1. **Kafka used for production implementation of new data adding to ML model**
2. **Validation curve to test Hyperparameters and higher degree of polynimoals**
3. **Bias and variance to test model on greater number of data**

**The Coefficient of Determination, r-squared**

SSR=∑ni=1(^yi−¯y)2

SSE=∑ni=1(yi−^yi)2

SSTO=∑ni=1(yi−¯y)2

SSTO = SSR + SSE

 SSR is the "regression sum of squares" and quantifies how far the estimated sloped regression line, ^yi, is from the horizontal "no relationship line," the sample mean or ¯y

 SSE is the "error sum of squares" and quantifies how much the data points, yi, vary around the estimated regression line, ^yi

 SSTO is the "total sum of squares" and quantifies how much the data points, yi, vary around their mean, ¯y.

**R2 =SSR/SSTO=1−SSE/SSTO**

* Since *r*2 is a proportion, it is always a number between 0 and 1.
* If *r*2 = 1, all of the data points fall perfectly on the regression line.The predictor *x* accounts for *all* of the variation in *y*!
* If *r*2 = 0, the estimated regression line is perfectly horizontal. The predictor *x* accounts for *none* of the variation in *y*!

**r2 ×100 percent of the variation in y is 'explained by' the variation in predictor x**

Students often ask: "what's considered a large r-squared value?" It depends on the research area. Social scientists who are often trying to learn something about the huge variation in human behavior will tend to find it very hard to get r-squared values much above, say 25% or 30%. Engineers, on the other hand, who tend to study more exact systems would likely find an r-squared value of just 30% merely unacceptable. The moral of the story is to read the literature to learn what typical r-squared values are for your research area!

**(Pearson) Correlation Coefficient r**

The correlation coefficient *r* is directly related to the coefficient of determination *r*2 in the obvious way. If *r*2 is represented in decimal form, *e.g.* 0.39 or 0.87, then all we have to do to obtain *r* is to take the square root of *r*2:

r=±√r2

The sign of *r* depends on the sign of the estimated slope coefficient *b*1:

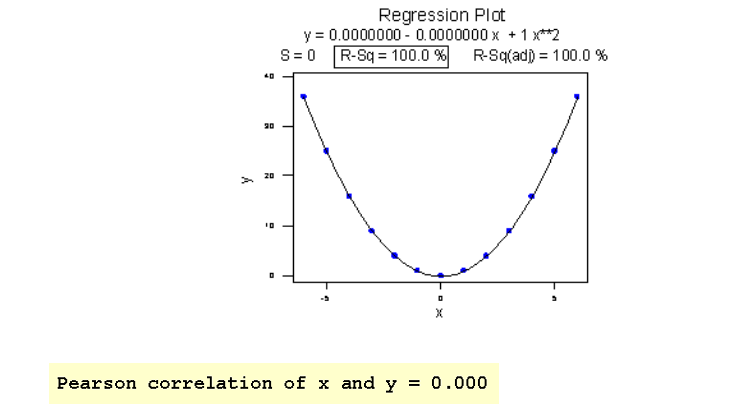
* If *b*1 is negative, then *r* takes a negative sign.
* If *b*1 is positive, then *r* takes a positive sign.

That is, the estimated slope and the correlation coefficient *r* always share the same sign. Furthermore, because *r*2 is always a number between 0 and 1, the correlation coefficient *r* is always a number between -1 and 1.

* If *r* = -1, then there is a perfect negative linear relationship between *x* and *y*.
* If *r* = 1, then there is a perfect positive linear relationship between *x* and *y*.
* If *r* = 0, then there is no linear relationship between *x* and *y*.

