V}' = \{1, 3, 4, 5, 7\}$ is 4, while its median is also 4. On the other hand, if the last number in this set is changed from 7 to 12 to yield the set $\mathcal{V}'' = \{1, 3, 4, 5, 12\}$, then the median remains 4, while the mean changes to 5. Thus, the median cares more for the number of values to its left and right rather than the actual values to its left and right. For the set $\mathcal{V}''' = \{1, 1, 3, 4, 5, 12\}$, the mean is $\frac{13}{3}$, whereas the median is the average of 3 and 4, which is 3.5. In general, for a symmetric histogram, the arithmatic mean equals the median. For a longer (shorter) right tail, the arithmatic mean is greater (smaller) than the median.

In most applications, mean is preferred over median as a measure of center. However, when the data is very skewed, median is preferred over mean as a measure of center. For example, while computing summary statistics for incomes, median is often preferred over mean, since you do not want a few very huge incomes to affect your measure of center. Similarly, median is preferred over mean as a measure of center of housing prices.

Thus,

1. the first single point estimate of the data set is the **mean**. This is denoted by $\overline{y} = \sum_{i=1}^{n} y_i/n$. For example, for the above table, it may be that the mean $\overline{y}$ is 58.6 kgs.

2. **Median** is that value y_{med} such that there are as many items above it as there are below. In other words, if we were to sort the list, then $y_{med} = y_{n/2}$. For the data-set for Vasai in Figure 4, the median is 403.

3. The **mode** of a dat-set is the value which occurs the most number of times. For a data-set which has a lot of distinct possibilities, the mode has no real significance. However, e.g., if (y_i) were the number of children in a household, the mode would be important. For the data-set in Figure 4, a reasonable mode could be read from the histogram and it would be 250, which is of course, the middle value of the interval $[200, 300]$. A mode

could also be a *local maxima* in the number of occurences of a data-item (or a band of data items).

4. Existence of two or more modes may point to two or more phenomena resposible for the data, or some *missing information*. Consider for example, the weights of students in a classroom. Upon plotting the histogram, we may notice two peaks, one in the range 43-45 and another in the range 51-53. Now, it may be that the class is composed of students from two distinct cultural groups, with students from one group weighing more, on the average. Or even simpler, the girls may be lighter than the boys. Thus, the data seems to point that an additional item, e.g., community or sex, should have been recorded while recording y_i.

   Example 2 *Suppose that we are given data* (y_i) *as above. Suggest a mechanism of estimating the two expected mean weights for the two communities/sexes.*

Another often encountered measure is *percentile*. The k^{th} percentile of a set of values of a variable is the value (or score) of the variable below which k percent of the data points may be found. The 25^{th} percentile is also known as the *first quartile*; the 50^{th} percentile happens to be the median.

4.2 Standard measures of spread and association

The measures of center discussed thus far do not capture how spread out the data is. For example, if the average height of a class is 5 feet, 6 inches ($5'\ 6''$), it could be that everyone in the class has the same height or that someone in the class is just $4'\ 5''$ and the tallest student is $6'\ 3''$. The *interquartile range* is an illustrative but rarely used measure of the spread of data. It is defined as the distance between the $25th$ percentile and the 75^{th} percentile. Generally, smaller the interquartile range, smaller will be the spread of the data.

An often used measure of spread is the *standard deviation* (SD), which measures the typical distance of a data point from the arithmatic mean. It is computed as the root mean square[7] (rms) of deviations from the artihmatic mean. That is, given a variable V that assumes values given by the set $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, if μ is the arithmatic mean of $\mathcal{V}$, then the standard deviation σ of $\mathcal{V}$ is

$$\sigma = \sqrt{\frac{\displaystyle\sum_{i=1}^{N} v_i^2}{N}}$$

[7] The root mean square (rms) measures the typical '*size*' or '*magnitude*' of the numbers. It is the square root of the mean of the squares of the numbers. For example, the rms of the set $\{1, 3, 4, 5, 7\}$ is $\sqrt{20} = 4.47$.

The SD of the set $\mathcal{V}' = \{1, 3, 4, 5, 7\}$ is 2, which is a typical distance of any number in $\mathcal{V}'$ from the mean $\mu = 4$. Formulated by Galton in the late 1860s, the standard deviation remains the most common measure of statistical spread or dispersion. The square of standard deviation is called *variance.*

Thus,

1. The **variance**, $\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n}$ is denoted by σ^2. The **standard deviation** is simply the square-root of the variance and is denoted by σ. Note that the units of σ are the same as that of y_i, which in this case, is kilos.

   Lemma 3 *If $z_i = ay_i + b$, where a, b are constants, then $\overline{z} = a\overline{y} + b$, and $\sigma(z) = a\sigma(y)$.*

   The variance is the first measure of *randomness* or indeterminacy in the data. Note that the variance is a sum of non-negative terms whence the variance of a data set is zero iff each entry y_i is equal to $\overline{y}$. Thus, even if one entry deviates from the mean, the variance of the data set will be positive.

2. Much of quantitative research goes into the *analysis of variance*, i.e., the reasons by which it arises. Fo example, if (y_i) were the weights of 1-year-old babies, then the reasons for their variation will lead us to malnutrition, economic reasons, genetic pool and so on. A high variance will point to substantial deviations in the way that these children are raised, maybe the health of the mothers when they were born, and so on. A higher variance is frequently a cause for worry and discomfort, but sometimes is also the basis of many industries, e.g., life insurance. If our mortality was a fixed number with zero variance then the very basis of insurance will disappear.

   Example 4 *Let there be two trains every hour from Kalyan to Kasara, one roughly at xx:10 and the other roughly at xx:50. Suppose that roughly 10 customers arrive at Kalyan bound for Kasara every minute and suppose that the discomfort in a train is proportional to the density, what is the average discomfort?*

   Solution: Well, for the xx:10 train, there will be 200 customers and for the xx:50 train, there will be 400 customers. Whence the density at xx:10 is 200 and that for xx:50 is 400. Thus the average density is $(200 * 200 + 400 * 400)/600 = 2000/6 = 333$. Thus, we see that, on the average there is train every 30 minutes and thus the average density should be 300, however, since this the variance is high, i.e., the departure times are 20 and 40 minutes apart, the average discomfort rises. It is for this reason that irregular operations of trains cause greater discomfort even though the average behaviour may be unchanged. □

Example 5 *For a given data-set (y_i), minimize the function $f(\lambda) = \sum_i (y_i - \lambda)^2$.*

Example 6 *Consider the census data set for Thane and for each taluka, compute the mean, variance and standard deviation for the number of house-holds in each village.*

3. Sometime you need to be careful with computing the means. Here is an example. Part II data of the census, lists for each village, whether or not its people have access to tap water. Thus, let $y_i = 1$ if the i-th village has access to tap-water and $y_i = 0$ otherwise. If we ask, what fraction of the people of Thane have access to tap-water then we would be tempted to compute $\overline{y} = \sum_i y_i/n$ and we would be wrong, for different villages may have different populations. Whence we need the data as a tuple (w_i, y_i), where w_i is the population of the i-th village and thus the correct answer would be:

$$\mu = \overline{y} = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

Thus, one needs to examine if there is a weight associated with each observation y_i. Similarly, the variance for this weighted data is similarly calculated as:

$$\sigma^2 = \frac{\sum_i w_i (y_i - \overline{y})^2}{\sum_i w_i}$$

4.2.1 Effect of change of scale

What is the effect of modifying the data $\mathcal{V}$ on the summary statistics of the data such as arithmatic mean, median and standard deviation? The effect of some data modifications have been studied in the past and are enumerated below.

1. *Adding a constant to every number of the data:* The effect is that arithmatic mean and median go up by that constant amount, while the standard deviation remains the same. This is fairly intuitive to see.

2. *Scaling the numbers in data by a positive constant:* The effect is that the arithmatic mean, the median and the standard deviation get scaled by the same positive constant.

3. *Multiplying numbers in data by -1:* The average and the median get multiplied by -1, whereas standard deviation remains the same.

4.3 The Chebyshev Inqequalities

1. Two-sided: $N(S_k)$ = number of items such that $|x_i - \overline{x}| < ks$

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}$$

Proof:

$$\begin{aligned}(n-1)s^2 &= \textstyle\sum_i (x_i - \overline{x})^2 \\ &\geq \textstyle\sum_{i:|x_i - \overline{x}| > ks} (x_i - \overline{x})^2 \\ &\geq (n - N(S_k))k^2 s^2\end{aligned}$$

$$\Rightarrow \frac{n-1}{nk^2} \geq (1 - \frac{N(S_k)}{n})$$

2. One-sided: $N(k)$ = number of items such that $x_i - \overline{x} \geq ks$

$$\frac{N(k)}{n} \leq \frac{1}{1+k^2}$$

Limits on how 'far' data points can be from mean. Usually data sets are more bunched than Chebyshev.

4.4 Correlation coefficient

In Section 3.2, we discussed a method of studying the association between two variables (x and y). The natural question is if there is a measure of how related are the x_i's with the y_i's. There are indeed metrics for this and the simplest are **covariance** and **correlation**.

Correlation coefficient (r) measures the strength of the linear relationship between two variables. If the points with coordinates specified by the values of the two variables are close to some line, the points are said to be strongly correlated, else they are weakly correlated. More intuitively, the correlation coefficient measures how well the points are clustered around a line. Also called, *linear association*, the correlation coefficient r between sets of N values $\mathcal{X}$ and $\mathcal{Y}$ assumed by variables x and y respectively can be computed using the following three steps.

1. Convert the values in $\mathcal{X}$ and $\mathcal{Y}$ into a set of values in standard unit, *viz.*, $\mathcal{X}_{su}$ and $\mathcal{Y}_{su}$ respectively. Computing standard units requires knowledge about the mean and

standard deviation and could therefore be an expensive step. More precisely, $\mathcal{X}_{su} = \left\{\frac{x_i-\mu_x}{\sigma_x}\,|x_i \in \mathcal{X}\right\}$ and $\mathcal{Y}_{su} = \left\{\frac{y_i-\mu_y}{\sigma_y}\,|y_i \in \mathcal{Y}\right\}$.

2. Let $\mathcal{P}_{su} = \{p_{su} = x_{su}y_{su}\,|x_{su} \in \mathcal{X}_{su}, y_{su} \in \mathcal{Y}_{su}\}$.

3. Let μ_{su} be the arithmatic mean of values in $\mathcal{P}_{su}$. The the correlation coefficient $r = \mu_{su}$.

Thus, if μ_x and μ_y are the means of x and y, and σ_x and σ_y are the respective standard deviations[8]

$$r = \frac{\displaystyle\sum_{x_i \in \mathcal{X}, y_i \in \mathcal{Y}} \frac{x_i - \mu_x}{\sigma_x} \times \frac{y_i - \mu_y}{\sigma_y}}{N} \tag{1}$$

The sample scatter plot of Figure 5 is reproduced in Figure 10 with four regions marked, which are all bordered by the average lines. Points (x_i, y_i) in regions (1) and (3) contribute as positive quantities in the summation expression for r, whereas points (x_i, y_i) in regions (2) and (4) contribute as negative quantities. The correlation coefficient has no units and is always between -1 and $+1$; if r is $+1$ (-1), the points are on a line with positive (negative) slope. A simple case for which $r = 1$ is when all values assumed by y are scalar multiples of the corresponding values of x. If $r = 0$, the variables are uncorrelated. Two special but (statistically) uninteresting cases with $r = 0$, are when either of the variables always takes a constant value. Other interesting cases with $r = 0$ are when the scatter plot is symmetrical with respect to any horizontal or vertical line.

As an example, the correlation coefficient between the marks in years 1 and 2, for the data in Figure 5 is a positive quantity 0.9923. On the other hand, the correlation coefficient between the weight and mileage of cars is generally found to be negative. O-rings are one of the most common gaskets used in machine design. The failure of an O-ring seal was determined to be the cause of the Space Shuttle Challenger disaster on January 28, 1986. The material of the failed O-ring was a fluorinated elastomer called FKM, which is not a good material for cold temperature applications. When an O-ring is cooled below its glass transition temperature (Tg), it loses its elasticity and becomes brittle. In fact, the correlation coefficient between the extent of damage to the O-ring and temperature has been found to be negative.

[8] Note that, while for the Chebyshev's inequality we assumed $\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_x)^2}{n-1}}$, generally, we will assume that $\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_x)^2}{n}}$

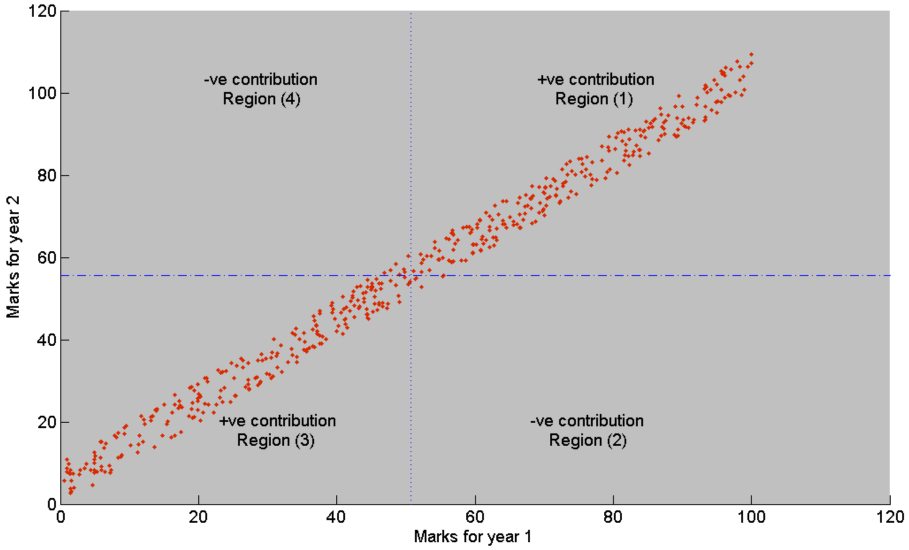

Figure 10: The sample scatter plot of Figure 5 reproduced with four regions marked based on positive and negative contributions to the correlation coefficient.

There are some words of caution that one should exercise while interpreting and applying scatter plots:

1. Extrapolation is generally not a good idea for determining exact values. This is because, outside the range of values considered, the linear relationship might not hold.

2. Even if you cannot do extrapolations, scatter plots can be informative and could give us hints about general trends (such as whether the value of one variable will increase with increase in value of the other variable).

While correlation measures only the strength of the linear association between two variables, the relationship could also be non-linear. In such cases, the scatter plot could show a strong pattern that is not linear (as an example, the scatter plot could assume the shape of a boomerang) and therefore, the quantity r is not as meaningful.

A word of caution before we move on; correlation does not imply causation, while it could definitely make one curious about a possible causal relationship. Just because the GPAs of students and their incomes are positively correlated, we cannot infer that high GPAs are caused by high incomes or vice verca. There could be latent cause of both observations, resulting in the positive correlation. For example, there is generally a positive correlation between the number of teachers and the number of failing students at a high school. But this mainly because generally, larger a school, larger is the number of teachers, greater is the student population and consequently, more are the number of failing students. Therefore, instead of treating the scatter diagram or the correlation measure as a proof of causation, these could be used as indicators that might possibly signal causation.

4.5 Covariance

For a paired data (x_i, y_i), where μ_X and μ_Y are the means of the individual components, the *covariance* of X, Y, denoted as $cov(X, Y)$ is defined as the number

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{n}$$

It can be shown that the correlation coefficient r, also denoted by $corr(X, Y)$ is:

$$corr(X,Y) = \frac{cov(X,Y)}{\sqrt{cov(X,X)cov(Y,Y)}}$$

Lemma 7 *We have $cov(X, Y) = cov(Y, X)$ and that $cov(aX + b, cY + d) = ac \cdot cov(X, Y)$ and $corr(aX + b, cY + d) = corr(X, Y)$. Furthermore, $-1 \leq corr(X, Y) \leq 1$.*

The first part is a mere computation. The second part is seen by recalling the property of the inner product on n-dimensional vectors, which says that $a \cdot b = \|a\| \cdot \|b\| \cdot \cos(\theta)$, where θ is the angle between the two vectors.

We see that the correlation of `(P-06/TOT-P, P-LIT/TOT-P)` is -0.76 while that between `P-06/TOT-P` and `, no-HH` is -0.16. A correlation close to 1 or -1 conveys a close match between X and Y. The correlation between `(p-06/TOT-P)` with `(P-ST/TOT-P)` is 0.57 thus indicating that the fraction of children is more tightly correlated with literacy than with being tribal. Scilab allows a 3-way plot and we plot the fraction of children with that of ST and LIT in Fig. 3.2 below.

Example 8 *Show that $cor(X, Y) = 1$ (or -1) if and only if $Y = aX + b$ with $a > 0$ (or $a < 0$). This exercise shows that if the coorelation of two variables is ± 1 then all points of the scatter plot lie on a line. Furthermore the sign of the slope is determined by the sign of the correlation. Thus, the correlation measures the dependence of X on Y (or vice-versa).*

4.5.1 Effect of change of scale

The effects of change of scale on correlation are far simpler than they happened to be for arithmatic mean and standard deviation. We will list some effects on SD of changing a single variable. The effects of changing values of both variables can be derived by systematically considering effects produced by changing value of each variable.

1. When a constant is added or subtracted from every value of any single variable, the correlation coefficient stays the same, since such an operation involves translating all points and average lines by the same constant value along the corresponding axis. Consequently, the relative positions with respect to the closest line (or the standard units) remain the same.

2. When every value of any single variable is multiplied by the same constant, the correlation coefficient remains the same, since the standard units of the points remain the same (since the average and SD get scaled by the same amount as the values).

3. When every value of any single variable is multiplied by -1, the signs of the values of each variable in standard units change (the value and mean change signs, whereas, the SD does not). Thus, r gets multiplied by -1. However, if each value of both the variables is multiplied by -1, the overall correlation coefficient will remain unchanged.

4. When the values of x are switched with the values of y, the correlation coefficient stays the same, since the terms within the summation expression for r in (1) remain the same.

4.6 Ecological Correlation

In contrast to a correlation between two variables that describe individuals, *ecological correlation* is a correlation calculated based on averages (or medians) of subgroups. The subgroups could be determined based on properties of the data and ecological correlation is just the correlation between means of the subgroups. For instance, the subgroups of students within a class could be determined by sections within the class or by zipcode of the residential area (which is indicative of the affluence) of the students. The ecological correlation between the incomes and the grades of students in a class could then be the standard correlation coefficient between the arithmatic means of the incomes and grades of students within each section or zipcode category. Some researchers suggest that the ecological correlation gives a better picture of the outcome of public policy actions [**?**]. However, what holds true for the group may not hold true for the individual and this discrepancy is often called the *ecological fallacy*. It is important to keep in mind that the ecological correlation captures the correlation between the values of the two variables across the subgroups (such as the zip code of residence) and not across individual students. The ecological correlation can help one draw a conclusion such as '*Students from wealthier zip codes have, on average, higher GPAs*'. A recurring observation is that correlations for subgroup averages are usually larger than correlations for individuals.

5 Linear regression

Regression is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable y (response variable) and of a vector of independent variables $\mathbf{x}$ (explanatory variables). The dependent variable in the regression equation is modeled as a function of the independent variables, corresponding parameters ("constants"), and an error term. However, the relationship between y and need not be causal (as in the case of correlation). Regression is used in several ways; one of the most often used ways is to estimate the average y value corresponding to a given $\mathbf{x}$ value. As an example, you might want to guess the inflation next year, based on the inflation during the last three years.

Consider we have a 2-attribute sample (x_i, y_i) for $i = 1, \ldots n$, e.g., where x_i was the ST population fraction in village i and y_i was the population fraction below 6 years of age. Having seen the scatter plots, it is natural to determine if the value of x determines or explains y to a certain extent, and to measure this extent of explanation. The simplest functional form, of course, is the linear form $y = bx + a$, where the constants b, a are to be determined so that a measure of error is minimized. The simplest such measure is

$$E(b,a) = \sum_{=1}^{n} (y_i - (bx_i + a))^2$$

Since $E(b,a)$ is a continuous function of two variables, its minimization must be obtained at a derivative condition:

$$\frac{\partial E}{\partial a} = 0 \quad \frac{\partial E}{\partial b} = 0$$

These simplify to:

$$\begin{array}{rcl} 2\sum_{=1}^{n}(y_i - (bx_i + a)) & = & 0 \\ 2\sum_{=1}^{n} x_i(y_i - (bx_i + a)) & = & 0 \end{array}$$

This gives us two equation:

$$\begin{bmatrix} \sum_i 1 & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}$$

These are two linear equations in two variables. An important attribute of the matrix is (where μ_X is the mean):

$$\begin{array}{rcl} det\left(\begin{bmatrix} \sum_i 1 & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}\right) & = & n\sum_i x_i^2 - (\sum_i x_i)^2 \\ & = & n\sum_i (x_i - \mu_X)^2 + 2n\mu_X \sum_i x_i - n^2\mu_X^2 - n^2\mu_X^2 \\ & = & n\sum_i (x_i - \mu_X)^2 \end{array}$$

This shows that the determinant is actually non-zero and positive and in fact, $n\sigma^2$. By the same token:

$$\begin{array}{rcl} det\left(\begin{bmatrix} \sum_i 1 & \sum_i y_i \\ \sum_i x_i & \sum_i x_i y_i \end{bmatrix}\right) & = & n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i) \\ & = & n\sum_i (x_i - \mu_X)(y_i - \mu_Y) + n\mu_Y \sum_i x_i + n\mu_X \sum_i y_i - n^2\mu_X\mu_Y - n^2\mu_X\mu_Y \\ & = & n\sum_i (x_i - \mu_X)(y_i - \mu_Y) \end{array}$$

Thus, the slope of the line, viz., b is:

$$b = \frac{\sum_i (x_i - \mu_X)(y_i - \mu_Y)}{\sum_i (x_i - \mu_X)^2}$$

which is a close relative of the correlation `correl(x,y)`. It is easy to check (how?) that the value of b, a as obtained above, actually minimize the error. Thus, our *best linear model* or **linear regression** is $y = f(x)$ is now totally defined. Also observe that $f(\mu_X) = \mu_Y$, i.e., the linear regression is mean-preserving. This is seen by the first defining equation $\frac{\partial E}{\partial a} = 0$, which gives us $\sum_i (y_i - (bx_i + a)) = 0$, and which implies that $\sum_i y_i - f(x_i) = 0$, and which is exactly what we have claimed.

