Urbana-Champaign Data Science User Group: A Beginner's Guide to Sample Size Calculations in A/B Testing

Ryan Muraglia

Oath Champaign

August 4, 2017

Problem Statement

To improve user engagement (as measured by daily time spent in app in seconds per user), you and your team just finished developing a new news article recommendation algorithm for your app.

Before rolling out the change to all of your users, you've been tasked with "proving" that the new algorithm does indeed improve user engagement.

Your manager says that you'll be able to submit a request to try out the new algorithm on some of the userbase, but to be careful with how much you request, since your request is more likely to get denied as you ask for more and more test users.

Ring a bell?

A/B testing, hypothesis testing, statistical power, confidence intervals, $\mathcal{X}^2$ test vs Z-test vs student's t-test vs ANOVA, paired vs unpaired, one-tailed vs two-tailed, type I and type II errors...

Searching for statistical significance

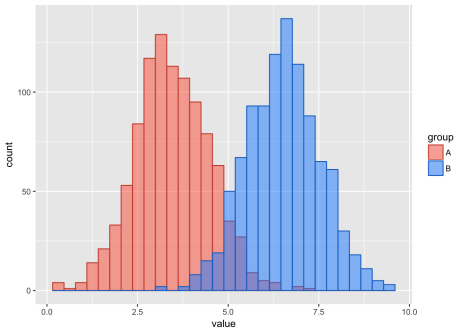
Null and alternative hypotheses

- H_0: The control and experimental groups are the same
- H_1: The control and experimental groups are different

Searching for statistical significance

Null and alternative hypotheses

- H_0: The control and experimental groups are the same
- H_1: The control and experimental groups are different



Searching for statistical significance

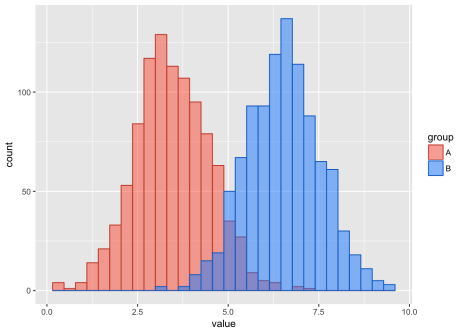
Null and alternative hypotheses

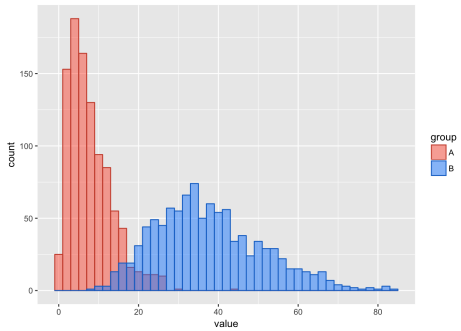
- H_0: The control and experimental groups are the same
- H_1: The control and experimental groups are different

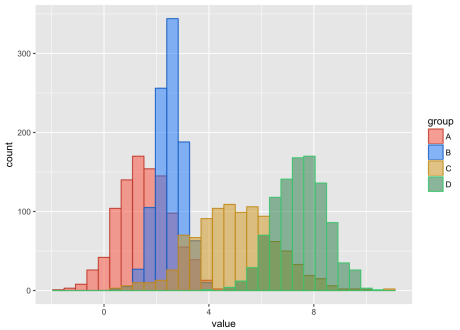
Searching for statistical significance

Null and alternative hypotheses

- H_0: The control and experimental groups are the same
- H_1: The control and experimental groups are different



