

JMP Academic Case Study 014

Kerrich: Is a Coin Fair?

Inference for One Proportion

Produced by

Dr. DeWayne Derryberry, Idaho State University
Department of Mathematics

Kerrich: Is a Coin Fair?

Inference for One Proportion

Key Ideas

Practical Importance Versus Statistical Significance, Low Power

Background

John Kerrich was an English mathematician who found himself in a prison camp in Denmark during WW II. This freed up his calendar considerably. To help occupy his time, he performed a series of experiments in probability.

Most statisticians claim that coins are fair, and this gave Kerrich a chance to test this claim. He had the time to flip a coin 10,000 times, resulting in 5,067 heads and 4,933 tails (*The Practice of Statistics in the Life Sciences*, 2nd Edition, Baldi and Moore, Page 209).

The Task

Determine whether this is a fair coin. “Fair” means that heads and tails are equally likely to occur.

The Data **kerrich.jmp**

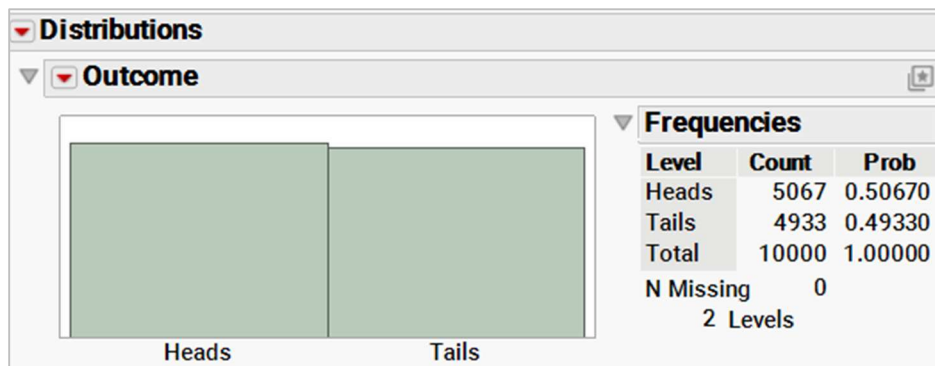
The data set contains the outcome and the counts for each outcome.

Outcome	Heads or Tails
Count	The number of heads and tails in the 10,000 flips

Analysis

We start by exploring the frequency of heads and tails in the 10,000 flips.

Exhibit 1 Distribution of Outcome



(Analyze > Distribution; select Outcome as Y, Columns and Count as Freq, and click OK. For a horizontal layout select Stack under the top red triangle.)

Certainly, the coin *appears* to be fair – there are roughly the same number of heads and tails in the sample.

The 95% confidence interval (Exhibit 2) indicates that the most plausible probabilities of heads range from 0.4969 to 0.5165.

Exhibit 2 Confidence Intervals for Outcome

▼ Confidence Intervals					
Level	Count	Prob	Lower CI	Upper CI	1-Alpha
Heads	5067	0.50670	0.4969	0.516494	0.950
Tails	4933	0.49330	0.483506	0.5031	0.950
Total	10000				

Note: Computed using score confidence intervals.

(From the Distribution output, click on the lower red triangle and select Confidence Interval > 0.95.)

These data are consistent with results we'd see with a fair coin. However, we CAN'T state that the coin is fair - we cannot affirm the null hypothesis.

The hypothesis test is a goodness-of-fit test:

- Ho: The coin is fair.
- Ha: The coin is not fair.

The goodness-of-fit tests (Exhibit 3) both give p-values of 0.1802. We do not have evidence that the coin is not fair. Again, the double negative is required to avoid affirming the null hypothesis.

Exhibit 3 Goodness-of-Fit Test for Outcome

Test Probabilities			
Level	Estim Prob	Hypoth Prob	
Heads	0.50670	0.5	
Tails	0.49330	0.5	
Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	1.7957	1	0.1802
Pearson	1.7956	1	0.1802

Method: Fix hypothesized values, rescale omitted

(From the Distribution output, click on the lower red triangle and select Test Probabilities. Enter 0.5 under Hypoth Prob for both Heads and Tails, and click Done.)

