STOCKHOLM UNIVERSITY
Department of Statistics

Michael Carlson
Emma Pettersson
2019-10-15

Solutions to Plenary Exercises: Plenary Exercise 7

Basic Statistics for Economists, 15 ECTS, STE101

EXERCISE 1

There are two samples, one taken before and one after the police car. Assume that the speed of the cars are independent of each other.

The samples are not independent since it's the same car being measured at two different locations. The two samples show paired observations for the speed before (X) and the speed after (Y).

Assume that the difference $D_i = X_i - Y_i$ is a random variable with observations that are independent and identically normally distributed $D_i \sim N(\mu_D, \sigma_D)$.

The mean difference $\mu_D = \mu_X - \mu_Y$ and the variance σ_D^2 are unknown and are estimated with the sample mean $\bar{d}$ and sample variance s_d^2.

Hypotheses:

$$H_0: \mu_D = \mu_X - \mu_Y = 0$$

$$H_A: \mu_D = \mu_X - \mu_Y > 0$$

$$\alpha = 0.05$$

Test statistic:

$$t = \frac{\bar{d} - \mu_0}{s_d/\sqrt{n}} \sim t\ (d.f = n-1)$$

Decision rule: we reject the null hypothesis if $t_{obs} > t_c$.

The critical value:

$$t_c = t_{n-1;\alpha} = t_{11;0.05} = [table\ 3] = \mathbf{1.796}$$

Calculations and the observed value of the test statistic:

Car no.	1	2	3	4	5	6	7	8	9	10	11	12	**Sum**
$\boldsymbol{x_i}$	83	88	106	131	84	87	129	97	92	91	124	99	
$\boldsymbol{y_i}$	61	84	95	121	83	79	92	88	69	90	115	100	
$\boldsymbol{d_i}$	22	4	11	10	1	8	37	9	23	1	9	-1	134
$\boldsymbol{d_i^2}$	484	16	121	100	1	64	1369	81	529	1	81	1	2848

Mean:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{134}{12} = 11.167$$

Variance

$$s_d^2 = \frac{\sum_{i=1}^{n} d_i^2 - n\bar{d}^2}{n-1} = \frac{2848 - 12(11.167^2)}{11} = 122.879$$

$$t_{obs} = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{11.167}{\sqrt{122.879/12}} = \mathbf{3.49}$$

Conclusion:

We reject the null hypothesis since $t_{obs} > t_c$. The observed average speed after seeing the police car is significantly less than the average speed before the police car.

EXERCISE 2

Let X represent car model 1 and Y represent car model 2.

Assumptions:
Two independent samples, each one with $n_x = n_y = 64$ independent observations from unknown distributions with means μ_X and μ_y and variances σ_X^2 and σ_Y^2 respectively.

Because both sample are large ($n_X = n_Y > 30$), according to the CLT, $\bar{X}$ and $\bar{Y}$ are approximately normally distributed. $\bar{X} \sim N(\mu_X, \sigma_X^2/n_X)$ and $\bar{Y} \sim (\mu_Y, \sigma_Y^2/n_Y)$.

Note that the text doesn't specify whether the variances given are the true population variances (σ_X^2 and σ_Y^2) or if they are the sample variances (S_X^2 and S_Y^2). However, since the samples are both large, we can use a z test, not knowing whether the variance is known or unknown isn't a problem.

Hypotheses:

$$H_0: \mu_X - \mu_Y = 0$$

$$H_A: \mu_X - \mu_Y \neq 0$$

$$\alpha = 0.05$$

Test variable:

$$Z = \frac{\bar{X} - \bar{Y} - D_0}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \sim N(0,1)$$

Decision rule:
We reject the null hypothesis if $|z_{obs}| > z_{\alpha/2}$.

The critical value:

$$z_{\alpha/2} = [table\ 2] = 1.96$$

The observed value:

$$z_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} = \frac{29.5 - 27.25}{\sqrt{\frac{6.37}{64} + \frac{5.44}{64}}} = 5.2378$$

Since $z_{obs} > z_{\alpha/2}$, we reject the null hypothesis. The observed average difference in the breaking power of cars 1 and 2 is significant.