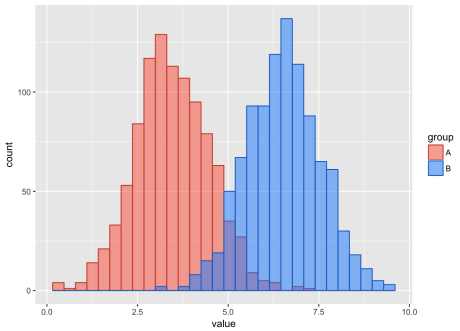


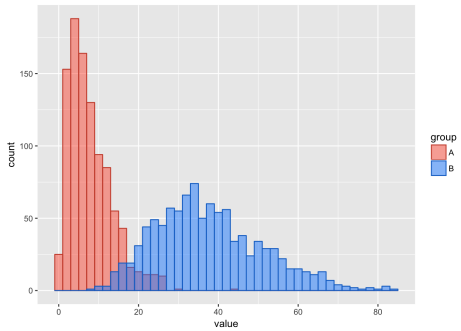


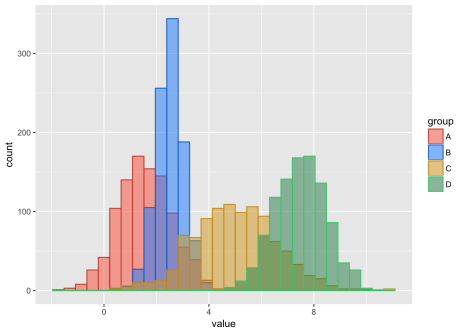
Searching for statistical significance

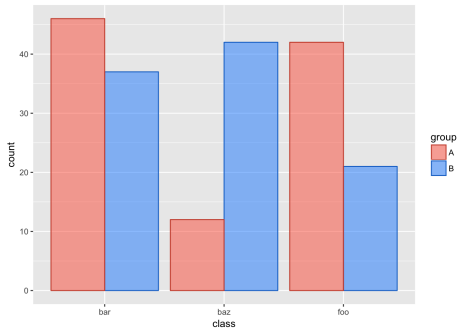
Null and alternative hypotheses

- H_0: The control and experimental groups are the same
- H_1: The control and experimental groups are different





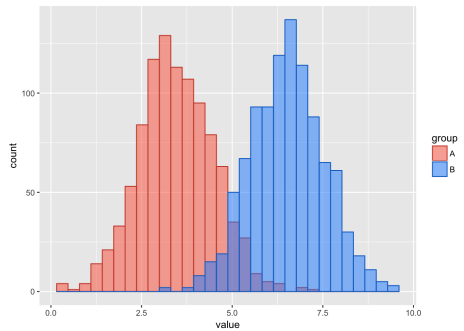


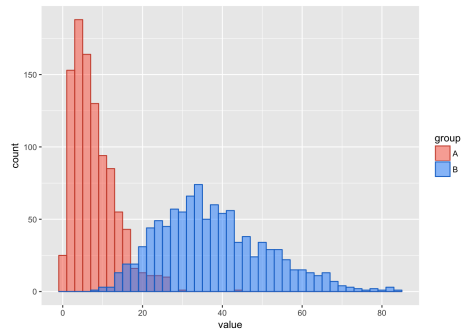


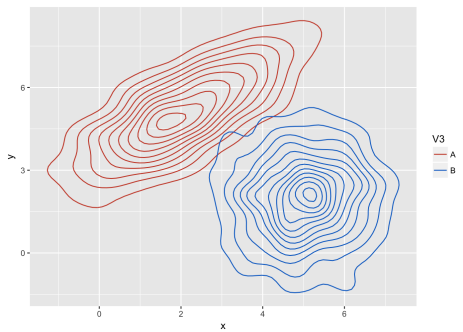
Searching for statistical significance

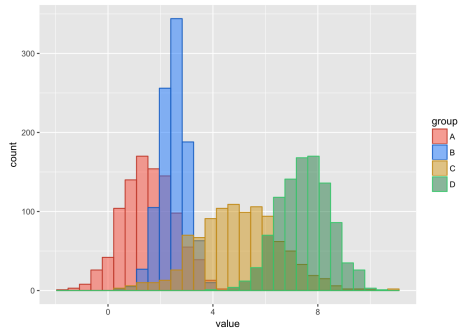
Null and alternative hypotheses

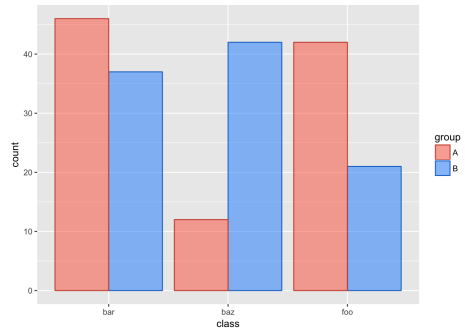
- H_0: The control and experimental groups are the same
- H_1: The control and experimental groups are different







