Activity Sheet:

1. **In the morning session, you were introduced to the concepts of Hypothesis testing- Importance of confidence intervals, Formulating the hypothesis, using appropriate test to test hypotheses, types of errors etc.**
2. **In today's activity, we shall be working on solving the problems related to Hypothesis testing that involves working with concepts learnt in the morning session.**
3. **This activity has two parts:**
   a. **In class assignment: Solve these problems manually (you may use calculators/R to perform computations required to solve the problem)**
   b. **Homework assignment: There are functions available in R to solve these problems. Explore and solve these problems using R functions**

Problems:

1. A random sample of 100 items is taken, producing a sample mean of 49. The population SD is 4.49. Construct a 90% confidence interval to estimate the population mean.

   n = 100, sample $\overline{x}$ = 49, population standard deviation σ = 4.49;

   90% confidence interval, z = -1.64 to +1.64; R: qnorm (0.05) and qnorm (0.95)

   C.I. = $\overline{x} \pm z*\sigma/\sqrt{n}$
   C.I. = 49 - 1.64 x 4.49/√100; 49 + 1.64 x 4.49/√100
   C.I = (48.2, 49.7)

   Interpretation: If we repeatedly draw several samples with sample size n=100, there is 0.95 probability (95%) that the above C.I. will contain the true value of the population mean.

   Note: The sample mean can be approximated to the population mean as per the CLT. However, it is only an approximation and not exact. Therefore we provide an interval

2. A random sample of 35 items is taken, producing a sample mean of 2.364 with a sample variance of 0.81. Assume x is normally distributed and construct a 90% confidence interval for the population mean.

   n =35, sample mean = 2.364, sample variance = 0.81

   90% confidence interval, z = -1.64 to +1.64;

R: qnorm (0.05) and qnorm (0.95)

C.I. = 2.364 $\pm$ (1.64) x (0.9) / $\sqrt{35}$

3. State the null and alternative hypotheses to be used in testing the following claims and determine generally where the critical region is located: (a) The mean snowfall at Lake George during the month of February is 21.8 centimeters. (b) No more than 20% of the faculty at the local university contributed to the annual giving fund.

   (a) H_0: $\mu = 21.8$, H_1: μ # 21.8; critical region in both tails.
   (b) H_0: $p \leq 0.2$, H_1: $p > 0.2$; critical region in right tail.

4. Suppose a car manufacturer claims a model gets 25 mpg. A consumer group asks 40 owners of this model to calculate their mpg and the mean value was 22 with a standard deviation of 1.5. Is the manufacturer's claim supported?

H_0: $\mu \geq 25$
H_a: $\mu < 25$
$\overline{x} = 22$
$\mu = 25$
$s = 1.5$
$n = 40$
$z = (\overline{x} - \mu) / (s/\sqrt{n})$

$z = (22 – 25) / (1.5 / \sqrt{40}) = -3 / 0.23 =$ **-12.64**, P (z = -12.43) = $\mathbf{6.35 \times 10^{-37}}$**.** This is the probability value of the test statistic. In R: pnorm (-12.64)

Significance level (α) or the critical value = **0.05**; the z-score corresponding to this critical value of 0.05 is **-1.64.**

We reject the null hypothesis because the test statistic falls way beyond the critical region. The decision can be made based on either the Z-scores or probabilities of the test statistic and critical region; i.e. the manufacturer's claim is suspicious.

5. The CEO of a large electric utility claims that 80 percent of his 1,000,000 customers are very satisfied with the service they receive. To test this claim, the local newspaper surveyed 100 customers, using simple random sampling. Among the sampled customers, 73 percent say they are very satisfied. Based on these findings, can we reject the CEO's hypothesis that 80% of the customers are very satisfied? Use a 0.05 level of significance.

H_0: $p \geq 80\% = 0.8$ (i.e. 80% of the customers are satisfied with the services)
H_a: $p < 0.80$ (less than 80% of the customers are satisfied with the services)

Since the alternate hypothesis has "<" symbol we use lower tail

p =0.8

p̂ = 0.73

H_0: $p_o \geq$ 80% =0.8 (i.e. 80% of the customers are satisfied with the services)

H_a: p_o < 0.80 (less than 80% of the customers are satisfied with the services)

Since the alternate hypothesis has "<" symbol we use lower tail

p_o = 0.8

p̂ = 0.73

The formula for proportions with the continuity correction factor is as follows (please refer slide 93 of the presentation). Note that in proportions we divide the number by n and therefore the 0.5 correction factor should also be divided by 100

$$Z = \frac{\hat{p} + \frac{0.5}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(0.73 + \frac{0.5}{100}) - 0.80}{\sqrt{\frac{0.8*0.2}{100}}} = \frac{(0.73 + \frac{0.5}{100}) - 0.80}{\sqrt{\frac{0.8*0.2}{100}}} = -1.625$$

P (Z = -1.625) = 0.052

Since 0.052 > 0.05, we cannot reject null hypothesis.

