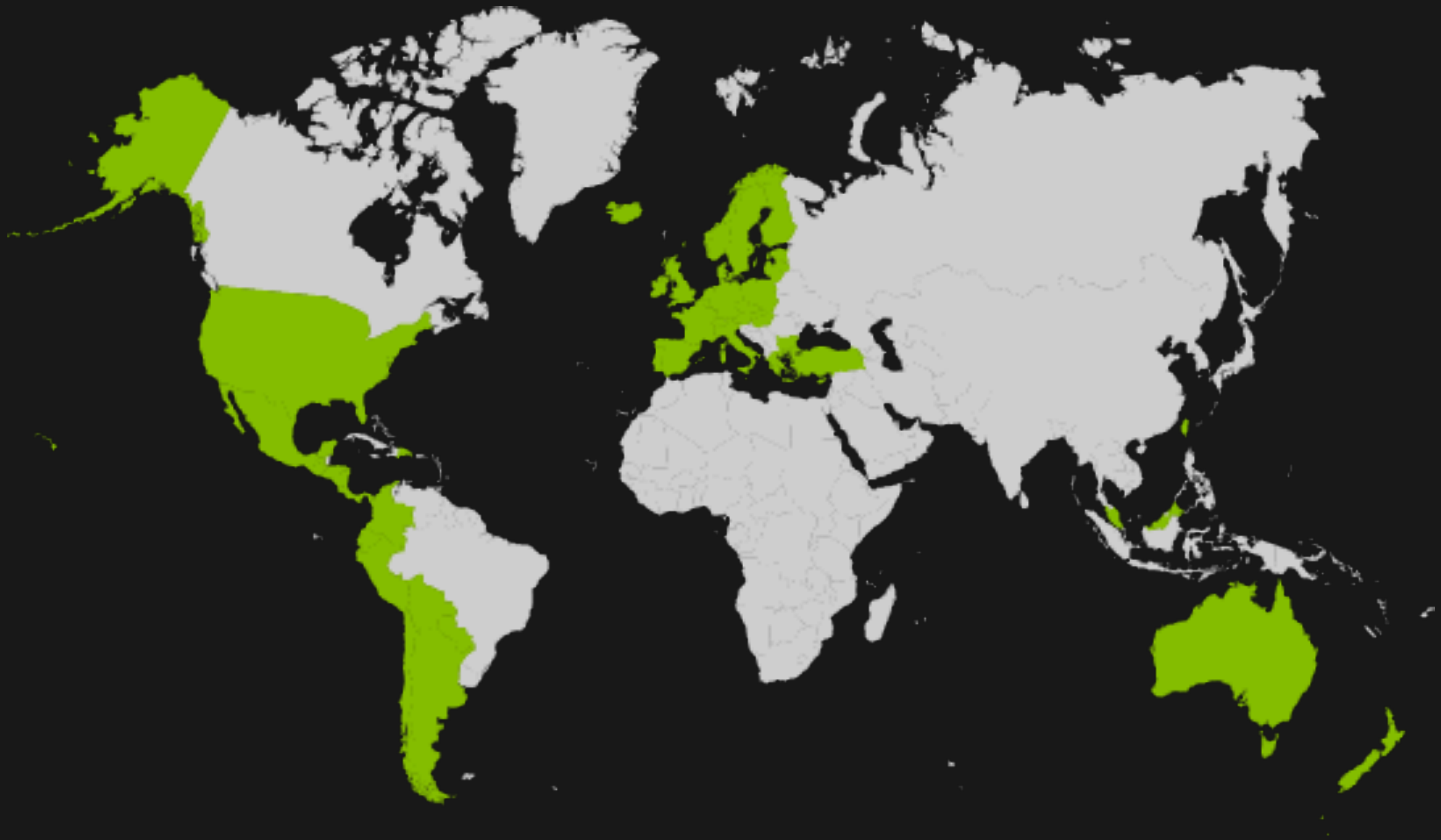
A/B Testing: Avoiding Common Pitfalls

Danielle Jabin

Spotify®

März 6, 2014

Make all the world's music available instantly to everyone, wherever and whenever they want it



Over 24 million active users

Access to more than 20 million songs

Discover

You have been listening to a lot of **House**. Try this song by **Lovebirds**?

Want You In My Soul - Original Mix

Lovebirds

Recommended for you. **In-Grid**.

You listened to **Cristiano Araújo** and **Jorge & Mateus**. Here's an album you might like.

Curtição

João Bosco e Vinícius

You listened to **Maya Jane Coles**.

Tony Lazarew has been listening to a lot of **Матрёшка** this week.

