Data Analysis and Interpretation

Milind Sohoni
CTARA & CSE, IIT Bombay
sohoni@cse.iitb.ac.in

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.

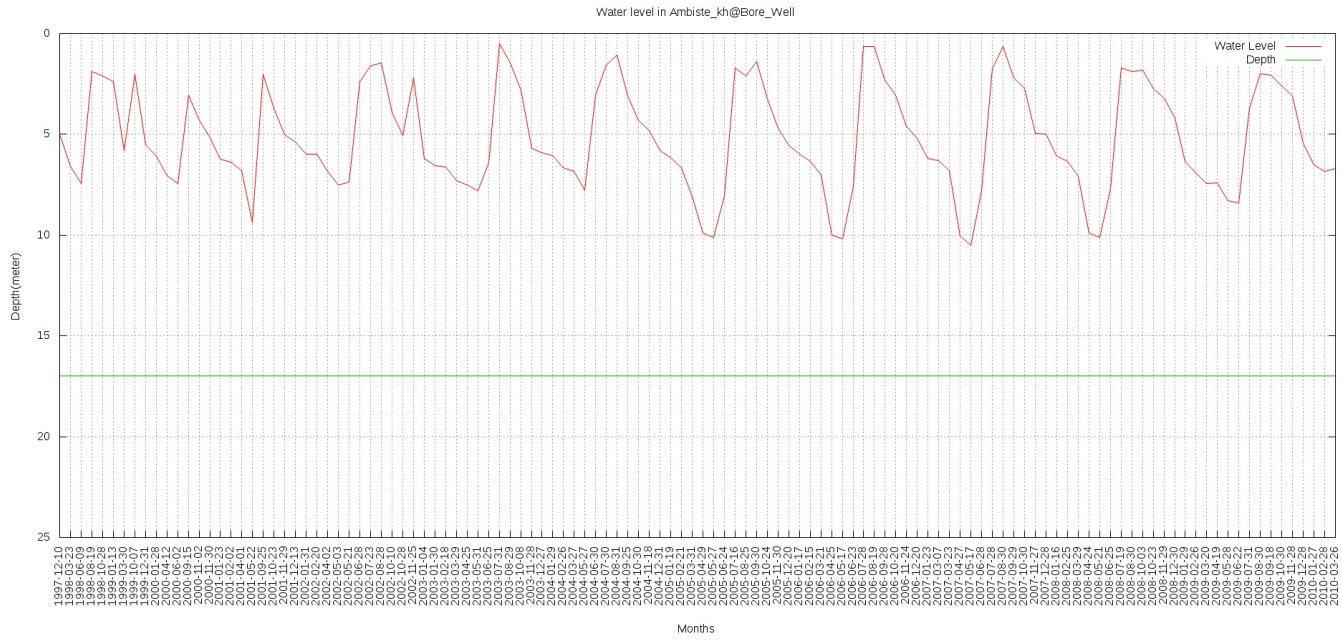

Figure 1: The water levels in a borewell (Courtesy GSDA)

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 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 attributes that we will associate with data. These are:

- *Qualitative vs. Quantitative*: This is about the variable in question and whether it is completely described by numeric quantities. Typical quantitative attributes would be *weight (in kgs.) and location (in latitude, longitude)*. However, there are many which are not, e.g., *Satisfaction with Service in a Hotel*. Frequently, such attributes 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. A third type of data item is the *discrete*, e.g., the names of *talukas* in a district. The discrete set of values is generally regarded as quantitative since its measurement is usually unambiguous.

- *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*, but 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 attributes 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 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 of statistics such as the *sample* and attrbutes such as the *sample mean* and *sample variance*. We will use the Thane census 2001 data-set through out for this part. 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 examples 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 The Thane census dataset

Our main dataset for the course will be the Thane district census 2001 dataset. This is available at `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 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:
  - `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.

Here is the data for Pimpalshet of Jawhar taluka, Thane.

3 Elementary properties of data

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

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

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

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. The first single point estimate of the data set is of course, 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.

The next useful computation is that of the **variance** and that is $\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n}$ and it 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 1 *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.

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 2 *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, se 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 3 *For a given data-set (y_i), minimize the function $f(\lambda) = \sum_i (y_i - \lambda)^2$.*

Example 4 *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.*

Sometime you need to be careful with computing the means. Here is an example. Part II data lists for each village if its people have access to tap water or not. 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 Data representation

Given a large set of data-items, say in hundreds, the mean μ and the variance σ^2 are but two attributes of the data. 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. Fox 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 and the standard deviation 583,

and the maximum size of a village is 3152 households. However, 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* below, Fig. 4. 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 $\mu \pm 3\sigma$.

At this point, another attribute of a data-set is the **median** which 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 above, it is 403. 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, this 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 current data-set, 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). 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 5 *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.*

Coming back to 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:

