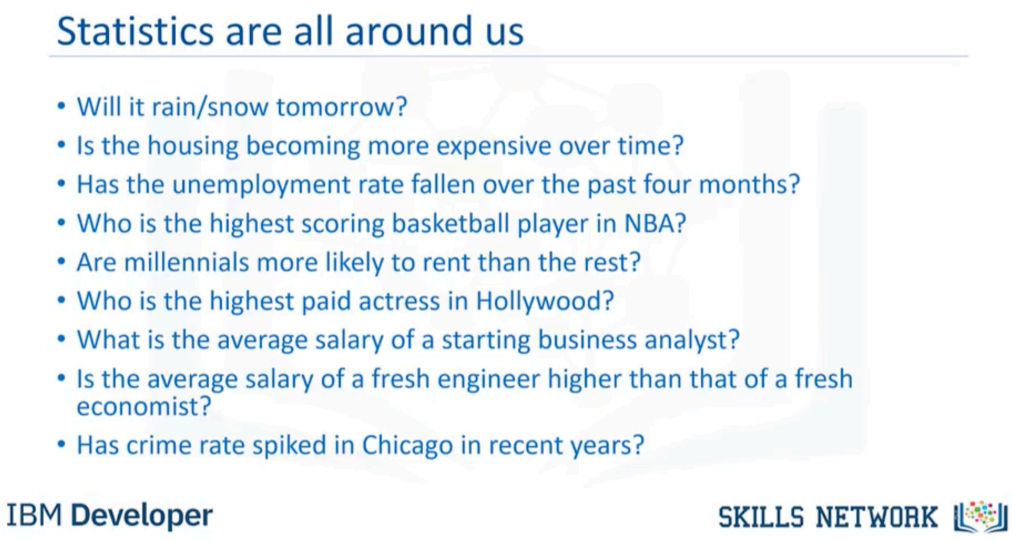
Statistics with Python

By: Vadhna Samedy Hun

I. Introduction



1.1. Measure of Central Tendency

Properties of the Mean

- Meaningful for interval and ratio data (continuous variables)
- Affected by unusually large or small observations (outliers)
  - Hence median is also useful
- The only measure of central tendency where the sum of the deviations of each value from the measure is zero; i.e.,