A DAY AGO

Матрёшка

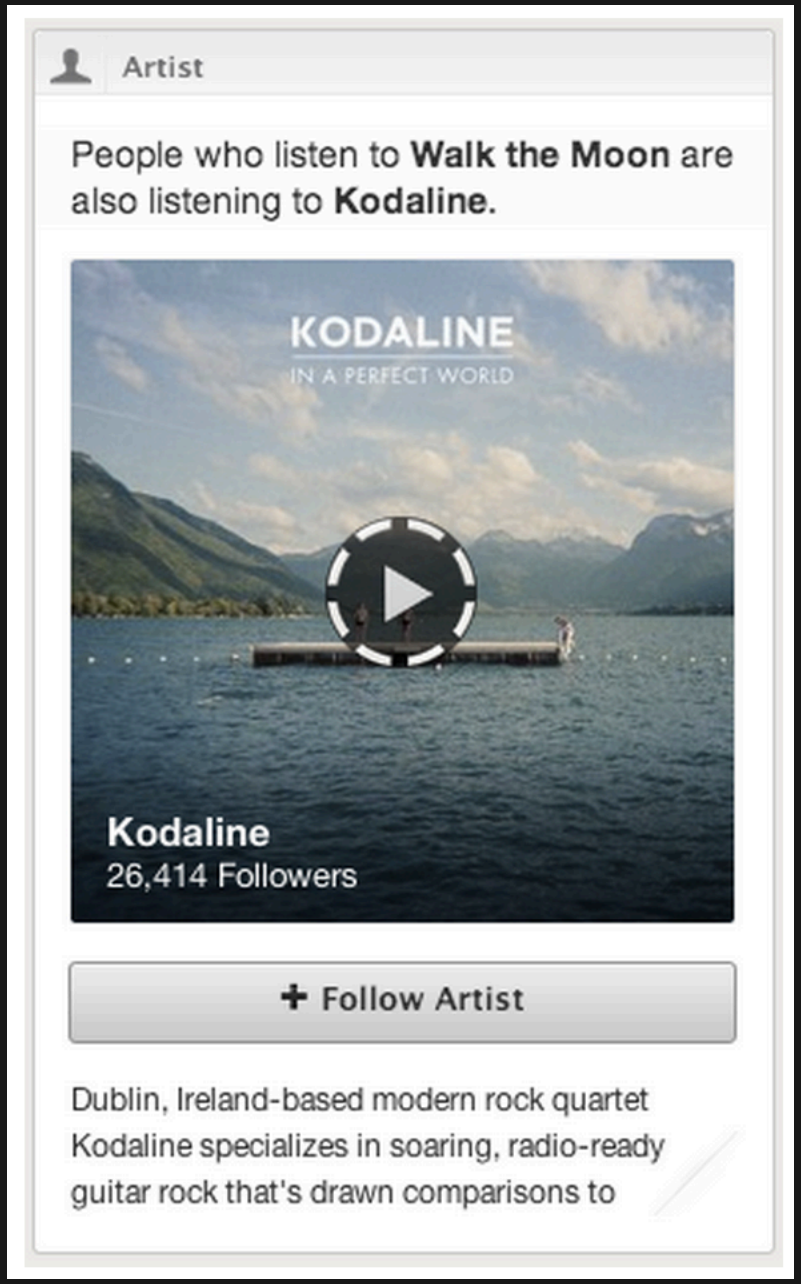
Ляпис Трубецкой

0 ▶ 0 Like

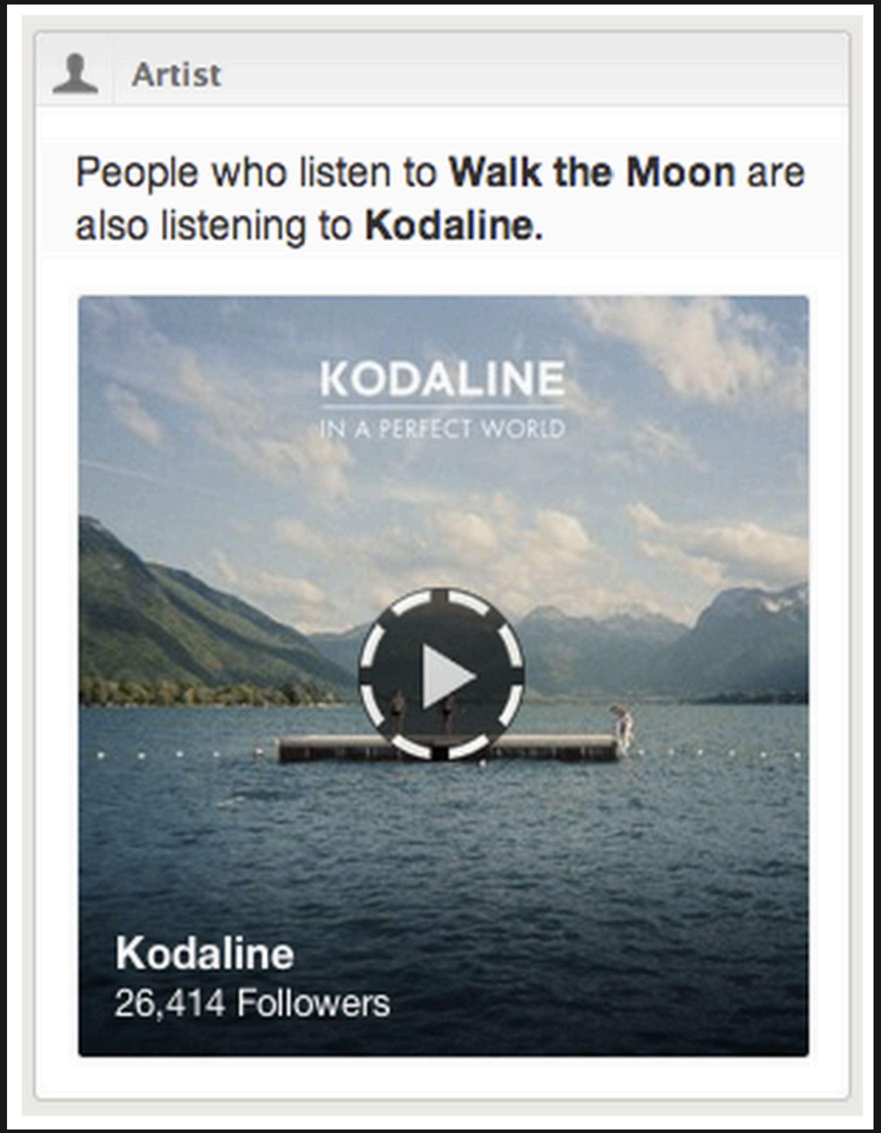
But can we make it even easier?

We can try…
…with A/B testing!

So…what’s an A/B test?

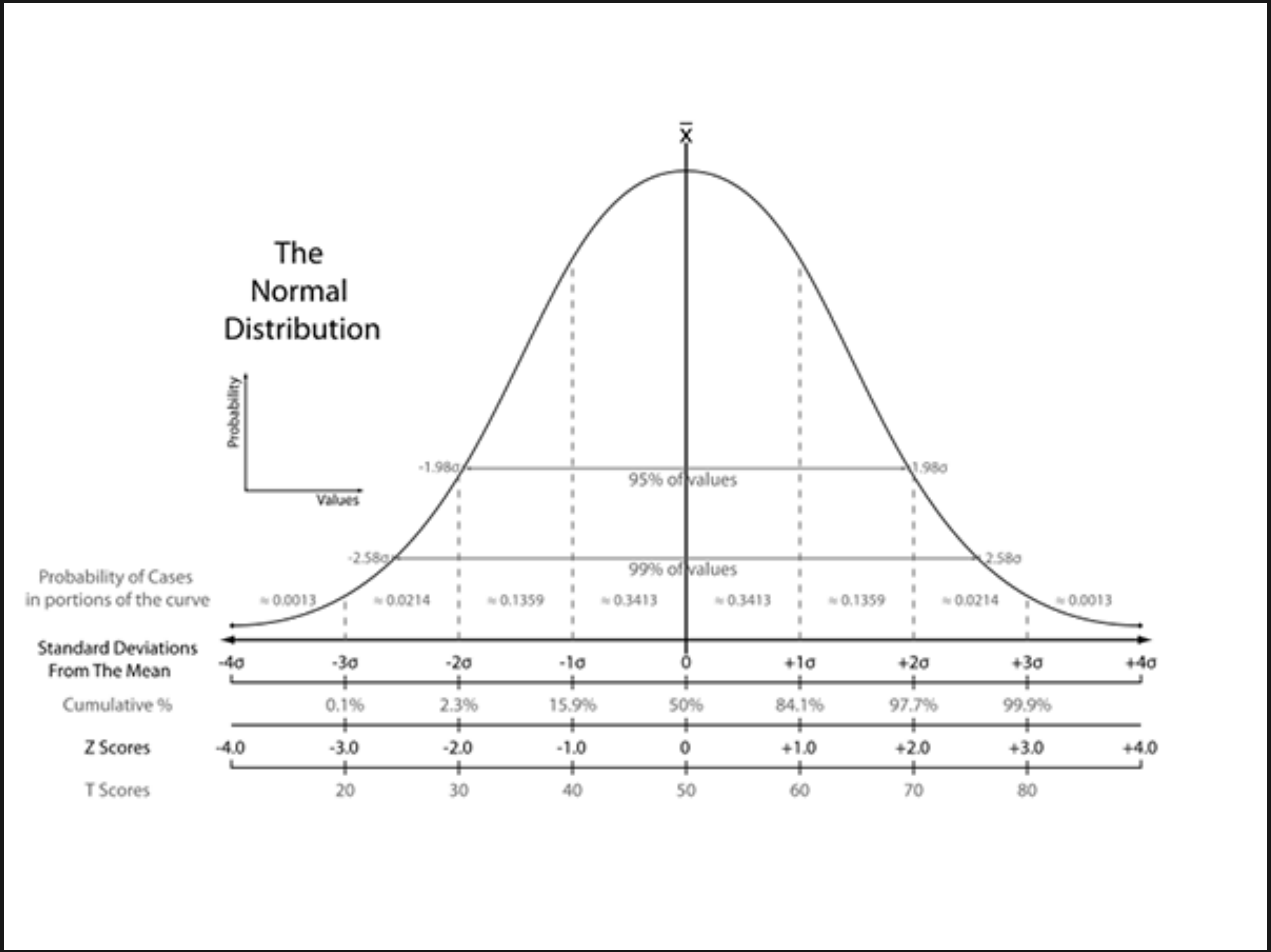

Artist
People who listen to **Walk the Moon** are also listening to **Kodaline.**
KODALINE
IN A PERFECT WORLD
Kodaline
26,414 Followers
+ Follow Artist
Dublin, Ireland-based modern rock quartet Kodaline specializes in soaring, radio-ready guitar rock that's drawn comparisons to

