Statistical Significance

An observed event is considered to be **statistically significant** when it is highly unlikely that the event happened by random chance. More specifically, an observed event is statistically significant when its *p*-value falls below a certain threshold, called the **level of significance**. Passing this threshold and achieving statistical significance often marks a decision or conclusion to be drawn from the results of a study.

EXAMPLE

A recent study of a cancer drug showed a 150 basis point increase in overall survival over the control group. This result had a *p*-value of 0.02, which is significant at the 0.05 level. As a result, the drug was approved for further trials.

A p**-value** is the probability that an event will happen that is *as extreme as or more extreme than* an observed event. This probability also comes with the assumption that extreme events occur with the same relative frequency as they do under normal circumstances. Put more simply, a *p*-value can be considered to be a measurement of how *unusual* an observed event is. The lower the *p*-value, the more unusual the event is.

EXAMPLE

The *p*-value of 0.02 in the previous example indicates that the control group (representing what would happen under normal circumstances, without the drug) would only have a 0.02 chance to have the same increase in overall survival rate.

There is a great amount of controversy in the use of *p*-values and statistical significance. This controversy stems partially from the practice of "*p*-hacking," applying bias in selecting data from a study to produce a more significant *p*-value. However, there also exists the much less malicious practice of misusing and misunderstanding statistical significance.

Contents

p-values

p-values come from running experiments and comparing the results to what one would expect under normal circumstances. The language of "as extreme or more

extreme" can be difficult to comprehend, but it becomes much more clear with a simple example.

EXAMPLE

A coin was flipped 5 times, and 4 of the flips were heads. An observer suspects that the coin might be weighted towards heads. Assuming the coin is fair, what is the probability that something as extreme or more extreme would happen again in 5 flips of the same coin?

It is important what the observer suspects, because this informs how to interpret the "as extreme or more extreme" clause. In this case, an event *as extreme* would be flipping 4 heads. An event *more extreme* would be flipping 5 heads.

Thus, the probability of observing an event as extreme or more extreme would be (applying the binomial distribution):

$$p = \binom{5}{4}\left(\frac{1}{2}\right)^5 + \binom{5}{5}\left(\frac{1}{2}\right)^5$$

$$= 0.1875. \square$$

The probability in the previous example can be considered to be a p-value. In particular, it is a **one-tailed p-value**, because extreme events were only considered in one direction (heads). Sometimes, it is more appropriate to consider extreme events in both directions.

EXAMPLE

A coin was selected from a newly-minted batch and flipped 5 times. 4 of the flips were heads. An observer suspects that the newly-minted batch might be weighted towards one side (the weighted side could be different for each coin). Assuming the coins are fair, what is the probability that something as extreme or more extreme would happen again by selecting another coin and flipping it 5 times?

Note how each coin could be weighted towards a different side than the last. Extreme events would be considered in both directions: "4 or more heads" or "4 or more tails." Thus, the probability of observing an event as extreme or more extreme would be twice as much as the previous example:

$$p = 2\left[\binom{5}{4}\left(\frac{1}{2}\right)^5 + \binom{5}{5}\left(\frac{1}{2}\right)^5\right]$$

$$= 0.375. \square$$

The probability in this example can be considered to be a **two-tailed p-value**, because extreme events were considered in two directions. The additional uncertainty that comes from considering both directions causes this *p*-value to be twice as much as the one-tailed *p*-value.

Of course, neither of the probabilities in these examples is particularly concerning. Both probabilities are large enough that the events can be attributed to regular variation. To obtain a more significant, smaller *p*-value, one would need to observe a more extreme event.

EXAMPLE

A six-sided die was rolled 10 times, and 8 of the rolls resulted in 6. An observer suspects that the die is weighted to show the 6 side more than the other sides. What is the probability that an event as extreme or more extreme would happen with 10 rolls of a fair die?

