[HTML]DB5800

Fatemeh Nadi
810101285
fatemehnadi@ut.ac.ir

Homework 4
Statistical Inference , Fall 1401

December 16, 2022

1

- a

  FALSE: Bootstrap distributions are created by resampling with replacement from the original sample, not from the population.

- b

  TRUE.

- c

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

- d

  TRUE

- e

  FALSE, In ANOVA, the More variance within the groups, the greater the noise (variance).

- f

  TRUE

- g

  TRUE

- h

  FAlSE, If you want to prevent the type II error, you must assign a large amount of .

2

- a

  $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{76+84+69+92+58+89+73+97+85+77}{10} = 80$

  $var = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = 137.11$

  $s = \sqrt{var} = 11.7$

- b

  can assume that $CI = 95\%$ so:
  Checking conditions to use CLT:
  1. Independence: random sample and n ¡ 10% of all students' university. 2. Sample size/skew: n 30 !!
  So, we can't assume that the sampling distribution of average scores of 10 students this university from samples of size 500 will be nearly normal.
  But we can use Student's t-distribution for this problem

  $$margin\ of\ error = |t^*_{df}\ SE|$$

  $df = n - 1 = 9$, $SE = \frac{s}{\sqrt{n}} = \frac{11.7}{\sqrt{10}} = 3.70$

  $CI = 95\% \rightarrow (10.95)/2 = 0.025$, $t^*_9 \quad = \quad > qt(0.025, df = 9) = -2.26$

  $$margin\ of\ error = |-2.26 \times 3.7| = 8.37$$

- c

  like part b we have:

  $$point\ estimate\ \pm margin\ of\ error = point\ estimate\ \pm t^*_{df}\ SE$$

  $df = n - 1 = 9$, $SE = \frac{s}{\sqrt{n}} = \frac{11.7}{\sqrt{10}} = 3.70$

  $CI = 90\% \rightarrow (10.90)/2 = 0.05$, $t^*_9 \quad = \quad > qt(0.05, df = 9) = -1.83$

  $point\ estimate = \bar{x} = 80$
  approximate 90% CI for μ: 80 $\pm$1.83$\times$SE:

  $$(73.21, 86.78)$$

- d

  We are 90% confidence that the mean point of students' score in this university is between 73.21 and 86.78.

3

- a

Let's define null hypothesis (H0) and alternative hypothesis (HA):

$$H_0 : \mu = 8 \quad on\ average,\ New\ Yorkers\ sleep\ for\ 8\ hours.$$
$$H_a : \mu < 8 \quad on\ average,\ New\ Yorkers\ sleep\ less\ than\ 8\ hours\ a\ night.$$

- b

Checking conditions to use CLT:
1. Independence: random sample and n ¡ 10% of all New Yorkers. 2. Sample size/skew: n 30 !!
So, we can't assume that the sampling distribution of average duration of sleep per night of 25 New Yorkers will be nearly normal.
But we can use Student's t-distribution for this problem
$df = n - 1 = 24$, $SE = \frac{s}{\sqrt{n}} = \frac{0.77}{\sqrt{25}} = 0.154$

$t = \frac{\bar{x}-\mu}{SE} = \frac{7.73-8}{0.154} = -1.75$

- c

$$\begin{aligned} p - value &= P\left(observed\ or\ more\ extreme\ outcome \mid H_0\ true\right) \\ &= P\left(\bar{x} \leq 7.73 \mid H_0 :\ \mu = 8\right) \\ &= P\left(t \leq -1.75\right) \end{aligned}$$

$> pt(1.75, df = 24, lower.tail = FALSE) = 0.046$
According to the value of the p-value, we can decide about the result of this test reject H0 or not.

- d

In this case, we reject the null hypothesis since the p-value is less than the alpha level of 0.05.
so, New Yorkers sleep less than 8 hours a night on average.

- e

$CI = 90\% \rightarrow 10.90/2 = 0.05$, t^*_{24} $=$ $> qt(0.05, df = 24) = -1.71$

$$\begin{aligned} point\ estimate\ \pm margin\ of\ error &= \bar{x} \pm t^*_{df}\ SE \\ &= 7.73 \pm 1.71 \times 0.154 \end{aligned}$$

approximate 90% CI for μ: 7.73 ±0.26:

$$(7.47, 7.99)$$

4

Comparing Two Means

Based on the probability distribution, establish the null hypothesis and alternative hypothesis.

