The art of A/B testing

Walk through the beautiful math of statistical significance

Sylvain Truong Follow
Oct 3, 2018 · 10 min read

A/B testing, aka. split testing, refers to an experiment technique to determine whether a new design brings improvement, according to a chosen metric.

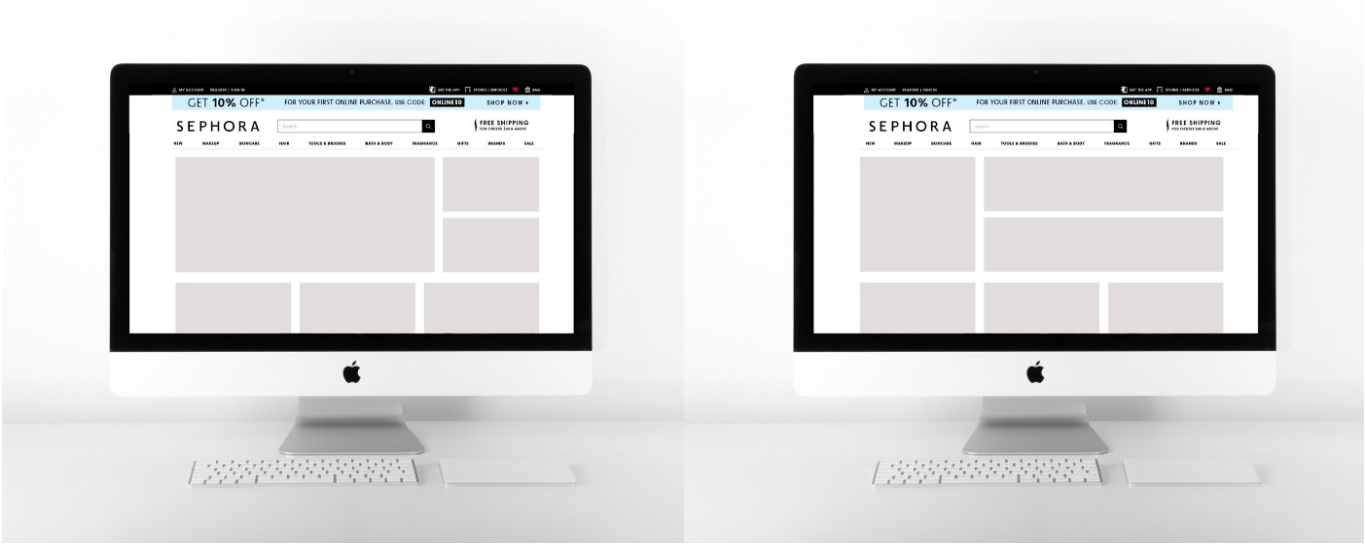
In web analytics, the idea is to challenge an existing version of a website (A) with a new one (B), by randomly splitting traffic and comparing metrics on each of the splits.

Exchanging on interface layouts

For the sake of example, let us look at beauty company Sephora South-East Asia (Sephora SEA, where I work), and in particular, its e-commerce platform.

The use case

Let us assume Sephora SEA is considering a landing page rearrangement for its Gold members.



Landing pages pane arrangement: version A (current) and version B (alternative to be tested)

The metrics that matter to the company are:

- Average time spent on the landing page per session
- Conversion rate, defined as proportion of sessions ending up with a transaction

An A/B test can be used to challenge to current arrangement.

Note that you can choose the split of traffic not to be 50–50 and allocate more traffic to version A, in case you are concerned about losses due to version B.

However, keep in mind that a very skewed split often leads to longer times before the A/B testing becomes *(statistically) significant*.

A/B testing as the only ultimate test

The thing with the development of new features on a website is that you hardly know beforehand how these are going to perform: they may actually hurt your business or make your profits skyrocket.

The knowledge of a User Experience (UX) designer is crucial in singling out feature suggestions that are likely to work. It would often follow best practices in UX or design examples that proved to be successful in other similar contexts.

However, no prior assumption can beat the real live test that is the A/B test.

