ST2137 Midterms

by Zong Xun

Describing Numerical Data

A **dataframe** is an object with rows and columns:
- The rows contain different observations or measurements.
- The columns contain the values of different variables.

A **parameter** is a numerical summary of the population. It is unknown.

A **statistic** is a summary of the sample taken from the population.
- Descriptive: numerical or graphical representation e.g. variance, histograms.
- Inferential: hypothesis test, estimation, regressions.

Summarizing Data

The data can mostly be summarized with **location** and **variability** of the data. **Location** is a simple summary of the data. **Variability** is mean squared deviation from the mean.

Generally,
- For a dataset, when the mean is the same (or approximately the same) as median, the data is said to be close to symmetric.
- Mean is sensitive to outlier while median is not.
- When the mean is much larger than the median, data is positively skewed, vice versa.

Skewness

If the distribution is unimodal, we should also check if the distribution is **skewed** or **symmetric**. We say that the distribution is positively skewed if the right tail is longer, and negatively skewed if the left tail is longer.

Given a sample size of n, the sample skewness is

$$\frac{\sqrt{n(n-1)}}{n-2} \times \frac{m_3}{m_2^{3/2}}$$

where

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ and } m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

- If skewness = 0, then the distn is perfectly symmetrical.
- If skewness is greater than $|1|$, then the distn is highly skewed.
- If skewness is $> |\frac{1}{2}|$ and $< |1|$, then the distn is moderately skewed.
- If skewness is between $-\frac{1}{2}$ and $\frac{1}{2}$, then the distn is approximately symmetric.

Kurtosis

Higher values of kurtosis indicate a higher, sharper peak; lower values indicate a lower, less distinct peak.

Given a sample of size n, the sample kurtosis is:

$$\frac{n-1}{(n-2)(n-3)}[\frac{(n+1)m_4}{m_2^2} - 3(n-1)]$$

where

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ and } m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Association between two variables

An association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variables.

Quantifying with correlation

Let X and Y be variables from a set with n objects. The correlation between these two variables is defined as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{X_i - \bar{X}}{s_X})(\frac{Y_i - \bar{Y}}{s_Y})$$

where $\bar{X}, \bar{Y}$ are the sample means. s_X and s_Y are the sample standard deviations of the two variables.
- Note that the two variables have the **same correlation**, no matter which one is treated as the response or explanatory variable.
- A positive correlation coefficient **does not** indicate a linear relationship.

Graphical Summaries

Nothing beats a picture.

Histogram

Used to portray the frequencies or relative frequencies of possible outcomes for **quantitative variables**.

Look out for:
- The overall pattern. Do the data cluster together, or is there a gap in which one or more data deviate from the rest?
- Is the data single modal, bimodal or multimodal?
- Is the distribution skewed? Any suspected outliers?

Boxplots

Provide a skeletal representation of a distribution. Well suited for showing quantitative categorical data. A boxplot is able to help us identify the median, lower, upper quantiles and outlier(s).
- Define max whisker to be $Q_3 + 1.5IQR$, vice versa. Points is out of range from the min to max whisker are considered outliers.
- Define upper whisker to be the maximum point of the data.
- Define extreme outlier to be larger than $Q_3 + 3IQR$ or $Q_1 - 3IQR$

If there are more than 1 variable,
- Is there a trend?
- Compare the range, median, and mean

QQ plots

A QQ-plot plots the **standardized sample quantiles (x-axis)** against the **theoretical quantiles (y-axis)** of a N(0; 1) distribution. If the points on the tail forms a trend deviating from the straight line, there is evidence that the data is not normal.
- Right tail is below the straight line: longer tan Normal.
- Right tail is above the straight line: shorter than Normal.
- Left tail is below the straight line: shorter than Normal.
- Left tail is above the straight line: longer than Normal.

Scatterplots

Scatterplots can help to visualize the association between two quantitative variable.
- Is there any relationship between the two variables?
- Is the association positve or negative? If so, is it linear or non-linear?
- Are there any observations that departs from the overall trend?
- Is the variance of the y-variable stable when the value of x-variable changes?

Robust Estimators

We make assumptions about the underlying distributions. However, outliers in the sample may cause the sample distribution to depart from the underlying distribution assumptions. The conclusion derived from this sample might not be reliable if the statistical methods used are not robust.

