# R Assignment:02

Keya Adhyaru

## Question 1:

David recently purchased a Valve Steam Deck and wants to know if the average purchase price of the games he is playing has changed since he bought it. He randomly samples game prices from before the purchase and again randomly samples game prices from after the purchase.

Note: *No* in the After_Steam_Deck column is one category and *Yes* is the other category.

a) Perform a parametric test to answer this question at the $\alpha$= 0.08 significance level. Assume all conditions are satisfied and do not check assumptions. State the test that you are using.

**Answer:**

Let $\mu_{diff}$ be the average differences between of prices of game before steam deck purchase and after steam deck purchase, $\mu_{diff} = \mu_{before}$ - $\mu_{after}$.

$H_0 : \mu_{diff} = 0$

$H_A : \mu_{diff} \neq 0$

Assuming that all conditions are satisfied, we further perform independent t-test for two samples, by following the parametric approach.

```
#differentiating targeted columns from the given data
After_price_vector <- c(16.09,6.43,18.07,12.98,13,3.38,3.83)
Before_price_vector <- c(15.45,4.79,13.55,6.2,19.65,3.04,6.2,6.77,25.29)

#calculating mean
xbar_after_price = mean(After_price_vector )
xbar_before_price = mean(Before_price_vector)

#checking for equal variance
sd_before <- sd(Before_price_vector)
sd_after <- sd(After_price_vector)
sd_ratio <- sd_before/sd_after
sd_ratio
```

```
## [1] 1.288488
```

```
#sd_ratio is less than 2, so we consider it as same variances
#performing independent two sample t-test

t.result <- t.test(Before_price_vector, After_price_vector,
                   alternative = "two.sided",conf.level = 0.92, mu = 0, var.equal = TRUE)
t.result
```

```
##
##  Two Sample t-test
##
## data:  Before_price_vector and After_price_vector
## t = 0.19185, df = 14, p-value = 0.8506
## alternative hypothesis: true difference in means is not equal to 0
## 92 percent confidence interval:
##  -5.970773  7.321884
## sample estimates:
## mean of x mean of y
##  11.21556  10.54000
```

From the test, we got our p-value= 0.8506133 $> \alpha$. So, we fail to reject null hypothesis. That means, from our analysis we find no difference between the average prices of the game console which were measured before david bought Valve Steam Deck and after his purchase.

b) Construct the 92% parametric confidence interval.

**Answer:**

Constructing 92% confidence interval

```
interval <- t.result$conf.int
```

We are 92% confident that the true $\mu_{diff}$ lies is in the interval (-5.970773, 7.321884).

c) Perform a non-parametric test to answer this question using $\alpha = 0.08$.

**Answer:**

For non-parametric approach, we will perform permutation test.

Let $\mu_{diff}$ be the average differences between of prices of game before steam deck purchase and after steam deck purchase, $\mu_{diff} = \mu_{before}$ - $\mu_{after}$.

$H_0 : \mu_{diff} = 0$

$H_A : \mu_{diff} \neq 0$

```
library(wPerm)
perm_result <- perm.ind.loc(x= Before_price_vector, y= After_price_vector, parameter= mean,
             stacked= TRUE, alternative = "two.sided", R = 10000)
```

From the permutation test, we get the p-value= 0.996 $> \alpha$. So we fail to reject null hypothesis.

