1) Research Question

The client plans to set up a wine shop that specializes in selling wines from USA. His expertise is in marketing, thus he would like to hire an expert wine reviewer to help him identify USA wines that are worth stocking. Unfortunately, many wine reviewers claim that they are experts in wines of many countries, and he is unable to easily tell the difference.

After many interviews, the client is about to hire, Joe Czerwinski, who claims to specialize in USA and French wine. He asserts that USA wines are superior to French wines and that he is the best person for the job. The client would like to know if this is true.

2) Plan and Collect Data

A data set of Joe Czerwinski's reviews of USA and French wines have been collected from: https://www.kaggle.com/zynicide/wine-reviews

There is a total of 200 rows of data with the columns:

- Country
- Points (rating of wine)
- Taster Name (the reviewers name)
- Title (name of the wine)

country	points	taster_name	title
US	86	Joe Czerwinski	Elk Cove 2001 Riesling (Willamette Valley)
US	86	Joe Czerwinski	Glenora 2002 Dry Riesling (Finger Lakes)
US	87	Joe Czerwinski	Mayo 1998 Sangiacomo Vyds Pinot Noir (Carneros)
US	86	Joe Czerwinski	Barefoot Cellars 1998 Reserve Pinot Noir (Sonoma County)
US	81	Joe Czerwinski	Wagner 1998 Ice Wine Vidal Blanc (Finger Lakes)
US	92	Joe Czerwinski	Panther Creek 1998 Bednarik Vineyard Pinot Noir (Willamette Valley)
US	91	Joe Czerwinski	Siduri 1998 Archery Summit Vineyard Pinot Noir (California)

However, this format is not suitable for mini-tab and I converted it to:

USA	French
86	91
86	90
87	90
86	88
81	87
92	87
91	89

$$n_{usa\_wine} = 100,$$

$$n_{french\_wine} = 100$$

3) Analyze Data

The samples n_{usa_wine} and n_{fre_wine} are independent because the values from each sample are not related or paired to the values from the other sample even though it is the same wine taster that did the reviews of all the wines from both samples. It is not a comparison of two different methods of measurements or two different treatments i.e. both samples are not from one population and measured twice, hence they are independent.

$$n_{usa\_wine} = 100,$$

$$n_{fre\_wine} = 100$$

$$\sigma = unknown$$

Hypothesis

Let μ_1 be the mean rating of USA wines

Let μ_2 be the mean rating of French wines

Status Quo: USA wines are not more superior to French wines (as rated by Joe Czerwinski)

$$h_0: \mu_1 - \mu_2 = 0$$

Alternative: USA wines are superior to French wines (as rated by Joe Czerwinski)