The A/B test measures performance live, with real clients. Provided it is well executed, with no bias when sampling populations A and B, it gives you the best estimate of what would happen if you were to deploy version B.

Need for statistical formalism

After doing some prior market study, Sephora SEA decides it would be interesting to live test version B, with the following traffic split:

Version	Traffic Split
A	0.6
B	0.4

Let us assume that after 7 days of A/B testing, the tracking metrics of the experiment are

Version	number of sessions	avg (time)	std (time)	conversion rate
A	6000	60 s	40 s	1.50%
B	4000	62 s	45 s	2.00%

Just from looking at these outcomes, some questions arise:

- Because version B exhibits higher CR, does it mean version B brings improvement? Similarly, can we conclude on the influence on the average time spent?
- If so, with what level of confidence?
- Did a higher CR/lower average time spent of version B happen by chance?

Before jumping into conclusions, what you need to keep in mind is that

> The raw results we have are only samples of bigger populations. Their statistical properties vary around the ones of the populations they come from.

Therefore, statistically modelling these outputs is necessary. Introducing the concept of *statistical significance* in *hypothesis testing* also is.

A primer on Hypothesis Testing

For in-depth explanation of hypothesis testing, I would recommend this great post from William Koehrsen.

As for people looking for a quick statistics refresher, you may want to look at this article from Cassie Kozyrkov.

Here, I will go through the main lines of hypothesis testing: a **tool to compare** the distributions of 2 populations, based on samples from them.

What can be compared are either parameters of their distributions (*eg.* mean of time spent) or the distributions themselves (*eg.* the binary distribution of conversion rate).

The process starts in stating a *null hypothesis* ***Ho*** about the populations. In general, it is the *equality hypothesis*: *eg.* "the two populations have the same mean".

The *alternative hypothesis* ***H1*** negates the null hypothesis: *eg.* "the mean in the second population is higher than in the first".

The test can be summarised in two steps:

- **1.** Model ***Ho*** as a distribution on a single real-valued random variable (called the *test statistic*)
- **2.** Assess how likely the samples, or more extreme ones, could have been generated under ***Ho.*** This probability is the famous *p-value*. The lower it is, the more confident we can be in rejecting ***Ho.***

In the aforementioned post, @williamkoershen tests if the mean of hours of sleep in a university is lower than the national average, thus comparing an estimated value to a theoretical value. He uses a Z-test to do so.

Here, I suggest to extend that framework to A/B testings in the case of Sephora SEA.

. . .

In particular, I will show:

- how the Z-test can be applied to testing whether the clients experiencing B spend more time on average
- how the χ^2 test can be used to decide whether or not version B leads to a higher conversion rate
- how the Z-test can be adapted to test conversion rate of version B and if it yields the same conclusion as the χ^2 test

Let's get testing!

1 | Z-test for average time spent

The hypothesis to test are:

- ***Ho***: “the average time spent is the same for the two versions”
- ***H1***: “the average time spent is higher for version B”

The first step is to model H_0

The Z-test uses the Central Limit Theorem (CLT) to do so.

