



Percentage Comparison:

Hypothesis testing & Graphic Methods

Session 8

Programación Estadística con Python

Alberto Sanz, Ph.D
asanz@edem.es

- Hypothesis testing over the relationship of two nominal (categorical) variables.
 - Numeric methods:
 - Cross tabulations (Descriptive, sample level) +
 - Chi2 (Hypothesis testing, at the **population** level) +
 - Cramer's V (Strength of the association, at the **pop.** level)
(To be developed)
 - Graphic methods:
 - Grouped barplots.
 - Mosaic plots. (To be developed)

Our Dependent Variable

3

```
#Recoding DV for analysis
res = wbr.cnt.describe()
m = res[1]
sd = res[2]
n = res[0]

### Recode cnt to string
wbr.loc[(wbr['cnt']<(m-sd)), "cnt_str"] = "Low rentals"
wbr.loc[((wbr['cnt']>(m-sd)) & (wbr['cnt']<(m+sd))), "cnt_str"] = "Average rentals"
wbr.loc[(wbr['cnt']>(m+sd)), "cnt_str"] = "High rentals"
```

```
### Recode cnt to ordinal
my_categories = ["Low rentals", "Average rentals", "High rentals"]
my_rentals_type =
CategoricalDtype(categories=my_categories, ordered=True)
wbr["cnt_cat"] = wbr.cnt_str.astype(my_rentals_type)
wbr.info()
```

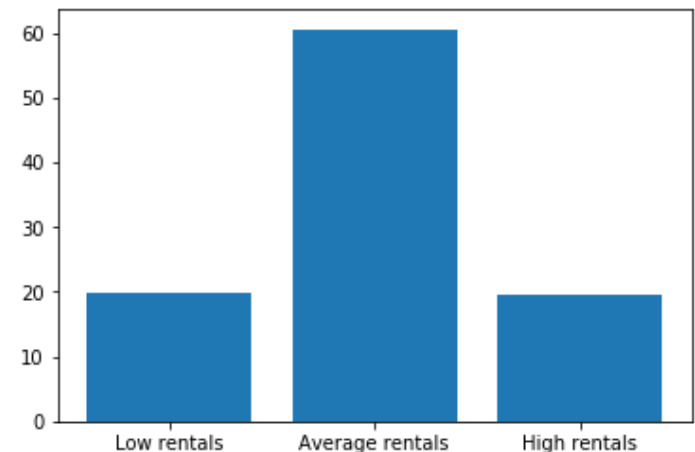
```
#Percentage table & barchart
mytable = pd.crosstab(wbr.cnt_cat, columns="count",
normalize='columns')*100

plt.bar(mytable.index, mytable['count'])
```

Table 1. Percentage of days with different rentals in Washington D.C.

Low rentals	19,8
Average rentals	60,6
High rentals	19,56
TOTAL	100,0
(n)=731	

Source: Own analyses over Fanaee, Hadi and Gama (2013) data



Percentage comparison (n groups)

4

1. Describe the two variables involved in the hypothesis separately.
Special attention to be paid at the distribution of the DV*
2. Describe the DV, by the levels in the IV
(Cross tabulation of DV by IV)
3. Perform the numeric test for **inference**: Chi² test
4. Graphic representation: Combined barplot
5. When possible, combine:
 1. Crosstabs + inference test (as footnote)
 2. Combined barplot + inference test (as text insert)

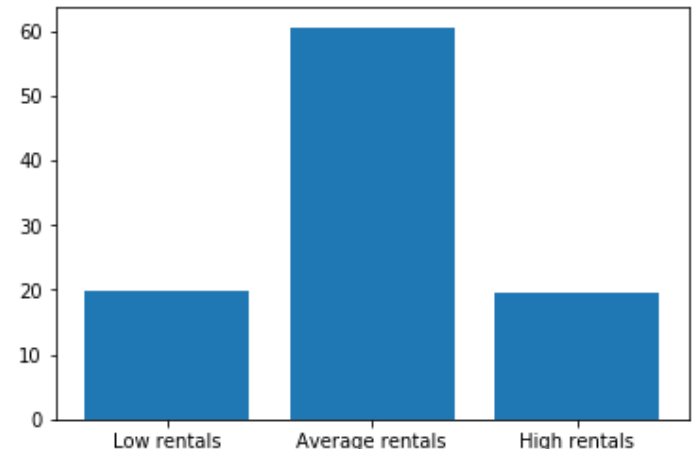
* DV stands for Dependent variable. IV stands for Independent Variable

Why some days are rent *more* bikes?

