Goodness-of-Fit Tests for Discrete Distributions

By Jim Frost — 15 Comments

Discrete probability distributions are based on discrete variables, which have a finite or countable number of values. In this post, I show you how to perform goodness-of-fit tests to determine how well your data fit various discrete probability distributions.

How do you recognize discrete distributions? For discrete distributions, you can create a table that contains all possible values and a non-zero probability for each value. The sum of all probabilities must equal 1. In contrast, continuous distributions are based on continuous variables and have an infinite number of possible values.

The following are examples of different types of discrete distributions.

- **Binary**: For each customer that enters a dealership, there are two possible outcomes—sale or no sale. Each outcome has a probability.
- **Poisson**: The number of cars that a dealership sells in a day can follow the Poisson distribution. You can create a table with the counts (0, 1, 2, 3, etc.) along with the probability for each daily count.
- **Categorical**: The color of the car is a categorical variable. You can list all possible colors along with their probabilities.

Each type of discrete probability distribution requires a different type of data and allows you to model different characteristics. Before you can use these distributions, you need to determine whether your data follows one of them.

Before proceeding, I need to clarify that there are two different approaches based on the type of discrete probability distribution you are testing:

- For binary data, you need to check the assumptions.
- For other types of discrete variables, you need to perform a goodness-of-fit test.

Related posts: Understanding Probability Distributions and Identify the Distribution of Your Continuous Data

Check the Assumptions for Discrete Distributions Based On Binary Data

If you want to use a discrete probability distribution based on a binary data to model a process, you only need to determine whether your data satisfy the assumptions. You don't need to perform a goodness-of-fit test. If you are confident that your binary data meet the assumptions, you're good to go!

I'll walk you through the assumptions for the binomial distribution. You use the binomial distribution to model the number of times an event occurs within a constant number of trials.

The binomial distribution has the following four assumptions:

1. **There are only two possible outcomes per trial.** For example, yes or no, pass or fail, etc.
2. **Each trial is independent.** The outcome of one trial does not influence the outcome of another trial. For example, when you flip a coin, the result of one flip doesn't affect the next flip.
3. **The probability remains constant over time.** For some cases, this assumption is true based on the physical properties, such as flipping a coin. However, if there is a chance the probability can change over time, you can use the P chart (a control chart) to confirm this assumption. For example, it's possible that the probability that a process produces defective products can change over time.

4. **The number of trials is fixed.** The binomial distribution models the frequency of events over a fixed number of trials. If you need to model a different characteristic, use a different distribution.

Typically, you must have good knowledge about the process, data collection methodology, and your goals to determine whether you should use the binomial distribution. If you can meet all four of these assumptions, you can use the binomial distribution.

Other distributions that use binary data

There are several other discrete distributions that use binary data. I list several of them below along with how they differ from the binomial distribution. Each distribution has assumptions or goals that vary a bit from the binomial distribution.

Distribution	Main differentiation from the binomial distribution
Negative binomial	Models the number of trials to produce a fixed number of events.
Geometric	Models the number of trials to produce the first event.
Hypergeometric	Assumes that you are drawing samples from a small population with no replacements, which causes the probabilities to change.

If you are working with binary variables, the choice of binary distribution depends on the population, constancy of the probability, and your goals. When you confirm the assumptions, there typically is no need to perform a goodness-of-fit test.

Example Use of the Binomial Distribution

In a future post, I will show you ways you can use the various discrete probability distributions for binary data. For now, I'll include an example use of only the binomial distribution to give you an idea. The graph below shows us that if the probability of a defective product is 1.5% and you are modeling a sample size of 30, you'd expect just over 60% of the samples to have zero defective products. Additionally, the binomial distribution predicts that about 7.4% of the samples will have two or more defective products.

Example use of the binomial distribution, which is one of the discrete distributions.

Related post: Learn more about the various discrete probability distributions for binary data

Performing a Goodness-of-Fit Test for other Discrete Distributions

If you are working with discrete data that are not binary data, chances are you'll need to perform a Chi-square goodness-of-fit test to decide if your data fit a particular discrete probability distribution. These tests compare the theoretical frequencies to the frequencies of the observed values. If the difference is statistically significant, you can conclude that your data do not follow that specific discrete distribution.

