

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

Lecture - 08
Statistics for Hypothesis Testing - Part 2

Welcome to this lecture. This is the second and final lecture on sampling distributions, I should say. Well, there is a part of the sampling distribution with respect to sample correlation that will discuss later, but more or less we would be covering all the sampling distributions that we require for hypothesis testing through this lecture. And again, in this lecture, what we are going to look at is a continuation of what we discussed previously.

(Refer Slide Time: 00:47)

Statistics for Hypothesis Testing - Part 2 References

Learning objectives

Sampling distribution of

- ▶ Difference in sample means
- ▶ Sample variance
- ▶ Ratio of sample variances
- ▶ Sample proportion
- ▶ Difference in sample proportions



Prof. Arun K. Tangirala, IIT Madras

Intro to Statistical Hypothesis Testing

2

We were going to look at sampling distributions of difference in sample means, sample variance, ratios of sample variances, sample proportions and also difference in sample proportion. Now, all of these are kind of in line with examples that we discussed in the motivation lecture. And let me also tell you up front that in this lecture mostly it is a statement of results, neither we will try to prove anything nor use many standard results that are available and so may be for a few will use some standard results, but will not attempt to prove anything. And, anyway in 10 hour course there is not much time to prove

anything, and we are not interested in that we want quickly get to hypothesis testing. And of course, just to reiterate it is important to study the sampling distributions for hypothesis testing because after all the entire test of hypothesis rests on the knowledge of the sampling distribution of the test statistics that we use. In the previous lecture, we looked at sampling distribution of mean under different conditions known variance unknown variance and so on.

(Refer Slide Time: 02:04)

Difference in sample means

It is frequently required to compare means of two different populations. In such cases, a statistic based on the difference in sample means is used.

Sampling distribution of a difference in sample means

Suppose we have two independent populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 . Further, if \bar{X}_1 and \bar{X}_2 are the sample means of two independent random samples of sizes n_1 and n_2 from these populations, then the sampling distribution of

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \quad (1)$$

is approximately standard normal, if the conditions of CLT apply.

If the two populations are normal, the $f(Z)$ is exactly standard normal.



Now, we look at difference as in sample means again I would advise you to go back and relate to the examples that we discussed in the motivation lecture. At least one example involved comparison of means, right, where we were looking at the nylon connector example. So, of course, this situation arises in many, many other applications as well. So, let us come to point here when I have 2 populations of different means, let us say μ_1 and μ_2 , I would like to know for example, if indeed they are different or not, and for this purposes we setup a test statistic. And this test statistic has to be commensurate with what I want to test, what I want to test is whether μ_1 which is the mean of population 1 and μ_2 which is that of population two are different from each other or not. So naturally the test statistic involves differences in the estimates of means and once again as I said in the previous lecture, we are going to use only the sample mean as an estimator of the mean; however, that should not preclude you from or prevent you from using other estimators.

So, coming back to the sample mean as an estimator. If \bar{X}_1 and \bar{X}_2 are sample means computed from samples of the respective populations of sizes n_1 and n_2 and known variances. So, we assume in the beginning a simpler situation, where the variances of the populations are known, namely σ_1^2 and σ_2^2 then this test statistic $\bar{X}_2 - \bar{X}_1 - \mu_1 + \mu_2$ divided by square root of $\sigma_1^2/n_1 + \sigma_2^2/n_2$ is approximately standard normal, if the conditions of the central limit theorem apply. What are the conditions of central limit theorem, well the observations have to follow out of an identical and independent distribution, right.

Now, as usual we usually talk of the distribution of the standardized one and Z is also the standardized one; recall in the previous lecture why we talk of this standardized test statistic. One advantage is we can use a standard distribution. Now, what if the variances are not known or the populations are normal and so on, this results that you seen equation one is of it for a general population that is X_1 and X_2 that is the population one and population two can be characterized by any distribution in this results. However, if the populations are characterized by Gaussian distributions then Z follows an exact standard normal distribution.

(Refer Slide Time: 05:12)

Unknown variances: Difference in means

When the population variances are unknown, the problem of determining $f(\bar{X}_1 - \bar{X}_2)$ is quite complicated, especially when $n_1 \neq n_2$.

Large sample sizes: Unknown variances

The statistic

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \quad (2)$$

is **approximately** standard normal.

The remaining two cases pertain to the **small sample sizes** with (i) $\sigma_1 = \sigma_2$ and (ii) $\sigma_1 \neq \sigma_2$.