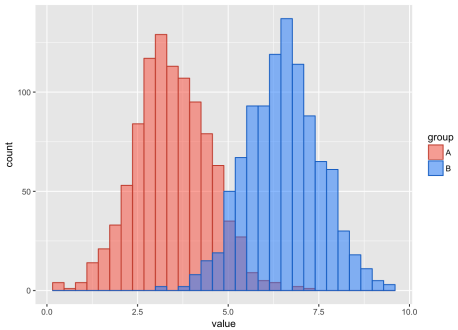


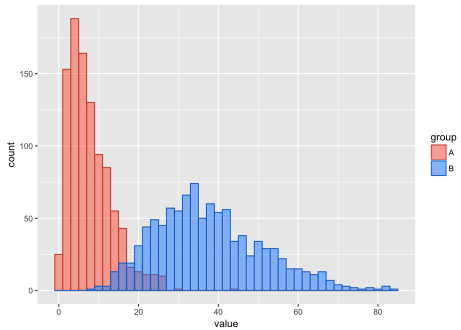


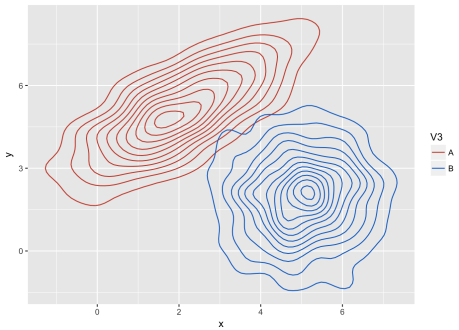
Searching for statistical significance

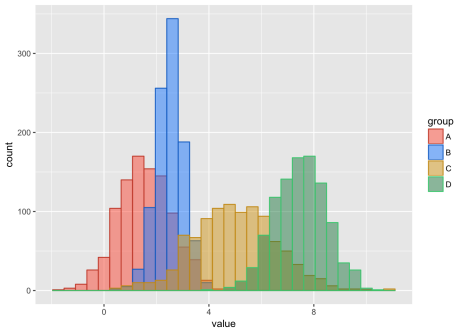
Null and alternative hypotheses

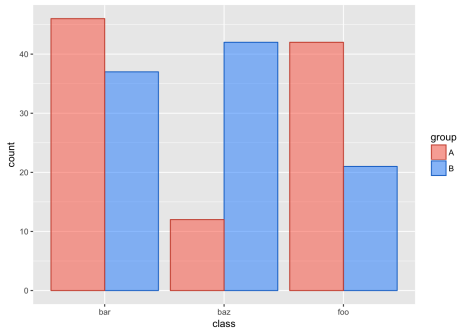
- H_0: The control and experimental groups are the same
- H_1: The control and experimental groups are different