Table 1. Percentage of days with different rentals in Washington D.C.

Low rentals	19,8
Average rentals	60,6
High rentals	19,56
TOTAL	100,0
(n)=731	

Source: Own analyses over Fanaee, Hadi and Gama (2013) data



- H0.: Percentage of days with *low/average/high* rentals is the same in *working days* vs. *not working days*.
- H1.: Percentage of days with *low/average/high* rentals differs in *working days* vs. *not working days*.

Percentage comparison (n groups)

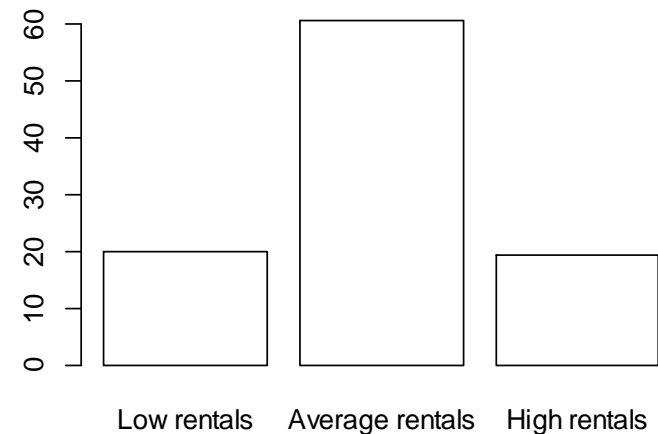
6

1. Describe the two variables involved in hypothesis

Working days



Rentals



Percentage comparison (n groups)

7

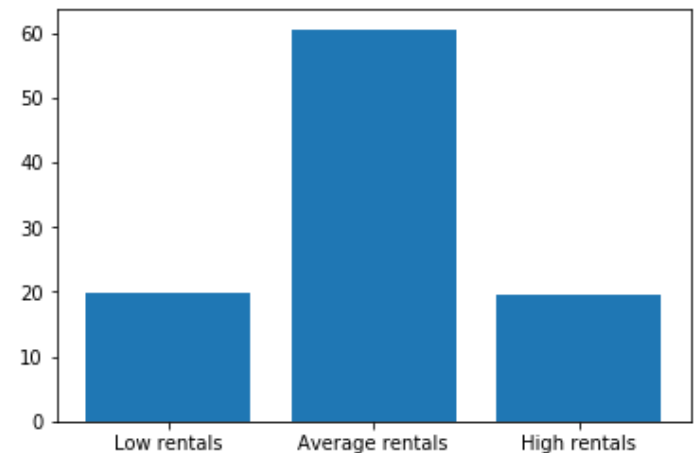
1. Special attention to the distribution of Dependent Variable

Rentals

Table 1. Percentage of days with different rentals in Washington D.C.

Low rentals	19,8
Average rentals	60,6
High rentals	19,56
TOTAL	100,0
(n)=731	

Source: Own analyses over Fanaee, Hadi and Gama (2013) data



Percentage comparison (n groups)

8

2. Describe the DV, by the factor levels in the IV (Cross tabulation of DV by IV)

```
pd.crosstab(wbr.cnt_cat, wbr.wd_cat, normalize='columns', margins=True)*100
```

	Non working days	working days	TOTAL
Low rentals	24.67532	17.63527	19.86301
Average rentals	57.14286	62.32465	60.68493
High rentals	18.18182	20.04008	19.45205
Sum	100.00000	100.00000	100.00000

Percentage comparison (n groups)

9

2. Describe the DV, by the factor levels in the IV (Cross tabulation of DV by IV)

Use DV as
reference

	Non working days	working days	TOTAL
Low rentals	24.67532	17.63527	19.86301
Average rentals	57.14286	62.32465	60.68493
High rentals	18.18182	20.04008	19.45205
Sum	100.00000	100.00000	100.00000