How the Goodness-of-Fit Tests Work

Both the Likelihood Ratio and the Pearson tests compare the observed counts (o_i) to the expected counts (e_i). In this case the observed counts are 5,067 for heads (o_{heads}) and 4,933 for tails (o_{tails}). The expected counts for both heads and tails (e_{heads} and e_{tails}) are 5,000.

You can verify (an exercise) that the likelihood ratio test statistic is:

$$LRT = 2 \left[o_{heads} \times \log \left(\frac{o_{heads}}{e_{heads}} \right) + o_{tails} \times \log \left(\frac{o_{tails}}{e_{tails}} \right) \right]$$

and the Pearson test statistic is:

Power and Sample Size
$$PRT = \frac{(o_{heads} - e_{heads})^2}{e_{heads}} + \frac{(o_{tails} - e_{tails})^2}{e_{tails}}$$

Although a basic analysis of the data in this situation is quite straightforward, it does provide an opportunity to explore some important ideas in a simple context. Let's consider two other data sets, with very different sample sizes.

Exhibit 4 A Small Sample (Small Count) and a Very Large Sample (Large Count)

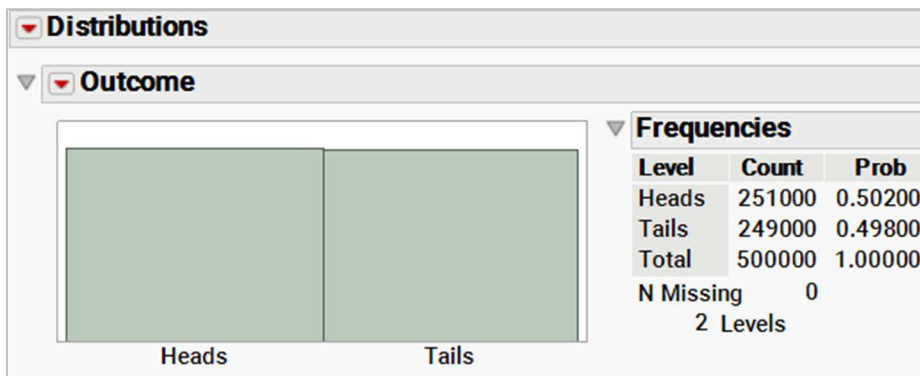
	Outcome	Count	Small Count	Large Count
1	Heads	5067	31	251000
2	Tails	4933	19	249000

For the small data set the sample proportion heads is 0.62, while the sample proportion is 0.502 for the large data set. Which will provide stronger evidence the coin is unfair? Until you become familiar with some aspects of hypothesis testing, the results might be surprising.

Analysis of the Large Sample

Certainly the coin looks fair (Exhibit 5). It is hard to tell from the graph which occurs more often, heads or tails. Heads actually occurred only 50.2% of the time. This coin appears, to the uninitiated, even fairer than the one John Kerrich flipped.

Exhibit 5 Distribution of Outcome, Large Sample



Inference tells a different story. In fact, there is strong evidence for the claim the coin is not fair (Exhibit 6). The p-value is 0.0047!

Exhibit 6 Inference for the Large Sample

Confidence Intervals						Test Probabilities				
Level	Count	Prob	Lower CI	Upper CI	1-Alpha	Level	Estim Prob	Hypoth Prob		
Heads	251000	0.50200	0.500614	0.503386	0.950	Heads	0.50200	0.5		
Tails	249000	0.49800	0.496614	0.499386	0.950	Tails	0.49800	0.5		
Total	500000									
Note: Computed using score confidence intervals.										
						Test	ChiSquare	DF	Prob>Chisq	
						Likelihood Ratio	8.0000	1	0.0047*	
						Pearson	8.0000	1	0.0047*	
Method: Fix hypothesized values, rescale omitted										

Is this a mistake? No. The confidence interval actually supports this claim. The interval estimate is that the coin lands heads, in the long run, between 50.006% and 50.34% of the time. The interval excludes 50%, so the coin is loaded to land heads.

