

Introduction to Statistical Hypothesis Testing
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 03
Probability and Statistics: Review – Part 2

Hello friends, welcome to the lecture on the Review of Probability and Statistics, as a part of the course on introduction to statistical hypothesis testing. This is part-2 of the review the previous lecture we learnt the concepts of probability basics of probability axioms and we also studied the notion of random variable, probability distribution functions and in particular probability density functions for continuous valued random variables.

Now, if I may connect with the overall purpose of this lecture and the previous lecture. What we are trying to do is we are trying to describe the truth or the population characteristics. Ultimately, we will not be working with the population because what we mean a population is a large sample space or the sample space that we have, we will not be able to work with that in practice. We collect subsets or what we call as a sample and then we work with the data resulting from that sampling and then draw inferences about the population. So, what we are trying to do is we are trying to understand, how truth is being described and what are the parameters of interest was and then gradually will move on to the practical or the practicing aspects of hypothesis testing.


So, in this review what we are going to do is, learn what are known as movements of a PDF.

(Refer Slide Time: 01:51).

Probability and Statistics: Review - Part 2 References

Learning objectives

- ▶ Moments of a p.d.f.
- ▶ Mean and Variance
- ▶ Joint p.d.f.
- ▶ Covariance and correlation



Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 2

In particular, the mean and variance and we will also study the bivariate case, that is until now, until the previous lecture, we have learnt how to describe a single random variable. But very often we will deal with more than one random variable; for example, in regression and so on. We will not the study the multivariable case, but will study what is known as a bivariate case. That is a case of 2 random variables, in which situation we come across the notion of a joint probability density function or joint probability mass functions, depending on whether you are looking at continues or describe valued random variables respectively. In that contexts, will study the ubiquitous the most frequently encountered measure of relationship or dependence between 2 random variables know as a correlation. So, let us begin our journey or continue journey with the univariate case. As I said in the previous lecture, we came across a notion of probability density function.

(Refer Slide Time: 02:58).


Probability and Statistics: Review - Part 2 References

Practical Aspects

In practice, knowing the pdf of a random variable is seldom possible theoretically. One has to conduct experiments and then try to fit a known pdf that best explains the behaviour of the RV.

- ▶ It may not be necessary to know the pdf in practice!
- ▶ What is of interest in practice is (i) **the most likely value and/or the expected outcome (mean)** and (ii) **how far the outcomes are spread (variance)**

Instead of working with p.d.f.s, it is convenient to work with moments (as in mechanics)



Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 3

Again, to study single random variable associated with the random phenomena. Now, as far as theory goes, the probability density functions or the distribution functions are very nice to work with. But, the question is, how do you obtain these probability density functions or even distributions? It is going to be quiet tough task. There are a few processes for which perhaps, you can derive the PDF's or the cumulative distribution functions from the physics or the process as I had mentioned the book by Ogunnaike, shows you how to derive the distribution functions, when you know something about the physics or the process. But, there are large classes of processes for which we will not be able to do that. We will we may not have the luxury to do that. So, how do you now deal with that? The natural recourse is to estimate or reconstruct those PDF's from data that we collect and we know, there it is impossible in general for a last class of processes to be able to obtain all possible out comes. So, we end up with working with subsets of data which we call as samples.

Now from the sample, we are suppose to estimate the PDF now; if you turn to the literature on the estimation of PDF's or the probability density or distribution functions. There exist no highly reliable methods for estimating this PDF's there are; it does not mean that there are no methods there are methods but they are not as reliable as those

methods that exist. For example, estimating averages variances and so on. So, can we work with those methods that, estimate averages or estimate variability and so on. And, that is what we are going to talk about briefly now. So, if you look at practically knowing the PDF of a random variable is not going to be possible theoretically. If at all it is possible in a few cases, it should be in the case then as I said, just know you will have to conduct experiments now the good news is in many situations, you may not have to reconstruct the PDF it may not be the necessary to know the full PDF as a simple example; suppose, I am visiting new city which have never visited and I need to make a decision on what clothes to pack in my bag.

What type of clothes; should it be in a woolen clothes or cotton clothes and so on? That is depends of course, on the temperatures that prevail in the city. So, what would I do as a lay man or even as a general user, traveler? I would actually go to the website or consult some book or some one and seek information about this temperature and obviously, the temperature in a city can be thought of is a random variable, because it is not predictable accurately, on any day. So, it is meaningful to treat it as a random variable. However, if you think of gen, very rarely do we ask for the PDF of the random variable it may sound as big silly to think of the PDF of the temperature in a particular city. Practically, what we would be interested in is some basic so called statistics the minimum, maximum average and so on. And, with these pieces of information we are able to make a decision at least to begin with some decent good decision that we can actually make on what kind of cloth to pack.

The question is how many pieces of information such information sufficient to begin with? So, again going back to the same example, I would like to know, I need to know. For example, what is the average temperature that is the first piece of information, that I would like to know if the average temperature. Let us say transfer to be 10 degree Celsius then going from a tropical country to such a city would mean, that I may have to pack cloths for or pack winter clothing in addition I would also need one information which is the range of temperatures that one gets to see in that season during which I am visiting, which we call as which we may either measure by minimum and maximum, which we will give us an estimate or we look at the variability or we what we call as a standard deviation. Once I know that let us say that this particular website tells me that

the standard deviation is about 10 degree Celsius so, that means I can expect temperatures anywhere between 10 plus or minus, may be two standard deviation; 20 and so on. So, that is a quite a lot of variation.

Nevertheless, never less is just an example. Not to worry about it and this 2 sigma interval that we look at is ideal, if you think of this random variable following a Gaussian distribution and so on. But, those are more detailed aspects of it the bottom line is I am glad or sorry I am fine with knowing with these two pieces of information. As far as my decision making is concern and that is the case usually in practice we are mostly interested in mean and variance and perhaps of course, it not mean median and so on. Which are representatives of the most likely value or the expected outcome and how far the outcomes are spread? Respectively; so, the mean tells me, what I can expect when I visit the city and the variance gives me an idea of what is the spread of the outcome, what are the extremes that I get to see. So, to summaries instead of working with PDF's it is convenient to work with this 2 kind of statistics you can say population statistics or 2 parameters of the PDF's and so on. Now, it turns out that the mean that we are talking of is nothing but the first moment of a PDF.

(Refer Slide Time: 09:31).


Probability and Statistics: Review - Part 2 References

First moment of a pdf: Mean

Mean: First moment of the pdf (analogous to the center of mass). It is also the **expected value** (outcome) of the RV.

Mean

The mean of a RV, also the **expectation** of the RV, is defined as

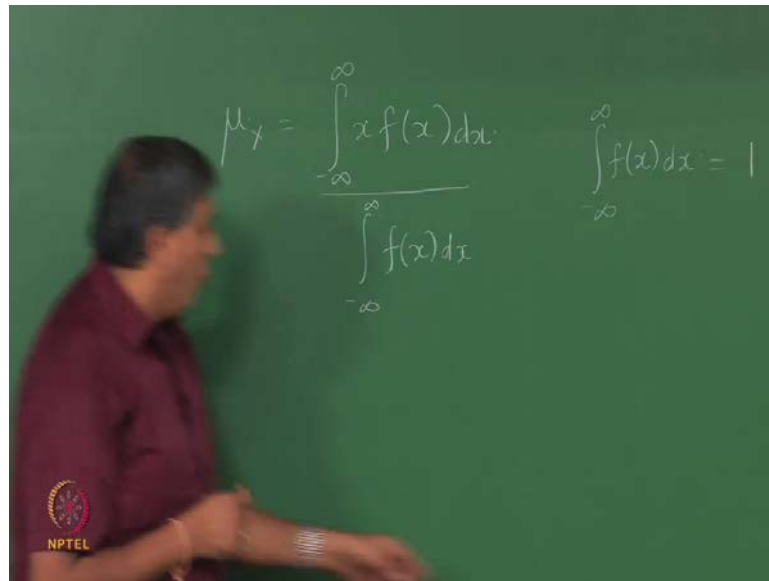
$$E(X) = \mu_X = \int_{-\infty}^{\infty} x f(x) dx \quad (1)$$


Anun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 4

Now, many a times people have difficulties understanding or interpreting this probability density functions but, it is convenient if you take an analogy in mechanics of the mass distribution or the mass density functions typically we talk of densities which is mass per unit volume but we can also talk of mass per unit length we call it as specific mass. So, this density function that we are looking at is pretty much analogous to the density functions that we see in mechanics and we know from mechanics. We also have moments and so on; which tell us it give us an idea of where the center of mass is for example, with we have learnt center of mass center of gravity and so on likewise here we have a set of outcomes spread over range and we would like to know where the statistical center is, that is what is a most likely outcome or what is an expected outcome as we got and for this purpose we introduce this term called expectation it is strictly an operator and typically denoted with the symbol E with. So, we write expectation of x as you see in equation one here, which is also conventionally denoted by the Greek symbol μ and it is defined as minus this integral here running from minus infinity to infinity x times f of x dx , now this minus infinity to infinity should not be looked upon numerically but more.