$H_0 : \mu_{intensive} - \mu_{paced} = 0 \quad on\ average,\ intensive\ tutoring\ is\ not\ statistically\ different\ from\ paced\ tutoring.$
$H_a : \mu_{intensive} - \mu_{paced} \neq 0 \quad on\ average,\ intensive\ tutoring\ is\ statistically\ different\ from\ paced\ tutoring.$

Conditions for inference for comparing two independent means:

1. Independence:

1.1 : within groups: sampled observations must be independent : random sample/assignment, if sampling without replacement, $n < 10\%$ of population(students).
1.2 : between groups: groups are tutored separately

2. Sample size/skew: The more skew in the population distributions, the higher the sample size needed.

Based on these conditions, we can assume that they are met.

$$\begin{aligned} SE_{(\bar{x}_{intensive} - \bar{x}_{paced})} &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= \sqrt{\frac{6.44^2}{12} + \frac{7.52^2}{10}} \\ &= 3.018 \end{aligned}$$

$df = min\,(n_1 - 1, n_2 - 1) = 9$

Assume the confidence interval is 95%, so:

$$t_9^* = qt(.025, 9) = -2.26$$

$$point\ estimate\ \pm margin\ of\ error$$

$$\bar{x}_{intensive} - \bar{x}_{paced} \pm t_{df}^*\ SE_{(\bar{x}_{intensive} - \bar{x}_{paced})}$$

$$3.52 \pm -2.26 \times 3.018$$

$$\pm 6.827$$

approximate 95% CI for μ: 3.52 $\pm$6.827: $(-3.307, 10.347)$

$$\begin{aligned} T_9 &= \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{SE} \\ &= \frac{46.71 - 42.79 - 0}{3.018} \\ &= 1.298 \end{aligned}$$

$pt(1.298, df = 9, lower.tail = FALSE) * 2 = 0.226 \not< 0.05$

$\rightarrow$ Can't reject H_0, Based on the confidence interval and t-test, we can say intensive tutoring isn't statistically different from paced tutoring (covering less material in the same amount of time).

5

Calculate the required sample size for 80% power:
$Effect\ size = 0.5, \quad power = 1 - \beta = 80 \rightarrow = \beta = 0.2, \quad \alpha = 0.05, \quad s = 2.2$

$$Effect\ size = \frac{\mu_{new} - \mu_{current}}{s} = 0.5 \rightarrow \mu_{new} - \mu_{current} = 1.1$$

Let's transform our sampling distribution under the null hypothesis to a standard normal distribution to make the calculations more straightforward.

1. The critical value at 0.05 significance is 1.96.

2. If we now consider the sampling distribution given the alternative hypothesis, then we want the area under the curve between -1.96 and 1.96 to equal 20% (for 80% power). Therefore, the critical value has to be a distance of 0.84 away from the mean.

So the total standardized difference between means must be 1.96 + 0.84 = 2.8.
This may be easier to understand graphically:

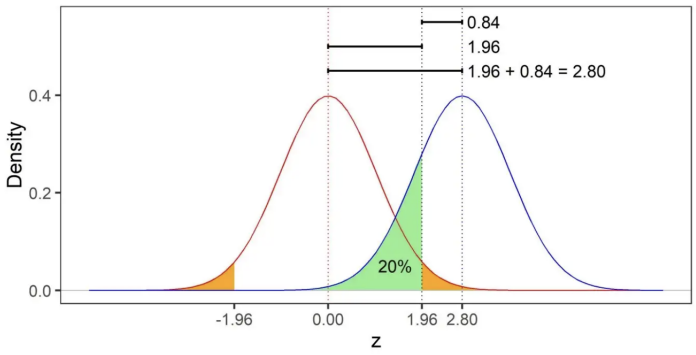

Figure 1

We also know that the real difference between means ($\mu_{new} - \mu_{current}$) is 1.1, therefore, the standardized difference equals $\frac{1.1}{SE}$. So we can construct an equality and solve for SE:

$$SE_{new-current} = \frac{1.1}{2.8} = 0.392$$

We know that:

$$SE_{new-current} = \sqrt{\frac{s_{new}^2}{n_{new}} + \frac{s_{current}^2}{n_{current}}}$$

