Hypothesis Testing: (Almost) Everything you need to know

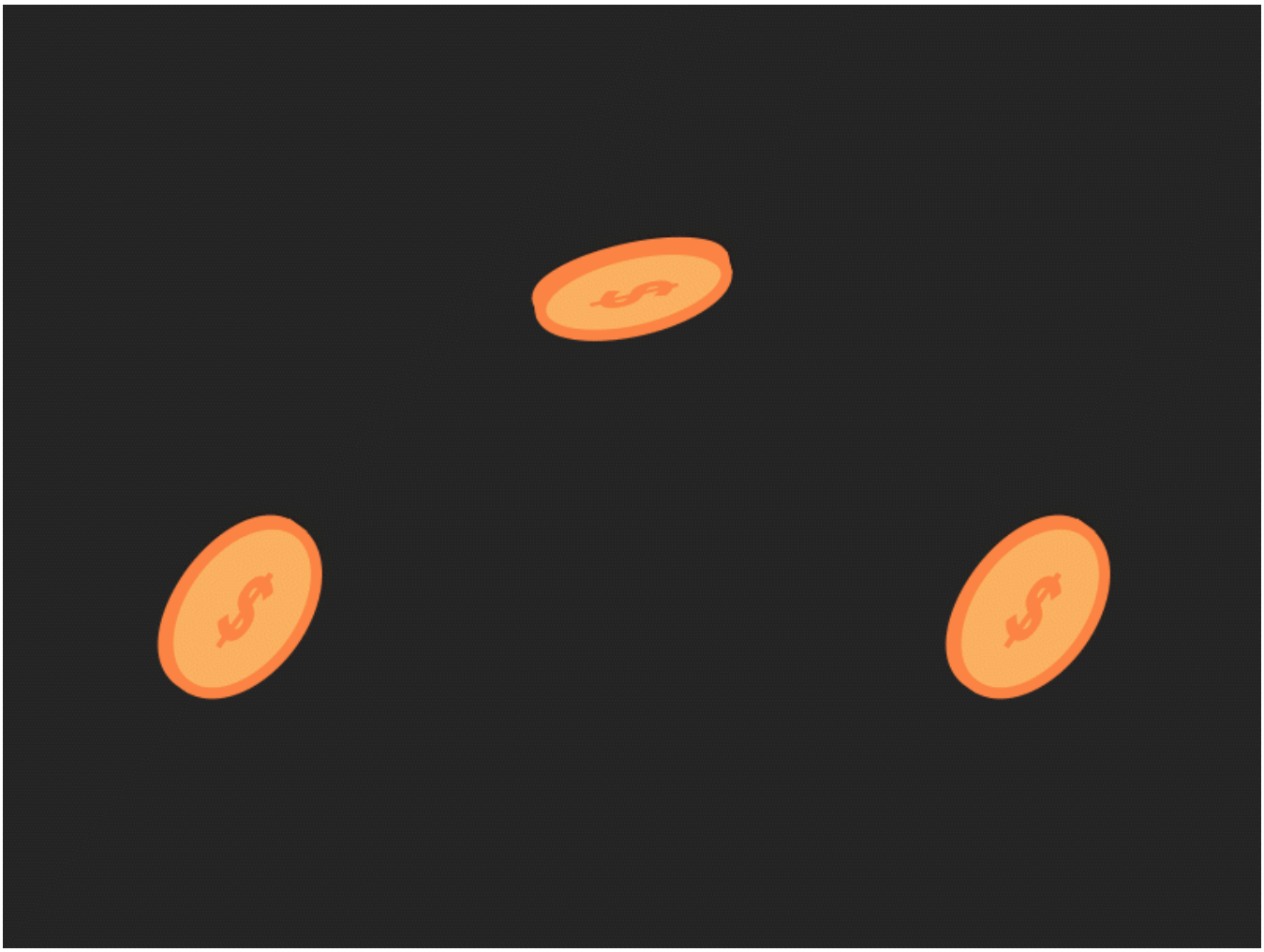
Intuition for hypothesis testing with simple examples

Lance Galletti · Follow

10 min read · Nov 25, 2022

Listen Share

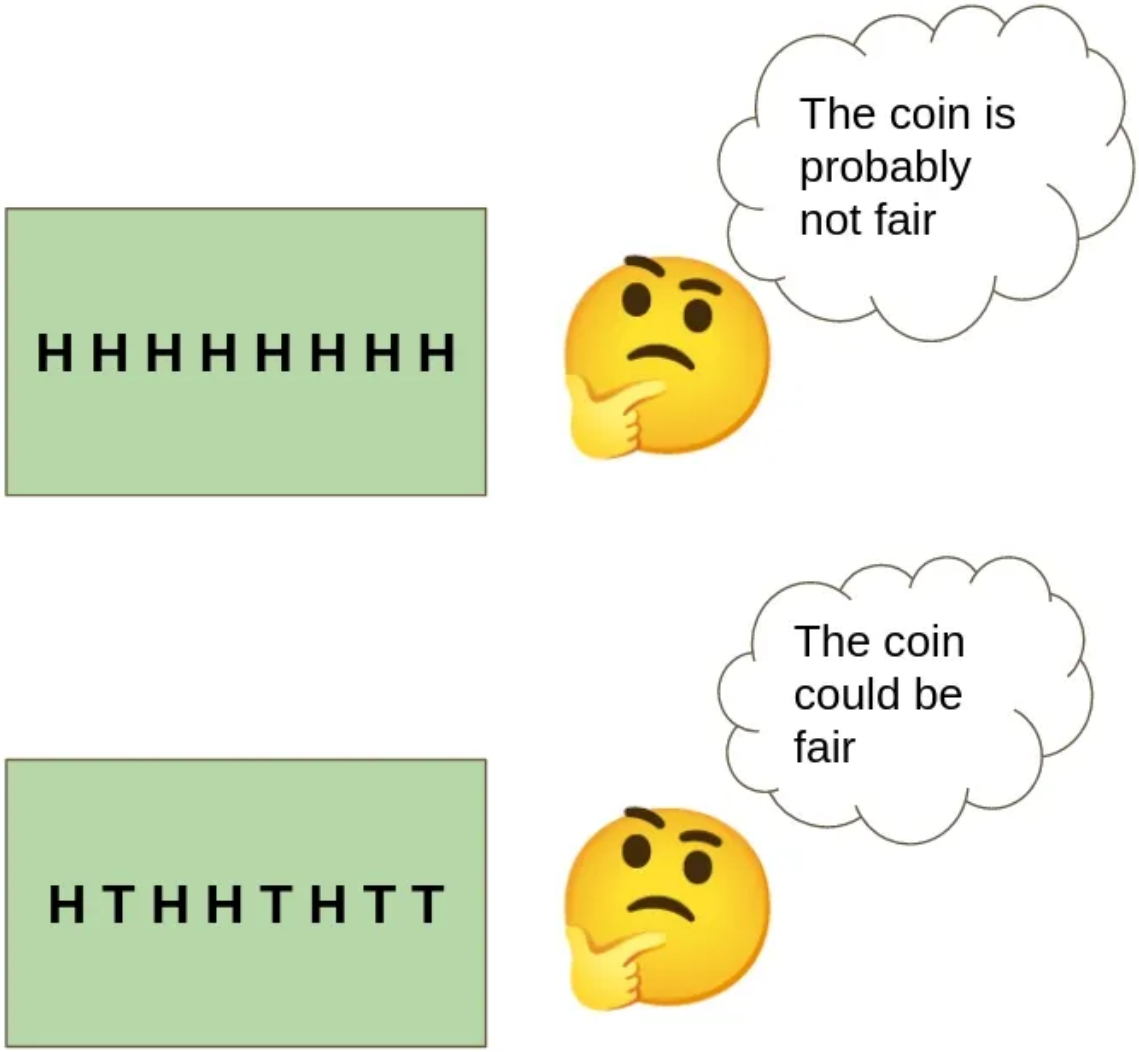
If you need a probability refresher, please read through the following article.