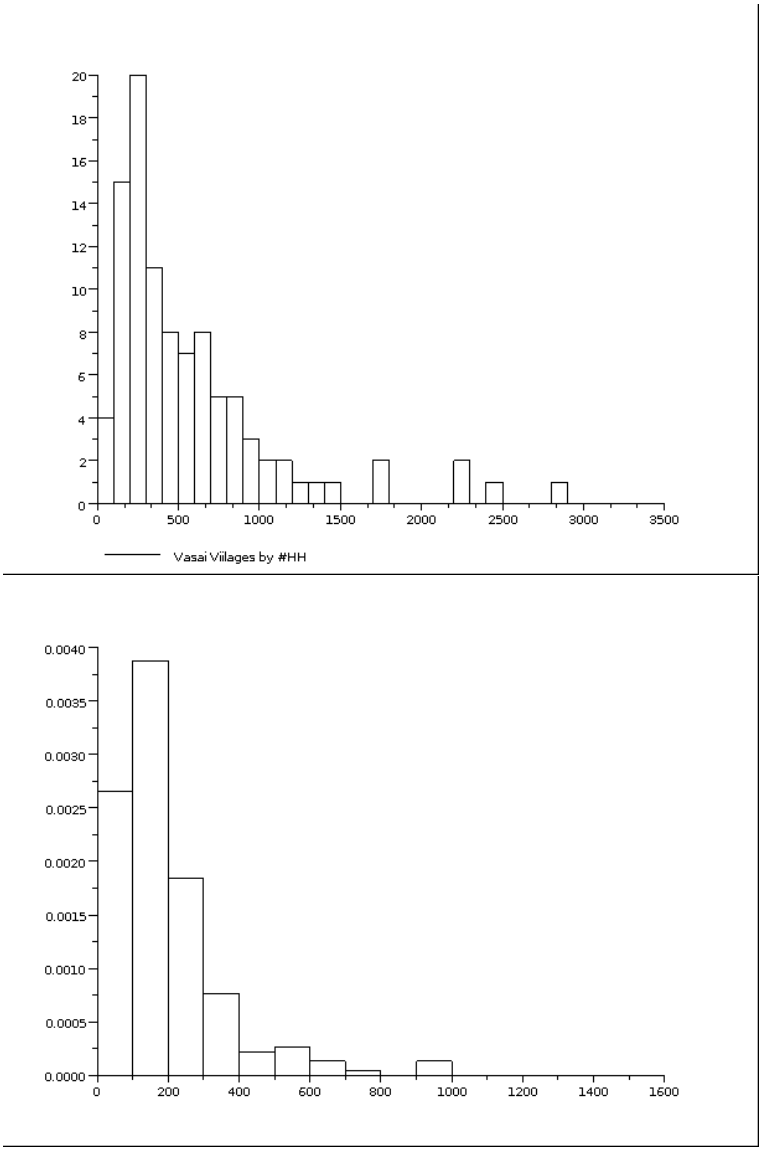

Figure 3: Number of households in villages in Vasai and Shahpur

	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.

The second important representation is the **scatter plot**. This is done 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 6 *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 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.*

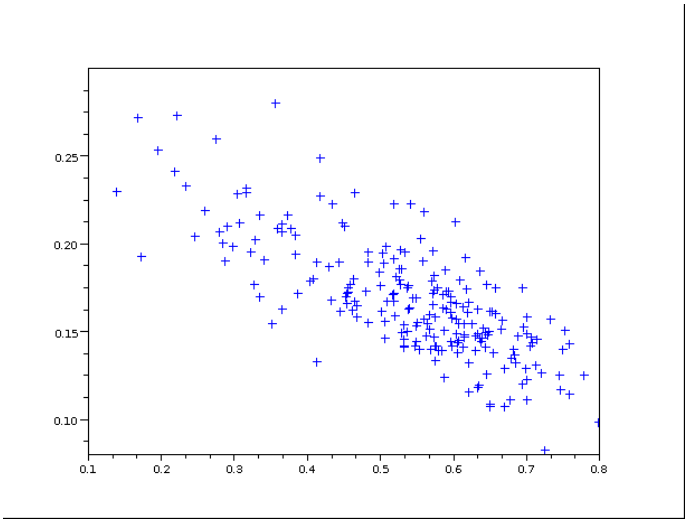

Figure 4: Population under 6 vs. literacy fractions for Shahpur

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. 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**. 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

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

The correlation, 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

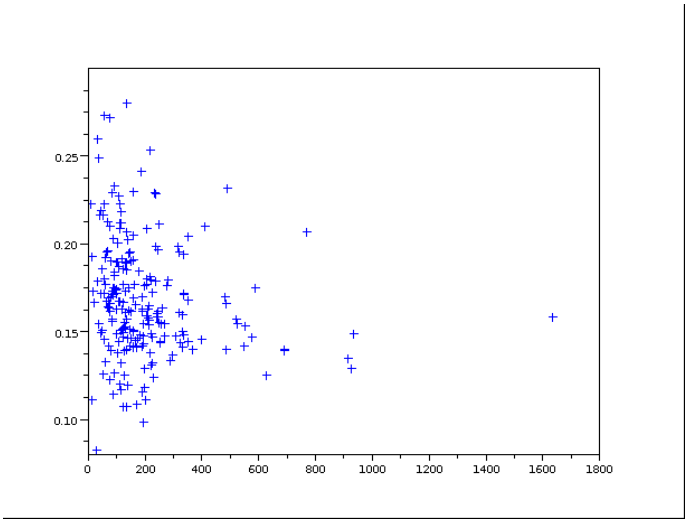

Figure 5: Population under 6 fraction vs. number of HH for Shahpur

and LIT in Fig. 4 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).*

5 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 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 $(X_i/X, y_i/Y)$.

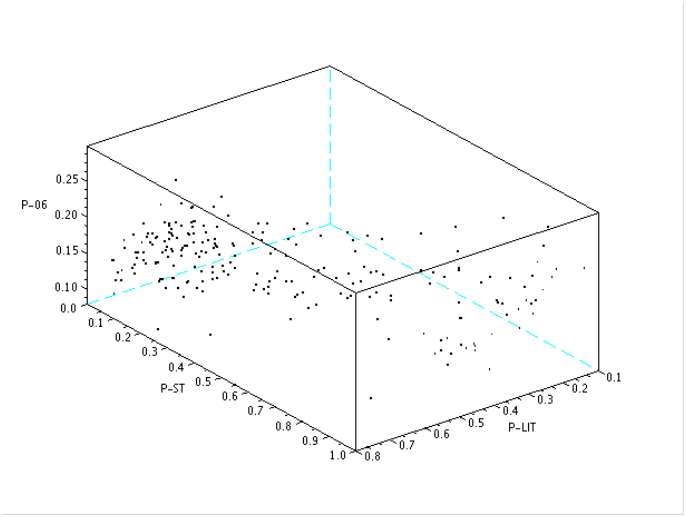

Figure 6: A 3-way plot for Shahpur

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 below:

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 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.