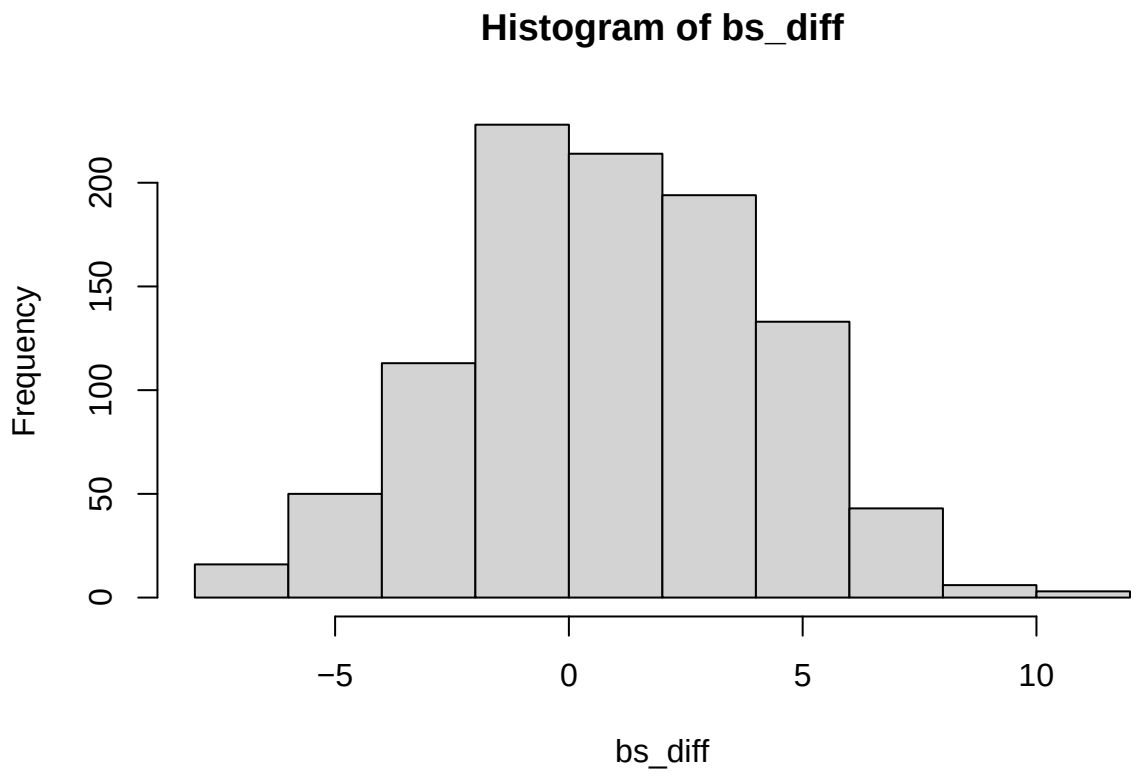
By concluding the test, we find no difference between the average prices of the game console which were measured before david bought Valve Steam Deck and after his purchase.

d) Construct the 92% confidence interval estimate using a percentile interval (non-parametric approach).

**Answer:**

Calculating 92% Confidence interval using percentile interval:

```
library(boot)

mean_wrapper <- function(x, index){
     mean(x[index])
}

bs_before <- boot(Before_price_vector, statistic = mean_wrapper, R = 1000)
bs_after <- boot(After_price_vector, statistic = mean_wrapper, R = 1000)

bs_diff <- bs_before$t - bs_after$t

hist(bs_diff)
```




```
perc_interval <- quantile(bs_diff, probs = c(0.04, 0.96))
```

The percentile interval for 92% confidence level is -4.804819, 6.4152127

e) What are the conditions of the parametric approach you used? Which test and confidence interval would you report to David to answer his question and why (i.e., should you have used the parametric or non-parametric approach)?

**Answer:**

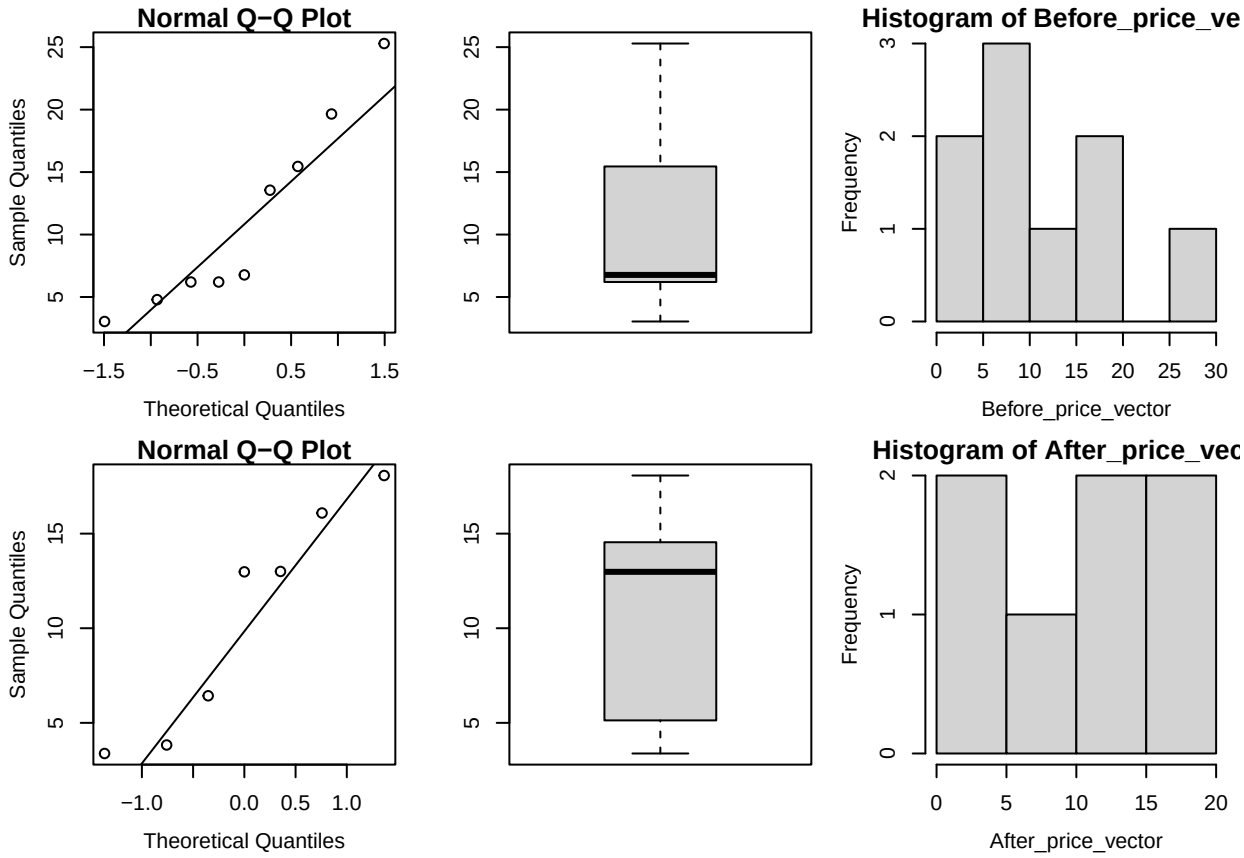
Conditions to approach parametric test are as follows:

1) Independence: We are told that data are randomly sampled.
2) Normality: Observations should be normally distributed
3) Skewness and Outliers: There must not be extreme outliers.

```
par(mfrow = c(2, 3), mar = c(4,4,1.2,0.5), mgp = c(2.5, 1,0))

#checking for skewness
qqnorm(Before_price_vector)
qqline(Before_price_vector)
 boxplot(Before_price_vector)

#checking for normality
hist(Before_price_vector)

#shapiro test for conforming normality status
shapiro.test(Before_price_vector)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Before_price_vector
## W = 0.88815, p-value = 0.1909
```

```
#checking for skewness
qqnorm(After_price_vector)
qqline(After_price_vector)
boxplot(After_price_vector)

#checking for normality
hist(After_price_vector)
```




```
#shapiro test for conforming normality status
shapiro.test(After_price_vector)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  After_price_vector
## W = 0.89483, p-value = 0.3008
```

The qqplots depicts normality upto some extent, while box-plots does not show extreme outliers, lastly we can hardly see the normality in the histogram but to confirm the distribution, we will perform shapiro test. The p-value from the shapiro-test for the observation is greater than 0.05. Hence, the distribution of the given data is not different from normal distribution significantly.

Since, the data follow all the required conditions I will suggest david to go with parametric test as it has more statistical power i.e more ability to detect an effect when an effect is present, while results can be less reliable if the distributions is violated.

## Question 2:

Required data: Provided in the table below. A marketing firm randomly selects customers of a local grocery store and asks: 1) what their favorite frozen meal is; and 2) what is their current housing situation. They want to know if housing type and frozen meal type are independent to help with their advertising campaign. Perform a test at the $\alpha = 0.06$ significance level.

a) Should you perform a parametric or non-parametric test? What is the name of the test (e.g., "t-test", "permutation test").

**Answer:**

Given the distribution, we shall use non-parametric approach by conducting the chi-square test for Independence.

b) Conduct the hypothesis test.

**Answer:**

Setting up the hypothesis

$H_0$ : Housing type and frozen meal type are independent.

$H_A$ : Housing type and frozen meal type are dependent.

Conditions:

1) Each observation is independent of the others.
2) There must be 5 or more expected observations in each cell.

We assume that data is randomly sampled.

```
 eg.as <- data.frame(housing = c(rep("rent", 67), rep("condo", 60),
    rep("town_house",67), rep("detached",70 )), frozenmeal = c(rep("pizza", 14),
    rep("hungary man", 25),rep("burrito", 20),rep("lasagna", 8),
  rep("pizza", 20), rep("hungary man", 18),rep("burrito", 10),rep("lasagna", 12),
  rep("pizza", 22), rep("hungary man", 10),rep("burrito", 5),rep("lasagna", 30),
  rep("pizza", 17), rep("hungary man", 5),rep("burrito", 6),rep("lasagna", 42)
  ))

ctab <- table(eg.as)

ctab
```

```
##             frozenmeal
## housing      burrito hungary man lasagna pizza
##   condo           10          18      12    20
##   detached         6           5      42    17
##   rent            20          25       8    14
##   town_house       5          10      30    22
```

```
chisq.test(ctab, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ctab
## X-squared = 62.578, df = 9, p-value = 4.258e-10
```

From the conducted test, P-value is $< 0.06$, so we reject null hypothesis. Hence, we can say that Housing type and frozen meal type are dependent.

c) Based on research across the country, the market researchers expect stores to sell Pizza, Hungry Man, Burritos, and Lasagna in the proportions: 25%, 20%, 15%, and 40% respectively. Perform a hypothesis test to check whether these proportions are true for the current store (please also state the name of the test).