Alternate method 1 (without using proportion)

Z = $\frac{73.5-80}{\sqrt{100*0.8*0.2}} = \frac{73.5-80}{\sqrt{100*0.8*0.2}} = \frac{-6.5}{4}$ = -**1.625** (Note: as you can see, we get the same z-score)

For large n the effects of the continuity correction factor is very small and will be omitted.

Alternate method 2 (directly using formula in R)

Normal: pnorm (q = 73.5, mean =80, sd = 4) = 0.052

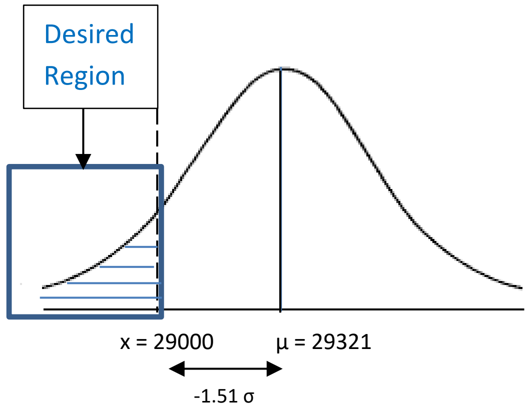
Binomial: pbinom (q = 73, size = 100, prob = 0.8) = 0.055

6. A population of 29 year-old males has a mean salary of $29,321 with a standard deviation of $2,120. If a sample of 100 men is taken, what is the probability their mean salaries will be less than $29,000?

Step 1: Insert the values into the z-formula: Z = (x - μ)/(σ/√n)

Z = (29,000-29,321)/ (2,120/√100) = -321/212 = -1.51

Step 2: Calculate Z-score, Z = -1.51 has an area of 6.55%. Hence, the probability of mean salaries less than $29,000 is 6.55%.



7. A large freight elevator can transport a maximum of 9,800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean μ = 205 pounds and standard deviation σ = 15 pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

We are given x = 9800, n = 49, μ = 205, σ = 15. The elevator can transport up to 9800 pounds. Therefore these 49 boxes will be safely transported if they weigh in total less than 9800 pounds.

First we find the Z score for 49 boxes, each of which weigh 205 pounds.
$Z = [9800-49(205)]/ [(15*49)/\sqrt{49})] = 245/105 = -2.33$

The probability that the total weight of these 49 boxes is less than 9800 pounds is
$P(Z < 9800) = P(Z < -2.33) = 0.0099$ or 0.9%; in R: pnorm(-2.33)

8. A student, to test his luck, went to an examination unprepared. It was a MCQ type examination with two choices for each questions. There are 50 questions of which at least 20 are to be answered correctly to pass the test. What is the probability that he clears the exam? If each question has 4 choices instead of two, What is the probability that he clears the exam?

Each event has a success or a failure - Binomial distribution. But it can be cumbersome to solve for all the values (from 20 right answers to 50 right answers). When n is larger, binomial can be approximated to normal distribution.

$\mu = np$; $\sigma^2 = npq$, where n is number of trials, p is success probability and q is failure probability.
Therefore, $\mu = 50*0.5 = 25$; $\sigma^2 = 50*0.5*0.5 = 12.5$

$Z = (19.5-25)/\sqrt{(12.5)} = -1.55$

$P(Z > -1.55) = 1 - 0.0606 = 0.9394$. There is about 93.94% probability that he passes the exam.

R: 1-pnorm(-1.55)

9. A marketing director of a large department store wants to estimate the average number of customers who enter the store every five minutes. She randomly selects five-minute intervals and counts the number of arrivals at the store. She obtains the figures 68, 42, 51, 57, 56, 80, 45, 39, 36 and 79. The analyst assumes the number of arrivals is normally distributed. Using this data, the analyst computes a 95% confidence interval to estimate the mean value for all five-minute intervals. What interval value does she get?