Control


Artist
People who listen to **Walk the Moon** are also listening to **Kodaline.**
KODALINE
IN A PERFECT WORLD
Kodaline
26,414 Followers

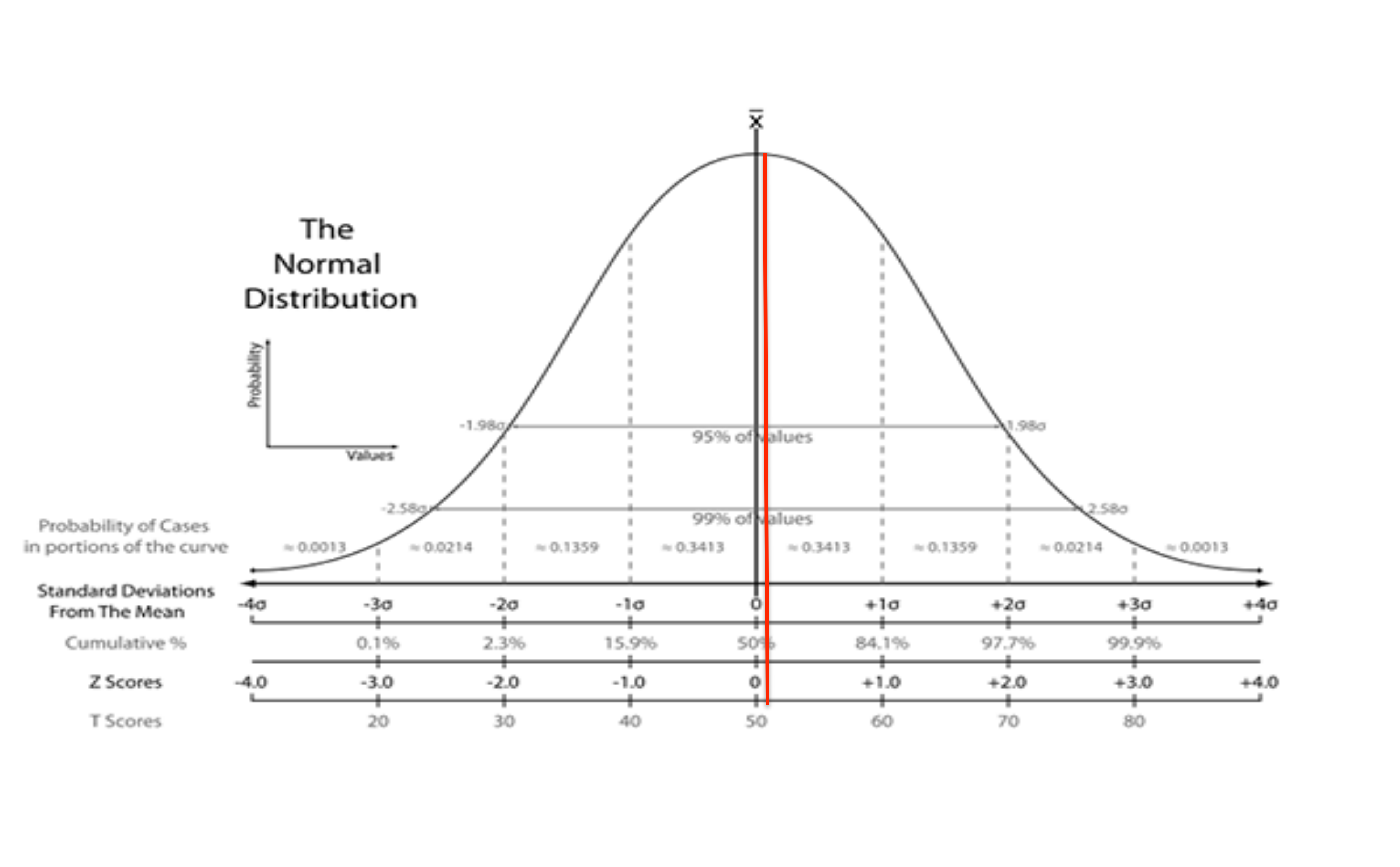
A

Pitfall #1: Not limiting your error rate

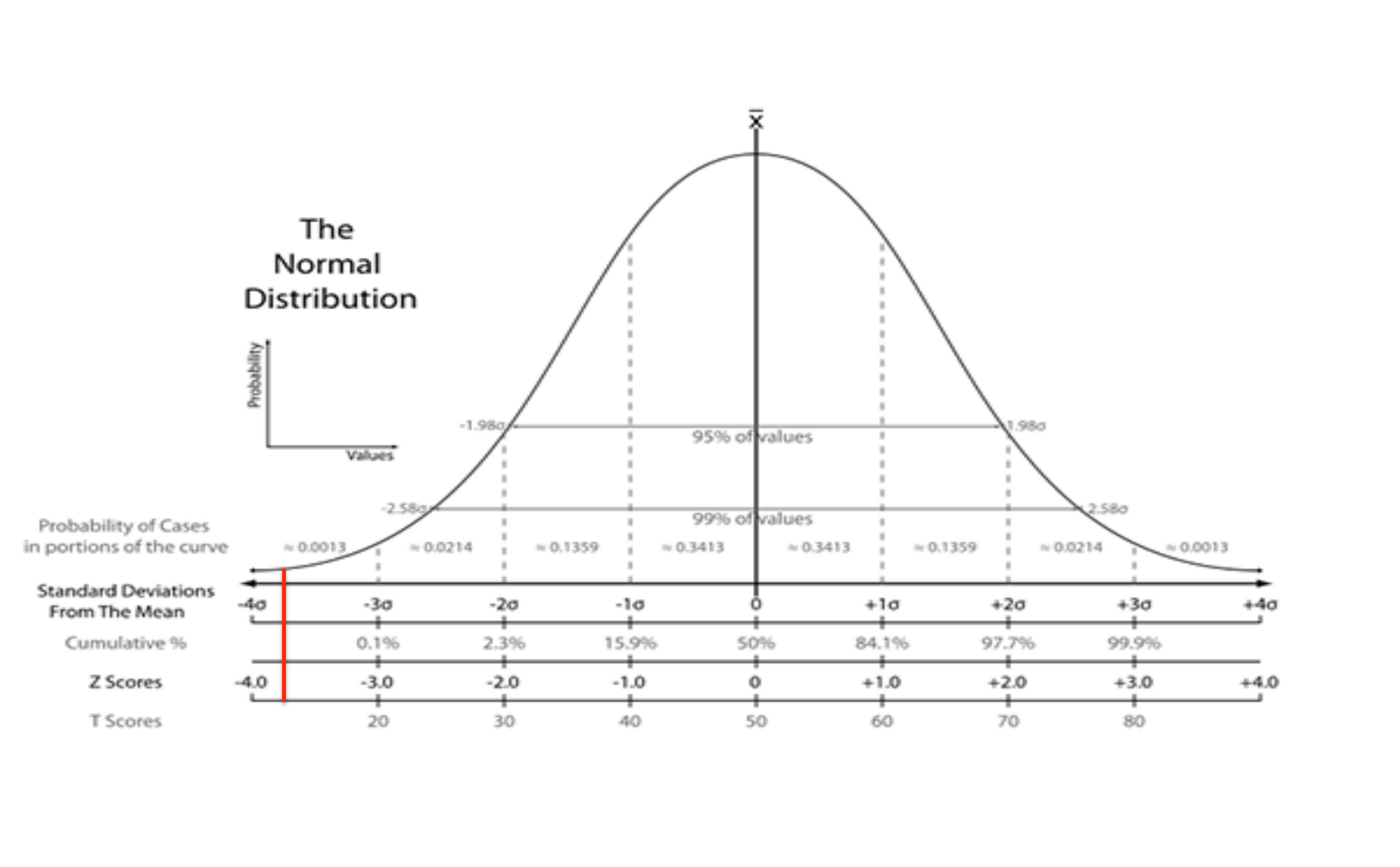


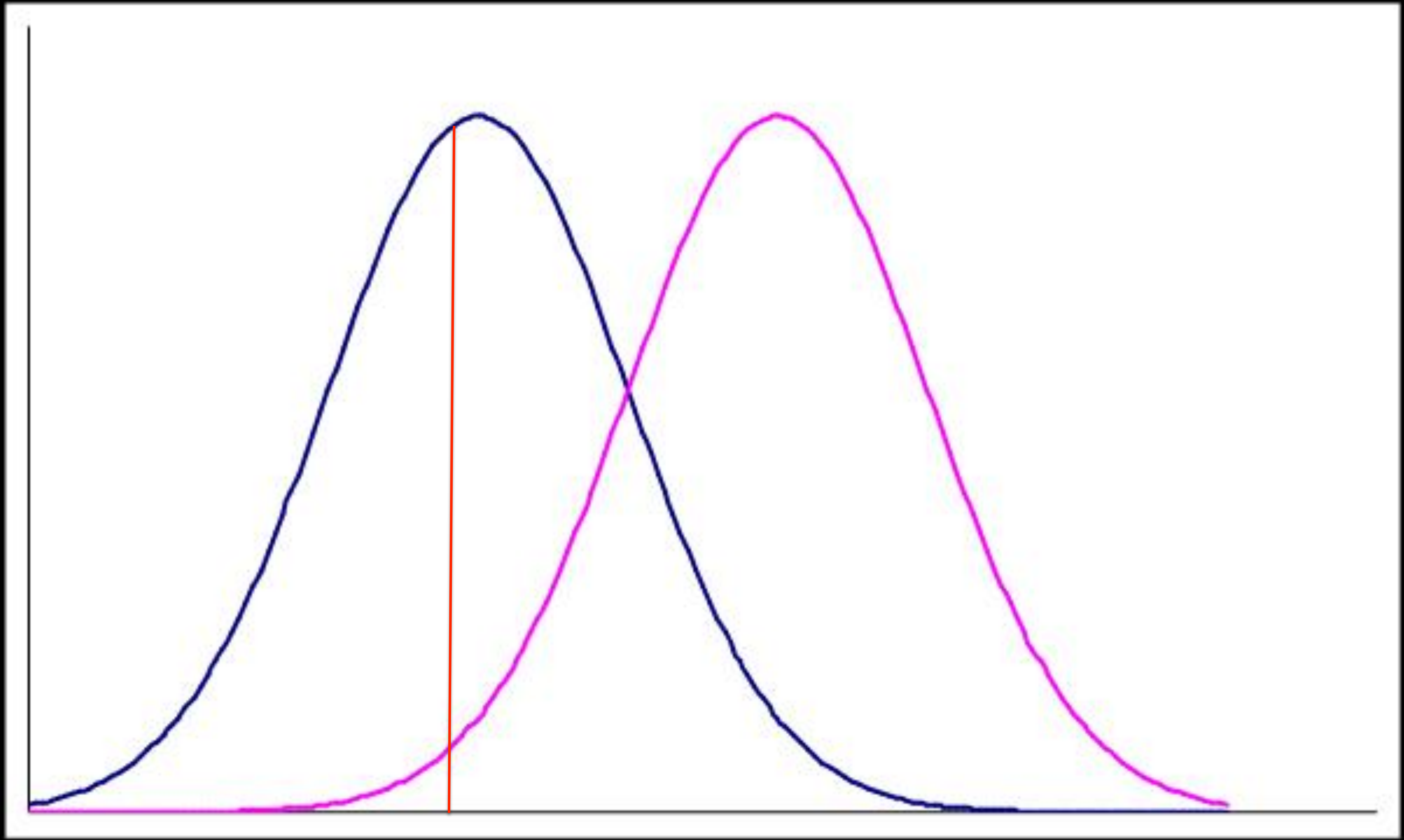
Source: assets.20bits.com/20081027/normal-curve-small.png

