**SDA INNOVATIVE EXAM REPORT**

**Application of central limit theorem in cholesterol**

Group Member :

1. Rishabh Tiwari – 57
2. Yoshit Verma – 58
3. Vedanti Wadekar – 59
4. Nidhi Worah – 60

**Title of Project :**

Application of central limit theorem in cholesterol

**Theoritical Background :**

This report will begin with a brief introduction, accompanied through the evaluation, and give up with recommendations for in addition analysing. The evaluation will introduce the central limit theorem in cholestrol, explain exclusive kinds of central limit theorem and give examples of its applications in cholesterol.

Cholesterol molecules are transported in blood by large macromolecular assemblies (illustrated below) called lipoproteins that are really a conglomerate of molecules including apolipoproteins, phospholipids, cholesterol, and cholesterol esters. This macromolecular carrier particles make it possible to transport lipid molecules in blood, which is essentially an aqueous system.

Different classes of these lipid transport carriers can be separated (fractionated)based on their density and where they layer out when spun in a centrifuge. High density lipoprotein cholesterol (HDL) is sometimes referred to as the "good cholesterol," because higher concentrations of HDL in blood are associated with a lower risk of coronary heart disease. In contrast, high concentrations of low density

lipoprotein cholesterol (LDL) are associated with an increased risk of coronary heart disease. The illustration on the right outlines how total cholesterol levels are classified in terms of risk, and how the levels of LDL and HDL fractions provide additional information regarding risk.

**Literature Survey :**

An unwritten assumption of much of the literature on the t-test is that all two-sample tests are effectively testing the same null hypothesis, so that it is meaningful to compare the Type I and Type II error rates of different tests. This assumption is frequently untrue, and testing for a difference in means between two samples may have different implications than testing for a difference in medians or in the proportion above a threshold. We defer until later a discussion of these other important criteria for selecting an estimator or test. Most of the literature on the assumptions of the t-test is concerned with the behavior of the t-test in relatively small samples, where it is not clear if the Central Limit Theorem applies. For linear regression, the statistical literature largely recognizes that heteroscedasticity may affect the validity of the method and non-Normality does not. The literature has thus largely been concerned with how to model heteroscedasticity and with methods that may be more powerful than linear regression for non-Normal data. These issues are outside the scope of our review

**Example :**

The blood cholesterol levels of a population of workers have mean 202 and standard deviation 14.

(a)

If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

(b)

Repeat (a) for a sample size of 64.

Solution

(a)

It follows from the [central limit theorem](https://www.sciencedirect.com/topics/mathematics/central-limit-theorem) thathttps://ars.els-cdn.com/content/image/3-s2.0-B9780123743886000077-fx_x-9780080922102.jpgis approximately normal with mean μ = 202 and standard deviation σ/n=14/36=7/3. Thus the standardized variable

W =X¯−2027/3

has an approximately [standard normal distribution](https://www.sciencedirect.com/topics/mathematics/standard-normal-distribution). To compute P{198≤ X¯ ≤206}, first we must write the inequality in terms of the standardized variable *W*. This results in the equality

P{198≤X¯≤206}=P {198−2027/3≤X¯−2027/3≤206−2027/3}=P {−1.714≤W ≤1.714}≈P {−1.714≤Z≤1.714}=2P {Z≤1.714}−1=0.913

where *Z* is a standard normal random variable and the final equality follows from Table D.1 in App. D (or from Program 6-1).

(b)

For a sample size of 64, the sample meanhttps://ars.els-cdn.com/content/image/3-s2.0-B9780123743886000077-fx_x-9780080922102.jpgwill have mean 202 and standard deviation 14/64=7/4. Hence, writing the desired probability in terms of the standardized variable

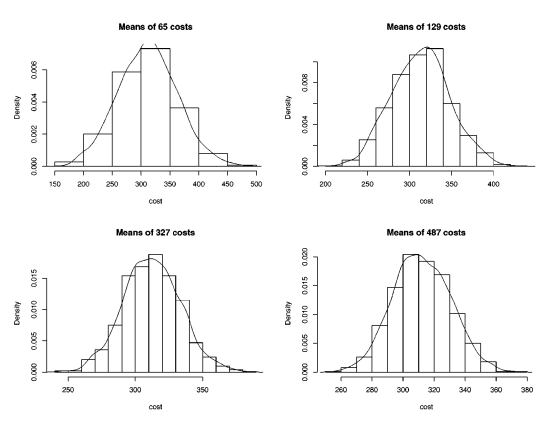
X¯−202

yields

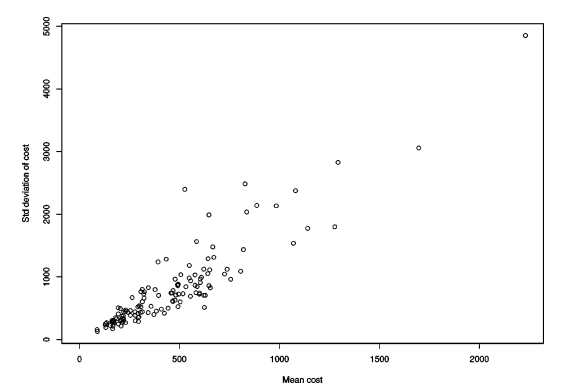
P{198≤X¯≤206}=P{198−2027/4≤X¯−2027/4≤206−2027/4}≈P{−2.286≤Z≤2.286}=2P {Z≤2.286}−1=0.978