Coin Flip by Paul Boici

Suppose you are given a coin and are asked to test whether this is a fair coin.

If you flip the coin 10 times and it falls on Heads every single time, it should be intuitive that the fairness of the coin should be put into question. Although it's not impossible for a fair coin to land 10 consecutive times on Heads, it's very unlikely (around 0.1% probability). **That unlikeliness should serve as evidence against the hypothesis that the coin is fair** because, under that hypothesis, the likelihood of observing the data we saw (these 10 coin flips all being Heads) is small.



We just went through the following steps (which we will refine and formalize later):

1. Formulate a falsifiable hypothesis: "The coin is fair"
2. Collect data: let's toss the coin 20 times
3. Determine the probability of having observed the data we observed under that hypothesis (this will serve as evidence against the hypothesis)

Reviewers of the experiment can then decide whether or not there is sufficient evidence to reject the hypothesis (by setting a specific threshold for 3).

Let's get into the details in the following examples.

Coin Flips

As above, we are looking to test whether a certain coin is fair. We flip the coin 5 times and get the following output:

```
flips = ['H', 'T', 'T', 'H', 'T']
```

Recall that a coin can be modeled by a **Bernoulli** R.V. characterized by a parameter **p** (the probability of Heads). For any **p**, what is the probability of observing the above dataset? Assuming independent and identically distributed coin flips, it should be:

P(H) P(T) P(T) P(H) P(T) = $p^2 (1 - p)^3$

Our hypothesis of a fair coin is that **p = 1/2**. Under that hypothesis, the probability of observing the above dataset is 1/32 ~ .031. This seems really small... Intuitively, if the coin was fair, seeing 2 Heads out of 5 coin flips seems spot on.

So let's compute that probability of seeing a certain number of Heads! Since there are 5C2 ways of getting 2 Heads out of 5 flips, the probability of seeing 2 Heads out of 5 coin flips is

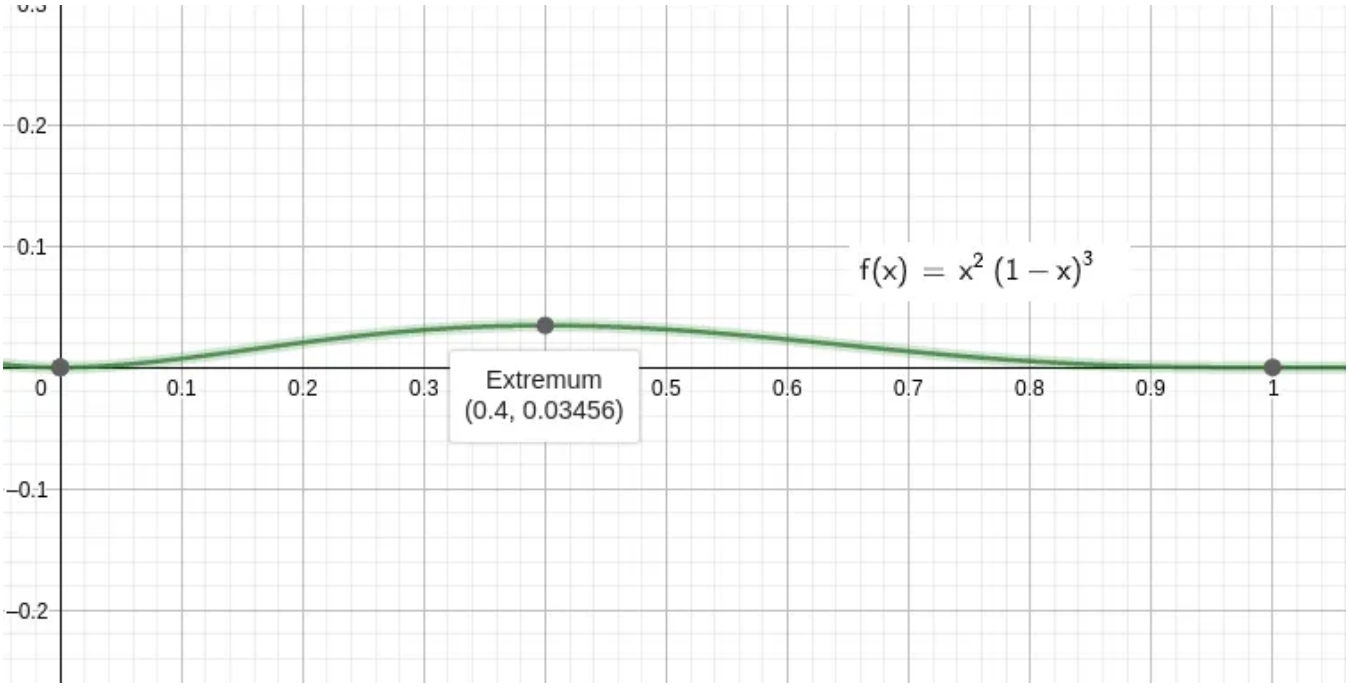
$$\binom{5}{2} p^2 (1-p)^{(5-2)}$$

which is 10 * 1/32 ~ .31 (so about 31%) — much better!

A different perspective

Given the above data, what would be a good estimate for **p**?

The best that we can do is to choose **p** that maximizes the probability of generating that dataset (this is the **Maximum Likelihood Estimation** (MLE) approach). Recall the probability of observing the data observed (assuming independent and identically distributed Bernoulli RVs) is $p^2 (1 - p)^3$.

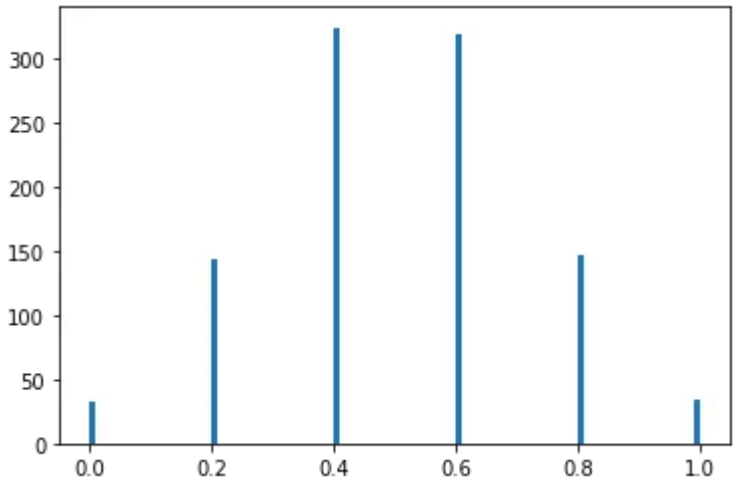


We can find the maximum of that function in the interval [0, 1] by taking the derivative and setting it equal to 0. This gives us **p = 2/5** (which is exactly the sample mean).

We can now ask, what is the probability that, under the hypothesis that the coin is fair and using this MLE approach, we would estimate **p** to be 2/5 on a sample of size 5?