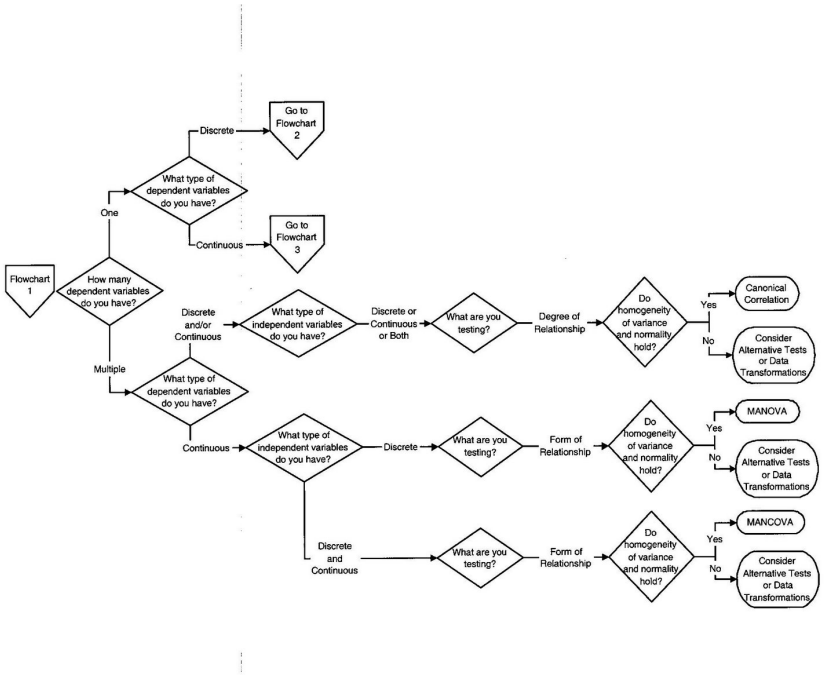


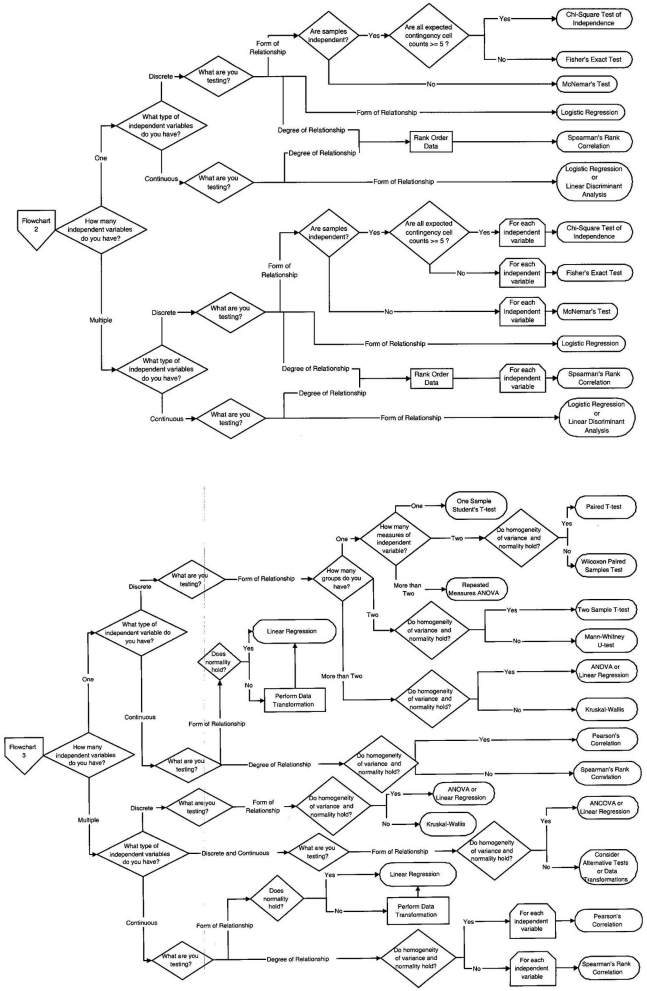


And more...

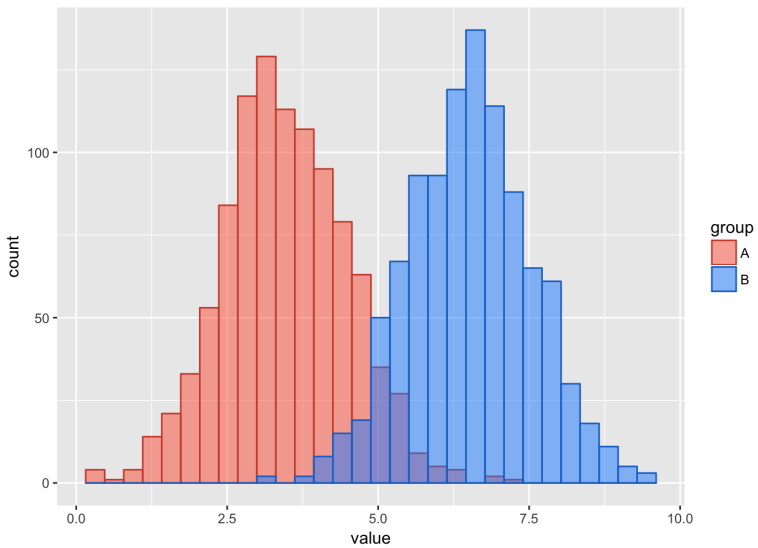
The meaning of "different" & choice of statistical test

It depends on what kind of data you have





A/B testing example



Hypotheses

- H_0: The new algorithm has no impact on average user engagement
- H_1: The new algorithm increases average user engagement

Choice of statistical test

In this case, we will use an unpaired, two-sample, one-tailed t-test.