Percentage comparison (n groups)

10

2. Describe the DV, by the factor levels in the IV (Cross tabulation of DV by IV)

	Non working days	working days	TOTAL
Low rentals	24.67532	17.63527	19.86301
Average rentals	57.14286	62.32465	60.68493
High rentals	18.18182	20.04008	19.45205
Sum	100.00000	100.00000	100.00000

Percentage comparison (n groups)

11

2. Describe the DV, by the factor levels in the IV (Cross tabulation of DV by IV)

	Non working days	working days	TOTAL
Low rentals	24.67532	17.63527	19.86301
Average rentals	57.14286	62.32465	60.68493
High rentals	18.18182	20.04008	19.45205
Sum	100.00000	100.00000	100.00000

Answer: In the **sample** of 731 days, yes. In non working days it seems that we rent more bikes. BUT,... what about in the **population**? Answer: Still don't know.

Percentage comparison (n groups)

12

3. Perform the numeric test for inference: Chi² test

	Non working days	working days	TOTAL
Low rentals	24.67532	17.63527	19.86301
Average rentals	57.14286	62.32465	60.68493
High rentals	18.18182	20.04008	19.45205
Sum	100.00000	100.00000	100.00000

```
# We apply the stats.chi2_contingency()  
over the original  
crosstab containing FREQUENCIES  
  
ct= pd.crosstab(wbr.cnt_cat, wbr.wd_cat)  
stats.chi2_contingency(ct)
```

Output:

```
(4.9833225686178624,  
0.082772343895498146,  
2,
```

This is the P. Value

CONCLUSION:

As P. Val > 0.05, we do NOT REJECT H0.:

In other words:

**Percentage days with high/mid/low rentals
do not significantly differ in Working days
and Non working days.**

Percentage comparison (n groups)

Table 2. Rental levels in Washington, by type of day. (In percentage points)

	Non working days	Working days	All days
Low rentals	24,7	17,6	19,9
Average rentals	57,1	62,3	60,7
High rentals	18,2	20,0	19,5
TOTAL	100	100	100
(n)=731			

$\chi^2=4.983$, $p\text{-value} = 0.083$. Source: Own analyses over Fanaee, Hadi and Gama (2013) data.

Conclusion: As $P\text{ Value} > 0.05$ Do not reject H_0 .

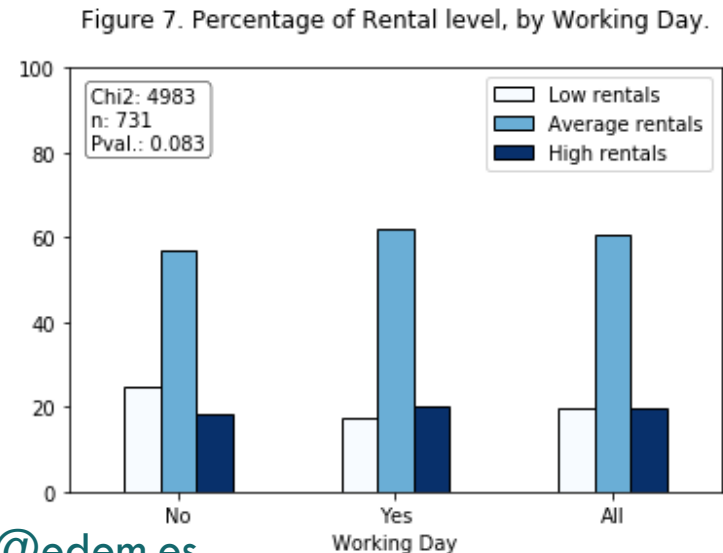
- ✓ ☐ H_0 .: Percentage of days with *low/average/high* rentals is the same in working days vs. not working days.
- ✗ ☐ H_1 .: Percentage of days with *low/average/high* rentals differs in working days vs. not working days.

