Assignment #2 (25pts) - Data Science and Management

SKKU 2010313470
KyoungLim Kwak

1. (5pts) The data below are the number of points scored in 30 games by the Portland Trailblazers.

Scores: 90, 95, 89, 71, 73, 96, 87, 95, 107, 89, 96, 80, 97, 95, 102, 97, 93, 101, 82, 83, 74, 91, 83, 98, 95, 111, 99, 120, 93, 84

a. Estimate the sample mean score. What does the quantity estimate?

The sample me score is 92.2 and it means the estimated population mean.

```
> data=c(90,95,89,71,73,96,87,95,107,89,96,80,97,95,102,97,93
,101,82,83,74,91,83,98,95,111,99,120,93,84)
> xbar=mean(data)
> xbar
[1] 92.2
```

b. Is the estimate in part(a) likely to equal the population parameter? Why or why not?

No. Because P-value is less than 0.05, H0(=Sample mean is equal to population mean) can be rejected.

```
> data=c(90,95,89,71,73,96,87,95,107,89,96,80,97,95,102,97,
93,101,82,83,74,91,83,98,95,111,99,120,93,84)
> xbar=mean(data)
> xbar
[1] 92.2
> t.test(data)

        One Sample t-test

data:  data
t = 46.647, df = 29, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 88.15754 96.24246
sample estimates:
mean of x
     92.2
```

c. Calculate the standard error for your sample estimate.

The standard error is 1.976529.

```
> SE <- sd(data)/sqrt(length(data))
> SE
[1] 1.976529
```

d. What does the quantity in part(c) measure?

The standard error means the standard deviation of the sampling distribution of statistic.

e. Calculate a 95% confidence interval for the population mean.

A 95% confidence interval for the population mean is (88.15754, 96.24246). We already know this through a t-test made in part(c).

f. **Provide an interpretation for the interval you calculated in part(e).**

Confidence interval can be expressed in terms of repeated samples. Were this procedure to be repeated on multiple samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter 90% of the time. Note that this does not refer to repeated measurement of the same sample, but repeated sampling. Thus, the interval calculated in part(e) include the population mean 95%.

2. **(5pts) Using the following data, test the null hypothesis that male and females have the same mean cholesterol concentrations. Include descriptive statistics, hypothesis testing (e.g. t-test) and 95% confidence intervals.**

Males: 220.1, 218.6, 229.6, 228.8, 222, 224.1, 226.5
Females: 223.4, 221.5, 230.2, 224.3, 223.8, 230.8

Because P-value is bigger than 0.05, H0(=Male and females have the same mean cholesterol concentrations) cannot be rejected. Descriptive statistics can be summarized in a simple table.

	Male	Female
Mean	224.2429	225.6667
Standard Deviation	4.254745	3.866609

Here is the outcome of hypothesis testing and 95% confidence intervals(-6.386747, 3.539128).

```
> male=c(220.1, 218.6, 229.6, 228.8, 222, 224.1, 226.5)
> female=c(223.4, 221.5, 230.2, 224.3, 223.8, 230.8)
> mean(male)
[1] 224.2429
> sd(male)
[1] 4.254745
> mean(female)
[1] 225.6667
> sd(female)
[1] 3.866609
> t.test(male, female, var.eqaul=T)

        Welch Two Sample t-test

data:  male and female
t = -0.63184, df = 10.942, p-value = 0.5405
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.386747  3.539128
sample estimates:
mean of x mean of y
 224.2429  225.6667
```

3. (5pts) A clinical trail was carried out to test whether a new treatment has an effect on the rate of recovery of patients. The null hypothesis "H0: the treatment has no effect" was rejected with a P-value of 0.04. The researchers used a significance level of 5%. State whether the following conclusions is correct. If not, explain why.

a. The treatment has only a small effect.

[False] From this test, we don't know the size of effect of the treatment. We just know whether this treatment has effect or not.

b. The treatment has some effect.

[True] H0 can be rejected in a signification level of 5%. So, It is true that this treatment has some effect.

c. The probability of committing a Type I error is 0.04.

[False] The probability of committing a Type I error would be 0.05, not 0.04.

d. The probability of committing a Type II error is 0.04.

[False] We can't derive a Type II error from this test.

e. The null hypothesis would not have been rejected if the significance level was α=0.01.

[True] Because the P-value is 0.04, the null hypothesis cannot be rejected in a significance level of 1%.

4. (5pts) The data below are volumes of red blood cells from two individuals. Test the hypothesis (using the Mann-Whitney test) that the red blood cells of person B are 1.5 times the volume of person A.

person A: 248, 236, 269, 254, 249, 251, 260, 245, 239, 255
person B: 380, 391, 377, 392, 398, 374

Because P-value is bigger than 0.05, H0(=The red blood cells of person B are 1.5 times the volume of person A) cannot be rejected.

```
> # pa = person A, pb = person B
> pa=c(248, 236, 269, 254, 249, 251, 260, 245, 239, 255)
> pb=c(380, 391, 377, 392, 398, 374)
> wilcox.test(1.5*pa,pb,alter="two.sided")

        Wilcoxon rank sum test

data:  1.5 * pa and pb
W = 16, p-value = 0.1471
alternative hypothesis: true location shift is not equal to 0
```

5. (5pts) What is the difference between the standard error of mean and the standard deviation? Provide example data that illustrates their difference.

Put simply, the standard error of mean is an estimate of how far the sample mean is likely to be from the population mean, whereas the standard deviation is the degree to which individuals within the sample differ from the sample mean.

If the population standard deviation is finite, the standard error of mean will tend to zero with increasing sample size, because the estimate of the population mean will improve, while the standard deviation will tend to the population standard deviation as the sample size increases.