If this seems like a bizarre outcome, then you may not be familiar with the impact of large samples on hypothesis tests. Statistical tests have the power to detect very small deviations from the null hypothesis when sample sizes are large. Here, the null hypothesis is that the coin is fair, and the coin appears not to be fair. Of course, the coin is not very unfair, but it does appear to have a slight bias toward heads. This may or may not be what you really mean by a fair coin. If you and your roommate are using this coin to determine who takes out the garbage, this coin may be perfectly fair. But, if a casino is using this coin to bet on heads, while patrons bet on tails, and there are a 100,000 bets of \$10 per hour, this may make the casino a fortune.

In statistics, the power of a hypothesis test is the probability that the test will reject the null hypothesis when it is false. When sample sizes are large, tests have a lot of power. In fact, when sample sizes are very large, rejecting the null hypothesis may be routine. This gets at the difference between practical importance and statistical significance. Measures indicating practical importance such as confidence intervals, R^2 , etc. may be more important than p-values when sample sizes are large.

Analysis for a Very Small Sample

Now, we consider the results for the small sample.

From the bar chart and frequency distribution in Exhibit 7, it seems like clear evidence the coin lands heads much more than tails. The coin actually landed heads 62% of the time in the sample, compared to 50.67% for Kerrich's data. But, as before, inference will tell a different story.

Exhibit 7 Distribution of Outcome, Very Small Sample



In fact, the p-values are both above the significance level of 0.05, so we cannot reject the null hypothesis that the coin is fair.

Exhibit 8 Inference for the Very Small Sample

Confidence Intervals						Test Probabilities				
Level	Count	Prob	Lower CI	Upper CI	1-Alpha	Level	Estim Prob	Hypoth Prob		
Heads	31	0.62000	0.481504	0.741372	0.950	Heads	0.62000	0.5		
Tails	19	0.38000	0.258628	0.518496	0.950	Tails	0.38000	0.5		
Total	50									
Note: Computed using score confidence intervals.										
						Test	ChiSquare	DF	Prob>Chisq	
						Likelihood Ratio	2.9083	1	0.0881	
						Pearson	2.8800	1	0.0897	
Method: Fix hypothesized values, rescale omitted										

Examination of the confidence interval in Exhibit 8 is again revealing. Because the sample size is small, the confidence interval is very wide. The interval for the long-run estimate of the proportion of the time the coin would land heads is from 0.48 to 0.74. This interval includes values above and below 0.50, so we really do not have enough evidence to rule out a fair coin.

The Moral

Sample size plays a big role in statistical inference. When sample sizes are large we may detect a difference, but the difference may not be of any practical importance. When sample sizes are small we may not be able to detect a difference (between our sample and the null hypothesis) even when it seems that something is going on. Large samples generally mean we have a lot of power to detect differences, while small samples usually result in low power.

An example using some randomly generated data, scatterplots and regression further illustrates this idea.

To generate the data in **Kerrich Power.jmp**:

1. First, 2,000 rows of random normal values (error) were produced using the JMP Formula Editor, with a mean of zero and a standard deviation of 1.
2. Next, x values (x_{big}) were generated. These values are random integers from 1 to 8.
3. Finally, the y values (y_{big}) were generated from the random normal values (errors) and the x values (x_{big}).