What if I flip a coin 100 times and get 51 heads?



What if I flip a coin 100 times and get 5 heads?

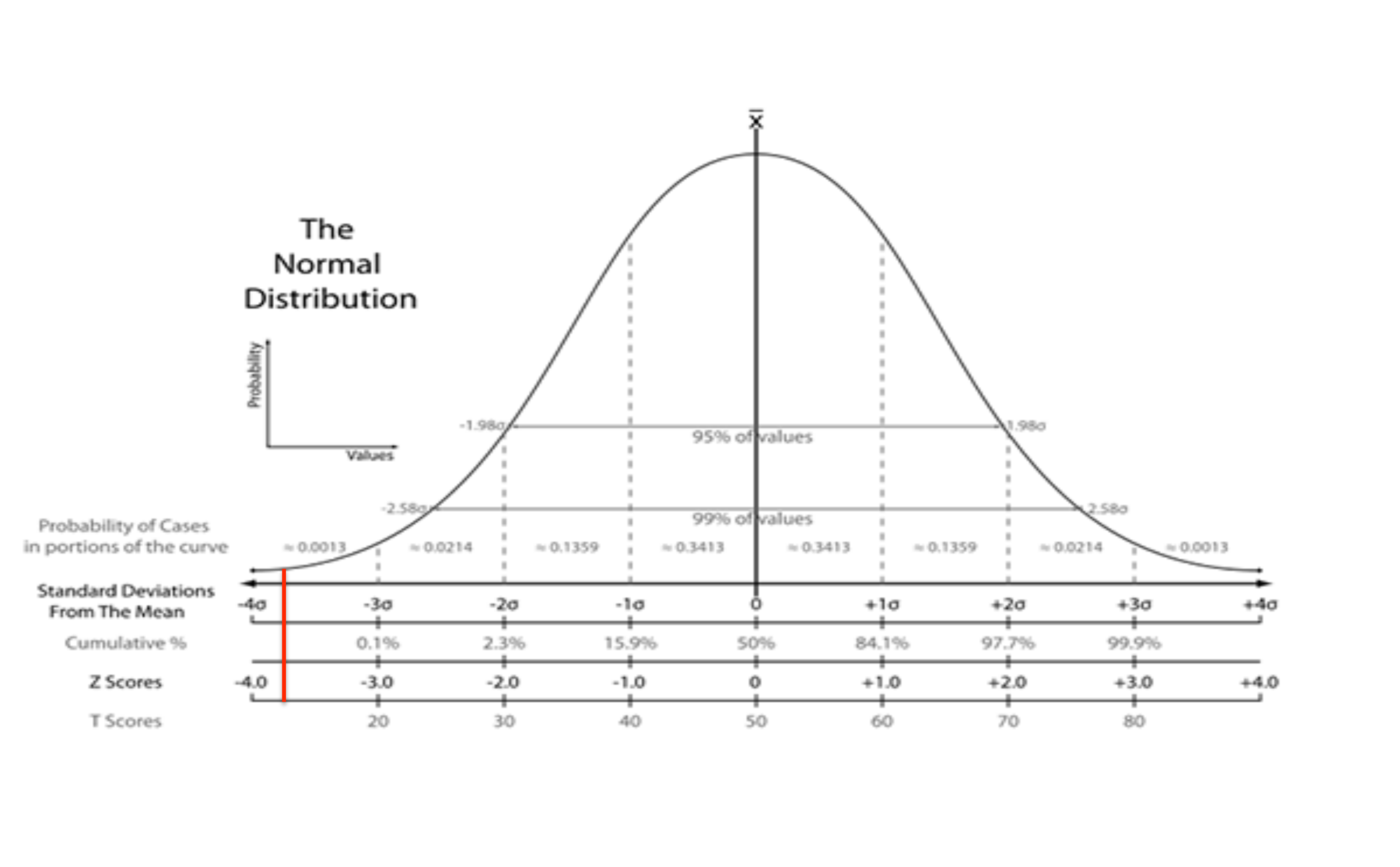




The likelihood of obtaining a certain value under a given distribution is measured by its p-value

If there is a low likelihood that a change is due to chance alone, we call our results statistically significant

What if I flip a coin 100 times and get 5 heads?