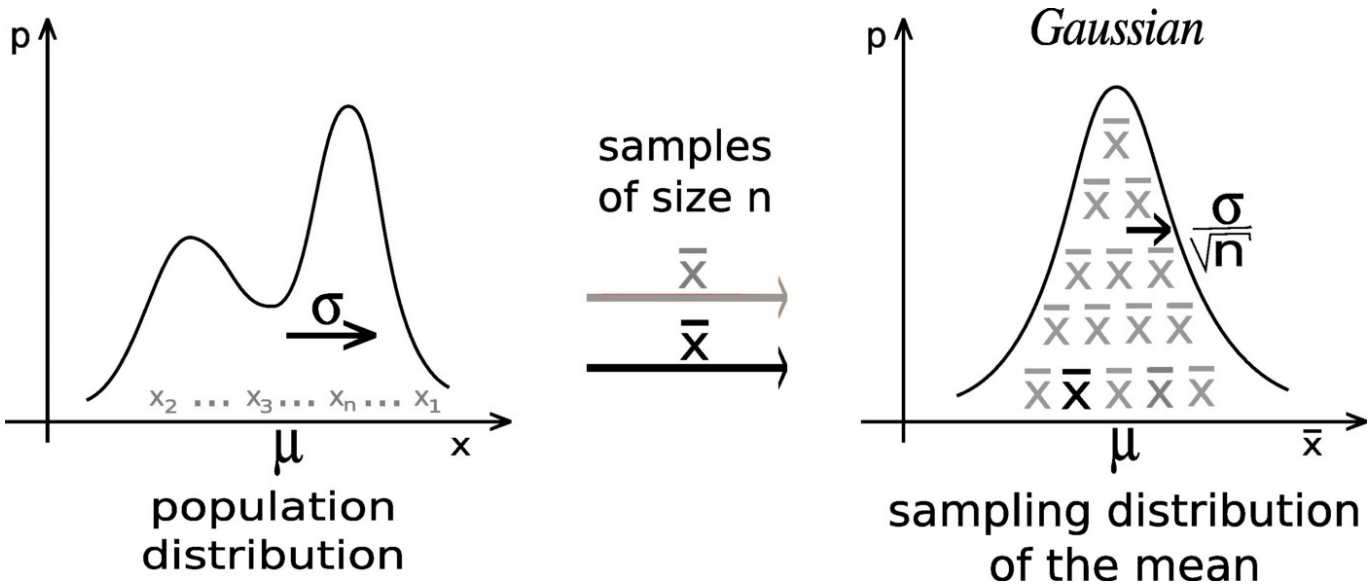


Illustration of the CLT (from Wikipedia)

> *The CLT establishes that:*
> *Given a random variable (rv)* ***X*** *of expectation* ***μ*** *and finite variance* σ^2,
> ***{X1,…,Xn}~X***, ***n*** *independent identically distributed (iid) rv, the following* ***approximation*** *on their average (also a rv) can be made*

$$\mu(n,X) = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

In our context, we model the time spent for each client session ***i*** as a realisation:

- a_i of rv $A_i \sim A$, if the client session belongs to the version A split
- b_i of rv $B_i \sim B$, else

We use the approximation provided by the CLT to derive that

$$\left(\quad, \sigma^2\right)$$

$$\mu(n_A, A) \sim \mathcal{N}\left(\mu_A, \frac{\sigma_A^2}{n_A}\right)$$

$$\mu(n_B, B) \sim \mathcal{N}\left(\mu_B, \frac{\sigma_A^2}{n_B}\right)$$

Hence the approximation for rv of the difference (wikipedia: sum of normally distributed rv),

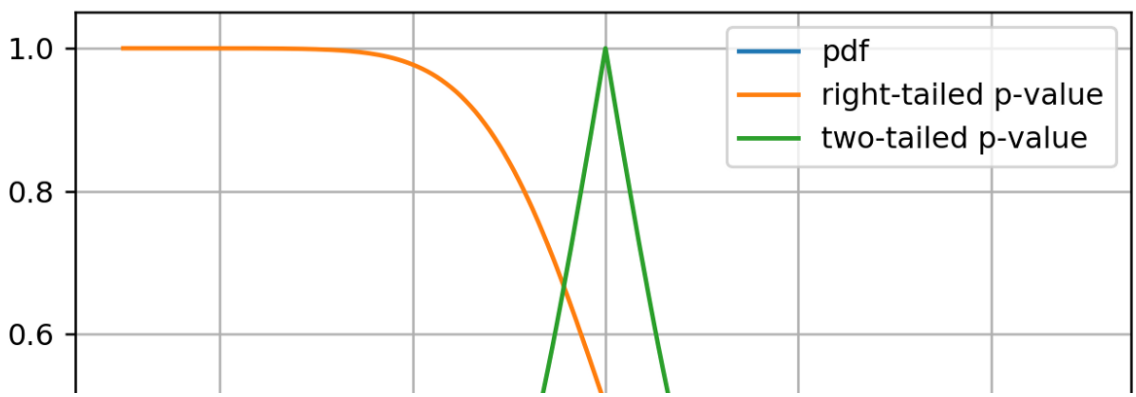
$$\mu(n_B, B) - \mu(n_A, A) \sim \mathcal{N}\left(\mu_B - \mu_A, \frac{{\sigma_B}^2}{n_B} + \frac{{\sigma_A}^2}{n_A}\right)$$