Suppose, I do not know the variances which is the more realistic case then what do I do, well I adopt the same approach that I adopt in the case of sample mean that is testing of single sample mean. We

replace the theoretical variances with the respective estimates denoted by S_1^2 and S_2^2 ; once again we have not talked about how to estimate sample, how to estimate variances yet we will do so shortly. So, the result is again more or less the same; again this is for the large sample case, you should remember. The test statistic that is given in equation two is approximately standard normal. The only difference between the test statistic in 1 and 2 is that, we have replaced the theoretical variances with the respective estimates. And, what will do next is look at the small sample case, but under two different situations; variances are equal and variances are unequal. This is possible that 2 populations can have same variances.

(Refer Slide Time: 06:19)

Small sample size, equal variances: Difference of means

Equal variances

The statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2) \quad (3)$$

where S_p is the pooled standard deviation

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad (4)$$

How does it make a difference? Let us look at this. So, now as you must have guessed like in the sample mean case when the variance was unknown, and when the samples were small, the test statistic followed T distribution there. Here also the test statistic based on the differences in sample means of course, the standardized test statistic follows T distribution with $n_1 + n_2 - 2$ degrees of freedom. So, you can say well, we lose one degree of freedom when we estimate sample mean from population one and another degree of freedom when we are estimating sample mean for another population. So, in total, we have two degrees of freedom loss, therefore the total degrees of freedom that we have when it comes to estimating the variance is $n_1 + n_2 - 2$.

We have assumed the variance to be equal or variances to be equal, and therefore, we construct what is known as a pooled variance. What happens is when the variances is equal, we are exploiting this situation to our advantage; from population one, we have sample of size n_1 ; from population 2, we have a sample of size n_2 . More the number of observations to estimate a variance (Refer Time: 07:42) situation. Therefore, since we are already given that the variances are equal; we pooled the data together to estimate the common variance and that is the expression given to you in equation 4 for the sample variance. Where you can see in the denominator we have $n_1 + n_2 - 2$ that is nothing but your degrees of freedom, alright.

So, this $n_1 - 1$ and $n_2 - 1$ we already know. What about S_1^2 and S_2^2 ? S_1^2 is a sample variance estimated from the first population; S_2^2 is a sample variance calculated from the second population. And of course, again I say here, we have not yet seen the expression for calculating S_1^2 and S_2^2 will do so very soon. Just to tell you that we are going at probably slightly rapid base because as I said early on in the lecture, we are looking at statement of results more than trying to prove anything. And, at some point this may get boring to you, but it is an inevitable devil you can say that we have to deal with. So, what about the small sample size, unequal variances case; again the same story, because it is a small sample size the test statistic would follows T distribution.

(Refer Slide Time: 09:20)

Small sample size, unequal variances: Difference of means

Unequal variances

The statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}} \quad (5)$$

is **approximately** t -distributed with $\nu = (n_{12} - 2)$ d.o.f., where

$$n_{12} = \left\lfloor \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1 + 1} + \frac{(S_2^2/n_2)^2}{n_2 + 1}} \right\rfloor \quad (6)$$

And in this case, here again we are looking at a non-Gaussian case and that is why it is an approximately T distributed. The difference between this and a previous cases the variances are unequal, therefore, I cannot pool the variance estimates to come with the common variance. And, the calculation is a bit more involve as you can see particularly in the calculation of degrees of freedom denoted by $n_1 + n_2 - 2$, the $n_1 + n_2$ calculation is quite complicated compare to the previous case. But, we have to leave it there were no worries I mean most of the times the computer calculates set for you nevertheless if you have to do it by hand then you have to use this expression no other choice. But in the end, what is important is to know also the type of distribution it is a T distribution that does not change fortunately. And, as usual we have $\bar{X}_1 - \bar{X}_2$ as the estimate of the difference in means.

So, in all of this something should evolve for you that what we are doing is we first identify the parameter of interest in this case the difference in means then identify the test statistic before between identifying the parameter and identifying test statistic there is an intermediate step which is choosing the statistic or the estimator itself. We can say the estimator here is $\bar{X}_1 - \bar{X}_2$ for $\mu_1 - \mu_2$, and the test statistic is what you see in each of this, so that is the routine that we will see in the remaining examples, therefore, you should not have any difficulty in terms of understanding the procedure. If any difficulty that you may have that would be in remembering the distributions, but that is all right in most of the hypothesis testing exercises. Typically, we do not have the necessarily remember the sampling distribution let us say in an exam I have to remember I do not have to I can be actually given that information on a simple sheet and perhaps we will do that for you in the exam, it is not a difficult thing to do.

What is important is to understand the concepts of hypothesis testing that there is a sampling distribution of the statistic that you dealing with and that goodness of the entire hypothesis test realize on the sampling distribution apart from a few other factors such as sample sizes and variances and so on. So, with this we kind of close the means case, but there is one more example, I am sorry, there is one more example that is going to come up and with that will close the means and that has got to do with the paired difference of means.