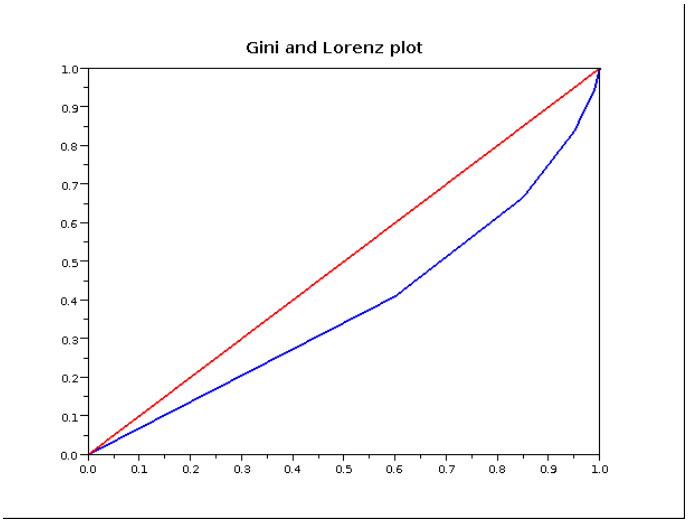

Figure 7: The Lorenz plot for the company data

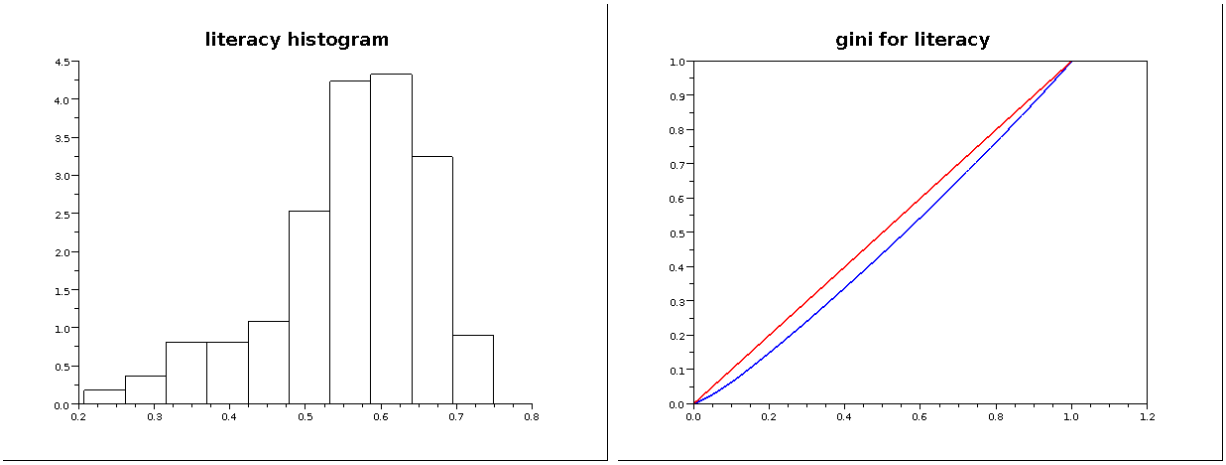

Figure 8: The Lorenz plot for Murbad literacy

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.

6 Linear regression

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:

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

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):

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

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

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

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

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

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 $0 \leq r^2 \leq 1$.

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

Thus, we see that the n-vectors (e_i) and $(f(x_i) - \mu_Y)$ are perpendicular, and sum to $(y_i -$