To answer this question we need to know the distribution of the estimates **p** for a sample of size 5. We could try to do the math or we could just run a simulation! Repeatedly generating samples of size 5 using a fair coin, we can record the estimate of **p** at each step:

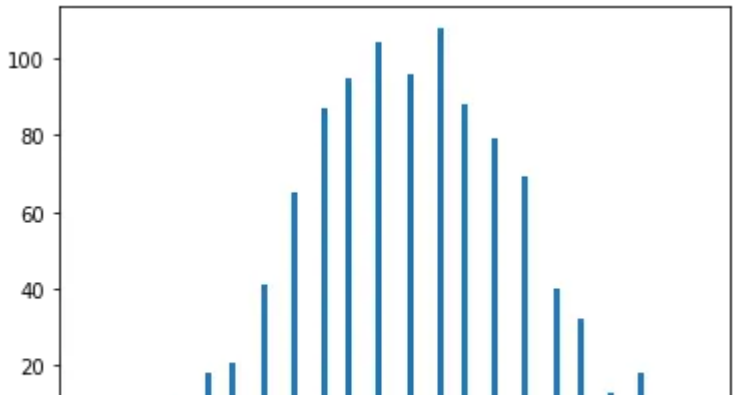
```
SAMPLE_SIZE = 5
p_est = []

for _ in range(1000):
    flips = [np.random.choice([0, 1]) for _ in range(SAMPLE_SIZE)]
    p_est.append(sum(flips) / SAMPLE_SIZE)

fig, ax = plt.subplots()
ax.hist(p_est, bins=100)
plt.show()
```



We can see from this simulation (without having done any math), that about 300 out of the 1000 simulations estimated **p** to be 2/5, which aligns with the previous computation.

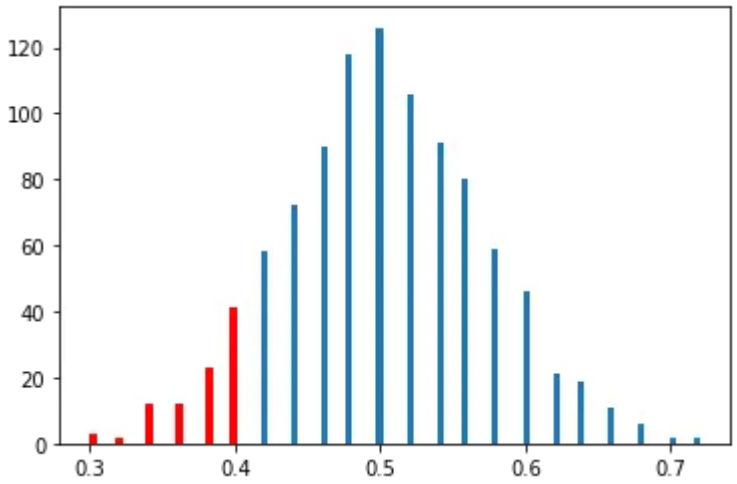
Generalizing

What happens as we increase our sample size (from 5 to say 50) and repeat our simulation?

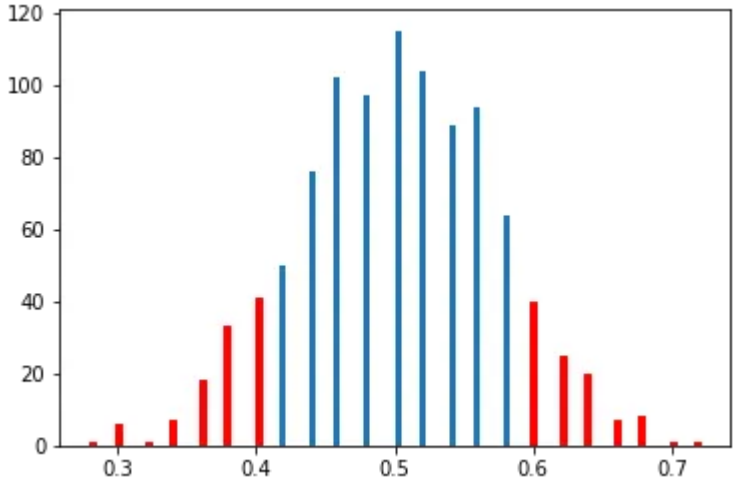


reject the hypothesis that the coin is fair… But we can intuit that seeing about 20 Heads out of 50 doesn't seem THAT impossible for a fair coin.

We also realize that as we increase the sample size, the probability of seeing a specific number of Heads becomes less and less meaningful. Instead of asking what the probability of observing a specific **p** is, we could ask for a range of values that would be considered as evidence against the hypothesis. For example we could ask for the probability that $\mathbf{p} \leq 2/5$ under our hypothesis of a fair coin:



But notice that an unfair coin need not only be defined by a low probability of observing $\mathbf{p \leq 2/5}$ under our fair coin hypothesis. It can also be defined by a low probability of observing $\mathbf{p \geq 3/5}$ under the fair coin hypothesis.



In this way, we are able to answer the question: what is the probability of observing an estimate of $\mathbf{p}$ at least as extreme as the one we observed (here $\mathbf{p = 2/5}$) under the hypothesis that the coin is fair? This probability is know as a p-value.

At this point we may be able to intuit that we not only need to define an initial hypothesis but also an alternate hypothesis used to determine what evidence qualifies. Formalizing the above, we had:

Initial Hypothesis: the coin is fair ($\mathbf{p = 1/2}$)

Alternate Hypothesis: the coin is not fair ($\mathbf{p \neq 1/2}$)

Which would justify gathering evidence for **both $\mathbf{p > 1/2}$ and $\mathbf{p < 1/2}$** under the assumption that the initial hypothesis is true. However we could set up a different

experiment where the alternate hypothesis is that **p < 1/2.** In which case we would only gather evidence for **p < 1/2** under the initial hypothesis.

Doing the math

We would like to accurately and quickly compute the probability (p-value) of observing a value of **p** at least as extreme as the one we observed. For this we need the probability distribution of the estimates of **p** under the hypothesis of a fair coin.

Going back to the fundamentals of probability, we know that the number of Heads in the sample of size **N** follows a Binomial Distribution. So we can compute the probability of observing a number of Heads (**Np**) at least as extreme as the one we observed by evaluating the Binomial Distribution on the range of values that would qualify as evidence against the initial hypothesis.

Using the example above where we are testing for a fair coin and observed 20/50 Heads, that probability would be:

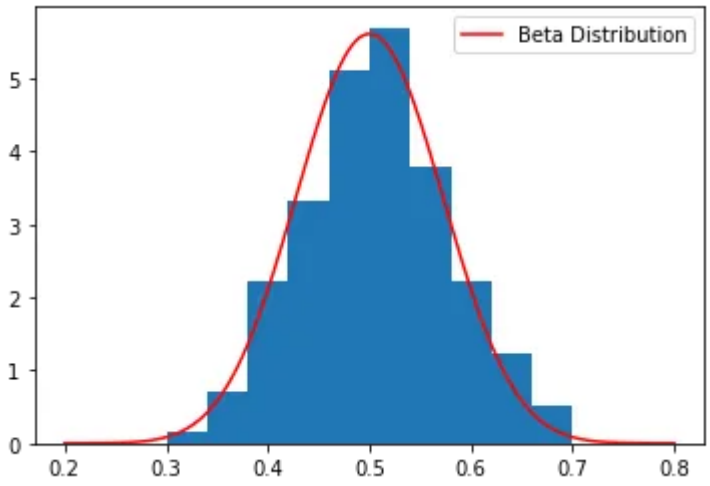
$$\sum_{k=1}^{20} \binom{50}{k} \frac{1}{2}^{k} \left(1-\frac{1}{2}\right)^{50-k} + \sum_{k=30}^{50} \binom{50}{k} \frac{1}{2}^{k} \left(1-\frac{1}{2}\right)^{50-k}$$

which evaluates to about 20%.