### Caution # 1

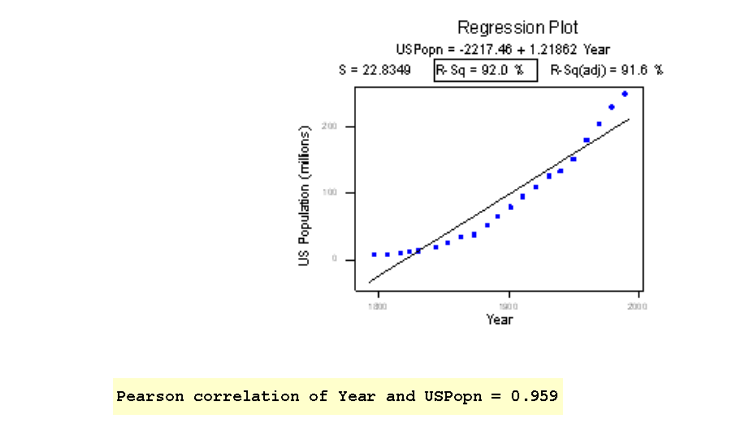
**The coefficient of determination r2 and the correlation coefficient r quantify the strength of a linear relationship. It is possible that r2 = 0% and r = 0, suggesting there is no linear relation between x and y, and yet a perfect curved (or "curvilinear" relationship) exists.**



### ****Caution # 2****

**A large r2 value should not be interpreted as meaning that the estimated regression line fits the data well. Another function might better describe the trend in the data.**

Consider the following example in which the relationship between year (1790 to 1990, by decades) and population of the United States (in millions) is examined:

****

### ****Caution # 3****

**The coefficient of determination r2 and the correlation coefficient r can both be greatly affected by just one data point (or a few data points).**

# *p*-Value

When you perform a hypothesis test in statistics, a *p*-value helps you determine the significance of your results. [Hypothesis tests](https://www.dummies.com/education/math/statistics/how-hypothesis-tests-are-used-in-statistics/) are used to test the validity of a claim that is made about a population. This claim that’s on trial, in essence, is called the [*null hypothesis*](https://www.dummies.com/education/math/statistics/how-to-determine-a-p-value-when-testing-a-null-hypothesis/)*.*

The *alternative hypothesis* is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it. All hypothesis tests ultimately use a *p*-value to weigh the strength of the evidence (what the data are telling you about the population). The *p*-value is a number between 0 and 1 and interpreted in the following way:

* A small *p*-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
* A large *p*-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
* *p*-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the *p*-value so your readers can draw their own conclusions.

P values evaluate how well the sample data support the devil’s advocate argument that the null hypothesis is true. It measures how compatible your data are with the null hypothesis. How likely is the effect observed in your sample data if the null hypothesis is true?

* High P values: your data are likely with a true null.
* Low P values: your data are unlikely with a true null.

A low P value suggests that your sample provides enough evidence that you can reject the null hypothesis for the entire population.

For example, suppose that a vaccine study produced a P value of 0.04. This P value indicates that if the vaccine had no effect, you’d obtain the observed difference or more in 4% of studies due to random sampling error.:

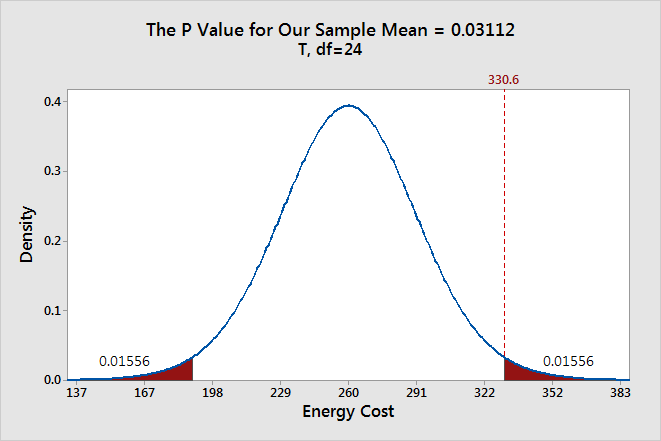
* **Correct:** Assuming that the vaccine had no effect, you’d obtain the observed difference or more in 4% of studies due to random sampling error.

## What Are P values?

P-values are the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis.

This definition of P values, while technically correct, is a bit convoluted. It’s easier to understand with a graph!

To graph the P value for our example data set, we need to determine the distance between the sample mean and the null hypothesis value (330.6 - 260 = 70.6). Next, we can graph the probability of obtaining a sample mean that is at least as extreme in both tails of the distribution (260 +/- 70.6).



In the graph above, the two shaded areas each have a probability of 0.01556, for a total probability 0.03112. This probability represents the likelihood of obtaining a sample mean that is at least as extreme as our sample mean in both tails of the distribution if the population mean is 260. That’s our P value!

When a P value is less than or equal to the significance level, you reject the null hypothesis. If we take the P value for our example and compare it to the common significance levels, it matches the previous graphical results. The P value of 0.03112 is statistically significant at an alpha level of 0.05, but not at the 0.01 level.

