Stat 615/415 (Regression) HW #1 – Solutions

Review of Estimation and Hypothesis Testing (handouts, your old notes, ...)

1. The manufacturer of a certain brand of household light bulbs claims that the bulbs produced by his factory have an average life of at least 2,000 hours. The mean and standard deviation of 20 light bulbs selected from the manufacturer's production process were calculated to be 2,160 and 142 hours, respectively.

   (a) Do the data represent sufficient evidence to support the manufacturer's claim? How can you interpret your answer?

   (b) Construct a 95% confidence interval for the mean lifetime of household light bulbs.

   SOLUTION. *Given:* $n = 20$, $\bar{X} = 2160$, $s = 142$.

   (a) Test $H_0 : \mu = 2000$ *vs* $H_A : \mu > 2000$.

   Test statistic $t_{stat} = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}} = \dfrac{2160 - 1000}{142/\sqrt{20}} = \underline{5.04}$.

   Pvalue $= \boldsymbol{P}\{t \geq 5.04\} < 0.0001$, *from the table of t-distribution with* $n - 1 = \underline{19}$ *d.f.*

   With such a low P-value, reject H_0*; there is sufficient evidence that the average live of these light bulbs exceeds than 2000 hours.*

   (b) A 95% confidence interval for the mean lifetime is

$$\bar{X} \pm t_{0.025}\frac{s}{\sqrt{n}} = 2160 \pm (2.093)\frac{142}{\sqrt{20}} = 2160 \pm 66.5 = \boxed{[2093.5, 2226.5]}$$

2. There are two manufacturing processes, old and new, that produce the same product. The defect rate has been measured for 20 days for the old process, and for 14 days for the new process, resulting in the following sample summaries.

	OLD	NEW
Average defect rate	4.7	2.3
Standard deviation	6.8	5.0

   The firm is interested in switching to the new process only if it can be demonstrated convincingly that the new process reduces the defect rate. Is there significant evidence of that? Use $\alpha = 5\%$; assume that the collected data represent two random samples from Normal distributions. Use the method of testing that is appropriate for this situation.

   SOLUTION. *Test* $H_0 : \mu_1 = \mu_2$ *against* $H_1 : \mu_1 > \mu_2$ *(that is, the defect rate of the old process is higher).*

   To choose the correct method, we need to compare variances by testing

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{vs} \quad H_A : \sigma_X^2 \neq \sigma_Y^2.$$

   The F-statistic is

$$F_{\text{obs}} = \frac{s_X^2}{s_Y^2} = \frac{6.8^2}{5.0^2} = 1.85,$$

   and the corresponding P-value for this two-sided test is

$$P = 2\boldsymbol{P}\{F > 1.85\} = 0.260,$$

   from F-distribution with $n_X - 1 = 19$ *and* $m - 1 = 13$ *d.f. (in R,* `2*(1-pf(1.85,19,13))` *).*

Thus, there is no evidence that variances are unequal, and we can use the pooled standard deviation

$$s_p = \sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}} = \sqrt{\frac{(19)(6.8)^2 + (13)(5.0)^2}{19 + 13}} = \sqrt{37.6} = 6.13$$

Now we are testing

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_1 : \mu_1 > \mu_2.$$

Under the assumption of equal variances, the test statistic is

$$t_{\text{STAT}} = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} = \frac{4.7 - 2.3}{6.13\sqrt{\frac{1}{20} + \frac{1}{14}}} = \underline{1.12}.$$

Compute the p-value

$$Pvalue = \boldsymbol{P}\{t \geq t_{\text{STAT}}\} = \boldsymbol{P}\{t \geq 1.12\} = \text{ more than 0.10 },$$

from the table of T-distribution with $20 + 14 - 2 = 32$ *d.f. (because 1.12 < 1.309). The exact P-value is 0.13558, obtained by R command* `1-pt(1.12,32)`. *Either way, the p-value is high, and* H_0 *is not rejected. The data do not provide sufficient evidence that the new manufacturing process reduces the defect rate.*

3. (Required for Stat-615, optional for Stat-415) An account on server A is more expensive than an account on server B. However, server A is faster. To see if whether it's optimal to go with the faster but more expensive server, a manager needs to know how much faster it is. A certain computer algorithm is executed 30 times on server A and 20 times on server B with the following results,

	Server A	**Server B**
Sample mean	6.7 min	7.5 min
Sample standard deviation	0.6 min	1.2 min

(a) Is there a significant difference between the two servers?

(b) Is server A significantly faster?

State the null and alternative hypotheses, use the data to select the most appropriate procedure, and conduct the test. What assumptions are we making for this test to be valid?

SOLUTION. *In this problem,* $n = 30$, $m = 20$, $\bar{X} = 6.7$, $\bar{Y} = 7.5$, $s_X = 0.6$, *and* $s_Y = 1.2$.

(a) To see if there is any significant difference between servers A and B, test

$$H_0 : \mu_A = \mu_B \quad \text{vs} \quad H_A : \underline{\mu_A \neq \mu_B}. \tag{1}$$

Standard deviations are unknown; they are estimated from the data. Should we use the pooled standard deviation or the Satterthwaite approximation method? It depends on the equality of population variances. So, test

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{vs} \quad H_A : \sigma_X^2 \neq \sigma_Y^2.$$

The F-statistic is

$$F_{\text{obs}} = \frac{s_Y^2}{s_X^2} = 4,$$

and the corresponding P-value for this two-sided test is

$$P = 2\left(\boldsymbol{P}\{F > 4\}\right) < 0.002,$$

using F-distribution with $n-1=29$ and $m-1=19$ d.f. (F-table with 29 and 19 d.f. gives $P(F > F_{obs}) < 0.001$).