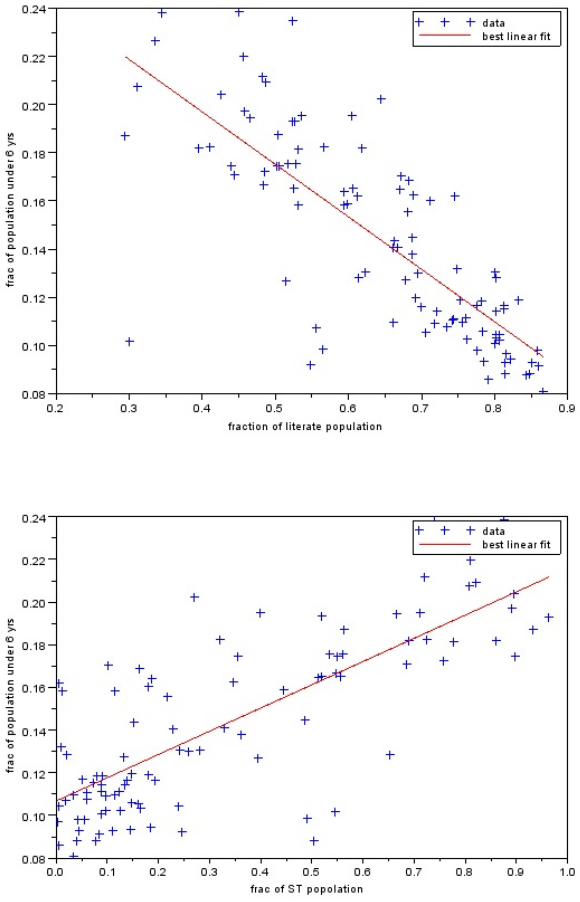

Figure 9: Regression: Population under 6 vs. literacy and ST fraction

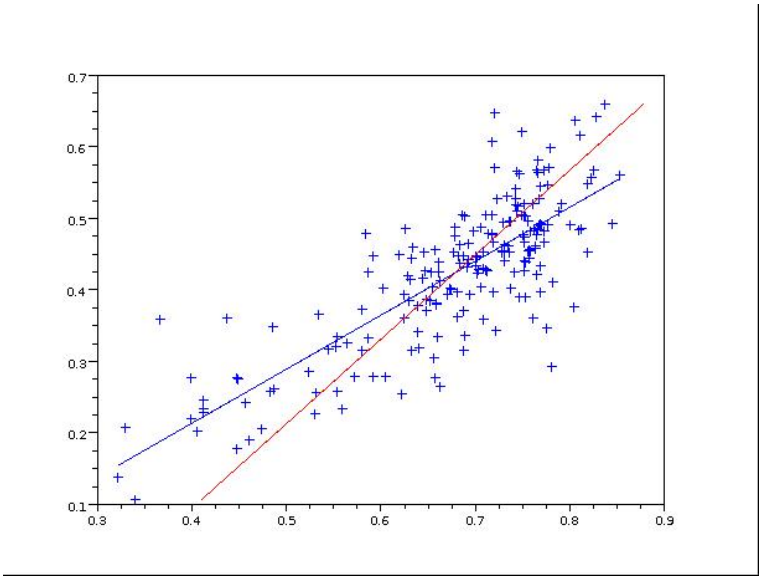

Figure 10: Vasai female vs. male literacy. Both way regression.

$f(x_i) + f(x_i) - \mu_Y) = (y_i - \mu_Y)$. 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).

7 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 approximating y is given by:

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

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:

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

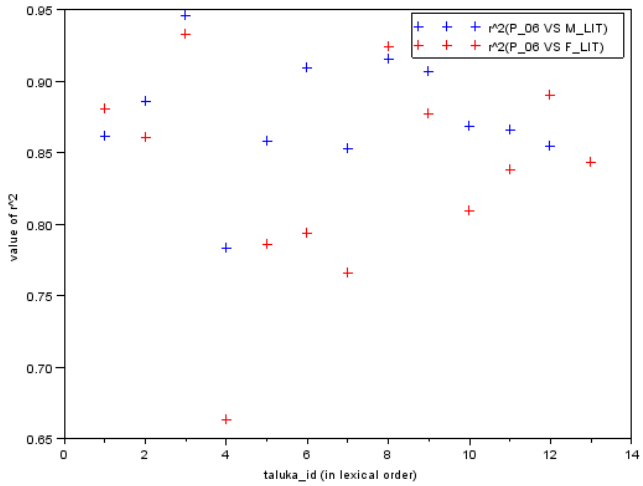

Figure 11: The male and female literacy and P-06

We also see that

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

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.

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 now 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. Now, we move to the probability function p. It must have the following properties: (P1) $p(A) \geq 0$ for all $A \in \mathcal{E}$, (P2) $p(\phi) = 0$ and $p(S) = 1$, and (P3) if $A \cap B = \phi$ then $p(A \cup B) = p(A) = p(B)$.

Example 11 ***The biased coin***. *Here we construct the random variable $C(q)$ corresponding to the biased coin. Let $S = \{H, T\}$, 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., 2^2) subsets of S, and let $0 < q < 1$ be a fixed real number. We define p 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 12 ***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 $\mathcal{E}_1 \times cal E_2$, and define $p(A \times B) = p_1(A)p_2(B)$.

Example 13 *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 14 *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 $f : S^n \to [n]$ 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 simoplest 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 $f : S \to \mathbb{R}$ 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 15 ***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 16 *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

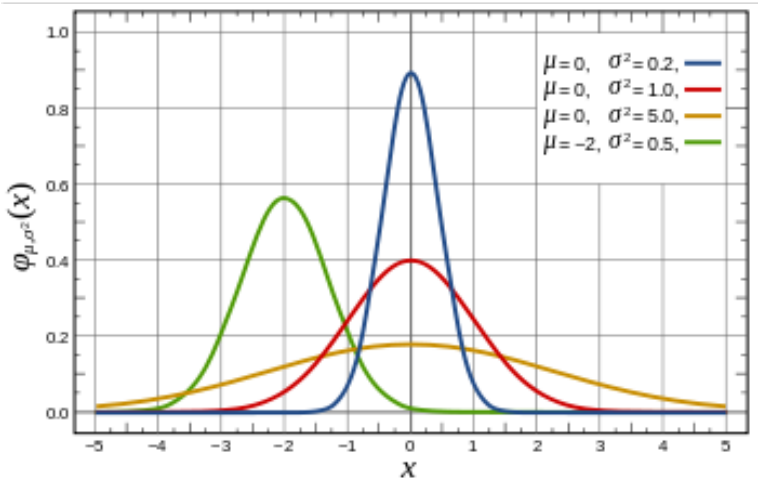

Figure 12: The Normal density function (from wikipedia)

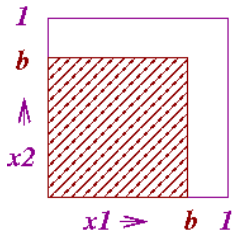

Figure 13: The cross-product of two uniform variables.

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 17 *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* $\int_0^1 \int_0^1 f(x_1, x_2)dx_1dx_2 = 1$*.*

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

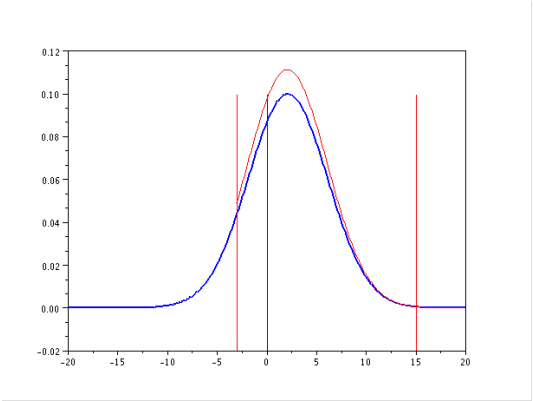

Figure 14: The temperature at Shimla as see by a thermometer

$f_Y(y) = \frac{1}{a} f(\frac{y-b}{a})$. 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 18 *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

