JMP® ACADEMIC CASE STUDY

JMP018: Mendel's Laws of Inheritance

χ^2 Goodness-of-Fit Test and χ^2 Test for Two-Way Tables

Produced by:

Dr. DeWayne Derryberry
Idaho State University Department of Mathematics

jmp® STATISTICAL DISCOVERY FROM SAS

Mendel’s Laws of Inheritance

$\chi 2$ Goodness-of-Fit Test and $\chi 2$ Test for Two-Way tables

Key ideas

Genetic linkage, independence, laws of probability.

Background

Gregor Mendel began breeding garden peas in about 1857. His scientific training prepared him to focus on experimentation and observation. He began to observe key laws of inheritance that could only be fully explained much later, when it was possible to link reproduction and chromosomes. (At around 1902 the chromosome theory of inheritance began to take form. Historical and biological information was taken from Biology, Fourth Edition, 1996, Neil Campbell.)

Mendel noticed that traits (observable characteristics) of peas, such as color and shape, seem to follow a regular pattern in reproduction. Mendel must have noticed the existence of dominant and recessive traits. A plant carries a pair of traits, and the outward appearance is based on the pair combination.

For example, for some of the peas studied by Mendel the colors were green or yellow. Yellow (Y) is dominant and green (y) is recessive. A plant can have any combination (YY, Yy, yY and yy). But YY, Yy and yY will all appear yellow, while only yy will appear green. In this sense yellow is dominant and green is recessive. The shapes were round (R) and wrinkled (r), with round the dominant trait, given our notation.

Suppose a plant carries the combination (Yy, Rr) and self-pollinates. In this case the seeds get random combinations of the pairs, which occur in approximately equal proportions (there are a lot of seeds). There are 16 possible combinations: YYRR, YYRr, YYrR, YYrr, YyRR, YyRr, YyrR, Yyrr, yYRR, yYRr, yYrR, yYrr, yyRR, yyRr, yyrR, yyrr.

The outward appearance follows the dominance-recessive pattern:

Underlying pair combination (genotype)	**Outward appearance (phenotype)**
YYRR, YYRr, YYrR, YyRR, YyRr, YyrR, yYRR, yYRr, yYrR	YR (Yellow, Round)
YYrr,Yyrr, yYrr	Yr (Yellow, Wrinkled)
yyRR, yyRr, yyrR	yR (Green, Round)
yyrr	yr (Green, Wrinkled)

If every combination is equally likely, we would expect the observed appearances YR, Yr, yR, and yr to appear in a ratio of 9:3:3:1.

This is actually based on a more basic idea. We expect any dominant-recessive pattern to appear as 3:1 or (3/4, 1/4). The reproduction process produces the overall pattern (9/16, 3/16, 3/16, 1/16), which is the multiplication of the probabilities, suggesting some relationship to independence. In other words, outward appearances follow probability rules for independent events: the Pr(Y) = 3/4 and Pr(R) = 3/4, so Pr(Y and R) = 3/4 x 3/4 = 9/16. In biology, this is called independent assortment.

The Task

We will assess whether Mendel's data support the 9:3:3:1 ratio and a second set of data to assess whether other traits display a 3:1 ratio between the dominant and recessive traits.

The Data   There Are Two Data Sets

Mendel1.jmp (two dominant-recessive pairs)

Type	The combination of traits observed in pea seeds (yellow or green, round or wrinkled).
Count	The observed count of seeds displaying each combination of traits.

Mendel2.jmp (several dominant-recessive pairs)

Character	One of a number of observed characteristics.
Trait	The observed appearance of the characteristic.
Dominant	Whether the trait was dominant or recessive.
Count	The observed counts.

The Count column in both JMP data tables has been assigned the "Freq" role.