Percentage comparison (n groups)

4. Graphic representation: Combined barplot

```
#Transpose and plot
my_ct2=my_ct.transpose()

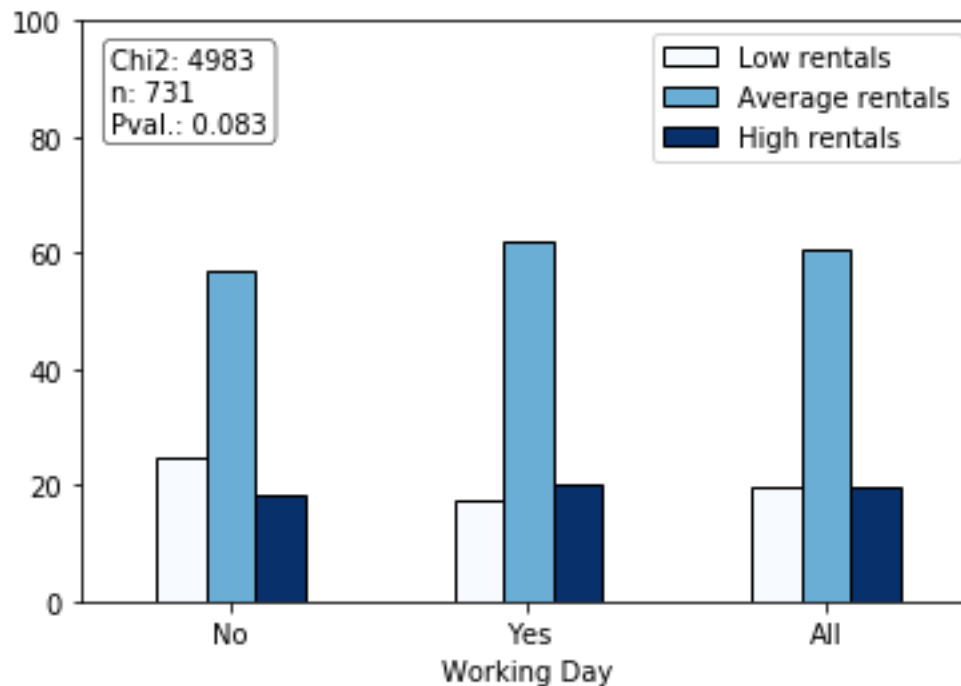
my_ct2.plot(kind="bar", edgecolor = "black", colormap='Blues')
props = dict(boxstyle='round', facecolor='white', lw=0.5)
plt.text(-0.4, 81, 'Chi2: 4983'\n'n: 731' '\n' 'Pval.: 0.083',      bbox=props)
plt.xlabel('Working Day')
plt.title('Figure 7. Percentage of Rental level, by Working Day.''\n')
plt.legend(['Low rentals', 'Average rentals', 'High rentals'])
plt.ylim(0,100)
plt.xticks(rotation='horizontal')
```



Percentage comparison (n groups)

4. Graphic representation: Combined barplot

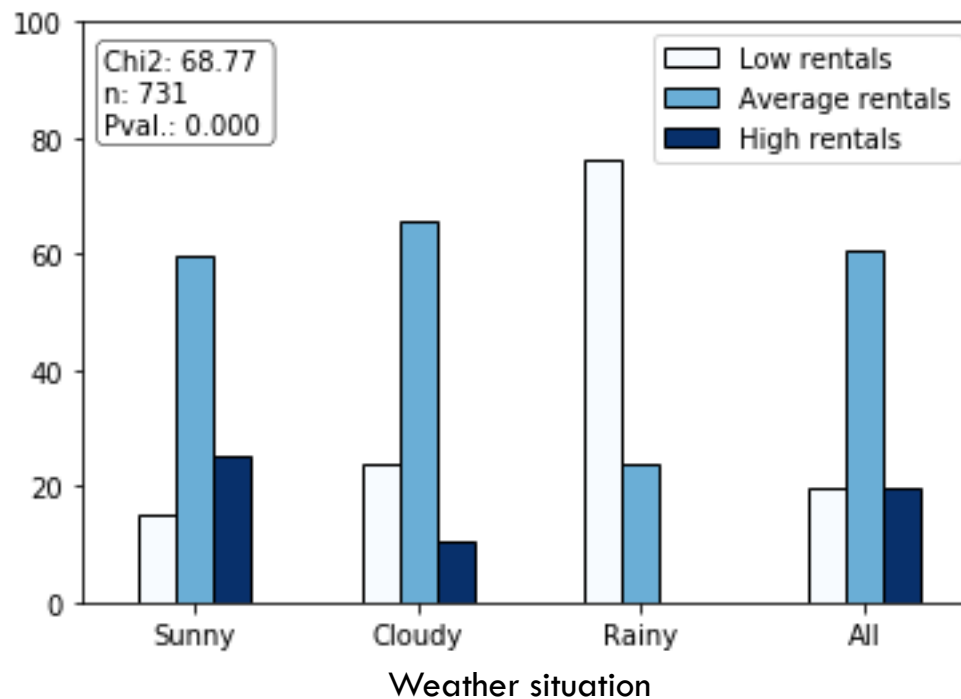
Figure 7. Percentage of Rental level, by Working Day.