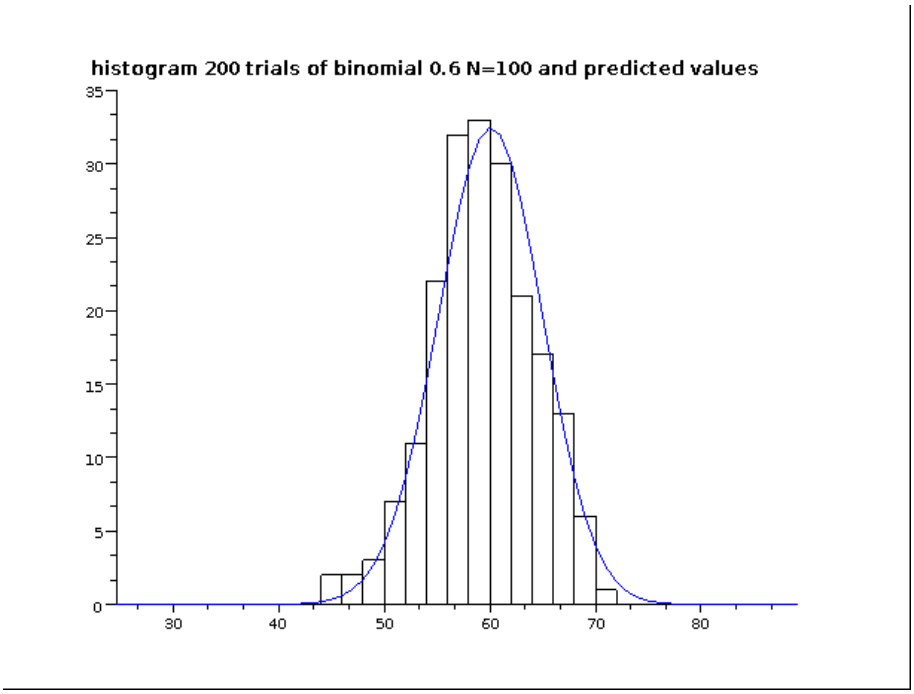

Figure 15: The binomial sample and expectation

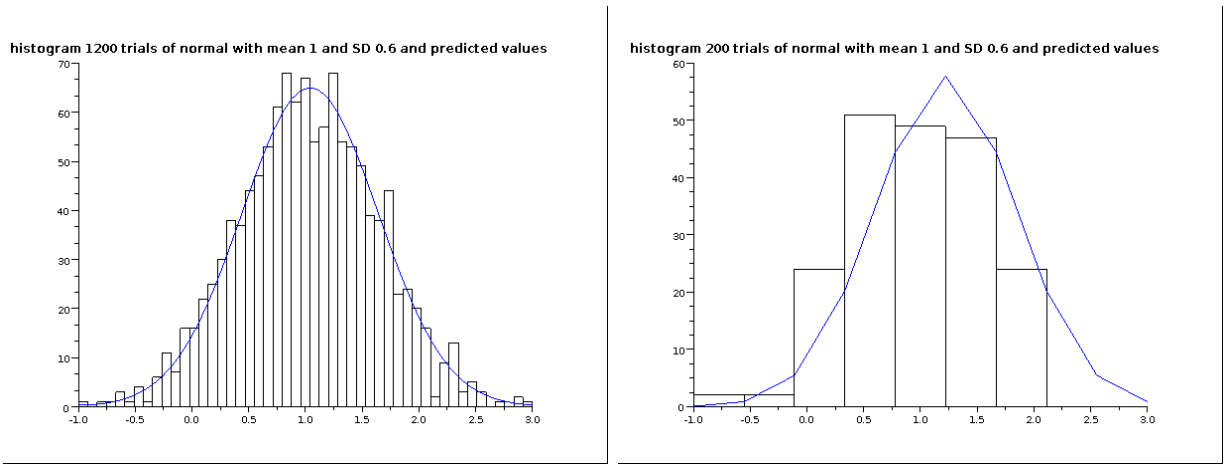

Figure 16: The normal trial and expectation

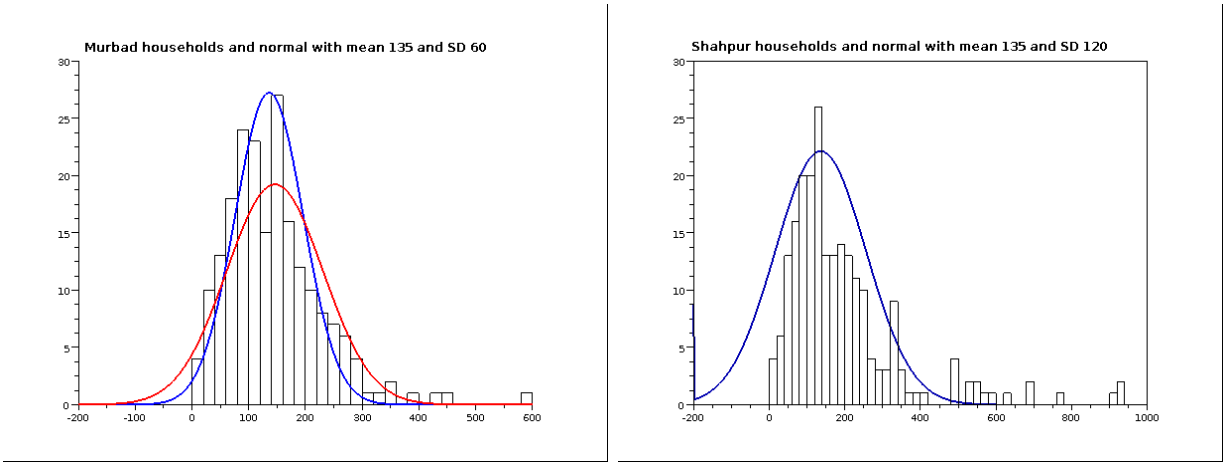

Figure 17: The normal fit to Murbad and Shahpur HH data

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 19 *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?*