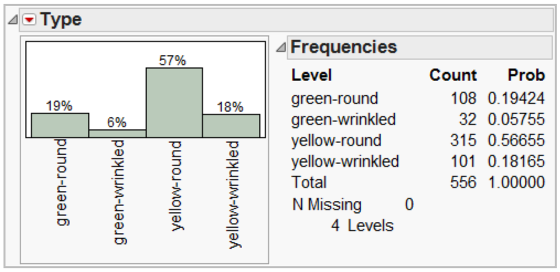
Data are from the following sources:

Peas: Experiments in Plant Hybridization (1865) by Gregor Mendel
The Meetings of February 8th, and March 8th, 1865 (found on Mendelweb, http://www.mendelweb.org/Mendel.html)

Analysis of Mendel1.jmp

We'll start by taking a look at the data (Exhibit 1). Does the yellow-green, round-wrinkled pea data follow the 9:3:3:1 ratio suggested by Mendel's laws?

Exhibit 1 Distribution of Type



Frequencies

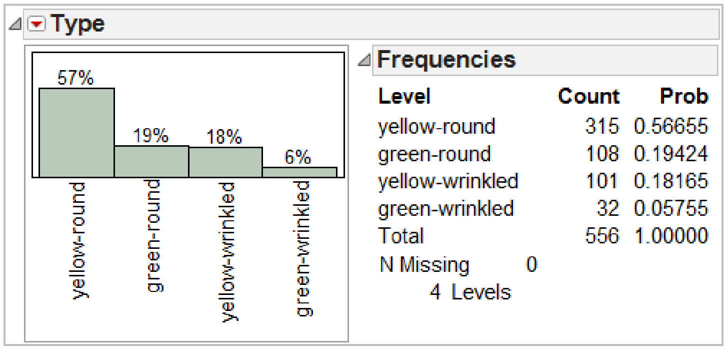
Level	Count	Prob
green-round	108	0.19424
green-wrinkled	32	0.05755
yellow-round	315	0.56655
yellow-wrinkled	101	0.18165
Total	556	1.00000

N Missing 0
4 Levels

(Analyze > Distribution; select Type as Y, Columns. Count, which was assigned the "Freq" role in the data table, appears automatically in the Freq field. Click OK. For a horizontal layout select Stack under the top red triangle. To add percent labels, select Histogram Options > Show Percents from the red triangle for Type.)

Note that the trait combinations are displayed in alphabetical order. It's a little easier to see the breakdown if we re-order the trait combinations (Exhibit 2).

Exhibit 2 Distribution of Type, Ordered by Trait Combination



Level	Count	Prob
yellow-round	315	0.56655
green-round	108	0.19424
yellow-wrinkled	101	0.18165
green-wrinkled	32	0.05755
Total	556	1.00000

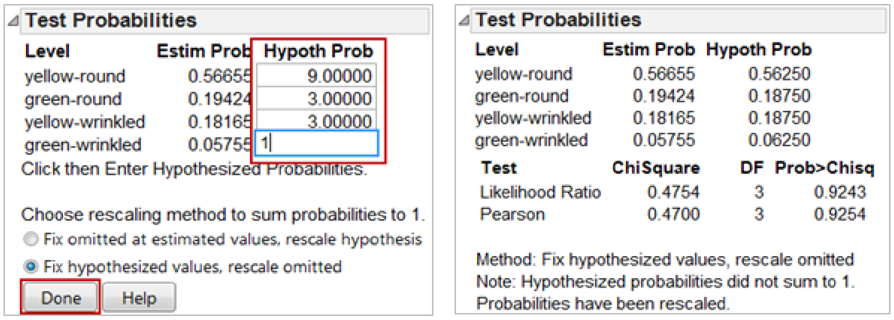
N Missing 0
4 Levels

(To re-order the trait combinations, right-click on the Type column in the data table and select Column Info. Under Column Properties select Value Ordering. Click on the values to move up or down as needed, and click OK.

Return to Analyze > Distribution, and click the Recall button under Actions to pull up previous specifications. Then click OK to re-run the analysis.)

The graph and the counts seem to show consistency with Mendel's Law. A □2 goodness-of-fit test (Exhibit 3) provides a formal test for the 9:3:3:1 pattern.

Exhibit 3 A Goodness-of-Fit Test for the 9:3:3:1 Pattern



Level	Estim Prob	Hypoth Prob
yellow-round	0.56655	9.00000
green-round	0.19424	3.00000
yellow-wrinkled	0.18165	3.00000
green-wrinkled	0.05755	1

Click then Enter Hypothesized Probabilities.