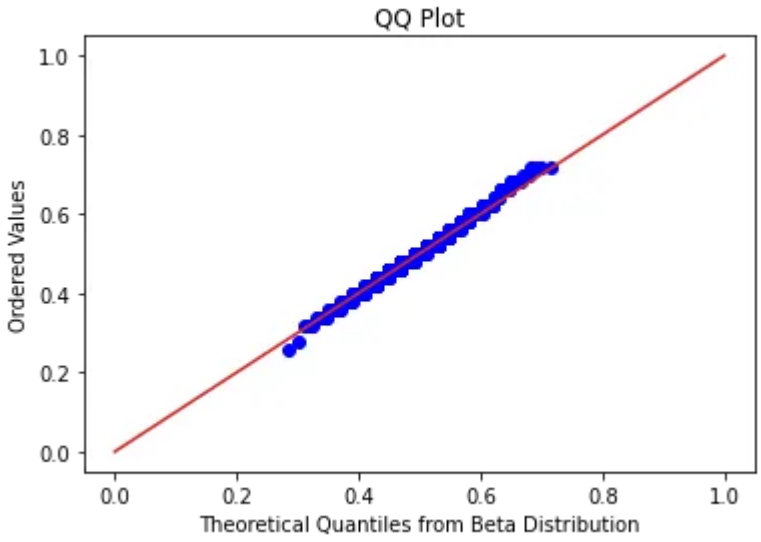
When the sample size is large enough

We can estimate the above probability using other distributions that are easier to compute. For example, we know that proportions are a continuous value, so provided a large enough sample size, we could model the proportion of Heads using a continuous variable. The distribution of the proportion of Heads can be approximated by a Beta Distribution with, in the case of a fair coin, a number of successes **Np** equal to the number of failures.

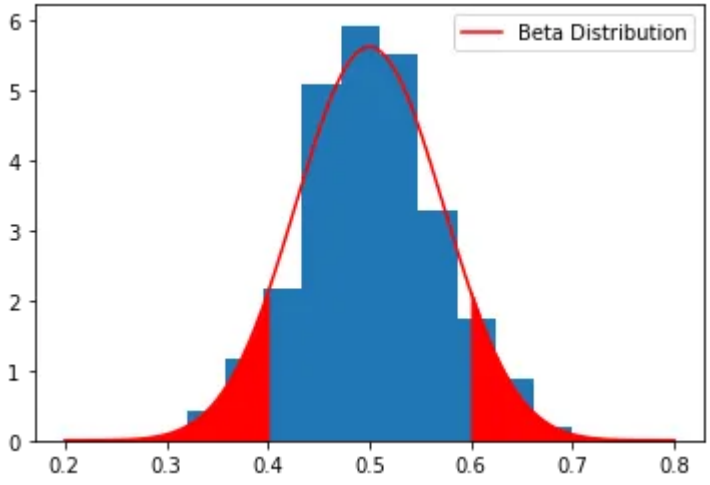
We can visually check that the distribution of the proportion of Heads from repeatedly (1000 times) sampling 50 coin tosses and recording the proportion of Heads follows a Beta distribution:



Another way would be to plot the Quantiles from the sample of 1000 proportions against the Quantiles of the Beta Distribution in what is called a QQ plot:



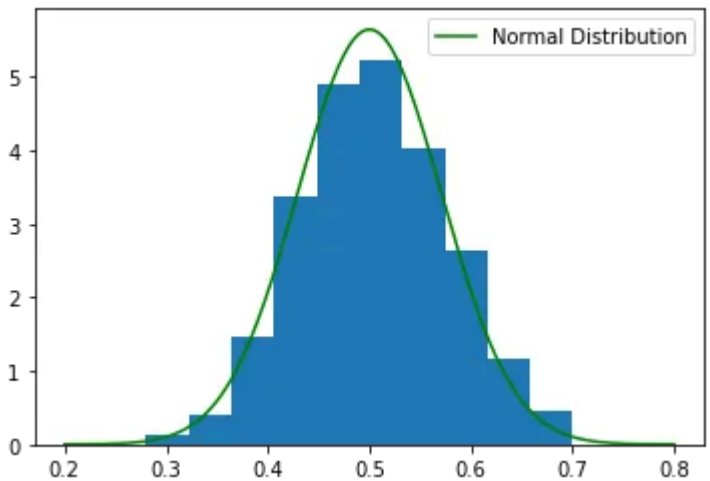
We can then answer the question “what is the probability that the proportion of Heads is less than 2/5 or greater than 3/5 in a sample of size 50?” by evaluating the area under the curve:



which gives us an estimated p-value of about .16 for a sample size of 50. We expect that the estimated p-value would get closer to the true p-value as the sample size increases.

Another option is to use the **Central Limit Theorem** that tells us that the distribution of the mean of **N** independent and identically distributed random samples can be approximated by a Normal Distribution (regardless of whether the samples come from a Normal Distribution) as **N** becomes large.

```
SAMPLE_SIZE = 50
p_est = []
p = 1/2
stdev = (p * (1 - p) / SAMPLE_SIZE) ** (1/2)

for _ in range(1000):
    flips = [np.random.choice([0, 1]) for _ in range(SAMPLE_SIZE)]
    p_hat = sum(flips) / SAMPLE_SIZE
    p_est.append(p_hat)

xs = np.linspace(.2,.8,1000)
fig, ax = plt.subplots()
ax.hist(p_est, density=True)
ax.plot(xs, norm.pdf(xs, p, stdev), color='green', label='Normal Distribution')
ax.legend()
plt.show()
```



Evaluating the area under the curve gives us an estimated p-value of .16.

Here is some code you can use to see how the p-values and estimated p-values change for different sample sizes, and estimates of **p** under the fair coin hypothesis:

```
from scipy.stats import norm, beta, binom

SAMPLE_SIZE = 50
p_hat = 2/5
p = 1/2
var = p * (1 - p) / SAMPLE_SIZE

print("p-value = ", 2 * sum([binom.pmf(k, SAMPLE_SIZE, p) for k in range(1, int
print("p-value estimate from Beta Distribution = ", 2 * beta.cdf(p_hat, p * SAM
print("p-value estimate from Normal Distribution = ",2 * norm.cdf(p_hat, p, (va
```

Hypothesis Testing Formalized

The above followed Fisher's approach (in contrast to Neyman–Pearson's) to hypothesis testing (often called significance testing), which can be summed up as follows:

1. Formulate a falsifiable hypothesis (called a null hypothesis)
2. Define an alternate hypothesis that defines what will qualify as evidence against the null hypothesis
3. Collect data
4. Report the *exact* probability of observing data at least as extreme as that observed in 3 under the assumption that the null hypothesis is true. This is known as a p-value.

The p-value should be reported exactly and not as "p < .01" which is sadly prevalent in scientific literature. The reported p-value represents the strangeness of the data observed conditional on the null hypothesis. Note that p-values are not to be confused with the probability of either hypothesis being true conditioned on the data observed:

$$P(\text{observed data} \mid H) \neq P(H \mid \text{observed data})$$

Caution should be exercised when making decisions based on the above p-value. Close attention should be paid to things like the set up of the experiment, prior experiments / knowledge — the p-value is only a small piece of what should be used for inference / decision making.