Two examples of the best fit lines are shown below, where we use the Census dataset for Vasai taluka. We map for each village, the fraction of people 6 years old or under as a function of (i) the literacy, and (ii) the fraction of tribal population in the village. Note that the sign of the slope matches that of the correlation.

If we denote $e_i = y_i - bx_i - a$, the error in the i-th place, then (i) $\sum_i e_i = 0$ and the total error squared is obviously $\sum_i e_i^2$. We will show later that $\sum_i e_i^2 < \sum_i (y_i - \mu_Y)^2$. A measure of the goodness of the fit is the ratio

$$r^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \mu_Y)^2}$$

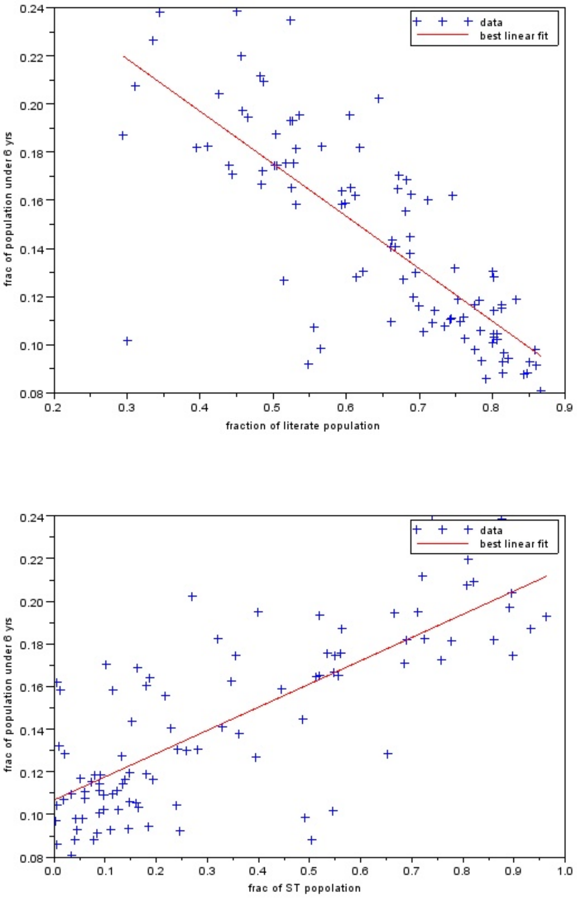

Figure 11: Regression: Population under 6 vs. literacy and ST fraction

The closer r^2 is to 1, the better is the fit. The difference $1-r^2$ is the *residual* or *unexplained* error. See for example, the two data-sets for Vasai: (i) ST-fraction vs. Population below 6, and (ii) male literate fraction vs. female literate fraction.

We now prove the claim that $0 \leq r^2 \leq 1$.

$$\begin{aligned}
\textstyle\sum_i e_i(f(x_i)-\mu_Y) &= b\textstyle\sum_i e_i x_i - a\textstyle\sum_i e_i - \mu_Y \textstyle\sum_i e_i \\
&= \textstyle\sum_i e_i x_i \\
&= 0 \quad \text{since this is the second basic equation}
\end{aligned}$$

Thus, we see that the n-vectors (e_i) and $(f(x_i)-\mu_Y)$ are perpendicular, and sum to $(y_i - f(x_i) + f(x_i) - \mu_Y) = (y_i - \mu_Y)$. Thus we must have $\sum_i e_i^2 \leq \sum_i (y_i - \mu_Y)^2$. In other words $0 \leq r^2 \leq 1$.

Another point to note is that if the input tuple were reversed, i.e., if x were to be explained as a linear function of y, say $x = b'y + a'$, then this line would be different from the best-fit line for y as a function of x. To see this, note that $bb' \neq 1$ in general. In fact:

$$bb' = \frac{\langle x, y\rangle^2}{\langle x, x\rangle\langle y, y\rangle}$$

and thus unless (x, y) are in fact linearly related $bb' < 1$ and thus the two lines will be distinct. See for example below, the two lines for the Vasai female literacy vs. male literacy. The blue line is the usual line while the red line inverts the role of X and Y. Note that the point of intersection is (μ_X, μ_Y).

6 The general model

The above linear regression is a special case of a general class of best-fit problems. The general problem is best explained in the inner product space $\mathbb{R}^n$, the space of all n-tuples of real numbers, under the usual inner product, i.e., for vectors $v, w \in \mathbb{R}^n$, we define $\langle v, w\rangle = \sum_{i=1}^n v_i w_i$. Note that $\langle v, v\rangle > 0$ for all non-zero vectors v and is the square of the length of the vector.

Let W be a finite subset of $\mathbb{R}^n$, say $W = \{w_1, \ldots, w_k\}$. Suppose we have an observation vector $y \in \mathbb{R}^n$. For constants $\alpha_1, \ldots, \alpha_k$, let $w(\alpha) = \sum_{j=1}^k \alpha_j w_j$. Thus $w(\alpha)$ is an α-linear combination of the vectors of W. A good measure of the error that $w(\alpha)$ makes in

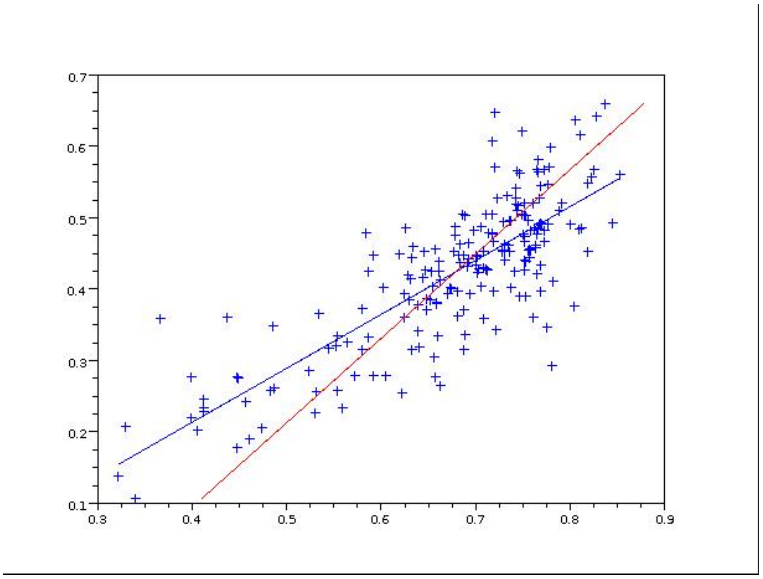

Figure 12: Vasai female vs. male literacy. Both way regression.

approximating y is given by:

$$\begin{aligned} E(\alpha_1, \ldots, \alpha_k) &= \langle y_i - w(\alpha)_i, y_i - w(\alpha)\rangle \\ &= \langle y - \sum_j \alpha_j w_j, y - \sum_j \alpha_j w_j\rangle \end{aligned}$$

The best possible linear combination is given by find those α_j which minimize the error $E(\alpha_1, \ldots, \alpha_k)$. This is done by the equations:

$$\frac{\partial E}{\partial \alpha_j} = 0 \text{ for } j = 1, \ldots, k$$

If we simplify this, we see that these equations reduce to:

$$\langle y - \sum_i \alpha_i w_i, w_j\rangle = 0 \text{ for } j = 1, \ldots, k$$

which in turn reduces to the system:

$$\begin{bmatrix} \langle w_1, w_1\rangle & \langle w_1, w_2\rangle & \ldots & \langle w_1, w_k\rangle \\ \langle w_2, w_1\rangle & \langle w_2, w_2\rangle & \ldots & \langle w_2, w_k\rangle \\ \vdots & & & \vdots \\ \langle w_k, w_1\rangle & \langle w_k, w_2\rangle & \ldots & \langle w_k, w_k\rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix} = \begin{bmatrix} \langle w_1, y\rangle \\ \langle w_2, y\rangle \\ \vdots \\ \langle w_k, y\rangle \end{bmatrix}$$

This matrix system is actually invertible (but we will not prove this) and this solves for the optimal values of the constants $\alpha_1, \ldots, \alpha_k$. Let $f = \sum_j \alpha_j w_j$ be this linear combination and let $e = y - f$ be the error.

Remark: *To see how our earlier linear case is a specialization, we see that for the tuple (x_i, y_i), our W consists of just two vectors, viz., the vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{1} = (1, 1, \ldots, 1)$. The general linear combination is precisely $\alpha_1 \mathbf{1} + \alpha_2 \mathbf{x}$, with the i-th entry $(\alpha_1 + \alpha_2 x_i)$, which after relabelling is $(a + bx_i)$.*

We see that if $\mathbf{1} \in W$, then the condition $\langle e, w_i \rangle = 0$ for all i says that:

$$\langle e, \mathbf{1} \rangle = 0 \Rightarrow \mu_Y = (\sum_i y_i)/n = (\sum_i f_i)/n = \mu_f$$

We also see that

$$\begin{array}{rcl} \langle y - f, f - \mu_f \mathbf{1} \rangle & = & \langle y - f, f \rangle + \mu_f \langle y - f, \mathbf{1} \rangle \\ & = & \sum_j \alpha_j \langle y - f, w_j \rangle + 0 \\ & = & 0 \end{array}$$

This implies that $y - f$ and $f - \mu_f \mathbf{1}$ are perpendicular and thus $\|e\|^2 \leq \|y - \mu_Y \mathbf{1}\|^2$, and thus the error in the approximation does not exceed the variance of the observations y and we may thus define r^2, the goodness of fit, and the residual error similarly.

One useful application of the above formulation is to construct the multi-variable regression. Suppose that we are given the tuples $(x_i, y_i, z_i)_{i=1}^n$ and we seek a regression of the type $z = ax + by + c$. This is computed by considering the set $W = \{(x_i), (y_i), \mathbf{1}\}$ and solving for a, b, c as:

$$\begin{bmatrix} \langle x, x \rangle & \langle x, y \rangle & \langle x, \mathbf{1} \rangle \\ \langle y, x \rangle & \langle y, y \rangle & \langle y, \mathbf{1} \rangle \\ \langle 1, x \rangle & \langle 1, y \rangle & \langle 1, \mathbf{1} \rangle \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \langle x, z \rangle \\ \langle y, z \rangle \\ \langle \mathbf{1}, z \rangle \end{bmatrix}$$

One example of the above is given below- expression of population fraction below 6 as a function of ST-fraction and literacy fraction for Shahpur gives us the coefficient of literacy as -0.2, that of ST fraction as -0.004 and the constant term of 0.227. This indicates that the ST fraction is actually *negatively correlated* with number of children, once literacy is accounted for. Another interesting statistic is the r^2 values for the fits of P-06 with male and female literacy separately. This is shown below for all the talukas of Thane.

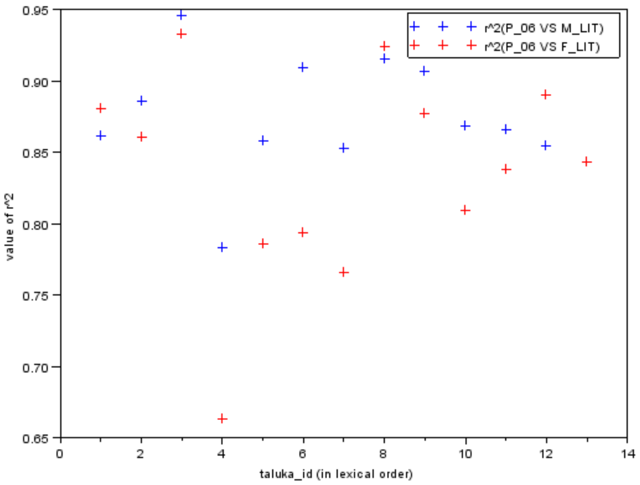

Figure 13: The male and female literacy and P-06

6.1 Illustration on some more examples

We will take three examples and discuss how 'good' the average estimate of the dependent variable could be.

1. In the case of the scatter diagram in Figure 14, one could make use of regression to estimate the average possible marks of a student in 'year 2', given her marks in 'year 1'. In this case, the correlation coefficient (0.993) is very close to 1. The mean values for x and y are 60.26 and 60.33 respectively while the corresponding standard deviations are 10.48 and 10.54 respectively. The graph that shows the average y value corresponding to each x value is called the *graph of averages*. It can be seen in Figure 14, that the average values guessed for y corresponding to different x values are very close to a line and have very little variation across the line. Such a line is called the *regression line*. The regression line is a smoothed version of the graph of averages.

2. In contrast, if the x variable represented a student's score in mathematics and the y variable respresented her score in physics, the correlation might not be as strong. For a typical scatter plot as in Figure 15 for such data, the variance across every average estimate of y for every x value is larger and the correlation coefficient is further away from 1 (0.801 to be precise).

3. In the case of two uncorrelated variables (*i.e.*, $r \approx 0$) such a person's height and his average scores in mathematics, the average value of y given any particular value of x will be more or less the same. The variance will also be large.

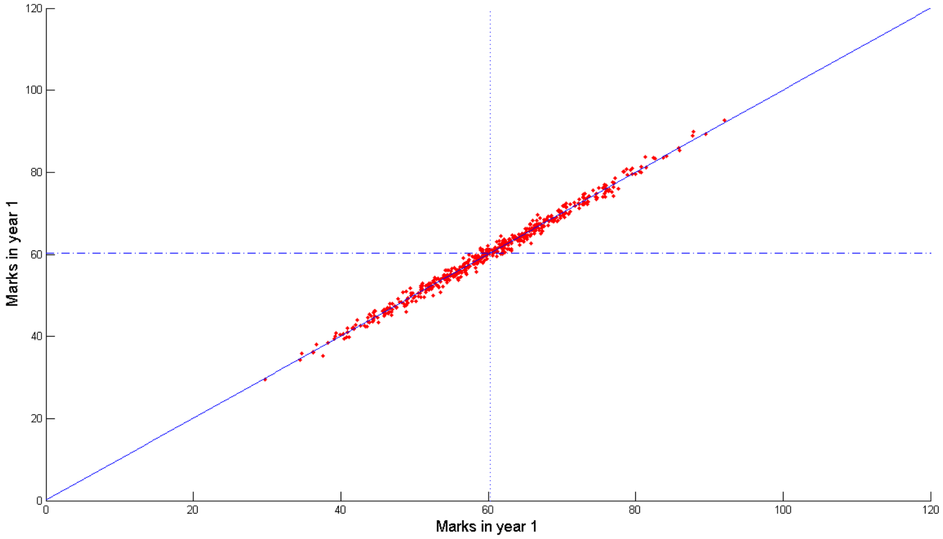

Figure 14: A scatter plot for data similar to that in Figure 5, along with the regression line. The mean values for x and y are 60.26 and 60.33 respectively while the corresponding standard deviations are 10.48 and 10.54 respectively. The correlation coefficient is 0.993.

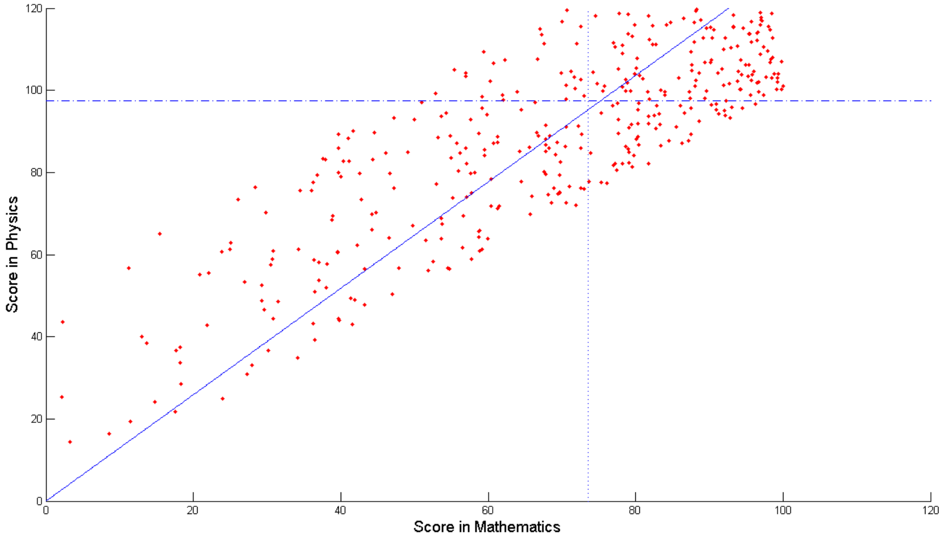

Figure 15: A scatter plot for the marks of students of some class in mathematics (x-axis) against physics (y-axis), along with the regression line. The mean values for x and y are 60.34 and 60.01 respectively while the corresponding standard deviations are 9.82 and 12.06 respectively. The correlation coefficient is 0.801.

It turns out actually that if x is one SD above its average μ_x, the corresponding predicted value of y is r SDs above its average μ_y. The point (μ_x, μ_y) is in fact termed as the *point of averages*. In the third example discussed above, with $r \approx 0$, the predicted y hard changes with the value of x. For the first example, the predicted y changes the most with the value of x.

The *linear regression method* concerns itself with finding points on the regression line, for a given table of data values with each row consisting of values of the independent variable x and the dependent variable y for a particular data point. The following three steps comprise the linear regression method.

1. Compute the correlation coefficient r between x and y. For the data in Figure 14, the correlation coefficient was 0.993.

2. Convert any given value of x into its standard units. Multiply this number by r to get the predicted value of y in its standard units. This step is essentially the *regression step*. For example, with respect to Figure 14, consider $x = 76$. This number in standard units will be $\frac{76-60.26}{10.48} = 1.50$. The predicted value of y in standard units will be 1.49.

3. Convert the value of y from standard units into normal units. The value of 0.993 for y in standard units will correspond to $0.993 \times 10.54 + 60.33 = 70.80$.

The regression line discussed above is for predicting value of y based on x (the usual notion of line). It is also possible to perform regression to predict value of x based on value of y, to yield another kind of regression line.

As an exercise, let us predict (using linear regression on the data in Figure 14) the marks for a student in year 2, given that she obtained 63 marks in year 1. The predicted value for marks in year 2 will be $60.33 + \left(0.993 \times \frac{63-60.26}{10.48}\right) \times 10.54 = 63.07$, which compares favourably with the actual value of $y = 63.53$.

As another exercise, with respect to the scatter plot in Figure 14, suppose a student is at 31^{st} percentile in year 1. What is a good guess for this student's percentile rank in year 2? We note that the scatter diagram is '*football shaped*' (or tilted oval) which is an indication that the two variables roughly follow a normal distribution[9] First we will determine the standard units for x corresponding to the 31^{st} percentile by looking up the value in standard units corresponding to an area of $\frac{100-31}{100} = 0.69$ in standard normal Table **??**. The value happens to be around 0.50. Multiplying this by the correlation coefficient $r = 0.993$, we get the standard units of the approximated y value to be 0.497. The corresponding percentile rank can be again obtained using the normal table to be $100 \times (1 - 0.6879) = 31.21$, *i.e.*,

[9] Later, we will discuss techniques that will help verify such assumptions.

around 31 percentile. Note that we really did not make use of any of estimates for mean or standard deviation, since all computations dealt only with standard units.

What about the spread of values around the average predicted by linear regression? How accurate is regression? For each point in the scatter diagram, the actual value of y will not typically be on the regression line. The *prediction error* at a point x is defined as the difference between the actual and predicted values of y at that point. Analogous to the definition of the standard deviation, the spread for linear regression is measured as the RMS of prediction errors and is called the *RMS error*. Linear regression should be suitable for datasets for which the RMS error is very close to sum of the deviations from average across all points. While RMS error can be used in the context of any prediction method, for the case of linear regression, the RMS error can be simplified to the following formula:

$$e_{rms} = \sigma_y \sqrt{1 - r^2} \tag{2}$$

where σ_y is the SD for y.

On the other hand, if the prediction method were to always predict the overall average μ_y of y regardless of the value of x, the RMS error would be the standard deviation σ_y of y (by definition). We can readily see that if $r \in (-1, 1)$, prediction using regression will always yield a lower value of e_{rms} than will a naive prediction using the average μ_y.

Mathematically, if $r = \pm 1$, all the points must be on a line (with slope r), resulting in 0 prediction errors and $e_{rms} = 0$. On the other hand, if $r = 0$ and $e_{rms} = \sigma_y$. This is because there is no linear relationship between x and y, implying that it does not help to know x while predictin y; you can always guess y to be μ_y and could do no better. For any other $r \in (-1, 0) \cup (0, 1)$, $e_{rms} \in (0, \sigma_y)$. If e_{rms} is close to 0, it means that the predictions are very accurate. Whereas, if e_{rms} is close to σ_y, the predictions must be far from accurate.