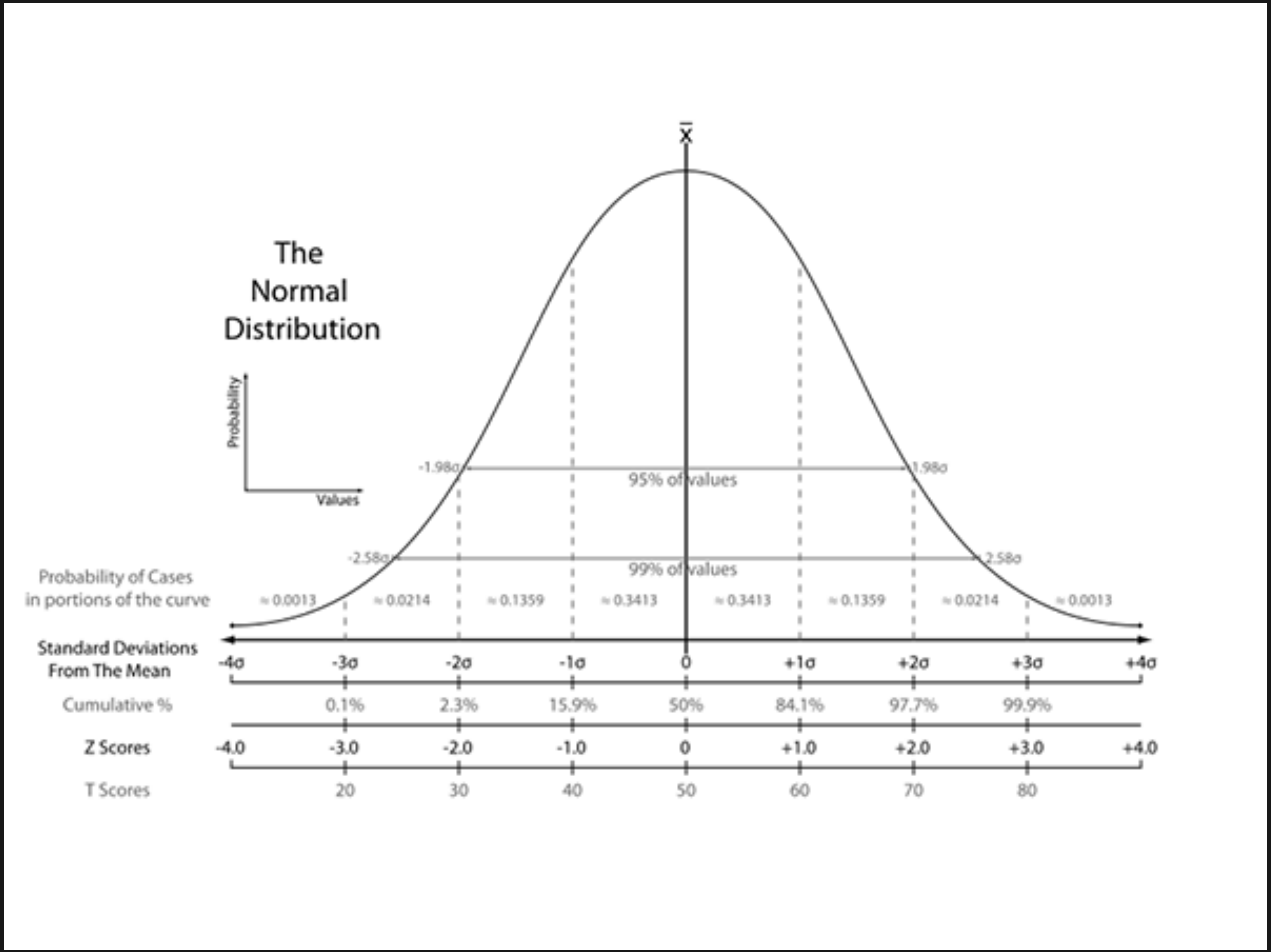
Statistical significance is measured by alpha

- alpha levels of 5% and 1% are most commonly used
  - Alternatively: P(significant) = .05 or .01

Each alpha has a corresponding Z-score

alpha	Z-score (two-sided test)
.10	1.65
.05	1.96
.01	2.58

The Z-score tells us how far a particular value is from the mean (and what the corresponding likelihood is)

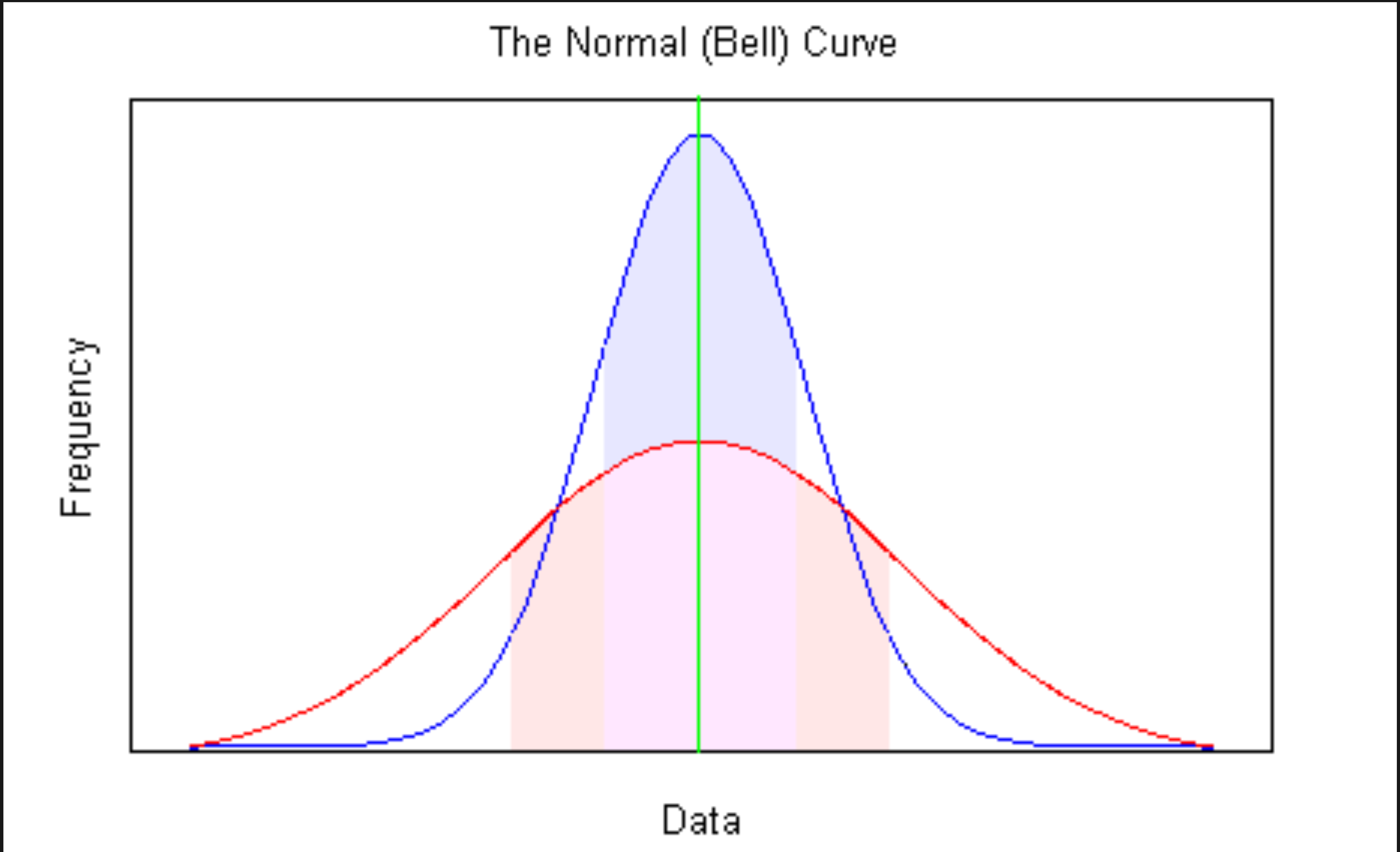


Source: assets.20bits.com/20081027/normal-curve-small.png

Compute the Z-score at the end of the test

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Standard deviation (σ) tells us how spread out the numbers are