Why this test?

unpaired: the samples in each group do not correspond to one another (they aren't before and after measurements)

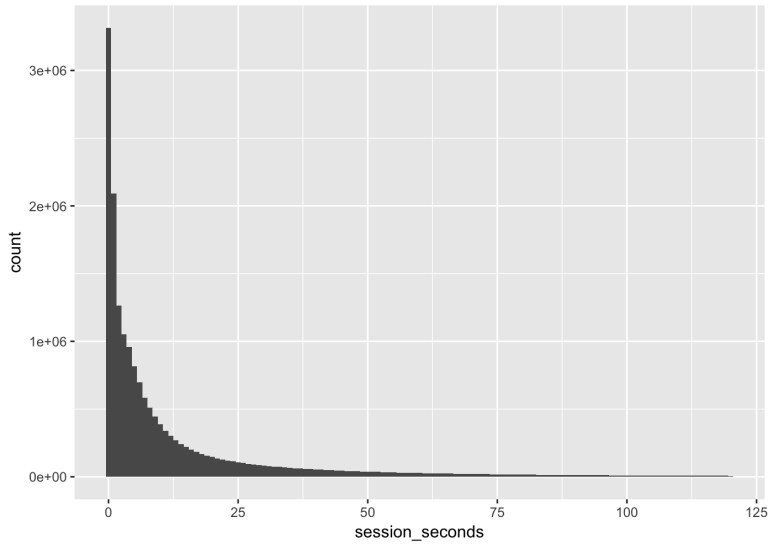
two-sample: we are comparing two population means, not one population mean to a known constant

one-tailed: we are looking for a directional change (increase), not just any deviation

t-test: appropriate for comparison of means of normally distributed data when variances are unknown

Note: when dealing with large sample sizes, the t and Z test are essentially equivalent, since degrees of freedom for t-test gets so large that the t-distribution approaches the standard normal.

But my data isn't normally distributed!

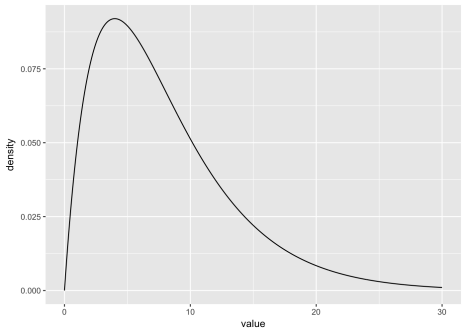


Central limit theorem

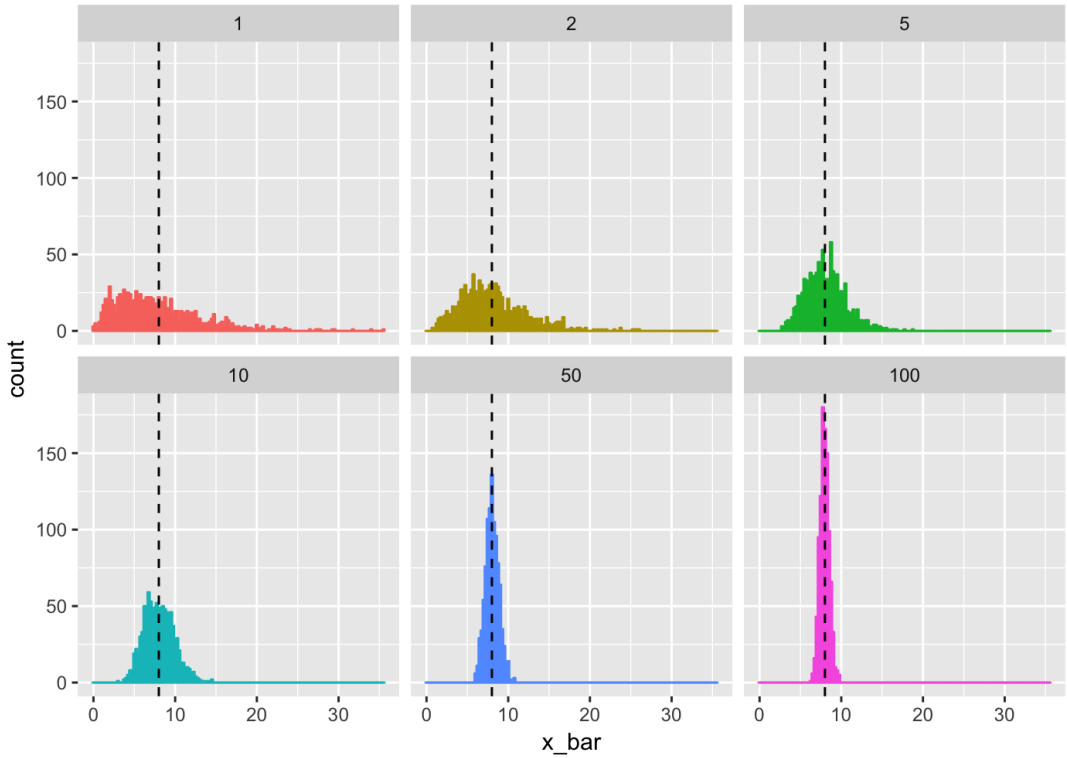
The average of a large number of iid samples will be normally distributed, as $\mathcal{N}(\mu, \sigma^2/n)$, where μ and σ^2 are the mean and variance of the underlying distribution, and n is the number of samples we are averaging over. This is true, regardless of the underlying distribution.