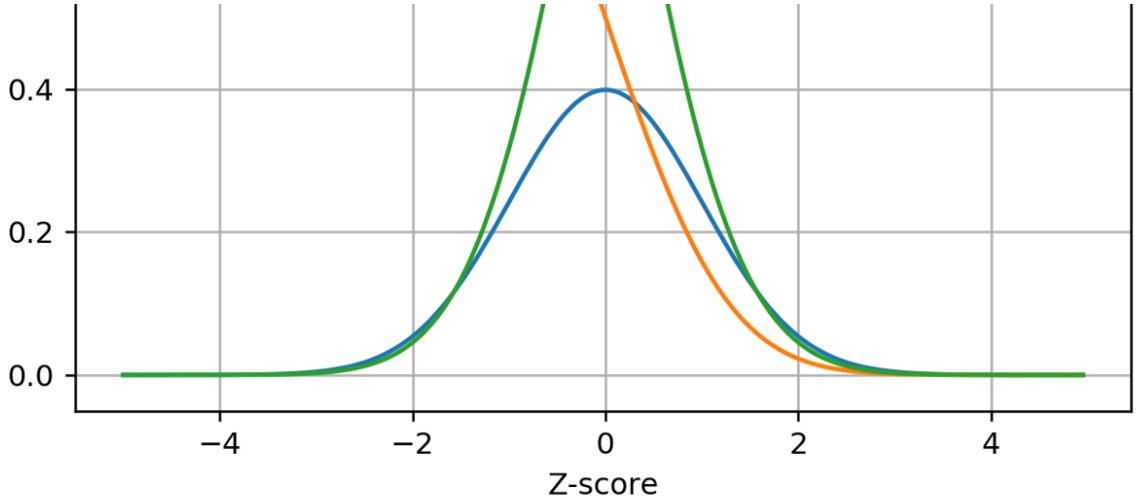
> *Under* **Ho**, *we have equality of the true means and therefore the model*

under $\mathcal{H}_0$,

$$\mu(n_B, B) - \mu(n_A, A) \sim \mathcal{N}\left(0, \frac{{\sigma_B}^2}{n_B} + \frac{{\sigma_A}^2}{n_A}\right)$$

$$\Leftrightarrow \frac{\mu(n_B, B) - \mu(n_A, A)}{\sqrt{\frac{{\sigma_B}^2}{n_B} + \frac{{\sigma_A}^2}{n_A}}} \sim \mathcal{N}(0, 1)$$





Curves about N(0,1): centrered and reduced Gaussian distribution, probability density function (pdf) and associated p-values

The second step is to see how likely our samples are under H_0

Note that true expectation and variance for ***A*** and ***B*** are unknown. We introduce their respective empirical estimators:

$$\hat{\mu}_A = \frac{1}{n_A}\sum_i a_i \text{ and } \hat{\sigma}_A = \frac{1}{n_A}\sum_i (\hat{\mu}_A - a_i)^2$$

$$\hat{\mu}_B = \frac{1}{n_B}\sum_i b_i \text{ and } \hat{\sigma}_B = \frac{1}{n_B}\sum_i (\hat{\mu}_B - b_i)^2$$

Our samples generated the following test statistic ***Z***, which needs to be tested against the reduced centered normal distribution:

$$Z = \frac{\hat{\mu}_B - \hat{\mu}_A}{\sqrt{\frac{\hat{\sigma}_B^{\,2}}{n_B} + \frac{\hat{\sigma}_A^{\,2}}{n_A}}}$$

Conceptually, ***Z*** represents the number of standard deviations the observed difference of means is away from 0. The higher this number, the lesser the likelihood of ***Ho***.