Choose rescaling method to sum probabilities to 1.

○ Fix omitted at estimated values, rescale hypothesis

● Fix hypothesized values, rescale omitted

Level	Estim Prob	Hypoth Prob
yellow-round	0.56655	0.56250
green-round	0.19424	0.18750
yellow-wrinkled	0.18165	0.18750
green-wrinkled	0.05755	0.06250

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.4754	3	0.9243
Pearson	0.4700	3	0.9254

Method: Fix hypothesized values, rescale omitted
Note: Hypothesized probabilities did not sum to 1.
Probabilities have been rescaled.

(Select Test Probabilities from the red triangle next to Type. Then, enter 9, 3, 3, and 1 under the hypothesized probabilities and click Done. The hypothesized probabilities will be rescaled to sum to 1.)

As the test shows, the data is consistent with Mendel's laws (p-value = 0.9243), perhaps too consistent. No discussion of this data would be complete without at least mentioning that R.A. Fisher was suspicious as to how well the data agrees with the proposed model. In this case, the test statistics are both unusually small, and the p-values are unusually large. It is possible for data to be too good to be true!

We only mention this controversy – we have no strong opinions. But something to consider is that Mendel performed many experiments and may have had a lot of data to choose from. If he hand-picked this data from among the various experiments he did, it may be the first known example of data dredging. If this is the case, I doubt that Mendel would have known that it was an error to do so. Data dredging is performing a number of statistical analyses and selecting the ones you like (it is a felony!).

Genetic linkage

The traits we have just analyzed follow the multiplication rule because they are independent. When traits are associated with genes on different chromosomes, they mix at random in reproduction, following Mendel's laws. When two traits are associated with genes on the same chromosome, they tend to appear together more than expected by Mendel's laws. An example of this will be given in the exercises.

The more frequently traits appear together in reproduction, the closer they are on the same chromosome. An understanding of this principle is part of the process of forming linkage maps.

Analysis of the Mendel2.jmp (Several Dominant-Recessive Pairs)

Next we consider the extent to which the data involving several traits displays a common 3:1 ratio indicative on a dominant-recessive ratio. We will perform two tests to form a complete analysis.

We really would just like to combine all of the data and test whether the dominance-recessive ratio is 3:1. However, we must perform a preliminary test to see if it makes sense to combine all of the data (Simpson's paradox is an example of inappropriately combining data). It only makes sense to combine the data if the pattern is the same in each group. So a preliminary test should be conducted to determine whether there is any relationship between Character and Dominance. If there is a relationship, we should not combine all of the data, because the ratios could vary from characteristic to characteristic.

The first test is a χ^2 test with the hypotheses:

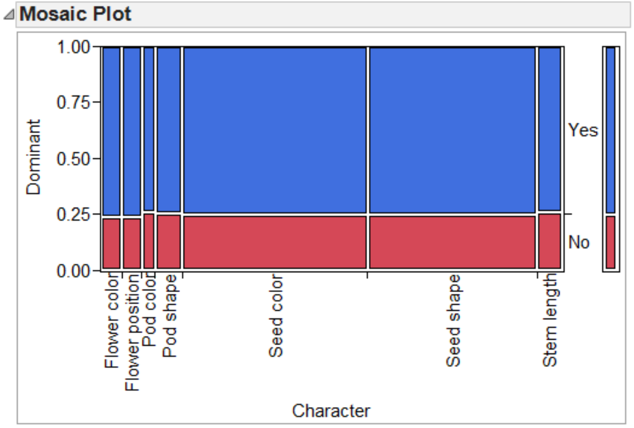
Ho: Characteristic and dominance are unrelated.
Ha: Characteristic and dominance are related.

If we get a large p-value for this test, we will ignore characteristic and test for a 3:1 ratio between dominance and recessive counts.

Ho: Dominance-recessive are in a 3:1 ratio.
Ha: Dominance-recessive are not in a 3:1 ratio.

It is easy to see (Exhibit 4) what we are testing and also why it is important to test the relationship between characteristic and dominance before combining the data. We should combine the data only if breaking the data up by characteristic provides no additional information.