$$\sum(x_i - \bar{x}) = 0$$

IBM Developer

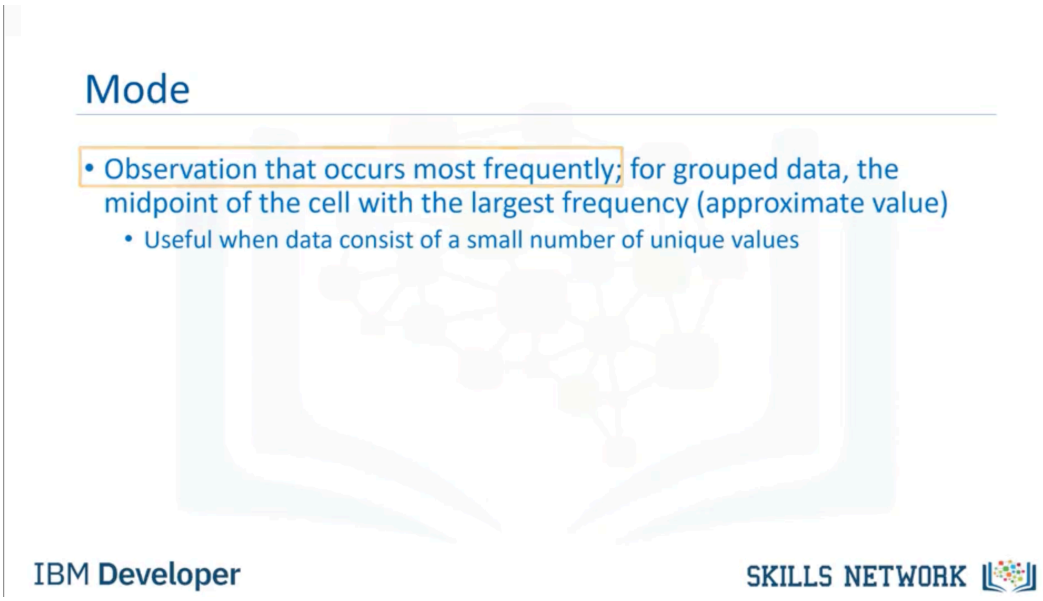
SKILLS NETWORK

Median

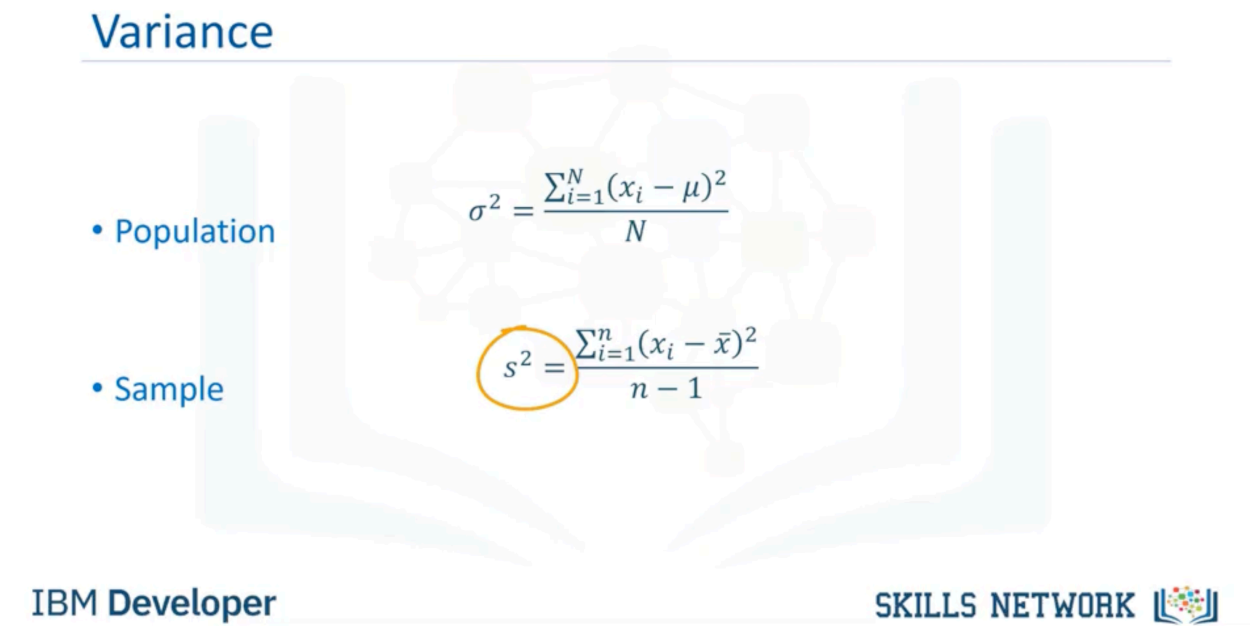
- Middle value when data are ordered from smallest to largest. This results in an equal number of observations above the median as below it
  - Unique for each set of data
  - Not affected by extremes
  - Meaningful for ratio, interval, and ordinal data

IBM Developer

SKILLS NETWORK



1.2. Measure of Dispersion



Standard Deviation

- The standard deviation is the square root of the variance
- The variance is in "square units" so the standard deviation is in the same units as x

- Population

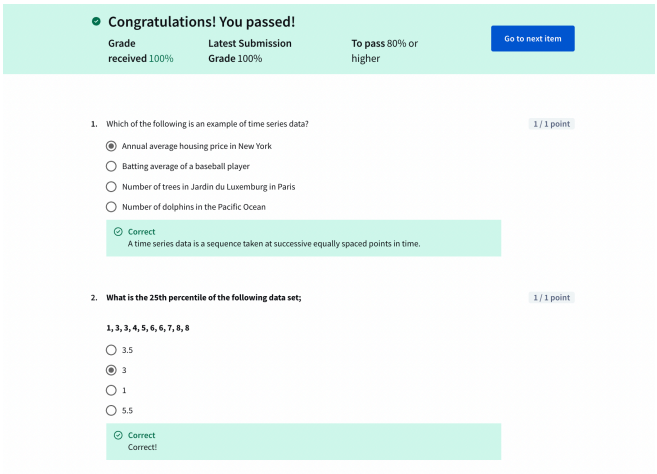
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

- Sample

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

IBM Developer SKILLS NETWORK

QUIZ 1:



Congratulations! You passed!