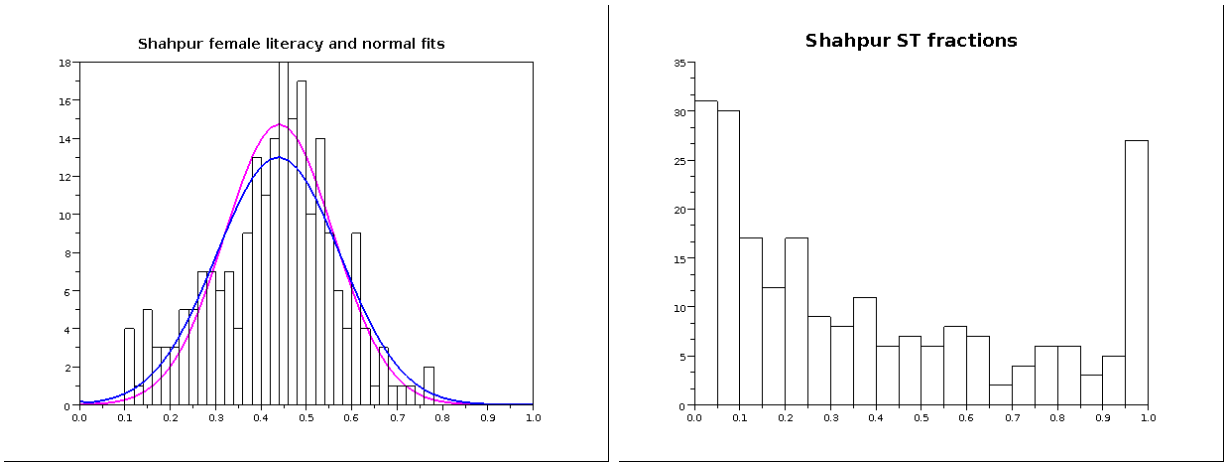

Figure 18: Shahpur female literacy fraction and ST fraction

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 20 *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 21 *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 22 *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 23 *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 $E(x) = \int_0^1 x dx = \left[\frac{x^2}{2}\right]_0^1 = 1/2$. 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:

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

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 24 *The* **transform** $\Phi_X(s)$ *of the random variable* X *given by the density function* f *is* $E(e^{-sX}) = \int f(x)e^{-sx}dx$.

In fact, the transform 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 $\Phi_X(s) = e^{\mu s + \frac{\sigma^2 s^2}{2}}$. We see this in the following steps:

$$\begin{aligned}
\Phi_X(s) &= \int_{-\infty}^{\infty} e^{-sx} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{2\sigma^2 sx - (x-\mu)^2}{2\sigma^2}} dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-(\mu+\sigma^2 s))^2}{2\sigma^2}} e^{\frac{s^2\sigma^4 + 2\mu\sigma^2 s}{2\sigma^2}} dx \\
&= e^{\frac{s^2\sigma^2}{2} + \mu s}
\end{aligned}$$

- Suppose that X_1 and X_2 are independent random variables with density functions $f_1(x)$ and $f_2(x)$, and transforms $\Phi_1(s)$ 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:

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

  Now differentiating under the inetgrals gives us:

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

  The transform of $f_Y(y)$ is the product $\Phi_Y(s) = \Phi_1(s) \cdot \Phi_2(s)$. This is seen by:

$$\begin{aligned}
\Phi_Y(s) &= \int_{y=-\infty}^{\infty} e^{-sy} \cdot \int_{x=-\infty}^{\infty} f_1(x)f_2(y-x)dxdy \\
&= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} e^{-sy} f_1(x)f_2(y-x)dydx \\
&= \int_{x=-\infty}^{\infty} e^{-sx} f_1(x) \left[\int_{y=-\infty}^{\infty} e^{-s(y-x)} f_2(y-x)dy\right] dx \\
&= \int_{x=-\infty}^{\infty} e^{-sx} f_1(x)\Phi_2(s)dx \\
&= \Phi_1(s)\Phi_2(s)
\end{aligned}$$

- 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

$$\Phi_Y = \prod_i e^{\frac{s^2\sigma_i^2}{2} + s\mu_i} = e^{\frac{s^2 \sum_i \sigma_i^2}{2} + s\sum_i \mu_i}$$

  This is clearly the transform of the normal random variable fo 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 25 *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.

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

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 26 *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 $\sigma^2_{Z_n}$ is $n\sigma^2_X/n\sigma^2_X = 1$. Thus Z_n has mean 0 and variance 1.

Theorem 27 ***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 $n\sigma^2_X$.*

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 $\sigma^2_X = 1/12$. Let us consider n trials and the variable

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

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 28 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.