$\overline{x}$ = 55.3, s = 15.9, n =10
Confidence interval = $\overline{x} \pm t_{\alpha/2,\, n-1} * \frac{s}{\sqrt{n}}$
t-value @ 95% confidence, and 9 dof = t (0.05/2,9) = 2.262
Confidence interval = 55.3 – 11.45, 55.3 + 11.45

10. Write the Null, and alternate Hypotheses and identify type I and type II errors in the following scenarios:
    a) An innocent person is sent to jail

       Null hypothesis: The person is innocent
       Alternate Hypothesis: The person is guilty
       Actual result: The person is actually innocent, but is sent to jail
       Error: The null hypothesis is actually true. But, they accepted alternate hypothesis (considered him guilty, and sent him to jail).
       i.e., **they (falsely) rejected the true null hypothesis. Hence it's Type I error**

    b) A manager sees some evidence that stealing is occurring but lacks enough confidence to conclude the theft, and he decides not to fire the employee

       Null Hypothesis: The employee is not stealing
       Alternate Hypothesis: The employee is stealing (and hence he should be fired)
       Actual Result: The employee is actually stealing, but manager does not have enough confidence to conclude the theft, and hence he didn't fire the employee

Error: The null hypothesis that he is not stealing is falsely accepted due to lack of confidence.
i.e., **they failed to reject the false null hypothesis. Hence it's a type II error**

11. A population of heights has a mean of 68 inches. What is the probability of selecting a sample of size 25 that has a mean of 70 or greater standard deviation of 4?
    a. State the Hypothesis. Is it a one tailed test or two tailed test.
    b. What is observed and critical t value?
    c. Do we have enough evidence to reject null hypothesis?

Hypothesis H_0: $\mu = 68$
H_1: $\mu >= 68$
Since we do not know the standard deviation of the population, we go for t-test.
Estimated standard error SE= 4/sqrt(25) = 4/5= 0.8
t= (x- μ)/SE = 70-68/0.8= 2/0.8= 2.5.
$t_{crit} = t_{\alpha,df} = t_{0.05,24}$= 1.711. Since t observed is greater than t critical we reject the null hypothesis that mean height=68 inches

12. A survey was conducted to examine the differences between younger and older adults in perceived happiness. For the questions, the scores range from 0 to 60 and higher score indicates higher perceived happiness. Ten younger and older adults were given the test and the scores are as given below.
    a. Formulate the hypothesis and perform appropriate test for your hypothesis
    b. What is the observed value of the test
    c. What is the critical value of the test
    d. Is there a significant evidence to reject null hypothesis

Older Adults	Younger Adults
45	34
38	22
52	15
48	27
25	37
39	41
51	24
46	19
55	26
46	36

Older Adults: mean=44.5, standard deviation= 8.682678, Variance=75.3889
Younger Adults: mean= 28.1, standard deviation= 8.543353492, Variance= 72.9889

1. Null Hypothesis: H_0: The perception of happiness is same for both younger and older adults
   Alternate Hypothesis: H_1: The perception of happiness is different for younger and older adults.
2. To validate/test out hypothesis we preform a two tailed t-test. We know that

   $s_p{}^2 = \frac{(n_1-1)s_1{}^2+(n_2-1)s_2{}^2}{(n_1-1)+(n_2-1)}$; $t = \frac{(\overline{x_1}-\overline{x_2})}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ with $(n_1 + n_2 - 2)$ df.

   $S_p{}^2$ = ((9*75.3889) + (9*72.9889))/(9+9)= 74.1889
   S_p= 8.6133

3. The observed t-value (test to identify if these two samples are derived from same population
   T_{obs}=(44.5-28.1)/8.6133*sqrt(0.2)= 4.257
4. t-critical is given by 2.101
5. p value corresponding to the t-value is 2*(1-pt(4.257,18))= 0.0004739. The means are significantly different and we can reject null hypothesis that "perception of happiness is same for both younger and older adults"

Population Parameter	Population Distribution	Conditions	Confidence Interval
μ	Normal	You know σ^2 n is large or small $\bar{X}$ is the sample mean	$(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}})$
μ	Non-normal	You know σ^2 n is large (> 30) $\bar{X}$ is the sample mean	$(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}})$
μ	Normal or Non-normal	You don't know σ^2 n is large (> 30) $\bar{X}$ is the sample mean s^2 is the sample variance	$(\bar{X} - z\frac{s}{\sqrt{n}}, \bar{X} + z\frac{s}{\sqrt{n}})$
p	Binomial	n is large p_s is the sample proportion q_s is $1 - p_s$	$(p_s - z\sqrt{\frac{p_s q_s}{n}}, p_s + z\sqrt{\frac{p_s q_s}{n}})$

	Sigma known	Sigma unknown
n≥30	$\bar{x} \pm z*\sigma/\sqrt{n}$	$\bar{x} \pm z*s/\sqrt{n}$
n<30	$\bar{x} \pm z*\sigma/\sqrt{n}$	$\bar{x} \pm t*s/\sqrt{n}$

Level of Confidence	Value of z
90%	1.64
95%	1.96
99%	2.58

.