Also notice that in the case the estimated expectations are actually different, (number of samples)↗, Z↗.

From the formula of Z, you can also get the intuition that the smaller the difference to prove is, the more samples you need.

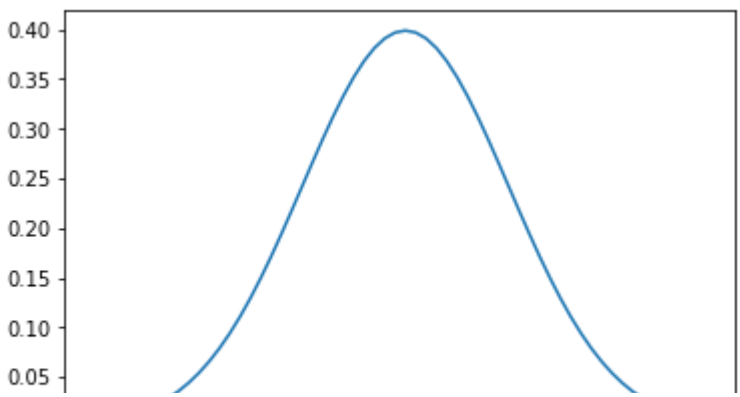
In Python, the calculation looks like

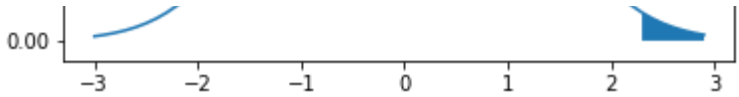
In [1]:
```
import numpy as np
from scipy.stats import norm

mu_B = 62
mu_A = 60

std_B = 45
std_A = 40

n_B = 4000
n_A = 6000

Z = (mu_B - mu_A)/np.sqrt(std_B**2/n_B + std_A**2/n_A)
pvalue = norm.sf(Z)

print("Z-score: {0}\np-value: {1}".format(Z,pvalue))
```

```
Z-score: 2.2749070654279993
p-value: 0.011455752709549046
```

ztest_1.ipynb hosted with ❤️ by GitHub view raw

In [2]:
```
import matplotlib.pyplot as plt

z = np.arange(-3, 3, 0.1)
plt.plot(z, norm.pdf(z))
plt.fill_between(z[z>Z], norm.pdf(z[z>Z]))
plt.show()
```





ztest_2.ipynb hosted with ❤️ by GitHub view raw

p-value calculation and graphical representation

There is a *pvalue* chance that a result as extreme as the one we observed could have happened under ***Ho.*** With a common go-to α criterion of 5%, we have *pvalue*$<\alpha$ and ***Ho*** can be rejected with confidence.

In cases where the sample size is not as big (< 30 per version), and the CLT approximation does not hold, one may take a look at Student's t-test.

2 | χ^2 test for conversion rate

The hypothesis to test are:

- ***Ho***: "the conversion rate is the same for the two versions"
- ***H1***: "the conversion rate is higher for version B"

Unlike the previous case, the outcome for each client session is not continuous but binary: either "not converted" or "converted".

The summary of the observed outcomes is the following

Version	number of sessions	converted	non converted
A	6000	$O_{A1}=90$	$O_{A0}=5910$
B	4000	$O_{B1}=80$	$O_{B0}=3920$

The χ^2 test compares distributions of multinomial outcomes but we will keep to the binary case in this example.