So, symbolically they represent the left and the right extreme values for the single random variable and we know that this integral is the first moment of the PDF, again from mechanics. Now, strictly speaking there should be a denominator in this and that is often avoided for a reason.

(Refer Slide Time: 11:37).



And let me tell you that, we have μ as $\int x f(x) dx$ in the expression; however, strictly speaking you are supposed to have an $\int f(x) dx$ also. Now, probably, you are in a much better position to relate to the expressions that you see, that we get to see in mechanics. Of course, the limits are the same; integration limits are the same on both integrals but, we know by definition of the probability density function that the area under the density function is unity and therefore, this is avoided. In fact, we talked about this any density function to be called as density function should satisfy this criterion. So, it is assumed that we have already done some kind of normalization. So, that this legitimate requirement is met and therefore, we do not need to write this anymore, it is understood, it is an invisible denominator which is unity, alright.

Now, of course we are talking about continuous valued random variables here and there is also another interpretation, rough interpretation that you can take from this integral.

(Refer Slide Time: 12:53)

$$\mu_x = \int_{-\infty}^{\infty} x f(x) dx$$
$$\int_{-\infty}^{\infty} f(x) dx = 1$$
$$\approx \sum x_i f(x_i) dx_i$$
$$E(x) \text{ OR } E(g(x)) = \int_{-\infty}^{\infty} g(x) f(x) dx$$

The chalkboard also includes a small interval notation $(x_i, x_i + dx_i)$ next to the summation formula. An NPTEL logo is visible in the bottom left corner of the image.

Approximately, if you think as the integral as a summation, after all integral has limiting case of the summation then and if you divide the outcomes space very finely let say, it is a continuous valued outcome space or sample space and you divide it very finely into a grid and call the i th point on the grid as x_i ; then you would say f of $x_i dx_i$ here. So, you have the integral being replaced by summation approximately. So, look at what this expression carefully. Now, you have x_i times f of x_i times dx_i this f of $x_i dx_i$ is a probability of this random variable taking on values in this infinitesimal interval x_i and $x_i + dx_i$. So, this is a probability that the random variable will take on values in this small or infinitesimal interval. So, what you are saying is that the average value is the particular outcome weighted by this probability and this, what one gets to see for discrete valued random variables for discrete variable, you will not see integrals we will see summations in that case. We replace f of $x_i dx_i$ by the probability of x taking on that value itself, but this is continuous valued case. Therefore, we tend to write this way. So, this kind of helps us remember or probably get a better interpretation of this integral.

Now, this expectation operator is a fundamental operator in all in the entire theory of a probability and one needs to be quite comfortable with computing expectations of course, in hypothesis testing it may not be required, but what is required is interpreting

this expectation in a correct manner whenever we compute expectation of a random variable x or expectation of any function of that random variable we are essentially doing is computing the statistical average or you can say the Ogunnaike average or the average across the population and so on. It is not the average in time at all there is it is often misconstrued and misconceived as being average in time there is no notion of time. Here, the expectation is an averaging operation that is being performed across the entire population space. In fact, the expectation of g of x is nothing but integral g of x times f of x dx . So, g any function of a random variable is also random variable that is the first point and therefore, it would have a statistical average and that is once again written this way and you can give similar interpretation as we have given here. Whenever x occurs you can calculate g of x and that is weighted by the probability. So, you're actually again doing the same thing we use this expression to compute or to come up with the definition of variance as we shall see shortly. Let us go through couple of examples on computing theoretical or the population mean.

(Refer Slide Time: 16:31).

Probability and Statistics: Review - Part 2 References


Example

Problem: Determine the mean of a RV that follows Gaussian distribution.

Solution: The Gaussian distributed RV has the pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right)$

Therefore,

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right) dx$$

$$= \mu$$


Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 5

Let us say I have a random variable that follows Gaussian distribution and I want to determine the mean. Now, the piece of information that we require to proceed is the density function and we know that the Gaussian density function has this expression; 1

over $\sigma^2 \phi$ exponential of $-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}$. Therefore, I plug in that expression for f of x in the definition of mean and if you work out the integral of course, you need a bit of math here and am skipping all of that eventually, the answer works out to be μ and that is why we have been using this symbol μ in the first place itself, right. So, that tells us we have used μ in the PDF.

And, going to the second example; suppose, I have a random variable that follows a uniform distribution. Now, we know that the uniform distributor random variable has the PDF in the interval a comma b as $\frac{1}{b - a}$ then the expectation is once again $\int x \times \frac{1}{b - a} dx$ running between running from a to b because those are the left and right extremes and it is clearly $\frac{b + a}{2}$ that is intuitive. So, what is important here? Is that the μ , that we are seeing is a statistical center it is not necessarily the geometrical center of the outcomes and I will talk about it briefly, but in these 2 distribution both Gaussian and uniform distribution are symmetric about their means therefore, the statistical center and the geometrical center workout to be the same.

(Refer Slide Time: 18:18).

Probability and Statistics: Review - Part 2 References


Remarks on Mean

- ▶ The integration in equation (1) is across the **outcome space** and NOT across any time space.
- ▶ The symbol E is the **expectation operator**. Applying the operator E to a random variable produces its "average" or expected value.
- ▶ There are other measures of the center of outcomes - for example, **median**
- ▶ Prediction perspective:

The mean is the best prediction of the random variable in the minimum mean square error sense, *i.e.*,

$$\mu = \min_c E((X - \hat{X})^2) \text{ s.t. } \hat{X} = c$$

where \hat{X} denotes the prediction of X .



Anun K. Tangirala, IIT Madras

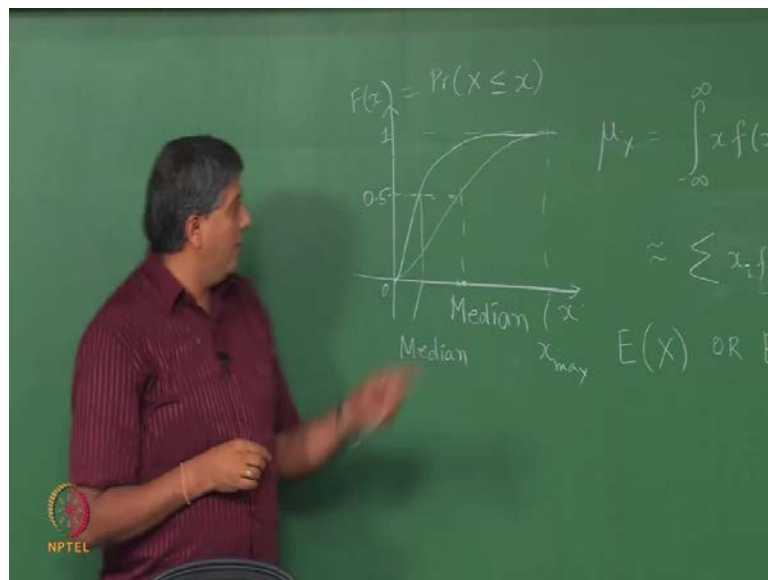
Intro to Statistical Hypothesis Testing

7

Now, as I mention. So, I just want to go through few make a few remarks on the mean itself and then I will again comeback to the statistical center and geometrical center part.

So, the integration in equation integral in equation one is across outcome space as I had mention earlier and not across the time or any other special or frequency domain and wherever you see expectation hence forth, you should think it as an averaging operation and of course, there is a reason why we have given that name expectation will talk about that. There are other measures for the center of outcomes mean is not the only measure of the center, remember the reason why we are working with mean is we want to know the point around which the outcomes are anchored and that will first help me establish. Where in the space of real numbers the outcomes anchored around like in the city visit example what is the center around which central or the temperature around which the temperatures during the season that I visit are anchored; so, another alternative is median. Now, we know that the definition of median is that value which divides the probability distribution into 2 equal halves. Let me explain that. So, the theoretical median the definition of median is defined as the center sorry, the point at which f of x hits the value of 0.5 which f here is the probability distribution not the density.