We say a statistical method is **robust** if it performs adequately even when the assumption is modestly violated.

Recall that the 95% CI for population mean has the following assumptions:

$$\bar{X} \pm t_{n-1,0.975} \times \frac{s}{\sqrt{n}}$$

- Obtained from randomization.
- Sample distribution must be approximately normal.

The procedure of using the t-distribution to compute confidence intervals is not robust to the randomization assumption.

Location Parameter

Robust estimators of Location Parameter include **trimmed mean** and **winsorized mean**.

Trimmed mean

The $100\alpha\%$ trimmed mean calculated by: discarding the lowest $100\alpha\%$ and highest $100\alpha\%$ and take the arithmetic mean of the remaining data. It is recommended that we use α from 0.1 to 0.2, $mean(x, trim = 0.2)$.

Winsorized mean

Winsorized mean is the mean of trimmed and replaced data. Winsorization replaces extreme data values with less extreme values. The winsorized mean is computed after all the $[n\alpha]$ smallest observations are replaced by x($[n\alpha]$+1), and the $[n\alpha]$ largest observations are replaces by x(n-$[n\alpha]$)). It is recommended that we use α from 0.1 to 0.2.

M-Estimators for Location Parameter

One can obtain more robustness by another function of error than the sum of their squares. We can find the estimator denoted by T - which is a function of $x_1, \ldots, x_n$ and this T is the minimizer of

$$\sum_{i=1}^{n} p(x_i - T)$$

where p is a **non-constant function** that is meaningful.

Examples:
- if we set the function $p(x) = x^2$, the minimizer of $\sum_{i=1}^{n}(x_i - T)^2$ is $\bar{x}$.
- if we set the function $p(x) = |x|$, the minimizer of $\sum_{i=1}^{n}|(x_i - T)|$ is the sample median.
- if we set the function

$$p(x) = \begin{cases} 1/2x^2 & \text{for } |x| \le k \\ k|x| - 1/2k^2 & \text{for } |x| > k \end{cases}$$

  then the estimator corresponds to a Winsorized mean
- if we set the function

$$p(x) = \begin{cases} 1/2x^2 & \text{for } |x| \le k \\ 1/2k^2 & \text{for } |x| > k \end{cases}$$

  then the estimator corresponds to a trimmed mean

Scale Parameter

Should remain robust even when a portion of the data points are replaced by arbitrary numbers.

Standard Deviation

Not robust, sensitive to outliers, and may not remain bounded when a single data point is replaced by an arbitrary number.

Interquartile Range

Better than standard deviation, but it is not a robust as well. For a normal distribution, the standard deviation, σ, can be estimated by dividing the interquartile range by 1.35.

$$IQR(X) = \sigma \times IQR(Z)$$

Median Absolute Deviation

The median absolute deviation (MAD) is a robust measure of the variability of a univariate sample of quantitative data.

$$MAD = median_i(|x_i - median_j(x_j)|)$$

where the inner median, $median_j(x_j)$ is the median of n observations and the outer median, $median_i$ - the median of the n absolute values of the deviations about the median. For a normal distribution, $1.4826 \times MAD$ can be used to estimate the standard deviation.

$$MAD(X) = \sigma \times MAD(Z)$$

Categorical Data Analysis

A variable is called a categorical variable if each observation belongs to one of a set of categories. To distinguish between quantitative and categorical variables, one can ask if there is a meaningful distance between any two points in the data.
- If the observations can be ordered, the variable is **ordinal**. Otherwise, the variable is **nominal**.

Single Categorical Variable

We can use a frequency table (barplot) to produce the proportion or percentage of categories. The categorary with the highest frequency is known as the **modal** category.

Two Categorical Variable

- Two categorical variables are **independent** if the population conditional distributions for one of them are identical at each category of the other.
- The variables are **dependent** if the conditional distributions are not identical.

Contingency Table

Important to identify which variable is the **response** variable and which one is the **explanatory** variable, so that the conditional proportion can be calculated properly. Explanatory variables are often placed along the rows.

Comparing Proportions

Let ϕ_1, ϕ_2 denote the probabilities of success for each row. Let p_1, p_2 denote the sample proportion of successes for each row.
In an ideal scenario, $\phi_1 = p_1$ and $\phi_2 = p_2$.

