Lab12: Sample Answer Simple Hypothesis Testing

Handout date: Wednesday, November 18, 2020

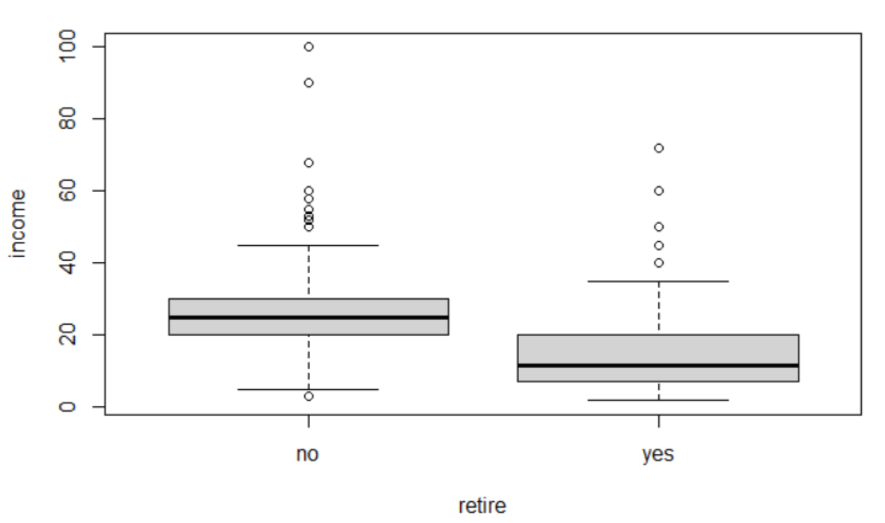
Due date: Wednesday, November 25, 2020 at the beginning of class

This lab counts 4 % toward your total grade

Task 1: Independent Sample Comparison of Means (1 point)

For the **Concord** data-frame in the workspace **TRAFFIC.RDATA** perform the following analyses:

[a] In a side-by-side boxplot compare the income for retired and non-retired households. Formulate a proper set of one-sided null and alternative hypotheses for the difference in means.



$$H_0: \mu_{retired} \geq \mu_{non-retired}$$
$$H_1: \mu_{retired} < \mu_{non-retired}$$

[b] Test for the equality of the variances in both groups.

```
var.test(income~retire, data = Concord)
      F test to compare two variances
data:  income by retire
F = 1.1098, num df = 349, denom df = 145, p-value = 0.4715
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8365818 1.4491927
sample estimates:
ratio of variances
          1.109794
```

The ratio of variances falls into the 95 percent confidence interval, so we could not reject the non-hypothesis, which means their variance are equal.

[c] Based on the outcome in [b] perform an independent samples *t*-test and properly interpret the results with respect to your formulated hypotheses.

```
t.test(income~retire,data = Concord,var.equal=TRUE,alternative = "greater")
      Two Sample t-test

data:  income by retire
t = 9.1465, df = 494, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 8.92994      Inf
sample estimates:
 mean in group no mean in group yes
         26.28286          15.39041

The p-value is much smaller than 0.05, so we could safely reject the non-
hypothesis, which means the average income of non-retired people is higher
than retired.
```

[d] Run a the regression model **`summary(lm(income~retire, data=Concord))`** and interpret the estimated intercept and slope parameters with respect to the mean income levels of retired and non-retired households.

```
summary(lm(income~retire, data=Concord))
Call:
lm(formula = income ~ retire, data = Concord)
Residuals:
    Min      1Q  Median      3Q     Max
-23.283  -8.283  -1.283   4.610  73.717
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.2829     0.6461  40.679   <2e-16 ***
retireyes   -10.8924     1.1909  -9.147   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 12.09 on 494 degrees of freedom
Multiple R-squared:  0.1448,   Adjusted R-squared:  0.1431
F-statistic: 83.66 on 1 and 494 DF,  p-value: < 2.2e-16
```

Intercept means the average income of non-retired folks is $26,283, the slope indicates the average income of retired households is 10.89 × $1,000 less than non-retired, that is, their income is around $15,390.

Task 2: Test for Differences in Traffic Volumes (2 points)

The workspace **Traffic.RData** holds two data-frames with identical measurements on the traffic volume (number of vehicles per day in thousands) in all four directions at 11 intersections in downtown Akron, Ohio, for the first Saturday in June in either 1970 or 1980.

It is generally assumed that the downtown traffic volume has decreased from 1970 to 1980. This is due to the opening of new retail locations in the suburbs, which pull the traffic away from the central city into the suburbs. We need to ignore other factors such that the overall traffic volume may have changed within the decade from 1970 to 1980 because we lack that information.

<u>Hint:</u> Make sure to study the online help of the test functions so you specify your hypothesis in the correct order.

[a] Test whether the variances of the traffic volume in 1970 is identical to the variance of the traffic volume in 1980. The function is **`var.test( )`**. Use an error probability of $\alpha = 0.05$.

```
var.test(Traffic~Year,data = TrafficByYear)
      F test to compare two variances

data:  Traffic by Year
F = 1.0503, num df = 43, denom df = 43, p-value = 0.873
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5730734 1.9247938
sample estimates:
ratio of variances
          1.050261
```

The p-value being substantially larger than $\alpha = 0.05$ indicates we fail to reject non-hypothesis, which means their variance are equal.