Grade received 100% | Latest Submission Grade 100% | To pass 80% or higher | Go to next item

1. Which of the following is an example of time series data? 1 / 1 point

- (•) Annual average housing price in New York
- ( ) Batting average of a baseball player
- ( ) Number of trees in Jardin du Luxemburg in Paris
- ( ) Number of dolphins in the Pacific Ocean

Correct
A time series data is a sequence taken at successive equally spaced points in time.

2. **What is the 25th percentile of the following data set;** 1 / 1 point

1, 3, 3, 4, 5, 6, 6, 7, 8, 8

- ( ) 3.5
- (•) 3
- ( ) 1
- ( ) 5.5

Correct
Correct!

3. Which of the following is a measure of variability? 1 / 1 point

- ○ **Mean**
- ○ Mode
- ◉ Variance
- ○ Median

✓ Correct
Correct!

4. Which of the following measures of central tendency will always change if a single value in the data changes? 1 / 1 point

- ○ All of the above
- ○ Median
- ◉ Mean
- ○ Mode

✓ Correct
Correct!

5. Which of the following data sets has a mean of 10 and standard deviation of 0? 1 / 1 point

- ○ 15, 15, 15
- ◉ 10, 10, 10
- ○ 0, 0, 0
- ○ 0, 10, 20

✓ Correct
Many data sets can have a mean of 10. However, if you force the standard deviation to be 0, you have only one choice: 10, 10, 10. A standard deviation of 0 means the average distance from the data values to the mean is 0. In other words, the data values don't deviate from the mean at all, and hence they have to be the same value.

6. What is meta data? 1 / 1 point

- ○ The data about metamorphism
- ◉ It's the data about data
- ○ The metabolism data in a clinical trial
- ○ Data about metal fatigue

✓ Correct
Correct!

7. Which of the following is an example of categorical data? 1 / 1 point

- ○ Number of fire hydrants in a city
- ○ Number of children at a kindergarten
- ○ Length of the river Nile
- ◉ Mode of travel to work

✓ Correct
Correct!

8. Median represents a value in the data set where: 1 / 1 point

- ○ Most observations are positive
- ○ Most observations are negative
- ◉ Half of the observations are above the median and the other half below it
- ○ Half of the observations are known and the other half not known

✓ Correct
Correct!

9. If the variance of a dataset is correctly computed with the formula using (n – 1) in the denominator, which of the following option is true? 1 / 1 point

- Data is a sample
- Data contains other variables with categorical data
- Data is from an unknown source
- Data is a population

