# Sample Answer Lab12: $\chi^2$-Test and Kernel Density

**Handout date:** Wednesday, November 20, 2019

**Due date:** Wednesday, December 4, 2019 at the beginning of the lecture as hardcopy

*This lab counts 4 % toward your total grade*

**Objectives:** In this lab you practice the $\chi^2$-test under different scenarios and generate spatial kernel density maps.

## Task 1: Impact of Cell Counts [1 point]

Below you find two $2 \times 3$ cross-tabulations of the variables "R" and "C". You can do the calculations with by entering these tables as matrices into .

| | C1 | C2 | C3 |
|---|---|---|---|
| R1 | 9 | 7 | 5 |
| R2 | 3 | 6 | 9 |

| | C1 | C2 | C3 |
|---|---|---|---|
| R1 | 18 | 14 | 10 |
| R2 | 6 | 12 | 18 |

***Table A*** ***Table B***

[a] For each table calculate the ***expected probabilities*** under the assumption independence between both variables' "R" and "C". Round the probabilities up to 3 significant digits.

| | C1 | C2 | C3 |
|---|---|---|---|
| R1 | 0.166 | 0.179 | 0.193 |
| R2 | 0.142 | 0.154 | 0. 166 |

***Table A***

| | C1 | C2 | C3 |
|---|---|---|---|
| R1 | 0.166 | 0.179 | 0.193 |
| R2 | 0.142 | 0.154 | 0. 166 |

***Table B***

[b] For each table calculate their $\chi^2$-test statistics and their associated *p*-values.

```
(mx1 <- matrix (c (9,3,7,6,5,9), nrow=2, ncol=3))
(chi1 <- chisq.test(mx1))

X-squared = 4.0128, df = 2, p-value = 0.1345

(mx2 <- matrix(c(18,6,14,12,10,18), nrow=2, ncol=3))
( chi1 <- chisq.test(mx2) )

X-squared = 8.0255, df = 2, p-value = 0.01808
```

[c] Interpret the magnitude of the $\chi^2$-test statistics and their associated *p*-values in the light that the counts in Table *B* are just twice as large as in Table *A*.

Given identical cross-tabulation probabilities, the $\chi^2$-test statistics is sensitive to the sample size. As the sample size doubles so does the $\chi^2$-test statistics. Thus, while the degrees of freedom remain constant, a $\chi^2$-test statistics for larger samples will become significant and rejects the independence null hypothesis

## Task 2: Violation of Assumptions and Recovery [1 point]

An instructor of an undergraduate statistics class is interested if there is a gender bias in the grade distribution in her class. The final grades by gender are

| | ***A*** | ***B*** | ***C*** | ***D*** | ***F*** |
|---|---|---|---|---|---|
| ***female*** | 6 | 17 | 10 | 3 | 2 |
| ***male*** | 7 | 12 | 6 | 2 | 2 |

You may enter this table as matrix into .

[a] Specify a proper ***null hypothesis*** and its associated alternative hypothesis.

$H_0$: Grades is nothing related with gender. $H_0: \pi_{grade|male} = \pi_{grade|female} \quad \forall$ grades

$H_1$: Grades would vary from gender. $H_1: \pi_{grade|male} \neq \pi_{grade|female} \quad$ for at least one grade

[b] Perform the $\chi^2$-test at the error probability $\alpha = 0.05$.

```
(mx3 <- matrix(c(6,17,10,3,2,7,12,6,2,2), nrow=5, ncol=2))
(chi3 <- chisq.test(mx3) )

X-squared = 0.94713, df = 4, p-value = 0.9177

Tips: usually we put class as the row and compared datasets as column.
```

[c] What assumptions for the $\chi^2$-test are not satisfied.

Because p-value is larger than 0.05, so we fail to reject the null hypothesis.

Expected:

```
        Female      Male
[1,]  7.373134  5.626866
[2,] 16.447761 12.552239
[3,]  9.074627  6.925373
[4,]  2.835821  2.164179
[5,]  2.268657  1.731343
```

[d] Explain how you could change the cross-tabulation to bring it into agreement with the assumptions. Re-enter the modified cross-tabulation.

$\chi^2$-test require expected cell frequencies are larger than 5, so we could aggregate class C D F together as a class with poor grade, so the table would as follow:

| | Good grade | Fair grade | Poor grade |
|---|---|---|---|
| Female | 6 | 17 | 15 |

| | | | |
|---|---|---|---|
| Male | 7 | 12 | 10 |

```
(mx4 <- matrix(c(6,17,15,7,12,10), nrow=3, ncol=2))
( chi1 <- chisq.test(mx4) )

X-squared = 0.74345, df = 2, p-value = 0.6895
```

[e] Perform another $\chi^2$-test on the modified cross-tabulation and interpret its outcome with regards to the instructor's concerns about a gender bias in the course grades.

```
( chi1 <- chisq.test(mx4) )
X-squared = 0.74345, df = 2, p-value = 0.6895
```

Again, we fail to reject the null hypothesis.