As before, we will tackle the problem in two steps:

The first step is to model H_0

In ***Ho,*** conversions in version A and version B follow the same binomial distribution ***B(1,p)***. We pool the observations in both version A and B and derive the estimator for CR

$$\hat{p} = \frac{O_{A1} + O_{B1}}{O_{A0} + O_{B0} + O_{A1} + O_{B1}}$$

and get `\hat{p} = 0.0170`

Thus, under ***Ho,*** the theoretical outcome table is

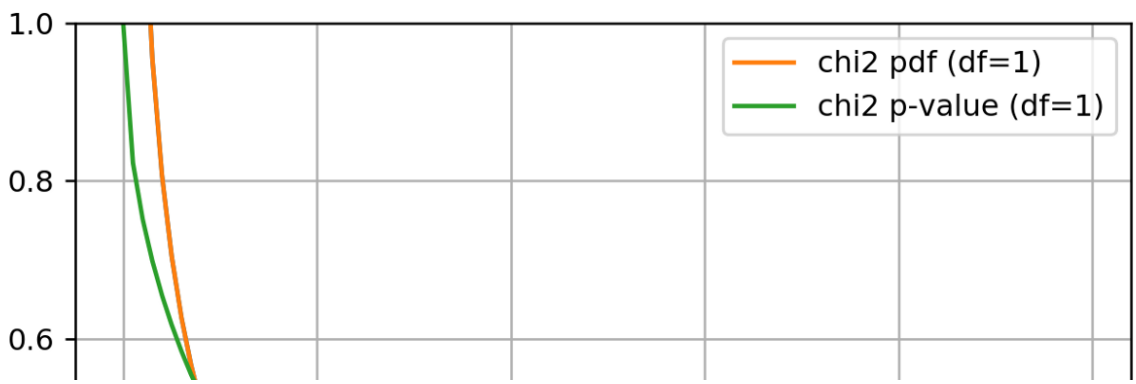
Version	number of sessions	converted	non converted
A	6000	T_{A1}=102	T_{A0}=5898
B	4000	T_{B1}=68	T_{B0}=3932

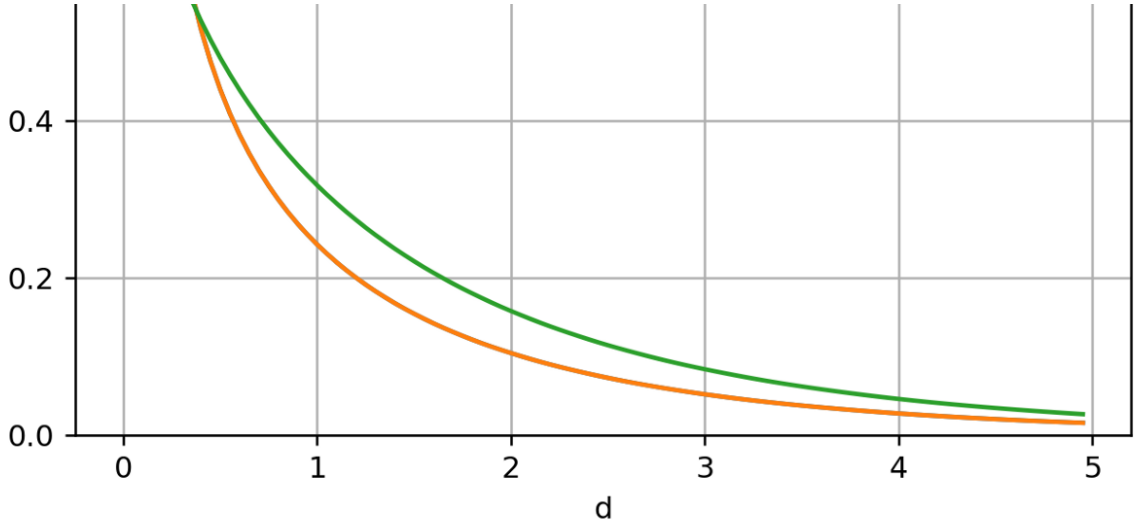
Let us look at the rv D, defined by

$$D = \sum_{version \in \{A,B\}, conv \in \{0,1\}} \frac{(O_{version,conv} - T_{version,conv})^2}{T_{version,conv}}$$

D represents a squared relative distance between the theoretical and the observed distributions.

According to Pearson's theorem, under ***Ho***, ***D*** follows a χ^2 probability law with 1 degree of freedom (df).