What about answer to a question such as '*what percentage of all values are within one unit of RMS error of the regression line for a fixed value of x*'? The answer will be around 68%, assuming that the points approximately follow a normal curve and that the average values of y for any given x approximately coincide with the point (x, rx) on the regression line, so that one unit of RMS error approximately corresponds to one standard deviation. For a football shaped scatter diagram, the normality condition can be assumed to hold for both x and y as well as for all points with a fixed value of x. Using this information, let us answer the following question: of all the students considered in Figure 14 who obtained a score of 65 in year 1, what percentage of them obtained over 67 in year 2? First of all, the average marks in year 2, for students who obtained 65 in year 1, predicted using linear regression will be $60.33 + \left(0.993 \times \frac{65-60.26}{10.48}\right) \times 10.54 = 65.06$. The RMS error can be computed using

(2) to be $\sqrt{1-(0.993)^2} \times 10.09 = 1.19$. The question then reduces to 'what percentage of values on a normal curve with mean 65.06 and SD 1.19 are above 67?' The answer to this can be easily obtained by looking up the standard normal table for the percentage area that is above 1.63 standard units, which is $100 \times (1 - 0.9484) = 5.16\%$.

Further, a football shaped scatter diagram is *homoscedastic*, *i.e.*, it has approximately the same spread around the regression line for every value of x. To cite a practical example, we would expect different divisions within a class to have the same mean and SD values for scores in a particular year, irrespective of their sizes. Of course, divisions with smaller sizes may have a smaller range of scores than larger divisions.

You can refer to the Appendix for a more detailed discussion on Regression including (a) regularization (b) drawbacks of regression.

6.2 The Regression Effect

The regression effect is also called 'regression to the mean'. This effect is essentially the fact that for $r \in (-1, 1)$, and both x and y in standard units, the predicted (based on linear regression) y value will be closer to the mean than than the x value. This is because, if x is 1 SD above average, then the predicted value of y will be $r \in (-1, 1)$ SDs above average. Thus, if a student performs exceptionally well in the first year, the predicted performance of the student in the second year will not be as good. A corollary of this effect is that if y_1 is the predicted average value of y variable for a value x_1 of the x variable, then the predicted average value of x_2 of the x variable based on value y_1 will be less than x_1.

6.3 SD Line

The SD line is a line that passes through the *point of averages* in a scatter plot and has slope equal to the ratio of standard deviation of y to that of x multiplied by $sgn(r)$. Points on the SD line are the same in standard units for both x and y. It can be proved that if $r = 1$, the SD and regression lines will coincide.

7 The Gini Coefficient

This is yet another interpretation of a tuple data (x_i, y_i) which is also used frequently as a measure of inequality. Suppose that the tuple is a *frequency data* for a variable y_i, e.g., the income. In other words, suppose that for each i, there were x_i persons with income y_i. Such data is frequently available, e.g., for professors in IIT-B and their scale of pay. The variable y_i need not always be economic, e.g., y_i could be from 1-15, denoting the number of years

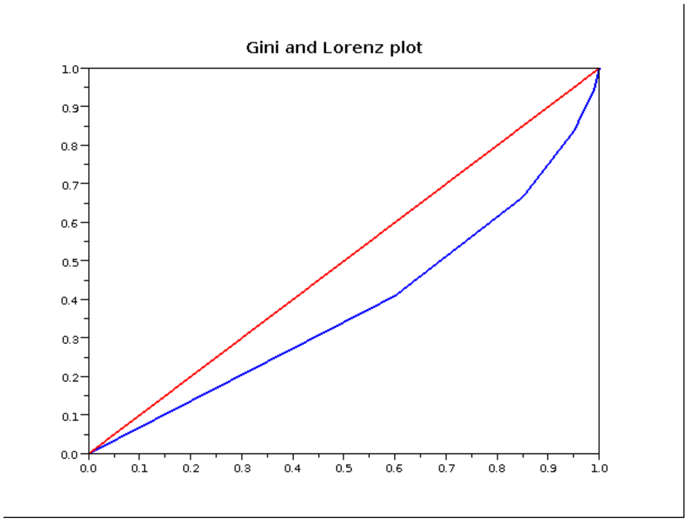

Figure 16: The Lorenz plot for the company data

for formal education and then x_i would be the number of people having i years of formal education.

Now, we would like to measure the *inequality* in the data. Our first step is to assume that the y_i's are sorted, i.e., $y_1 < y_2 < y_3 \ldots < y_n$. Next, let $X_i = \sum_{j=1}^{i} x_i$, in other words, X_i is the number of people with values less than or equal to y_i. Let $X = X_n$ be the number of people in the sample. Next, we define $Y_i = \sum_{j=1}^{i} x_j * y_j$, i.e., net value for the first i groups of people. Let $Y = y_n$, the total value of the population. The *Lorenz curve* is the plot which begins at $(0, 0)$ and plots $(X_i/X, y_i/Y)$.

Example 9 *A company has 100 employees at various levels. The number of employees at each level and their salaries are given below:*

No. of Employees	*60*	*25*	*10*	*4*	*1*
Pay (in lakh Rs.)	*1*	*1.5*	*2.5*	*4*	*8*

We thus see that $X = 100$, $Y = 146.5$ *and the plots for the Lorenz curve will have the following data:*

0.00	*0.60*	*0.85*	*0.95*	*0.99*	*1*
0.00	*0.41*	*0.67*	*0.84*	*0.95*	*1*

The curve is shown in Figure 16.

It is easy to see (show this as an exercise) that $y_i/Y < X_i/X$, i.e., the Lorenz curve always sits below the 45-degree straight line joining $(0, 0)$ with $(1, 1)$. Note that in the above

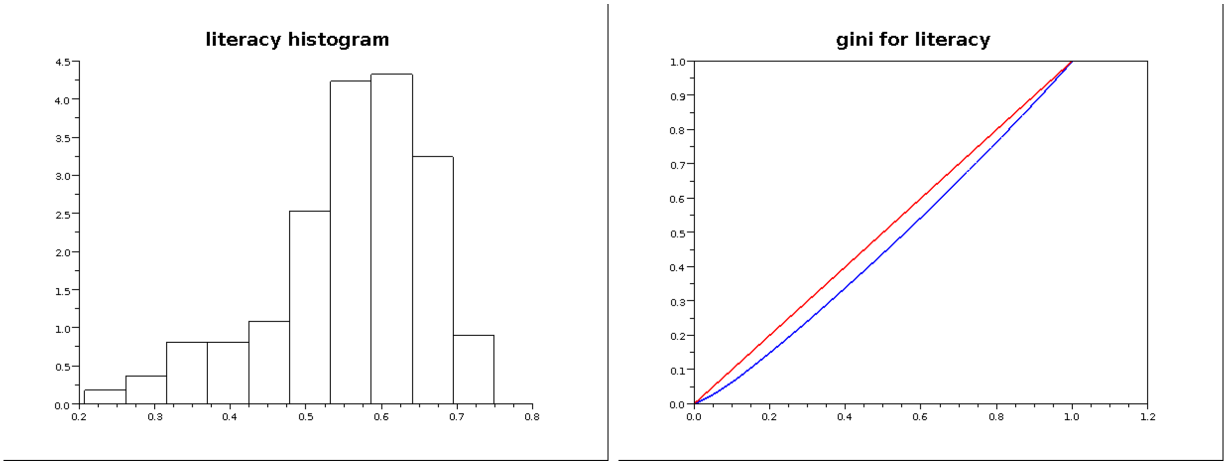

Figure 17: The Lorenz plot for Murbad literacy

example, if the salaries were more equal then the Lorenz curve will be closer to the 45-degree. The *Gini coefficient* is the ratio of the area A between the Lorenz curve and the 45-degree line to the area below the line. Since area under the line is 0.5, the Gini coefficient is exactly $2 \cdot A$. The Gini ceofficient is easily computed using the trapezium rule, as follows:

$$2 \cdot G = \sum_{i=1}^{n} \frac{x_i}{X} \frac{(x_i - Y_i) + (x_{i-1} - Y_{i-1})}{2}$$

This is available as a function `gini.sci` which inputs a matrix of two columns, where the first column are the x_i's and the second column are the y_i's. Make sure that the second column is increasing. It turns out that our company has a Gini coefficient of 0.245.

Let us try another example for aggregate data. For the Murbad taluka census data, we have for each village i, its population (TOT-P) and the number of literates (P-LIT). The i-th village literacy fraction y_i is then given by $PLIT_i/TOTP_i$. Let us denote x_i by $TOTP_i$. Let us understand what this tuple data and its Gini coefficient (x_i, y_i) would mean. Since the data is aggregated for each village, we will measure the inequality in the literacy levels *across* villages. This will smoothen out the education levels *within* the village, at the individual level. For Murbad, we see that the coefficient is 0.0878 which is quite small. This is also evident from the histogram which is bunched around the mean. The plots appear below.

Warning 10 *The Gini coefficient must be used with care. For aggregate data, it will tend to under-compute the inequality. You should try this for say part II data, e.g., total agricultural land. The Gini may be quite low but may hide that within each village, land may be concentrated in very few households. So unless household data is available, the inequality in land ownership cannot be measured.*

8 Probability

The notion of probability comes from a **random variable**, which is just an abstract data source. Think for example, of a cannon which may be fired repeatedly. Every firing i will yield a transit distance d_i of the cannon ball. Clearly, as there are variations in the sizes and weights of the cannon ball, variations in the wind conditions, and so on, we will have that the d_i's will not be all equal. All the same, a repeated observation will indeed give us an estimate of the range of the cannon.

We can now loosely define a random variable X as (i) an **outcome set** S, (ii) a collection $\mathcal{E}$ of subsets of S, called the **event set**, and (iii) a **probability function** $p : \mathcal{E} \to \mathbb{R}$, all with certain properties. For $\mathcal{E}$, we must have that (E1) $S \in \mathcal{E}$, (E2) if $A, B \in \mathcal{E}$, then so are $A \cap B$ and $\overline{A}$, i.e., the complement of A. These conditions say that the subsets in $\mathcal{E}$ are closed under boolean operations. More formally:

8.1 Basic Definitions

Definition 11 *Sample space ($\mathcal{S}$) : A sample space is defined as the set of all possible outcomes of an experiment. Example of an experiment would be a coin pair toss. In this case $S = \{HH,\ HT,\ TH,\ TT\}$.*

Definition 12 *Event ($\mathcal{E}$) : An event is defined as any subset of the sample space. Total number of distinct events possible is $2^{\mathcal{S}}$, where $\mathcal{S}$ is the number of elements in the sample space. For a coin pair toss experiment some examples of events could be*

$$\text{for at least one head, } \mathcal{E} = \{HH, HT\}$$
$$\text{for all tails, } \mathcal{E} = \{TT\}$$
$$\text{for either a head or a tail or both, } \mathcal{E} = \{HH, HT, TH, TT\}$$

Definition 13 *Random variable (X) : A random variable is a mapping (or function) from set of outcomes to a set of real numbers. Continuous random variable is defined thus*

$$X : \mathcal{S} \to \mathbb{R}$$

On the other hand a discrete random variable maps outcomes to a countable set (e.g. discrete real numbers)

$$X : \mathcal{S} \to \textit{Discrete } \mathbb{R}$$

Now, we move to the probability function Pr. Probability Pr is a function $P : 2^S \to \Re$. It must have the following properties: (P1) $Pr(A) \geq 0$ for all $A \in \mathcal{E}$, (P2) $Pr(\phi) = 0$ and $Pr(S) = 1$, and (P3) if $A \cap B = \phi$ then $Pr(A \cup B) = p(A) = p(B)$. More formally:

8.2 The three axioms of probability

The probability function $Pr(.)$ satisfies the following three axioms:

1. For every event $\mathcal{E}$, $Pr(\mathcal{E}) \in [0, 1]$
2. $Pr(\mathcal{S}) = 1$ where, $\mathcal{S}$ is the sample space. (Equivalently, $P(\emptyset) = 0$)
3. If $\mathcal{E}_1, \mathcal{E}2, \ldots, \mathcal{E}_n$ is a set of pairwise disjoint events, then

$$Pr(\bigcup_{i=1}^{n} \mathcal{E}_i) = \sum_{i=1}^{n} Pr(\mathcal{E}_i)$$

Example 14 ***The biased coin.*** *Here we construct the random variable $C(q)$ corresponding to the biased coin. Let $\mathcal{S} = \{H, T\}$, i.e,* `heads` *or* `tails`*, be the only possible outcomes of a coin toss. Let $\mathcal{E}$ be the set of all possible (i.e., $2^{\mathcal{S}}$) subsets of $\mathcal{S}$, and let $0 < q < 1$ be a fixed real number. We define Pr by the table below:*

set	ϕ	$\{H\}$	$\{T\}$	$\{H, T\}$
p	0	q	$1 - q$	1

This merely says that the probability of the coin falling H is q, of T is (obviously) $1 - q$, of not falling at all is zero, and of falling either H or T is 1.

Example 15 ***The cannon-ball.*** *Here, let $S = [100, 101]$, ie., the possible outcomes are all real numbers between 100 and 101. Let $\mathcal{E}$ be the collection of all sub-intervals, open or closed, of $[100, 101]$ and their unions. For an interval $[a, b]$ we define $p([a, b]) = b - a$. This random variable CB simulates the falling of a cannon ball. It says that the cannon ball will always fall between 100m and 101m from the cannon and the probability that a particular trial falls within the interval $[a, b]$ is in fact $b - a$. For example, the probability of the ball falling between $[100, 100.2]$ or $[100.5, 100.7]$ is equal and 0.2. In other words, every outcome between 100 and 101 is equally likely.*

Two random variables X and Y are called independent if the outcome of one do not affect the outcome of the other. Here are some *dependent* random variables. Let B be a box containing k red and $n-k$ black balls. Let us first draw one ball and note its (random) colour as X_1 and throw it away. Next, let us draw a second ball and denote its colour by the variable X_2. Note that as individual random variables, X_1 and X_2 are identical, viz., the probability of a red ball is k/n. However, they are certainly not independent. If we know the outcome of one then we do know something more about the outcome of the other. Another example is when X is the time that you will wait for your bus and Y is the time elapsed since the last bus, measured at the instant that you show up at the bus-stop. Another example is say the life-expectancy of one resident of a village with that of another in the same village.

We will not study **independence** formally but assume an informal understanding that one should be careful before assuming that two random variables are independent.

We will denote by $\mathcal{E}^0$ the collection of all open/closed intervals and their disjoint unions. Verify that it satisfies condition E1 and E2. When S is a finite set, we assume that $\mathcal{E}$ is the collection of all subsets of S. Note that p is then defined by specifying its value on singletons, i.e., $p(\{s\})$ (this we abbreviate as $p(s)$) for all $s \in S$. For if $A = \{s_1, \ldots, s_k\}$), then $p(A)$ is clearly $p(s_1) + \ldots + p(s_k)$.

Next, let us construct new random variables from old. The simplest is the **cross product**. If $(S_1, \mathcal{E}_1, p_1)$ and $(S_2, \mathcal{E}_2, p_2)$ are two random variables, then we can construct the *product*. We define $S = S_1 \times S_2$, $\mathcal{E}$ as the sets which include $\mathcal{E}_1 \times calE_2$, and define $p(A \times B) = p_1(A)p_2(B)$.

Example 16 *Lets look at $C(q) \times C(r)$. This corresponds to two independent coin throws, where one coin has bias q and the other r. We see that $S = \{HH, HT, TH, TT\}$ and $p(HH) = p_1(H)p_2(H) = qr$, while $p(HT) = p_1(H)p_2(T) = q \cdot (1-r)$, and so on.*

We may construct $CB \times CB$, i.e., the random variable corresponding to two independent ordered cannon ball firings. Clearly the outcome set is $[100, 101] \times [100, 101]$, i.e., the unit square situated at $(100, 100)$. The probability $p([100, 100.2] \times [100.3, 100.4]) = 0.2 \times 0.1 = 0.02$. Thus the probability of the first shot falling in the first named interval and the second in the second interval is 0.02.

There is another technique of constructing random variables. Let $R = (S, \mathcal{E}, p)$ be a random variable and let S' be another set and $f : S \to S'$ be a onto function. We define the new variable $R' = (S', \mathcal{E}', p')$, where S' is as above. We say that $A' \in \mathcal{E}'$ iff $f^{-1}(A) \in \mathcal{E}$, and when this happens, we define $p'(A') = p(f^{-1}(A))$.

Let us now construct our first important example and that is the **Binomial random variable** $Binom(q, n)$.

Definition 17 *The variable $Binom(q, n)$ has the outcome set $[n] = \{0, 1, \ldots, n\}$ with $p(\{k\}) = \binom{n}{k} q^k (1-q)^{n-k}$. The binomial random variable arises from the n-way repeated trials of $C(q)$,* *i.e., $C(q) \times \ldots \times C(q)$. Note that sample space of this product is S^n which is the collection of 2^n sequences in H and T, corresponding to the fall of the i-th coin. Now consider the map $f : S^n \rightarrow [n]$ where each sequence goes to the number of H's in it. For example, for $n = 4$, $f(HHTH) = 3$ while $f(TTHH) = 2$ and so on. Thus, the function f merely counts the number of heads. Now, if we consider any $k \in [n]$, then then $f^{-1}(k)$ is precisely the set of sequences with k heads, and the probability of the occurence of k heads in an n-toss of a biased coin then is precisely the number above.*

Here is an example where $Binom(q, n)$ will find use. Suppose that we must judge the fraction q of tribals in a large village. One test, if we are unable to survey the whole village, would be to take a sample of n people (more about sampling later), and count the number of tribals, say k. Whence, if q were this fraction, then the chance of our meeting exactly k tribals from a sample of n is exactly $\binom{n}{k}q^k(1-q)^{n-k}$. We will see later that k/n is a reasonable estimate of q.

9 Probability Density Functions

We now come to the important case of **probability density functions**. These arise, in their simplest form, when the outcome set S is a simple subset of $\mathbb{R}$, say an interval or the whole real line, and the event set is $\mathcal{E}^0$. Let $f : S \rightarrow \mathbb{R}$ be a smooth function such that (i) $\int_S f(x)dx = 1$, (ii) $f(x) \geq 0$ for all $x \in S$, and $f(x) = 0$ when $x \notin S$. We may define the probability of an interval I as $p(I) = \int_I f dx$, i.e., the area under the curve $f(x)$ over the interval I. When we construct a random variable in such a manner, f is called its probability density function. In a crude sense, the probability that an outcome of the random variable is between x and $x + dx$ is $f(x)dx$.

Example 18 ***The uniform random variable.*** *Let $S = [0, 1]$ and let $f(x) = 1$ for $x \in [0, 1]$ and zero otherwise. We see that for any sub-interval $[c, d]$, $p([c, d]) = \int_c^d 1.dx = d - c$. If we wished to construct the uniform random interval over another interval $[a, b]$, then $f(x) = \frac{1}{b-a}$ for $x \in [a, b]$ would do the job, and then, as expected, $p([c, d]) = \frac{d-c}{b-a}$.*

Example 19 *Here is a more interesting case. Let $S = [0, 1]$ and let $f(x) = 2x$ for $x \in S$ and zero otherwise. We see that $\int_S 2x.dx = (x^2)_0^1 = 1 - 0 = 1$. Also, $f(x) \geq 0$ for all x, and thus f defines the pdf of a random variable. We see that $p([0, 0.5]) = 1/4$ while $p([0.5, 1]) = 3/4$ and thus this random variable prefers higher values than lower ones.*

The Normal density function.

We now come to the famous Normal or Gaussian random variable. The outcome set for this is $\mathbb{R}$, the whole real line. Let

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

This is a curious function which arises from classical mathematics and is plotted as the red curve in the image below (from wikipedia). We see that the curve is smooth and symmetric. The integral $\int_{\mathbb{R}} f(x)dx$ is known to be 1. We see that the normal random variable allows for all real numbers as outcomes but prefers smaller numebrs (in absolute value) to bigger one. The integral values of $\int_a^b f(x)dx$ are rather hard to calculate analytically and are usually tabulated. We see for example that $p([-2,2]) = \int_{-2}^{2} f(x)dx = 0.95$, roughly. As can be seen from the graph below, most of the area under the red curve is indeed between -2 and 2. In terms of randomness, we see that the chance that the random outcome is in $[-2,2]$ is about 95%.

The above denisty function is usually denoted by $N(0,1)$. The general function is $N(\mu,\sigma)$ and is given by:

$$N(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-((x-\mu)^2)}{2\sigma^2}}$$

Assuming that $\int f(x)dx = 1$, it is easily shown that $N(x;\mu,\sigma)$ also gives a density function. This is called the **normal density function with mean μ and variance σ^2.** The figure shows some plots for various μ and σ^2. We see that μ decides the center value around which the random variable is symmetric. Increasing σ increases the spread of the outcomes. For example,

$$\int_{-2}^{2} N(x;0,2)dx = \int_{-1}^{1} N(x;0,1)dx = 0.65$$

Thus, the spread of $N(0,2)$ is more than $N(0,1)$.

The obvious question is: where and how do normal random variables arise? The answer is really from the Binomial case when n is large and x is taken to be $k/n - 0.5$. But more on that later.

The density function approach is an important analytic tool in understanding many other random variables. For example, we may wish to understand how is the maximum score in a quiz for a class distributed, or for example, the distribution of the mean of n repeated trials and so on.