## Task 3: Goodness of Fit $\chi^2$-Test [1 point]

In Task 1 of Lab10 you calculated the distribution of the grid-cell counts and evaluated the expected counts assuming the data would follow a Poisson-distribution.

[a] Formulate the null hypothesis and the alternative hypothesis whether the observed distribution of grid cell counts follows a Poisson distribution.

H0: The overserved data follow Poisson distribution (No variance exists)

H1: The overserved data do not follow Poisson distribution.

[b] Calculate the goodness of fit $\chi^2$-test statistic and its *p*-value. Make sure that the underlying assumptions are satisfied. Justify any aggregation of classes.

Raw:

| | Prob | Observed | Expected |
|---|---|---|---|
| 0 | 0.04978707 | 0 | 0.796593 |
| 1 | 0.1493612 | 4 | 2.389779 |
| 2 | 0.2240418 | 2 | 3.584669 |
| 3 | 0.2240418 | 5 | 3.584669 |
| 4 | 0.1680314 | 3 | 2.688502 |
| 5 | 0.1008188 | 0 | 1.613101 |
| 6 | 0.05040941 | 1 | 0.806551 |
| 7+ | 0.03350854 | 1 | 0.536137 |

Aggerated:

| | Observed | Expected |
|---|---|---|
| < = 2 | 6 | 6.77 |
| = 3 | 5 | 3.58 |

| | | |
|---|---|---|
| > = 4 | 5 | 5.64 |

The test statistic is $\chi^2 = \sum_{i=1}^{k} z_i^2$ with $z_i^2 = \frac{(f_i - F_i)^2}{F_i}$

$$\chi^2 = \frac{(6-6.77)^2}{6.77} + \frac{(5-3.58)^2}{3.58} + \frac{(5-5.64)^2}{5.64}$$
$$= 0.7234419$$

```
pchisq(q = 0.7234419,df = 1, lower.tail=FALSE)
```

```
0.3950172
```

[c] Explain what degrees of freedom you needed to use.

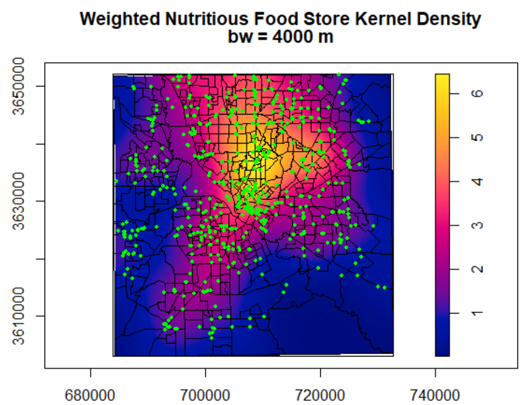
Degrees of freedom $= 3$ (for the number of classes) – 1 (for the estimated intensity $\lambda$) – 1 (so the observed number of counts match the expected number of counts)$= 1$

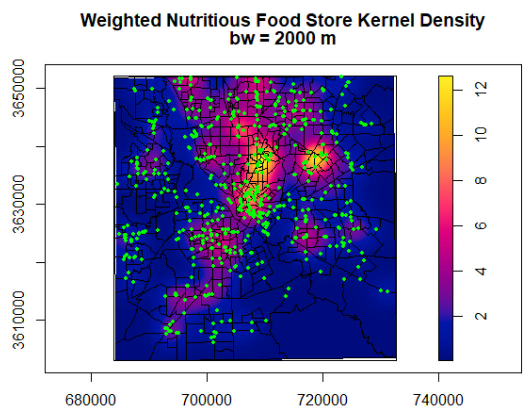
## Task 4: Spatial Kernel Density and Bandwidth [1 point]

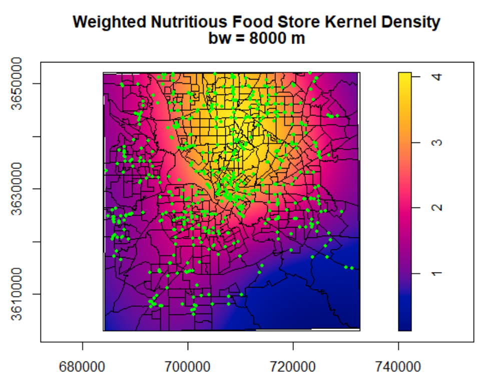
You already worked with kernel densities in the univariate setting in Task 4 of Lab03. Now you will apply kernel densities in a spatial setting. Go to the "User guides, package vignettes and other documentation" in the package **`DallasTracts`** and open the **`R-code`**. Copy the code into RStudio's editor and run the code up to line 148.

Generate and show the kernel density maps using the code in lines 142-148 with bandwidths of 2000, 4000 and 8000 meters. Hint: the bandwidth is set with the **`sigma`**-parameter.

Which is the most appropriate bandwidth describing the spatial coverage of the grocery stores in Dallas County?




The bandwidths with $\sigma = 4000$ meters is the best displays the spatial distribution of the grocery stores.

The bandwidths with $\sigma = 2000$ meters is too fragmented.

The bandwidths with $\sigma = 8000$ meters over-smooths the distribution of stores and makes it almost homogeneous throughout Dallas county.