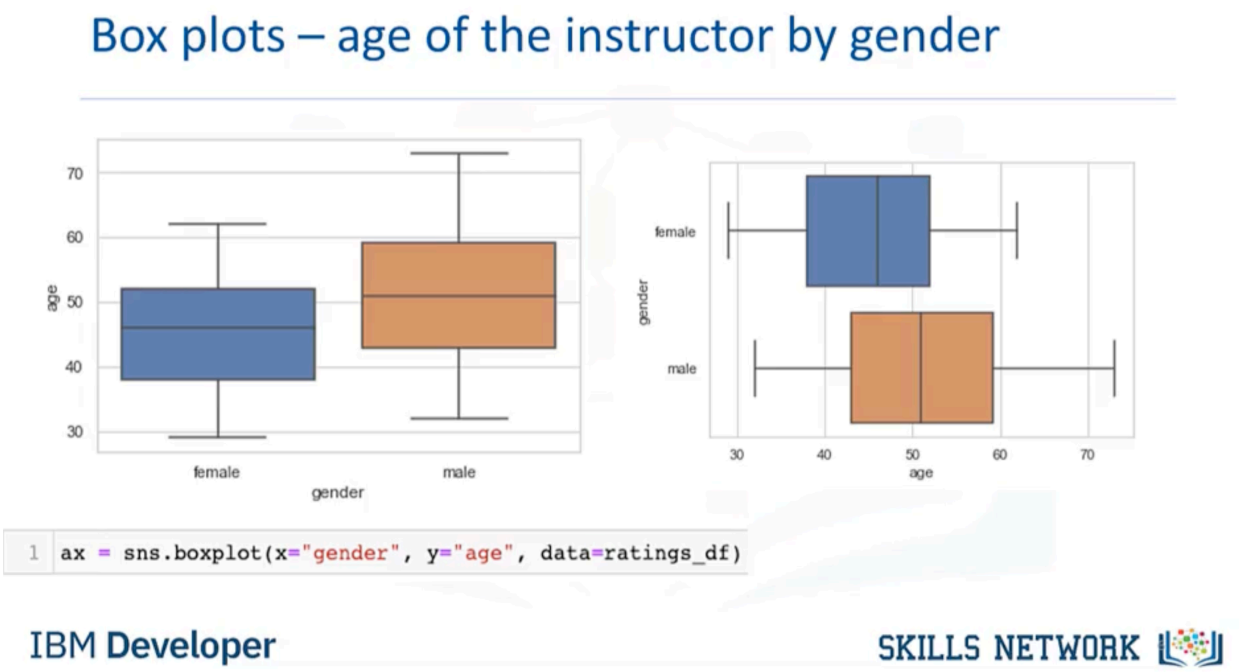
Correct
Correct!

10. Which of the following is NOT a descriptive statistic? 1 / 1 point

- t-test
- Mean
- Median
- Standard Deviation

Correct
Correct!

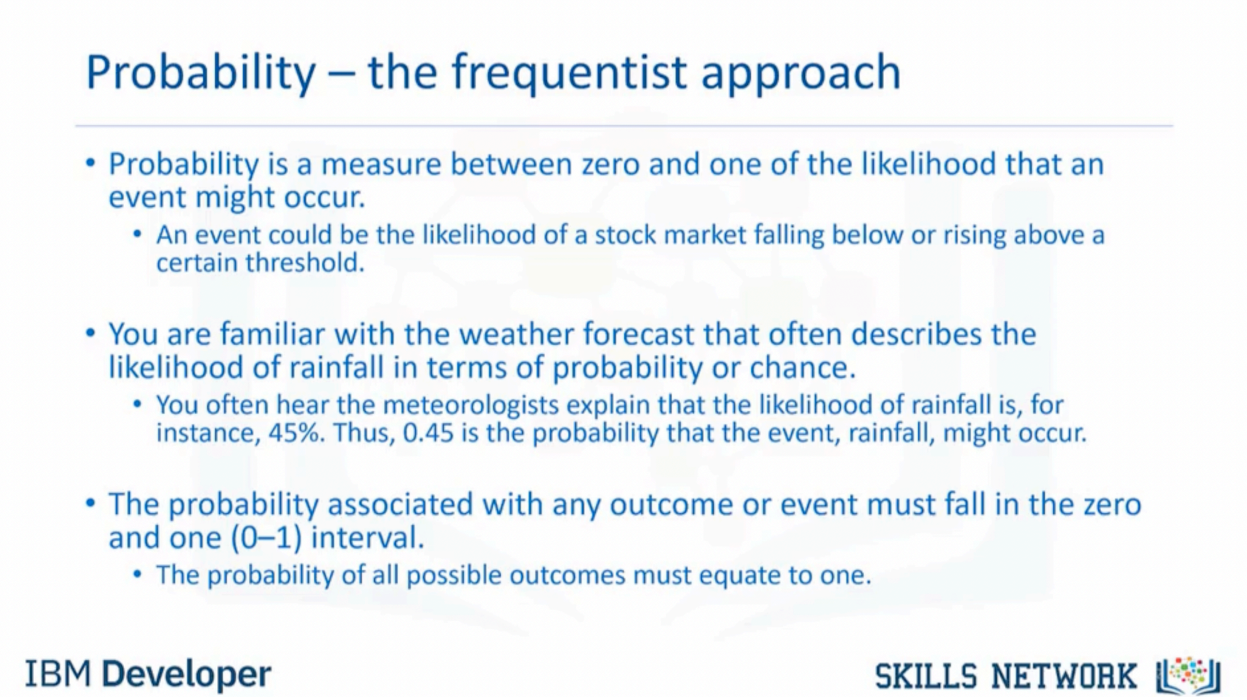
II. Fundamentals of Visualization




```
sns.distplot(ratings_df[ratings_df['English_speaker'] == 1]['beauty'], color='orang
sns.distplot(ratings_df[ratings_df['English_speaker'] == 0]['beauty'], color="blue
```

```
plt.show()
```