Let us look at the first problem. Let R_1, R_2 be two variables given by density functions f_1, f_2, then the outcome set of the cross-product is clearly (x,y) with $x, y \in \mathbb{R}$, or in other

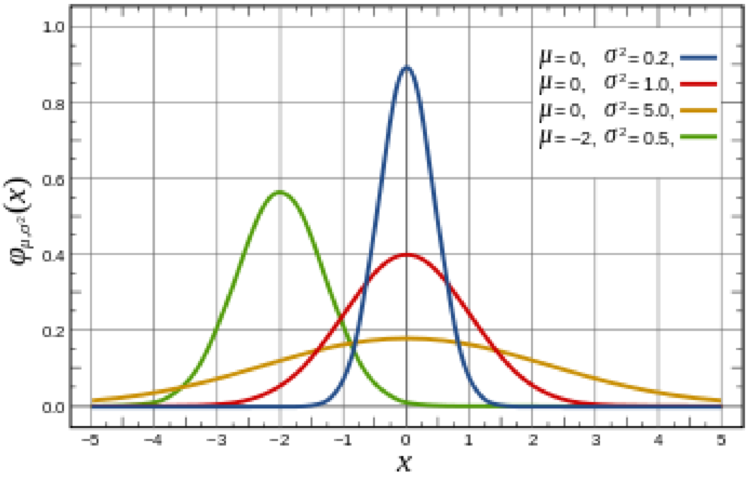

Figure 18: The Normal density function (from wikipedia)

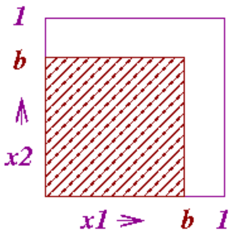

Figure 19: The cross-product of two uniform variables.

words, the plane $\mathbb{R}^2$. Whence, the proabability that $x \in I$ and $y \in J$ would be $\int_I f_1 dx \cdot \int_J f_2(y)dy$. Thus, the density function for the cross-product is merely $f(x,y) = f_1(x)f_2(y)$ with the outcome set $\mathbb{R} \times \mathbb{R}$.

Example 20 *Let us pick two random numbers uniformly between* 0 *and* 1*, say* x_1 *and* x_2*. Let* $x = \max(x_1, x_2)$*. What is the probability that* $0 \leq x \leq b$*? To solve this, let us look at the random variable* $z = (x_1, x_2)$ *where each* x *is uniform over* $[0,1]$*. Thus, the density function of* $z = (x_1, x_2)$ *is merely* $f(x_1, x_2) = f_1(x_1)f_2(x_2)$*, which is* $1 \cdot 1 = 1$*. Note that the function* f *is zero outside the unit square and that* $\int_0^1 \int_0^1 f(x_1, x_2)dx_1 dx_2 = 1$*.*

Next, we see that for the maximum of (x_1, x_2) *to be less than b, both* $x_1 \leq b$ *and* $x_2 \leq b$*, and thus, the probability of this event is* b^2*. See Fig 9 below.*

One common operation is a scale and translate of an existing random variable. Thus, for example, $Y = aX + b$, where $f(x)$ is the density function for X. In other words, $f(x)dx$ is the probability that X lies in the interval $[x, x + dx]$. Now, if $Y \in [y, y + dy]$ then $X \in [\frac{y-b}{a}, \frac{y-b}{a} + \frac{dy}{a}]$. Thus if $f_Y(y)$ is the probability density function of Y, then

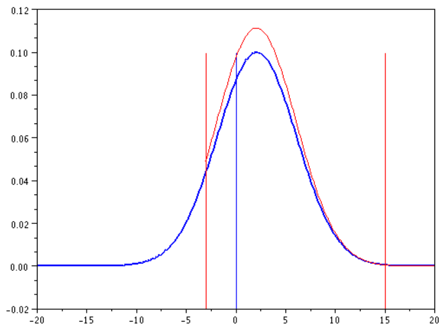

Figure 20: The temperature at Shimla as see by a thermometer

$f_Y(y) = \frac{1}{a} f(\frac{y-b}{a})$. We see for example, that

$$N(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-((x-\mu)^2)}{2\sigma^2}} = \frac{1}{\sigma} N(\frac{x-\mu}{\sigma}; 0, 1)$$

In other words, the $Y = N(\mu, \sigma)$ random variable is related to the the variable $X = N(0, 1)$ by $Y = \sigma X + \mu$.

Another common operation is **restriction**. Assume that X is a random variable with density function $f(x)$ and outcome set $S \subseteq \mathbb{R}$. Now consider the random variable Y, where Y only reports X if it lies in a sub-range $[a, b]$ of S. For example, Let X represent the temperature at Shimla on 1st of January over the years. However, our thermometer measures temperatures in the interval $[-3, 15]$ and reports an error if the temperature lies outside this interval. Let Y be the reported temperature by this thermometer, whenever an error *does not* occur. Thus Y is a restriction of X to the interval $[-3, 15]$. Now suppose that X was actually $N(2, 4)$, i.e., normal with mean 3 and standard deviation 4. What would be the density function of Y? If f_Y is the density function of Y, then clearly, it must be zero outside $[a, b]$. Next, it must mimic the *shape* of f within this interval, i.e., must be a multiple of f, i.e., $f_Y(x) = \alpha f(x)$ when $x \in [a, b]$, for a constant α. This is determined easily by requiring that $\int_a^b f_Y(x)dx = \alpha \int_a^b f(x)dx = 1$. Thus, we see that $\alpha = 1/\int_a^b f(x)dx$.

For our example, the Shimla temperature variable is shown in blue in Figure 9 below. The range $-3, 15]$ is marked in red. α turns out to be $1/0.896$ which is 1.11. Thus, f_Y is a scaled version of f in the interval $[-3, 15]$ and is shown in red.

10 Data and probability models

The basic use of probability models is to simulate real data and to predict the effect of certain interventions with a level of confidence. Here is a concrete example.

Example 21 *A literacy program was implemented in 120 revenue villages in the eastern part of Shahpur, which has a total of 222 revenue villages. The program entailed a teacher training program, introduction of new kits and so on. The program director wishes to a quick and economical mid-term appraisal of the program now that 1.5 years have elapsed. Please come up with a project plan for this task and list the technical outcomes.*

It is clear that this calls for understanding the status of the villages which were a part of the program and compare it with others in the taluka which were not. Next, perhaps, a sample of the 120 program villages will be taken up for a detailed (and expensive) survey. The selection of these villages is crucial to make a concrete assertion, with a level of confidence, on the success of the program. It is in this confidence assertion where known probability models become very useful, for here these calculations can be done *a priori* and a testing regime designed based on these assumptions.

The first task is of course, to check if the data that you have matches with some known probability density function. We shall briefly examine this question. The first point is to check that most standard density functions can be programmed on a computer and repeated trials generated. In other words, for any density function, we may produce a virtual cannon which will fire according to that density function. For the standard ones, such as *Binomial* or *normal*, Scilab provides ready-made function `grand` with appropriate arguments and inputs, see Section 18. Let us use `grand` to generate 200 random numbers distributed in the *Binomial* density function with $N = 100$ and $q = 0.6$. After generating this sample of 200, let us plot it as a histogram for a width of 2, i.e, $\{k, k+1\}$, for even k. Let us also plot the expected number of occurences, which will be $200 * (pr(k) + pr(k+1))$, where $pr(k)$ is the probability thaty the bionomial random variable of $q = 0.6$ and $N = 100$ will yield k. This combined plot is shown below in Fig. 10. We see fairly nice things in the plot that the number of actual outcomes fairly match with the predicted numbers. Moreover, the maximum is close to $60 = 0.6 * 100$.

We try it next with the normal density function with mean 1 and SD 0.6. We plot for 1200 trials and 200 trials as below in Fig. 10. We see the important outcome that as the number of trials increase, the observed numbers match with the predicted numbers much better.

We now consider the case of real data and checking if it matches known density functions. Let us start with the case of number of households per village in Murbad taluka. After

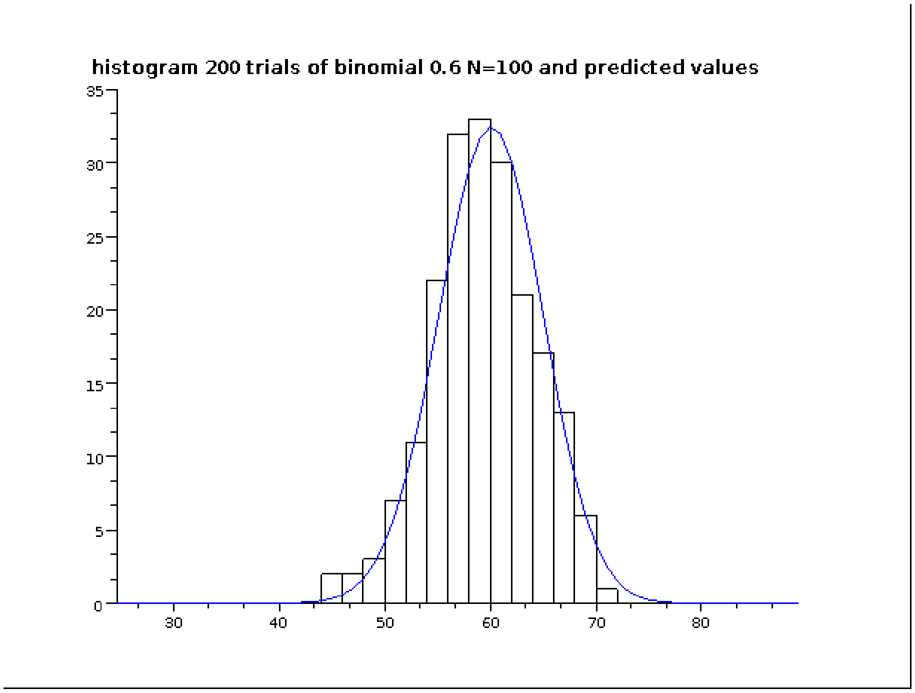

Figure 21: The binomial sample and expectation

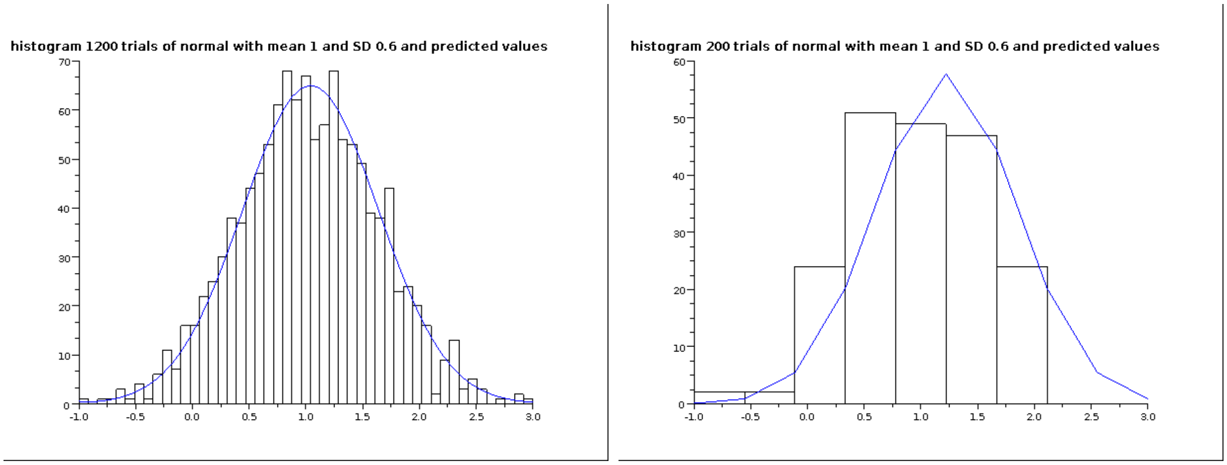

Figure 22: The normal trial and expectation

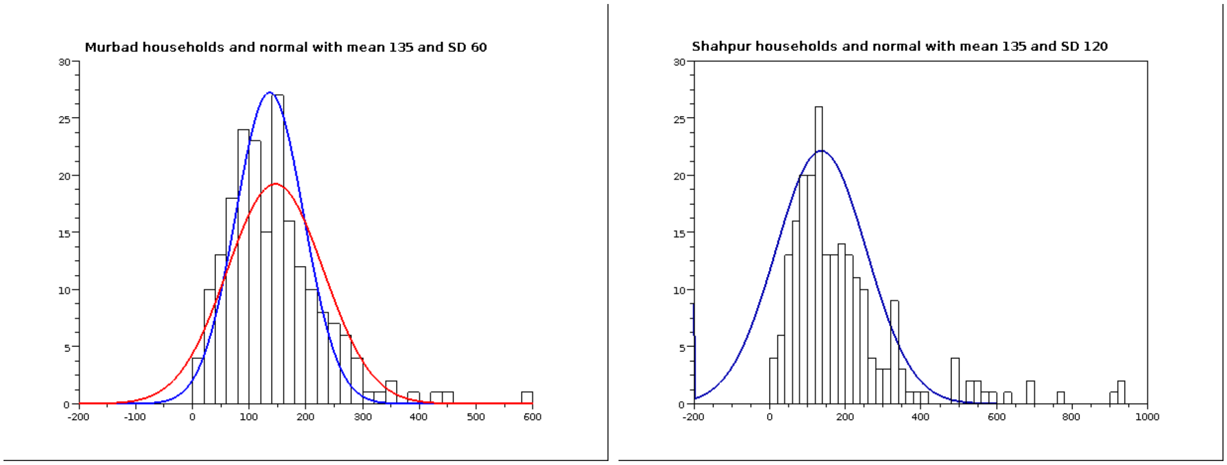

Figure 23: The normal fit to Murbad and Shahpur HH data

several attempts, we see that $N(135, 60)$, i.e., the normal denisty function with mean 135 and standard deviation 60 (plotted in blue) fits the data fairly well. The actual mean and SD of the data set are 145 and 85 respectively. We plot that in red. As we see, this is not as good for many reasons. Firstly, we see that the data naturally has a truncation effect, i.e., there cannot be any villages with negative number of households. This truncation also causes a change in the variation which is not very predictable. So, the question remain, *is the observed data from $N(135, 60)$ or not and with what confidence?* Such questions are important and are tackled through specific tests. One of them is the Kolmogorov-Smirnov test which we will discuss later. We also note that the Shahpur households dont quite fit the normal density function.

We may try the same with some other attributes. Below, in Fig. 10 we have the female literacy fraction for various villages of Shahpur. The mean and SD of the data are 0.428 and 0.136 respectively. This is plotted in blue. The best suited (according to my eyes) is with mean and SD 0.43 and 0.12 respectively. This is plotted in magenta. Of course, not all data sets are so *normalizable*. See for example, the ST-fraction for Shahpur. We see that far from being close to normal, it in fact shows bi-modal behaviour, i.e., with two peaks, at close to 0 and at close to 1. This indicates that Shahpur villages are fairly divided into those which are largely ST and those which are largely non-ST.

Example 22 *Write scilab code to obtain each of the above plots. Also, consider the question of verifying whether ST communities tend to have better sex-ratios than non-ST communities. How would you test the above proposition?*

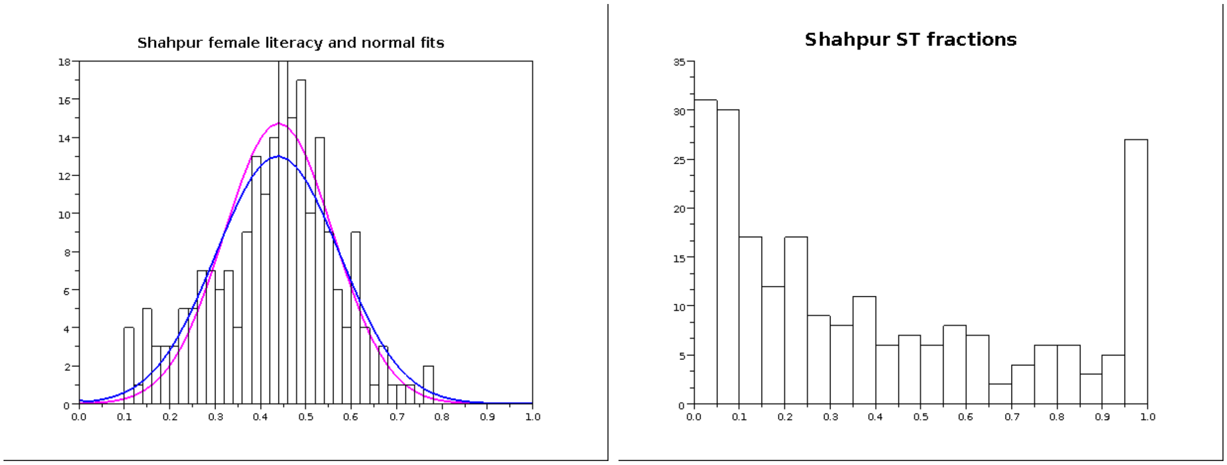

Figure 24: Shahpur female literacy fraction and ST fraction

11 Functions and expectation

In this section, we will delve deeper into the theory of random variables. For the purpose of this section, we will assume that the outcome set of our standard random variable is $\mathbb{R}$ and is given by density functions f and so on. In other words, for an interval I, we have $p(I) = \int_I f(x)dx$.

Frequently, we have a function $g : S \to \mathbb{R}$. This g may represent a value $g(s)$ that we attach to each outcome $s \in S$. For example, $S = \{HH, TH, HT, TT\}$, and $G(HH) = 4$ while $g(TT) = g(TH) = g(HT) = -1$. This may model the outcomes of a game of two coin tosses with two heads fetching Rs. 4 while any other outcome resulting in a loss of Rs. 1.

Definition 23 *Given such a function g on the outcomes of a random variable X, we define the* **expectation** *$E_X(g)$, or simply, $E(g) = \sum_s g(s)p(s)$, or as the integral $\int_S f(x)g(x)dx$.*

Example 24 *For the example above, for an unbiased coin, we have $p(HH) = p(HT) = p(TH) = p(TT) = 0.25$, whence $E(g) = 0.25$. Thus, the games is expected to benefit you Rs. 0.25 every time you play it.*

Example 25 *Let X be the uniform density function on $[0, 1]$ and let $Y = X \times X$. Thus $f_Y(x_1, x_2) = 1$ for all $x_1, x_2 \in I$. Let $g(x_1, x_2) = \max(x_1, x_2)$. Let us compute $E(g)$. We see that the set S may be divided into two halves along the diagonal. The first domain would be S_i where $x_1 \geq x_2$ and the other, where $x_2 \geq x_1$. Clearly*

$$E(g) = \int_S g(x_1, x_2)f(x_1, x_2)dx_1dx_2 = \int_{S_1} g(x_1, x_2)dx_1dx_2 + \int_{S_2} g(x_1, x_2)dx_1dx_2$$

By symmetry, both integrals must be equal and we evaluate the first one. We see that

$$\int_{S_1} g(x_1, x_2)dx_1dx_2 = \int_{x_1=0}^{1} \int_{x_2=0}^{x_1} x_1dx_1dx_2 = \int_{x_1=0}^{1} x_1^2dx_1 = 1/3.$$

Thus $E(g) = 2/3$. We should recall that the maximum of two uniform random variable is also a random variable Z with outcome set $[0, 1]$ and density function $2x$. In this case, $g(x) = x$ and the desired number of merely $E_Z(x)$ for the random variable Z. We see that $\int_{[0,1]} 2x \cdot xdx = 2/3$.

Let us note some elementary properties of expectation.

- $E(g_1 + g_2) = E(g_1) + E(g_2)$. This follows from the linearity of integration.
- If $Y = aX + b$ then $\mu_Y = a\mu_X + b$. This follows from the previous item above.

- If $Y = aX + b$, then $\sigma_Y^2 = a^2\sigma_X^2$. This is an honest calculation:

$$\begin{aligned}
\sigma_Y^2 &= \int f_Y(y)(y-\mu_Y)^2 dy \\
&= \tfrac{1}{a}\int f(\tfrac{y-b}{a})(y-\mu_Y)^2 dy \\
&= \int f(x)(ax+b-\mu_Y)^2 dx \text{ (after substituting } y = ax+b) \\
&= a^2\sigma_X^2
\end{aligned}$$

Definition 26 *The* **mean** μ_X *of a random variable X with outcome set contained in $\mathbb{R}$ is defines as $E(x)$, i.e., the expectation of the identity function $g(x) = x$. The quantity μ_X is a real number. The* **variance** σ_X^2 *is defined as $E((x-\mu_X)^2)$.*

Let us now compute the means and variances of the standard random variables.

- *Uniform.* Here $f(x) = 1$ on the outcome set $[0,1]$. We have $E(x) = \int_0^1 xdx = \left[\frac{x^2}{2}\right]_0^1 = 1/2$. This is expected. We have the variance as

$$\int_0^1 (x-\frac{1}{2})^2 dx = \left[\frac{(x-\frac{1}{2})^3}{3}\right]_0^1 = \frac{1}{12}$$

- *Binomial.* We have $p(k) = \binom{n}{k}q^k(1-q)^{n-k}$ and thus

$$\begin{aligned}
\mu &= \textstyle\sum_{k=0}^n k\cdot\binom{n}{k}q^k(1-q)^{n-k} \\
&= \textstyle\sum_{k=1}^n n\cdot\binom{n-1}{k-1}q^k(1-q)^{n-k} \\
&= nq\textstyle\sum_{j=0}^{n-1}\binom{n-1}{j}q^j(1-q)^{n-1-j} \\
&= nq
\end{aligned}$$

  This establishes the expected value nq as the mean. The variance is also similarly calculated and equals $nq(1-q)$.

- *Normal* $N(\mu,\sigma)$.By the linear combination result, we only need to prove this for $N(0,1)$, i.e., the standard normal. Now, $x\cdot\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}$ is an odd function, whence its integral must be zero. Thus the mean of the standard normal is indeed zero. The mildly harder case is the variance. We see this in the following steps:

$$\begin{array}{rcl}\sigma^2 & = & \int_{-\infty}^{\infty} x^2 \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx \\ & = & -\int_{-\infty}^{\infty} x \cdot \frac{d}{dx}(\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}) dx \\ & = & \left[-x \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx \\ & = & 1 \end{array}$$

Here is another expectation which is very important in the theory of random variables, esp. in repeated trials and the structure of the normal distribution.