(Refer Slide Time: 11:58)

Paired difference of means

In several applications, we are interested in comparing averages "before" (corresponding to X_1) and "after" (corresponding to X_2) a treatment or some processing step. In such cases, the random samples for X_1 and X_2 do not meet the usual independence requirement.

Define $D_i = X_{1,i} - X_{2,i}$. Subsequently, introduce

$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n}, \quad S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2 \quad (7)$$

The statistic

$$T = \frac{\bar{D} - (\mu_1 - \mu_2)}{S_D/\sqrt{n}} \sim t(n-1) \quad (8)$$



In the previous 2 or 3 situations, we had looked at differences in means from 2 different populations; assuming that these 2 populations are kind of independent of each other. But now, there are we are looking at a situation where that is not the case; where the 2 populations are not independent, and when thus the situation arise. A standard example that is given is a weight-loss program where I am looking at a set of individuals that is I want to test a weight-loss program and I want to test that it has been effective. How would I test that? Well, I look at the average weight before implementing the weight-loss program and then the average weight after the implementing the weight loss program. So, what I do is I randomly sample a set of individuals, before the weight loss-program, take their weights and get an average estimate of the average weight, and then put them through the weight-loss program and then again do the same calculation, take the weights and compute the average weight.

So, now, obviously, these are the same individuals that are going to be screened for testing right. And there is a high level of dependency and there better be a dependency otherwise, what is the meaning of a weight-loss program. When we conduct test like this, in these kinds of situations, it could be a weight-loss program, it could be a training program to see if there is an improvement in the performance of teachers or students and so on, there also you will end up with this pair difference of means case; where these obviously, the sample sizes are identical for both population, because we are looking at the same set of individuals or same set of specimens you may say. So, there is a one-on-one

correspondence, and you are looking at a pair you are looking at a pair difference therefore. In this case, the independence assumption is violated, and therefore it warrants a separate analysis.

So, what is the result here the first thing is you define this variable d_i as X_{1i} minus X_{2i} that is think of X_{1i} as a weight of the i th individual before the program and X_{2i} as a weight after the program. So, we are not actually testing, it basically amounts to testing the zero mean you can say for d or d_i , we are not individually testing for X_1 and X_2 you can think of it that way. So, now, having defined a variable d_i as X_{1i} minus X_{2i} , we can now more or less think of this as a single sample mean calculation or mean hypothesis testing case. But now in terms of d , so as we did in the single sample case; that means, single population case where we were testing for means here we compute the average mean of the differences, so that is your \bar{D} and then we calculate the variance in a way that we would calculate for a single population.

So, what we have done is essentially taken 2 populations, but because they are dependent on each other they; they have a strong correlation between them. We have clubbed that into single population now described in terms of this difference variable D_i that is all. Now, the result that is given here is for the small sample case; there is no need to specify the large sample case because we know the T distribution tends to Gaussian distribution. So, we do not need to state that separately. The rest of the story is the same. Now, the test statistic $\bar{D} - \mu_1 - \mu_2$ by your S over root n follows T distribution with $n - 1$ degrees of freedom where n is the sample size of the individual population. Here as we said earlier, both populations have to be of identical sorry, both samples have to be of identical size and that is your n , very good. Please keep these distributions handy of course, I will also bring up the respective distributions, when we take up the eight examples that we had discussed in the motivation lecture for hypothesis testing - illustrating the hypothesis test, alright.

(Refer Slide Time: 16:42)

Estimation of variance

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

The theoretical $f(S^2)$ depends on the underlying distributions.

Gaussian distributed $X_i \sim \mathcal{N}(\mu, \sigma^2)$:

$$C = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

General case: Nothing can be said about $f(S^2)$. However, it is an unbiased estimate (average of all possible estimates is the true value):