(Refer Slide Time: 20:16).

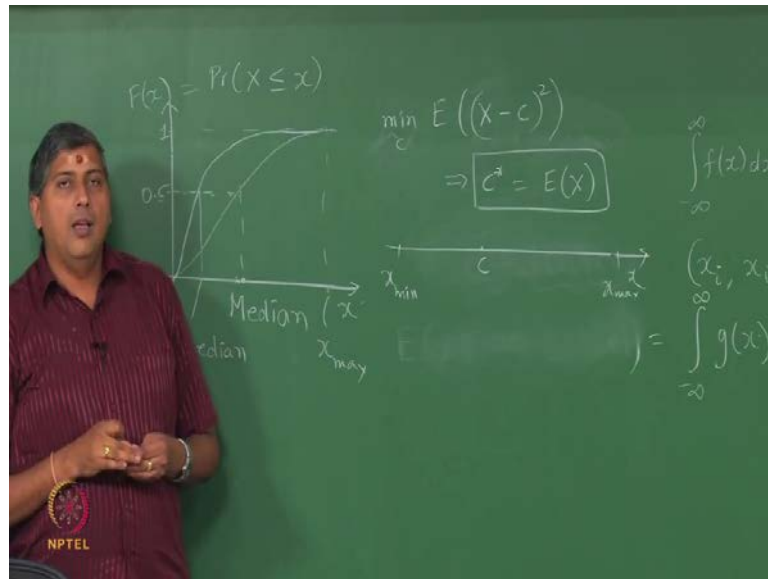


We know that by definition f of x is the probability that x takes on less than or x takes on value less than or equal pre specified values x of x . So, here is a median which need not be it depends on the distribution. Whether this median is exactly suppose here i have the

extreme value now let us say 0 and let say here this is x_{\max} this median need not coincide with the geometrical center. So now, we are talking of statistical center and geometrical center. So, if you look at it exactly biased on the schematic that i have drawn here the geometrical center is here and let us assume that here is the distribution which coincides. That is the median coincide with the geometrical center.

Now, this occurs for symmetric distribution for example, for Gaussian distribution and so on. But you can have another kind of a PDF which for example, here that goes this way and the 0.5 probability is achieved here. So, here is a median, right, this is also legitimate probability distribution function, but now the median is here in this case. The median is not coinciding with the geometrical center and same explanation can be given for mean as well the mean is just a statistical center. Now, it is best understood from a prediction view point suppose your visiting now going back to the city visit example suppose you are predicting in your mind essential that is what we are trying to do we are trying to predict; what would be the temperature on the day I land? For example, what is that single number that I can use as a prediction for the temperature on the day I land and that single number if you call as c , what kind of prediction do you want to make? You want to make best prediction, but best in what sense; well we want to predict in such a way that we want to predict in such a way that we minimize the average error from all possible outcomes.

(Refer Slide Time: 22:33)



For example, here this is the (Refer Time: 22:43) for x and let us say outcomes are spread any were between this x_{\min} and x_{\max} and that outcomes along this continuous up and I want to predict in such way that is c is prediction. Let us say, it is a somewhere here; I do not know where it is. The c is to be in such a way that it is at a minimum distance from all possible outcomes on an average but, which average statistical average. So, this we call as a minimum means square error criteria. So, the mean has very nice interpretation because now the solution to this you can actually solve this you say find c that minimizes this objective or cause function, you will find that c^* (Refer Time: 23:37) nothing but the expectation of x just using standard optimizing techniques you can arrive at this solution.

So, what this tell us when I do not have any other information at all about the random variable and I know, let us all the outcomes or let us say have given me all the outcomes then I simply take the average of this outcome and that is the best prediction it does not mean that, that will be the value of the temperature on the day I land; it is not, but I minimizing the risk that is all that actual realization that will occur, will only occur when I land and that typically will be different from c^* but, if were to look back on the day and ask you know, I have, if I wish, I had made a better prediction, very good prediction

in the minimum mean square error. Since this is the best prediction that is why intuitively perhaps we tend to work with averages, but having said that it is best in some sense that is this mean is best prediction in some sense and not best in some other sense or so but, we will not go into that the purpose of this discussion is only to give you a better feel of the mean, right. As I said, you could also work with medians and so on and there are distribution for which mean is a much mean is much better suited then median and vice versa. Now, for symmetric for Gaussian distribution the mean and median coincide that is something to keep in mind.

(Refer Slide Time: 25:15).

Probability and Statistics: Review - Part 2 References


Mean of discrete RVs

If X is a discrete RV with probability mass function $f(x)$,

$$\mu_X = E(X) = \sum_x xf(x) \quad (2)$$

Example

If X is a binomial RV(n trials, p success probability)

$$E(X) = \sum_i E(X_i) = np \quad (3)$$


Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 8

Let us move on to the case of discrete valued random variables. As I said earlier all that we do in the discrete case is that we replace expectation with sorry the integral in the expectation with the summation; obviously, because now we are not dealing with continuous valued random variables there is no notion of density but rather we have a probability mass function f of x . So, this you should not get confuse with the same notation that you are using you have to interpret f of x in the context if it is a discrete valued random variable f of x is a probability mass function if it is continuous valued random variable. Then it is a density function as an example suppose i am looking at a binomial random variable that is a random variable which can only take which comes out

of a trail, right. What is a binomially distributed random variable? Recall from the previous lecture. I conduct n Bernoulli trials, where I have n trials and in each trial there are only 2 outcomes that are possible with the probability of the success as we call as you label being p and then of course, the other assumptions that this probability remains the same and each trail is independent of the other and so on. Then the number of successes x , successes in n trials with p being the probability of success follows a binomial distribution. So, for such a random variable the average is $n p$ it is fairly straight forward all you do is you invoke the definition that is expectation of x is simply sum of expectation of the individual trails and that comes out to be n times p you can work out the math, if you have difficulty you can ask us this is a standard example that you come across.


(Refer Slide Time: 27:16).

Probability and Statistics: Review - Part 2 References

Variance

Variance: An important quantity in decision making, error analysis of parameter estimation, experimental design and all other forms of data analyses.

Variance
It is the average spread of outcomes around its mean,

$$\sigma_X^2 = E((X - \mu_X)^2) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx \quad (4)$$


Arun K. Tangirala, IIT Madras

Intro to Statistical Hypothesis Testing

9

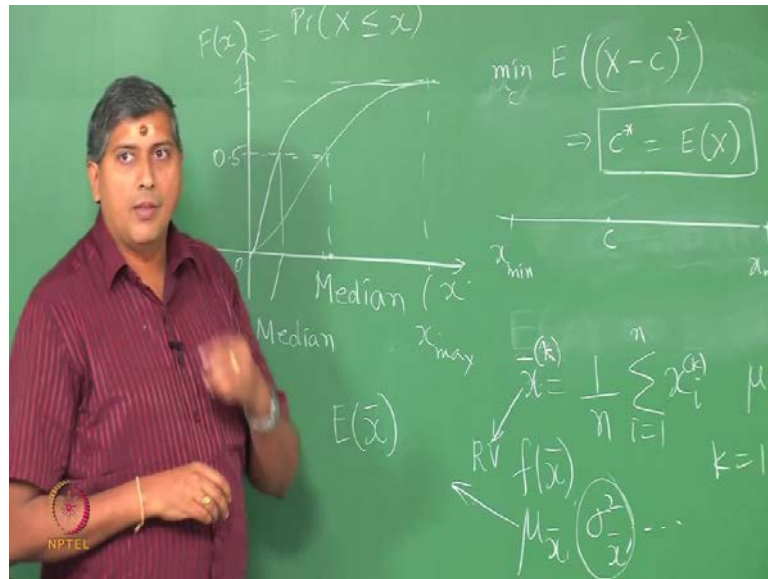
So, mean is the as we say you can say the first not statistic, but say you can say the first population statistic you can say vital statistic of the PDF and so on; that we would be interested in and that is easy to estimate we have of course, gone through only the theoretical definitions. The expression for estimating mean will be given in the following lecture the next quantity of interest was as we have discussed earlier, is a measure of the spread of a outcomes we have learnt how to characterize the center of the outcomes.

Now, we want to know the spread, how far the outcomes are spread around the center and that quantity is variance and this is a fundamental quantity in decision making. Now, at this point I also want to make another aspect of hypothesis testing or statistical inferencing clear, which that until now we have been talking of random variables and we have been thinking of random variables being associated with the random phenomenon; however, random variables can arise in many other situations also.