In considering events "as extreme or more extreme," the die would need to show 6 eight or more times. Using the binomial distribution, this probability is:

$$p = \binom{10}{8}\left(\frac{1}{6}\right)^8\left(\frac{5}{6}\right)^2 + \binom{10}{9}\left(\frac{1}{6}\right)^9\left(\frac{5}{6}\right)^1 + \binom{10}{10}\left(\frac{1}{6}\right)^{10}\left(\frac{5}{6}\right)^0$$

$$\approx 1.945 \times 10^{-5}. \square$$

The *p*-value in the previous example is so small that it suggests something suspicious is happening. A reasonable observer would conclude that it is *highly unlikely* that the die is *not* weighted. Even so, the observed result could be attributed to chance, no matter how unlikely that chance is.

Calculating a p-value

- Run an experiment and record the observed event.
- Consider whether to calculate a one-tailed *p*-value or a two-tailed *p*-value. If extreme events are only suspected to happen in one direction, choose a one-tailed *p*-value. Otherwise, choose a two-tailed *p*-value.

- Consider which events would be as extreme as or more extreme than the observed event.
- A *p*-value is a conditional probability; it assumes that extreme events happen with the same relative frequency as what happens under normal circumstances. Given this assumption, compute the probability of events as extreme as or more extreme than the observed event.

TRY IT YOURSELF

A slot machine at a casino is designed to pay out $\frac{1}{5}$ of the games played. During a recent trip to the casino, Estelle played 8 games on a slot machine, but she only won once. She complained to the casino manager, who then explained to her the *p*-value of her experience. What *p*-value did the casino manager tell Estelle? Round your answer to three decimal places.

Reveal the answer

Level of Significance

A challenge in interpreting data with statistics is that a result can *always* be attributed to random chance, even a result with an extremely low *p*-value. Applying a level of significance is a way to set a standard for when to stop attributing results to chance.

A **level of significance**, denoted by α, is a numerical threshold that is compared to a *p*-value. When the *p*-value of an observed event passes below the level of significance, the observed event is considered to be **statistically significant**. Statistical significance often leads to a decision being made or a conclusion being drawn from the results of an experiment.

A level of significance is chosen somewhat arbitrarily, but there are a number of considerations that would affect one's choice.

Considerations in choosing a level of significance

- The most commonly chosen level of significance is $\alpha = 0.05$.
- A smaller level of significance will ensure a more conservative interpretation of the results.
- A smaller level of significance is chosen when an incorrect conclusion can be harmful.
- A smaller level of significance often requires much more collection of data.
- A larger level of significance will ensure conclusions are more easily drawn from the results of an experiment.

- A larger level of significance is chosen when the potential benefits of a conclusion outweigh the potential impacts of an incorrect conclusion.

TRY IT YOURSELF

A group of medical researchers is developing a new flu vaccine. They are about to begin a human trial to test the efficacy of the vaccine. There are a number of concerns about beginning this study:

- If approved, the vaccine will be widely distributed. It is important that the researchers are sure of their results.
- Money is not an object in determining how much data should be collected for their study.
- There is concern among advocacy groups that the vaccine could have some harmful effects. The researchers would like to allay these fears as much as possible with solid data.

Based on these concerns, which is the most appropriate level of significance for the researchers to use in this study?

- ○ $\alpha = 0.001$
- ○ $\alpha = 0.05$
- ○ $\alpha = 0.10$
- ○ $\alpha = 0.25$

Reveal the answer

Hypothesis Tests

Main Article: Hypothesis Testing

p-values and statistical significance are used in hypothesis tests. There are a multitude of different types of hypothesis tests, each with a different way to compute the *p*-value. Below are a few examples.

EXAMPLE

One-sample mean test, population standard deviation known

A nutrition researcher wishes to know if customers of a certain fast food restaurant are a higher weight than average. The researcher selected a random sample of 25 adult males who described themselves as customers of the fast food restaurant. He found their mean weight to be 82 kg. The average weight of an adult male is 78 kg with a standard deviation of 13 kg. Determine if these results are statistically significant at the 0.05 level.