$$E(S^2) = \sigma^2$$



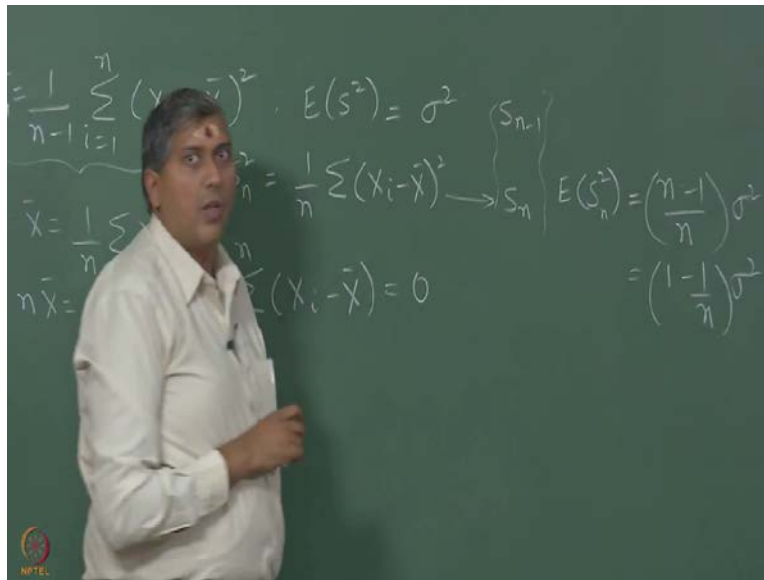
Now, let us move on to the estimation of variance. In all of this, we have seen wherever the variance was unknown I have to estimate it from the sample. Once again, what is the best estimator for variance? There are many different variance of estimating variance, I could look at sample variance as you see on the screen right now or I could use range for example, as an estimator of variance and so on. Again given many estimators which estimated do I pick is always the question that has been of interest for several decades in estimation theory and continues to be so. For some situations, there are results that are established, but in general, there are set of criteria as we discussed in the previous lecture, for choosing an estimator. One criteria is that we want an accurate or an unbiased estimator and two that we want an efficient estimator; an estimator which gives me as low error as possible, and thirdly consistent estimator which is that as the sample size increases the estimate converges to the truth. So, these are some of the 3 very important criteria in choosing an estimator.

And a sample variance satisfies all of this it is an accurate estimator which means it is unbiased the expectation of S square is nothing but sigma square of course, under some conditions. And then it is an efficient estimator, it is not the most efficient estimator, but it is fairly efficient; it would not be give me estimates with large errors. And thirdly, it is a consistent estimator as n grows large, the variance estimates will convert to the truth very good. So, with these points in mind, we will stick to the sample variance as an estimator of the variance and will work with those with this estimator only.

Now of course, the question of interest is what is a distribution of this sample variance; can we derive this theoretical? Yes, we can, if you recall one of the standard results that we discussed yesterday, there was one result where we said when I take random variables - standardized random variables that fall out of a Gaussian distribution and square them and add them up, then the resulting random variable follows a chi square distribution. If I am adding up n such variables then the resulting random variable has a chi square distribution with n degrees of freedom. How does that result apply here? Well, look at the expression for the sample variance. What am I doing there I am actually adding up squared variables of course, there are deviations; are these deviation variables normally distributed? Well, yes, if you assume that the distributions are following out of a Gaussian distribution. If they are not, what about it will discuss that shortly, but assume that the samples are following out of a Gaussian distribution.

Then we know \bar{X} also follows a Gaussian distribution, difference of two Gaussian distributed variables also follows a Gaussian distribution. Therefore, I can apply that result here and come to the conclusion that S^2 follows a chi square distribution with here with $n - 1$ degrees of freedom that is the only point that we have to watch out for why $n - 1$ degrees of freedom, why not n degrees of freedom. Because, the result that we had seen earlier says that the resulting random variable follows a chi square distribution the n degrees of freedom, but the catch here is that although we are adding n terms here, we do not really have n independent terms. And, let me actually explain that to you quickly.

(Refer Slide Time: 20:53)



Adding up n such terms and the point of contention is whether all these terms are independent. What this means says let us look at linear dependence part from linear algebra view point, we say that n variables are linearly independent; if you cannot find linear combination that yields 0, or in other words if you can express one as a linear combination of the other. Now, if you take X_i minus \bar{X} , we know that \bar{X} itself has been computed from X_i right, which means that there is a result what this means is there is a relation between X_i and \bar{X} . How does that come about from this we can show that these deviations variable are not completely independent; there is a linear dependence between them. How do we do that? We have $n \bar{X} = \sum X_i$, right and I have n such \bar{X} and straight away I can see that $\sum (X_i - \bar{X}) = 0$. So, which means there is a non trivial combination of X_i minus \bar{X} that gives me 0, which means truly this n deviation quantities are not linearly independent. Simply said if I know $n - 1$ deviation quantities, I can construct the n th one; it does not matter whether it is last one or it is a first one and so on.