So, for example, let us say I want to estimate the mean we have seen the theoretical definition of mean how would I go about estimating the mean; I would collect some data. So, if it average temperature that I want to estimate then I would let us say, in this room that I want to measure the estimate the average temperature I would come with the sensor and maybe install it at suitable location collect data, maybe for half an hour and then let us say, I have about 30 samples or maybe I collect every half a minutes I have 120 sorry 60 samples and then am going to take average of all of that because intuitively, we feel that the simple average will give me a good estimate of the true average which I do not know or the population average.

Now, how did we construct the simple average we have constructed this by simply averaging adding up all the observations that I have; however, we should not forget that each observation is a random variable in itself. So, what I am doing is actually adding up all the random variables that I have or all the 60 observations that I have and coming up with the new random variables. So, the sample mean as we call is also random variable. So, that we will also have it is own PDF, it will have it is own mean it will have it is own variances and all the other movements that you can think of. So, now, we should broaden our notion of a random variable from the simply as a outcomes of random phenomenon to operations on random variables, to include operations on random variables as well; when we do such operation as we will learn in the next lecture, the sample mean; the sample mean that we are talking of, \bar{x} .

(Refer Slide Time: 30:21).



Let us say, I have n observations. We just now said that this sample mean as it is normally known sample average is being constructed from n random variables. This also has a same dna as this, that is what is a dna quality of that randomness. So, \bar{x} inherits it is randomness from the observations therefore, this also should be treated as a random variable and it has its own PDF f of \bar{x} or $\mu_{\bar{x}}$ and so on or may be even variance as we shall define and so on. So, it has its own random variable in its own, right.

Now, very often what we are interested in is this $\sigma^2_{\bar{x}}$ as will become very clear in the next lecture because this is a measure of how much error we are incurring in \bar{x} with respect to the truth; the truth is μ_x and the entire purpose of going through this is to get an estimate of μ and $\sigma^2_{\bar{x}}$ is a measure. In fact, the square root of that is called the standard error, is a measure of how far \bar{x} is away from its truth and $\mu_{\bar{x}}$ is how accurate is an \bar{x} as an estimate of μ . So, $\mu_{\bar{x}}$ is also important. In fact, if the mean of \bar{x} , what is mean of \bar{x} ? Very simple, from one experiment I have n observations; from another experiment I have another data record of n observations and so on, so, for each such experiment. So, let us say I conduct r experiments and in the big r and this is the k th experiment. So, this $x_i^{(k)}$ is nothing but the observation from the k th experiment and I have r such experiments.

So, which means for each experimental record I can construct \bar{x} and I have as a result r \bar{x} bars. Now, extend this r to infinity; that means, I have performed infinite numbers of experiments. Why have I performed infinite number of experiments? May be to see all possible outcomes perhaps and if I take the average of all such \bar{x} bars that is here μ which we call as the expectation of the \bar{x} bar, right. So, that is the interpretation of μ \bar{x} bar of course, will talk more about this later on.

So, the variance is define as the average of the distance square distance of the mean from the outcome or outcomes from the mean whichever were you, look at it. So, we say that it σ^2 measures how far the outcomes are spread around the center. So, it is actually a central movement, it is not a pure movement. A pure movement would have been pure second order movement, would have been $\int x^2 f(x) dx$. This is a central second order movement meaning that we are measuring we are calculating this movement around the center now of course, these are very common movements and cumulants in statistics and this σ^2 for random variable if it is with respect coming out of a random phenomena, then it is a measure of the spread of outcomes or as we have just discussed if this random variable corresponds to some value or number that am calculating again out of the observations which we call as statistics. Then, it will measure or it is the measure of the error in average error in the number in \bar{x} bar for example, with respect to the truth now an often when you turn to estimation like we just discussed for example, $\sigma^2 \bar{x}$ would be a measure of how precise an \bar{x} bar is as an estimate of μ and this variance is very important because lower the variance lower is a spread of outcomes.


In fact, the limiting cases when $\sigma^2 \bar{x}$ goes to 0, then what happen to the randomness in \bar{x} ? Well, the random \bar{x} ceases to a random variable and become deterministic again here this variance should not be confused ever with the variability that you see in time for any signal or any data that is variability computed across time here. The variability is being computed across the Ogunnaike across the sample space or the population alright and in defining the variance. We have used the previous expression that I gave you expectation of g of x is $\int g(x) f(x) dx$ and the positive square root of this standard variance is known as a standard deviation.

(Refer Slide Time: 35:36)

Probability and Statistics: Review - Part 2 References

Points to note

- ▶ As (4) suggests, σ_X^2 is the **second central moment** of $f(x)$. However, it can be rewritten as
$$\sigma_X^2 = E(X^2) - \mu_X^2 \quad (5)$$
- ▶ The variance definition is in the space of outcomes. **It should not be confused with the widely used variance definition for a series or a signal (time samples).**
- ▶ Large variance indicates far spread of outcomes around its statistical center. Naturally, in the limit as $\sigma_X^2 \rightarrow 0$, X becomes a deterministic variable.



Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 10

And, you can also re-write the variance as expectation of x square minus μ square clearly telling you that if there is a random variable; which has 0 mean, the second order movement which is expectation of x square and variance coincide and as I said earlier large variance indicates that the outcomes are spread widely far up, far away from its center.

(Refer Slide Time: 36:08)

Probability and Statistics: Review - Part 2 References

Variance of discrete RVs


If X is a discrete RV with probability mass function $f(x)$, then

$$\sigma_X^2 = E((X - \mu_X)^2) = \sum_x (x - \mu_X)^2 f(x) \quad (6)$$

Example

If X is a binomial RV (n trials, p success probability)

$$E((X - \mu_X)^2) = n \sum_i \text{var}(X_i) = np(1 - p) \quad (7)$$

 NPTEL

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 11

Now, as usual what is the variance; we talked about the discrete random variable case. Again, we replace the integrals with summation otherwise, nothing changes you have instated of probability density function probability, mass functions and as an example; if x is a binomial random variable again, where you have n trials and you are looking at x number of successes with p being the probability of success, then you can work out the math expectation of x minus μ whole square is n time sigma variance of the reason. We are simply adding them up is because the trials are independent, you should remember that and then it works out to be $n p$ times i minus p when you work out the math. I am not going through the math because the purpose of this lecture is only to review.

(Refer Slide Time: 36:59)

Probability and Statistics: Review - Part 2 References


Example 1: Variance

Example

Problem: Determine the variance of a RV that follows Gaussian distribution

Solution: The variance is found using (4),

$$\begin{aligned}\sigma_X^2 &= E((X - \mu_X)^2) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx \\ &= \int_{-\infty}^{\infty} (x - \mu_X)^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right) dx \\ &= \sigma^2\end{aligned}$$


NPTEL

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 12

As an example of the variance for the continuous random variable case, you can look at again the Gaussian distribution and when you workout the math it turns out to be sigma square and again, that is the sigma square is nothing but sigma square that we see in the PDF. So, again, we now understand the reason for which we have used this Greek symbol sigma square early on in the PDF.

(Refer Slide Time: 37:31).

Probability and Statistics: Review - Part 2 References


Example 2: Variance

Example

Problem: Determine the variance of a RV that follows a Laplace distribution, the pdf of which is given by

$$f(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right) \quad (8)$$

Solution:

$$[t]\sigma_X^2 = E((X - \mu_X)^2) \quad (9)$$
$$= \int_{-\infty}^{\infty} (x - \mu_X)^2 \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right) dx \quad (10)$$
$$= 2b^2 \quad (11)$$


Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 13

Now, in the case of a Laplace distribution that we have seen earlier the variance works out to be $2b^2$. So, it may be a good idea to go through these examples more in detail for your own insights and for your own learning, these are just examples to make you comfortable with expectations. In general, we are more interested in the statistics and the hypothesis testing part in this course, in a full probability and statistics course we would have more time to go through the detailed derivation of the answers.


(Refer Slide Time: 38:04)

Probability and Statistics: Review - Part 2 References

Mean and Variance of scaled RVs

In statistical data analysis including estimation, identification and prediction, we encounter scaled random variables. It is useful to know how the properties of the scaled variables are related to those of the original ones.