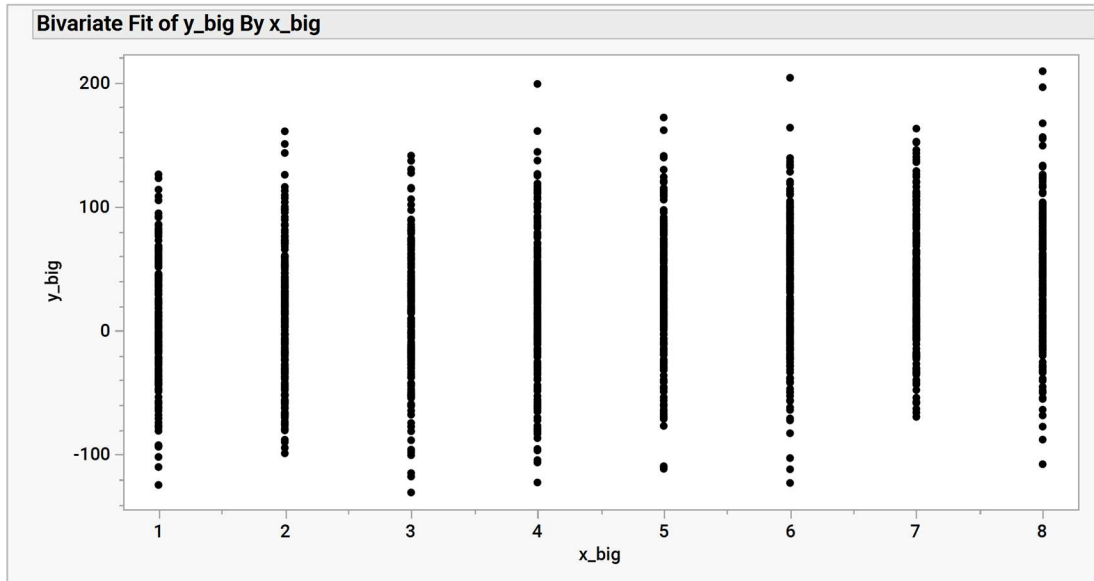
Random Normal()

Random Integer(8)

$2 + 5 * x_{big} + 50 * error$

It may not look like a line would fit this data, but big data sets ($n = 2,000$) produce some surprises. Do you see a trend in the scatterplot in Exhibit 9?

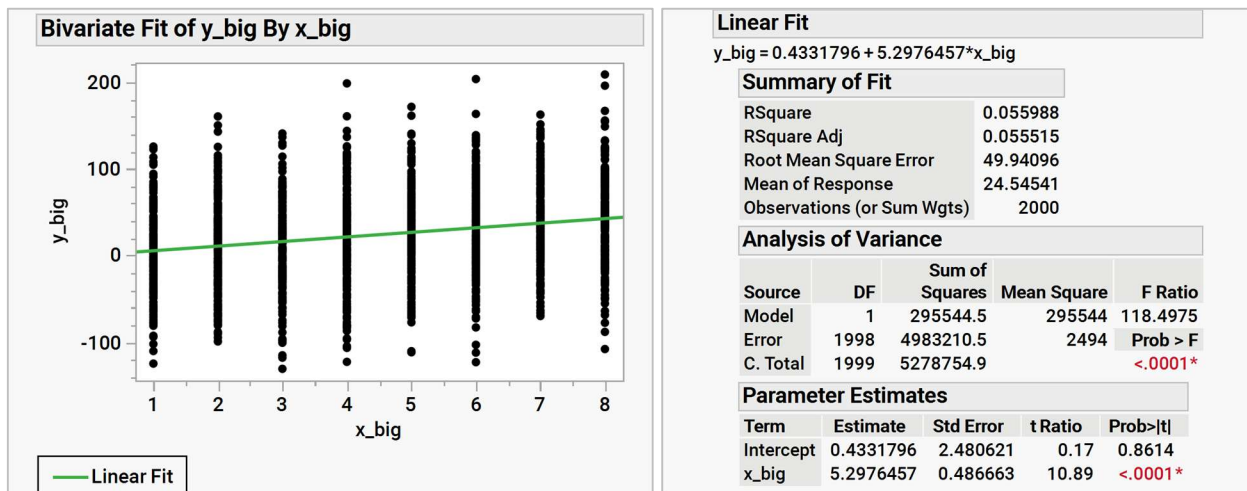
Exhibit 9 Scatterplot for a Large Data Set (y_big and x_big)



(Analyze, Fit Y by X; use y_big as Y, Response and x_big as X, Factor.)