$s_{new} = s_{current} = 2.2, \quad n_{new} = n_{current} = n$, so we can construct an equality and solve for n:

$$n = \sqrt{\frac{2 \times (2.2)^2}{(3.92)^2}} = 62.72$$
$$\geq 63$$

6

Analyzing paired data

This two sets of observations(platelet aggregation before and after smoking) have this special correspondence. they are said to be paired.
To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations:

$$diff = after \ - \ before$$

before	after	diff
25	27	2
25	29	4
27	37	10
44	56	12
30	46	16
67	82	15
53	57	4
52	61	9
53	80	27
60	59	-1
28	43	15

$\bar{x}_{diff} = 10.27, \quad var_{diff} = 63.61, \quad s_{diff} = 7.97, \quad n = 11, \quad \alpha = 0.05$

Let's define null hypothesis (H_0) and alternative hypothesis (H_A) for paired means:

$H_0 : \mu_{diff} = 0 \quad on\ average,\ There\ is\ no\ difference\ between\ before\ and\ after\ platelet\ aggregation.$
$H_a : \mu_{diff} \neq 0 \quad on\ average,\ There\ is\ difference\ between\ before\ and\ after\ platelet\ aggregation.$

Calculate the test statistic and the p-value for this hypothesis test.

$df = n - 1 = 10, \quad SE = \frac{s}{\sqrt{n}} = \frac{7.97}{\sqrt{10}} = 2.52$

$$T = \frac{\bar{x}_{diff} - \mu_{diff}}{SE} = \frac{10.27 - 0}{2.52} = 4.07$$

Now we must calculate the probability of obtaining a random sample of 11 individuals where the average difference between before and after platelet aggregation is at least 10.27 (in either direction) if the actual

average difference is 0.

$P_{value} = p\,(T \geq 4.07) * 2 = pt(4.075397, df = 10, lower.tail = FALSE) * 2 = 0.0022 < 0.05$

Reject $H_0 \rightarrow$ There is difference between before and after platelet aggregation.

7

- a

Let's define null hypothesis (H_0) and alternative hypothesis (H_A):

$H_0 : \mu_{diff} = 0 \quad on\ average,\ There\ is\ no\ adifference\ between\ the\ height\ of\ three\ groups\ of\ men\ in\ these\ thre$
$H_a : \mu_{diff} \neq 0 \quad on\ average,\ There\ is\ adifference\ between\ the\ height\ of\ three\ groups\ of\ men\ in\ these\ three\ c$

- b

		DF	Sum SQ	Mean SQ	F-Value
Groups	Class	df_G	SSG	MSG	F
Error	Residuals	df_E	SSE	MSE	
	Total	df_T	SST		

$SST = \sum_{i=1}^{n=24} (y_i - \bar{y})^2 = 660503.5$
y_i: value of the response variable for each observation
$\bar{y}$: grand mean of the response variable.

	n	Mean	sd
US	8	177.375	4.299
UK	8	177.5	5.979
India	8	176.875	5.599
Overall	24	177.25	5.348

$SSG = \sum_{j=1}^{k=3} n_j (\bar{y}_j - \bar{y})^2 = 1.75$

n_j: number of observations in group j
$\bar{y}_j$: mean of the response variable for group
$\bar{y}$: grand mean of the response variable.

$SSE = SST - SSG = 660501.75$

$df_T = n - 1 = 23$, $df_G = k - 1 = 2$, $df_E = df_T - df_G = 21$

$MSG = \frac{SSG}{df_G} = 0.875$, $MSE = \frac{SSE}{df_E} = 31452.464$

F statistic: Ratio of the average between group and within group variabilities:

$$F = \frac{MSG}{MSE} = 2.7819759750825792e - 05$$

		DF	Sum SQ	Mean SQ	F-Value
Groups	Class	2	1.75	31452.464	2.7 e-05
Error	Residuals	21	660501.75	31452.464	
	Total	23	660503.5		

- c

$p_{value} = Pr\,(> F) = pf(2.78e - 05, 2, 21, lower.tail = FALSE) = 0.99 \simeq 1 > \alpha$
Can't reject H_0. On average, There is a difference between the height of men in these three countries

8 R

Please see this file: "Q8-R.Rmd" and "Q8-R.html" and "Q8-R.R"
Code and explanation are provided.