Percentage comparison (n groups)

4. Graphic representation: Combined barplot (ex. II)

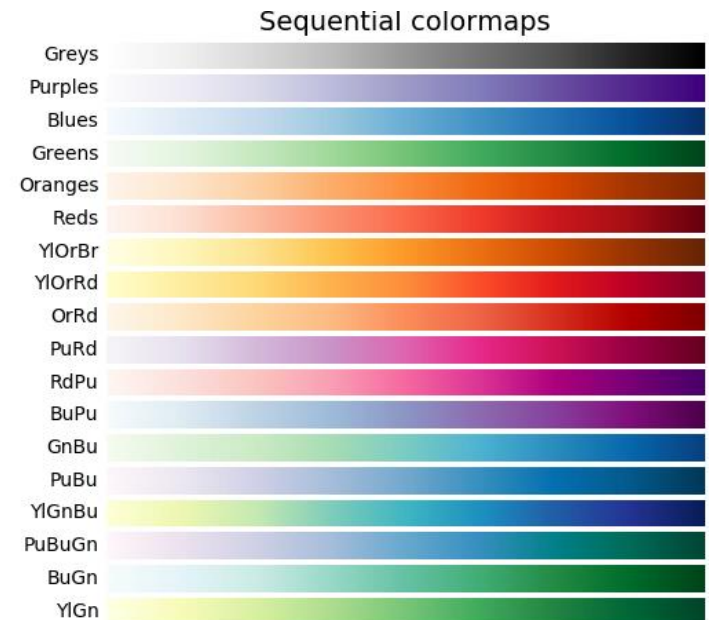
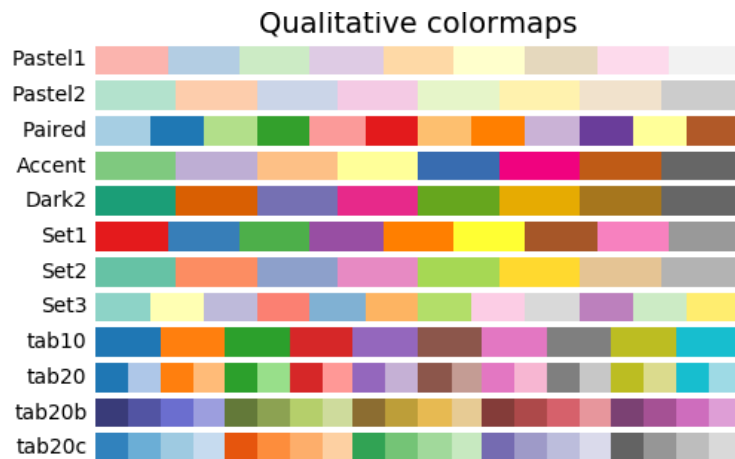
Figure 8. Percentage of Rental level,
by Weather situation



You may want to explore:

□ The matplotlib Colormaps

https://matplotlib.org/2.0.2/examples/color/colormaps_reference.html



- General Remainder:
 - Always **describe/explore your data** (numerically + graphically) prior to perform any statistical analysis.
- Main Numeric Procedure:
 - Crosstabulation with Column percentages
 - Chi2: test
- Main Graphic Procedure:
 - Combined Barplot

Questions?

Thank you !

Alberto Sanz
asanz@edem.es