- ▶ Adding a constant to a RV simply shifts its mean by the same amount. The variance remains unchanged (since addition merely shifts the mean and variance is a central measure of spread)
- ▶ **Multiplication:**

$$Y = \alpha X + \beta, \alpha \in \mathcal{R} \implies \mu_Y = \alpha \mu_X + \beta \quad (12)$$
$$\sigma_Y^2 = \alpha^2 \sigma_X^2 \quad (13)$$


Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 14

Now, very often we may work with scaled random variables that is, I may have a variable x and I may construct a new variable y ; which is $\alpha x + \beta$. You will see this often in linear regression for example, where I have 2 random variables x and y and I think of y being linearly related to x . It is not exactly linear, this is called an affine relation linear would be simply y equals αx . So, when I know the mean of x and variance of x and of course, knowing α and β ; can I know the mean of y and variance of y , of course; I can. So, simply apply the expectation operator and remember that the expectation operator is a linear operator as a result μ_y is $\alpha \mu_x + \beta$, all you have to do is apply the expectation operator the expectation of y is α times expectation of x plus β , that is that gets you the answer in 12 equation; in 12 and likewise. The variance of y can be obtain as $\alpha^2 \sigma_x^2$, now what should be observe is that the variance in y is not dependent on β at all because what β is doing is simply causing a mean shift in x , you can say your shifting the mean of course, scale by α as you can see from equation 12, but variances are anyway measures around the center therefore, changing the center should not cause a change in variance that is how you should interpret the results.

(Refer Slide Time: 39:39)


Probability and Statistics: Review - Part 2 References

Properties of Normally distributed variables

The normal distribution is one of the most widely assumed and studied distribution for two important reasons:

- ▶ It is completely characterized by the mean and variance
- ▶ Central Limit Theorem

▶ If x_1, x_2, \dots, x_n are uncorrelated normal variables, then $y = a_1x_1 + a_2x_2 + \dots + a_nx_n$ is also a normally distributed variable with mean and variance

$$\mu_y = a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n$$
$$\sigma_y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2$$


Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 15

And finally, there are many situations where we may add up variables that are Gaussian distributed in a linear way as you can see here suppose I have y . In fact, the notation should have been upper case just note that. So, if I have n random variables a big x 1 a big x 2 up to big x n and I am constructing new random variable which is a linear combination. Assume now that, this n random variables x is falling out of a Gaussian distribution and so, are called unrelated will talk about correlation shortly. Then y is also a non Gaussian or normally distributed variable with the mean being again you just have to extend the results that we have seen in the previous slide where we talked about scale random variables. So, if you simply extend those results you get μ_y as being a linear combination of the means of the respective random variables in the same proportion as y and σ_y^2 being sum of $a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2$; this is quite useful for us later on when we are trying to derive the distribution of \bar{x} .

For example, we saw \bar{x} as being an average of n observations suppose, those n observation are falling out of a Gaussian distribution, using this expression I can calculate the PDF of \bar{x} or of course, once I know the mean and \bar{x} . First, I know if each x is falling out of a Gaussian distribution \bar{x} has a Gaussian distribution and

secondly, the mean can be calculated knowing the means of the respected observations and variance can also be computed in effect I know that \bar{x} then has a Gaussian PDF with the mean given by μ and variance given by σ^2 as shown in slide and variance given by σ^2 again as shown in this slide. So, you see straight away we are able to use this result to compute what is known as a sampling distribution this term is something that will talk about this more in detail in the next lecture

(Refer Slide Time: 42:02)

Probability and Statistics: Review - Part 2 References

Central Limit Theorem

The central limit theorem is one of the classical results in statistics. It is widely used to support the assumption of Gaussian distribution for many random phenomena and is also used to derive distributions of parameter estimates.

Central Limit Theorem

Let X_1, X_2, \dots, X_m be a sequence of independent identically distributed random variables each having finite mean μ and finite variance σ^2 . Let

$$Y_N = \sum_{i=1}^N X_i, \quad N = 1, 2, \dots$$

Then, as $N \rightarrow \infty$, the distribution of

$$\frac{Y_N - N\mu}{\sigma\sqrt{N}} \rightarrow \mathcal{N}(0, 1)$$

One of the popular applications of the CLT is in deriving the distribution of sample mean, which is simply the average of

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 16

And, that also takes us to the very familiar central limit theorem which essentially say, if I have n random variables, but the relaxation here is this n random variables need not come from a Gaussian distribution. Earlier, we had seen for a specifically for a Gaussian distribution, now we are saying what if the observations or these random variables are falling out of a non Gaussian distribution, but all of them falling out of let us say identical distribution same distribution and that they are independent. So, there are 2 relaxation here; one, we are moved from a Gaussian to some general distribution, but they are identical and two, we are moved from uncorrelated to independent. Both of these terms will become clear shortly, when I will take such random variables and add them up to produce a new random variable y subscript n , then as n goes to infinity that is the number of such random variables I am adding up goes to infinity the distribution of y

minus n by $\sigma \sqrt{n}$ follows a standard Gaussian distribution that is very easy first thing for you to show is expectation of y , that is average of y is nothing but n times the average of x because all x 's are falling out of an identical distribution and again you show that the variance of y is nothing but n square sorry, n times σ^2 therefore, the standard deviation y is simply \sqrt{n} times σ .

And we know that, of course, the only thing that is not obvious which is the theorem is telling us is that y when n becomes very large, this new variable y follows a Gaussian distribution that all and that is what we have used in written this theorem of course, the proof of this is available in any standard text in the literature every where this is a result that we will use later on to derive what are known as sampling distribution or distribution on \bar{x} and other statistics.

(Refer Slide Time: 44:06).

Probability and Statistics: Review - Part 2 References

Bivariate analysis


Quite often we will be required to analyze two or more variables simultaneously. Of particular interest would be to examine the presence of linear dependencies, i.e., correlations, and develop linear models.

In all such situations, we start to think of **joint probability density functions**, which aid in the computation of probabilities as in the 1-D case.

We shall restrict ourselves to the two-variable case.

Examples:

- ▶ Height and weight of an individual
- ▶ Temperature and pressure of a gas



Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 17

So, with that we come to a close on the review of the probability and the movements of a PDF. There are other movements of a PDF that one can think off such as, a Skewness; which is a third order central moment or Kurtosis skewness is the measure of symmetry for example. And, for a Gaussian distributions skewness is 0, because it is a symmetric distribution and kurtosis is a measure of another feature of PDF and so on.

But, we do not worry ourselves with higher order movements because it is sufficient to know and in this 10 hour course that is all we can discuss. Now, we move on to the bivariate analysis. As I had mention early on in the lecture very often, we may be dealing more than one random variable as a linear regression for example, in thermodynamics we know that temperature and pressure of an ideal gas are related at a fixed volume. So, I have temperature measurements, I have pressure measurements; both have randomness in them and I want to analyze them jointly. I want to see if temperature is linearly has a linear relation with pressure as predicted by the ideal gas law for example, then what would I do? I have to now analyze this temperature and pressure jointly or it is believe that height and weight of an individual in bio medical field and in medical field you know that the height and weight of individual are believed to be related to each other, so I have an individual for which I have taken height and weight am looking at an entire population. I would like to know how they are related; again, I have to analyze them jointly. When can I analyze these individually? When there is no relation between them, but we are not interested in such situations anyway.

(Refer Slide Time: 45:55).

Probability and Statistics: Review - Part 2 References

Joint probability density function

Consider two continuous-valued random variables X and Y . The probability that these variables take on values in a rectangular cell is given by

$$Pr(x_1 \leq x \leq x_2, y_1 \leq y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(x, y) dx dy$$

Associated with this joint probability (density function), we can ask two questions:

- What is the probability $Pr(x_1 \leq X \leq x_2)$ regardless of the outcome of Y and vice versa? (**marginal density**)
- What is the probability $Pr(x_1 \leq X \leq x_2)$ given Y has occurred and taken on a value $Y = y$? (**conditional density**)
 - ▶ Strictly speaking, one cannot talk of Y taking on an exact value, but only of values within an infinitesimal neighbourhood of y .

Anun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 18

So, for this purpose we introduce typically work with what is joint probability PDF because now, it is like two students working on a project together, we say that they have

probably linearly influenced thoughts then we have to evaluate them jointly becomes difficult to evaluate them independently. So, the same case here. Now, we are dealing with continuous valued random variables, all the notions of density, distribution and so on carry forward except that, in place of univariate we have joint functions, that is all.