CLT in action

Say your underlying distribution is skewed like this...



Watch as the CLT works its magic and makes the distribution of the sampled mean increasingly normal and narrow as the number of samples increases:



Bringing it all together

What we already have:

- Normally distributed data
- A statistical test, and its associated test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

What we're still missing:

- Error rates
- Expected effect size
- Required sample size

Back to hypothesis testing

Type I error

Falsely rejecting H_0 when H_0 is true.
α denotes the probability of committing a type I error.

Type II error

Falsely accepting H_0 when H_1 is true.
β denotes the probability of committing a type II error.

As part of your experiment design, you need to choose error rates you're willing to live with.
Common values are $\alpha = 0.05$ and $\beta = 0.1$, which correspond to critical test statistic values of 1.65 and -1.28.

Look these critical values up in a Z-table – remember that we're in a large sample size regime where the t-distribution is basically normal, so we don't need to mess around with the degrees of freedom in a t-table.

Expected effect size

Effect size (Cohen's d)

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

Effect sizes help to contextualize the severity of a difference in means by scaling it by the standard deviation.

For your experiment, you will have to have some idea of the effect size you are expecting to see – a pilot experiment is a good way to get a feel for this.

Alternatively, you can calculate required sample sizes for a range of effect sizes.

Two conditions to satisfy

Recall that we have a test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Let's assume that $s_1 = s_2$ and $n_1 = n_2$ and simplify:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{2s^2/n}}$$

Type I error

We require the computed t-statistic to be $\leq$ the critical test statistic, t_α. For the null hypothesis, $\Delta = 0$.

$$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{2s^2/n}} \leq t_\alpha$$

Type II error

We require the computed t-statistic to be $\geq$ the critical test statistic, t_β. For the alternative hypothesis, $\Delta = \mu_1 - \mu_2$

$$\frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{2s^2/n}} \geq t_\beta$$

A threshold value to satisfy both conditions

This gives us a little system of equations we can solve for n:

$$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{2s^2/n}} \leq t_\alpha$$

$$\bar{x}_1 - \bar{x}_2 \leq t_\alpha \sqrt{2s^2/n}$$

Note: by construction, Δ is negative here, which is why we flip the direction of the inequality in the penultimate line.

A threshold value to satisfy both conditions

This gives us a little system of equations we can solve for n:

$$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{2s^2/n}} \leq t_\alpha$$

$$\bar{x}_1 - \bar{x}_2 \leq t_\alpha \sqrt{2s^2/n}$$

$$\frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{2s^2/n}} \geq t_\beta$$

$$t_\alpha - \frac{\Delta}{\sqrt{2s^2/n}} \geq t_\beta$$

$$t_\alpha - t_\beta \geq \frac{\Delta\sqrt{n}}{\sqrt{2s^2}}$$

$$\frac{(t_\alpha - t_\beta)\sqrt{2s^2}}{\Delta} \leq \sqrt{n}$$

$$2 * \left(\frac{(t_\alpha - t_\beta) * s}{\Delta}\right)^2 \leq n$$

Note: by construction, Δ is negative here, which is why we flip the direction of the inequality in the penultimate line.