IV. Probability Distributions



The Math behind Normality

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2pi}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

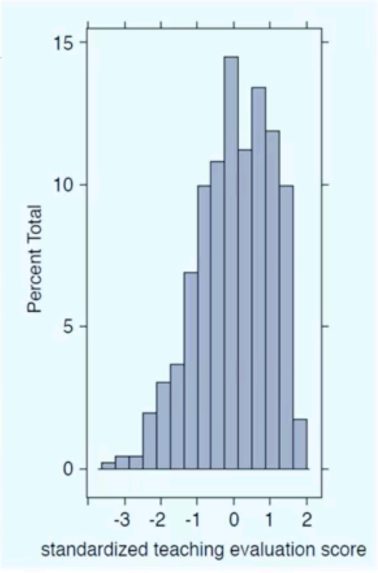
IBM Developer

SKILLS NETWORK

Standardization

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{4.5 - 3.998}{0.554} = 0.906$$



IBM Developer

SKILLS NETWORK

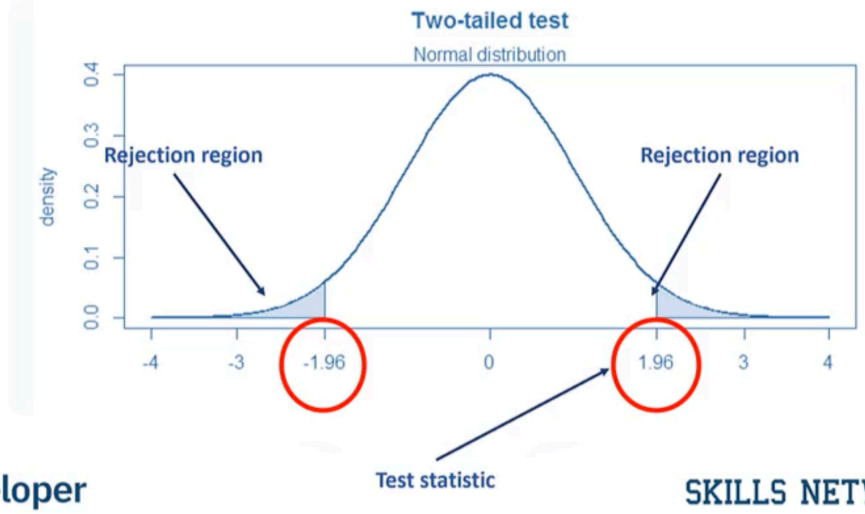
Comparing means – 4 cases

Comparing sample mean to a population mean when the population standard deviation is known	• Use Z test
Comparing sample mean to a population mean when the population standard deviation is not known	• Use T Test
Comparing the means of two independent samples with unequal variances	• Always use T Test
Comparing the means of two independent samples with equal variances	• Always use T Test

IBM Developer

SKILLS NETWORK

Normal distribution and rejection regions



IBM Developer

SKILLS NETWORK

QUIZ:

1. Using the teacher's rating data, is there an association between native (native English speakers) and the number of credits taught? What test will you use? — 1 point

- ○ Chi-Square Test for Association
- ○ Z-test
- ◉ ANOVA
- ○ T-test

Incorrect
Incorrect! ANOVA is used to check if there is a difference between two or more group means