Thus, there is an evidence that variances are unequal, and we should use the Satterthwaite approximation for the two-sample t-test (1).

The test statistic for test (1) is

$$t = \frac{6.7 - 7.5}{\sqrt{\frac{(0.6)^2}{30} + \frac{(1.2)^2}{20}}} = -2.7603.$$

It must be compared against critical values from the T-table. What degrees of freedom should we use? Satterthwaite approximation gives

$$\nu = \frac{\left(\frac{(0.6)^2}{30} + \frac{(1.2)^2}{20}\right)^2}{\frac{(0.6)^4}{30^2(29)} + \frac{(1.2)^4}{20^2(19)}} = 25.4 \text{ degrees of freedom.}$$

Then, the P-value for this two-sided test is

$$P = 2\boldsymbol{P}\{t > |-2.7603|\} \text{ is between 0.01 and 0.02 .}$$

We conclude that there is a significant difference between servers A and B at a level of 2% or higher, and the difference is not significant at a level of 1% or lower.

(b) *A faster server should have a shorter execution time. Thus we test $H_0 : \mu_A = \mu_B$ vs $H_A : \mu_A < \mu_B$. For this one-sided test, the P-value equals*

$$P = \boldsymbol{P}\{t < -2.7603\} \text{ is between 0.005 and 0.01.}$$

This is rather significant. At a 1% level of significance and any level above that, we have a significant evidence that server A is faster than server B.

4. **Micro-project**. Data on 522 recent home sales are available on our Blackboard web site The following variables are included.

Column	Variable
1	Identification number 1–522
2	Sales price of residence (×\$1000 dollars)
3	Finished area of residence (square feet)
4	Total number of bedrooms in residence
5	Total number of bathrooms in residence
6	Air conditioning: present or absent
7	Number of cars that garage will hold
8	Swimming pool: present or absent
9	Year property was originally constructed
10	Quality of construction: high, medium, or low
11	Architectural style. Three styles are coded as 1, 2, and 3
12	Lot size (square feet)
13	Location near a highway: yes or no

Use software to find out if there is significant evidence that:

(a) The sales price depends on the air conditioner in the house.

(b) On the average, homes with an air conditioner are more expensive.

(c) On the average, homes with an air conditioner are larger.

(d) The sales price depends on the proximity to a highway.

(e) On the average, homes are cheaper when they are close to a highway.

(f) On the average, homes are cheaper when they are far from a highway.

SOLUTION.

(a) *There is significant evidence that the sales price depends on the air conditioner in the house.* $t = 10.304$, $p < 2.2 \cdot 10^{-16}$.

(b) *There is significant evidence that on the average, homes with an air conditioner are more expensive.* $t = 10.304$, $p < 2.2 \cdot 10^{-16}$.

(c) *There is significant evidence that on the average, homes with an air conditioner are larger.* $t = 7.756$, $p = 4.817 \cdot 10^{-13}$.

(d) *There is no significant evidence that the sales price depends on the proximity to a highway at any significance level* $\alpha < 0.0901$. $t = -1.8552$, $p = 0.0901$.

(e) *Borderline case. There is significant evidence that on the average, homes are cheaper when they are close to a highway, at any significance level* $\alpha > 0.045$. *There is no significant evidence that on the average, homes are cheaper when they are close to a highway, at any significance level* $\alpha < 0.045$. $t = -1.8552$, $p = 0.045$.

(f) *There is no significant evidence that on the average, homes are cheaper when they are far from a highway.* $t = -1.8552$, $p = 0.9549$.

R code

```
> H = read.csv("HOME_SALES.csv")
> names(H)
 [1] "ID"              "SALES_PRICE"     "FINISHED_AREA"   "BEDROOMS"
 [5] "BATHROOMS"       "GARAGE_SIZE"     "YEAR_BUILT"      "STYLE"
 [9] "LOT_SIZE"        "AIR_CONDITIONER" "POOL"            "QUALITY"
[13] "HIGHWAY"
> attach(H)
> table(AIR_CONDITIONER)
AIR_CONDITIONER
 NO YES
 88 434
> t.test(x=SALES_PRICE[AIR_CONDITIONER=="YES"], y=SALES_PRICE[AIR_CONDITIONER=="NO"])
   t = 10.304, df = 241.5, p-value < 2.2e-16
> t.test(x=SALES_PRICE[AIR_CONDITIONER=="YES"], y=SALES_PRICE[AIR_CONDITIONER=="NO"],alternative="greater")
   t = 10.304, df = 241.5, p-value < 2.2e-16
> t.test(x=FINISHED_AREA[AIR_CONDITIONER=="YES"], y=FINISHED_AREA[AIR_CONDITIONER=="NO"],alternative="greater")
   t = 7.756, df = 160.14, p-value = 4.817e-13
> table(HIGHWAY)
HIGHWAY
 NO YES
511  11
> t.test(x=SALES_PRICE[HIGHWAY=="YES"], y=SALES_PRICE[HIGHWAY=="NO"])
   t = -1.8552, df = 11.178, p-value = 0.09011
> t.test(x=SALES_PRICE[HIGHWAY=="YES"], y=SALES_PRICE[HIGHWAY=="NO"], alternative="less")
   t = -1.8552, df = 11.178, p-value = 0.04506
> t.test(x=SALES_PRICE[HIGHWAY=="YES"], y=SALES_PRICE[HIGHWAY=="NO"], alternative="greater")
   t = -1.8552, df = 11.178, p-value = 0.9549
```