- The **sample difference**, $p_1 - p_2$ is used to estimate the difference between $\phi_1 - \phi_2$. If this difference is significant, we can infer association between the two variables.
- **Relative risk**: The ratio p_1/p_2 is used to estimate ϕ_1/ϕ_2. If the ratio is significantly different from 1, we can infer association between the two variables.

Odds ratio

On top of what was mentioned, we have **odds ratio**.

$$\theta = \frac{\phi_1/(1-\phi_1)}{\phi_2/(1-\phi_2)}$$

When the two variables are independent, $\phi_1 = \phi_2$, so $\theta = 1$. The further it is away from 1, the stronger the association. Suppose OR = 1.51, then we say that the odds of **success** given **row** is 1.51 times the odds of **success** given **row 2**.

If the order of the rows **or** columns is reversed, the new value of θ is inverse of the original value.

On the other hand, if the table orientation reverses (row become column, and vice versa), the odds ratio **does not** change. This is unlike RR or difference of proportions, whose values depends on **each row**.

Confidence interval for Odds Ratio

The $100\%(1-\alpha)$ confidence interval is formed by

$$exp\{log\hat{\theta} \pm z_{\alpha/2} \times ASE(log\hat{\theta})\}$$

where

$$ASE(log\hat{\theta}) = \sqrt{\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{n_{ij}}}$$

If we want to get a 95% CI for the population OR then we'll use α = 0.05, and $z_{\alpha/2}$ = 1.96. If the CI contains 1, that means that the population OR might be 1, hence the two variables might be independent!

Prospective vs Retrospective Studies

For **prospective studies**, sample subjects are picked randomly from the population. Record the explanatory variable status. Follow the subjects over time to see if they "succeeded" with relevance to the study.
For **retrospective studies**, sample a group of success and failure cases (along columns). Further split the success and failure into the explanatory variable status (along rows).

- Cheap, quick and fewer subjects are involved.
- Cannot obtain a valid estimate of ϕ_1 and ϕ_2.
- Can only use **odds ratio**.

Chi-squared (χ^2) Test

An indication of the degree of evidence for an association. Do note that χ^2 test **does not depend** on the order in which the rows and columns are listed. Thus, they ignore some information when there is an ordinal variable.

For 2x2 Tables

We have the following hypotheses:

H_0 : The two variables are independent
H_1 : The two variables are dependent

We compare the **expected counts** to the **observed counts**. For a particular cell,

$$\text{Expected Count} = \frac{\text{Row total} \times \text{Column Total}}{\text{Total sample size}}$$

The formula χ^2 test statistic (with continuity correction) is:

$$\chi^2 = \sum \frac{(|\text{observed count} - \text{expected count}| - 0.5)^2}{\text{expected count}}$$

p-value is calculated from χ^2 distribution with degree of freedom 1. The smaller p-value will give stronger evidence against h_0.

Fisher exact test: You should use this if there is at least one expected count less than 5. More generally, when more than 25% of the cell counts have expected values less than 5.
McNemar Test: Test for dependence. Check if the same set of samples is used + before and after comparison is made. The test statistic:

$$\chi_1^2 = \frac{(|b-c|-1)^2}{b+c}$$

only subtract 1 in the numerator if the sample has a small cell count.

For RxC Tables

Generally, χ^2 test can be extended to tables larger than 2 by 2. The only **difference** is the distribution now follows (r - 1)(c - 1) degree of freedom.

Standardized Residuals

Test statistics and p-value describe the evidence against H0.

- cell-by-cell comparison of observed and estimated expected frequencies

Define **standardized residuals**

$$r_{ij} = \frac{n_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}(1-p_{i+})(1-p_{+j})}}$$

where n_ij is the **observed count**, μ_{ij} is the **expected count**, p_{i+} is the **marginal probability** of row i and p_{+j} is the **marginal probability** of column j. The denominator is the estimated standard error of $(n_{ij} - \mu_{ij})$ under H_0.

- If H_0 is true, then each r_{ij} has a large-sample standard normal distribution.
- If $|r_{ij}|$ in a cell exceeds 2, then it indicates lack of fit of H_0 in that cell. Positive means more than expected, vice versa.