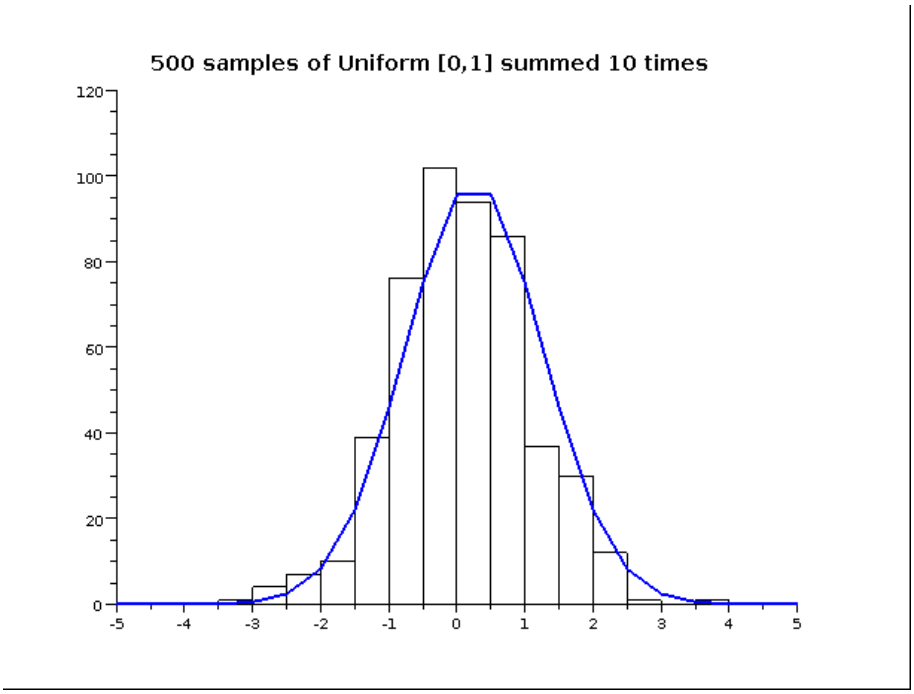

Figure 19: 500 trials of a 10-uniform-sum

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-The Binomial Case

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 $p(3) = \binom{10}{3} q^3(1-q)^{10-3}$ 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.

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

Thus $q = k/n$ is where the derivative is zero. It is easy to check that this is a maxima. Thus

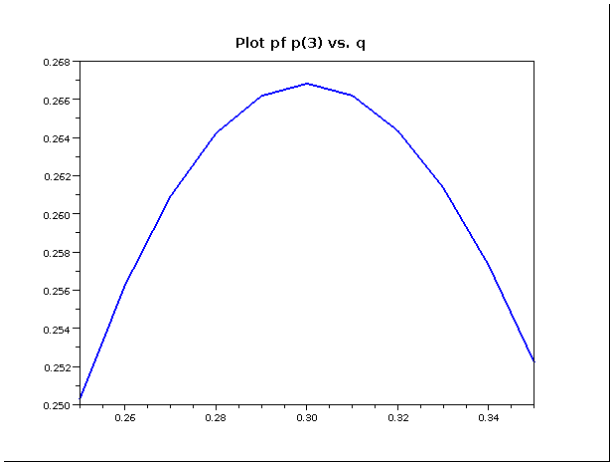

Figure 20: Estmating q when k=3 and n=10

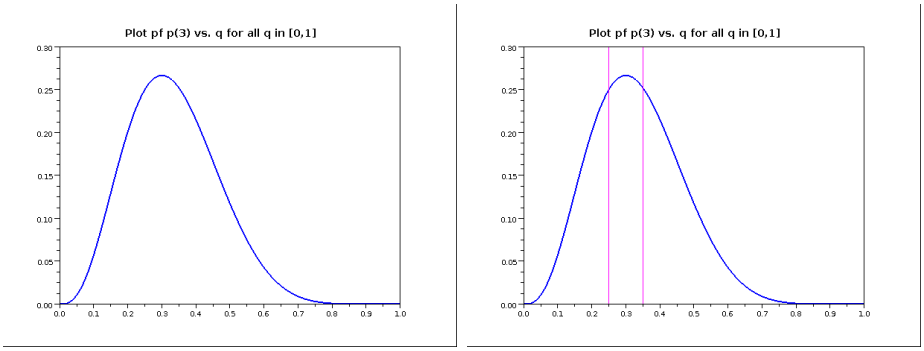

Figure 21: p(3) for all q and the confidence interval $[0.25, 0.35]$

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.

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

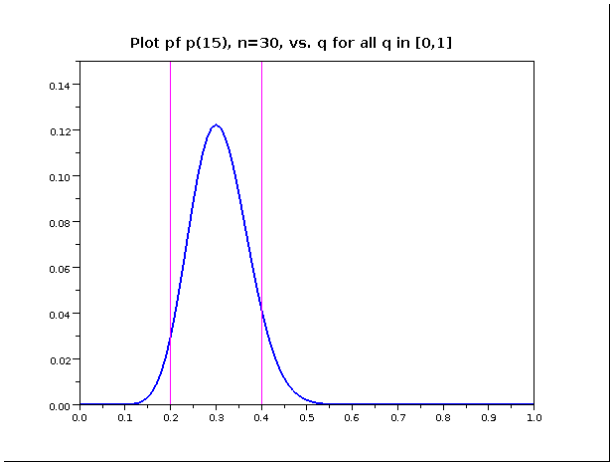

Figure 22: p(16) and n=50 for all q and the confidence interval $[0.2, 0.3]$

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 $q \in [0, 1]$.

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 $[0.37n, 0.43n]$.

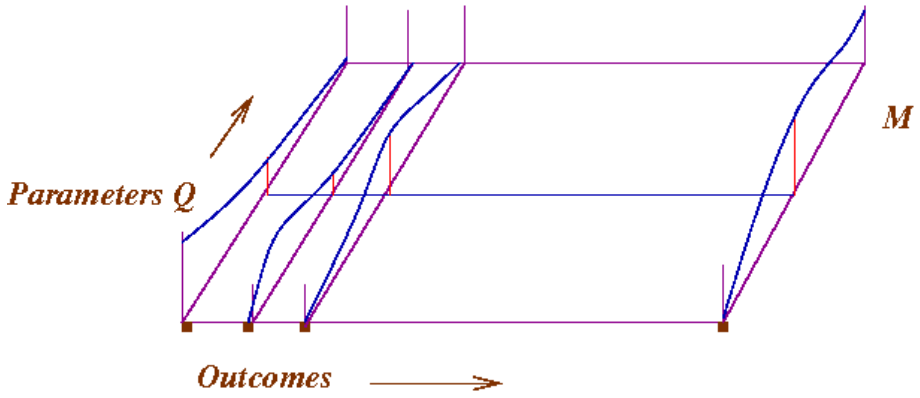

Figure 23: The matix M of Q vs. Outcomes

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 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 29 *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 30 *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

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

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

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 31 *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? (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 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 has a parameter n and is denoted by χ^2_n. 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^2_n is the density function