Definition 27 *The* **moment generating function** *(mgf) $\phi_X(t)$ of the random variable X given by the density function f is $E(e^{tX}) = \int f(x)e^{tx}dx$.*

In fact, the mgf of a function f determines (more or less) determines it uniquely. We present three results on the transform.

- If X is normal with mean μ and variance σ^2 then $\phi_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$. We see this in the following steps:

$$\begin{array}{rcl}\phi_X(t) & = & \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \\ & = & \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{2\sigma^2 tx - (x-\mu)^2}{2\sigma^2}} dx \\ & = & \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-(\mu+\sigma^2 t))^2}{2\sigma^2}} e^{\frac{t^2\sigma^4 + 2\mu\sigma^2 t}{2\sigma^2}} dx \\ & = & e^{\frac{t^2\sigma^2}{2} + \mu t} \end{array}$$

- Suppose that X_1 and X_2 are independent random variables with density functions $f_1(x)$ and $f_2(x)$, and mgfs $\phi_1(s)$ and $\phi_2(s)$. Let $Y = X_1 + X_2$, then the density function f_Y is given by $f_Y(y) = \int_{-\infty}^{\infty} f_1(x)f_2(y-x)dx$. This is called the **convolution** of f_1 and f_2. This is readily seen by considering the random variable $X_1 \times X_2$ with density function $f_1(x_1)f_2(x_2)$. Let $F_Y(y)$ denote the probability that $x_1 + x_2 \leq y$. We see that:

$$\begin{array}{rcl}F_Y(y) & = & \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{y-x_1} f_1(x_1)f_2(x_2)dx_1dx_2 \\ & = & \int_{x_1=-\infty}^{\infty} f_1(x_1)dx_1 \int_{x_2=-\infty}^{y-x_1} f_2(x_2)dx_2 \end{array}$$

Now differentiating under the inetgrals gives us:

$$\begin{array}{rcl}f_Y(y) = \frac{d}{dy}(F_Y(y)) & = & \int_{x_1=-\infty}^{\infty} f_1(x_1)dx_1 \frac{d}{dy}\left[\int_{x_2=-\infty}^{y-x_1} f_2(x_2)dx_2\right] \\ & = & \int_{x_1=-\infty}^{\infty} f_1(x_1)dx_1 f_2(y-x_1) \end{array}$$

The mgf of $f_Y(y)$ is the product $\phi_Y(t) = \phi_1(t) \cdot \phi_2(t)$. This is seen by:

$$\begin{array}{rcl}
\phi_Y(t) & = & \int_{y=-\infty}^{\infty} e^{ty} \cdot \int_{x=-\infty}^{\infty} f_1(x) f_2(y-x) dx dy \\
& = & \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} e^{-sy} f_1(x) f_2(y-x) dy dx \\
& = & \int_{x=-\infty}^{\infty} e^{tx} f_1(x) \left[\int_{y=-\infty}^{\infty} e^{t(y-x)} f_2(y-x) dy \right] dx \\
& = & \int_{x=-\infty}^{\infty} e^{tx} f_1(x) \phi_2(t) dx \\
& = & \phi_1(t)\phi_2(t)
\end{array}$$

- If for $i = 1, \ldots, n$, the variables X_i are normal with mean μ_i and variance σ_i^2 then so is the variable $Y = X_1 + \ldots + X_n$ and it has mean $\sum_i \mu_i$ and variance $\sum_i \sigma_i^2$. This directly follows from the above two facts. We see that

  $$\phi_Y = \prod_i e^{\frac{t^2 \sigma_i^2}{2} + t\mu_i} = e^{\frac{t^2 \sum_i \sigma_i^2}{2} + t \sum_i \mu_i}$$

  This is clearly the transform of the normal random variable for the said mean and variance.

12 Repeated trials and normality

Let us now consider a random variable X and for $i = 1, \ldots, n$, let X_i be an independent trial of X. This corresponds to, e.g., a repeated firing of a cannon, or a sampling of a few villages of Murbad and so on. Let $Y = \sum_i X_i$ and $\overline{X} = \frac{\sum_i X_i}{n}$.

Lemma 28 *The mean μ_Y of Y equals $n\mu_X$ and its variance $\sigma_Y^2 = n \cdot \sigma_X^2$. For $\overline{X}$, we have $\mu_{\overline{X}} = \mu_X$ and $\sigma_{\overline{X}}^2 = \sigma_X^2 / n$.*

.

The linearity of expectation explains most things. The only calculation is the calculation of the variance of the sum C of two independent random variables, say A and B, which we do now.

$$\begin{aligned}
\sigma_C^2 &= E((c-\mu_C)^2) \\
&= E((a+b-\mu_A-\mu_B)^2) \\
&= E((a-\mu_A)^2)+E((b-\mu_B)^2)+2E((a-\mu_A)(b-\mu_B)) \\
&= \sigma_A^2+\sigma_B^2+\int_A\int_B f_A(a)f_B(b)(a-\mu_A)(\_\mu_B)dadb \\
&= \sigma_A^2+\sigma_B^2+\{\int_A f_A(a)(a-\mu_A)da\}\{\int_B f_B(b)(\_\mu_B)db\} \\
&= \sigma_A^2+\sigma_B^2
\end{aligned}$$

Thus, we see that the variance of the variable $\overline{X}$ diminishes with n while its mean remains invariant. This, in fact, is the basis of much of sampling. Let us try this in an example.

Example 29 *A team of CTARA students studied 12 randomly chosen villages of Shahpur. In that exercise, they observed the mean female literacy of the 12 villages to be 0.36. Given that the census data puts female literacy as normal with mean 0.43 and standard deviation 0.13, what is the probability that the mean of 12 independent samples should come out to be 0.36 or below?*

We see that $\overline{X} = \frac{X_1+\ldots+X_{12}}{12}$ should be normal with mean 0.43 and variance $0.13/\sqrt{12} = 0.038$. We see that $0.43 - 0.36$ is 0.07, i.e., $1.8 \cdot \sigma_{\overline{X}}$. We use `cdfnor(-1.8,0,1)` in Scilab to get 0.035. In other words there was a 3.5% chance that if the census data was correct, the team would have the above observations from 12 villages. Thus this puts into grave doubt either the census data or the methodology used by the team.

Consider next $Z_n = \frac{X_1+\ldots+X_n-n\mu_X}{\sigma_X\sqrt{n}}$, i.e., the sum of independenat repeated trials of a variable X scaled and translated by some constants. We see that $\mu_{Z_n} = 0$ and $\sigma_{Z_n}^2$ is $n\sigma_X^2/n\sigma_X^2 = 1$. Thus Z_n has mean 0 and variance 1.

Central Limit Theorem. For a wide class of random variables X, as $n \to \infty$, the variable Z_n approaches the standard normal $N(0,1)$. Thus, the simple repeated sum $\sum_{i=1}^{n} X_i$ also approaches the normal density function with mean $n\mu_X$ and variance $n\sigma_X^2$.

The good thing about the above theorem is that it applies to a wide variety and almost certainly to most commonly occuring density functions.

Let us conduct an experiment to verify the Central Limit Theorem. Let X be the simplest of all random variables, viz., with the uniform random variable with outcome set $[0,1]$. We see that $E(X) = 0.5$ and $\sigma_X^2 = 1/12$. Let us consider n trials and the variable

$$Z_n = \frac{X_1+\ldots+X_n-n\mu_X}{\sigma_X\sqrt{n}} = \frac{X_1+\ldots+X_n-n/2}{\sqrt{n/12}}$$

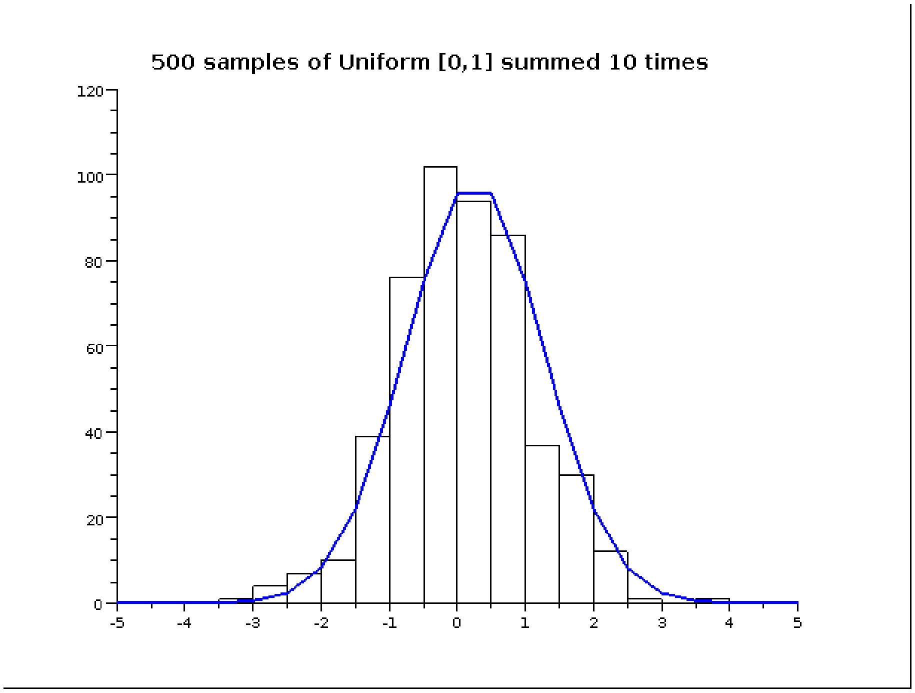

Figure 25: 500 trials of a 10-uniform-sum

We make 500 trials and plot the observed frequencies for $n = 10$, i.e., Z_{10}. The blue line is the expected frequencies for the normal curve. We see a close match.

Example 30 The basis for assuming normality in social data. *Scientists studied for Thane, the passing percentages of girls and boys in their school years and considered all factors such as economic conditions, social status, distance from school and so on, and came out with the following probability estimates for a girl/boy to pass the 10th standard exam:*

Xth passing		
	ST	*non-ST*
Boy	*0.13*	*0.33*
Girl	*0.21*	*0.26*

Next, consider the village of Dhasai with population structure given below. Let X be the random variable modelling the number of Xth standard pass adults.

Dhasai Adult Population		
	ST	*non-ST*
Male	*123*	*312*
Female	*133*	*286*

It is clear that if $X_{11}X_{12}, X_{13}, X_{14}$ are random variables expressing if a given boy/girl who is ST/non-ST is Xth pass, then X is merely the sum of repeated trials 123 copies of X_{11}, 312 copies of X_{12} and so on. Now if Y_{ij} are these repeated sums then the theorem says that each of these is close to being normal. Thus X, the sum of the Y_{ij}'s is also almost normal.

This settles the argument that the number of Xth pass (or its fraction) in Dhasai should be normal. However, it does not answer why should this quantity for another village Mhasa be distributed by the **same mean and variance** *as Dhasai. This is argued as follows. Suppose that the number of adults N_{ij} in Murbad taluka of various categories is known. Suppose next that a village has some n number of adults. Then we may assume that the composition of this village by various categories is obtained by n* **independent** *trials on the whole Murbad taluka population. If that is valid, then a further counting of Xth pass may proceed along earlier lines, giving an argument why the Xth pass fractions across all villages be distrbuted by a common normal random variable.*

This is partly the basis in assuming many of these social variables as normal. There are of course, serious limitations to this approach. First is the non-independence of many attributes of individuals with those in his/her village, community etc., as pointed out earlier. Secondly, as we saw in Shahpur the ST-fraction in villages is **not** *normally distributed. In fact, there is a divergence towards the extremes of 0 and 1. All the same, the literacy fractions do show some match with a common normal variable. This may be due to some other mechanisms at work which are common to both ST and non-ST.*

13 Estimation and Hypothesis testing

Let us now to the question of estimating a parameter of a random variable of a known type. The simplest example is when the elements of a population P may be divided into two disjoint parts, say A and B and we are required to estimate $q = |A|/|P|$. Standard examples include estimating the fraction of ST people in Murbad, literate people in a village and so on. Note that the parameter space for q is $Q = [0, 1]$ and we must estimate the correct q by conducting some experiment. The standard procedure would be to sample n items of P and count the number k of elements who actually belong to A. The the outcome set S of our experiment is $S = \{0, 1, \ldots, n\}$. Now we devise the **estimator** $e : S \to Q$ as $e(k) = k/n$. In other words, if there were k on n elements in A, then our estimate of q would be k/n. Let us try and understand this process in more detail, through an example.

Consider the situation when we have made 10 trials and observed 3 successes. For various possible values of q, let us calculate and plot the probability of the event of k successes happening. This is clearly the Binomial density function $Bin(q, 10)$ and computing

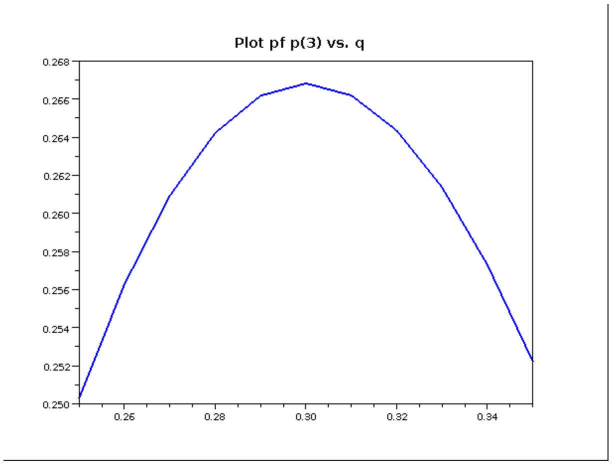

Figure 26: Estmating q when k=3 and n=10

$p(3) = \binom{10}{3} q^3 (1-q)^{10-3}$ for various values of q. The plot in Figure 13. We see that the probability of the event $k = 3$ is indeed maximized when $q = 0.3$, although the probability itself is only about 0.266. Moreover, for $q = 0.25$, the probability of the event $k = 3$ is about 0.25 which is not far from 0.266.

Let us first prove the simple fact that $q = k/n$ is indeed where the probability $p(k)$ is maximum. Let us differentiate $\binom{n}{k} q^k (1-q)^{n-k}$ and equate this to zero to obtain q.

$$\begin{aligned}
\tfrac{d}{dq}\left[\tbinom{n}{k} q^k (1-q)^{n-k}\right] &= 0 \\
\tbinom{n}{k}\left[kq^{k-1} - (n-k)(1-q)^{n-k-1}\right] &= 0 \\
kq^{k-1}(1-q)^{n-k} - (n-k)q^k(1-q)^{n-k-1} &= 0 \\
k(1-q) - (n-k)q &= 0 \\
k - nq &= 0
\end{aligned}$$

Thus $q = k/n$ is where the derivative is zero. It is easy to check that this is a maxima. Thus our function $e : S \to Q$ with $e(k) = k/n$ actually estimates a q such that the probability of the outcome k is maximized. Such an estimator is called the parameter $q \in Q$ is called as the **maximum likelihood estimator** of q.

The next matter is of **confidence**. Suppose that, a priori, we had no guidance on the possible values of q and that every $q \in [0, 1]$ was equally possible. We then plot $p(k)$ for all values of $q \in [0, 1]$. This is plotted in Fig. 13. We may well assert that $q = 0.3$, but there is no reason to doubt that $q = 0.28$, in fact. Let us quantify our assertion that $q = 0.3$ by looking at the area under the curve in the interval $[0.25, 0.35]$. We see that this is roughly 31% of the total area. Thus, assuming that all values of q were equally likely, we may assert that we have 31% confidence in our assertion.

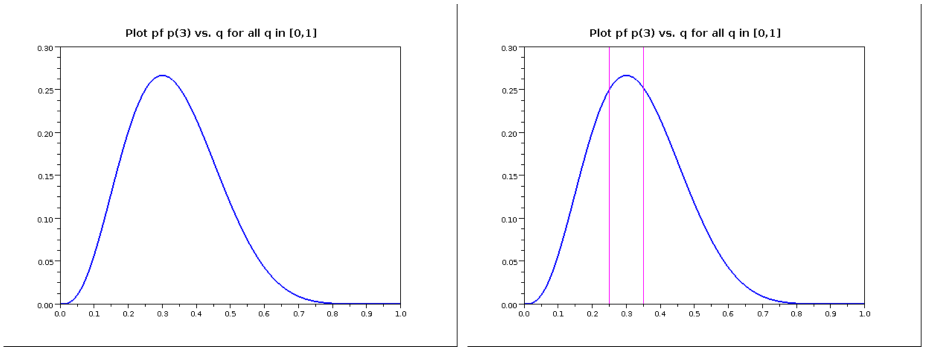

Figure 27: p(3) for all q and the confidence interval $[0.25, 0.35]$

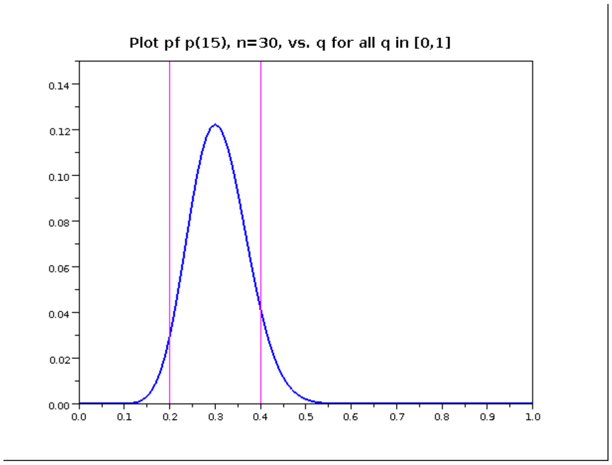

Figure 28: p(16) and n=50 for all q and the confidence interval $[0.2, 0.3]$

How do we strengthen our assertion? The first option is to widen the interval. For example, we check that for the interval $[0.2, 0.4]$ we have a larger confidence of 56%. The other, and wiser, option is to increase the number of trials. Suppose now that $n = 50$ and $k = 15$ and thus $q = 0.3$. Thus the estimated value remains the same. However, the q-plot changes dramatically, as seen in Fig. 13. Also, now the confidence in the interval $[0.2, 0.4]$ goes to roughly 91%.

All of this crucially depends on the fact that all $q \in [0, 1]$ were equally likely. Suppose, a priori, we knew that q is in fact in the interval $[0.2, 0.8]$. In which case, our confidence in our assertion would increase to $area(0.2, 0.3)/area(0.2, 0.8)$ which is 93%. In general, we have a general *a priori* probability density f on $[0, 1]$ for q. In such a situation, the confidence for the interval $[a, b]$ when we have observed k successes for n trials would be:

$$\frac{\int_a^b f(q)\binom{n}{k}q^k(1-q)^{n-k}dq}{\int_0^1 f(q)\binom{n}{k}q^k(1-q)^{n-k}dq}$$

Such an analysis is called a **Bayesian** analysis since it bases its estimate of q by conditioning on the case for each $q \in [0,1]$.

13.1 Bayesian Estimation

13.1.1 Conjugate prior

Suppose we have a multivariate bernoulli distribition of the μ's and let $Pr(\mu = \mu_1) = Pr(\mu = \mu_2) = \ldots Pr(\mu = \mu_k) = \frac{1}{k}$. As an example condider the toss of a dice. Suppose at ∞, all observations are say a particular value V_i then, we will have $Pr(\mu = \mu_1) = 0, \ldots Pr(\mu = \mu_i) = 1 \ldots Pr(\mu = \mu_k) = 0$

Using Bayes rule

$$Pr([\mu_1 \ldots \mu_k]|D) = \frac{Pr(D|\bar{\mu})Pr(\bar{\mu})}{\sum_{\bar{\mu'}} Pr(D|\bar{\mu'})Pr(\bar{\mu'})}$$

If $Pr(\mu)$ has a form such that $Pr(\mu|D)$ has the same form, we say that $Pr(\mu)$ is the *conjugate prior* to the distribution defining $Pr(D|\mu)$.

Some of the conjugate priors that we will see are

Dirichlet and Multivariate Bernoulli

Beta and Bernoulli

Gaussian and Gaussian

13.1.2 Dirichlet prior

Prior $P(\mu) \in [0,1] \& \int_0^1 p(\mu)d\mu = 1$
Now, $P(\mu|X_1, X_2, ..X_n) = \frac{P(X_1,X_2,..X_n|\mu).P(\mu)}{\int_0^1 P(X_1,X_2,..X_n|\mu).P(\mu)d\mu}$
$= \frac{\mu^{n_1}(1-\mu)^{n_2}}{\int \mu^{n_1}(1-\mu)^{n_2}d\mu}$
$= \mu^{n_1}(1-\mu)^{n_2}\frac{(n_1+n_2+1)!}{n_1!n_2!}$
HW. If $P(\mu)$ had the form $\frac{\mu^{a-1}(1-\mu)^{b-1}(a=b-1)!}{(a-1)!(b-1)!}$ what will be the form of $P(\mu|D)$? $P(\mu|X_1, X_2, ..X_n) = \frac{\mu^{a-1+n_1}(1-\mu)^{b-1+n_2}(n_1+n_2+a+b-1)!}{(n_1+a-1)!(n_2+b-1)!}$ $Beta(\mu|a+n_1, b+n_2)$
so $P(\mu = 1)$ $Beta(\mu|1,1)$ Expected value of Beta $E(Beta(\mu|a,b)) = \frac{a}{a+b}$
Why it is reasonable:

- $E[\mu]_{Beta(\mu|a,b)} = \frac{a}{a+b}$ is intuitive
- a=b=1 gives uniform distribution
- Form of the posterior and prior are the same.