**Example for Linear Regression**

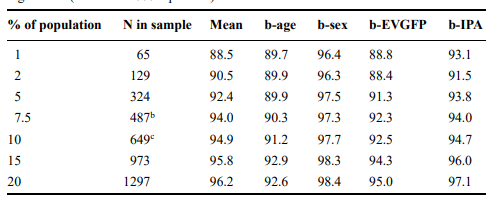
Medical costs usually have the very non-Normal distribution we see here, but transformations are undesirable as our interest is in total (or mean) dollar costs rather than, say, log dollars (11). We considered the 6918 subjects to be the population of interest and drew samples of various sizes to determine whether the test statistics of interest had the distribution that was expected. In addition, there is substantial heteroscedasticity and a somewhat linear relation between the mean and variance. In Figure 3 we divided subjects into groups by age and sex and calculated the mean and standard deviation of cost for each group. It is clear that the standard deviation increases strongly as the mean increases. The data are as far from being Normal and homoscedastic as can be found in any real examples. We used these data to determine how large a sample would be needed for the Central Limit Theorem to provide reliable results. For example, as illustrated on the first line of Table 1, we drew 1000 1% samples, of average size 65, from the population. For each sample we calculated the regression of cost on age, sex, self-rated health, and HMO (IPA = 0) versus Fee for Service (IPA = 1). For each parameter in the regression model we calculated a 95% confidence interval and then checked



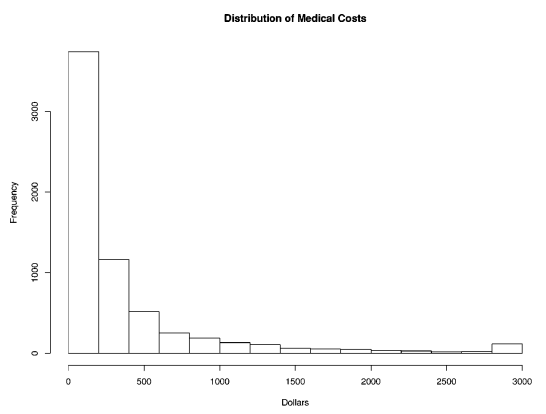
to see whether the confidence interval contained the true value. The percent of times that the confidence interval included the value computed from the entire population of 6918 is an estimate of the true amount of confidence (coverage) and would be 95% if the data had been Normal to start with. For samples of size 65 and 129, some of the confidence interval coverages are below 90%. That means that the true alpha level would be 10% or more, when the investigator believed it to be 5%, yielding too many significant regression coefficients. Note that for sample sizes of about 500 or more, the coverage for all regression coefficients is quite close to 95%. Thus, even with these very extreme data, least-squares regression performed well with 500 or more observations. These results suggest that cost data can be analyzed using least-squares approaches with samples of 500 or more. Fortunately, such large samples are usually the case in cost studies. With smaller samples, results for variables that are highly significant (p < .001, for example) are probably reliable. Regression coefficients with p-values between .001 and .10, say, might require additional analysis if they are important. For data without such long tails much smaller sample sizes suffice, as the examples in the literature review indicate. For example, at one time a popular method of generating Normally distributed data on a computer was to use the sum



**Methodology :** The literature summarized above and our simulations illustrate that linear regression and the t-test can perform well with data that are far from Normal, at least in the large samples usual in public health research. In this section we examine alternatives to linear regression. In some disciplines these methods are needed to handle small samples of non-Normal data, but in reviewing their appropriateness for public health research we focus on other criteria. These methods usually come with their own sets of assumptions and they are “alternatives” to least-squares methods only when no specific summary statistic of interest can be identified, as we discuss in the next section. We examine the Wilcoxon rank-sum test as an alternative to the t-test and the logistic and proportional odds models as alternatives to linear regression.

****

**Example :** The Central Limit Theorem depends on the sample size being “large enough,” but provides little guidance on how large a sample might be necessary. We explored this question using the cost variable in the Washington Basic Health Plan data. Annualized outpatient cost has a very long right tail, as shown in Figure 1. We truncated the histogram at $3000 so that the distribution for lower values could be seen, but use the full distribution in the following analysis. The actual costs ranged from $0 to $22, 452, with a mean of $389. The standard deviation is $895, standardized skewness is 8.8, and standardized kurtosis is 131. Figure 2 shows the sampling distribution of 1000 means of random samples of size 65, 129, 324, and 487 from this very non-Normal distribution (approximately 1%, 2%, 5%, and 7.5% of the population). The graph shows a histogram and a smooth estimate of the distribution for each sample size. It is clear that the means are close to Normally distributed even with these very extreme data and with sample sizes as low as 65.



**Future Scope :** Conversely, the sampling distributions of the mean for smaller sample sizes are much broader. For small sample sizes, it’s not unusual for sample means to be further away from the actual population mean. You obtain less precise estimates.

In closing, understanding the central limit theorem is crucial when it comes to trusting the validity of your results and assessing the precision of your estimates. Use large sample sizes to satisfy the normality assumption even when your data are nonnormally distributed and to obtain more precise estimates!

**Conclusion :** The t-test and least-squares linear regression do not require any assumption of Normal distribution in sufficiently large samples. Previous simulations studies show that “sufficiently large” is often under 100, and even for our extremely non Normal medical cost data it is less than 500. This means that in public health research, where samples are often substantially larger than this, the t-test and the linear model are useful default tools for analyzing differences and trends in many types of data, not just those with Normal distributions. Formal statistical tests for Normality are especially undesirable as they will have low power in the small samples where the distribution matters and high power only in large samples where the distribution is unimportant.

**References :**

<https://myweb.uiowa.edu/pbreheny/4120/s20/notes/3-10.pdf>

<https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability13.ht> ml