Linear-by-Linear test

For ordinal data. To detect a trend, assign scores to categories and measure the degree of linear trend or correlation. The scores should have
1. same ordering
2. reflect the distances between categories

The null hypothesis is: two variables are independent; the alternative hypothesis is: the two variables are dependent. The test statistic is calculated by

$$M^2 = (n-1)r^2$$

where r is sample correlation between X and Y,
$\bar{u} = \sum_i u_i p_{i+}$ is the sample mean of row scores,
$\bar{v} = \sum_j v_j p_{+j}$ is the sample mean of the column scores.

$$r = \frac{\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})p_{ij}}{\sqrt{[\sum_i (u_i-\bar{u})^2 p_{i+}][\sum_j (v_j - \bar{v})^2 p_{+j}]}}$$

where $p_{ij} = n_{ij}/n$; $p_{i+} = n_{i+}/n$; $p_{+j} = n_{+j}/n$.
For large samples, test statistic M^2 has approximately a χ^2 distribution with 1 df.

General Notes

- If the $sd(data)$ is much larger than the estimation by MAD, and it is also smaller than the estimation by IQR/1.35. This suggests that there could be a large value(s) in the dataset that inflate the sd. Hence, the better estimate of sigma is using MAD.
- If the trimmed mean and Winsorized mean are much smaller than the arithmetic mean, which indicates that there is/are some large values that pull up the mean. Hence, the "central" location is better measured by its median, or estimate by the trimmed mean with alpha increased to 0.3.
- Plot different categories (female vs male) of the data on different scatter plots of quantitative data (height vs weight) to determine if there is an association between the quantitative data. It can be that the relation only hold for one of the category.
- Between two dataset, compare their range, median, mean, quantiles and extreme values.

R Code snippets

Reading files

```
data2<-read.table("ex_1.txt",
    header = FALSE, col.names = c("name1", "name2")
data3<-read.table("ex_1_comma.txt",
    header = FALSE, sep = ",")
```

Write files

```
write.table(data,"C:/Data/saved_file.txt")
write.csv(data,"C:/Data/saved_file.csv")
```

Matrix operations

```
solve(m) # inverse
m %*% n # matrix multiplication
diag(2) # create diagonal matrix, 2x2
rbind(a, b); cbind(a, b) # concatenate rows and cols
```

Skewness

```
skew <- function(x){
    n <- length(x)
    m3 <- mean((x-mean(x))^3)
    m2 <- mean((x-mean(x))^2)
    sk = m3/m2^(3/2)*sqrt(n*(n-1))/(n-2)
    return(sk)
}
```

Kurtosis

```
kurt <- function(x){
    n <- length(x)
    m4 <- mean((x-mean(x))^4)
    m2 <- mean((x-mean(x))^2)
    kurt =
        (n-1)/((n-2)*(n-3))*((n+1)*m4/(m2^2)-3*(n-1))
    return(kurt)
}
```

Histograms

```
# To include lowest endpoint set include.lowest=TRUE.
hist(data$y, freq = FALSE, main ="Histogram")
# xpoint <- seq(0 ,30,0.05)
x = seq((min(data$y)- 5), (max(data$y)+5),
    length.out = length(data$y))
yy <- with(data, dnorm(x, mean(data$y), sd(data$y)))
lines(x, yy, col = "red")
lines(density(data$y))

# To specify that 3 graphs in one column in one page
par(mfrow=c(1,3))
# To get back to 1 graph in one page.
par(mfrow=c(1,1))
```

QQ plots

```
## datax places the sample quantiles on the x-axis
qqnorm(mark, datax = TRUE, ylab = "Sample Quantiles",
    xlab="Predicted Quantiles", main="QQ", pch=20)
qqline(mark,datax = TRUE)
```

Scatterplot

```
male <- data[which(gender == "M"), ]
female <- data[which(gender == "F"), ]
plot(male$A, male$B,
    col="red", pch=2, xlab = "A", ylab = "B")
points(female$A, female$B, # overlay
    col="green", pch=20)
legend(x= "topright", # legend
    legend=c('male', 'female'),
    pch=c(19, 19), col=c('red', 'green'))
cor(A, B) # correlation
```