Curves about the χ^2 law (df=1)

The second step is to see how likely our samples are under H_0

It consists in computing the observed ***D*** and deriving its corresponding p-value according to the χ^2 law.

This is how it can be done in Python:

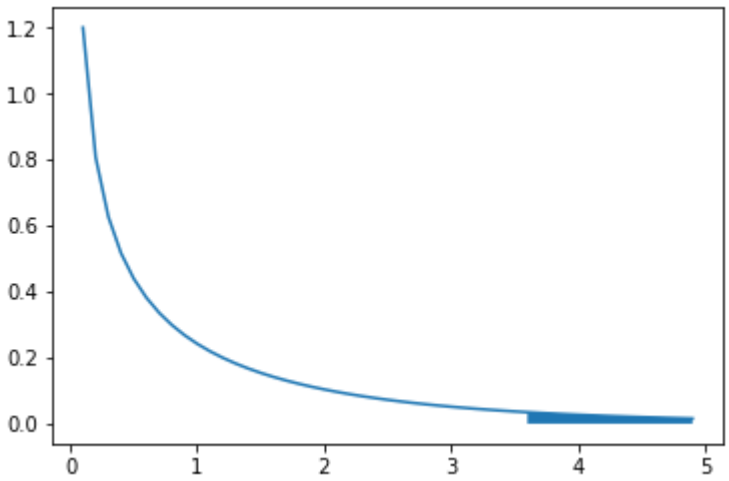
In [1]:
```
from scipy.stats import chi2
import numpy as np

T = np.array([102, 68, 5898, 3932])
O = np.array([90, 80, 5910, 3920])

D = np.sum(np.square(T-O)/T)

pvalue = chi2.sf(D, df=1)

print("distance d: {0}\np-value: {1}".format(D,pvalue))
```

```
distance d: 3.590449404583807
p-value: 0.05811252579106675
```

chi2_1.ipynb hosted with ❤ by GitHub view raw

In [2]:
```
import matplotlib.pyplot as plt
```

```
d = np.arange(0, 5, 0.1)
plt.plot(d, chi2.pdf(d, df=1))
plt.fill_between(d[d>D], chi2.pdf(d[d>D], df=1))
plt.show()
```



chi2_2.ipynb hosted with ❤️ by GitHub view raw

p-value calculation and graphical representation

There is a *pvalue* chance that a result at least as distant from the theoretical distribution as our observation would have happened under ***Ho.*** With a common go-to α criterion of 5%, we have *pvalue*$>\alpha$ and ***Ho*** cannot be rejected.

3 | Z-test for conversion rate

The Z-test could be adapted to conversion rate by modelling conversion as an rv which realisations are in {0,1}:

- 1 for a conversion
- 0 else

We keep the same notations as before and model conversion for each client session i as a realisation:

- $a_i \in \{0,1\}$ of rv $A_i \sim A$, if the client session belongs to the version A split
- $b_i \in \{0,1\}$ of rv $B_i \sim B$, else

The first step is to model H_0

Under Ho, $\mu(A) = \mu(B)$ and we have

$$\frac{\mu(n_B, B) - \mu(n_A, A)}{\sqrt{\frac{{\sigma_B}^2}{n_B} + \frac{{\sigma_A}^2}{n_A}}} \sim \mathcal{N}(0,1)$$

The corresponding test statistic

$$Z = \frac{\hat{\mu}_B - \hat{\mu}_A}{\sqrt{\frac{{\hat{\sigma}_B}^2}{n_B} + \frac{{\hat{\sigma}_A}^2}{n_A}}}$$

This time, with binary rvs, it can be shown that the estimators for the standard deviations are functions of the expectations:

$${\hat{\sigma}_A}^2 = \hat{\mu}_A(1 - \hat{\mu}_A)$$
$${\hat{\sigma}_B}^2 = \hat{\mu}_B(1 - \hat{\mu}_B)$$