[b] Perform a ***matched pairs difference-of-mean test*** the null hypothesis $H_0 : \mu_{80} - \mu_{70} \geq 0$ against the alternative hypothesis $H_A : \mu_{80} - \mu_{70} < 0$ and ***interpret*** the results. Use an error probability of $\alpha = 0.05$.

Why are you performing a one-sided test?

```
traffic70 <- TrafficByYear$Traffic[TrafficByYear$Year==1970]
traffic80 <- TrafficByYear$Traffic[TrafficByYear$Year==1980]
t.test(traffic80, traffic70,alternative='less',paired=TRUE)

      Paired t-test

data:  traffic80 and traffic70
t = -4.0999, df = 43, p-value = 9.002e-05
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.9894095
sample estimates:
mean of the differences
              -1.677045
```

Since the p-value are much smaller than $\alpha = 0.05$, we could safely reject the null-hypothesis, which means $\mu_{80} < \mu_{70}$.

You want to show that the traffic volume has dropped within the decade between 1970 and 1980.

[c] Perform a ***two independent samples difference-of-means test*** for $H_0 : \mu_{80} \geq \mu_{70}$ against $H_A : \mu_{80} < \mu_{70}$ and ***interpret*** the results of this test. Use an error probability of $\alpha = 0.05$.

Decide based on task [a] under which assumption you are performing the test.

```
t.test(Traffic~Year,data = TrafficByYear,alternative='less')
      Welch Two Sample t-test

data:  Traffic by Year
t = 1.2023, df = 85.948, p-value = 0.8837
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 3.996462
sample estimates:
mean in group 1970 mean in group 1980 
          13.07727           11.40023 
```

P-value is larger than $\alpha = 0.05$, so we fail to reject the null-hypothesis.

[d] Is the ***matched pairs*** test scenario or the ***two independent samples*** test scenario more appropriate for the given data? ***Justify*** your choice.

Why do the test outcomes differ and which test is more powerful for the particular scenario?

The matched test design is substantially more powerful because at each intersection the baseline traffic flow is controlled for, thus focusing precisely on the change of flow at each intersection. In the independent samples design the comparison is whether the traffic flow after averaging over all intersections has changed.

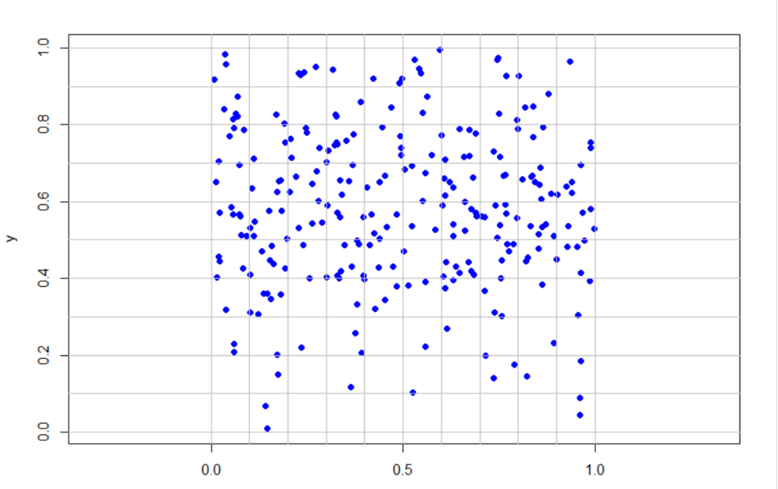
As explained above, since the matched pairs testing use additional information, so the result is quite different.

Task 4: Quadrat Analysis (1 point)

Pages 401-402 and 537-541 in Burt, Barber and Rigby discuss the quadrat analysis for spatial point patterns. The spatial pattern of trees is available in the file `trees.txt` and is processed with the code `trees.R`.

[a] Interpret the observed point pattern. Are the trees spatially cluster, random or disperse (uniformly) distributed?

From the first look, those points are likely randomly distributed.



[b] What does the ***variance-mean ratio*** tell you?

$VMR = \frac{Var(X)}{EX(X)} = 1.904296$, which indicates there may have cluster patterns.

[c] What does the *t*-test tell you about whether the observed point pattern comes from a random point pattern population?

The p-value is much smaller than 0.05, which means we could safely reject the null-hypothesis, the VMR is significantly greater than 1, cluster patterns are discovered.

[d] Use R or manually perform a χ^2-test to evaluate whether the observed point pattern is Poisson distributed. The tail counts have been properly aggregated for low expected frequency counts. Use the correct degrees of freedom. You can evaluate the significance of the χ^2-statistics with the function **`dchisq( )`**.

# of Observed Points j	# of Cells E_j	Expected Poisson Cell Count E_j	$\frac{(O_j - E_j)^2}{E_j}$
0	21	7.65	23.3
1	18	19.67	0.14
2	16	25.28	3.40
3	14	21.65	2.70
4	12	13.91	0.26
5	8	7.15	0.10
6+	11	4.69	8.4
Sum	100	100	38.4

dchisq(38.4,df = 5)

1.451549e-07

The probability is quite small, which indicates the given point pattern is not Poisson distributed as it would be assumed under complete spatial randomness.