On the one hand, there may be bias post-experiment, to favor a stronger or weaker threshold with which to reject the hypothesis. On the other hand, there should never be a single, standard threshold with which all hypotheses are rejected (i.e. a golden rule like $p < .01$). This is because all experiments are different and while we can say that a smaller p-value represents stronger evidence against the null hypothesis than larger ones, this is only the case for similar experiments. In general, with different experiments, on different populations, with different sample sizes etc. we cannot say that a p-value of .0123 in one experiment represents the same strength in evidence against the null hypothesis as a p-value of .0123 in another experiment.

Conclusion

The main criticism of Fisher's approach is that it pays too much attention to the outcome of the experiment and not enough attention to the experiment itself (and what outcomes it would generate long term if repeated many times).

A different approach to hypothesis testing is the Neyman–Pearson approach which aims to design experiments based on pre-defined error thresholds such that, long term, the conclusions of the experiment are not too often wrong were we to repeat the experiment many times. In Fisher's approach, one can set a threshold α with which to conclude rejecting the null hypothesis given the reported p-value (a property of the experiment's outcome). This approach defines α upfront as a property of the procedure / experiment. If α defines the threshold at which the null hypothesis is rejected in an experiment, it also defines the probability of falsely rejecting the null hypothesis (also known as a type I error). Then, β can define the probability of falsely failing to reject the null hypothesis (also known as type II error). The experiment is then carefully designed to match the defined α and β.

Hypothesis Testing | Probability | Statistics | Statistical Significance | P Value

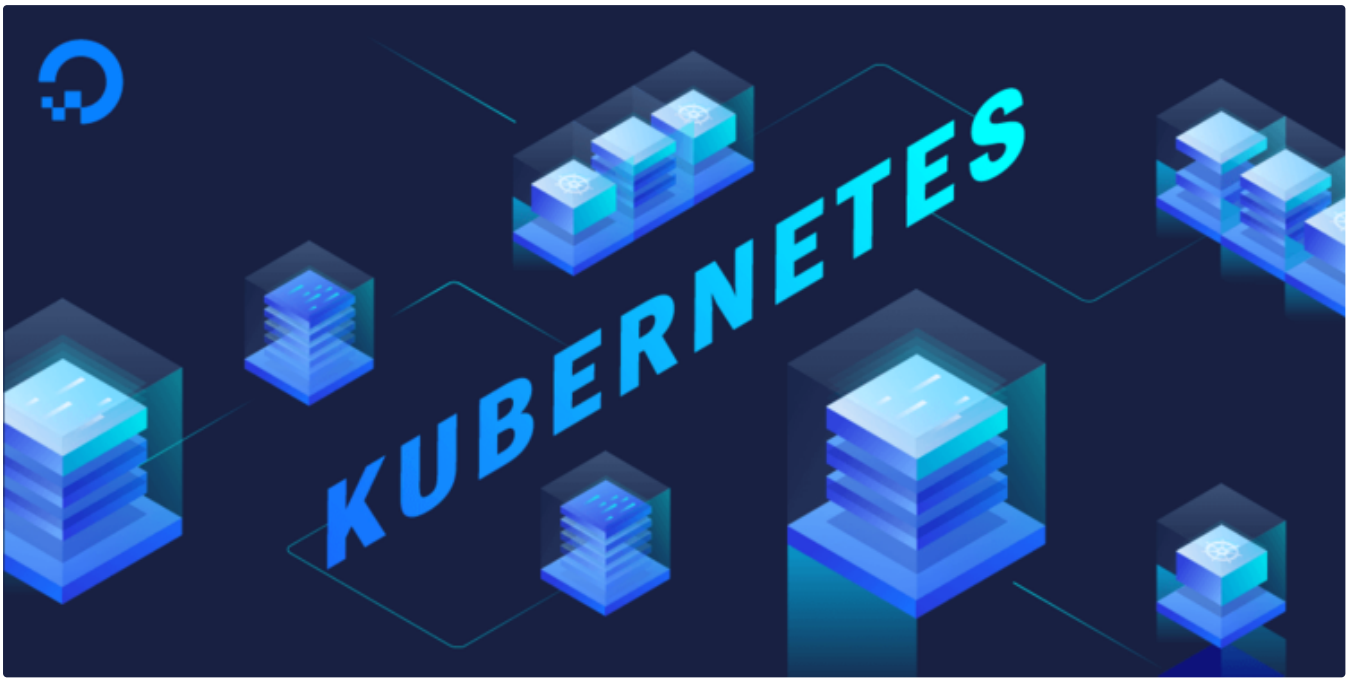
Follow

Written by Lance Galletti

94 Followers

Lecturer @Boston University. Principal Software Engineer @Red Hat. Always refining.

More from Lance Galletti



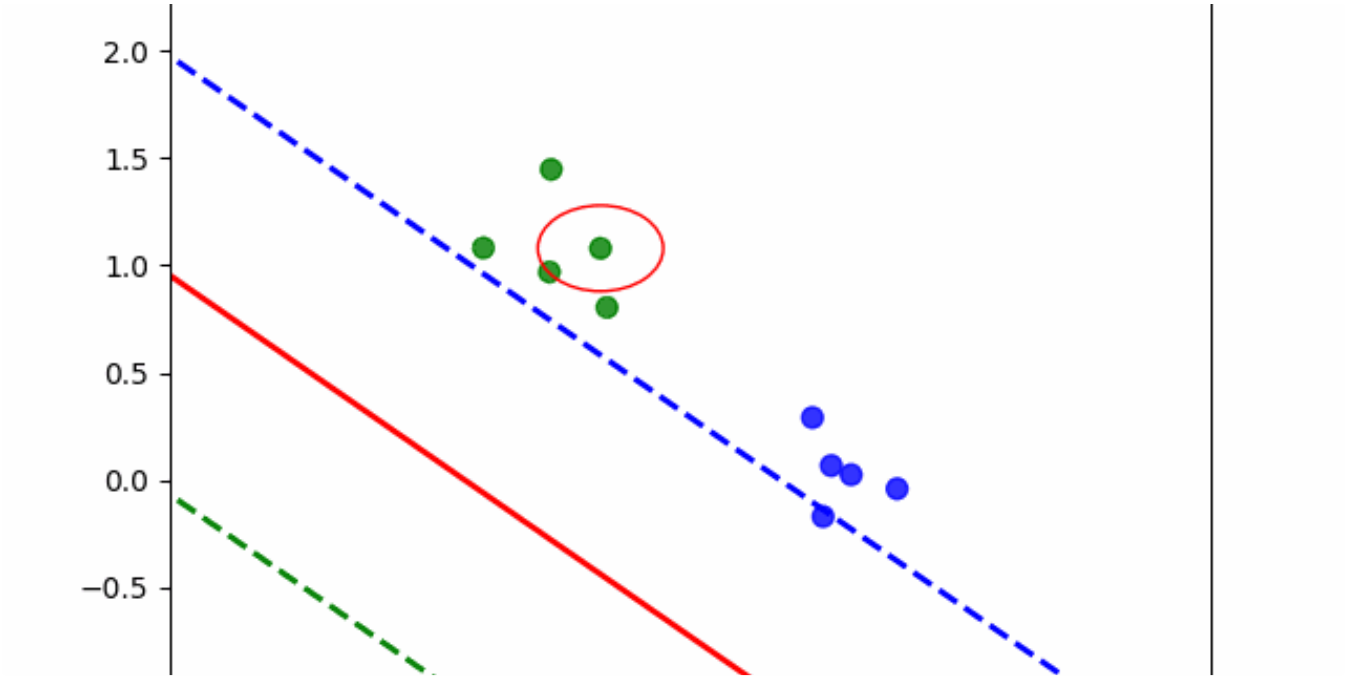
Lance Galletti

10 Things You Should Know Before Writing a Kubernetes Controller

The controllers in this article respond to APIs defined by Custom Resource Definitions (as opposed to Aggregated API servers). We will be...