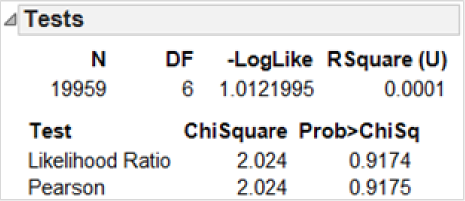
Exhibit 4 Is There a Relationship Between Character and Dominance?



(Analyze > Fit Y by X, select Dominant as Y, Response and Character as X, Factor. Count, which was previously assigned the "Freq" role in the data table, appears automatically in the Freq field. Click OK.)

It appears that the pattern for Dominant (Yes or No) is the same for all characteristics, but we verify with a formal test (Exhibit 5).

Exhibit 5 Formal Test for a Relationship Between Character and Dominance



Tests

N	DF	-LogLike	RSquare (U)
19959	6	1.0121995	0.0001

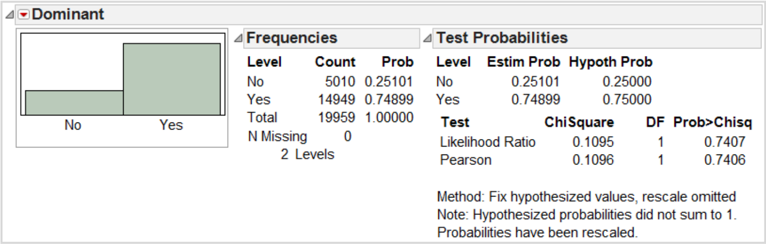
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	2.024	0.9174
Pearson	2.024	0.9175

(Scroll to the bottom of the Fit Y by X analysis window – the test results display by default.)

These tests are consistent with the claim of no relationship between dominance and the characteristic (p-value = 0.9174). As before, the data is a little too good to be true, but let's not explore this further.

We can now perform an analysis ignoring characteristic. This time, we test for a 3:1 ratio between dominant and recessive counts.

Exhibit 6 Test for 3:1 Ratio Between Dominant and Recessive



Frequencies

Level	Count	Prob
No	5010	0.25101
Yes	14949	0.74899
Total	19959	1.00000
N Missing	0	
2 Levels		

Test Probabilities

Level	Estim Prob	Hypoth Prob
No	0.25101	0.25000
Yes	0.74899	0.75000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.1095	1	0.7407
Pearson	0.1096	1	0.7406

Method: Fix hypothesized values, rescale omitted
Note: Hypothesized probabilities did not sum to 1. Probabilities have been rescaled.

(Use Analyze > Distribution; select Dominant as Y, Columns. Count should appear automatically in the Freq field. Click OK. Select Test Probabilities from the red triangle next to Dominant. Then, enter 1 and 4 under the hypothesized probabilities and click Done.)

The data fits a 3:1 model very well (p-value = 0.7407).

Note: We have been flirting with affirming the null hypothesis several times in this analysis. It is often the case that, when using a goodness-of-fit test, people are tempted to affirm the null hypothesis. There are a number of reasons for this: we usually believe the hypothesis is true; we usually have other scientific reasons for believing the hypothesis is true; we usually know of lots of other data sets that seem to support the hypothesis. However, from a statistical point of view, it is more appropriate to simply say that the data is consistent with the null hypothesis (i.e., the model).

