

# EXPERIMENT 1

Submitted by--Surosh Khan

1. What are the categorical variables in this dataset?

Ans.

Hospt, Treat, Outcome and Gender.

Code.

```
depression=read.csv('C:/Users/Surosh/Downloads/OneDrive_1_7-5-2020/depression.csv')
depression
```

2. What are the quantitative variables in this dataset?

Ans.

Time, AcuteT and Age.

Code

```
depression=read.csv('C:/Users/Surosh/Downloads/OneDrive_1_7-5-2020/depression.csv')
depression
```

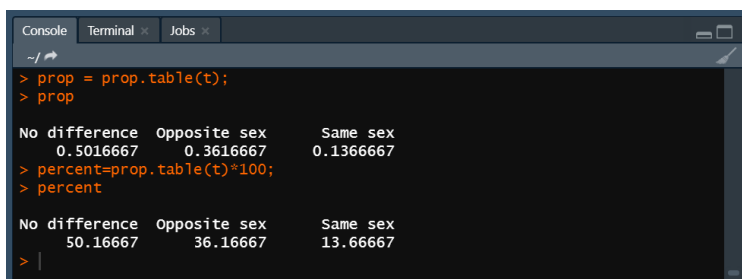
3. Describe the distribution of the variable "friends" in dataset - Survey that asked 1,200 U.S. college students about their body perception

Ans.

About 50% of the students find it as easy to make friends with the opposite sex as with the same sex. Among the remaining 50% of the students, the majority (36.2%) find it easier to make friends with people of the opposite sex, and the remainder (13.7%) find it easier to make friends with people of their own sex.

Code

```
friends=read.csv('C:/Users/Surosh/Downloads/OneDrive_1_7-5-2020/friends.csv')
friends
t = table(friends$Friends);
t
prop = prop.table(t);
prop
percent=prop.table(t)*100;
percent
```



```
Console Terminal Jobs
~/
> prop = prop.table(t);
> prop
No difference Opposite sex Same sex
0.5016667 0.3616667 0.1366667
> percent=prop.table(t)*100;
> percent
No difference Opposite sex Same sex
50.16667 36.16667 13.66667
>
```

4. Describe the distribution of the ages of the Best Actor Oscar winners. Be sure to address shape, center, spread and outliers (Dataset - Best Actor Oscar winners (1970-2013))

Ans.

Shape: the distribution is skewed right. This means that most actors receive the best acting Oscar at a relatively younger age (before age 48), and fewer at an older age.

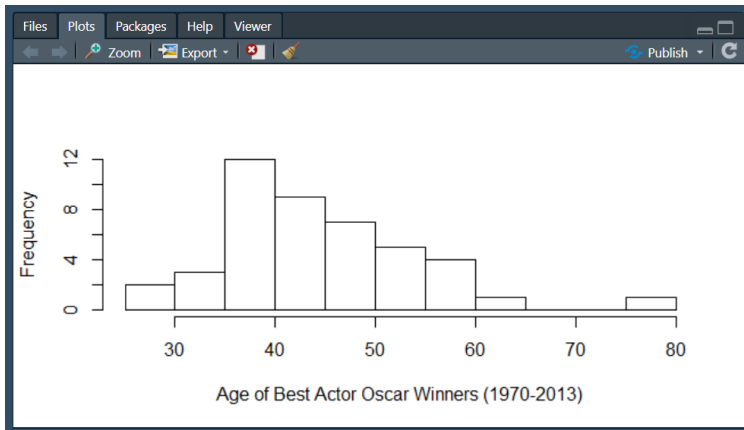
Center: The distribution seems to be centered at around 42-43. This means that about half the actors are 42 or younger when they receive the Oscar, and about half are older.

Spread: The age distribution ranges from about 30 to about 75. The entire dataset is covered, then, by a range of 45 years. It should be noted, though, that there is one high outlier at around age 75, and the rest of the data ranges only from 30 to 60.

Outliers: As mentioned above, there is one high outlier at around age 75.