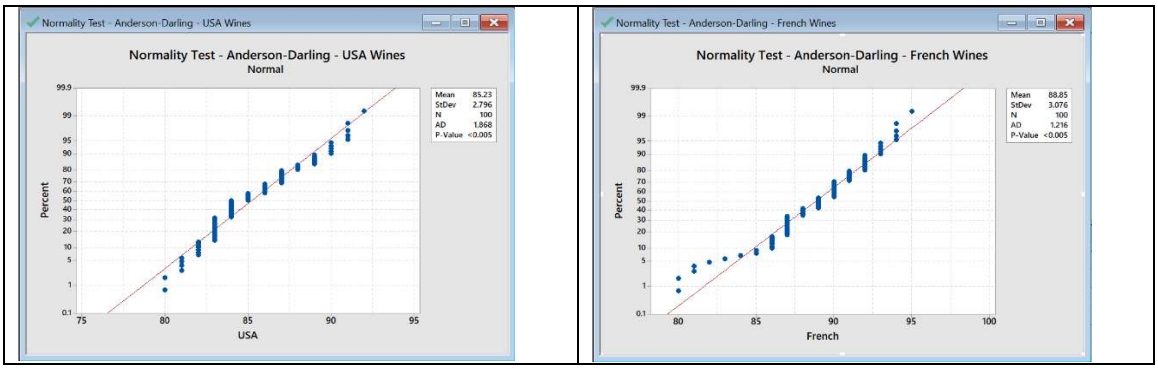
$$h_0: \mu_1 - \mu_2 > 0$$

$$\alpha = 0.05$$

Perform Normality Test

H0: Data follow a normal distribution

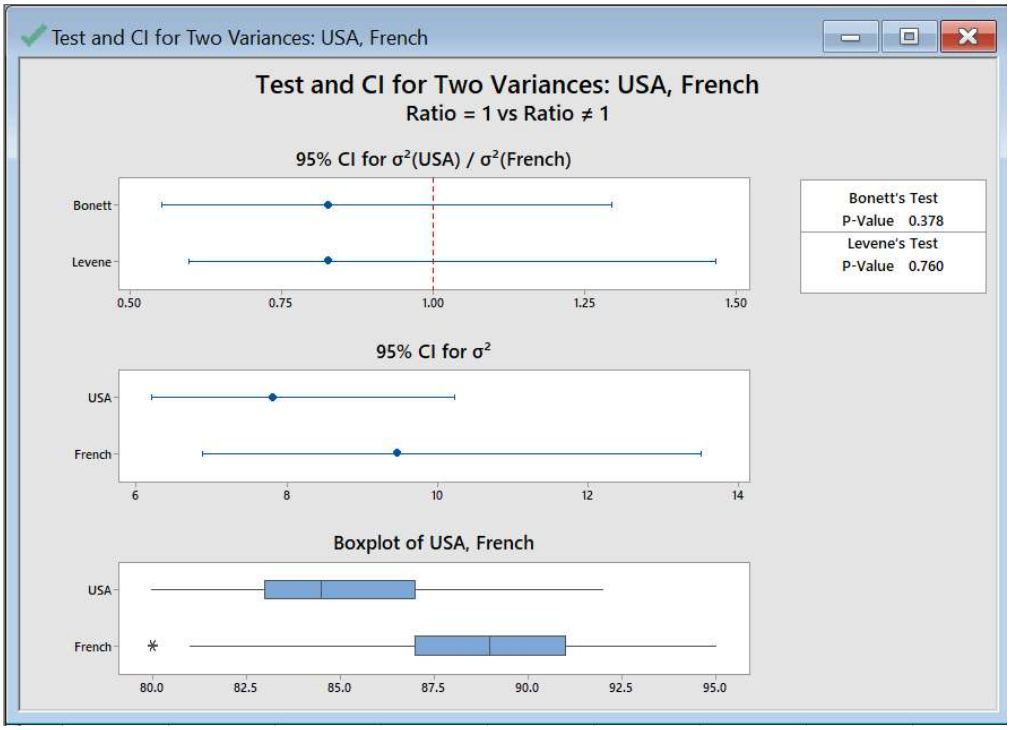
H1: Data do not follow a normal distribution



The P-Value for both Anderson-Darling Normality Tests are < 0.005, hence we reject H1 and conclude that both samples do not follow a normal distribution.

However, based on http://blog.minitab.com/blog/understanding-statistics-and-its-application/what-should-i-do-if-my-data-is-not-normal-v2 , several test are "robust to the assumption of normality, including t-test (1-sample, 2-sample and paired t-tests), hence even though normality is an underlying assumption for the tests above, they should work for non-normal data as well as if the data (or residuals) were normal.

Test for equality of variance



Null hypothesis	$H_0: \sigma_1^2 / \sigma_2^2 = 1$
Alternative hypothesis	$H_1: \sigma_1^2 / \sigma_2^2 \neq 1$
Significance level	$\alpha = 0.05$

P-value of both Bonett and Levene is 0.378 and 0.760 respectively and are both > $\alpha = 0.05$, hence we do not reject H_0 and conclude that the variances of both samples are equal. Hence, subsequently, we select independent two-sample t-test with equal variances at $\alpha = 0.05$

(see next page)

Analyse Samples

Two-Sample T-Test and CI: USA, French

Method

μ_1: mean of USA
μ_2: mean of French
Difference: $\mu_1 - \mu_2$

Equal variances are assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
USA	100	85.23	2.80	0.28
French	100	88.85	3.08	0.31

Estimation for Difference

Difference	Pooled StDev	95% Lower Bound for Difference
-3.620	2.939	-4.307

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 > 0$

T-Value	DF	P-Value
-8.71	198	1.000

4) Interpret Results

P-Value is $1 > \alpha = 0.05$, indicates that it is not rare to get a difference of sample mean wine ratings of -3.620 if the null hypothesis $\mu_1 - \mu_2 = 0$ is true.

So we do not reject H_0 at $\alpha = 0.05$

Hence, we do not reject the status Quo and conclude that there is no evidence to reject Joe Czerwinski's assertion that USA wines are superior to French wines.

The client wishes to focus on selling USA wines and wants a USA wine expert and since Joe shows a bias towards French wines by on average rating them higher than USA wines (wine taste is subjective and this shows bias), based on our analysis, the client decide not to hire Joe as his wine consultant and to continue searching for the right candidate.