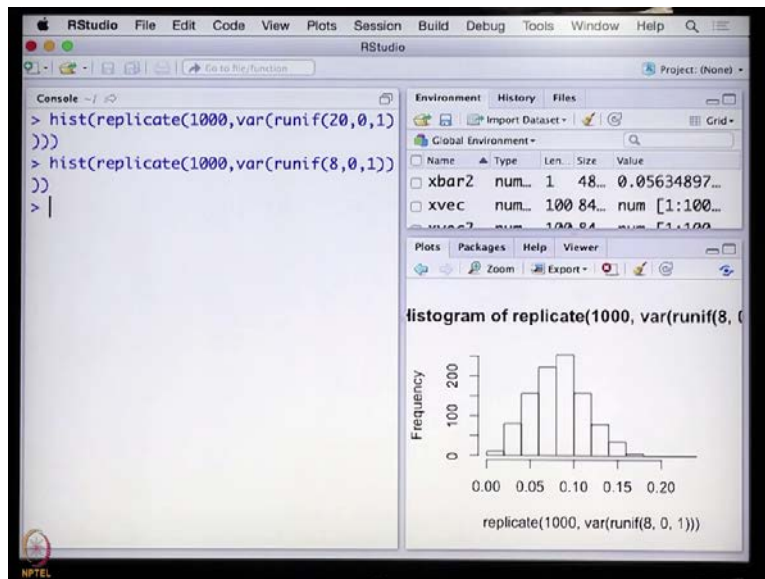
So, this is another way of looking at the loss of decrease of freedom and that also explains why a need an $n - 1$ here although I have n terms, so that I get a unbiased estimator. And on your calculators, normally we see S_{n-1} , and also an S_n , this S_{n-1} is nothing but the positive square root of S^2 based on $n - 1$ in the denominator. Very often, you will also see the literature a one over n in place of $n - 1$. So, we call as S^2 subscript n . The rest of the expression looks the same.

And the positive square root of this leads to S_n . Clearly if this is unbiased, this is the biased estimator, because $E[s^2] = \frac{n-1}{n} \sigma^2$ or you can say that this is $(1 - \frac{1}{n}) \sigma^2$. So, for finite n , this is a biased estimator and you can see the bias here. However when n becomes very large then $\frac{1}{n}$ goes to 0; as result, it becomes what is non as aesthetically unbiased estimator of variance. And therefore, for a large n whether you use s^2 or S^2 it should not make so much of a difference, but for small n let say 20, 30 and so on, it can make a bit of a difference.

In hypothesis testing, typically we work with this, because we do not know whether we have large samples or small samples. But if I know for sure that I have large samples typically the preference is for this. The reason is you can show theoretically that although this has a bias, this is more efficient S^2 although it is biased; it is more efficient than s^2 , which means it has a lower error than the s^2 . And for large n , typically does not matter whether you use this or not, so we prefer to use this. There are other reasons also which have beyond this scope of this course; the bottom line is we will use this, but there is also this competing estimator and on you are calculators and an a good calculator, you will see both of this, alright.

Now, we are convinced that there are $n-1$ degree of freedom only, although we have n terms in the expression for the sample variance. So under the Gaussian distribution assumption for the observation, we can now say that the sample variance follows as a chi square distribution or typically we work with standardized statistics. So, we say that the $(n-1) S^2 / \sigma^2$ follows a chi square distribution with $n-1$ degree of freedom. Now in a general case, when x size fall out of some arbitrator distribution non Gaussian distribution what can be said well nothing can be said theoretically about the distribution, may be for the large sample case we can say it follows a Gaussian distribution or something like that. But now you can resort to the Monte Carlo simulation approach and check what kind of distributions S^2 would have for a non Gaussian kind of distribution. Again, you can go back here and ask this question through simulations.

(Refer Slide Time: 26:35)



So, let say I do this, say replicate it is a 1000 times, what do I want to replicate? The distribution of the calculation of variance, so, we use the variance command here and I leave it to you to check in R whether this variance command uses S^2/n or $S^2/(n-1)$ in R. At the moment, we do not worry about it. Now what do I have to supply, I will let us give a non Gaussian distribution let say uniform distribution. And let us take about 20 samples, so that we are looking at a small sample case and standard uniform distribution all right. So, you see seems to follow Gaussian distribution. What about even fewer samples may be about 8, so easy slowly it is getting skewed; more or less, it is a chi square distribution, but perhaps with the different degrees of freedom. So, the distribution stills seems to be chi square.

Now chi square distribution has this feature that it is defined only over non-negative values. Clearly, because as you can see from the construction of the chi square variable it has some square, so some squares can only be a non zero, non-negative value, right. Therefore, always chi square distributed variables are defined over the intervals 0 to infinity. We can go back to the large sample case and you should expect a Gaussian distribution. So, we can take here 200 samples and see what happens you get a nice Gaussian distribution. Everything in his world seems to be standing to Gaussian right; at least when it comes to test statistics, so that is a nice thing at least we know in the large sample case everything tends to the Gaussian one fix distribution that I can always assume, good.