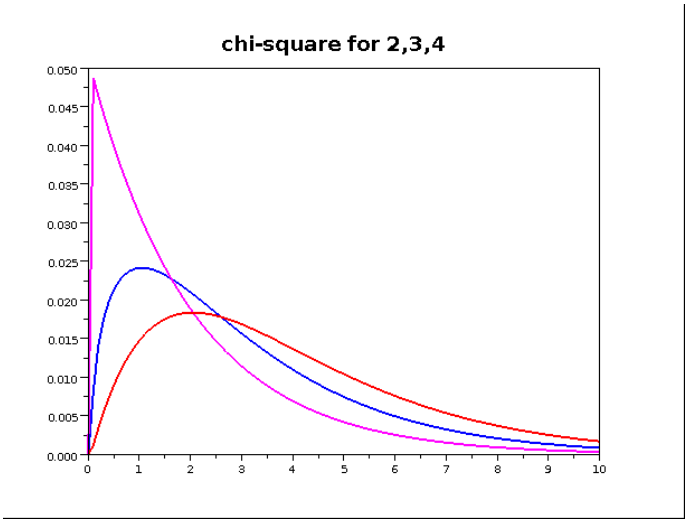

Figure 24: The χ^2_n density function for various n

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).

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

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 32 *A sample of* 10 *fractional literacy levels in 10 villages was the sequence* $[0.82, 0.73, 0.70, 0.69, 0$ *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.*

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:

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

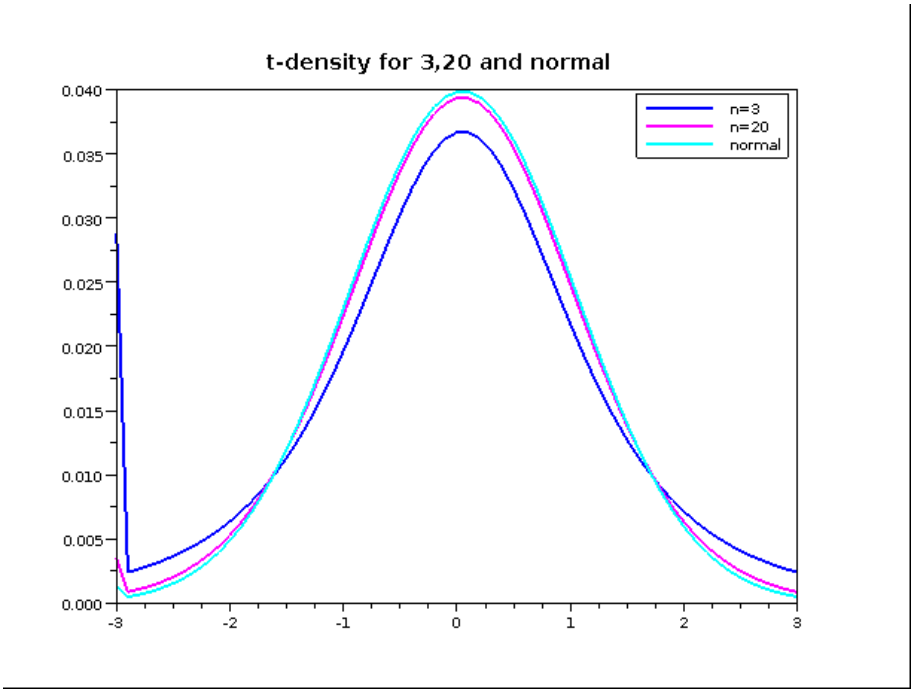

Figure 25: The t-density

The variable T is distributed by a classical distribution called the t-distribution of parameter n. 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 comes closer to the 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 33 *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 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* $n-1 = 9$ *degrees 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 uncertainty 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 can be found by* `cdft("PQ", -0.901,9)` *which is* 0.196*. Thus* $1 - 2 * p = 0.609$ *and we reject the claim with a mere 60.9% confidence.*

18 A few scilab commands and code

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 34 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:

```
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)
//
```

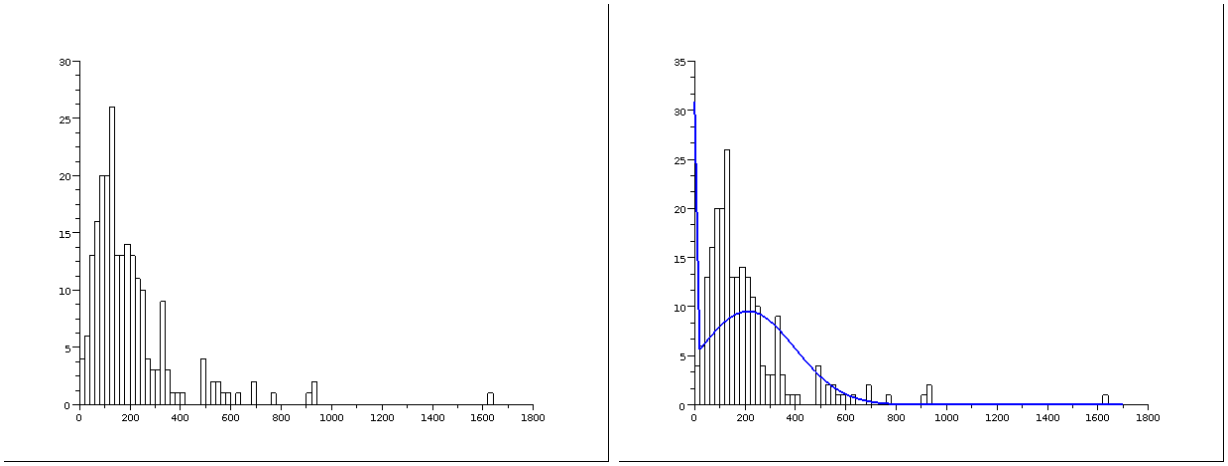

Figure 26: The households in Shahpur

```
$\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
```