If we stick to a significance level of 0.05, we can conclude that the average energy cost for the population is greater than 260.

A common mistake is to interpret the P-value as the probability that the null hypothesis is true

Brinell Hardness Scores

An engineer measured the Brinell hardness of 25 pieces of ductile iron that were subcritically annealed. The resulting data were:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 170 | 167 | 174 | 179 | 179 | 187 | 179 | 183 | 179 |
| 156 | 163 | 156 | 187 | 156 | 167 | 156 | 174 | 170 |
| 183 | 179 | 174 | 179 | 170 | 159 | 187 |  |  |

The engineer hypothesized that the mean Brinell hardness of all such ductile iron pieces is greater than 170. Therefore, he was interested in testing the hypotheses:

H0 : μ = 170  
HA: μ > 170

The engineer entered his data into Minitab and requested that the "one-sample t-test" be conducted for the above hypotheses. He obtained the following output:

#### Descriptive Statistics

| **N** | **Mean** | **StDev** | **SE Mean** | **95% Lower Bound** |
| --- | --- | --- | --- | --- |
| 25 | 172.52 | 10.31 | 2.06 | 168.99 |

*μ*

: mean of Brinelli

#### Test

Null hypothesis    H₀: *μ*

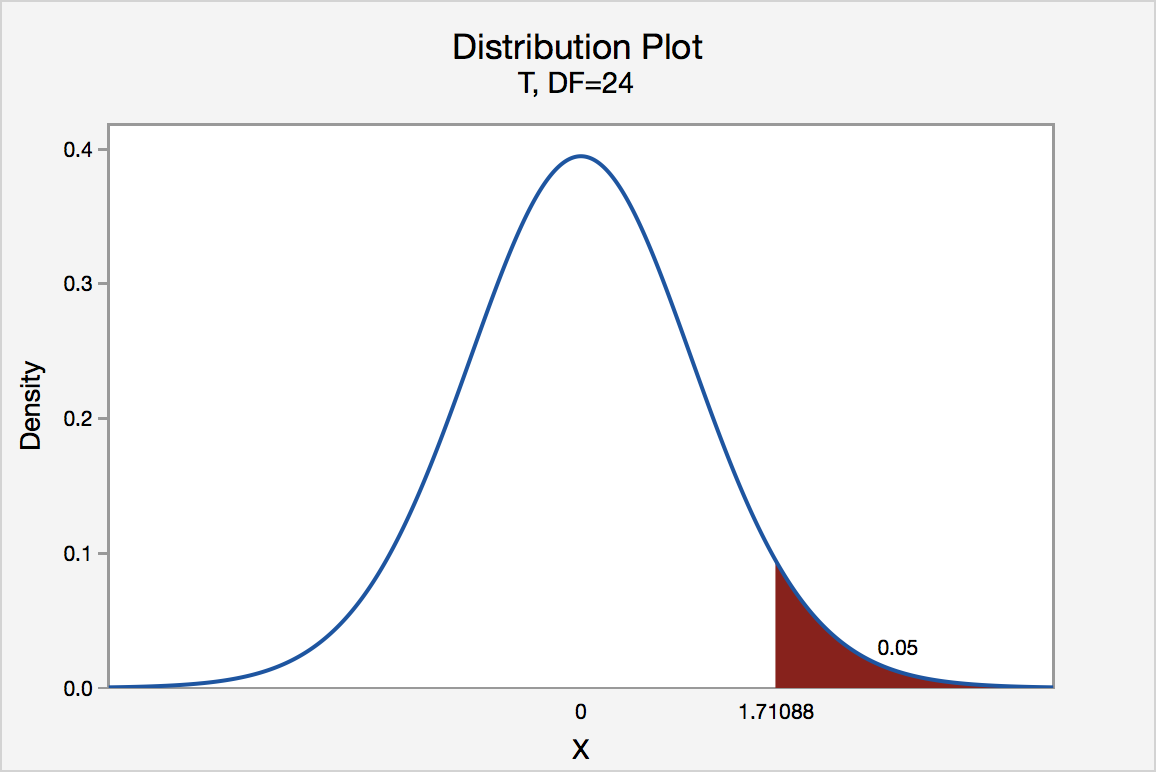
= 170  
Alternative hypothesis    H₁: *μ*

> 170

| **T-Value** | **P-Value** |
| --- | --- |
| 25 | 172.52 |

The output tells us that the average Brinell hardness of the n = 25 pieces of ductile iron was 172.52 with a standard deviation of 10.31. (The standard error of the mean "SE Mean", calculated by dividing the standard deviation 10.31 by the square root of n = 25, is 2.06). The test statistic t\* is 1.22, and the P-value is 0.117.