2. If I wanted to test for association using chi-square test, whether there is an association between gender (Male or Female) and tenure-ship (tenured or not tenured), what will be my degree of freedom? — 1 / 1 point

1

Correct
Formula for degree of freedom for chi-square if (r-1) *(c-1)

3. Consider a normally distributed data set with mean μ = 63.18 inches and standard deviation σ= 13.27 inches. What is the z-score when x = 91.54 inches? (To 3 decimal places) — 1 / 1 point

2.137

Correct

4. Battery life of smartphones is of great concern to customers. A consumer group tested four brands of smartphones to determine the battery life. Samples of phones of each brand were fully charged and left to run until the battery died. The table above displays the number of hours each of the batteries lasted. What test will be be using to test the difference in means?

- ○ T-test
- ○ Chi-square Test
- ○ Pearson Correlation Test
- ◉ ANOVA

Correct
Correct! there are more than two groups

5. A room in a laboratory is only considered safe if the mean radiation level is 400 or less. When a sample of 10 radiation measurements were taken, the mean value of the radiation was 414 with a standard deviation of 17. There are concerns that mean radiation is above 414. Radiation levels in the lab are known to follow a normal distribution with standard deviation 22. We will like to conduct a hypothesis test at the 5% level of significance to determine whether there is evidence that the laboratory is unsafe.

1 / 1 point

What will be the appropriate test?

- (●) z-test
- ( ) t-test
- ( ) ANOVA
- ( ) Chi-square

Correct
Correct! We use a z-test when the population standard deviation is known

6. The mineral content of a particular brand of supplement pills is normally distributed with mean 490 mg and variance of 400. What is the probability that a randomly selected pill contains at least 500 mg of minerals?

1 / 1 point

- (●) 0.3085
- ( ) 0.2023
- ( ) 0.0525
- ( ) 0.7967

Correct
Correct!

7. The P-value for a normally distributed right-tailed test is P=0.042. Which of the following is **INCORRECT**? 1 / 1 point

- ○ The z-score test statistic is approximately z=1.73
- ○ The P-value for a two-tailed test based on the same sample would be P=0.084
- ○ We will reject H0 at α=0.05, but not at α=0.01
- ◉ The P-value for a left-tailed test based on the same sample would be P= -0.042

Correct
Correct! P-values are proportion and range from 0 to 1. The left-tail test for this will also be 0.042

8. The time X taken by a cashier in a grocery store express lane to complete a transaction follows a normal distribution with mean 90 seconds and standard deviation 20 seconds.What is the first quartile of the distribution of X (in seconds)? 1 / 1 point

- ○ 88.0
- ○ 73.8
- ◉ 76.6
- ○ 81.2

Correct
Correct!

9. A man accused of committing a crime is taking a polygraph (lie detector) test. The polygraph is essentially testing the hypotheses

1 / 1 point

H0: The man is telling the truth vs. Ha: The man is not telling the truth.

Suppose we use a 5% level of significance. Based on the man's responses to the questions asked, the polygraph determines a P-value of 0.08. We conclude that:

- (●) We fail to reject the null hypothesis as there is insufficient evidence that the man is not telling the truth.
- ( ) We reject the null hypothesis as there is sufficient evidence that the man is telling the truth.
- ( ) The probability that the man is not telling the truth is 0.08.
- ( ) The probability that the man is telling the truth is 0.08.

Correct
Correct! p-value is greater than 0.05

10. The average hourly wage at a fast-food restaurant is $5.85 with a standard deviation of $0.35. Assume that the wages are normally distributed. The probability that a selected worker earns more than $6.90 is

1 / 1 point

- ( ) 0.9987
- (●) 0.0013
- ( ) 0.4987
- ( ) 0

Correct
Correct!