- As $n_1 + n_2 \leftarrow \infty$, spread of the distribution $\leftarrow 0$, a and b becomes immaterial. $(\frac{n_1+a}{n_1+n_2+a+b} = \frac{n_1}{n_1+n_2})$
  $\underset{B(\mu|n_1+a,n_2+b)}{E(\mu)} \leftarrow \mu_{\hat{ML}}$
  $\mu_{\hat{MAP}} = \underset{\mu}{argmax} P(\mu|D) = \frac{a+n_1-1}{a+n_1+b+n_2-2}$
  As $n_1, n_2 \leftarrow \infty, \mu_{\hat{MAP}} = \mu_{\hat{ML}}$

$P(X_{n+1}, ..X_{n+t}|\mu) = \prod_{i=1}^{t} P(X_{i+1}|\mu)$
$P(X_{n+1}|X_1..X_n) = \int P(X_{n+1}|\mu)P(\mu|X_1..X_n)d\mu = \int \mu^{X_n+1}(1-\mu)^{1-X_{n+1}} P(\mu|X_1..X_n)d\mu$
$= E[\mu]$ if $X_{n+1} = 1$
$= 1 - E[\mu]$ if $X_{n+1} = 0$
$Beta_{(\mu_1,\mu_2|a,b)} = \frac{\mu_1\mu_2(n_1+n_2-1)!}{(a-1)!(b-1)!}$
$Dir(\mu_1, \mu_2, ..\mu_k|a_1, a_2, ..a_k) = \frac{\prod_i \mu_i^{a_i-1}\sqrt{\sum_j a_j}}{\prod_j \sqrt{a_j}}$
$E(\mu)_{Dir} = [\frac{a_1}{\sum_i a_i} \frac{a_2}{\sum_i a_i} .. \frac{a_k}{\sum_i a_i}]^T$
Expression for $\mu_{\hat{Bayes}} = E(\mu) = [\frac{a_1+n_1}{\sum_i a_i+n_i} \frac{a_2+n_2}{\sum_i a_i+n_i} .. \frac{a_k+n_k}{\sum_i a_i+n_i}]^T$
Expression for $\mu_{\hat{ML}} = E(\mu) = [\frac{n_1}{\sum_i n_i} \frac{n_2}{\sum_i n_i} .. \frac{n_k}{\sum_i n_i}]^T$
Expression for $\mu_{\hat{MAP}} = E(\mu) = [\frac{a_1+n_1-1}{(\sum_i a_i+n_i)-K} \frac{a_2+n_2-1}{(\sum_i a_i+n_i)-K} .. \frac{a_k+n_k-1}{(\sum_i a_i+n_i)-K}]^T$
$P(X_{n+1}|X_1..X_n) = [E(\mu)]$ if $X_{n+1} = V_i$
$Dir_{l,j}(\mu^l_{1_{ij}}..\mu^l_{l_{ij}}|a^l_{1_{ij}}..a^l_{1_{ij}}) = \frac{\prod_i (\mu^l_{i,j})^{a^l_{i,j}-1}}{\prod_i \Gamma(a^l_{i,j})}\Gamma(a^l_{i,j} + (a^l_{2,j}) + ..(a^l_{k,j})$
$E(\mu)_{Dir} = [\frac{a^l_{1,j}}{\sum_i a^l_{i,j}} \frac{a^l_{k,j}}{\sum_i a^l_{i,j}} .. \frac{a^l_{k,j}}{\sum_i a^l_{i,j}}]^T$
$(\mu^l_j)_{Bayes} = [\frac{a^l_{1,j}+n^l_{1,j}}{\sum_i a^l_{i,j}+n^l_{i,j}} \frac{a^l_{2,j}+n^l_{2,j}}{\sum_i a^l_{i,j}+n^l_{i,j}} .. \frac{a^l_{k,j}+n^l_{k,j}}{\sum_i a^l_{i,j}+n^l_{i,j}}]^T$
$(\mu^l_j)_{MAP} = [\frac{a^l_{1,j}+n^l_{1,j}}{\sum_i(a^l_{i,j}+n^l_{i,j})-K_l} \frac{a^l_{2,j}+n^l_{2,j}}{(\sum_i a^l_{i,j}+n^l_{i,j})-K_l} .. \frac{a^l_{k,j}+n^l_{k,j}}{(\sum_i a^l_{i,j}+n^l_{i,j})-K_l}]^T$
$(\mu^l_j)_{Bayes} = [\frac{a^l_{1,j}+n^l_{1,j}}{\sum_i a^l_{i,j}+n^l_{i,j}} \frac{a^l_{2,j}+n^l_{2,j}}{\sum_i a^l_{i,j}+n^l_{i,j}} .. \frac{a^l_{k,j}+n^l_{k,j}}{\sum_i a^l_{i,j}+n^l_{i,j}}]^T$

Assume X is the event of a coin toss. Let X1=0 (TAILS say), X2=1, X3=0, X4=1, X5=1. We are interested in predicting the event X6=1 given the above. This can be calculated by different approaches. The ML, MAP and the Bayes Estimator are called the pseudo Bayes, and Bayesian estimator is called the pure Bayes.

Maximum likelihood

$\mu_{\hat{ML}}$ is the probability of $X = 1$ from the data.

$$P(X6|X1..X5) = \mu_{\hat{ML}}^{X6}(1 - \mu_{\hat{ML}})^{(1-X6)}$$

MAP

$\mu_{\hat{MAP}}$ is the probability of $X = 1$ from the data.

$$P(X6|X1..X5) = \mu_{\hat{MAP}}{}^{X6}(1 - \mu_{\hat{MAP}})^{(1-X6)}$$

Bayes Estimator

$\mu_{\hat{bayes}}$ is the probability of $X = 1$ from the data.

$$P(X6|X1..X5) = \mu_{\hat{bayes}}{}^{X6}(1 - \mu_{\hat{bayes}})^{(1-X6)}$$

Bayesian method

$$P(X6|X1..X5) = \int_0^1 \mu^{X6}(1-\mu)^{(1-X6)}P(\mu|X1..X5)\mathrm{d}\mu$$

The explanation for this equation is as follows:

$$P(X6|X1..X5) = \int \frac{P(X6|\mu, X1, ..X5)P(\mu|X1..X5)P(X1..X5)\mathrm{d}\mu}{P(X1..X5)}$$

this marginalises on the probability μ. Simplifying futher,

$$P(X6|X1..X5) = \int P(X6|\mu, X1, ..X5)P(\mu|X1..X5)\mathrm{d}\mu$$

Thus

$$P(X6|X1..X5) = \int_0^1 \mu^{X6}(1-\mu)^{(1-X6)}P(\mu|X1..X5)\mathrm{d}\mu$$

13.2 Hypothesis Testing

13.2.1 Basic Idea

Let us now turn the tables and assume that a claim has been made, say that $q = q_0$. It is our task to check the validity of the claim. Such a claim is called the ***null hypothesis*** and is denoted by H_0. Our task is to design an experiment with outcome set S and based on the outcome, either reject or accept the hypothesis. There are clearly two types of error we can make and this is given in the table below:

H_0	Our assertion	Type of Error
True	False	Type I
False	True	Type II

Our strategy will be as follows. We will design an experiment and specify an event set $E_0 \subseteq S$. If the outcome of the experiment $o \in E_0$ then we assert with some confidence that H_0 is false. This takes care of Type I errors of labelling something as false when it was actually true. Now consider Type II errors. We construct another hypothesis H_1 so that both H_0 and H_1 cannot simultaneously hold. For H_1 we construct an event $E_1 \subseteq E_0$ such that if the outcome $o \in E_1$ then we can claim with confidence that H_1 is true. Since H_1 is true, H_0 is certainly false, and we would have concluded from our experiment that H_0 is false. Thus, the correct task is to design the experiment so that if H_0 were false then E_1 should be as large as possible.

Thus, the task is to design an experiment and an event E_0 and conduct the experiment. Next, we observe the outcome o. Based on whether the outcome $o \in E_0$ or not:

- for a fixed and small α conclude that H_0 is **false** with a confidence $1 - \alpha$.
- produce another hypothesis H_1 and an event $E_1 \subseteq E_0$ which contradicts H_0 and a small number β, $o \in E_1$ asserts that H_1 holds with confidence $1 - \beta$.
- remain silent and plan for further experimentation.

Let us suppose that the null hypothesis is $H_0 \equiv q_0 = 0.4$. We are now supposed to built an event set E_0 which will help us refute the hypothesis. Let us suppose that we intend to conduct 100 trials and observe k, the number of successes. Thus $S = \{0, 1, \ldots, 100\}$. We see that if the hypothesis is true then the sum $\sum_{i=30}^{50} B(100, 0.4)(i) = 0.96$, thus we chooose E_0 as the event set $[0, 29] \cup [51, 100] \subseteq S$, and $\alpha = 5\%$. Clearly if the outcome $o \in E_0$, then we can reject the claim H_0 with confidence $1 - \alpha$, for if the hypothesis were true than $o \in E_0$ is a very unlikely event. Next we set $E_2 = [0, 20]$, $\beta = 1\%$ and H_1 as the hypothesis that $q_0 < 0.35$. We see that if for example, the outcome is 20, then using our earlier theory of estimatation, we can claim with 99% confidence (check this) that $q_0 < 0.35$. If the outcome is lower than 20, then the confidence in fact strengthens. Thus we have:

- $H_0 \equiv q_0 = 0.4$, $E_0 = [0, 29] \cup [51, 100]$ and $\alpha = 5\%$.
- $H_1 \equiv q_0 < 0.35$, $E_1 = [0, 20]$ and $\beta = 1\%$.
- However, if the outcome is in the set $[30, 50]$, i.e., the complement to $E_0 \cup E_1$, then we are forced to remain silent.

What do we do when $o \in [30, 50]$? Well we could conduct a fresh experiment with an additional 900 trials to get a total of 1000 trials. We see that the set E_0 in fact swings closer to the the number $0.4n$ and the forbidden set, where we cannot draw any conclusion becomes smaller. In fact, for $n = 1000$, the inconclusive set becomes $[0.37n, 0.43n]$.

13.2.2 Hypothesis Testing: More formally

Given a random sample $(X_1, X_2, \dots X_n)$ for a random variable X, the function $S : (X_1, X_2, \dots X_n) \rightarrow R$ is called *statistic* (or *sample statistic*). For instance, the *sample mean* $\frac{\sum X_i}{n}$ and sample variance $\frac{\sum_i \left(X_i - \frac{\sum X_i}{n}\right)^2}{n-1}$ are statistics.

Let us consider hypotheses H_0 and H_1 defined by:

$$H_0 : (X_1, X_2, \dots X_n) \in C$$

$$H_1 : (X_1, X_2, \dots X_n) \notin C$$

where, C is some tolerance limit also called the confidence region. The C is generally defined in terms of some statistic S.

The following types of errors are defined as a consequence of the above hypotheses:

- Type I error: Probability of rejecting H_0, if H_0 was actually true.
  This is given by: $Pr_{H_0}(\{X_1, X_2, \dots X_n\} \notin C)$
- Type II error: Probability of accepting (or not-rejecting) H_0, if H_0 was actually false.
  This is given by: $Pr_{H_1}(\{X_1, X_2, \dots X_n\} \in C)$

Given a significance level $\alpha \in [0, 1]$ (some bound on the Type I error[10]), we want to determine a C such that,

$$Pr_{H_0}(\{X_1, \dots X_n\} \notin C) \leq \alpha \qquad \text{Type I error}$$

Here, C is the set of all possible "interesting" random samples. Also,

$$Pr_{H_0}(\{X_1, \dots X_n\} \notin C^{'}) \leq Pr_{H_0}(\{X_1, \dots X_n\} \notin C) \qquad \forall C^{'} \supseteq C$$

Thus, we are interested in the "smallest" / "tightest" C. This is called the critical region C_α. Consequently,

$$Pr_{H_0}(\{X_1, \dots X_n\} \notin C_\alpha) = \alpha$$

14 The abstract estimation problem

The abstract estimation problem is the following. Let X be a random variable with its density function $f(q; x)$, depending on a parameter from the set Q. We design an experiment with

[10] We can defer discussion of the Type II error, i.e. $Pr_{H_1}(\{X_1, \dots X_n\} \in C)$.

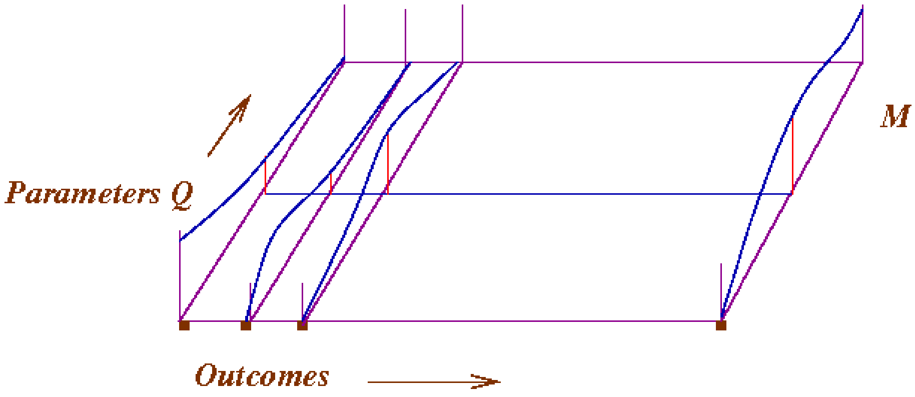

Figure 29: The matix M of Q vs. Outcomes

outcome set S. In our earlier case, $Q = [0,1]$ and $S = \{0, 1, \ldots, n\}$. We construct a $Q \times S$ *matrix* M where $M(q,x) = f(q;x)$. We see that each row of the matrix M, i.e., when $q \in Q$ is fixed, is merely the density function. While, for a fixed outcome $x \in S$, we see the dependence of the parameter on the oucome x. For our example, we see that for the outcome k, the column function is a smooth function with variable q, while the row function is the discrete probability $Bin(q,n)$ with a discrete outcome set S.

For the problem of estimation, since the outcome of the experiment is known, it is the column function which assumes importance. Thus, for Type II error analysis, the column function must be understood. The Type I error analysis is about a particular hypothesis on the parameter and thus it is the row function, i.e., the ordinary density function which must be understood.

15 The mean of a normal distribution with known variance

Suppose next that X is a normal variable with an unknown mean but with a known variance σ. The first question is of course, to ask where do such situations arise? These arise when an *additive* intervention is made on a subset A of a normal population whose mean and variance is known. It is expected that the mean of the members of A shifts to an unknown new mean.

Example 31 *The government decides to impose an additional tax of Rs. 400 per tonne of steel. Consequently, while some of the tax is absorbed by the industry, the remaining part is passed on to the consumer. Given the price of steel in open market as a time series, estimate the fraction which was passed on to the consumer.*

This is possibly an example where the mean and the variance of the price data is a normal *random variable. By observing this before the intervention, this old μ and σ may be accurately estimated. The economic mechanism suggests that the tax will merely cause a shift in the mean price from μ to $\mu + \delta$ without affecting σ.*

Example 32 *Karjat tribal block is a fairly homogenous sub-taluka of about 200 habitations with child literacy fraction normally distributed with mean $\mu = 0.68$ and $\sigma = 0.14$. Since distances to school coould be an important factor, an intervention was designed to serve a region of about 120 habitations by school rickshaws. The mechanism of literacy suggests that the intervention will move σ without significantly changing σ.*

Our task is to estimate μ of an unknown normal random variable X with known variance σ^2. We define our experiment as an n-way repeated trial with the outcome set $X_1 \times X_2 \times \ldots \times X_n$. The parameter set $Q = \mathbb{R}$ is the set of possible μ values, i.e., the set of real numbers. We define the estimator

$$e : X_1 \ldots \times X_n \rightarrow \mathbb{R}$$

$$e(x_1, \ldots, x_n) = \frac{x_1 + \ldots + x_n}{n}$$

Note that this is merely the mean of the observations. We see that if each X_i were indeed independent normal $N(\mu, \sigma)$ then the expectation $E(e)$ would merely be $\frac{n \cdot \mu}{n} = \mu$. Thus the estimator is **unbiased**, i.e., its expected value is indeed the correct value, if there is one.

We will next show that it is also a **maximum likelihood estimator**. To see this, the probability of an n-observation sitting within $[x_1, x_1 + \delta] \times \ldots \times [x_n, x_n + \delta]$ is proportional to $f(x_1) \cdot \ldots \cdot f(x_n)$, where $f(x) = \phi(\mu, \sigma; x)$, the normal density function. We may write this as:

$$\begin{aligned} Pr([x_1, x_1 + \delta] \times \ldots \times [x_n, x_n + \delta]) &= f(x_1) \cdot \ldots \cdot f(x_n) \delta^n \\ &= (\tfrac{1}{\sigma\sqrt{2\pi}})^n e^{\frac{-\sum_i (x_i - \mu)^2}{2\sigma^2}} \delta^n \end{aligned}$$

Now let us assume that σ and δ are fixed, and $x_1, \ldots, x_n$ are given observations, and that we would like to determine the best possible μ which will maximize the RHS. Next, we see that the RHS is maximized iff its log is maximized. But the log of the RHS as a function of μ, and upto constants, is merely $\sum_i -(x_i - \mu)^2$. Thus the RHS is maximized when $\sum_i (x_i - \mu)^2$ is minimized. This is easily seen by choosing $\mu = \frac{\sum_i x_i}{n}$. This proves that $e(x_1, \ldots, x_n) = \frac{\sum_i x_i}{n}$ is indeed the maximum likelihood estimator.

Let us denote $\frac{\sum_i x_i}{n}$ as $\overline{x}$, i.e., the observation, while $\frac{\sum_i X_i}{n}$ by $\overline{X}$, the random variable. We know that $\overline{X}$ is also normal with mean μ and variance $\overline{\sigma}^2 = \sigma^2/n$. The decrease in the variance of $\overline{X}$ is the key. We see right away that if μ were the unknown mean and $\overline{x}$ was the observation, then the abstract matrix M has

$$M(\mu, \overline{x}) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(\overline{x}-\mu)^2}{2\overline{\sigma}^2}}$$

Thus, both the row and the column functions have the same behaviour, which makes things much easier. We see that:

$$Pr(\mu - 2\overline{\sigma} \leq \overline{x} \leq \mu + 2\overline{\sigma}) \geq 0.95$$

We may rearrange this (using our observation on M) to get:

$$Pr(\overline{x} - 2\overline{\sigma} \leq \mu \leq \overline{x} + 2\overline{\sigma}) \geq 0.95$$

Example 33 *Suppose that X is the random variable denoting the child literacy in a village of Karjat tribal block. Suppose that it is known to be normal with an unknown mean but a known $\sigma = 0.14$. Suppose a team visits 10 villages and finds $\overline{x} = 0.76$. (i) What is the assertion we can make with 99%, 95% and 90% confidence? (iii) Suppose we have prior belief that the child literacy rate is around 0.5, how do we incorporate this prior information (ii) Suppose that an expert asserts that $\mu = 0.68$. With what confidence can you refute the claim?*

Let us solve (i) first. Firstly, we see that the effective standard deviation is only $0.14/\sqrt{10} = 0.044$. Next, We see that for a both-sided interval around 0.76, using `cdfnor`, we see that the intervals as a multiple $k\overline{\sigma}$, we have $k(0.99) = 2.58, k(0.95) = 1.96$ and $k(0.9) = 1.65$. Thus, we see that these intervals are $[0.65, 0.87]$, $[0.67, 0.85]$ and $[0.69, 0.83]$.

For (ii), we need to make use of Gaussian prior (which happens to be conjugate for mean of a Gaussian) on μ such that the mean of the gaussian prior is 0.5. Refer to the handwritten class notes for more details.

For (iii), we see that $(0.76 - 0.68)/0.044 = 1.82$. Again, using `cdfnor`, we see that the event of $\overline{x} = 0.76$, assuming that $\mu = 0.68$ is in the (one-sided) 4% and lower. Thus, we refute the claim with 96% confidence.

16 The variance of a normal distribution

Our next situation is to estimate the variance of a random variable which we know is normal. This arises frequently in engineering, pollution, ethnography and so on. Before we go on, we need to understand a new density function called the **chi-squared** density function which

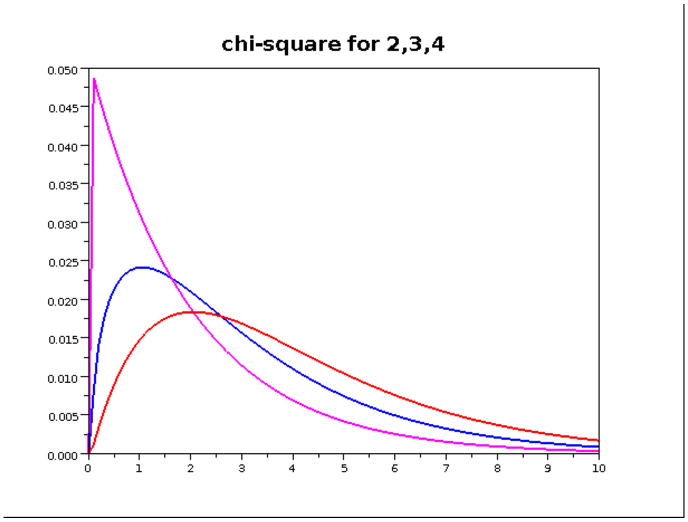

Figure 30: The χ_n^2 density function for various n

has a parameter n and is denoted by χ_n^2. This arises most commonly as the square of the distance of a random point. Let $X_1, \ldots, X_n$ be independent normal random variables with mean 0 and variance 1, i.e., $N(0,1)$. Let $Y = X_1^2 + \ldots + X_n^2$, then χ_n^2 is the density function of Y. Clearly $E(Y) = \sum_i E(X_i^2) = n \cdot 1 = n$. See the plots below in Fig. 16 (use `cdfchi ("PQ",x,n*ones(1,m))`).