The Normal (Bell) Curve
Frequency
Data

To lock in error rates before you start, fix your sample size

What should my sample size be?

Represents the desired power (typically .84 for 80% power).

Sample size in each group (assumes equal sized groups)

$$n = \frac{2\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

Standard deviation of the outcome variable

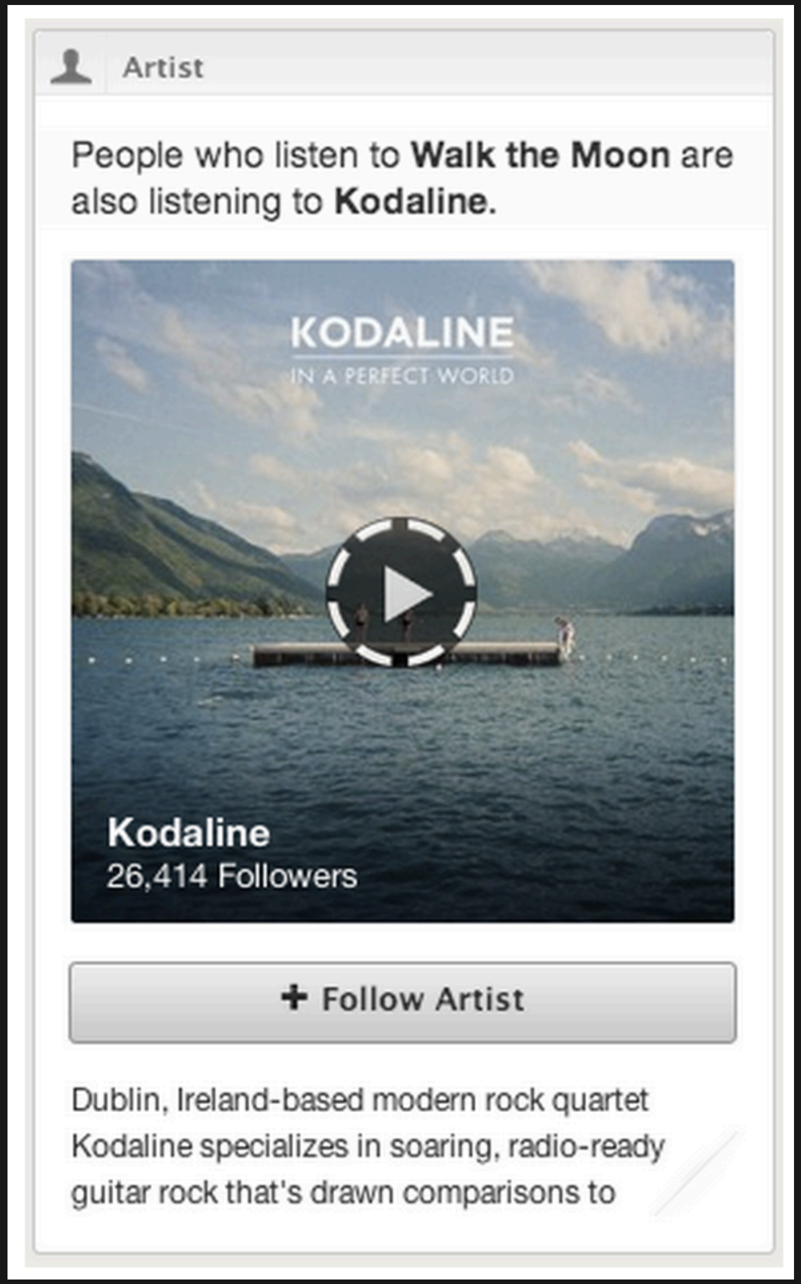
Effect Size (the difference in means)

Represents the desired level of statistical significance (typically 1.96).

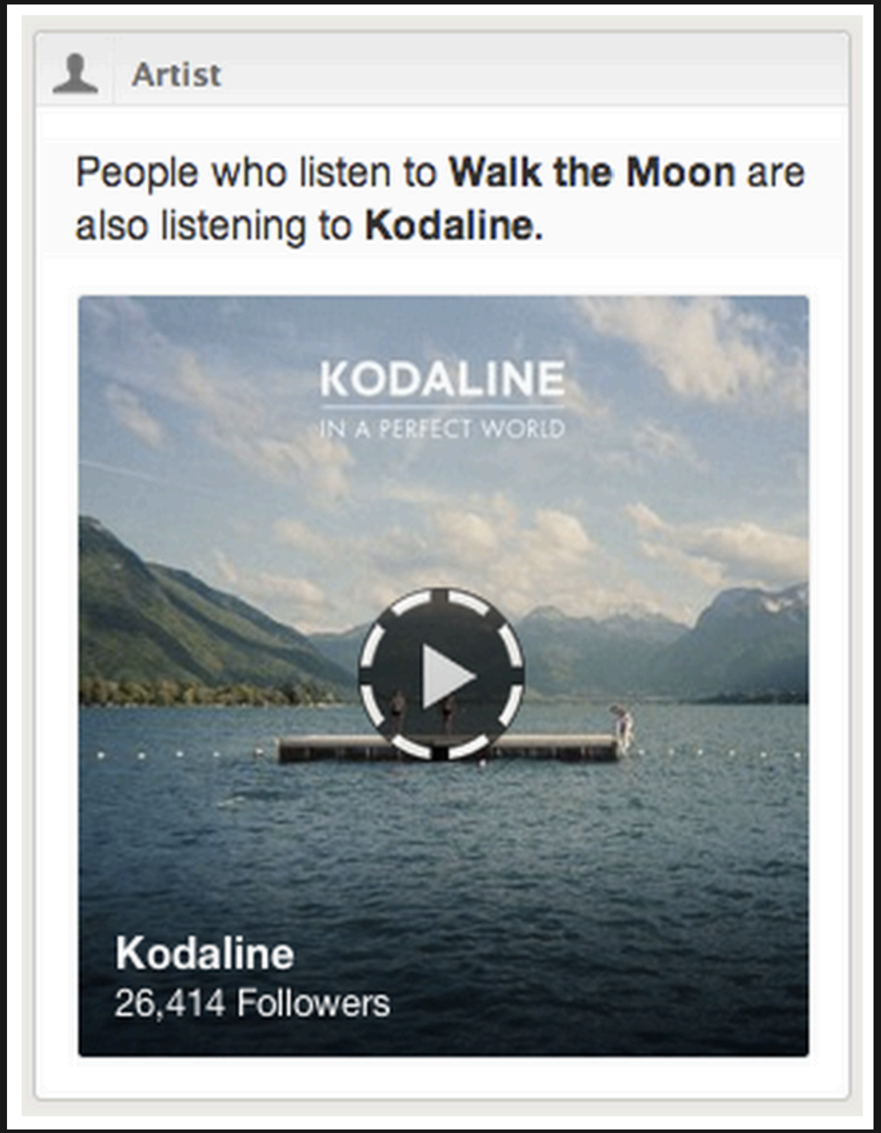
Source: www.stanford.edu/~kcobb/hrp259/lecture11.ppt

Recap: running an A/B test

- Compute your sample size
  - Using alpha, beta, standard deviation of your metric, and effect size
- Run your test! But stop once you've reached the fixed sample size stopping point
- Compute your z-score and compare it with the z-score for the chosen alpha level


Artist
People who listen to **Walk the Moon** are also listening to **Kodaline.**
KODALINE
IN A PERFECT WORLD
Kodaline
26,414 Followers
Follow Artist
Dublin, Ireland-based modern rock quartet Kodaline specializes in soaring, radio-ready guitar rock that's drawn comparisons to

Control


Artist
People who listen to **Walk the Moon** are also listening to **Kodaline.**
KODALINE
IN A PERFECT WORLD
Kodaline
26,414 Followers

A

Resulting Z-score?