**Answer:**

Given the proportions, we will be conducting Goodness of fit test.

$H_0$ : Proportions of Frozen meal type are accurate. $H_A$ : Housing type and frozen meal type are not accurate

```
p1 <- c(0.25,0.2,0.15,0.4)
obs <-c(73,58,41, 92)
chisq.test(x = obs, p = p1, correct = FALSE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 3.0556, df = 3, p-value = 0.3831
```

We got p-value $>\alpha$, so we fail to reject null hypothesis. So, we can say that proportions of frozen meal are accurate.

## Question 3:

Food researchers developed a new hotsauce with the goal of raising the perceived spiciness of a food item by 3 points (out of 10), but not more. In order to test how spicy a hotsauce is perceived to be, the researchers randomly sampled 37 individuals and had them:

- eat half of a taco without hotsauce;
- rate the perceived spiciness (out of 10);
- added 1 teaspoon of hotsauce to the remaining half of the taco; and
- rate the perceived spiciness (out of 10).

*Note: treat this spiciness rating as a continuous numeric quantity for this question*

Test whether this hotsauce performed according to their goal at the $\alpha = 0.10$ significance level and construct a 90% confidence interval.

**Answer:**

Setting up the hypothesis,

We are interested in determining if the average spiceness level of sauce is increased by 3 point.

Let $\mu_{diff}$ be the mean differences between spiciness level of sauce and plain taco, $\mu_{diff} = \mu_{sauce}$ - $\mu_{plain}$.

$H_0 : \mu_{diff} = 3$

$H_A : \mu_{diff} \neq 3$

We will be performing paired t-test

```
hot_sauce <- c(6.6,7.8,3.9,6.1,7.3,5,4.5,7.5,6.9,9,6.6,5.7,6.9,5.9,6.2,7.9,
6.3,7.8,6.6,6.5,5.4,4.2,5.8,5,4.1,7.1,6.9,7.4,5.6,6.1,5.9,
4.8,5.7,6.7,5.7,6.3,6.2)
plain_taco <- c(4.1,4.6,1,2.7,3.9,2.2,2.2,4,4.1,5.6,2.5,2.8,2.7,2.9,2.5,4.5,
2.7,4.7,3.3,3.5,2.8,1,2.8,2.3,1,3.4,3.4,4.2,2.5,2.7,2.5,1.7,3.0,4.2,3.1,3.1,2.6)

t.test(hot_sauce, plain_taco,
       alternative = "two.sided", mu = 3, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  hot_sauce and plain_taco
## t = 2.3273, df = 36, p-value = 0.02569
## alternative hypothesis: true mean difference is not equal to 3
## 95 percent confidence interval:
##  3.021195 3.308534
## sample estimates:
## mean difference
##        3.164865
```

By performing the test, we get the p-val= 0.02569 $< \alpha$. So, we reject the null hypothesis. We can say that hotsauce did not performed according to researcher's goal that means spiciness of sauce was not exactly exceeded by 3 points from that of plain tacos.

Calculating 90% confidence interval:

```
  t_sauce_result <-t.test(hot_sauce, plain_taco,
       alternative = "two.sided",conf.level = 0.9, mu = 3, paired = TRUE)
  t_sauce_result
```

```
##
##  Paired t-test
##
## data:  hot_sauce and plain_taco
## t = 2.3273, df = 36, p-value = 0.02569
## alternative hypothesis: true mean difference is not equal to 3
## 90 percent confidence interval:
##  3.045266 3.284463
## sample estimates:
## mean difference
##        3.164865
```

We are 90% confident that mean spiciness level of the sauce lies between (3.0452664, 3.2844633)

## Question 4:

a) What is the difference between a $x^2$-test of homogeneity and a $x^2$-test of independence?

**Answer:**

Mathematically, both the test produces the same outcome, however it does have a main difference, $x^2$-test of independence is used for measuring two variables in one population and $x^2$-test of homogeneity is used when we have to measure one variable for two or more than two population.

b) When should you use a paired sample test for means?

**Answer:**

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

We can identify a paired test when, when the size of both observations are same, Often these are "before and after" measurements on the same experimental units (people or objects).

c) Why do you need to check if each cell contains at least 5 expected (or observed) counts when conducting a $x^2$-test?

**Answer:**

The Chi-square statistic follows a chi-square distribution asymptotically with df=n-1. That means we can use the chi-square distribution to calculate an accurate p-value only for large samples. (That's where the asymptotically comes in). For small samples, it doesn't work.

d) What theorem are we relying on in order to conduct parametric hypothesis tests so far this semester? State the theorem's name and the colloquial definition.

**Answer:**

Parametric tests are conducted on the basis of having prior knowledge of the population distribution (i.e, normal), or if not then we can easily approximate it to a normal distribution which is possible with the help of the Central Limit Theorem.

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.