12 min read · Dec 6, 2021

56 3



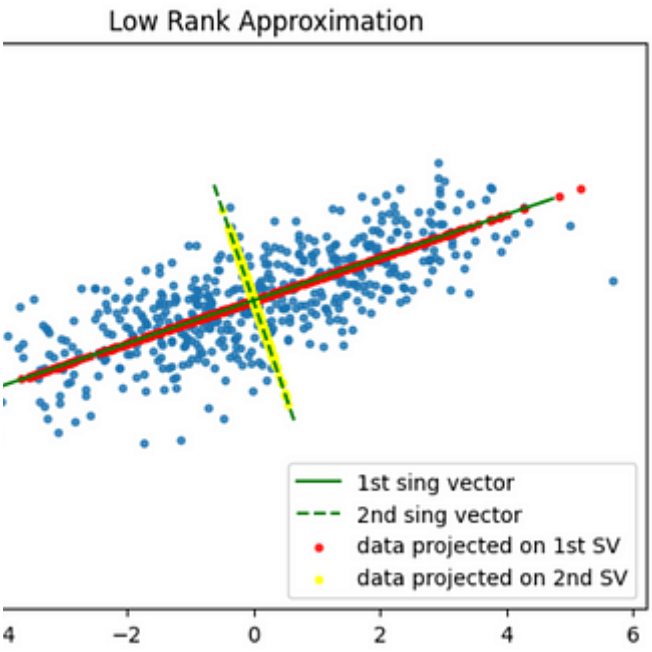
Lance Galletti in MLearning.ai

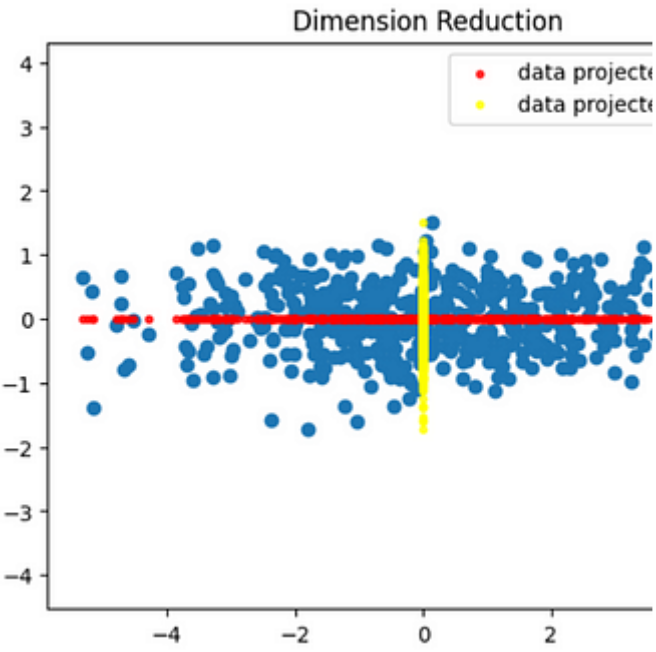
Support Vector Machines From Scratch

Using the perceptron algorithm

5 min read · Oct 31, 2022

89



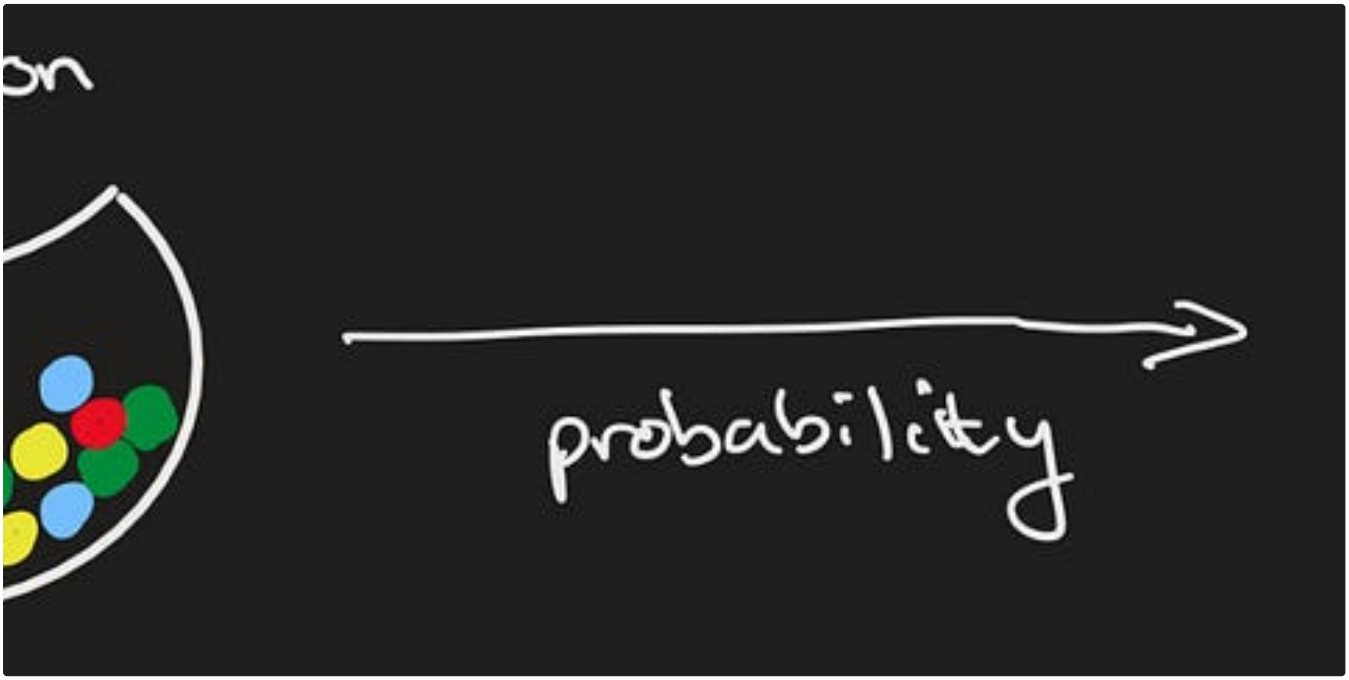


Lance Galletti

Singular Value Decomposition

A simplified walk-through of one of the most foundational tools in Data Science

8 min read · Oct 18



Lance Galletti

Probability: (Almost) Everything You Need To Know

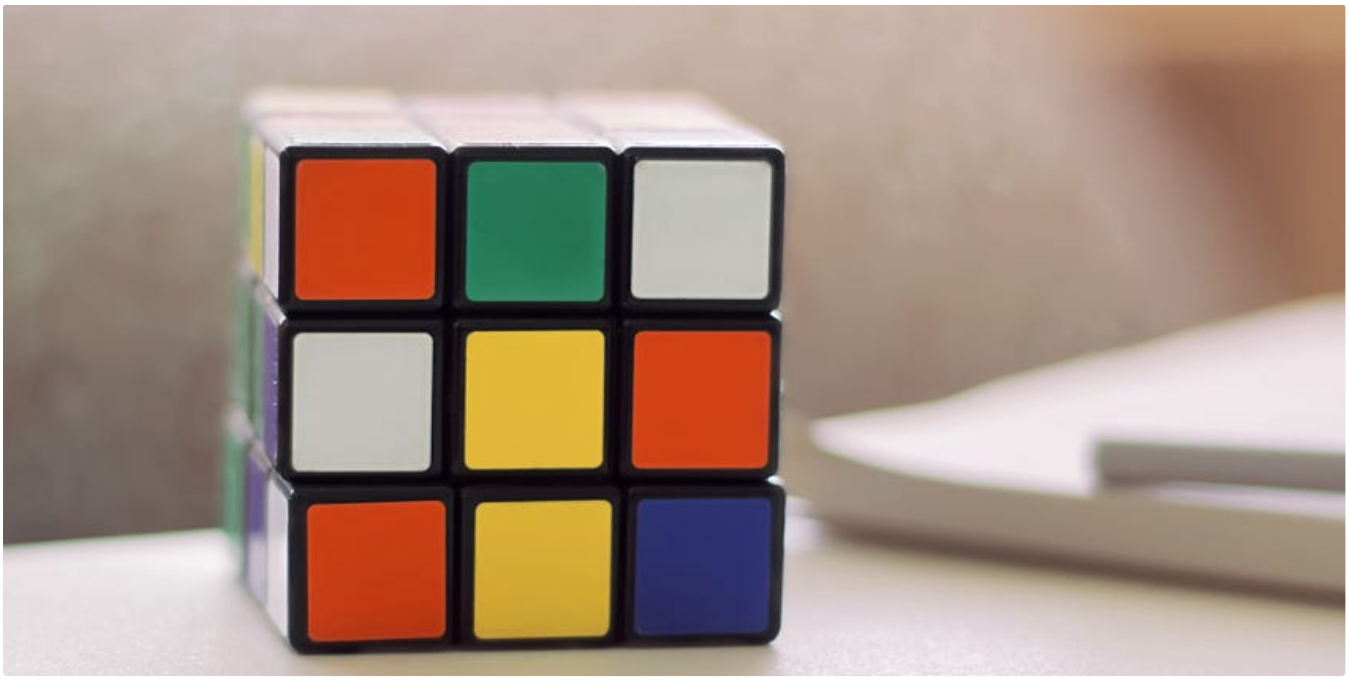