33.3

Pitfall #2: Stopping your test before the fixed sample size stopping point

Sample size for varying alpha levels

- With σ = 10, difference in means = 1

	Two-sided test
alpha = .10, beta = .80	1230
alpha = .05, beta = .80	1568
alpha = .01, beta = .80	2339

Let's see some numbers

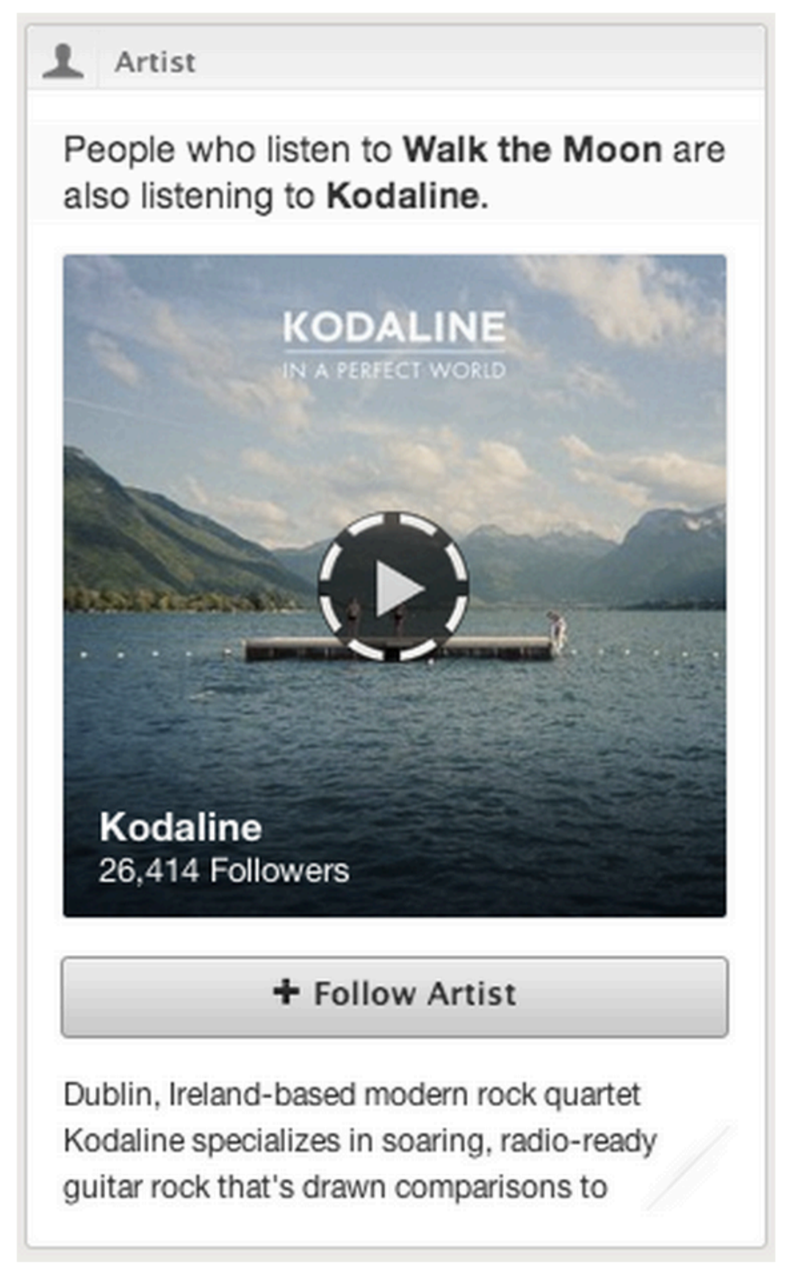
- 1,000 experiments with 200,000 fake participants divided randomly into two groups both receiving the exact same version, A, with a 3% conversion rate

	Stop at first point of significance	Ended as significant
90% significance reached	654 of 1,000	100 of 1,000
95% significance reached	427 of 1,000	49 of 1,000
99% significance reached	146 of 1,000	14 of 1,000

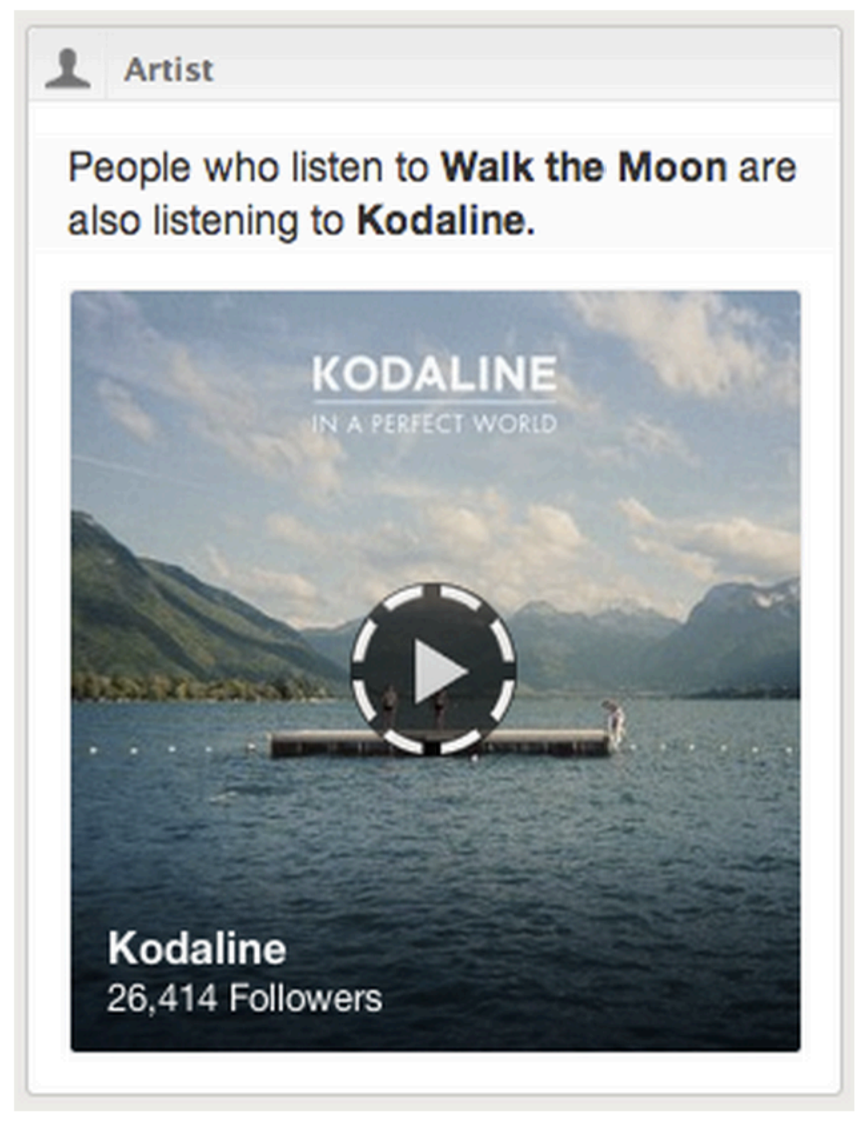
Source: destack.home.xs4all.nl/projects/significance/