The formula for 2-sample t-test

Minimum required sample size per group for 2-sample t-test

$$n = 2 * \left(\frac{(t_\alpha - t_\beta) * s}{\Delta} \right)^2$$

$$n = 2 * \left(\frac{(t_\alpha - t_\beta)}{d} \right)^2$$

- n is the # of samples
- t_α, t_β are critical test statistic values
- s is the sample std dev
- Δ is the diff of group means

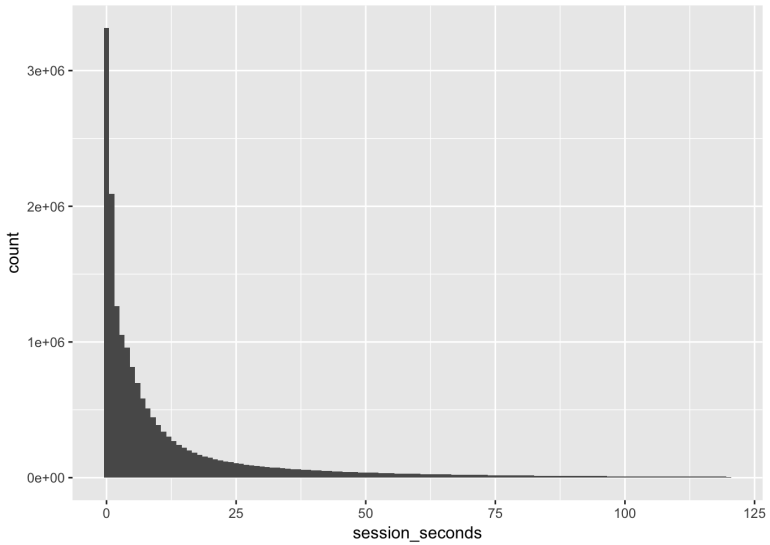
Note that the following all cause n to increase:

- Decrease in effect size, d (Δ decreasing, or s increasing)
- Decrease in leniency for type I error (t_α increasing)
- Decrease in leniency for type II error (t_β decreasing)

Note that the CLT plays a two-fold role: it both normalizes our data and decreases the variance (s^2) as the sample size increases.

Our data and test parameters

The distribution of time spent for 18.4 million users looks like:



Super skewed, super high standard deviation:
$\bar{x} = 27.848$
$s = 168.73$

We will be somewhat relaxed about our type I and type II error requirements:

- $\alpha = 0.1 \rightarrow t_\alpha = 1.28$
- $\beta = 0.2 \rightarrow t_\beta = -0.84$

We aren't sure what sort of effect size to expect, so we'll compute required sample sizes for a range and see what we think we can get away with.