Code.

```
actor_age=read.csv('C:/Users/Surosh/Downloads/OneDrive_1_7-5-2020/actor_age.csv')
actor_age
hist(actor_age$Age, breaks=8, xlab="Age of Best Actor Oscar Winners (1970-2013)", main="")
```



5. Getting information from the output: a. How many observations are in this data set? b. What is the mean age of the actors who won the Oscar? c. What is the five-number summary of the distribution? (Dataset - Best Actor Oscar winners (1970-2013))

Ans.

- a. There are  $n = 44$  observations in the data set (representing the age of the Best Actor Oscar winners of the 44 years from 1970 through 2013).
- b. Mean = 44.98
- c. The five-number summary is: min = 29, Q1 = 38, M = 43.5, Q3 = 50.5, Max = 76

Code

```
actor_age=read.csv('C:/Users/Surosh/Downloads/OneDrive_1_7-5-2020/actor_age.csv')
actor_age
mean(actor_age$Age)
length(actor_age$Age)
min(actor_age$Age)
max(actor_age$Age)
```

6. Get information from the five-number summary: a. Half of the actors won the Oscar before what age? b. What is the range covered by all the actors' ages? c. What is the range covered by the middle 50% of the ages? (Dataset - Best Actor Oscar winners (1970-2013))

Ans.

- a. Half the actors won the Oscar before age 43.5 (the median).
- b. The range covered by all the ages is: Range = Max - min = 76 - 29 = 47.
- c. The range covered by the middle 50% of the ages is: IQR = Q3 - Q1 = 50.5 - 38 = 12.5

Code

```
actor_age=read.csv('C:/Users/Surosh/Downloads/OneDrive_1_7-5-2020/actor_age.csv')
actor_age
median(actor_age$Age)
```

```
mn=min(actor_age$Age)
mx=max(actor_age$Age)
mx-mn
Q1=quantile(actor_age$Age, 0.25)
Q3=quantile(actor_age$Age, 0.75)
Q3-Q1
```

7. What are the standard deviations of the three rating distributions? Was your intuition correct? (Dataset - 27 students in the class were asked to rate the instructor on a number scale of 1 to 9)  
Ans.

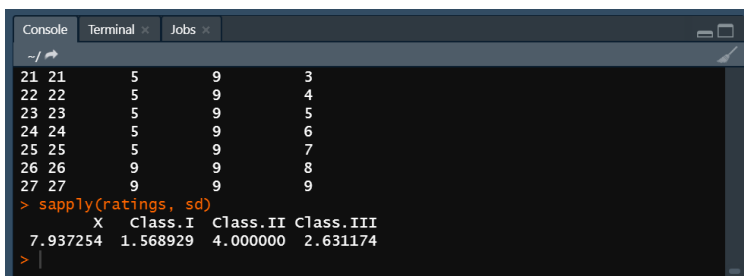
Here are the three standard deviations: Class I: 1.6 Class II: 4.0 Class III: 2.6 Note that through this example, we also learn that the number of distinct values represented in a histogram does not necessarily indicate greater variability.

Code

```
ratings=read.csv('C:/Users/Surosh/Downloads/OneDrive_1_7-5-2020/ratings.csv')
```

```
ratings
```

```
sapply(ratings, sd)
```



```

> sapply(ratings, sd)
      X      class.I      class.II      class.III
7.937254 1.568929 4.000000 2.631174

```

8. Assume that the average rating in each of the three classes is 5 (which should be visually reasonably clear from the histograms), and recall the interpretation of the SD as a "typical" or "average" distance between the data points and their mean. Judging from the table and the histograms, which class would have the largest standard deviation, and which one would have the smallest standard deviation? Explain your reasoning (Dataset - 27 students in the class were asked to rate the instructor on a number scale of 1 to 9)

Ans. In class I, almost all the ratings are 5, which is also the mean. The average distance between the observations and the mean, then, would be very small. In class II most of the observations are far from the mean (at 1 or 9). The average distance between the observations and the mean in this case would be larger. Class III is the case where some of the observations are close to the mean, and some are far, so the average distance between the observations and the mean would be somewhere in between class I and II. This observation would lead me to conclude that the standard deviation would be ranked (from smallest to largest): Class I, Class III, Class II.