V. Regression Analysis

Notation

- Dependent variable is denoted as y
- Explanatory variables are denoted as x
- y is explained by x or y is a function of x
- Mathematically:
  - $y = f(x)$
- Statistically:
  - $y = constant + weight_x(x) + error$
  - $y = \beta_0 + \beta_1 x + \epsilon$
  - If there are more than one explanatory variables:
    - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

IBM Developer SKILLS NETWORK

Regression in place of T-test

```
1 import statsmodels.api as sm
```

	female
130	0
173	0
357	0
457	1
17	1
254	0
411	0
121	1

```
1  import statsmodels.api as sm
2
3  ## X is the input variables (or independent variables)
4  X = ratings_df['female']
5  ## y is the target/dependent variable
6  y = ratings_df['eval']
7  ## add an intercept (beta_0) to our model
8  X = sm.add_constant(X)
9
10 model = sm.OLS(y, X).fit()
11 predictions = model.predict(X)
12
13 # Print out the statistics
14 model.summary()
```

IBM Developer SKILLS NETWORK

Regression in place of T-test

Dep. Variable:	eval	R-squared:	0.022
Model:	OLS	Adj. R-squared:	0.020
Method:	Least Squares	F-statistic:	10.56
Date:	Thu, 03 Sep 2020	Prob (F-statistic):	0.00124
Time:	14:50:47	Log-Likelihood:	-378.50
No. Observations:	463	AIC:	761.0
Df Residuals:	461	BIC:	769.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0690	0.034	121.288	0.000	4.003	4.135
female	-0.1680	0.052	-3.250	0.001	-0.270	-0.066

Omnibus:	17.625	Durbin-Watson:	1.209
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18.970
Skew:	-0.496	Prob(JB):	7.60e-05
Kurtosis:	2.981	Cond. No.	2.47

IBM Developer

SKILLS NETWORK

ANOVA in Python

Does beauty score for instructors differ by age?

	age_group	beauty		
		count	mean	std
0	40 years and younger	113	0.336196	0.913748
1	57 years and older	122	-0.245777	0.740720
2	between 40 and 57 years	228	-0.035111	0.686637

```
ratings_df.loc[(ratings_df['age'] <= 40), 'age_group'] = '40 years and younger'
ratings_df.loc[(ratings_df['age'] > 40)&(ratings_df['age'] < 57), 'age_group'] = 'between 40 and 57 years'
ratings_df.loc[(ratings_df['age'] >= 57), 'age_group'] = '57 years and older'
```

```
f_statistic, p_value = scipy.stats.f_oneway(forty_lower, forty_fiftyseven, fiftyseven_older)
print("F_Statistic: {0}, P-Value: {1}".format(f_statistic,p_value))
```

```
F_Statistic: 17.597558611010122, P-Value: 4.3225489816137975e-08
```

IBM Developer

SKILLS NETWORK

Regression for ANOVA

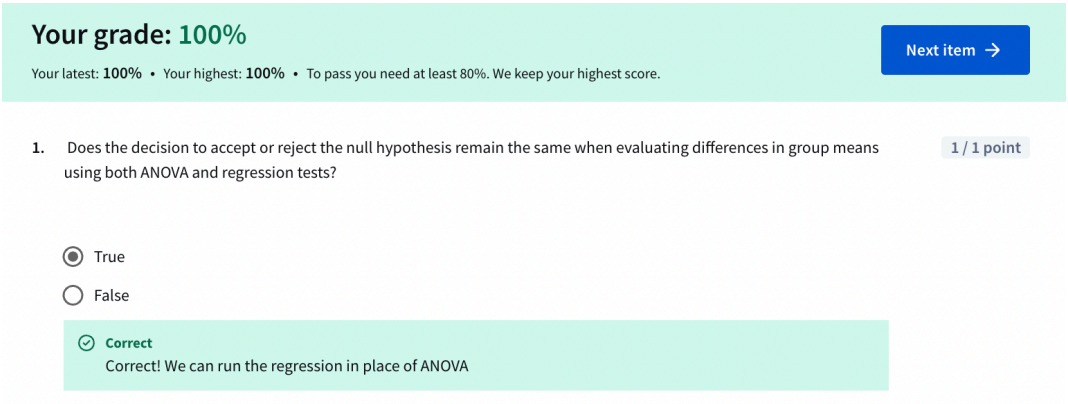
```
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