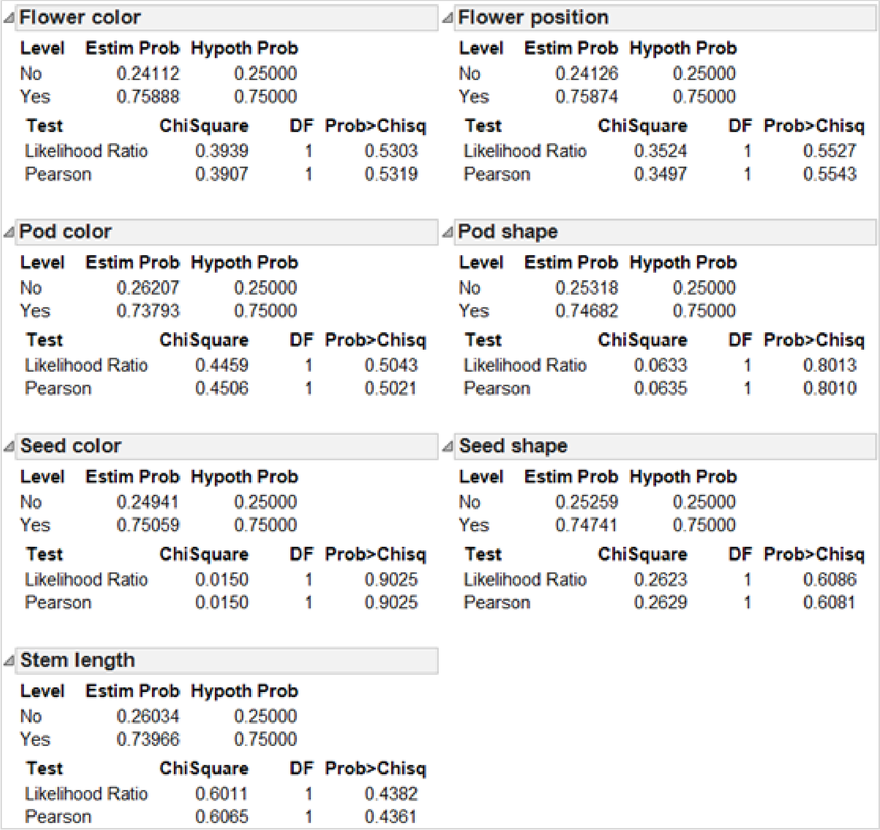
Nevertheless, we have gone further in supporting the null hypothesis in this analysis. We started by performing a test for differences in dominance across the characteristics. The p-value was large, so we then performed a second test. In proceeding to the second test, we acted as if the null hypothesis for the first test were true. Perhaps there is a more rigorous way to approach this problem.

Analysis of the Mendel2.jmp: A More Formal Approach (Optional)

When performing a number of independent χ^2 tests, it is possible to get an overall test statistic by summing the individual χ^2 values and treating the resulting test statistic as a χ^2 with degrees of freedom equal to the sum of the degrees of freedom of all the individual tests.

For example, we could look at the data characteristic-by-characteristic (there are seven of them), test the hypothesis
Ho: The ratio is 3:1 versus Ha: The ratio is not 3:1
for each case, and then combine the resulting test statistics. Each test has a χ^2 with one degree of freedom, so the sum is a χ^2 with seven degrees of freedom.

Exhibit 7 Results of Seven Tests



Flower color

Level	Estim Prob	Hypoth Prob
No	0.24112	0.25000
Yes	0.75888	0.75000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.3939	1	0.5303
Pearson	0.3907	1	0.5319

Flower position

Level	Estim Prob	Hypoth Prob
No	0.24126	0.25000
Yes	0.75874	0.75000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.3524	1	0.5527
Pearson	0.3497	1	0.5543

Pod color

Level	Estim Prob	Hypoth Prob
No	0.26207	0.25000
Yes	0.73793	0.75000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.4459	1	0.5043
Pearson	0.4506	1	0.5021

Pod shape

Level	Estim Prob	Hypoth Prob
No	0.25318	0.25000
Yes	0.74682	0.75000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.0633	1	0.8013
Pearson	0.0635	1	0.8010

Seed color

Level	Estim Prob	Hypoth Prob
No	0.24941	0.25000
Yes	0.75059	0.75000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.0150	1	0.9025
Pearson	0.0150	1	0.9025

Seed shape

Level	Estim Prob	Hypoth Prob
No	0.25259	0.25000
Yes	0.74741	0.75000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.2623	1	0.6086
Pearson	0.2629	1	0.6081

Stem length

Level	Estim Prob	Hypoth Prob
No	0.26034	0.25000
Yes	0.73966	0.75000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.6011	1	0.4382
Pearson	0.6065	1	0.4361

(Use Analyze > Distribution; select Dominant as Y, Columns and Character as By, and click OK. To select Test Probabilities for all characteristics at once, hold the Control key before selecting the option.)