So, now let us get back to the discussion although nothing can be said theoretically about the distribution, we have to use simulations to do that. What we can definitely say is that this S square that we have defined here on the screen is always an unbiased estimator, of course, assuming that the x is a fall in out a random sample, they belong to a random sample that means, they are all independent of each other. When x size are correlated then it becomes a complicated thing to analyze, and we will not worry about that in this course.

(Refer Slide Time: 29:20)

Ratio of variances

Often it is required to compare the variances of two different populations, having the same variance σ^2 (e.g., two different manufacturing processes, control schemes, etc.).

Sampling distribution

When two random samples of sizes n_1 and n_2 respectively are drawn from two **Gaussian** populations having the **same variance**, the statistic

$$F = \frac{S_1^2}{S_2^2} \sim \mathcal{F}(\nu_1, \nu_2) \quad \nu_1 = n_1 - 1; \nu_2 = n_2 - 1$$

where S_1^2 and S_2^2 are the sample variances of the respective random samples.



The next situation of interest is ratio of variances and I think we talked about this again whether in the oxide layer thickness for the semi conducted manufacturing process. We wanted to see if one mixture of gases gave me a lower variance than another mixture of gases. So, typically, we do not compute differences here, we compute ratios of variances. And this ratio of variances follow us and f distribution again with the standard assumption in place; we assume that we have taken to random samples meaning collection of observations of sizes n_1 and n_2 respectively, and also assume that the samples are falling out of a Gaussian distribution. So, you see as compare to the mean the result available for sample variances and ratio of variances are quite restrictive, whereas the sample mean always seems to follow Gaussian distribution or T distribution more or less depending on other sample size is large or small. But, when it comes to the variance we have theoretical results available predominantly only for the Gaussian case. There are of course, may be a few research papers that you can look up depending

on you are needs. What kind of distribution this ratio of variances follows; when they fall out of that the samples fall out of a non Gaussian distribution.

We will in this course assume that is samples fall out of a Gaussian distribution. This F distribution is characterized by 2 degrees of freedom. Again, let me tell you what degrees of freedom means in statistics; degrees of freedom means in how many independent ways the variability of the random variable is affected. When it comes to the F variable, because it is a ratio of 2 random variables S_1^2 and S_2^2 , we say that the F distribution is characterized by 2 degrees of freedom ν_1 and ν_2 . And, ν_1 is $n_1 - 1$, ν_2 is $n_2 - 1$. The reason for $n_1 - 1$ minus $n_2 - 1$ is now obvious from the previous discussion; S_1^2 has $n_1 - 1$ degrees of freedom; S_2^2 has $n_2 - 1$ degrees of freedom. So, you can expect therefore, the F distribution to have $n_1 - 1$ comma $n_2 - 1$ degrees of a freedom.

The analytical expression for the F distribution is quite intimidating it can actually be really intimidating. However, we do not have to worry about it; normally, we do not give out the analytical expressions, either we look up a table in a book or the more modern way of doing that would be to just use the computer software package like R, where you turn to the `R F` to generate a randomly sample from an F distribution or `P F` to compute the probability of a random variable which as F distribution and so on. So, go ahead and actually generate a probability density function the probability density function of a random variable that has F distribution with some pre specified numerator and denominator degrees of freedom. Again, this is a result we will use this result in hypothesis testing.

(Refer Slide Time: 32:44)

Sample proportion

Point (sample) estimator for proportion (from a random sample of size n):

$$\hat{P} = X/n \quad (9)$$

- ▶ X is the number of observations belonging to the "success" class.
- ▶ Population is fairly large (possibly infinite) with p probability of success.

Then,

$$\hat{P} \sim \mathcal{N}(p, p(1-p)/n) \text{ for large } n \text{ and if } p \text{ is not too close to } 0 \text{ or } 1.$$

Note: This approximation is valid for $np > 5$ and $n(1-p) > 5$.



Now, the final thing to be discussed in this lecture is that of the sample proportion. Again recall the example that we discussed in the motivation lecture, there is a manufacture of a controllers for automobile application, and the manufacture claims that the proportion of defective items is not more than a particular value and we want to test that hypothesis. So, what do we do here, of course, now the random variable that we are looking at is a binomial distributed random variable, because now we are looking at proportions were of the success or failure? And once again, you should go back to the definition of a binomial random variable, and recap all the conditions under which a random variable follows a binomial distribution. And one of the parameters, the only parameter that characterized as a binomial distribution is the proportion or the probability of success, which is denoted by p .

So, assume that now I actually take a sample of size n from a population which follows a binomial distribution the probability success being P , X is a number of observation belonging to the success class than P would be X by n . So, what I do is, I collect the sample of size n and count the number of success. So, in the case of defective items, I would randomly sample from a lot and through an experiment or through inspection and determine what is non-defective and call that as X . Then the proportion and estimate of the proportion of success probability success is given by X by n naturally, where X is the number of success cases by n which is the total number of cases that you have in your sample. We call that has \hat{P} as an estimate of the proportion sorry probability of success. The small p

is true probability of success.