- a

```
#preprocessing
diet <- diet %>%
  mutate(Diet = recode_factor(Diet, `1` = "A", `2` = "B", `3` = "C")) %>%
  mutate_at(c("gender"),
            ~ recode_factor(.x, `0` = "Female", `1` = "Male"))


diet$weight.loss <- diet$pre.weight - diet$weight6weeks
diet$Diet  <- factor(diet$Diet,levels=c("A","B","C"))
diet$gender      <- factor(diet$gender,levels=c("Male","Female"))



#draw bpx plot
boxplot(diet$weight.loss  ~ diet$Diet, ylab = "Weight loss (kg)", xlab = "Diet type",
        main = "side-by-side boxplots")
```

Figure 2: Q8-a code

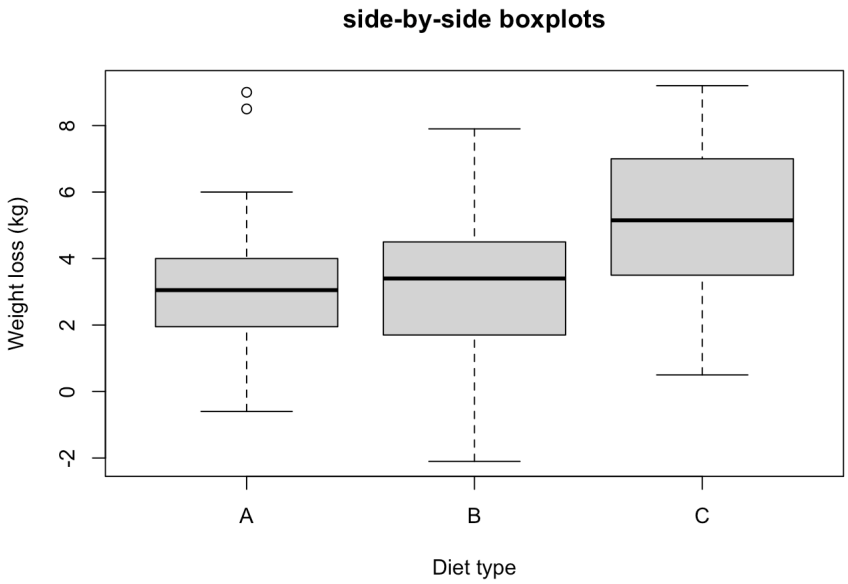

Figure 3

- b
  Conditions for ANOVA:

  1. Independence:

  1.1 within groups: sampled observations is independent.
  1.1.1 random sample / assignment.
  1.1.2 each n_j less than 10% of respective population.

  1.2 between groups: the groups is independent of each other (non-paired).

  2. Approximate normality: distributions should be nearly normal within each group.

  3. Equal variance: groups should have roughly equal variability.

  Based on these conditions, we can assume that they are met.

  Based on the probability distribution, establish the null hypothesis and alternative hypothesis.

$$H_0 : \mu_A - \mu_B = 0 \quad On\ average\ the\ mean\ of\ all\ group\ means\ are\ equal.$$
$$H_a : \mu_A - \mu_B \neq 0 \quad At\ least\ one\ group\ mean\ is\ different\ from\ the\ rest.$$

```
fisher  = aov(weight.loss~Diet,data=diet)
```

Figure 4: Q8-b ANOVA

- c

```
summary(fisher)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Diet         2   69.3   34.67    6.32 0.00284 **
## Residuals   79  433.3    5.49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5

Reject H_0 so at least one group mean is different from the rest.

- d

  Based on the probability distribution, establish the null hypothesis and alternative hypothesis.

  $H_0 : \mu_A - \mu_B = 0 \quad On\ average,\ There\ isn't\ a\ difference\ in\ weight\ loss\ between\ the\ type\ of\ diet\ A\ and\ B.$
  $H_a : \mu_A - \mu_B \neq 0 \quad On\ average,\ There\ is\ a\ difference\ in\ weight\ loss\ between\ the\ type\ of\ diet\ A\ and\ B.$

  $Var\,(A) = 5.22, \quad Var\,(b) = 6.16$, not equal therefore, use pairwise comparisons