Like any statistical hypothesis test, Chi-square goodness-of-fit tests have a null hypothesis and an alternative hypothesis.

- H_0: The sample data follow the hypothesized distribution.
- H_1: The sample data do not follow the hypothesized distribution.

For goodness-of-fit tests, small p-values indicate that you can reject the null hypothesis and conclude that your data were not drawn from a population with the specified distribution. Consequently, goodness-of-fit tests are a rare case where you look for high p-values to identify candidate distributions.

I'll show you how to test whether your discrete data follow the Poisson distribution and a distribution based on a categorical variable. You can download the CSV file that contains the data for both examples: DiscreteGOF.

Testing the Goodness-of-Fit for a Poisson Distribution

The Poisson distribution is a discrete probability distribution that models the count of events or characteristics over a constant observation space. Values must be integers that are greater than or equal to zero. For example, the number of sales per day in a store can follow the Poisson distribution. If these data follow the Poisson distribution, you can use this distribution to make predictions.

I’ll use an accident count example to show you how to determine whether you data follow the Poisson distribution.

Suppose a safety inspector needs to monitor the number of car accidents per month at a specific intersection. The inspector enters the number of monthly accidents in a worksheet like this:

Example worksheet that contains the number of accidents per month in each cell.

Each value denotes the count of accidents in one month. The actual dataset has 50 values that cover 50 months.

To determine whether these data follow the Poisson distribution, we need to use the Chi-Squared Goodness-of-Fit Test for the Poisson distribution. The statistical output for this test is below.

The statisticals results for the goodness-of-fit test for the Poisson distribution.

This test compares the observed counts to the expected counts based on the Poisson distribution. The p-value is larger than the common significance level of 0.05. Consequently, the test result suggests that these data follow the Poisson distribution. You can use the Poisson distribution to make predictions about the probabilities associated with different counts. You can also use analyses that assume the data follow the Poisson distribution. These analyses include the 1- and 2-sample Poisson rate analyses, and the U Chart.

Categorical Variables and Discrete Distributions

A categorical variable also has a discrete probability distribution of values. Each level of the categorical variable is associated with a probability. To determine whether the distribution of categorical data follows the values that you expect, you can perform the Chi-Square Goodness-of-Fit test. This test is very similar to the Poisson version except that you must specify the test proportions.

I’ll walk you through an example. It is fairly easy to perform this test.

Car color example of a discrete distribution

PPG Industries studied the paint color of new cars bought in 2012 for the entire world. We want to assess whether the distribution of car colors in our local area follows the global distribution. In this example, the PPG data are real but I'm making up our local data. The car color is our categorical variable and the levels are the individual colors.

After gathering a random sample of the color of cars sold in our state, we enter the observed data and global proportions in a worksheet like this:

Worksheet that contains the data for discrete distribution of car colors.

The OurState column contains the tally for each color that we observed. The PPG Industries data are in the Global Proportions column. We'll perform the Chi-square goodness-of-fit test to determine whether our local distribution is different than the global distribution. We'll use the PPG proportions as the test proportions.

The Chi-square goodness-of-fit test results

Chi-squared goodness of fit test results for the discrete distribution of car colors.

This goodness-of-fit test compares the observed proportions to the test proportions to see if the differences are statistically significant. The p-value is less than the significance level of 0.05. Therefore, we can conclude that the discrete probability distribution of car colors in our state is different than the global proportions.

The Contribution to Chi-squared column tells us which paint colors contribute the most to the statistical significance. Gray and Red are the top two colors, but we don't know the nature of how they contribute to the difference.

Let's look at the observed and expected values chart to see how these values are different.

Chart of observed and expected values for the discrete distribution of car colors.

The chart indicates that the observed number of gray cars is higher than expected. On the other hand, the observed number of red cars is less than expected.

Related post: Learn how the Chi-squared test of independence establishes whether there is a statistically significant relationship between categorical variables.

Closing Thoughts

There are several different types of discrete variables than can produce different types of discrete probability distributions. The process by which you test your data to determine whether it follows a specific distribution depends on the type of discrete variable.

In summary:

- **Binary data**: Check the assumptions.
- **Count data**: Use the Poisson Goodness-of-Fit Test.
- **Categorical variable**: Use the Chi-Square Goodness-of-Fit Test and designate the test proportions.

