Homework 4

Statistical Inference, Spring 2021

1-

a. False, the standard deviation of the original sample will be much larger than that of the bootstrap distribution. This is because the standard deviation of the sample includes single observations whereas the bootstrap distribution contains sample means. Single observations will vary much more than sample means.

b. False, the resamples should be the same sample size as the original sample for it to be a representative distribution.

c. False, the bootstrap distribution is created by resampling with replacement from the original sample.

d. True

e. True

f. True

g. True

h. True

i. False, If you care more about the type II error, you must assign a large amount of α.

2- R Question

3-

a.

$n = 25$

$\bar{y} = 25.028$

$sd = 1.34$

$H_0: \mu = 24.3, H_a: \mu \neq 24.3$

$$SE = \frac{sd}{\sqrt{n}} = \frac{1.34}{\sqrt{25}} = 0.268$$

$$T = \frac{\bar{y} - \mu}{SE} = \frac{25.028 - 24.3}{0.268} = 2.713$$

$$df = 25 - 1 = 24$$

$$\alpha = 0.05 \rightarrow p_{value} = P(T \geq 2.713) * 2 \approx 0.012$$

$$\rightarrow p_{value} < 0.05 \ \rightarrow We\ reject\ null\ hypothesis$$

So the body temperature of crabs exposed to ambient air temperature is different than the ambient air temperature.

b.

$$\alpha = 0.05 \rightarrow t^*_{df} = 2.064$$

$$C.I. = \bar{y} \pm t^*_{df} SE = 25.028 \pm 2.064 * 0.268 = (24.474, 25.582\,)$$

4-

$\bar{x}_{men} = 14.95, \quad s_{men} = 6.84, \quad n_{men} = 13$
$\bar{x}_{women} = 22.29, \quad s_{women} = 5.32, \quad n_{women} = 10$

$$SE_{(\,\bar{x}_{men} - \bar{x}_{women}\,)} = \sqrt{\frac{s^2_{men}}{n_{men}} + \frac{s^2_{women}}{n_{women}}} = \sqrt{\frac{6.84^2}{13} + \frac{5.32^2}{10}} = 2.536$$

$$df = \min(n_{men} - 1, n_{women} - 1) = \min(9, 12) = 9$$

$$H_0: \mu_{men} - \mu_{women} = 0$$

$$H_a: \mu_{men} - \mu_{women} \neq 0$$

$$T = \frac{(\bar{x}_{men} - \bar{x}_{women}) - 0}{SE_{(\,\bar{x}_{men} - \bar{x}_{women})}} = \frac{(14.59 - 22.29) - 0}{2.536} = -2.895$$

$$p_{value} = P(T \leq -2.895) * 2 \approx 0.018$$

$$\rightarrow p_{value} < 0.05 \ \rightarrow We\ reject\ null\ hypothesis$$

So there is a major difference between men and women body fat percentage.

5-

$H0$: μ = 0.30

HA: μ > 0.30

A high power means the lowest probability of making the Type II error.

She's dealing with the sample size and what is the true proportion of customers that would buy coffee?

The sample size is under her control but the true proportion isn't. Then she can change the sample size.

The general principle is the higher the sample size, the higher the power, so you want the highest possible sample size and u r going to have a higher power if the true proportion is further from your null hypothesis.

Choice (D) is correct

6-

H_0: μ = 7.5

H_A: μ < 7.5

n = 55

s = 1.6

$\bar{y}$ = 7.22

a.

Condition check:

$random\ sample$d and n < 10% $of\ total\ employees\ of\ the\ company$ → independence condition

n = 55 ≥ 30 and d$istribution\ close\ to\ normal$ → sample size condition

$$\sigma = 1 \cdot 6 \to SE = \frac{1.6}{\sqrt{55}} = 0.2157$$

$$Z = \frac{\bar{x}\ - \mu}{SE}\ \frac{7.22 - 7.5}{0.2157} = -1.2978$$

$$pvalue = P(Z < -1.2978) = 0.0975$$

Since $pvalue$ = 0.0975 > α =0.05 → we fail to reject H_0. Then we can say the average of annual paid vacation days is equal to the national average.

b.

$\mu_a = 7.4$

α =0.05

qnorm(0.05) ---> -1.644854

$$Power = P\left(\frac{\bar{x} - 7.5}{0.2157} > -1.645 \middle| \bar{x} \sim N(\mu = 7.4, SE = 0.2157)\right)$$

$$= P\left(\bar{x} > 7.1451 \middle| \bar{y} \sim N(\mu = 7.4, SE = 0.2157)\right) =$$

$$= p(Z > \frac{7.1451 - 7.4}{0.2157} = -1.1817) = 0.119$$

pnorm(-1.181488) ---> 0.1187045

β = 1 – power =1- 0.119= 0.881

7-

a.

$$H_0: There\ is\ no\ difference\ between\ the\ three\ levels$$

$$H_a: There\ is\ difference\ between\ the\ three\ levels$$

b.

	Class	DF	Sum SQ	Mean SQ	F Value
Group	Class	2	152477.7	76238.85	1.395
Error	Residual	15	819833.3	54655.5	
	Total	17	972311		

Degrees of freedom associated with ANoVA:

$$Total:\ df_T = n - 1 = 18 - 1 = 17$$

$$Group:\ df_G = k - 1 = 3 - 1 = 2$$

$$Error:\ df_E = df_T - df_G = 17 - 2 = 15$$

Sum of Squares Total (SST):

	n	mean
overall	18	817.8

$$y_i: value\ of\ the\ response\ variable\ for\ each\ observation$$

$$\bar{y}: grand\ mean\ of\ the\ response\ variable$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{18} (y_i - 817.8)^2 = 972311$$

Sum of Squares Groups (SSG):

$$mean_i = \frac{\sum_{i=1}^{6} y_i}{6}$$

	n	mean
Normal	6	938.3
Osteopenia	6	800
osteoporosis	6	715

$$n_j: numer\ of\ observations\ in\ group\ j$$

$$y_j: mean\ of\ the\ response\ variable\ for\ group\ j$$

$$\bar{y}: grand\ mean\ of\ the\ response\ variable$$

$$SSG = \sum_{j=1}^{k} n_j\left(y_j - \bar{y}\right)^2 = \sum_{j=1}^{3} n_j\left(y_j - 817.8\right)^2 = 152477.7$$

Sum of Squares Error (SSE):

$$SSE = SST - SSG = 819833.3$$

Mean square Error:

$$Group: MSG = \frac{SSG}{df_G} = 76238.85$$

$$Error: MSE = \frac{SSE}{df_E} = 54655.55$$

$$F = \frac{MSG}{MSE} = 1.395$$

c.

$$p_{value} = \Pr(> F) = 0.278 > 0.05$$

$$\rightarrow There\ is\ no\ difference\ between\ the\ three\ levels$$