Again, we make n trials $X_1, \ldots, X_n$ to obtain samples $x_1, \ldots, x_n$ and the sample mean $\overline{x} = (\sum_i x_i)/n$. The estimator of the variance is $S^2 = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}$. The curious term is of course, the denominator. To understand this, let us look at a related summation as a function on $X_1, \ldots, X_n$ (where μ is the unknown mean).

$$\begin{aligned}
\textstyle\sum_i (X_i - \mu)^2 &= \textstyle\sum_i ((X_i - \overline{X}) + (\overline{X} - \mu))^2 \\
&= \textstyle\sum_i (X_i - \overline{X})^2 + \sum_i (\overline{X} - \mu)^2 + 2\sum_i (X_i - \overline{X})(\overline{X} - \mu) \\
&= \textstyle\sum_i (X_i - \overline{X})^2 + \sum_i (\overline{X} - \mu)^2 + 2(\overline{X} - \mu)\sum_i (X_i - \overline{X}) \\
&= \textstyle\sum_i (X_i - \overline{X})^2 + \sum_i (\overline{X} - \mu)^2 \\
&= \textstyle\sum_i (X_i - \overline{X})^2 + n \cdot (\overline{X} - \mu)^2
\end{aligned}$$

Taking expectations on both sides, we see that:

$$n\sigma^2 = E(\sum_i (X_i - \overline{X})^2) + n \cdot \frac{\sigma^2}{n}$$

Thus, we see that $E(\sum_i (X_i - \overline{X})^2) = (n-1)\sigma^2$, and thus $E(S^2) = \sigma^2$. Thus, S^2 is an **unbiased estimator**.

Lets start with the last equality:

$$\sum_i (X_i - \mu)^2 = \sum_i (X_i - \overline{X})^2 + n \cdot (\overline{X} - \mu)^2$$

and divide everything by σ^2 to obtain:

$$\sum_i \left(\frac{X_i - \mu}{\sigma} \right)^2 = (n-1)\frac{S^2}{\sigma^2} + \left(\frac{\overline{X} - \mu}{\sigma\sqrt{n}} \right)^2$$

Since the LHS is a variable of density χ^2_n and the second term of the RHS χ^2_1, by a leap of faith, the variable $(n-1)\frac{S^2}{\sigma^2}$ is distributed by the χ^2_{n-1} density function, i.e., a known density function. Note that this does not need us to assume knowledge of μ at all. Let us now apply this in an example.

Example 34 *A sample of* 10 *fractional literacy levels in 10 villages was the sequence*

$$[0.82, 0.73, 0.70, 0.69, 0.67, 0.56, 0.45, 0.44, 0.43, 0.43]$$

Give 90% and 99% confidence interval estimates for σ^2*. With what confidence will you refute the claim that the SD is* 0.1*?*

We see that $S^2 = 0.0217$*. The variance is* 0.0195 *and the sample SD is* 0.140*. Since* $n = 10$*, we are dealing with* χ^2_9 *with expected value* 9*. We will find intervals* $[a, b]$ *around* 9 *such that* $Pr_{\chi^2_9}([a, b]) = 1 - \alpha$ *for* $\alpha = 0.1$ *and* 0.01*. We use* `cdfchi("PQ",x,9*ones(1,m))` *and get these intervals as* $[3.3, 18.9]$ *and* $[1.8, 24]$*. Thus, we see that:*

$$\begin{aligned} Pr(3.3 \leq 9 \cdot \tfrac{0.0217}{\sigma^2} \leq 18.9) &= 0.9 \\ Pr(2.727 \geq \tfrac{\sigma^2}{0.0217} \geq 0.476) &= 0.9 \\ Pr(1.651 \geq \tfrac{\sigma}{0.147} \geq 0.69) &= 0.9 \\ Pr(0.242 \geq \sigma \geq 0.101) &= 0.9 \end{aligned}$$

Thus, we can claim with 90% confidence that σ *lies in the interval* $[0.101, 0.242]$*. A similar (but larger) interval may be found for our 99% confidence assertion.*

Next, we move to refuting the $H_0 \equiv \sigma = 0.1$*. We see that* $9 \cdot \frac{0.0217}{0.01} = 1.953$*.* `cdfchi("PQ",1.953,9)` *gives the answer* 0.0078*, which is outside 1%. Thus, the observed* S^2 *is outside the 1% chance and thus we can claim with 99% confidence that* $\sigma = 0.1$ *is false.*

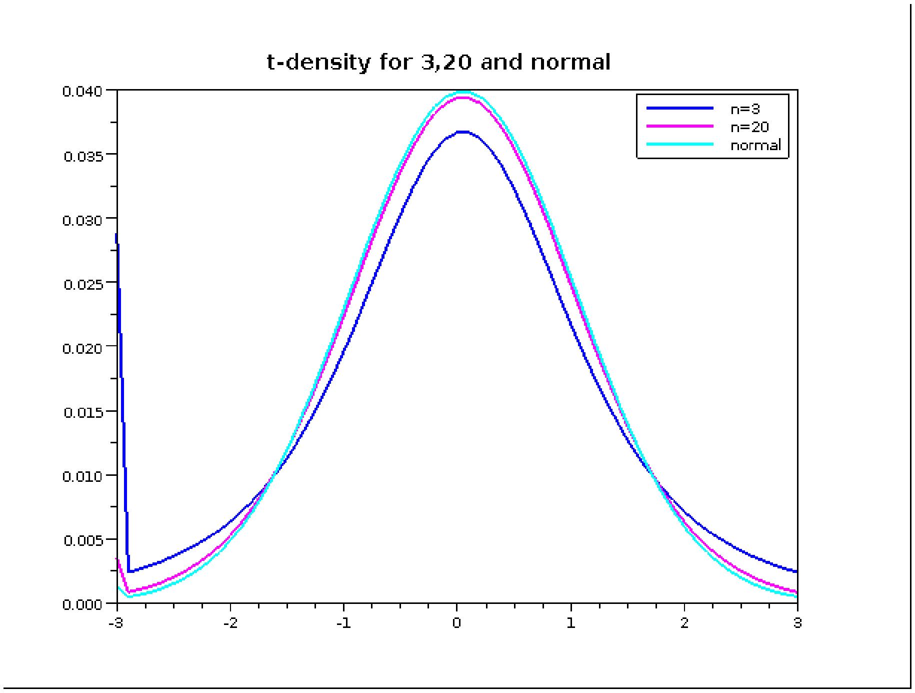

Figure 31: The t-density function

17 Normal with both mean and variance unknown

We now take up the common case that the only information we know about a data set that it is normal, without knowing its mean of variance. Again, the experiment is a repeated trial $X_1, \ldots, X_n$ followed by a computation of the sample mean $\overline{X}$ and sample variance S^2. Suppose that the mean μ were known, and consider the function:

$$T = \frac{(\sqrt{n})\overline{X} - \mu}{\sqrt{S^2})}$$

The variable T is distributed by a classical distribution called the t-distribution of parameter $n-1$. This marks the number of trials. Let us plot the t-density function along with the normal $N(0,1)$. Note that each of the curves is symmetric about the origin, and as expected bell-shaped. Also note that as n gets larger, the t-density function approaches the normal distribution. This is because as n increases, the denominator S^2 approaches the variance σ^2.

The problem here is to estimate μ or to test assertions on it. Again, we do this through an example.

Example 35 *A sample of* 10 *fractional literacy levels in 10 villages was the sequence*

$$[0.82, 0.73, 0.70, 0.69, 0.67, 0.56, 0.45, 0.44, 0.43, 0.43]$$

Give 90% and 99% confidence interval estimates for μ. With what confidence will you refute the claim that μ is 0.55?

We see that the sample mean is 0.592 *and* $S^2 = 0.0217$. *Thus, we have the variable* $T = \frac{0.592-\mu}{\sqrt{S^2/10}} = \frac{0.592-\mu}{0.0466}$. *Next, we use* `cdft` *(with* $n-1=9$ *degree of freedom) to compute the intervals for 90% and 99%. This we get by using the command:*
`cdft("T",[9 9],[0.9 0.99],[0.1 0.01])` *to get* $1.383, 2.827$. *We start with the first problem, i.e., 90%. We have that:*

$$Pr(-1.383 \leq \frac{0.592-\mu}{0.0466} \leq 1.383) \geq 0.9$$

By rearranging, we see that:

$$Pr(0.527 \leq \mu \leq 0.656) \geq 90\%$$

This gives us the confidence interval for the 90% as $[0.527, 0.656]$. *Note that this interval is larger than what would have been for the normal case with* $\sigma = 0.0466$. *The above interval would correspond to a confidence of 91.66% in the normal case. This is because there is an inherent uncertainly about the variance and that causes the t-density to be more broad than the normal case.*

Next, we see that $\frac{0.592-0.55}{0.0466} = 0.901$. *The p-value for this can be found by* `cdft("PQ",-0.901,9)` *which is* 0.196. *Thus, we can reject the claim with a mere* $1-2*0.196 = 0.609$, *i.e., 60.9%.*

18 A few scilab commands and code

Scilab code associated with each and every chapter of the primary textbook (Sheldon M. Ross) can be found at `http://www.cse.iitb.ac.in/~IC102/code/scilab_code.zip`. A local copy of the scilab code description can be found here `http://www.cse.iitb.ac.in/~IC102/code/scilab_code_description.pdf`. The code has been reformated and reorganized chapterwise for compatibility with Aakash tablets in both tar.gz and zip formats at `http://www.cse.iitb.ac.in/~IC102/code/aakash_code`.

In what follows, we only briefly describe

- some functionalities in scilab and some additional functionalities that we have shared for reading and processing data from xls files and
- some of the basic functionalities in scilab. Somewhat detailed introductory notes on scilab can be downloaded from `http://www.cse.iitb.ac.in/~IC102/code/Scilabnotes.pdf`.

For any detailed information on performing operations based on probability and statistics on datasets, you need to refer to the scilab code repository (and its documentation) pointed to above.

Reading a .xls file: In the beginning there is an .xls file. To input it into your scilab session, you need to use the `readxls` command, such as:

`murbad=readxls("thane_murbad_census_I.xls")`

This creates a copy of the .xls file in your session and the file is called `murbad`. These will have as many sheets as your original file had and these are refered as `murbad(1), murbad(2)` and so on. So lets do the following.

```
mu=murbad(1) // this picks out the first sheet
size(mu) // should output 211. 64.
mu.value // will list out the numeric part of the sheet
// and put a NaN (not a number) where it sees text
mu.text // does it for the non-numeric data
```

We see that columns 56, and 10 onwards are numeric, while the others are text. Now, let us select all the rows which correspond to VILLAGE (column 7) and all the numeric columns. This is done as follows:

```
I=[]; for i=1:211 if mu(i,7)=="VILLAGE" I=[I i]; end; end;
murbadnumeric=mu(I,[10:64]);
```

```
size(murbadnumeric) // should give you 205. 55.
save murbadnumeric // now a load will get this back for us
```

Now, we load all the index names. This is done by `exec "index.sci"`. What this will do is to define variables such as TOT_P and NON_WORK_M and put the correct column index for them, which are 11 and 63 respectively. Remember that while creating `murbadvillage` we have deleted the first 9 columns and hence `murbadvillage(:,TOT_P-9)` will be the column vector of the total populations of all villages in Murbad. Just for fun, we extract the population fraction under 6 as follows:

```
for i=1:205 y(i)=murbadnumeric(i,P_06-9)/murbadnumeric(i,TOT_P-9); end;
```

Next, let us list a few scilab functions.

- `mean(X)` returns the mean of the entries of the matrix X. Example `mean([1 2; 3 4])` returns 2.5.
- `nanstdev(X)` returns the standard deviation of the argument X.
- `variance(X,1), variance(X,1,1), variance(X,2)`: This computes the variance of the matrix X. If the second arument is 2 then it computes the variance of each row, while if it is 1 (default), then it does it for each column. The normalization is either (default) $m-1$ (where m is the appropriate dimension) of m. The option of m, which you would normally require, is obtained by adding a third argument 1. Example: `variance([1 2 3],1), variance([1 2 3],1,1), variance([1 2 3],2), variance([1 2 3],2,1)` returns `error`, $[0,0,0]$, 1 and 0.66 respectively.
- `covar(X,Y,eye(n,n))` returns the covariance of the two (row or column) vectors X and Y of equal length. Here n is the size of X (or Y). Example `covar([1 2 1],[2 2 3],eye(3,3))` returns -0.111. Instead of $eye(3,3)$ you could feed in the frequency matrix f, where $f(i,j)$ would be the number of times that you have observed the tuple (x_i, y_j).
- `correl(X,Y,eye(n,n))` returns the correllation of the two (row or column) vectors X and Y of equal length. Here n is the size of X (or Y). Example `correl([1 2 1],[2 2 3],eye(3,3))` returns -0.5. As above, instead of $eye(3,3)$ you could feed in the frequency matrix f, where $f(i,j)$ would be the number of times that you have observed the tuple (x_i, y_j).

- `histplot(M,X)`: plots a histogram of the entries in X. M is either an integer or a row-vector of values $M = [m_1, m_2, \ldots, m_k]$. If M is an integer, the produced figure has M divisions. If M is a vector, then the plots are for frequencies in $[m_{i-1}, m_i]$. he Y-axis is normally fraction of entries. Use `histplot(M,X,normalization=%f)` for frequencies.

- `plot2d(x,y)`: x and y should be vectors of the same size. This will plot a poly-line connecting (x_i, y_i) to $x_{i+1}, y_{i+1})$ for each i. `plot2d(x,y,'r+')` will not draw the line, but only the points. These will be marked red and with a "+" sign.

- `title("my title")` will add a title to your graph. `legend("my legend"), xlabel("mylabel"), ylabel("mylabel")` will add the labels and legends to your plot.

- `grand(m,n,"type",param-list)`: is the basic random number generator.
    - `grand(m,n,"bin",N,q)`: generates an $m \times n$ matrix of numbers in $[0, N]$ with the binomial density function.
    - `grand(m,n,"nor",mu,sig)`: generates an $m \times n$ matrix of reals drawn from the normal density function with mean mu and SD sig.
    - `grand(m,n,"unf",Low,High)`: generates an $m \times n$ matrix of reals drawn from the uniform denisty function for the interval $[Low, High]$.

- `X=binomial(q,n)` produces a vector X of size $n+1$, where $X(k+1)$ is the probability that the outcome of the binomial density function *Binom(q,n)* is k. In other words, $X(k+1) = \binom{n}{k} q^k (1-q)^{n-k}$.

- `XX=cdfnor("PQ ",X,`μ`,`σ`)`. The matrices X, μ, σ must be of the same dimensions and so will the output be.

$$XX(i,j) = \int_{-\infty}^{X(i,j)} \phi(\mu(i,j), \sigma(i,j); x) dx$$

  where ϕ is the gaussian function. Thus `cdfnor` implements the **cumulative density function**.

Example 36 Drawing histograms for actual and predicted frequencies *Consider the case when we have an array of values HH, which has, say, the number of households of all the villages in Shahpur taluka. Let us draw a histogram for this number and compare it with the ideal normal for the same mean and variance as the data HH.*

Here is a sample code fragment., with the output indicated after the % sign:

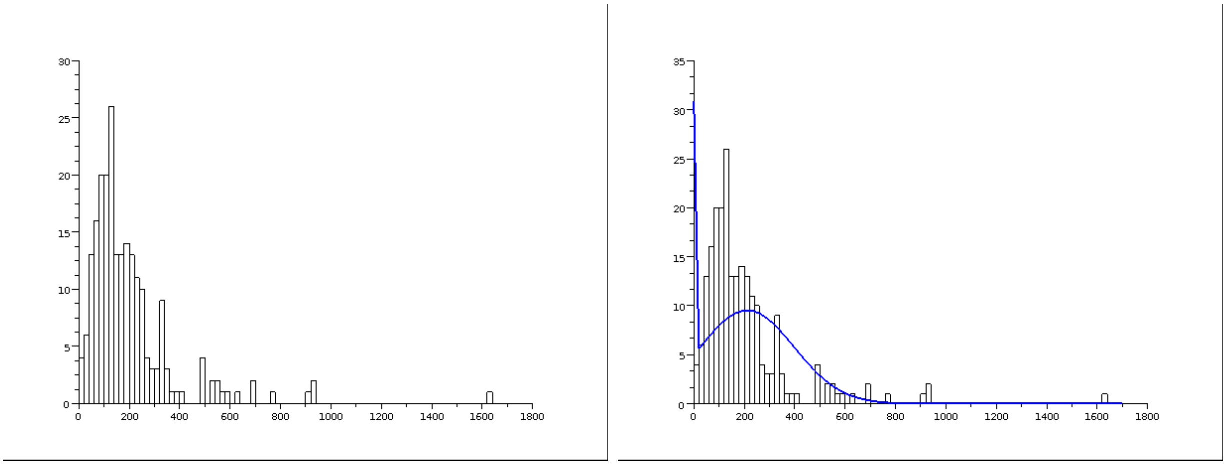

Figure 32: The households in Shahpur

```
mu=mean(HH) // 201
variance(HH,1) // 34484
max(HH) //  1635
sig=sqrt(varr) // 186
xx=linspace(0,1700,86) // this creates an array of 86 equally spaced point from
//0 to 1700, i.e., 20 apart
histplot(xx,HH,normalization=%f) // creates the histogram below

// now we set about creating the expected normal frequencies

size(HH) // is 222
cdf=cdfnor("PQ",xx,mu*ones(1,86),sig*ones(1,86));
// this produces the vector in cdf for all the stopping points
// of the histogram
pdf=differ(cdf)*222 // this is what we want
// differ is our function to compute input(i+1)-input(i)
//
$\chi^2_n $ density function for various $n$}
plot(xx,pdf) // does the job by plotting the normal on the histogram

// the first flick to 30 corresponds to the number which
// should have been there less than zero
```

19 Appendix on Regression

Suppose there are two sets of variables $\mathbf{x} \in \Re^n$ and $\mathbf{y} \in \Re^k$ such that $\mathbf{x}$ is independent and y is dependant. The regression problem is concerned with determining y in terms of $\mathbf{x}$. Let us assume that we are given m data points $\mathcal{D} = \langle \mathbf{x}_1, \mathbf{y}_1 \rangle, \langle \mathbf{x}_2, \mathbf{y}_2 \rangle, .., \langle \mathbf{x}_m, \mathbf{y}_m \rangle$. Then the problem is to determine a function f^* such that $f^*(\mathbf{x})$ is the best predictor for $\mathbf{y}$, with respect to $\mathcal{D}$. Suppose $\varepsilon(f, \mathcal{D})$ is an error function, designed to reflect the discrepancy between the predicted value $f(\mathbf{x}')$ of $\mathbf{y}'$ and the actual value $\mathbf{y}'$ for any $\langle \mathbf{x}', \mathbf{y}' \rangle \in \mathcal{D}$, then

$$f^* = \underset{f \in \mathcal{F}}{\arg\min}\ \varepsilon(f, \mathcal{D}) \tag{3}$$

where, $\mathcal{F}$ denotes the class of functions over which the optimization is performed.

19.1 Multivariate Nonlinear Example: Curve Fitting

Learn $f : X \to Y$ such that $E(f, X, Y_1)$ is minimized. Here the error function E and form of the function to learn f is chosen by the modeler.

Consider one such form of f,

$$f(x) = w_0 + w_1 x + w_2 x^2 + ... + w_t x^t$$

The sum of squares error is given by,

$$E = \frac{1}{2} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

So the expression is,

$$\underset{w=[w_1, w_2, ... w_t]}{\arg\min} \frac{1}{2} \sum_{i=1}^{K} [(w_0 + w_1 x + w_2 x^2 + ... + w_t x^t) - y_1(i)]^2$$

If there are m data points, then a polynomial of degree $m-1$ can exactly fit the data, since the polynomial has m degrees of freedom (where degrees of freedom=no. of coefficients)

As the degree of the polynomial increases beyond m, the curve becomes more and more wobbly, while still passing through the points. Contrast the degree 10 fit in Figure 34 against the degree 5 fit in Figure 33. This is due to the problem of overfitting (overspecification)

Now E is a convex function. To optimize it, we need to set $\nabla_w E = 0$. The ∇ operator is also called gradient.

Solution is given by,

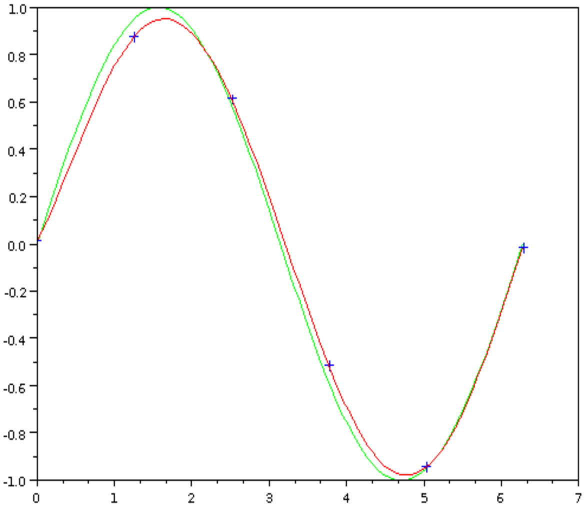

Figure 33: Fit for degree 5 polynomial.

$$X = (\phi^t \phi)^{-1} \phi^t Y$$

If $m << t$ then

- ϕ becomes singular and the solution cannot be found OR
- The column vectors in ϕ become nearly linearly dependent

RMS (root mean sqare) error is given by :

$$RMS = \sqrt{\frac{2E}{k}}$$

Generally, some test data (which potentially could have been part of the training data) is held out for evaluating the generalized performance of the model. Another held out fraction of the training data, called the validation dataset is typically used to find the most appropriate degree t_{best} for f.