In Exhibit 10, we fit a regression line to the data. Does a line fit the data?

Exhibit 10 Regression for a Large Data Set (y_big and x_big)



(From the Bivariate Fit output, click on the red triangle and select Fit Line.)

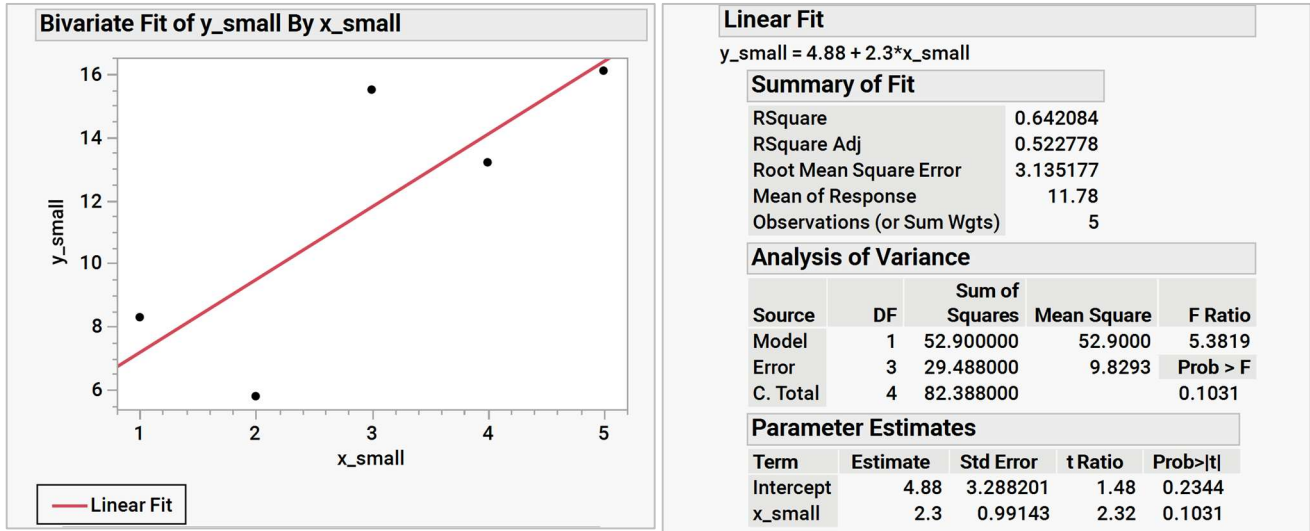
While R^2 is a mere 0.056, indicating only 5.6% of the variation in y_big is explained by the regression using x_big, the p-value is less than 0.0001 (the null hypothesis is no trend in the data, the alternative is that there is a trend in the data).

This is the power of large samples; a tiny trend is detected. However, the trend may not be of any practical importance. Certainly, if this is the rate as which my salary is going up, I would not be impressed!

Now, let's look at a very small sample, with y and x values drawn from the big sample. Columns y_small and x_small contain five observations.

In the plot in exhibit 5, it appears there is a strong linear trend in the data. But, keep in mind that the clear pattern in the data, but the sample size is very small (n = 5).

Exhibit 10 Regression for a Very Small Data Set (y_small and x_small)



I am sure you may have guessed where this is all going.

While the R^2 is 64.2%, indicating that much of the variation in y_small can be explained by the regression on x_small, the p-value is 0.1031. This means there is only weak evidence (at best) that a line fits the data. So what is happening?