We will also make an assumption that population is fairly large that is important. And sample size is also going to satisfy certain conditions given at the bottom of the slide. Now the result of interest was is the on the sampling distribution of \hat{P} , and the results says that \hat{P} also follows a Gaussian distribution which is very nice; provided certain conditions are made. And one of the key conditions sees that the true probability of success is not too close to 0 or 1; that means, here population should not have too many success cases or too few success cases. There should be some reasonable balance between success and failure; in that case, the \hat{P} follows a Gaussian distribution. In fact, I would say that go to your R package, and check if this result holds good by means of simulation again using the same idea repeat experiments using replicate and plot a histogram.

What about the mean and variance of the Gaussian distribution the mean is np , if you recall from the discussion on the statistics that we had, we said that the expected value of a binomially distributed random variable is np . And variance is given by $np(1-p)$ that is for that case. But here as you can see X follows a binomial distribution, but \hat{p} had therefore, also follows a binomial distribution, but it is X/n and therefore, you would have here normal distribution with variance $p(1-p)/n$.

So, now we move on to the last item in this lecture, where we discuss the case of a sample proportion. And again, you should recall the corresponding example that we used in the motivating lecture, where we were looking at the number of defective controllers in controllers applied by manufacturer for automobile applications. And in general, of course, they supplies to all those cases where we are looking at the defective, non-defective or success, failure, head and tail in a toss - coin toss and so on. So, we the setting as is as follows, we have a population characterized by binomial distribution and probability of success being p ; and from which, we randomly sample observation of size n . The natural estimator of the proportion of success - the number of success in a sample of size n is denoted by X in the slide. And an estimate of the probability of success is naturally X/n that is very much easy to understand.

If I want to know what is a probability of getting a head in a toss of a coin. What do I do, I repeatedly conduct those trials the coin toss trials and count the number of heads in those trials let say I conduct

about hundred trials and I count and I get about let us say 55 heads then the probability of getting a head in any coin toss in general would be 55 by 100 of course, that would give you 0.55. And, as you conduct more and more trials that is as a number of observation increase or the sample size increase, we should expect to the estimate to converts to the truth. So, here an unbiased estimator of the probability of success is \hat{P} given by X by n . Why is it unbiased? It is very easy to see expectation of \hat{P} is expectation X by n and we know from theory that if X has is a binomially distributed random variable, its averages $n p$, therefore, expectation of \hat{P} is p , alright.

Now with that estimator question is, what is the sampling distribution of \hat{P} under the assumption that the population is fairly large, and that the probability of success p remains constant and that a couple of more assumptions \hat{P} approximately follows a Gaussian distribution with expected value p showing that it is an unbiased estimator and the variance being p times 1 minus p by n . How do you get this? Well, expectation of \hat{P} is expectation of X by n , so that any ways establish that the mean of \hat{P} is p which is the probability of success. Variance of \hat{P} is variance of X by n square; right.

And, we know once again from the lecture go back to the lecture on statistics that variance of a binomially distributed random variable is $n p$ times 1 minus p . Therefore, variance of \hat{P} would be $n p$ times 1 minus p by n square, where you get p , therefore, the variance of \hat{P} to be p times 1 minus p by n . Now, this result is valid for large n ; that means, in the coin toss example, you would have conducted many trials or in the defective controller example, you would have collected enough items not 5 or 10 items in your random sample lot and also that the probability of success is neither to low not to large. That means, the in the original population, you should not have 2 few defective items in terms of proportion or too large as number of defective items or non-defective items as well. And, there is a typical guideline that is given here this that is approximation is valid when $n p$ is greater than 5 and the other condition that is given as well.

(Refer Slide Time: 41:26)

Sample proportion

... contd.

If n is large, the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \quad (10)$$

is **approximately** standard normal.



So, we assume that more or less this condition hold good at least in this course for the general case I urge you to refer to the literature. Now as usual we work with standard or standardized statistics, since \hat{P} follows a Gaussian distribution, the variable Z , which is \hat{P} minus p by σ , should also follow a standard normal distribution. And that is what your equation tells you here that \hat{P} minus p by square root of the variance of \hat{P} follows an approximately standard normal distribution, because \hat{P} itself follows an approximate Gaussian distribution. We use this in the hypothesis testing of the proportions.