So, the density; joint density now once again helps me in computing the probability that these 2 variables, again it should have been X and Y ; big X and big Y taking on values on between x_1, x_2 and y_1, y_2 respectively. So, now, we are looking at a cell we are not we are looking in two-dimensional plain, we moved from a single dimensional real axis to two-dimensional real axis assuming they are all real value. Now, we can ask many questions in this situation among which 2 are of interest to us. First of all we can ask what is a probability that x can take on values between x_1 and x_2 , regardless of the outcome of; I do not care what is a weight on individual, I randomly pick individual what is a probability that the height of the individual is between 5 and 6 feet for example, I can ask that, it is a very valid question to ask and vice versa. Here is where we come across a notion of marginal density. On the other hand, I can also ask what is a probability that given, that the weight of an individual is 50 kgs, that the height is between 5 and 6 feet. Now, obviously, that is going to be different from the earlier one, where we are looking at unconditional probability.

So, in the first case we are talking of unconditional probability, now we are talking of what is known as conditional probability. If you recall we have talked about this in the previous lecture as well. When would these two be identical, the unconditional, conditional probability? Intuitively, when x and y are independent of each other; when height and weight for example, have no influence on each other then the conditional-unconditional would coincide. Of course, there are a strict theoretical considerations when you talk of saying given that the weight of individual is exactly 50 kgs, but strictly speaking we are not suppose to be talking that way, because as we said in the previous lecture, where for continuous random variables we do not talk of probabilities that a random variables takes on an exact value. The way this conditional probability should be interpreted therefore, is what is a probability that the height of an individual is between 5 and 6 feet; given that the weight of an individual is in very small vicinity of 50 kgs, but we do not keep saying that, but you have remember that in your minds.

So, with that we introduce a notion of Marginal density and Independence, right; we have just spoken of that.

(Refer Slide Time: 48:59)

Probability and Statistics: Review - Part 2 References

Marginal density and Independence


The marginal density is arrived at by walking across the outcome space of the "free" variable and adding up the probabilities of the free variable taking values within infinitesimal intervals.

Definition

The marginal density of a RV X with respect to another RV Y is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad (14)$$

Two random variables are said to be **independent** if and only if

$$f(x, y) = f_X(x)f_Y(y) \quad (15)$$


Arum K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 19

Marginal density is like on your page of book is two dimensional, we have margins. So, remember this. When we walk along with one dimension, and adding up all that we encounter across another dimension we end up with marginal density. So, the marginal density certificantly denoted; for example, the marginal density of x is denoted by f of x subscript x . That subscript x is telling you that it is marginal density, it is not necessarily the density of x alone. There is a difference between marginal density and just PDF of x ; the difference is that the marginal density is obtained by adding up or integrating the density along the y access alone whereas, the density of x is taking into account probably all other factors also that are related to x . So, there is a difference and therefore, it is important to maintain that notation, f of x subscript x . Likewise, you can define marginal density of y as well.

Now, we say that two random variables are independent, if the joint PDF can be written as a product of the marginal density, respective marginal density. This has a strong relationship or resemblance to what we wrote in terms of probability, we say 2 events are independent if probability of A and B occurring together is, probability of A times

probability of B. Once again, do not confuse your independence with mutual exclusive events, that is different. Mutual exclusive events, probability 0, that both will occur together. And, as we spoke of conditional density earlier were we said, given the weight of an individual is 50 kgs; what is a probability that the height is between 5 and 6 feet, such probabilities can be calculated by the use of what is known as conditional densities.

(Refer Slide Time: 50:57)

Probability and Statistics: Review - Part 2 References

Conditional density


The conditional density is used in evaluating the probability of outcomes of an event given the outcome of another event

Example: What is the probability that Rahul will carry an umbrella given that it is raining?

Conditional Density

The conditional density of Y given $X = x$ (strictly, between x and $x + dx$) is

$$f_{Y|X=x}(y) = \frac{f(x, y)}{f(x)} \quad (16)$$



Arun K. Tangaia, IIT Madras Intro to Statistical Hypothesis Testing 20

And this conditional density expression is arrive by using base theorem or base result, which says that it the joint PDF divided by the marginal density; again, there should have been a subscript there, marginal density of x. Likewise, you can. This is the conditional density of y given x, alright. Likewise, you can write f of x given y, which is f of x comma y by f of y. And now, we can go back and relate this to independence that is suppose 2 variables are independent, then we just said earlier that the conditional probability and the unconditional probability would be identical; likewise, conditional density and unconditional densities would also be identical. I am just stating this without any proofs and so on. So, f of y given x would be the same as f of y or f of x given y would be the same as f of x; now, using this definition and the other relation that we just said is straight forward to show that f of x comma y would be simply f of x times f of y.

So just combine both and we will get that. So, that is the other definition of that is the definition of independence that we saw in equation 15.

(Refer Slide Time: 52:29).

Probability and Statistics: Review - Part 2 References

Covariance

One of the most interesting questions in bivariate analysis and prediction theory is if the outcomes of two random variables (linearly) influence each other, if they co-vary.

The statistic that measures the co-variance between two RVs is given by

$$\sigma_{XY} = E((X - \mu_X)(Y - \mu_Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_X)(y - \mu_Y)f(x, y) dx dy \quad (17)$$

Alternatively,

$$\sigma_{XY} = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - E(X)E(Y) \quad (18)$$

- ▶ Clearly, if $Y = X$, we obtain variance of X , the spread of outcomes of X
- ▶ Covariance is a measure of joint spread of X and Y

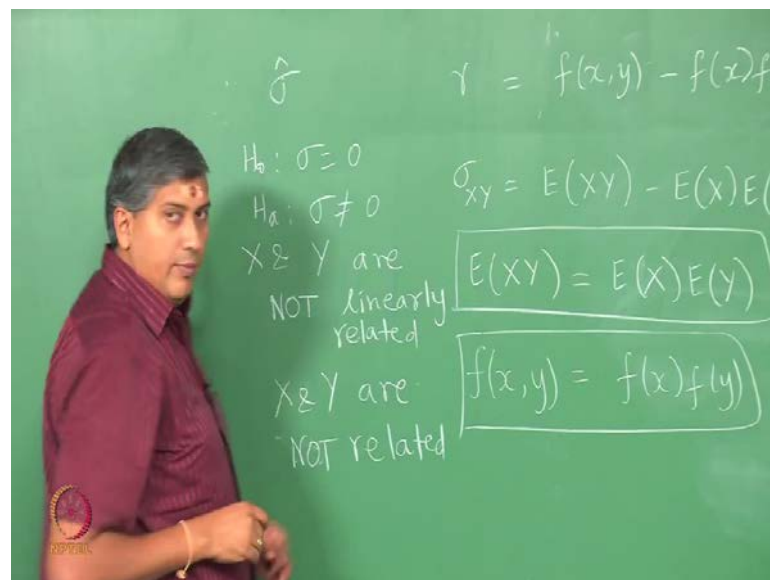
Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 21

Now, for the final part of this lecture; we talk about co-variance or correlation which is the most measure that is used in data analysis. So, if you are going to continue in data analysis then you should be really clear about the notion of correlation, what it does? What was it originally devised for? And, how is it defined? Right; the basis for the definition of correlation is co-variance. So, we have talked about independence earlier. Independence rules out any form of relation between x and y . And, we measure theoretically at least we say that, we measure that by looking at the joint density and the difference of that from the product of the marginal densities. If the difference is 0, then they are independent; if the difference is not 0, then there is some dependence. So, the question is, what form of dependence exists between for example, height and weight of an individual or in the examples that we discussed the highway mileage and the engine capacity? So, so many other examples and so on.

In this course and in general, we are interested at least to begin with linear dependence; is there a linear relation between two random variables. The advantage is, if there is a

linear relation then I can build linear models and the math associated with the estimation of linear models is so easy compared to the math associated with estimation of non-linear models. So, the entire modeling or the regression literature is very rich for linear models therefore. A co-variance is a measure of the linear dependence between two random variables, x and y . You may not prove that, but will discuss that a bit in detail. First, the definition of co-variance; it is now the second order central moment of the joint PDF, earlier we talked of variance and in fact, you can see from the definition in 17 for the co-variance; if I set y equals x , then it simplifies to the definition of variance and you can see it is a double integral, again minus infinity and plus infinity are only symbolic, they are the respective extreme values for x and y . And, μ_x and μ_y are the means obtained from marginal densities of x and y , respectively. Alternatively, you could write if you do not like those double integrals in equation 17, you could write σ_{xy} as expectation of xy minus expectation of x times expectation of y . Now, this is very nice resemblance between this and the expression that I spoke earlier in terms of PDF's let me point that out.