With a small data set, while a line does seem to fit the data, we cannot rule out chance as an alternative explanation for the apparent pattern. P-values are a way of assessing the probability that chance could have produced a pattern like this one. Although there appears to be a pattern, there is a 10.04% chance we could have gotten a pattern this strong, or even stronger, if the points were just produced at random with no true relationship between x_small and y_small.

Summary

Statistical Insights

Large samples can detect subtle patterns, sometimes these patterns are not of any practical importance. Very small p-values do not always mean something important is happening, particularly when the sample size is large. Small samples may indicate a relatively clear pattern, but we cannot be sure the pattern is real and not produced by chance variation. Large p-values do not always mean nothing is happening, particularly when the sample size is small.

Implications

Some of the most informative results in science occur when the results seem to contradict what we have stated about power and sample size. Often a pilot study with a very small sample size finds an observable effect. In such cases scientists are often encouraged even though the p-value may not be especially small, because something was found despite the low power of the test. Such results are often worth pursuing further with a larger study.

On the other hand, sometimes several studies involving huge sample sizes fail to find an effect. Although we can never affirm the null hypothesis, when several really large studies all produce no evidence of an effect, at some point we should be very skeptical of the claim.

For example, many (if not most) basketball fans believe some players have hot and cold shooting streaks. A group of cognitive psychologists used several large data sets from the NBA to test various versions of this claim. They found no evidence of streak shooting ("The Cold Facts About the 'Hot Hand' in Basketball," Amos Tversky and Thomas Gilovich, *Chance*, Vol. 2, 1989, Pages 16-21). Although we cannot say streak shooting does not occur (that would mean affirming the null hypothesis), the number of different data sets, the large sample size of the data sets, augmented with their explanation for why people misperceive random patterns combine to produce extreme skepticism toward streak shooting as a phenomenon.

When sample sizes are small (and as a result power is low) finding any evidence of an effect can be important. When sample sizes are large (so power is high) failure to find an effect is worth noting.

JMP® Features and Hints

This case used the Distribution platform to produce bar charts and frequency distributions, create confidence intervals for proportions, and conduct goodness-of-fit tests.

The formula editor was used to generate random data for illustration of the relationship to sample size and power in a regression situation. We used Fit Y by X to generate scatterplots and fit regression lines.

Exercises

1. For the Kerrich coin flipping exercise (Kerrich.jmp), what is the population and what is the sample? Is the sample a simple random sample? If the sample is not a simple random sample, do you think it is a representative sample? Why or why not? Do you think the coin flips are independent?
2. Karl Pearson, a famous statistician for whom the Pearson test statistic is named, flipped a coin 24,000 times, resulting in 12,012 heads. Create a data table in JMP, and perform an analysis of this data similar to that done above. What is your null hypothesis? What conclusion(s) can you draw?
3. In the large sample example we observed 251,000 heads and 249,000 tails. We would have expected to see (for a fair coin) 250,000 heads and 250,000 tails.
 - a. Verify the Likelihood ratio and Pearson test statistics using the formulas given:

$$LRT = 2 \left[o_{heads} \times \log \left(\frac{o_{heads}}{e_{heads}} \right) + o_{tails} \times \log \left(\frac{o_{tails}}{e_{tails}} \right) \right]$$

$$PRT = \frac{(o_{heads} - e_{heads})^2}{e_{heads}} + \frac{(o_{tails} - e_{tails})^2}{e_{tails}}$$

- b. Many textbooks use the following test statistic to perform a t test involving one proportion (for a fair coin):

Ho: $p = p_0$ versus

Ha: $p \neq p_0$

For this test, the test statistic is:

$$z = \frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Where \hat{p} is the proportion of heads in the sample, p_0 is the hypothesized proportion of heads, and n is the sample size.

Compute z and verify $z^2 = PRT$. In other words, confirm that this is just another form of the Pearson test statistic (you can also verify the p-values are the same).