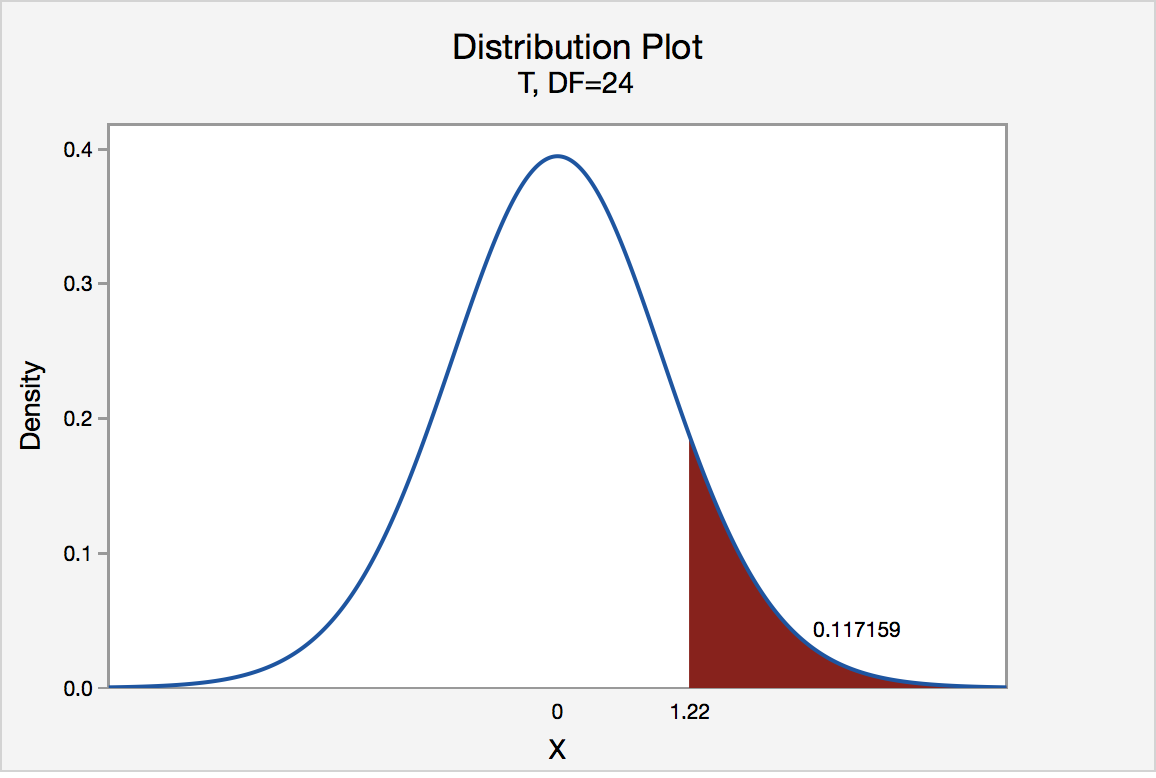
If the engineer set his significance level α at 0.05 and used the critical value approach to conduct his hypothesis test, he would reject the null hypothesis if his test statistic t\* were greater than 1.7109 (determined using statistical software or a t-table):



Since the engineer's test statistic, t\* = 1.22, is not greater than 1.7109, the engineer fails to reject the null hypothesis. That is, the test statistic does not fall in the "critical region." There is insufficient evidence, at the *α*

= 0.05 level, to conclude that the mean Brinell hardness of all such ductile iron pieces is greater than 170.

If the engineer used the P-value approach to conduct his hypothesis test, he would determine the area under a tn - 1 = t24 curve and to the right of the test statistic t\* = 1.22:



In the output above, Minitab reports that the P-value is 0.117. Since the P-value, 0.117, is greater than *α*

= 0.05, the engineer fails to reject the null hypothesis. There is insufficient evidence, at the *α*

= 0.05 level, to conclude that the mean Brinell hardness of all such ductile iron pieces is greater than 170.