(Refer Slide Time: 55:16)

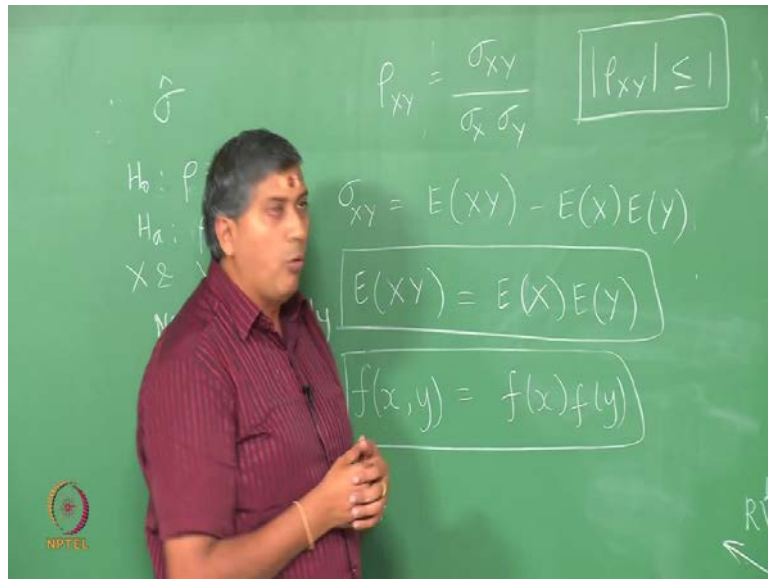


Well, introducing co-variance, that independence theoretically can be thought of as a difference between $f(x, y)$ minus $f(x)$ times $f(y)$; if this is 0, so let us say that

there is gamma measure which does this, let us just (Refer Time: 55:38) purposes of discussion, this is not really used in practice. On the other hand, looks at the difference between expectation of the product and the product of expectations and we say that two variables have a 0 co-variance when σ_{XY} is 0 that is when expectation of X times Y is a same as product of expectations. Look at the resemblance of this, they are not identical because here you are looking at a product this is just a joint PDF, but I am just trying to point out the similarities between these 2 expressions. If this is the case, then X and Y are not linearly related. If this is the case, then X and Y are not related at all; obviously, this is a much stronger condition compare to this. But in practice since we are varied about linear dependency we worked with this, so this is of course, theory in practice will have to use estimators for co-variance and see if that estimate is very small and subject to hypothesis test that is exactly what we are going to do in hypothesis testing. I am going to estimate sigma and I get sigma hat in fact, and then conduct the hypothesis test whether of, whether the true sigma is 0, again say alternative sigma is not equal to 0.

Now, there is a problem with this kind of hypothesis test bit primarily, because if you look at the definition of sigma it depends on the units of X and Y that is the main problem, right. And, second problem is, I do not know what is high sigma, low sigma and so on. But, we will not worry about so much as much as it is dependence on the units and it is with that purpose that, normally one works with correlation which is a normalized co-variance.

(Refer Slide Time: 58:05)



So row XY is a normalized co-variance which address 2 issues in one shot. One, that now this has become unit independent of X and Y; so if you look at temperature and pressure, whatever units you use for temperature and pressure the correlation comes out to be the same value whereas, sigma would work out to be a different value depending on the units. Moreover, you can show that correlation is bounded above by unity in magnitude. So, the maximum possible value for correlation, magnitude of correlation is unity. And then, with this correlation; I know when sigma is 0, row is 0. So, conducting a hypothesis test on sigma 0, can be replaced with row being 0 and alternate hypothesis being row naught equal to 0, that is exactly the kind of hypothesis test that we are going to study later on, alright. And, we will talk about a bit more about this correlation; what does equality, what does it mean when correlation becomes equal to 1.

(Refer Slide Time: 59:12)


Probability and Statistics: Review - Part 2 References

Correlation

Two issues are encountered with the use of covariance in practice:

- Covariance is sensitive to the choice of units for the random variables under investigation. Stated otherwise, it is **sensitive to scaling**.
- It is **not a bounded measure**, meaning it is not possible to infer the degree of the strength of the linear relationship from the value of σ_{XY} .

To overcome these issues, a normalized version of covariance known as **correlation** is introduced:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (19)$$


Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 22

We just spoke of this correlation as a normalized measure.

(Refer Slide Time: 59:15).

Probability and Statistics: Review - Part 2 References

Properties of correlation

Correlation enjoys all the properties that covariance satisfies,
In addition, correlation is a bounded measure, i.e., $|\rho_{XY}| \leq 1$

Boundedness

For all bivariate distributions with finite second order moments,

$$|\rho_{XY}| \leq 1 \quad (20)$$

with equality if, with probability 1, there is a linear relationship between X and Y .

Result can be proved using **Chebyshev's inequality**

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 23

And, correlation enjoys all the properties that the co-variance satisfies. In other words, if the co-variance is a measure of linear dependence, correlation is also a measure of linear dependence and is inequality, that correlation is bounded above in magnitude by unity can be proved by Chebyshev's inequality; again the proof is available quite widely.

(Refer Slide Time: 59:37)

Probability and Statistics: Review - Part 2 References

Unity correlation

Correlation measures linear dependence. Specifically,

- $\rho_{XY} = 0 \iff X$ and Y have no linear relationship (non-linear relationship cannot be detected)
- $|\rho_{XY}| = 1 \iff Y = \alpha X + \beta$ (Y and X are linearly related with or without an intercept)

Assume $Y = \alpha X$. Then, $\mu_Y = \alpha \mu_X$; $\sigma_Y^2 = \alpha^2 \sigma_X^2$

$$\begin{aligned} \rho_{XY} &= \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} \\ &= \frac{\alpha(E(X^2) - (E(X))^2)}{|\alpha| \sigma_X^2} \\ &= \frac{\alpha}{|\alpha|} \\ &= \pm 1 \end{aligned}$$

NPTEL
Aran K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 24

And in particular, we are interested in asking, what does correlation hitting a value of one mean? Correlation, when it takes a value of 0 we know X and Y are not linearly dependent which means, some non-linear dependence may exist. So, then one has to conduct tests of non-linear independence, will not worry about that. The fact is when correlation hits a value of unity in magnitude, then this is only possible when Y and X are linearly related, that is, this is a both necessary and sufficient condition. I am only showing one part of it here for your convenience, that is if Y is related linearly or affine way alpha X plus beta 2 X then, correlation is 1. Of course, without loss of generality for simplicity sake will assume beta to be 0. So, assume Y is alpha X and then you work out the expectation we have already seen if Y is alpha X, mu Y is alpha mu X, sigma square Y is alpha square, sigma square X. This is something that we have seen earlier. So, we are just using a known result and plugging in for row. The expectation of X comma y is alpha times expectation of X to the whole square minus, sorry, it should be expectation X

square minus expectation of X whole square, divided by the product of sigma X times alpha times root X but positive alpha because we know that sigma Y is the positive square root of the variance of Y. So, when you work out all of that, you get row to be plus or minus 1 depending on the sign of alpha. So, the magnitude works out to be 1. Of course, I have only shown that if Y is alpha X or alpha X plus beta then row is 1 but, you can show the other way round also, row equals 1 implies that Y is alpha X plus beta. So, this is a fundamental result in data analysis because it tells you that when correlation hits a value of 1 or estimates are very high, very close to 1 then a linear model will do a very good job of predicting Y given X; that means, you can live with the linear model as far as prediction of Y is concerned using X.

If correlation values are low then what does it mean? Well, what it means is that there is a deviation from the linearity assumption and that deviation could be presences of noise, presences of non-linearity and so on, we do not discuss that at this movement. This is just to introduce a review, sorry, review concepts of correlation. And, we have just said that two variables are uncorrelated, if we is to retreat, if the correlation is 0 or expectation of XY is the same as product of expectations.

(Refer Slide Time: 62:30)

Probability and Statistics: Review - Part 2 References


Uncorrelated variables

Uncorrelated variables

Two random variables are said to be **uncorrelated** if $\sigma_{XY} = 0 \implies \rho_{XY} = 0$.
 Alternatively, since $\sigma_{XY} = E(XY) - E(X)E(Y)$, the condition also implies

$$E(XY) = E(X)E(Y) \tag{21}$$

- ▶ Uncorrelated condition merely rules out the presence of linear relationship between X and Y .
- ▶ Determining the absence of non-linear dependencies requires the test of **independence**.



Arun K. Tangirala, IIT Madras

Intro to Statistical Hypothesis Testing

25

But, this is a weaker requirement compare to independence.