Remedies

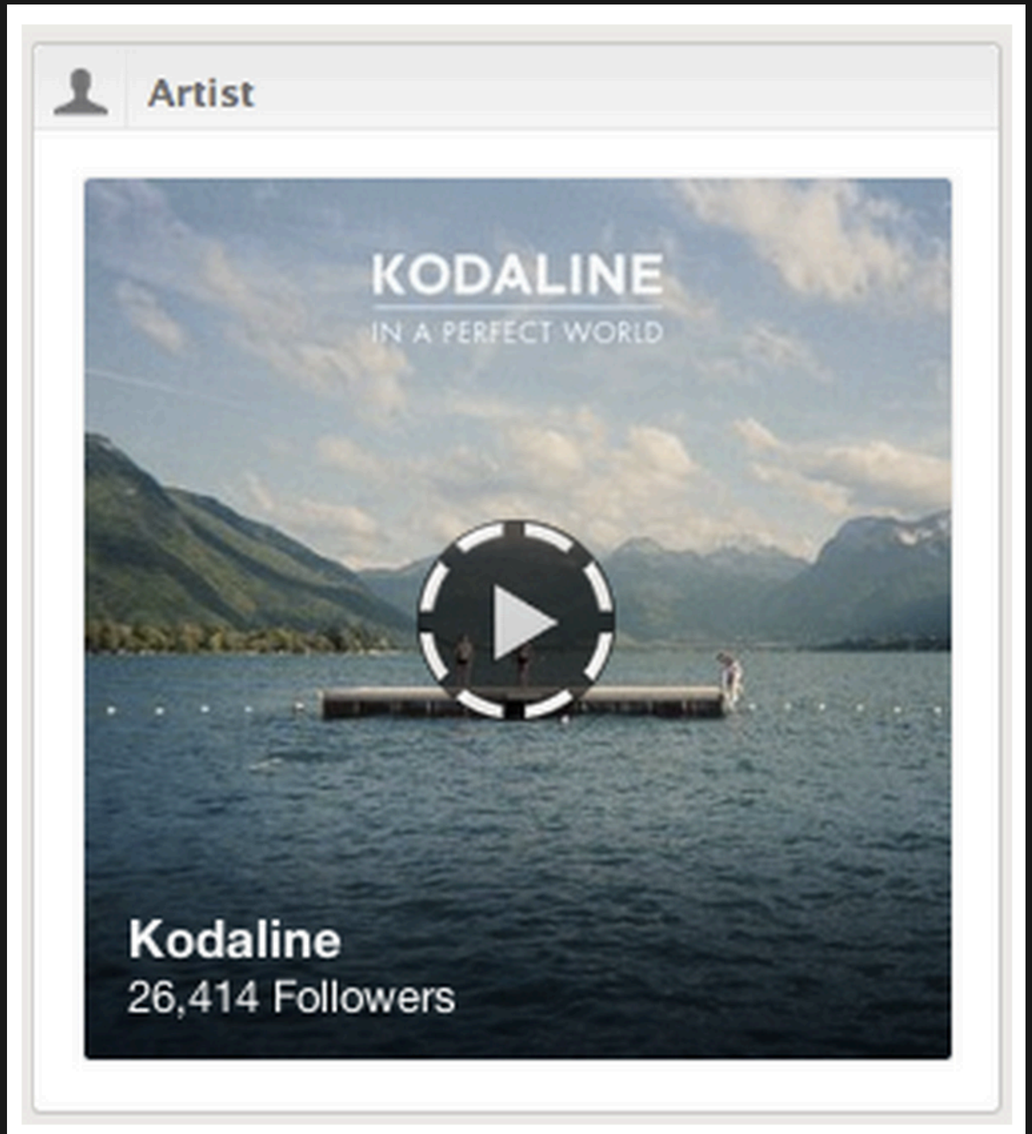
- Don’t peek
- Okay, maybe you can peek, but don’t stop or make a decision before you reach the fixed sample size stopping point
- Sequential sampling


Artist
People who listen to **Walk the Moon** are also listening to **Kodaline.**
KODALINE
IN A PERFECT WORLD
Kodaline
26,414 Followers
+ Follow Artist
Dublin, Ireland-based modern rock quartet Kodaline specializes in soaring, radio-ready guitar rock that's drawn comparisons to

Control


Artist
People who listen to **Walk the Moon** are also listening to **Kodaline.**
KODALINE
IN A PERFECT WORLD
Kodaline
26,414 Followers

A


Artist
KODALINE
IN A PERFECT WORLD
Kodaline
26,414 Followers

B

Pitfall #3: Making multiple comparisons in one test

A test can be one of two things: significant or not significant

- **P(significant) + P(not significant) = 1**
- Let's take an alpha of .05
  - P(significant) = .05
  - P(not significant) = 1 – P(significant) = 1 - .05 = .95

What about for two comparisons?

- P(at least 1 significant) = 1 - P(none of the 2 are significant)
- P(none of the 2 are significant) = P(not significant)*P(not significant) = .95*.95 = .9025
- P(at least 1 significant) = 1 - .9025 = .0975

What about for two comparisons?

- **That's almost 2x (1.95x, to be precise) your .05 significance rate!**

And it just gets worse…☹

	P(at least 1 signifcant)	An increase of…
5 variations	1 – (1-.05)^5 = .23	4.6x
10 variations	1 – (1-.05)^10 = .40	8x
20 variations	1 – (1-.05)^20 = .64	12.8x

How can we remedy this?

- **Bonferroni correction**
  - Divide P(significant), your alpha, by the number of variations you are testing, n
  - alpha/n becomes the new level of statistical significance

So what about two comparisons now?

- Our new P(significant) = .05/2 = .025
- Our new P(not significant) = 1 - .025 = .975
- P(at least 1 significant) = 1 - P(none of the 2 are significant)
- P(none of the 2 are significant) = P(not significant)*P(not significant) = .975*.975 = .951
- P(at least 1 significant) = 1 - .951 = .0499

P(significant) stays under .05 ☺

	Corrected alpha	P(at least 1 signifcant)
5 variations	.05/5 = .01	1 – (1-.01)^5 = .049
10 variations	.05/10 = .005	1 – (1-.005)^10 = .049
20 variations	.05/20 = .0025	1 – (1-.0025)^20 = .049

Questions?

Appendix

A/B test steps:

1. Decide what to test
2. Determine a metric to test
3. Formulate your hypothesis
    1. Select an effect size threshold: what change of the metric would make a rollout worthwhile?
4. Calculate sample size (your stopping point)
    1. Decide your Type I (alpha) and Type 2 (beta) error levels and the corresponding z-scores
    2. Determine the standard deviation of your metric
5. Run your test! But stop once you've reached the fixed sample size stopping point
6. Compute your z-score and compare it with the z-score for your chosen alpha level

Type I and Type II error

- Type I error: incorrectly reject a true null hypothesis
  - alpha
- Type II error: incorrectly accept a false null hypothesis
  - beta
  - Power: 1 - beta

Z-score reference table

alpha	One-sided test	Two-sided test
.10	1.28	1.65
.05	1.65	1.96
.01	2.33	2.58

Z-score for proportions (e.g. conversion)

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$