Boxplots

```
type = factor(type)
boxplot(energy~type) #quantitative~categorical
```

Winsorized mean

```
winsor<-function(x, alpha = 0.2) {
    n = length(x)
    xq = n * alpha
    x = sort(x)
    m = x[(round(xq)+1)]
    M = x[(n - round(xq))]
    x[which(x<m)] = m
    x[which(x>M)] = M
    return(c(mean(x),var(x)))
}
```

Median Absolute Deviation

```
median(abs(x - median(x))) # MAD
mad(x) # estimate of \sigma, = 1.4826*MAD
```

Odds ratio, RR, and Prop diff

```
## 2-sample test for equality of proportions without
## continuity correction:
test<-prop.test(chest.pain, correct=FALSE)
RR<-(test$estimate[1])/(test$estimate[2])
odds<-test$estimate/(1- test$estimate)
OR<-odds[1]/odds[2]
```

```
# Function for finding OR and CI of OR:
OR<-function(x, pad.zeros = FALSE, conf.level=0.95){
  if(pad.zeros){if(any(x==0)) {x<-x+0.5}}
  theta<-x[1,1]*x[2,2]/(x[2,1]*x[1,2])
  ASE<-sqrt(sum(1/x))
  CI<-exp(log(theta) +
    c(-1,1)*qnorm(0.5*(1+conf.level))*ASE)
  list(estimator=theta, ASE=ASE,conf.interval=CI,
      conf.level=conf.level)
}
```

Chiq test

```
chest.pain<-matrix(c(46,474,37,516), ncol=2, byrow=2)
dimnames(chest.pain)<-list(Gender=c("Male", "Female"),
                           CP=c("Yes","No"))
chisq.test(chest.pain)
chisq.test(chest.pain)$stdres # standardized residuals
```

Fisher Exact test

```
claritin<-matrix(c(4,184,2,260), ncol=2, byrow=2)
fisher.test(claritin,alternative = "two.sided")
```

McNemar test

```
x = matrix(c(25,1,17,7), nrow = 2, byrow = TRUE)
mcnemar.test(x, correct = TRUE)
```

Manual calculation

```
nc1<-c(17066,14464,788,126,37);## Column 1
nc2<-c(48,38,5,1,1);## Column 2

rsum<-nc1+nc2;   ## Row sums
csum<-c(sum(nc1),sum(nc2)); ## Column sums
n<-sum(csum)   ## total cell counts

rowp<-rsum/n   ## margin prob for rows
colp<-csum/n   ## margin prob for columns

pc1<-rsum*csum[1]/n;   ## expected value for Column 1
pc2<-rsum*csum[2]/n;   ## expected value for Column 2

############  Chi square test
## Chi-squared test statistics
X2<-sum((nc1-pc1)^2/pc1)+sum((nc2-pc2)^2/pc2)

## degrees of freedom for chi-squared test
df<-(5-1)*(2-1)
p_value<-1-pchisq(X2,df)

## adjusted residuals for Column 1
```

```
rc1<-(nc1-pc1)/sqrt(pc1*(1-rowp)*(1-colp[1]))
## adjusted  residuals for Column 2
rc2<-(nc2-pc2)/sqrt(pc2*(1-rowp)*(1-colp[2]))

########## Linear-by-Linear Association test:
v<-c(0,1);  ## scores for columns
u<-c(0,.5,1.5,4,7.0);  ## scores for rows

ubar=sum(u*rowp);  ## weighted avg scores for rows
vbar<-sum(v*colp);  ## weighted avg scores for columns
CV<-sum(c(
        sum((u-ubar)*nc1/n),
        sum((u-ubar)*nc2/n))*(v-vbar)
        )  ## weighted covariance

## weighted variance for rows' scores
V1<-sum((u-ubar)^2*rsum/n);
## weighted variance for columns' scores
V2<-sum((v-vbar)^2*csum/n);

r<-CV/sqrt(V1*V2)  ## weighted correlation
M<-sqrt(n-1)*r  ## Normalized test statistic

# one-sided p-value (+ve association)
pnorm(abs(M), lower.tail = FALSE)

# 2 sided p-value
# Two variables are dependent
pchisq(M^2,1, lower.tail = FALSE)
```