We can also calculate the p-value:

$$\text{p} - \text{value} = 2\,P(Z > |z_{obs}|) = 2\,P(Z > 5.2378) = 2\big(\,1 - P(Z \leq 5.2378)\big)$$
$$= [\text{from excel using function NORM. S. DIST}(5.2378, \text{TRUE})]$$
$$= 2(1 - (0.999999918749)) = 0.000000162502$$

We can also solve this problem in a more conservative way using a t-test:

Assumptions:

$X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim (\mu_Y, \sigma_Y^2)$. Also assume that $\sigma_Y^2 = \sigma_X^2$; we have two estimates for the same variance, so we pool s_x^2 and s_y^2 to estimate s_p^2.

The test variable:

$$t_{126} = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t(df = n_x + n_y - 2)$$

Decision rule:
We reject the null hypothesis if $|t_{obs}| > t_c$.

The critical value:

$$t_c = t_{n_x + n_y - 2; \alpha/2} = t_{126;0.025} = [table\ 3] = 1.979$$

The observed value:

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \frac{(63)(6.37 + 5.44)}{126} = \frac{6.37 + 5.44}{2} = 5.905$$

$$t_{obs} = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = \frac{29.5 - 27.25}{\sqrt{5.905}\sqrt{2/64}} = 5.2378$$

Conclusion:

We reject the null hypothesis since $t_{obs} > t_c$. Observe that we got the same value for t_{obs} and z_{obs}. This is because the size of the both samples were equal, if they had been different we would have gotten different values. The only difference is that the critical value of the t-test is larger, the t-test is more conservative as it makes it more difficult to reject the null hypothesis.

The p-value becomes larger (using the function =2*T.DIST.RT(5.2378,126) in Excel):

$$p - value = 2\,P(t > |t_{obs}|) = 2(t > 5.2378) = 0.0000006614$$

EXERCISE 3

The proportions we want to test:

Car	$P_C = \frac{1}{2}$
Public Transport	$P_P = \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$
Bicycle	$P_B = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$

Goodness of fit test:

Hypotheses:

$H_0: P_C = \frac{1}{2}, P_P = \frac{1}{3}, P_B = \frac{1}{6}$

H_A: The null hypothesis distriubtion does not hold, at least one of the statements above is false.

$\alpha = 0.05$

Test statistic:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-1}$$

Decision rule:
We reject the null hypothesis if $\chi^2_{obs} > \chi^2_{K-1;\alpha}$.

The critical value:

$$\chi^2_{K-1;\alpha} = \chi^2_{2;0.05} = 5.991$$

The observed value:

i	Car	Public Transport	Bicycle	Sum
P_i	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	1
$E_i = nP_i$	25	16.667	8.333	50
O_i	22	14	14	50
$(O_i - E_i)$	-3	-2.6667	5.66667	0
$(O_i - E_i)^2$	9	7.11111	32.1111	
$\frac{(O_i - E_i)^2}{E_i}$	0.36	0.42667	3.85333	$\boldsymbol{\chi^2_{obs} = 4.64}$

Conclusion:

We cannot reject the null hypothesis $\chi^2_{2;0.05} > \chi^2_{obs}$. The observed frequencies are not significantly different from the hypothesized distribution.

EXERCISE 4

Hypotheses:

$$H_0\text{: The data collection method and non response are } \textbf{independent}$$

$$H_A\text{: The data collection method and non response are } \textbf{dependent}$$

Test statistic:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \sim\chi^2(df=(r-1)(c-1))$$

$E_{ij} = \frac{R_i C_j}{n}$, Where R_i is the row sum and C_j is the column sum. r is the number of rows, and c is the number of columns.

We reject the null hypothesis if $\chi^2_{obs} > \chi^2_{2;0.05}$.

The critical value:

$$\chi^2_{2;0.05} = 5.991$$

The observed value:

$\boldsymbol{O_{ij}}$	MM	CATI	PAPI	Sum
Response	264	297	320	881
Non-Response	110	72	145	327
Sum	374	369	465	1208

$\boldsymbol{E_{ij}}$	MM	CATI	PAPI	Sum
Response	272.7599	269.1134	339.1267	881
Non-Response	101.2401	99.88659	125.8733	327
Sum	374	369	465	1208

$\boldsymbol{\left(O_{ij}-E_{ij}\right)^2/E_{ij}}$	MM	CATI	PAPI	Sum
Response	0.281333	2.889718	1.078738	4.24979
Non-Response	0.757965	7.785448	2.906326	11.44974
				15.69953

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} = 15.69953$$

Conclusion:

We reject the null hypothesis on the 5% significance level since the observed value is larger than the critical value. The level of non-response is dependent on the interview type.