Share this:

Tweet

Related

Understanding Probability Distributions
Understanding Probability Distributions
In "Basics"

Maximize the Value of Your Binary Data with the Binomial and Other Probability Distributions
Maximize the Value of Your Binary Data with the Binomial and Other Probability Distributions
In "Basics"

How to Identify the Distribution of Your Data
How to Identify the Distribution of Your Data
In "Hypothesis Testing"

Filed Under: Hypothesis Testing

Tagged With: analysis example, distributions, interpreting results

Comments

Ali Al-Khafaji says

June 10, 2019 at 10:26 pm

Hi, thank you for this great website! I am using multiple linear regression to predict a certain response. Now, I am trying to conduct statistical significance tests to assess the relative importance of each independent variable on that response. My issue is that I don't know what tests I should use. I have read a quite few papers in my field and still haven't figured it out. They all mention ANOVA, but they don't say how.
What I know is that Regression is used when both the input and output data are continuous. On the other hand, ANOVA is used when the input is discrete and the output is continuous, so how can I use ANOVA to assess the relative importance of a continuous independent variable (the input) and that test is not even made for continuous inputs? Thank you again!!

★ Loading...

Reply

Jim Frost says

June 10, 2019 at 10:49 pm

Hi Ali,

I have written a blog post that covers this topic exactly! Identifying the Most Important Predictors.

I think that post will answer your questions. And, because your study involves regression analysis, you might consider getting my ebook about regression because it covers this topic and many in more detail!

Best of luck with your analysis!

★ Loading...

Reply

phani says
April 18, 2019 at 10:09 pm

Dear Jim,
You are a life saver. Keep the good work coming. I am definitely gonna buy you books !
You make stat look like fun
:)) Cheers!

★ Loading...

Reply

Jim Frost says
April 18, 2019 at 11:31 pm

Thank you so much! I strive to make stats fun, so your comment means a lot to me!

★ Loading...

Reply

Neha says
February 17, 2019 at 10:01 am

While finding the tabulated value of chi square distribution, how to determine the degree of freedom for different distributions in goodness of fit test? ?

★ Loading...

Reply

Sanket Agrawal says
October 4, 2018 at 1:55 pm

Hi, while testing for the goodness of fit, if the count involves people (suppose) or any other entity that should be a whole number, should we round off our expected frequencies to the nearest integer.
I have seen many of my classmates doing so and arguing that number of people should be a whole number, however I am rather skeptical about this approach.
What is your suggestion on this?

★ Loading...

Reply

Jim Frost says
October 5, 2018 at 9:52 am

Hi Sanket,

Personally, I would not bother doing so. The expected counts with decimal places for the Poisson distribution represent the theoretical distribution. Rounding these values of can actually increase or bias the error. I don't see any reason to risk that when you can just leave the values alone!

★ Loading...

Reply

Sanket Agrawal says
October 4, 2018 at 1:45 pm

When testing the goodness of fit for discrete or categorical distributions for example Poisson in this case, do we have to round off the expected frequencies to the nearest integers, since Poisson distribution takes values on the range of positive integers only?
I have seen many of my classmates rounding off the expected frequencies for questions

that involve people's count arguing that they have to be whole numbers, however I am myself skeptical about this proposal.
What do you think about this?

★ Loading...

Reply

Stan Alekman says
January 15, 2018 at 1:57 pm

Thanks for the prompt reply to my question. Sterile drug manufacturers trend the count data they generate for microbial observations. Most of the time, the counts are zero. Occasionally or when there is a problem, counts are observed, which are evidence of contamination. These data are trended following zero inflated models.

★ Loading...

Reply

Stan Alekman says
January 15, 2018 at 1:17 pm

Often discrete data have so many zero values for events, the data do not fit the distributions and are addressed as over-dispersed fits. Can you comment and write about these data sets.

★ Loading...

Reply

Jim Frost says
January 15, 2018 at 1:50 pm

Hi Stan, I've heard about this issue mainly in the context of count data where there are too many zeros for the Poisson distribution. Zero inflated regression models account for this. I do write a little about this problem and potential solution in a post about Choosing the Correct Type of Regression Analysis. I'm not sure if you're working in the regression context. I can add this to my list of things to write about in more detail!