The second step is to see how likely our samples are under H_0

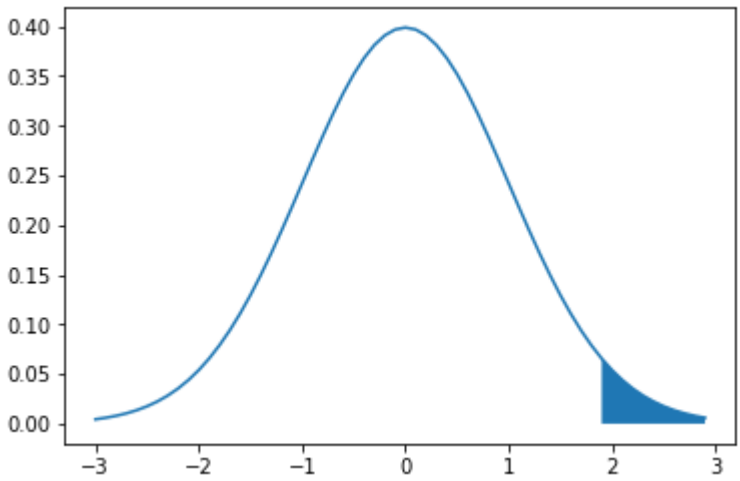
To this end, we compute the Z-score and the corresponding right-tailed p-value:

In [1]:
```
import numpy as np
from scipy.stats import norm

mu_B = 0.02
mu_A = 0.015

var_B = mu_B * (1-mu_B)
var_A = mu_A * (1-mu_A)

n_B = 4000
n_A = 6000

Z = (mu_B - mu_A)/np.sqrt(var_B/n_B + var_A/n_A)
pvalue = norm.sf(Z)

print("Z-score: {0}\np-value: {1}".format(Z,pvalue))
```

```
Z-score: 1.8427115179918694
p-value: 0.03268557071858785
```

ztest_categorical_1.ipynb hosted with ❤️ by GitHub view raw

In [2]:
```
import matplotlib.pyplot as plt

z = np.arange(-3, 3, 0.1)
plt.plot(z, norm.pdf(z))
plt.fill_between(z[z>Z], norm.pdf(z[z>Z]))
plt.show()
```



ztest_categorical_2.ipynb hosted with ❤️ by GitHub view raw

p-value calculation and graphical representation

With this modelling, the p-value output is slightly lower than with the χ^2 test. With the same $\alpha=0.05$ criterion, we would have rejected the null hypothesis (!!!).

This difference may be explained by a slight weakness of the Z-test, which does not acknowledge here the binary nature of the rv: $\mu(B)-\mu(A)$ is actually bounded in $[-1,1]$ and the observation is therefore attributed a lower p-value.

Always question your tests

and never make assumptions. A/B testing is indeed a great way to alleviate human bias when deciding on relevance of new features. However, do not forget that A/B testing still relies on a model of truth: as we have seen, there are different possible models.

In the case of large samples, they tend to converge to similar conclusions. In particular, the CLT approximation holds better than with small sample sizes.

In the latter cases, one may explore Student's t-test, Welch's t-test and Fisher's exact test. You may also explore the realm of Reinforcement Learning in order to maximise gains while testing (Multi-armed bandits and the Exploitation vs Exploration dilemma).

Not only should you be strict in your interpretations of results but also be aware of contextual effects of your A/B test:

- time of the year/month/week, the weather, the economic context can affect the nature of your audience
- even if after two days of A/B testing your results are significant, they may not be over the course of a week

Main take-home messages

- Hypothesis testing is about modelling a null hypothesis ***Ho*** and assessing how likely it is, given the samples you got from the A/B test
- The key is in the ***Ho*** model and we have seen, it can be derived from the CLT (Z-test) or Pearson's theorem (χ^2 test)

. . .

This concludes our statistics tour: I hope you enjoyed the ride as much as I enjoyed writing it! Comments, suggestions, corrections are much appreciated.

. . .

Credits:
Photos by rawpixel and *Louis Reed on Unsplash*

Data Science Statistics Marketing Testing User Experience