EXERCISE 5

Hypotheses:

$$H_0\text{: App experience and phone brand are independent}$$

$$H_A\text{: App experience and phone brand are dependent}$$

$$\alpha = 0.05$$

Test statistic:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2(df = (r-1)(c-1))$$

where $E_{ij} = \frac{R_i C_j}{n}$, R_i is the row sum and C_j is the column sum. r is the number of rows and c is the number of columns.

Decision rule:
Reject H_0 if $\chi^2_{obs} > \chi^2_c$

Critical value:

$$\boldsymbol{\chi^2_c} = \chi^2_{(r-1)(c-1),\alpha} = \chi^2_{1,0.05} = \mathbf{3.841}$$

Observed value:

O_{ij}		Has iPhone Yes	Has iPhone No	Σ
Likes the app	Yes	80	25	105
	No	80	15	95
Σ		160	40	200

E_{ij}		Has iPhone Yes	Has iPhone No	Σ
Likes the app	Yes	84	21	105
	No	76	19	95
Σ		160	40	200

$\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$		Has iPhone Yes	Has iPhone No	
Likes the app	Yes	0.19048	0.76191	
	No	0.21053	0.84211	$\Sigma = \boldsymbol{\chi^2_{obs}} = \mathbf{2.005}$

Since $\chi^2_{obs} < \chi^2_c$ we cannot reject H_0 on the 5% significance level. The app experience and the phone brand are independent.

EXERCISE 6

a. The first step is to calculate the moving averages. Because we have quarterly data ($s = 4$), we calculate 4-point moving averages. Then we calculate the average of two adjacent averages.
For example, for $t = 3$ we first calculate $x_{2.5}^*$ and $x_{3.5}^*$, then x_3^*.

$$x_{2.5}^* = \frac{x_1 + x_2 + x_3 + x_4}{4} = \frac{16913.6}{4} = 4228.4$$

$$x_{3.5} = \frac{x_2 + x_3 + x_4 + x_5}{4} = \frac{16988.3}{4} = 4247.1$$

$$x_3^* = \frac{x_{2.5}^* + x_{3.5}^*}{2} = \frac{\frac{x_1 + X_2 + x_3 + x_4}{4} + \frac{x_2 + x_3 + x_4 + x_5}{4}}{2} = \frac{x_1}{8} + \frac{x_2}{4} + \frac{x_3}{4} + \frac{x_4}{4} + \frac{x_5}{8}$$
$$= 4237.738$$

We do this for all observations except the first two ($t = 1$ and 2) and last two in the series ($t = 11$ and 12), those are instead marked with an '$*$' symbol in the table below. In total you should calculate 12 moving averages from x_3^* to x_{14}^*.

After that, the next step is the calculate the ratio between the observed values and the moving averages. We want to see if a seasonal effect causes a consistent percentage increase or decrease in the different quarters.

$$q_t = 100 \cdot \frac{x_t}{x_t^*} \quad \text{for } t = 3, \ldots 10$$

Year/quarter	t	x_t	x_t^*	q_t
2014Q1	1	4 119.8	*	*
2014Q2	2	4 239.7	*	*
2014Q3	3	4 329.6	4 237.7375	102.1677
2014Q4	4	4 224.5	4 254.2000	99.3019
2015Q1	5	4 194.5	4 267.4375	98.2908
2015Q2	6	4 296.7	4 285.0500	100.2719
2015Q3	7	4 378.5	4 307.8375	101.6403
2015Q4	8	4 316.5	4 333.5750	99.6060
2016Q1	9	4 284.8	4 356.1375	98.3624
2016Q2	10	4 412.3	4 373.3375	100.8909
2016Q3	11	4 443.4	*	*
2016Q4	12	4 389.2	*	*