Intuition for common probability concepts in a continued example style

17 min read · Oct 7, 2021

44

See all from Lance Galletti

Recommended from Medium



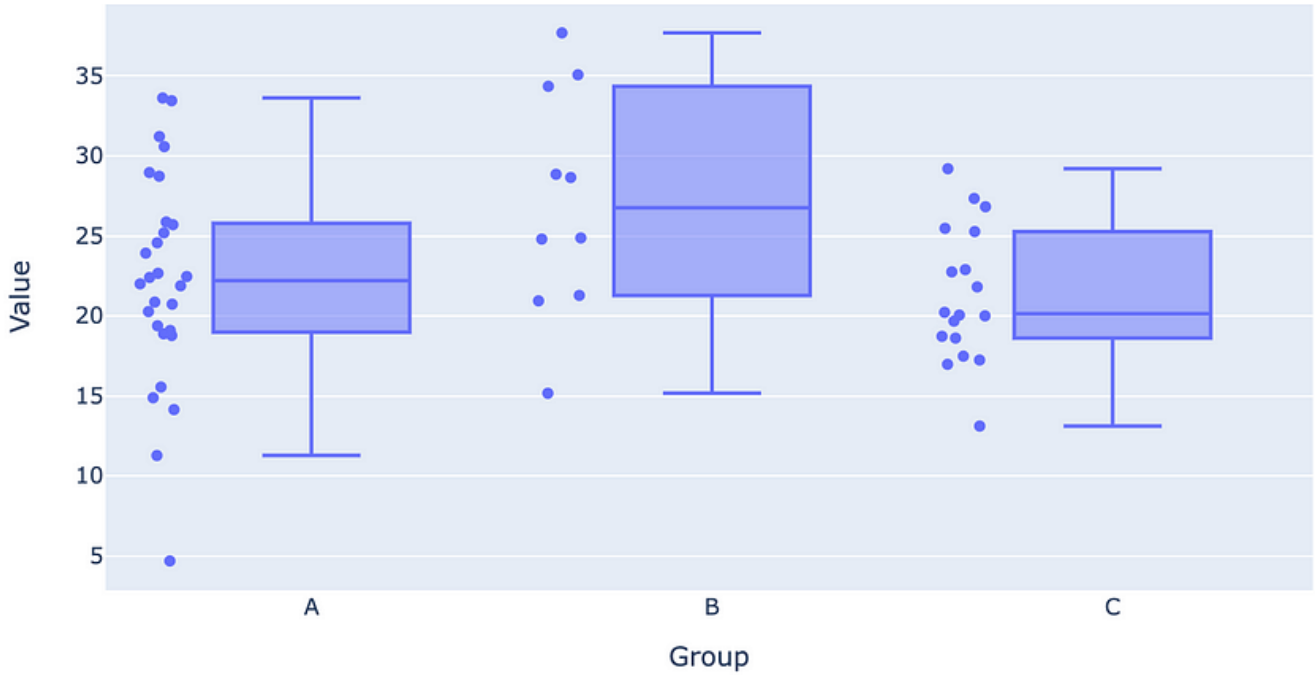
Mehul Gupta in Data Science in your pocket

Permutation testing explained with an example

Moving beyond hypothesis testing

4 min read · Jun 6

90



Dave Currie

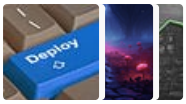
Statistical Stories - ANOVA (ANalysis Of VAriance)

Explaining how the ANOVA statistical test works using a story, its assumptions, formula, alternative tests, and some use cases.