```
gp <- unique(diet$Diet)

A <- diet$weight.loss[which(diet$Diet==gp[1])]
B <- diet$weight.loss[which(diet$Diet==gp[2])]

diet.t.test <- t.test(A , B)
diet.t.test
```

Figure 6: Q8-d code

```
##
##  Welch Two Sample t-test
##
## data:  A and B
## t = 3.1999, df = 54.724, p-value = 0.002289
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7507534 3.2678181
## sample estimates:
## mean of x mean of y
##  5.070000  3.060714
```

Figure 7: Compare the weight loss between the groups of A and B

$$\alpha^* = \frac{\alpha}{K} = \frac{0.05}{3} = 0.016, \alpha = 0.05, K = \binom{3}{2} = 3$$

$\alpha^* = 0.0022 < 0.016 \rightarrow RejectH_0$ mean of type A is different from the mean of type B.

And confidence interval is (0.7507534, 3.2678181)

9

Please see this file: "Q9-R.Rmd" and "Q9-R.html" and "Q9-R.R"
Code and explanation are provided.

- a

```
bootstrap <- replicate (1000, sample (house$Area..Meter., replace=T))
mean <-mean(bootstrap)
SE <- sd(bootstrap)/sqrt (length (mean))
```

```
B = 1000
n = nrow(house)
boot.samples = matrix(sample(house$Area..Meter., size = B * n, replace = TRUE),B, n)
boot.statistics = apply(boot.samples, 1, mean)

require(ggplot2)
ggplot(data.frame(meanprice = boot.statistics),aes(x=meanprice)) +
geom_histogram(binwidth=2,aes(y=..density..)) +
geom_density(color="red")
```

Figure 8: Q9-a code

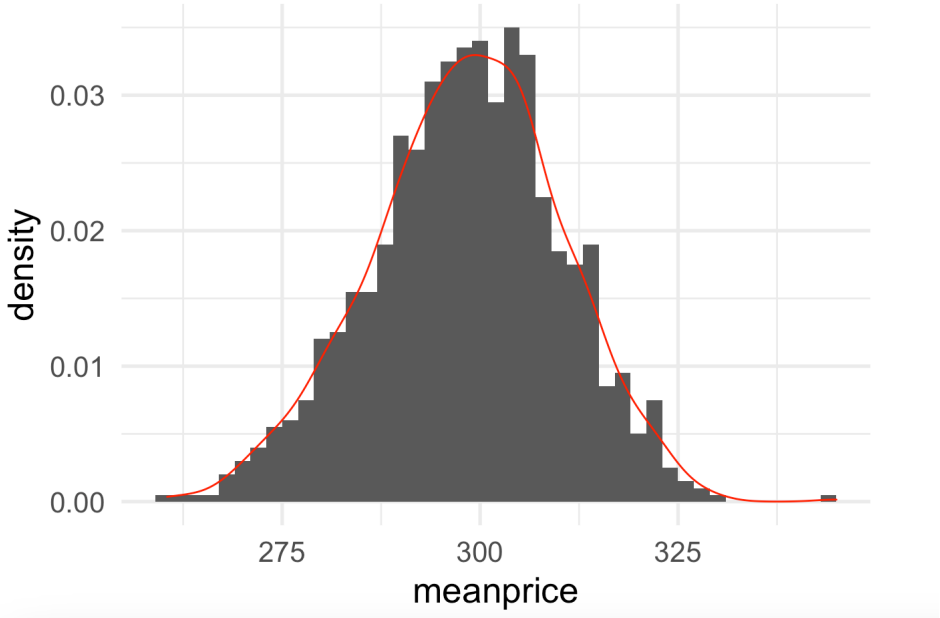

Figure 9: bootstrap distribution of houses

- b

- c