Note that the engineer obtains the same scientific conclusion regardless of the approach used. This will always be the case.

Test the independence of two categorical variables

Do you remember how to test the independence of two categorical variables? This test is performed by using a **Chi-square test of independence**.

Chi-Square Test Statistic

*χ*2=∑(*O*−*E*)2/*E*

where *O* represents the observed frequency. *E* is the expected frequency under the null hypothesis and computed by:

*E*=row total×column totalsample size

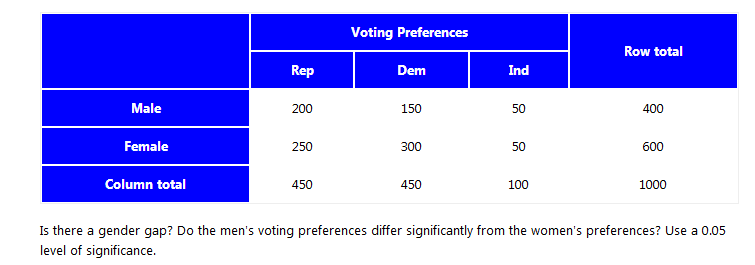
Now find the **P-value** that is the probability of observing a sample statistic as extreme as the test statistic

## Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](https://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

**Problem**

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the [contingency table](https://stattrek.com/Help/Glossary.aspx?Target=Contingency%20table) below.



**Solution**

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the [null hypothesis](https://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an alternative hypothesis.

Ho: Gender and voting preferences are independent.

Ha: Gender and voting preferences are not independent.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a [chi-square test for independence](https://stattrek.com/Help/Glossary.aspx?Target=Chi-square%20test%20for%20independence).
* **Analyze sample data**. Applying the chi-square test for independence to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the [degrees of freedom](https://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom), we determine the [P-value](https://stattrek.com/Help/Glossary.aspx?Target=P-value).

DF = (r - 1) \* (c - 1) = (2 - 1) \* (3 - 1) = 2

Er,c = (nr \* nc) / n  
E1,1 = (400 \* 450) / 1000 = 180000/1000 = 180  
E1,2 = (400 \* 450) / 1000 = 180000/1000 = 180  
E1,3 = (400 \* 100) / 1000 = 40000/1000 = 40  
E2,1 = (600 \* 450) / 1000 = 270000/1000 = 270  
E2,2 = (600 \* 450) / 1000 = 270000/1000 = 270  
E2,3 = (600 \* 100) / 1000 = 60000/1000 = 60

Χ2 = Σ [ (Or,c - Er,c)2 / Er,c ]   
Χ2 = (200 - 180)2/180 + (150 - 180)2/180 + (50 - 40)2/40  
    + (250 - 270)2/270 + (300 - 270)2/270 + (50 - 60)2/60  
Χ2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60  
Χ2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2

where DF is the degrees of freedom, r is the number of levels of gender, c is the number of levels of the voting preference, nr is the number of observations from level *r* of gender, nc is the number of observations from level *c* of voting preference, n is the number of observations in the sample, Er,c is the expected frequency count when gender is level *r* and voting preference is level *c*, and Or,c is the observed frequency count when gender is level *r* voting preference is level *c*.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.

We use the [Chi-Square Distribution Calculator](https://stattrek.com/Tables/ChiSquare.aspx) to find P(Χ2 > 16.2) = 0.0003.

* **Interpret results**. Since the P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

**The General Linear F-Test**

The "**general linear F-test**" involves three basic steps, namely:

1. Define a larger **full model**. (By "larger," we mean one with more parameters.)
2. Define a smaller **reduced model**. (By "smaller," we mean one with fewer parameters.)
3. Use an ***F-*statistic** to decide whether or not to reject the smaller reduced model in favor of the larger full model.

We use the general linear *F*-statistic to decide whether or not:

* to reject the null hypothesis *H*0: The reduced model
* in favor of the alternative hypothesis *H*A: The full model

In general, we reject *H*0 if *F*\* is large — or equivalently if its associated *P*-value is small.

In general, to test that all of the slope parameters in a multiple linear regression model are 0, we use the overall F-test reported in the analysis of variance table.