In the next step. we group the ratios (q_t) according to quarter and calculate the median ratio for each quarter. Then, to guarantee that the mean of the seasonal index will be 100%, we need to make adjustments so that the de-seasoned values aren't too large or small. The sum of the medians was 400.00266. so we adjust using:

$$\text{Index}_\text{Q} = \text{Median} \cdot \frac{400}{400.00266}$$

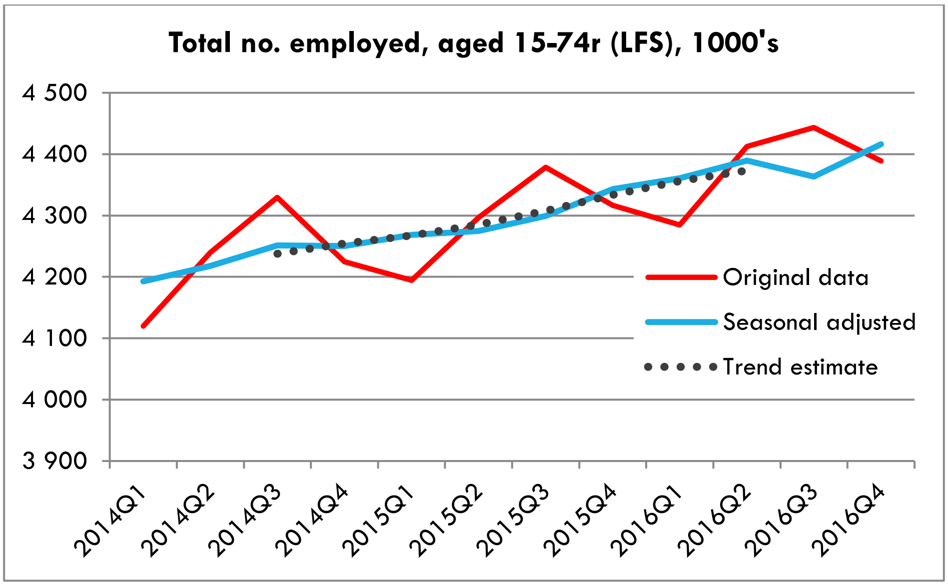
	2014	2015	2016	Median	Index
Q1	*	0.98291	0.98362	0.98327	0.98261
Q2	*	1.00272	1.00891	1.00581	1.00515
Q3	1.02168	1.01640	*	1.01904	1.01836
Q4	0.99302	0.99606	*	0.99454	0.99388
			Sum	4.00266	4

The last step is to calculate the de-seasoned values x_t^{adj} by multiplying the original x_t with 100 and dividing by the seasonal index.

$$x_t^{adj} = x_t \times \frac{100}{\text{Index}_\text{Q}}$$

Year/Quarter	t	x_t	Index_Q	De-seasoned
2014Q1	5	4119.8	98.261	4 192.70
2014Q2	6	4239.7	100.515	4 218.00
2014Q3	7	4329.6	101.836	4 251.53
2014Q4	8	4224.5	99.388	4 250.52
2015Q1	9	4194.5	98.261	4 268.72
2015Q2	10	4296.7	100.515	4 274.70
2015Q3	11	4378.5	101.836	4 299.55
2015Q4	12	4316.5	99.388	4 343.09
2016Q1	13	4284.8	98.261	4 360.62
2016Q2	14	4412.3	100.515	4 389.71
2016Q3	15	4443.4	101.836	4 363.28
2016Q4	16	4389.2	99.388	4 416.23

b. A time series diagram with both the original and the de-seasoned series.



We see that the seasonally adjusted series is almost a straight line with a positive trend but with small random fluctuations towards the end of the series where there is an unexpected decrease for quarter 3 in 2016.

Comments, what we have done:

First we assumed the series can be described using a multiplicative model:

$$X_t = T_t S_t I_t$$

The first step was to calculate a moving average that gives an estimated of the trend T_t for 12 time points $t = 3. \dots 14$

$$\hat{T}_t = x_t^* = \frac{x_{t-2}}{8} + \frac{x_{t-1}}{4} + \frac{x_t}{4} + \frac{x_{t+1}}{4} + \frac{x_{t+2}}{8}$$

By dividing the observed values with the moving average we are left with the seasonal and irregular parts.

$$\frac{x_t}{x_t^*} = \frac{x_t}{\hat{T}_t} = \frac{T_t S_t I_t}{\hat{T}_t} \approx S_t I_t$$

In the next step we gathered the values according to quarter and the median for each quarter was used as an estimate of the seasonal component $\hat{S}_Q$. We then adjusted these medians so the mean of the seasonal index is 1 (or 100%).

When the seasonal indices for the different quarters are ready we can then calculated the seasonally adjusted values that show the trend and irregular component.

$$\frac{x_t}{\hat{S}_Q} = \frac{T_t S_Q}{\hat{S}_Q} \approx T_t I_t$$