Save the testscores file in your directory

In SAS:

LIBNAME tests 'C:\Users\Etudiant\Desktop\gilwang';

Load the data to create dataset

**DATA** ts;

SET tests.Testscores; #tests is the name of the lib

Get the variables we are interested:

KEEP Gender SQTscore; # use DROP if want to get almost all variables

**run**;

Create 2 datasets (man and woman) and assign each instance to one of them to know the number of men and women .

**DATA** woman man;

SET tests.Testscores;

if gender = 'Male' then output man;

else output woman;

**run**;

Better option:

**proc** **freq** data = tests.Testscores;

tables gender;

**run**;

Some statistics; by default it provides number of non missing values, mean, std, min, max

**proc** **means** data=tests.Testscores mean min max median std stderr range kurtosis missing var;

var SATScore;

**run**;

**proc** **means** data=tests.Testscores mean median Q1 Q3 range QRANGE;

var SATScore;

**run**;

plot histogram

**proc** **univariate** data=tests.Testscores;

var SATScore;

histogram;

**run**;

To add normal distribution plot for comparison

**proc** **univariate** data=tests.Testscores;

var SATScore;

histogram SATScore / normal(mu=est sigma=est);

**run**;

22/11/2018

Create a histogram for male and another for female

**proc** **univariate** data=tests.Testscores;

class gender; #one different histogram for each class inside gender

var SATScore;

histogram;

**run**;

change the default bin size (from 800 to 1600 and width = 200):

with SAS you cannot define bins with different sizes!!!!!!!!

**proc** **univariate** data=tests.Testscores;

class gender;

var SATScore;

histogram / endpoints=(**800** to **1600** by **200**);

#with midpoint 800 would be the center of the first bin; with #endpoint 800 will be the low delimiter of the first bin

**run**;

Let the program estimate c by minimizing the MISE and show the normal curve to compare:

**proc** **univariate** data=tests.Testscores;

class gender;

histogram SATScore / kernel(c = MISE) endpoints=(**800** to **1600** by **200**) normal (mu=est sigma=est);

**run**;

**reg script**

To see the real names of the columns:

**proc** **print** data=stat1.ameshousing3;

**run**;

%let create a macro group of variables:

%let categorical=House\_Style Overall\_Qual Overall\_Cond Year\_Built

Fireplaces Mo\_Sold Yr\_Sold Garage\_Type\_2 Foundation\_2

Heating\_QC Masonry\_Veneer Lot\_Shape\_2 Central\_Air;

By default the proc freq does not plot the graphics, that’s why we use ods graphics;

**proc** **univariate** data=STAT1.ameshousing3 noprint;

var &interval;

histogram &interval / normal kernel;

inset n mean std / position=ne; #add it to the legend (ne = north east)

title "Interval Variable Distribution Analysis";

**run**;

T-test (student test):

**proc** **ttest** data=STAT1.ameshousing3

plots(shownull)=interval

H0=**135000**; #null hypothesis

var SalePrice;

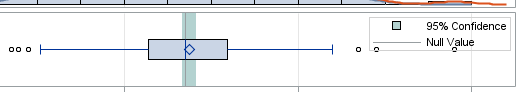
title "One-Sample t-test testing whether mean SalePrice=$135,000";

**run**;

It is for 1 variable and 1 sample. we assume that we have Gaussian distribution

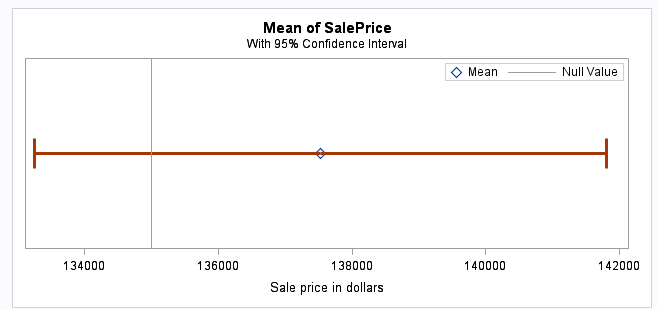
The normal and kernel curves are very similar so we can accept the assumption of gaussianity

DF = degrees of freedom = N -1

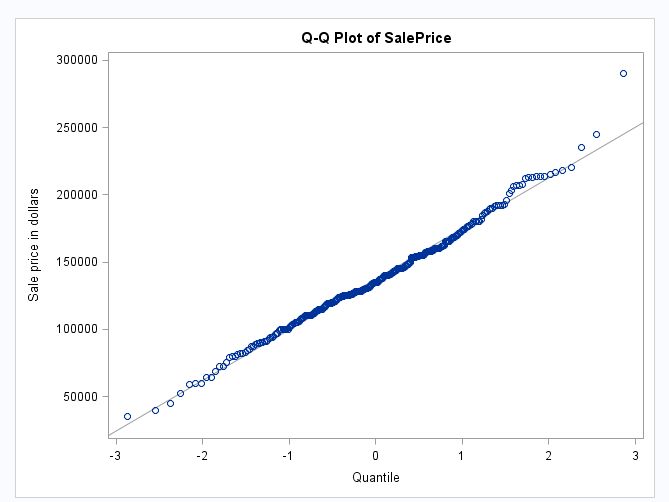


Since the vertical line is inside the confidence interval (green) we accept the null hypothesis. We also find: pvalue = 0.24602 higher than alpha = 0.05

ampliation of the green area:



QQ plot: other way to check the gaussianity of your data. If the points are exactly on the line or almost on it, you can consider your data follows a Gaussian distribution.



With class we make the analysis according to a binary variable separatedly:

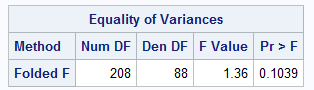
**proc** **ttest** data=STAT1.ameshousing3 plots(shownull)=interval;

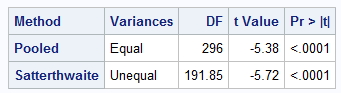
class Masonry\_Veneer;

var SalePrice;

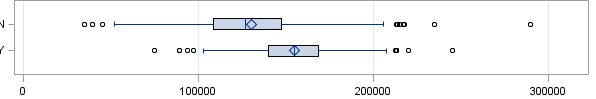
title "Two-Sample t-test Comparing Masonry Veneer, No vs. Yes";

**run**;

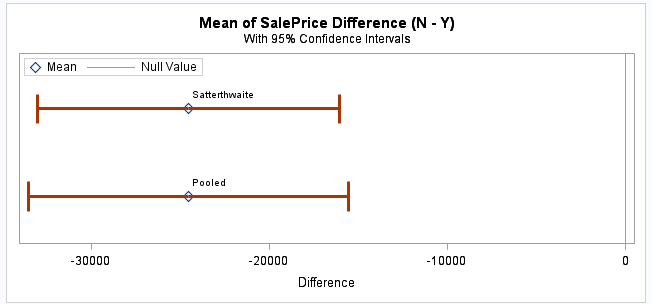
 Can we accept that the variances of both samples are equal?? since Pr = 0.1039 higher than alpha=0.05 we accept the equality. So we use the data in the first row in the following table (about the equallity of the expectation or mean):



since pvalue 0.0001 lower than alpha we reject it. We also see it in the following:



and here:



the vertical line on the right is not inside the red range.