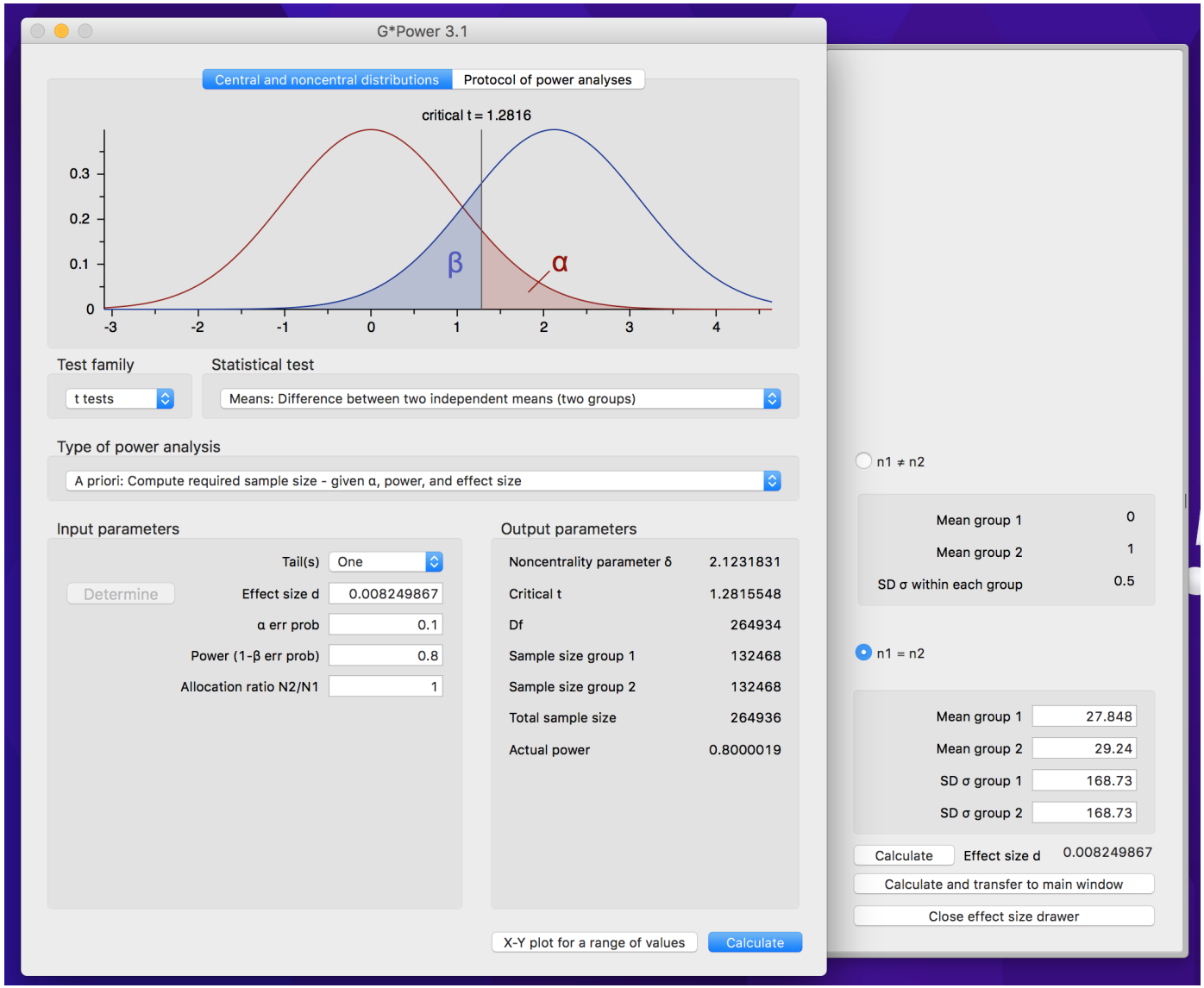
Formula reminder

$$n = 2 * \left(\frac{(t_\alpha - t_\beta) * s}{\Delta} \right)^2$$

The following parameters are invariant: $t_\alpha = 1.28$, $t_\beta = -0.84$, $\bar{x} = 27.848$ and $s = 168.73$.

% $\nearrow$ sess. time	$\lvert\Delta\rvert$	n (each group)	% of users to request
0.1	0.028	329,988,124	3,595
0.5	0.14	13,199,525	143.8
1	0.28	3,299,881	36
2	0.56	824,970	9
5	1.39	131,995	1.4
10	2.78	32,999	0.36

Using G*Power



In R

```
[R]> mu1 <- 27.848
[R]> mu2 <- mu1 * 1.05
[R]> sigma <- 168.73
[R]>
[R]> alpha <- 0.1
[R]> beta <- 0.2
[R]>
[R]> # in base R:
[R]> power.t.test(delta = (mu2 - mu1),
+                 sd = sigma,
+                 sig.level=alpha,
+                 power=(1-beta),
+                 type='two.sample',
+                 alternative='one.sided')

     Two-sample t test power calculation

              n = 132391
          delta = 1.3924
             sd = 168.73
      sig.level = 0.1
          power = 0.8
    alternative = one.sided

NOTE: n is number in *each* group
```

```
[R]> # or with the pwr package:
[R]> library(pwr)
[R]> pwr.t.test(d = (mu1-mu2)/sigma,
+               sig.level=alpha,
+               power=(1-beta),
+               type='two.sample',
+               alternative='less')

     Two-sample t test power calculation

              n = 132391
              d = -0.008252237
      sig.level = 0.1
          power = 0.8
    alternative = less

NOTE: n is number in *each* group
```

Useful links

Recommended tools:

- Standalone software: `http://www.gpower.hhu.de/en.html`
- pwr package for R: `https://cran.r-project.org/web/packages/pwr/index.html`

References for learning:

- `https://onlinecourses.science.psu.edu/stat414/node/306` (check the preceding lessons too)
- `http://www.ysumathstat.org/faculty/chang/class/s5817/L/L5817_1_2_PowerSampleSize_n.pdf`
- `https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Common_test_statistics`
- `https://github.com/rmuraglia/dsug/blob/master/201708_sample_size/dsug_sample_size_201708.pdf`