(Refer Slide Time: 62:32)

Probability and Statistics: Review - Part 2 References


Independence

Independent variables

Two random variables are said to be **independent** if and only if the joint pdf is factorizable into marginal pdfs

$$f(x, y) = f(x)f(y) \quad (22)$$

An alternative statement is in terms of the conditional pdf.
Two variables are independent if and only if

$$f_{Y|X=x}(y) = f_Y(y) \quad (23)$$


Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 25

What this means is, independence implies uncorrelatedness, but uncorrelatedness does not imply necessarily imply independence. Now, the result is that, if X and Y follow a Gaussian distribution that is a bivariate Gaussian distribution.


(Refer Slide Time: 62:51)

Probability and Statistics: Review - Part 2 References

Independence vs. Uncorrelated variables

Independence \implies Uncorrelated condition but NOT vice versa.
Thus independence is a stronger condition.

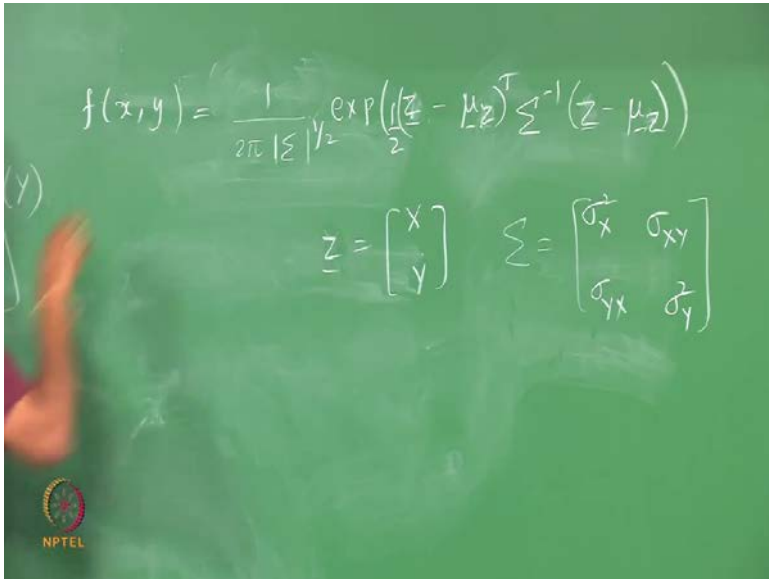

If the variables X and Y have a bivariate Gaussian distribution, Independence \iff Uncorrelated condition.
Therefore, in all such cases, independence and lack of correlation are equivalent.



Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 27

The bivariate Gaussian distribution has a different expression from what we generally think of is an individual Gaussian, there is similarity but they are a bit different and I will just write that for you.

(Refer Slide Time: 63:04)


$$f(x, y) = \frac{1}{2\pi |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\underline{z} - \underline{\mu}_z)^T \Sigma^{-1} (\underline{z} - \underline{\mu}_z)\right)$$
$$\underline{z} = \begin{bmatrix} x \\ y \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}$$


So, the bivariate Gaussian distribution; if two random variable x and y have a joint Gaussian distribution their ϕ times determinant of σ , I will tell you what is this σ ; square root of a determinant of σ times exponential. Let us say, this there is a vector X ; I am going to tell you what is that vector X and this is a vector mean, transpose σ inverse times x minus μ_x and there is half here. So, what is this σ and what is this X ? First, let me write what X is for you. X is nothing but the vector of random variable X and Y or you may think of this as Z , if you do not like X . So, Z is as a vector of random variables X and Y and σ is the variance co-variance matrix that normally one sees in the vector random variable case. So, it has the variance of the individual random variables along with a diagram and the co-variance on the off diagonal of the matrix.

So, now, this is a symmetric positive definite matrix. It is symmetric because σ_{XY} is same as σ_{YX} , co-variance is insensitive to the ordering of the subscripts that means, it does not know whether X causes Y or Y causes X ; it simply looks at the dependence between X and Y ; therefore, it is a symmetric matrix and further it is positive definite. So, you can guarantee therefore σ inverse exist. That means, its determinant is not 0 unless σ^2_X or σ^2_Y is 0 and some what, we will rule out such pathological cases. Then, if $f(x, y)$ is this then X and Y are say, this of course is a determinant, then we say x and y have a joint Gaussian distribution. Now, a settle point is if X and Y are individually Gaussian distributed, it does not necessarily mean that joint distribution is also Gaussian; examples can be given, but this is just a point for your remembrance.


(Refer Slide Time: 65:49)

Probability and Statistics: Review - Part 2 References

Independence vs. Uncorrelated variables

Independence \implies Uncorrelated condition but NOT vice versa.
Thus independence is a stronger condition.

If the variables X and Y have a bivariate Gaussian distribution, Independence \iff Uncorrelated condition.
Therefore, in all such cases, independence and lack of correlation are equivalent.



Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 27

So, now when that is the case when X and Y have a joint Gaussian distribution and they are independent; and they are uncorrelated. Then, you can show that they are also independent; that is what happens when X and Y are uncorrelated, the variance covariance matrix that big sigma that we have written on the board, is a diagonal matrix; that means, σ_{XY} is 0, in which case it becomes a diagonal matrix, right. Now, you can show that, if that is the case you can factorize f of X comma Y as f of X times f of Y and that means, X and Y are also independent but this is only true for the bivariate Gaussian distributed case. In fact, you can extend this to a multi-variate Gaussian distributed also. So, any set of random variables that have a joint Gaussian distribution and are uncorrelated that means, there is no linear relationship between them; then it also means there is no non-linear relation, absolutely, no relation between them at all and that is the specialty that you say about a Gaussian distribution. It does not apply to any other distribution necessarily. So, with that we come to the closure of the review of probability and statistics. To summaries what we have looked at is in concepts of probability.

(Refer Slide Time: 67:06).

Probability and Statistics: Review - Part 2 References

Bibliography I

-  Bendat, J. S. and A. G. Piersol (2010). *Random Data: Analysis and Measurement Procedures*. 4th edition. New York, USA: John Wiley & Sons, Inc.
-  Johnson, R. A. (2011). *Miller and Freund's: Probability and Statistics for Engineers*. Upper Saddle River, NJ, USA: Prentice Hall.
-  Montgomery, D. C. and G. C. Runger (2011). *Applied Statistics and Probability for Engineers*. 5th edition. New York, USA: John Wiley & Sons, Inc.
-  Ogunnaike, B. A. (2010). *Random Phenomena: Fundamentals of Probability and Statistics for Engineers*. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group.
-  Tangirala, A. K. (2014). *Principles of System Identification: Theory and Practice*. CRC Press, Taylor & Francis Group.



Arun K. Tangirala, IIT Madras

Intro to Statistical Hypothesis Testing

28

We have looked at axioms of probability, then we learnt; what is a notion of a random variable, probability distribution function, density function, mass function and then we said it is hard to work with PDF in practices; therefore, we will work with the moments of the PDF, mean and variance which is, what we have we have seen also in the examples, is that average, variability this is what we spoke of; in none of the examples in hypothesis testing, that we looked at we tested the PDF, we talked about the PDF if you recall. And, that is the case also in reality.

And then, we went on to learn or what are the interpretation of mean and variance in different context and finally, we looked at the bivariate case where we learnt that, for the bivariate case or even the multivariate case one has to work with the joint PDF. Again, the same story in the univariate case we said the PDF's are hard to work with therefore, we went to the movements case. So, in the bivariate case from the joint PDF we straight away went to the second order movement, which is co-variance and learnt that the co-variance or correlation, which is a normalized measure, is a measure of linear dependence; is a very good measure of linear dependence because when it hits a value of unity that is the correlation, when it hits the value of unity then there is a perfect linear relationship between those two variables.

Any deviation from that would mean that there is something else in Y which X cannot linearly explain, that something else in Y could be noise and or some non-linear function of X and finally, we also talked about independence which is a measure of lack of any relation ruling out all forms of dependence between Y and X . But, we will not worry about that any more. In this course as far as testing is concerned, however the notion of independence is very important because in the next lecture when we talk of what is known as the random sample and there we would talk of the joint PDF of the observations that we have obtained and a random sample would mean that the outcome of the first observation will not affect the outcome of any other observation. So, those n observations that we collect in random sample are going to be independent and with this, with that idea we have reviewed the notion of independence early on.

So, that is it. There is a tutorial that will also put up, go through the tutorial and then as far as the theory, the lecture are concerned we will get back again with the concepts of, what are known as statistics and sampling distribution, where we will talk of sample mean, sample variance, sample proportion and so on. And, sample correlation, alright, then.

So, hope to see you again soon in the next lecture.