★ Loading...

Reply

Jerry Tuttle, FCAS says

May 18, 2017 at 12:23 pm

Hi. One of the assumptions in the car accidents problem that the data follows a Poisson distribution is that the mean equals the variance. If the variance is greater than the mean, perhaps another distribution should be used such as Negative Binomial.

★ Loading...

Reply

Jim Frost says

May 18, 2017 at 3:02 pm

Hi Jerry, you raise an excellent point, and I agree with it. It's not uncommon to have count data that don't follow the Poisson distribution for the reason you state. The Poisson Goodness-of-Fit test should detect this condition. In other words, if you have count data where the variance is greater than the mean (or less than), you should get a statistically significant test result for the Poisson GOF test. This tells you that your data don't follow the Poisson distribution and you should consider a different distribution.

Jim

★ Loading...

Reply

Ricardo Garza-Mendiola says
May 16, 2017 at 11:01 pm

Hello, can you show me (remember) how do you estimated the expected value to calculate chi-square. Greetings.

★ Loading...

Reply

Jim Frost says
May 17, 2017 at 12:49 am

Hi Ricardo,

Thanks for the great question. The expected value depends upon which Chi-squared test you perform. In both cases, the expected values are the hypothesized values for the null hypothesis. You're testing to see if your actual values are significantly different from the hypothesized values.

For the Poisson goodness-of-fit test, the expected values are based on the Poisson distribution. If your data followed the Poisson distribution exactly, these are the values you'd observed in your data.

For the Chi-squared goodness-of-fit test for the categorical variable, the expected values are based on the values that you specify. In this case, I entered the proportions that PPG found in their study. The software uses these proportions and applies them to the sample size to calculate the expected values.

So, the expected values really depend upon the distribution you are testing.

I hope this helps!
Jim

★ Loading...

Reply

Leave a Reply

Robert Bosch Engineering & Bu

Sorry, you don't have permission to v

Website blocked

Not allowed the use of this Social Network site

WordPress

Need help? Contact our support team at rbeigcitinfra.support@in.bosch.com

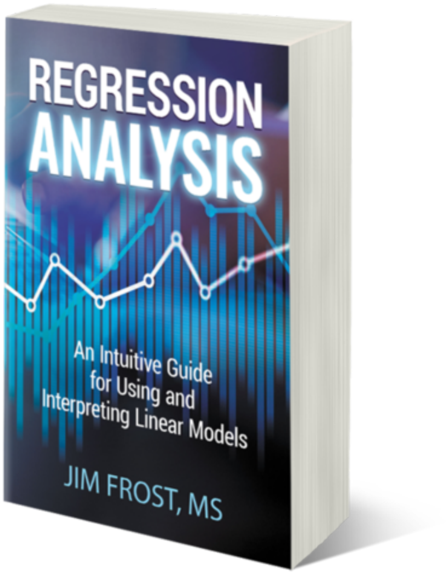
Your organization has selected Zscaler to protect you from i

Meet Jim

I'll help you intuitively understand statistics by focusing on concepts and using plain English so you can concentrate on understanding your results.

Read More…

Buy My eBook!



Search this website

Subscribe via Email!

Enter your email address to receive notifications of new posts by email.

Email Address

Subscribe

Follow Me

Facebook

RSS Feed

Twitter

Popular Latest

How To Interpret R-squared in Regression Analysis

How to Interpret P-values and Coefficients in Regression Analysis

Understanding Interaction Effects in Statistics

How to Interpret the F-test of Overall Significance in Regression Analysis

Measures of Central Tendency: Mean, Median, and Mode

Multicollinearity in Regression Analysis: Problems, Detection, and Solutions

The Importance of Statistics

Recent Comments

Jim Frost on How to Interpret Regression Models that have Significant Variables but a Low R-squared

Jim Frost on The Monty Hall Problem: A Statistical Illusion

Brent on The Monty Hall Problem: A Statistical Illusion

Mohammad Mohaghegh Faghih on Comparing Regression Lines with Hypothesis Tests

pradip on How to Interpret Regression Models that have Significant Variables but a Low R-squared