Python code snippets

Reading files

```
import pandas as pd
data = pd.read_csv (r"path_to_file.csv", sep="")
# changing the columns' name
data.columns = ['Obs', 'time', 'cases', 'distance']
```

Write files

```
#  Write the file 'text' to a csv file called 'test2'
text = pd.DataFrame(text)
text.to_csv("test2.csv")
```

Dataframe operations

```
# [id, test, grade]
grade = pd.DataFrame(grade, columns=['grade'])
data_test = pd.concat([test_scores, grade], axis = 1)
# [id, ...., test, grade]
data_new = pd.merge(data,data_test, on = 'id')

# select two columns
data_text = pd.DataFrame(text,
    columns = ['Subject', 'CA2'])

# mimic attach in R
def attach(df):
    for col in df.columns:
        globals()[col] = df[col]
```

Skewness

```
def skew(x):
    n = len(x)
    y = [0]*n
    z = [0]*n
    for i in range(n):
        y[i] = (x[i] - mean(x))**2
        z[i] = (x[i] - mean(x))**3
    m2 = mean(y)
    m3 = mean(z)
    sk = (m3/pow(m2,3/2))*pow(n*(n-1),1/2)/(n-2)
    return(sk)
```

Kurtosis

```
def kurt(x):
    n = len(x)
    y = [0]*n
    z = [0]*n
    for i in range(n):
        y[i] = (x[i] - mean(x))**2
        z[i] = (x[i] - mean(x))**4
    m2 = mean(y)
    m4 = mean(z)
    kur = (n-1)/((n-2)*(n-3))*
        ((n+1)*m4/(m2**2) - 3*(n-1))
    return(kur)
```

Barplots

```
import matplotlib.pyplot as plt
fig, axes = plt.subplots(1, 3, figsize=(15,4))
colors = ['tab:red', 'tab:blue', 'tab:orange']
plots = ['travel', "drivelic", "gender"]
for i, (ax, plot) in enumerate(zip(axes.flatten(), plots)):
  types = data[plot].unique()
  counts = [data[data[plot] == type][plot].count() for
type in types]
  ax.bar(types, counts, color=colors[i])
  plt.xlabel('type')
plt.show()
```

Histograms

```
import scipy.stats as scst
import matplotlib.pyplot as plt
l = list(np.arange(0,30,0.5))
y = scst.norm.pdf(l,  # this equivalent to qnorm in R
    loc = mean(x),scale = st.stdev(x))

# Add plots to base layer
plt.plot(l, y) # normal overlay
# density overlay
data['mark'].plot(kind = 'density', xlim = [0, 28])
plt.hist(data['mark'],
    bins=11, range=None, density=True, color='C4')
plt.title('Histogram of Midterm marks')
plt.xlabel('Values')
plt.ylabel('Probability')
plt.show()

### Plotting multiple histograms
fig, axes = plt.subplots(1, 3,
    figsize=(8,3), dpi=100, sharex=True, sharey=True)
colors = ['tab:red', 'tab:blue', 'tab:orange']

for i, (ax, type) in enumerate(
        zip(axes.flatten(), bats.type.unique())
    ):
    x = bats.loc[bats.type==type, 'energy']
    ax.hist(x, alpha=0.5, bins=30,
        density=True, stacked=True,
        label=str(type), color=colors[i])
    ax.set_title(type)

plt.subtitle('Probability Histogram by types',
    y=1.05, size=16)
ax.set_xlim(0, 50); ax.set_ylim(0, 1);
plt.tight_layout();
```

QQ plots

```
import pylab
import scipy.stats as scst
# Take note of the axis swap
scst.probplot(x, dist="norm", plot=pylab)
pylab.title('QQ Plot')
pylab.show()
```

Scatterplot

```
import matplotlib.pyplot as plt

groups = data.groupby("type")
for name, group in groups:
```

```
    plt.plot(group["x"], group["y"],
        marker="o", linestyle="", label=name)

# Alternatively
colors = ['tab:red', 'tab:blue']
for type in data.x11.unique():
    performance = data.y[data.x11==type]
    weight = data.x10[data.x11==type]

    plt.scatter(weight, performance,
        color=colors[type], label=type)

# End off
plt.legend()
plt.show()
```