19.2 Linear regression and method of least squares error

Depending on the function class we consider, there are many types of regression problems. In Linear regression we consider only linear functions, functions that are linear in the basis function. Here $\mathcal{F}$ is of the form $\{\sum_{i=1}^{p} w_i \phi_i(\mathbf{x})\}$. $\phi_i : \mathbb{R}^n \rightarrow \mathbb{R}^k$ Here, the ϕ_i's are called the **basis functions** (for example, we can consider $\phi_i(x) = x^i$, *i.e.*, polynomial basis functions) .

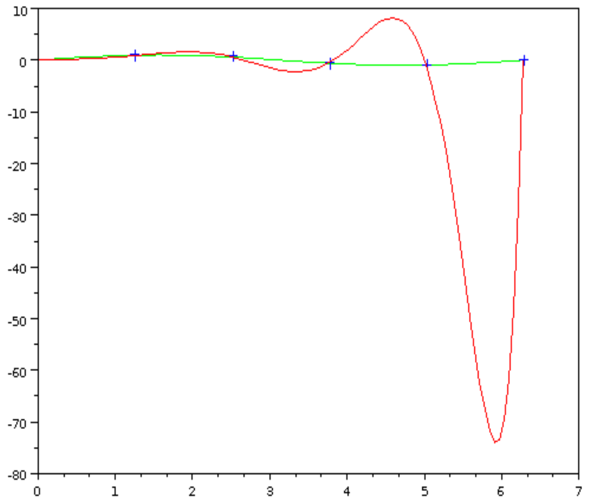

Figure 34: Fit for degree 10 polynomial. Note how wobbly this fit is.

Any function in $\mathcal{F}$ is characterized by its parameters, the w_i's. Thus, in (3) we have to find $f(\mathbf{w}^*)$ where

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \varepsilon(\mathbf{w}, \mathcal{D})$$

Least square solution

The error function ε plays a major role in the accuracy and tractability of the optimization problem. The error function is also called the **loss function**. The squared loss is a commonly used loss function. It is the sum of squares of the differences between the actual value and the predicted value.

$$\varepsilon(f, \mathcal{D}) = \sum_{\langle \mathbf{x}_i, y_i \rangle \in \mathcal{D}} (f(\mathbf{x}_i) - y_i)^2$$

So the least square solution for linear regression is given by

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{j=1}^{m} \Big(\sum_{i=1}^{p} (w_i \phi_i(x_j) - y_j \Big)^2$$

The minimum value of the squared loss is zero. Is it possible to achieve this value ? In other words is $\forall j, \ \sum_{i=1}^{p} w_i \phi_i(x_j) = y_j$ possible ?

The above equality can be written as $\forall u, \ \phi^T(x_u)\mathbf{w} = y_u$
or equivalently $\phi\mathbf{w} = \mathbf{y}$ where

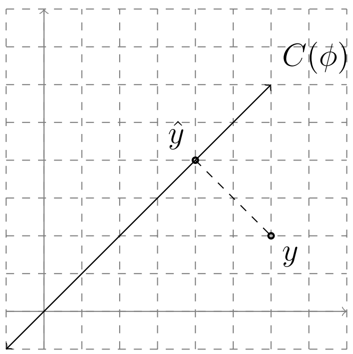

Figure 35: Least square solution $\hat{y}$ is the orthogonal projection of y onto column space of ϕ

$$\phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \cdots & \phi_p(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ \phi_1(\mathbf{x}_m) & \cdots & \phi_p(\mathbf{x}_m) \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

It has a solution if $\mathbf{y}$ is in the column space (the subspace of $\mathbb{R}^n$ formed by the column vectors) of ϕ. It is possible that there exists no $\mathbf{w}$ which satisfies the conditions? In such situations we can solve the least square problem.

Geometrical interpretation of least squares

Let $\hat{\mathbf{y}}$ be a solution in the column space of ϕ. The least squares solution is such that the distance between $\hat{\mathbf{y}}$ and $\mathbf{y}$ is minimized. From the diagram it is clear that for the distance to be minimized, the line joining $\hat{\mathbf{y}}$ to $\mathbf{y}$ should be orthogonal to the column space. This can be summarized as

1. $\phi\mathbf{w} = \hat{\mathbf{y}}$
2. $\forall v \in \{1, ..p\},\ (\mathbf{y} - \hat{\mathbf{y}})^T\phi_v = 0$ or $(\hat{\mathbf{y}} - \mathbf{y})^T\phi = 0$

$$\begin{aligned} \hat{\mathbf{y}}^T\phi \quad &= \mathbf{y}^T\phi \\ ie,\ (\phi\mathbf{w})^T\phi \quad &= \mathbf{y}^T\phi \\ ie,\ \mathbf{w}^T\phi^T\phi \quad &= \mathbf{y}^T\phi \\ ie,\ \phi^T\phi\mathbf{w} \quad &= \phi^T\mathbf{y} \\ \therefore\ \mathbf{w} &= (\phi^T\phi)^{-1}\mathbf{y} \end{aligned}$$

In the last step, please note that, $\phi^T\phi$ is invertible only if ϕ has full column rank.

Theorem: If ϕ has full column rank, $\phi^T\phi$ is invertible. A matrix is said to have full column rank if all its column vectors are linearly independent. A set of vectors $\mathbf{v}_i$ is said to be linearly independent if $\sum_i \alpha_i \mathbf{v}_i = 0 \Rightarrow \alpha_i = 0$.

Proof: Given that ϕ has full column rank and hence columns are linearly independent, we have that $\phi\mathbf{x} = 0 \Rightarrow \mathbf{x} = \mathbf{0}$.

Assume on the contrary that $\phi^T\phi$ is non invertible. Then $\exists \mathbf{x} \neq \mathbf{0} \ni \phi^T\phi\mathbf{x} = \mathbf{0}$.

$\Rightarrow \mathbf{x}^T\phi^T\phi\mathbf{x} = 0$

$\Rightarrow (\phi\mathbf{x})^T\phi\mathbf{x} = ||\phi\mathbf{x}||^2 = 0$

$\Rightarrow \phi\mathbf{x} = \mathbf{0}$. This is a contradiction. Hence the theorem is proved.

19.3 Regularised solution to regression

We previously derived solution for the regression problem formulated as a solution to the least-squares objective, that is, by minimizing the rms error over observed data points. We also analysed conditions under which the obtained solution was guaranteed to be a global minima. However, as we observed, increasing the order of the model yielded larger rms error over test data, which was due to large fluctuations in the model learnt and consequently due to very high values of model coefficients (weights). In this lecture, we discuss how the optimization problem can be modified to counter very large magnitudes of coefficients. Subsequently, solution to this problem is provided through lagrange dual formulation followed by discussion over obtained solution and impact over test data.

Problem formulation

In order to discourage coefficients from becoming too large in magnitude, we may modify the problem and pose a constrained optimization problem. Intuitively, for achieving this criterion, we may impose constraints on the magnitude of the coefficients. Any norm for this purpose might provide a good working solution. However, for mathematical convenience, we start with the euclidean (L_2) norm. The overall problem with objective function and constraint goes as follows:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & (\Phi\mathbf{w} - Y)^T(\Phi\mathbf{w} - Y) \\ \text{such that} \quad & ||\mathbf{w}||_2^2 \leq \xi \end{aligned} \tag{4}$$

As observed in last lecture, the objective function, namely $f(\mathbf{w}) = (\Phi\mathbf{w}-Y)^T(\Phi\mathbf{w}-Y)$ is strictly convex. Further to this, the constraint function, $g(\mathbf{w}) =\| \mathbf{w} \|_2^2 -\xi$, is also a convex function. For convex $g(\mathbf{w})$, the set $S = \{\mathbf{w}|g(\mathbf{w}) \leq 0\}$, can be proved to be a convex set by taking two elements $w_1 \in S$ and $w_2 \in S$ such that $g(w_1) \leq 0$ and $g(w_2) \leq 0$. Since $g(\mathbf{w})$ is a convex function, we have the following inequality:

$$\begin{aligned} g(\theta w_1 + (1-\theta)w_2) &\leq \theta g(w_1) + (1-\theta)g(w_2) \\ &\leq 0; \forall\theta \in [0,1], w_1, w_2 \in S \end{aligned} \tag{5}$$

As $g(\theta w_1+(1-\theta)w_2) \leq 0$; $\forall\theta \in S$, $\forall w_1, w_2 \in S$, $\theta w_1+(1-\theta)w_2 \in S$, which is both sufficient and necessary for S to be a convex set. Hence, function $g(\mathbf{w})$ imposes a convex constraint over the solution space.

Bound on λ in the regularized least square solution

As discussed earlier, we need to minimize the error function subject to constraint $\|\mathbf{w}\| \leq \xi$. Applying the first order necessary conditions of minimality to this problem, if $\mathbf{w}^*$ is a global optimum then from the first first order necessary conditions for minimality, we get,

$$\nabla_{\mathbf{w}^*}(f(\mathbf{w}) + \lambda g(\mathbf{w})) = 0 \tag{6}$$

where, $f(\mathbf{w}) = (\Phi\mathbf{w} - Y)^T(\Phi\mathbf{w} - Y)\,\text{and}\,g(\mathbf{w}) = \|\mathbf{w}\|^2 - \xi$
Solving we get,

$$2(\Phi^T\Phi)\mathbf{w}^* - 2\Phi^T - 2\lambda\mathbf{w}^* = 0$$

i.e.

$$\mathbf{w}^* = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y} \tag{7}$$

We will also obtain the following conditions from the Karush Kuhn Tucker necessary optimality conditions;

$$\|\mathbf{w}^*\|^2 \leq \xi \tag{8}$$

and

$$\lambda \geq 0 \tag{9}$$

and

$$\lambda \|\mathbf{w}^*\|^2 = \lambda \xi \tag{10}$$

Thus values of $\mathbf{w}^*$ and λ which satisfy all these equations would yield an optimal solution. Consider equation (7),

$$\mathbf{w}^* = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T \mathbf{y}$$

Premultiplying with $(\Phi^T \Phi + \lambda I)$ on both sides we have,

$$(\Phi^T \Phi + \lambda I)\mathbf{w}^* = \Phi^T \mathbf{y}$$

$$\therefore (\Phi^T \Phi)\mathbf{w}^* + (\lambda I)\mathbf{w}^* = \Phi^T \mathbf{y}$$

$$\therefore \|(\Phi^T \Phi)\mathbf{w}^* + (\lambda I)\mathbf{w}^*\| = \|\Phi^T \mathbf{y}\|$$

By triangle inequality,

$$\|(\Phi^T \Phi)\mathbf{w}^*\| + (\lambda)\|\mathbf{w}^*\| \geq \|(\Phi^T \Phi)\mathbf{w}^* + (\lambda I)\mathbf{w}^*\| = \|\Phi^T \mathbf{y}\| \tag{11}$$

Now , $(\Phi^T \Phi)$ is a nxn matrix which can be determined as Φ is known .
$\|(\Phi^T \Phi)\mathbf{w}^*\| \leq \alpha \|\mathbf{w}^*\|$ for some α for finite $|(\Phi^T \Phi)\mathbf{w}^*\|$. Substituting in the previous equation,

$$(\alpha + \lambda)\|\mathbf{w}^*\| \geq \|\Phi^T \mathbf{y}\|$$

i.e.

$$\lambda \geq \frac{\|\Phi^T \mathbf{y}\|}{\|\mathbf{w}^*\|} - \alpha \tag{12}$$

Note that when $\|\mathbf{w}^*\| \to 0, \lambda \to \infty$. This is obvious as higher values of λ would focus more on reducing values of $\|\mathbf{w}^*\|$ than on minimizing the error function.

$$\|\mathbf{w}^*\|^2 \leq \xi$$

Eliminating $\|\mathbf{w}^*\|$ from the equation (14) we get,

$$\therefore \lambda \geq \frac{\|\Phi^T \mathbf{y}\|}{\sqrt{\xi}} - \alpha \tag{13}$$

This is not the exact solution of λ but the bound (15) proves the existance of λ for some ξ and Φ.

RMS Error variation

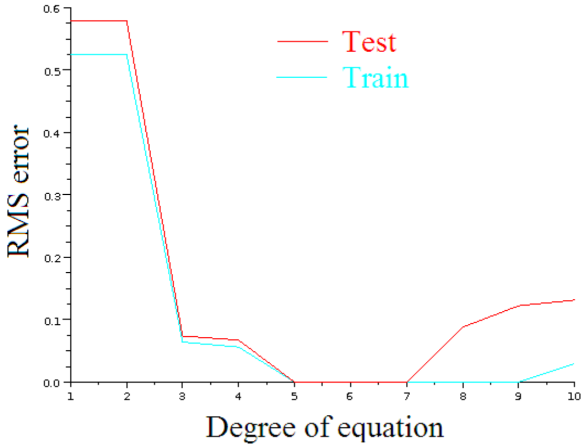

Figure 36: RMS error vs. degree of polynomial for test and train data

Recall the polynomial curve fitting problem we considered in earlier lectures. Figure 36 shows RMS error variation as the degree of polynomial (assumed to fit the points) is increased. We observe that as the degree of polynomial is increased till 5 both train and test errors decrease. For degree > 7, test error shoots up. This is attributed to the overfitting problem (The datasize for train set is 8 points.)

In Figure 37, the variation in the RMS error with variations in the Lagrange multiplier λ has been explored (keeping the polynomial degree constant at 6). Given this analysis, what is the optimum value of λ that must be chosen? We have to choose that value for which the test error is minimum (Identified as the point of optimum in the figure.).

Alternative objective function

Consider equation (6). If we substitute $g(\mathbf{w}) = \|\mathbf{w}\|^2 - \xi$, we get

$$\nabla_{\mathbf{w}^*}(f(\mathbf{w}) + \lambda \cdot (\|\mathbf{w}\|^2 - \xi)) = 0 \tag{14}$$

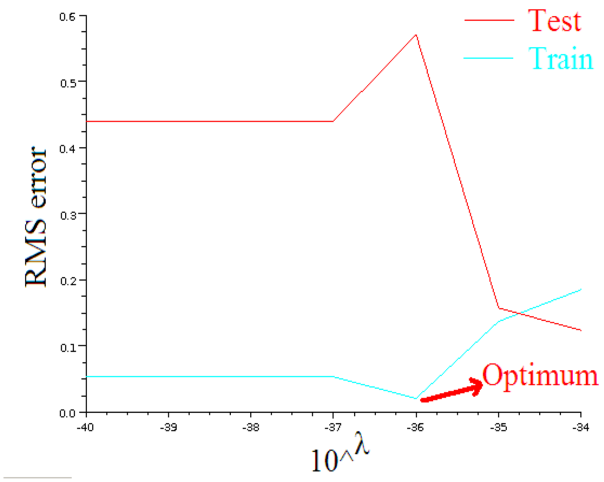

Figure 37: RMS error vs. $10^{\lambda}$for test and train data (at Polynomial degree = 6)

This is equivalent to finding

$$\min(\| \Phi\mathbf{w} - \mathbf{y} \|^2 + \lambda \| \mathbf{w} \|^2) \tag{15}$$

For the same λ, these two solutions are the same. This form or regression is known as Ridge regression. If we employ the L_1 norm it is called 'Lasso'. Note that the $\mathbf{w}^*$ form that we derived is valid only for the L_2 norm.

20 Appendix on Gaussian and Uniform Distributions

20.1 Information Theory

Let us denote I(X=x) as the measure of information conveyed in knowing value of X=x.

Question: Consider the two graphs above. Say you know probability function $p(x)$. When is knowing value of X more useful (that is, carries more information)?
Ans: It is more useful in the case(2), because more information is conveyed in Figure 38 than in Figure 39.

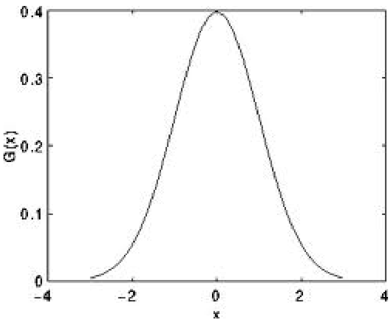

Figure 38: Figure showing curve where Information is not distributed all along.

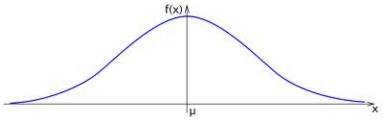

Figure 39: Figure showing curve where Information is distributed.

Expectation for I(X=x):

- If X and Y are independant random variables from the same distribution.

$$I(X = x, Y = x) = I(X = x) + I(Y = y) \tag{16}$$

One way of expressing the above is:

$$I(x, y) = I(x) + I(y) \tag{17}$$

where P(x),P(y) are the probability functions respectively.

- If $p(x)>P(y)$, then

$$I(x)<I(y)$$

There is only one function which satisfies the above two properties.

$$I(p(x)) = -c\log(p(x)) \tag{18}$$

- The Entropy in the case of discrete random variable can be defined as:

$$E_P\left[I(p(x))\right] = \sum_x -c\log[p(x)] \tag{19}$$

- In the case of continuous random variable it is,

$$E_P\left[I(p(x))\right] = \int_x -c\log[p(x)] \tag{20}$$

The constant 'C' in the above two equations is traditionally 1.

Observations:

- For a discrete random variable (with countable domain), the information is maximum for the uniform distribution.
- For Continuous random variable (with finite mean and finite variance), the information is maximum for the Gaussian Distribution.

Finding $\underset{p}{argmax}\ E_p$ in an infinite domain, subject to

$$\int xp(x)dx = \mu$$

and

$$\int (x-\mu)^2 p(x)dx = \sigma^2$$

The solution would be

$$p(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

Properties of gaussian univariate distribution

- If $X \sim N(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \ where - \infty < x < \infty$$

then $w_1X + w_0 \sim N(w_1\mu + w_0, w_1^2\sigma^2)$
(can prove this using moment generating function)

$$\Phi(N(\mu,\sigma^2)) = E_{N(\mu,\sigma^2)}[e^{tx}] = e^{\mu t + \frac{(\sigma t)^2}{2}}$$

<u>*Recall*</u>
$E(X) = \frac{d\phi(p)}{dt}$
$var(x) = \frac{d^2\phi(p)}{dt^2}$

$$E_{N(\mu,\sigma^2)}[e^{t(w_1 x + w_0)}] = (w_1\mu t + w_0 t + \frac{(\sigma t)^2}{2} \times w_1^2) \sim N(w_1\mu + w_0, w_1^2\sigma^2)$$

- Sum of i.i.d $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$ is also normal (gaussian)

$X_1 + X_2 + ... + X_n \sim N(n\mu, n\sigma^2)$

In genaral if $X_i \sim N(\mu_i, \sigma_i^2) \Longrightarrow \sum_{i=1}^n X_i \sim N(\sum \mu_i, \sum \sigma_i^2)$

- Corollary from (1) If $X \sim N(\mu, \sigma^2)$

$$z = \frac{X-\mu}{\sigma} \sim N(0,1) \text{ (Useful in setting interval estimate)}$$

(take $w_1 = \frac{1}{\sigma}$ *and* $w_0 = \frac{\mu}{\sigma}$)

<u>Note:-</u> If $X_1, X_2, ...X_m \sim N(0,1)$

1. $y = \sum_i X_i^2 \sim \chi_m^2$. That is, y follows the chi-square distribution with m-degrees of freedom.
2. $y = \frac{z}{\sum X_i^2} \sim t_n$. (where $z \sim \mathcal{N}(0,1)$)). That is, y follows the *students-t* distribution.

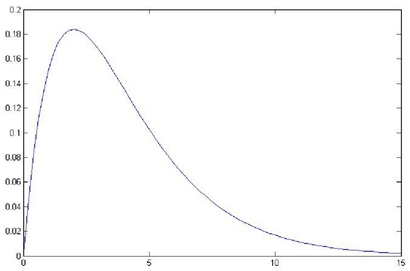

Figure 20.1 : Figure showing the nature of the $(chi - square)$ distribution with 5 degrees of freedom

- Maximum Likelihood estimate for μ and σ^2

Given $X_1, X_2, ...X_m$..... Random Sample.

$$\hat{\mu}_{MLE} = argmax_{\mu} \prod_{i=1}^{m} [\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(X_i-\mu)^2}{2\sigma^2}}]$$

$$= argmax_{\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\sum(X_i-\mu)^2}{2\sigma^2}}$$

$\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{m} X_i}{m}$ = sample mean

- With out relaying on central limit theorem Properties (2) and (1)

i.e. Sum of i.i.d's $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$

$\hat{\mu}_{MLE} = N(\mu, \frac{\sigma^2}{m})$

$Similarly$

$\hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^{m}(X_i - \hat{\mu}_{MLE})^2}{m}$ is χ^2 distrbution

$\sim \chi^2_m$

- Coming up with conjugate prior of $N(\mu, \sigma^2)$

  Case (1) σ^2 is fixed and prior on μ

  $\Rightarrow \mu \sim N(\mu_0, \sigma_0^2)$

  Case (2) μ is fixed and σ^2 has prior

  $\Rightarrow \sigma^2 \sim \Gamma$

  case (3) if μ and σ^2 both having the prior

  $\Rightarrow (\mu, \sigma^2) \sim$ Normal gamma distribution $\sim$ Students-t distribution