✦ · 5 min read · Oct 5

57

Lists



Predictive Modeling w/ Python

20 stories · 595 saves

Practical Guides to Machine Learning

10 stories · 675 saves

Business

38 stories · 41 saves

Medium Publications Accepting Story Submissions

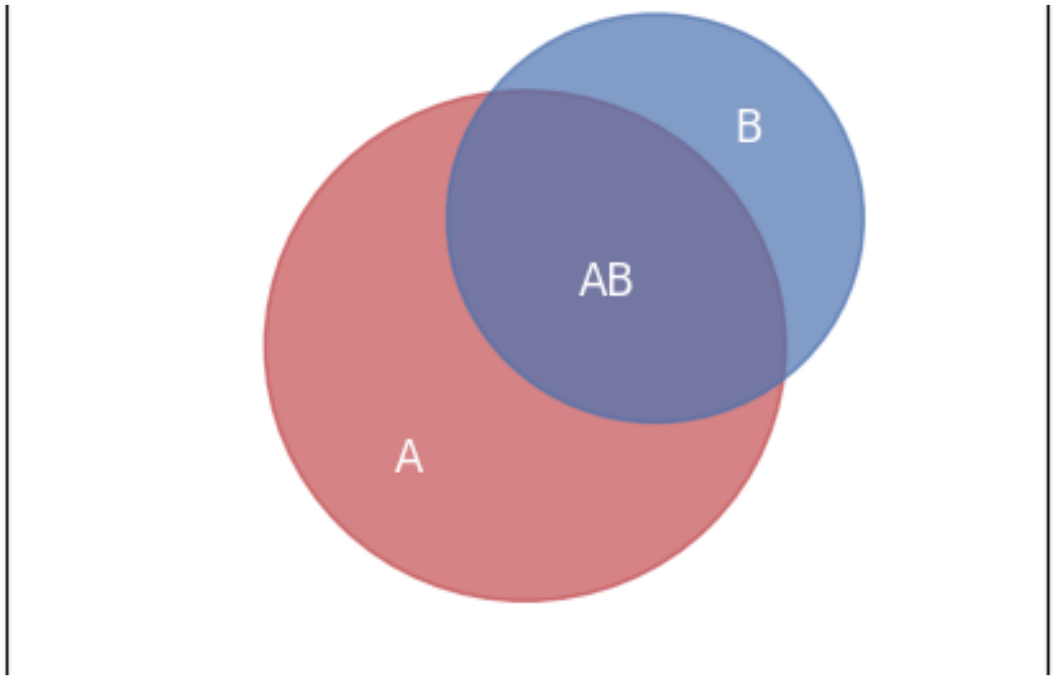
154 stories · 1004 saves

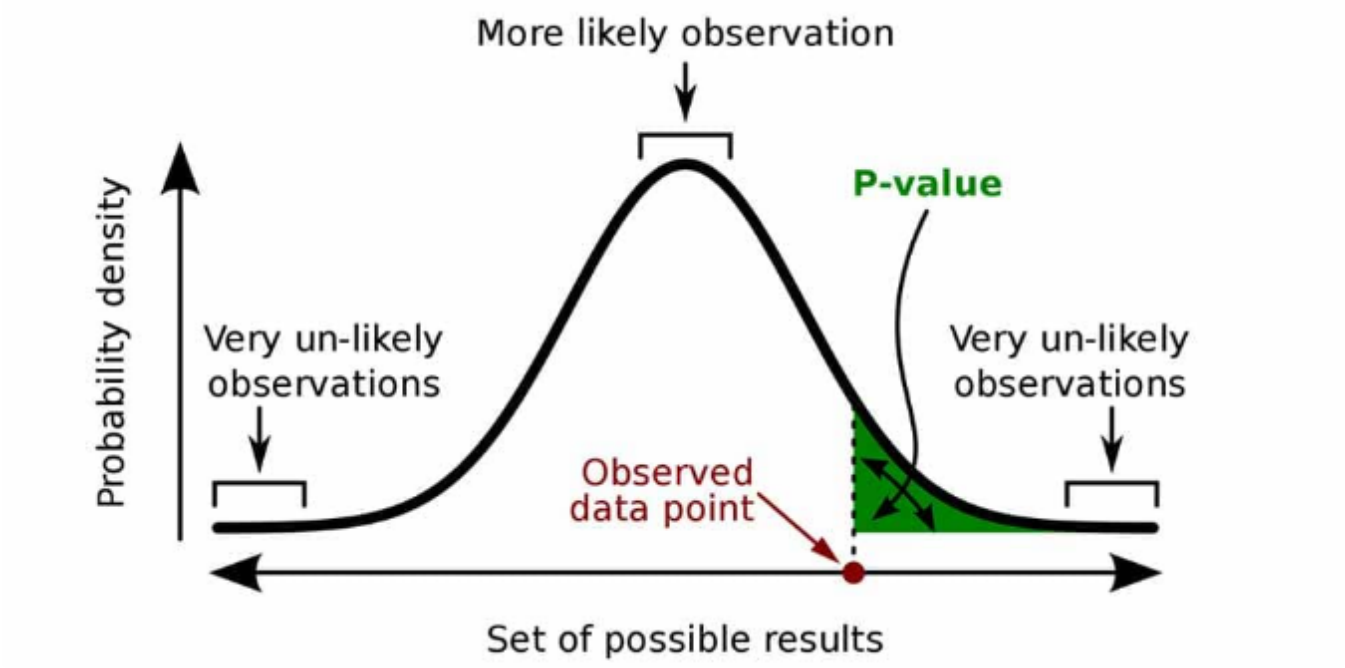
Ignacio de Gregorio

OpenAI Just Killed an Entire Market in 45 Minutes

The Story Everyone Should Have Seen Coming

✦ · 6 min read · Nov 9



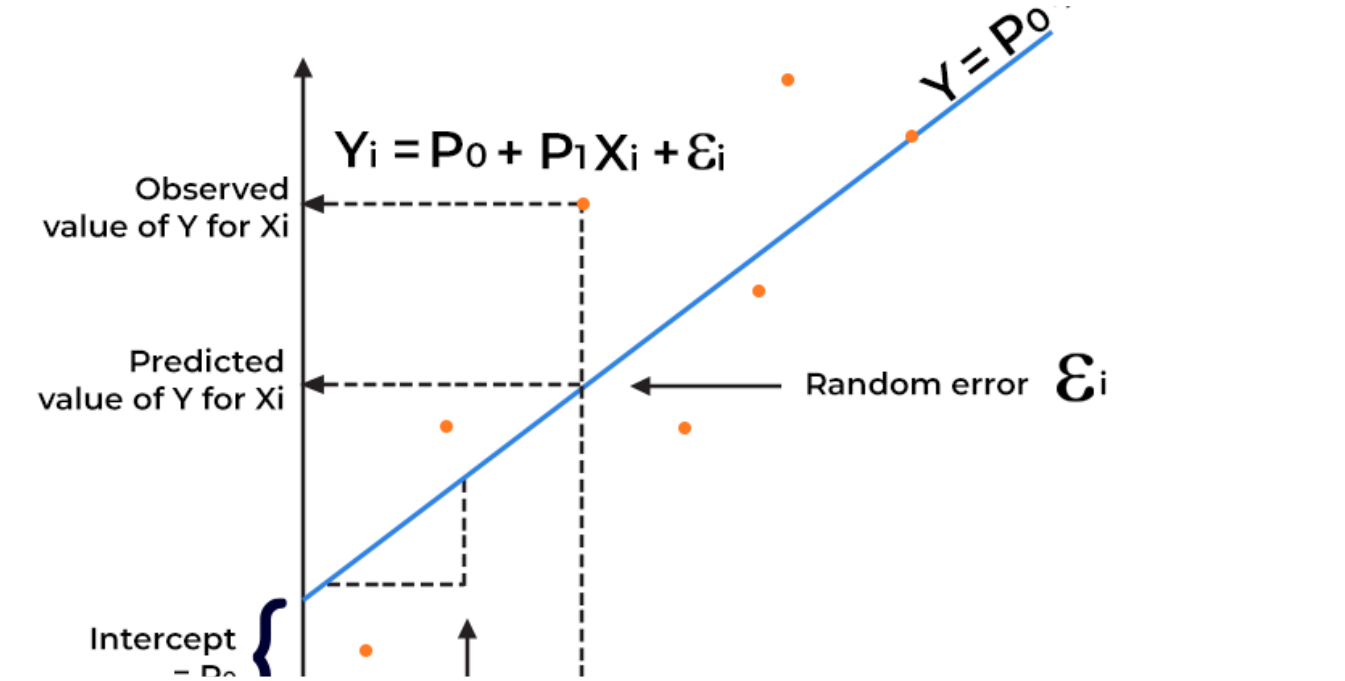


NitinKumar Sharma

P-value: Explained

6 min read · Sep 13

2



Yukio

Linear Regression: Multicolinearity, they taught you wrong

A FAMOUS MISCONCEPTION

4 min read · Jun 4

157 3

See more recommendations