H_0 : The average weight of fast food customers is the same as the populat
H_a : The average weight of fast food customers is more than the populati

The following values are given:

$$\begin{aligned} \text{Sample Mean:} &\quad \overline{x} = 82 \\ \text{Population Mean:} &\quad \mu = 78 \\ \text{Population Standard Deviation:} &\quad \sigma = 13 \\ \text{Sample Size:} &\quad n = 25. \end{aligned}$$

Since the population standard deviation is known, a *z*-score is computed:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{92 - 88}{13/\sqrt{25}}$$

$$\approx 1.54.$$

Looking up this *z*-score on the normal distribution table gives the *p*-value of $p \approx 0.06178$. This is greater than the level of significance, so the result is not statistically significant. The researcher would conclude that customers of the fast food restaurant are not heavier than the population.

EXAMPLE

One-sample proportion test

An industrial statistician is keeping track of the daily number of defects along the assembly line. Today, she discovered 6 defects out of a random sample of 100 products. The expected proportion of defects is 2%. Determine if this result is statistically significant at the 0.01 level.

H_0 : The proportion of defects today is the same as expected.
H_a : The proportion of defects today is more than expected.

The following values are given or computed:

$$\begin{aligned} \text{Sample Proportion:} &\quad \hat{p} = \tfrac{6}{100} = 0.06 \\ \text{Population Proportion:} &\quad p_0 = 0.02 \\ \text{Sample Size:} &\quad n = 100 \\ \text{Population Standard Deviation:} &\quad \sigma = \sqrt{p_0(1-p_0)} = 0.14. \end{aligned}$$

A *z*-score is computed:

$$z = \frac{\hat{p} - p_0}{\sigma/\sqrt{n}}$$

$$= \frac{0.06 - 0.02}{0.14/\sqrt{100}}$$

$$\approx 2.86.$$

Looking up this *z*-score on the normal distribution table gives the *p*-value of $p \approx 0.002118$. This is statistically significant at the 0.01 level. The statistician would conclude that there are more defects than usual today, and would likely feel compelled to do something about it. □

Controversy

The methods outlined on this page are certainly not perfect. As was mentioned before, *any* observed event can be attributed to simple random chance. In spite of all of our exhaustive attempts to remove chance from the analysis as much as possible (by reducing *p*-values as much as possible), the fact that these methods aren't perfect leads to much controversy.

Each year, approximately 2.5 million scholarly research papers are published. With such a vast competition, a researcher certainly does not want to spend extensive time and effort on a study, only to come to the conclusion that their results were not statistically significant. This desire to achieve big, statistically significant results can drive some to manipulate data in an underhanded way. One way to accomplish this is through p-hacking. The practice of **p-hacking** involves sorting through data and looking for a statistically significant pattern before a hypothesis is even drawn up. Then, once a statistically significant pattern is found, the hypothesis is written after the fact.

A potentially bigger issue with statistical analysis is that it is widely misunderstood and misused. Consider the following statements:

1: When the result of a study is statistically significant, it is highly likely that the alternative hypothesis is correct.
2: When the result of a study is statistically significant, it is highly unlikely that the null hypothesis is correct.

Many would not see much difference between these statements. Both statements seem to be expressing the same thing in a slightly different way. However, one of these statements gives the *correct* interpretation of statistical significance, and the other is a common misconception about statistical significance. Can you guess which statement is correct?

Show Answer

Statement 2 is correct, and statement 1 contains a common misconception about statistical significance.

Surprisingly, statistical significance tells us nothing about the accuracy of the alternative hypothesis. It only tells us how unlikely it is for the null hypothesis is to be accurate. Statisticians will often say, "reject the null hypothesis," as opposed to "accept the alternative hypothesis," when a study has a statistically significant result. This language might seem awkward, but it is important to distinguish the meaning of statistical significance.