In this case the overall Pearson test statistic is 0.3907 + 0.3497 + 0.4506 + 0.0635 + 0.0150 + 0.2629 + 0.6065 = 2.1389. This produces a p-value greater than 0.95 (using a Chi Square table with seven degrees of freedom). So the data is quite consistent with the model. (We summed the Pearson test statistics, but it would have been sound to sum the likelihood ratio test statistics, and would have resulted in a nearly identical result).

Compared to the previous analysis, this second approach (conducting seven individual tests) is superior to the previous analysis in that it will detect any individual counts that deviate from the model. However, it would be hard to find a data set where the two approaches get different results.

Summary

Statistical Insights

The data considered seems to closely follow the patterns suggested by Mendel. Certain traits, those found on different chromosomes, follow the multiplication rule of independent events. This mirrors the physical reality that the traits combine at random in reproduction.

A second analysis showed that many traits display the classic 3:1 ratio associated with dominant and recessive genes.

(Optional section) The second analysis was also a simple example of meta-analysis.

Implications

When several independent studies test the same set of hypotheses, it is usually possible to combine the test statistics or p-values to produce one overall p-value for the combined data.

JMP® Features and Hints

In this case the Distribution platform was used to test Mendel's law for the two data sets. In the second case, Fit Y by X was first used to test the relationship between Character and Dominance. The "By" field was used to conduct tests for several different characteristics from the Distribution platform.

All graphs display labels in alphanumeric order – Value Ordering was used to change label ordering. The Control key was used prior to selecting Test Probabilities to "broadcast" the command to all characteristics.

Exercises

1. The following data was taken from an old introductory statistics textbook (the original data was cited as collected by E.W. Lindstrom of the Cornell Agricultural Experiment Station and published in 1918). The plant involved was maize.

Characteristics	Type	Probability	Count
green – solid	dominant – dominant	9/16	773
gold – solid	recessive – dominant	3/16	231
green – stripped	dominant – recessive	3/16	238
gold – stripped	recessive – recessive	1/16	59

   a. Enter the data in JMP and remember to preset count to "Freq."
   b. Perform a χ^2 goodness-of-fit test and determine if these data are consistent with Mendel's laws. State a conclusion.
   c. Lindstrom commented that deviations from the Mendel model could be due to the relative lack of vigor or the recessive combinations, especially the double recessive group. The idea being that, while there may have been the proper ratios in the original seeds, fewer seeds of the recessive type survived until they could be classified/observed.

   d. “Exclude” the last row and retest for just the ratio 9:3:3 in the other three groups. State a conclusion.

2. Not all traits are independent. An example taken from *Biology, Fourth Edition*, by Neil Campbell illustrates linkage. We are considering the color and size of wings on fruit flies. Color can be gray (G, dominant) or black (g), and size can be normal (N, dominant) or vestigial (n).

   If a male with gray-normal wings actually carries both dominant and recessive genes and a female has all recessive genes, if traits were independent, all four combinations of outward appearance would be equally likely. In this case the male is (GgNn) appearing GN (gray and normal, also called the “wild type”) and the female is (ggnn), appearing gn (black and vestigial).

Fruit fly type	Probability	Count
gray-normal	0.25	965
black-vestigial	0.25	944
gray-vestigial	0.25	206
black-normal	0.25	185

   a. Enter the data in JMP. Remember to set count to the “freq” role.
   b. Test the model using a goodness-of-fit test. State a conclusion.
   c. (Optional) If you have done similar exercises in your class, verify each type has a probability of 0.25 (25 percent).
   d. Continuing from part c, verify linkage. That is, verify the type of the parents is more prevalent in the data than would be predicted by independence.

3. Test statistics. For the χ^2 goodness-of-fit test the expected cell counts (E_j) are the expected probabilities times the total count, while the observed cell counts (O_j) are just the counts given.

   a. Compute the expected cell counts.
   b. The likelihood ratio test statistic is: $2\sum_{k=1}^{4} \mathrm{O_k} \times \log\,[\mathrm{O_k}/\mathrm{E_k}]$.
   Verify the likelihood ratio statistic given in Exhibit 3.
   c. The Pearson test statistic is: $\sum_{k=1}^{4} (O_k - E_k)^2/E_k$.
   Verify the Pearson test statistic given on the bottom right in Exhibit 3 (0.9254).