Boxplots

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv(r"path_to_data.csv")
fig, ax = plt.subplots(figsize=(7,5))
data.boxplot(column=['col'], by='type',
    ax=ax, color = 'b')
plt.show()
```

Odds ratio, RR, and Prop diff

```
data = {'Yes': [46,37], 'No': [474,516]}
df = pd.DataFrame(data, columns =['Yes', 'No'])
tab = [
    df['Yes']/(df['Yes'] + df['No']),
    df['No']/(df['Yes'] + df['No'])
    ]
tab = np.asmatrix([tab[0],tab[1]])
tab = np.transpose(tab)
prob = df['Yes']/(df['Yes'] + df['No'])
RR =  prob[0]/prob[1] #this is Relative Risk

odds = prob/(1-prob) # the odds of 'Yes'
OR = odds[0]/odds[1] # this is odds ratio
```

Chiq test

```
import scipy.stats as scst
obs = np.array([[46,474], [37,516]])
# this will return:
# test statistic; p-value, dof, and expected values
scst.chi2_contingency(obs, correction = True)
```

```
# rxc table
obs = np.array([[762,327,468], [484,239,477]])
scst.chi2_contingency(obs, correction = True)
```

Fisher Exact test

```
claritin = np.array([[4,184], [2,260]])
scst.fisher_exact(claritin, alternative='two-sided')
```

McNemar test

```
from statsmodels.stats.contingency_tables
    import mcnemar
x = np.array([[25,1], [17,7]])
# the McNemar test in R is equivalent to this
# test in Python but using approximate p-value.
test1 = mcnemar(x, exact=False, correction=True)

# Should be used when table has small cell count
test2 = mcnemar(x, exact=True, correction=True)
```

Linear by linear test

```
import statsmodels.api as sm
table = np.array([[17066, 14464, 788, 126, 37],
    [48, 38, 5, 1, 1]])
ct = sm.stats.Table(np.asarray(table))
# scores for 2 rows
row_scores = np.asarray([0,1])
# scores for 5 columns
col_scores = np.asarray([0,0.5,1.5,4,7])
ct.test_ordinal_association(row_scores=row_scores,
    col_scores=col_scores)
```

Manual calculation

```
def linear_by_linear_test(col1, col2, u, v):
    matrix = [col1, col2]
    row_total = [a + b for a,b in zip(col1, col2)]
    col_total = [sum(col1), sum(col2)]
    n = sum(row_total)

    row_p = [x / n for x in row_total]
    ubar = sum([a*b for a,b in zip(row_p, u)])
    col_p = [x / n for x in col_total]
    vbar = sum([a*b for a,b in zip(col_p, v)])

    numerator = 0
    for i in range(len(u)):
        for j in range(len(v)):
            ui = u[i]
            vj = v[j]
            pij = matrix[j][i] / n
            numerator +=
                (ui - ubar) * (vj - vbar) * pij


    denominator = (
        sum([row_p[i] * (u[i] - ubar)**2
            for i in range(len(u))]) * \
        sum([col_p[j] * (v[j] - vbar)**2
            for j in range(len(v))]))**0.5

    r = numerator / denominator
    M = (n-1) * r**2
    p_value = 1- scst.chi2.cdf(M, 1)
    return "M is {} and two sided
        p-value is {}".format(M,p_value)

def summary(data):
    print(
        'min: ', min(data),
        "\nQ1: ", np.quantile(data,0.25),
        "\nmean: ", st.mean(data),
        "\nQ2: ", st.median(data),
        "\nQ3: ", np.quantile(data, 0.75),
        "\nmax: ", max(data),
        "\nrange: ", min(data), max(data),
        "\nvar: ", st.variance(data),
        "\nsd: ", st.stdev(data),
        "\nIQR: ", st.iqr(data)
    )

np.corrcoef(x, y)[0, 1] # correlation
pd.crosstab(
    index=data["type"],
    columns=data["count"])  # contingency table

# this is MAD as defined in the lecture notes
scst.median_absolute_deviation(days) # estimate sigma

# to winsorize the original data (trim and replace)
from scipy.stats.mstats import winsorize
new_days = winsorize(days, limits=[0.2, 0.2])

scst.trim_mean(days, 0.2) #trimmed mean
```