Now in the next lecture of course, now will focus on hypothesis testing, but before I conclude in this lecture, I would like to reiterate that it is not just important to know these results, but what is equally important is to know the assumptions under which we have derived this sampling distribution, because when you have a case of hypothesis testing, you should first look at the assumptions and see whether those assumptions actually match with the assumptions that we are made here. If not, then you may have to turn to a literature or some other text book to see what test statistic is appropriate and what is the sampling distribution of the test statistic under the assumptions in which you have going to conduct the hypothesis test.

In this course, on the remainder of this course, we will assume that whatever assumptions we have made

here are the assumptions under that will apply to the hypothesis test that examples that we are going to look at.

(Refer Slide Time: 43:30)

Statistics for Hypothesis Testing - Part 2 References

Differences in proportions

Suppose two independent samples of (large) sizes n_1 and n_2 are taken from two populations, and the Binomial RVs X_1, X_2 represent the number of observations of a particular class (success).

If $\hat{P}_1 = X_1/n_1$ and $\hat{P}_2 = X_2/n_2$ are the estimators of respective proportions p_1 and p_2 , the statistic

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \quad (11)$$

has an approximate standard normal distribution.



So, the final thing that is of interest towards differences in proportions. And again, the story is the same very much similar to the difference in the samples mean case. We have two different populations; remember the soft drink example, we have 2 campuses and we wanted to test, if the proportion of students on each campus is same when it comes to preference for a soft drink. So, here we have two populations; from which, I have randomly drawn samples of sizes n_1 and n_2 , and as usual all the other assumptions holding good. I estimate the proportion of has \hat{P}_1 and \hat{P}_2 and the difference that is $\hat{P}_1 - \hat{P}_2$ approximately follows a normal distribution. As a result of which, Z which is a standardized statistics corresponding to difference in proportions also follows approximately standard Gaussian distribution.



Now, Z is not truly a statistic, because p is unknown in the denominator, we do not know p . So, what we do is as for as a denominator is concerned, any way $p_1 - p_2$ is something that we will postulate in a hypothesis test. So, in the numerator it is not of a concern; the denominator it is a concern why have the p_1 and p_2 appeared in the denominator because if you can see $p_1(1-p_1)$ over n_1 is nothing but the sigma square, the variance or the you can say variance of \hat{P}_1 and

likewise, the other term being variance of \hat{P}_2 . They have appeared there because we are standardizing the $\hat{P}_1 - \hat{P}_2$. The point it is now I need to know the denominator to be able to calculate the observed statistic $\hat{p}_1 - \hat{p}_2$ is specified by the user in a hypothesis test, but the denominator has to be supplied as well. If the denominator is not supplied, typically that is the case then what you do is you use the same approach that you use in the sample mean case. When we did not know the variance, we used the estimates of variance.

Here the variance is in terms of a proportions, so we use the respective estimates of proportions in this terms to arrive at the estimates of the respective variances so that is the approach that is used here. At this point, do not get confused that if I know p_1 and p_2 ; that means, if I am supplying p_1 and p_2 in the denominator then where the question of hypothesis is testing. We are not supplying the true p_1 and true p_2 ; we are saying I do not know the true p_1 and p_2 . One option is to say, I will not solve this problem and go home, and sleep happily. The other option is to be practical and say well let me get an estimate and one of the easy ways of doing that is to use the estimates of the proportion from the respective samples and plug that in to the expression for Z , calculate the observed statistics. \hat{P}_1 and \hat{P}_2 have calculated from the respective samples; $\hat{p}_1 - \hat{p}_2$ is specified by the user in the hypothesis test, and the denominator is again calculated based on your estimates that is \hat{P}_1 and \hat{P}_2 that is the story.

(Refer Slide Time: 47:11)

Bibliography I

-  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.

Again, you should remember that, this result is only valid under certain assumptions all right and that brings us to the end of this lecture. In fact, the end of the topic of sampling distributions as far as the univariate case is concerned, for the bivariate case, when we look at correlation we will talk about the sampling distribution of correlation in that respective lecture that is the respective lecture on hypothesis testing of correlations. The sampling distributions are useful in many different ways, but the 2 uses for us in this course is in hypothesis testing and two is in construction of confidence intervals that we will demonstrate later on.

So, the next lecture that is in order is; obviously, hypothesis testing the crux of this course; well, where we will introduced some key terminology type I error, type II error, p value, significance level, power of a hypothesis test and so on. And then, take the sampling distributions take the 8 examples that we took up in the motivation lecture, quickly go through all those examples. And we now have all the paraphernalia that is required for hypothesis testing, now it is just a matter of implementing the standard procedure, therefore, when it comes to using this informational example it is going to be fairly very easy thing and a breeze thing, alright.

So, see you in the next lecture. Hopefully, you are enjoying all these lectures.

Bye for now.