```
lm = ols('beauty ~ age_group', data = ratings_df).fit()
table= sm.stats.anova_lm(lm)
print(table)
```

	df	sum_sq	mean_sq	F	PR(>F)
age_group	2.0	20.422744	10.211372	17.597559	4.322549e-08
Residual	460.0	266.925153	0.580272	NaN	NaN

IBM Developer SKILLS NETWORK

Quiz:



2. Give the results of the regression analysis below, what is the correlation coefficient? 1 / 1 point

Dep. Variable:	eval	**R-squared:**	0.036
Model:	OLS	**Adj. R-squared:**	0.034
Method:	Least Squares	**F-statistic:**	17.08
Date:	Thu, 03 Sep 2020	**Prob (F-statistic):**	4.25e-05
Time:	16:36:25	**Log-Likelihood:**	-375.32
No. Observations:	463	**AIC:**	754.6
Df Residuals:	461	**BIC:**	762.9
Df Model:	1		
Covariance Type:	nonrobust		

- ○ 17.08
- ○ 0.036
- ◉ 0.19
- ○ 0.034

Correct
Correct!

3. Given the results for tenure-ship vs teaching evaluation, if our null hypothesis is that there is no difference in mean evaluation scores for professors who are tenured vs professors who are not tenured. What will be the conclusion of the t-test statistics? 1 / 1 point

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.1333	0.055	75.791	0.000	4.026	4.241
tenured_prof	-0.1732	0.062	-2.805	0.005	-0.295	-0.052

- ◉ P-value is less than 0.05, that means that there is a difference in mean values for professors who are tenures versus professors who are not tenured.
- ○ P-value is less than 0.05, we will fail to reject the null hypothesis.
- ○ There is no conclusive evidence in the results above.

Correct
Correct!

4. We run a regression analysis in place of a t-test to test if there is a difference in number of students enrolled in classes with professors who are visible minority(vismin = 1) vs professors who are not (vismin = 0). The table is shown below. What does the coefficient for vismin mean? 1 / 1 point

	coef	std err	t	P>|t|	[0.025	0.975]
const	58.0902	3.745	15.513	0.000	50.731	65.449
vismin	-21.0746	10.072	-2.092	0.037	-40.867	-1.282

- ○ We can't conclude because the error is too large and if factored could change the conclusion of the tests.
- ○ Professors who are visible minority get about 21 students more on average that professors who aren't visible minority.
- ○ Professors who are visible minority get about 58 students less on average that professors who aren't visible minority.
- ◉ Professors who are visible minority get about 21 students less on average that professors who aren't visible minority.

Correct
Correct!

5. Which of these are correct about correlation coefficient? (Select all that apply) 1 / 1 point

- ☑ The correlation coefficient (r) ranges from -1 to 1

Correct
Correct! Values can be positively and negatively related

- ☐ A correlation coefficient of -0.9 indicates a weak linear relationship?
- ☑ A correlation coefficient of -0.9 indicates a strong linear relationship?

Correct
Correct! The negative sign means they are strongly negatively correlated

- ☐ The correlation coefficient (r) ranges from 0 to 1

6. Which of these options is most likely to be the null hypothesis for testing correlation between two variables? 1 / 1 point

- ◉ There is no association between an instructor's looks and teaching evaluation score.
- ○ There is an association between an instructor's looks and teaching evaluation score.
- ○ There is a partial association between an instructor's looks and teaching evaluation score.

Correct
Correct!

7. If we ran a regression analysis between two continuous variables amount of time spent running on a treadmill vs the amount of calories burnt. If I get a coefficient of 0.33 for the amount of time running on the treadmill and an R-square value of 0.81. What is the correlation coefficient? 1 / 1 point

- (●) 0.9
- ( ) 0.66
- ( ) 0.81
- ( ) 0.77

Correct
Correct!

8. Which of the following best explains a scatter plot? 1 / 1 point

- (●) A two-dimensional graph of data values.
- ( ) A two-dimensional graph of a straight line.
- ( ) A two-dimensional graph of a curved line.
- ( ) A one-dimensional graph of randomly scattered data.

Correct
Correct! A scatter plot represents the relationship between two continuous data