```
## [1] 79157.54
```

```
conf95_max
```

```
## [1] 104367
```

2.5 Non-parametric methods

This is a way to analyze data without making an assumption about the distribution. At Google, it was observed that the analytical estimates of variance was often under-estimated, and therefore they have resorted to use empirical measurements based on A/A test to evaluate variance. If you see a lot of variability in a metric in an A/A test, it is probably too sensitive to be used. Rather than do several multiple A/A tests, one way is to do a large A/A test, and then do bootstrap to generate small groups and test the variability.

With A/A tests, we can

1. Compare result to what you expect (sanity check)
2. Estimate variance empirically and use your assumption about the distribution to calculate confidence
3. Directly estimate confidence interval without making any assumption of the data

In summary, different metrics have different variability. The variability may be high for certain metrics which makes them useless even if they make business or product sense. Computing the variability of a metric is tricky and one needs to take a lot of care.

For a lot of analysts, a majority of the time is spent is validating and choosing a metric compared to actually running the experiment. Being able to standardize the definitions was critical in the test. When measuring latency, are you talking about when the first byte loads and when a last byte loads. Also, for latency, the mean may not change at all. The signals (e.g. slow/fast connections or browsers) causes lumps in the distribution, and no central measure works. One needs to look at the right percentile metric. The key thing is that you are building intuition, you have to understand data, and the business, and work with the engineers to understand how the data is being captured.

3 Designing an Experiment

1. Choose subject: What are the units in the population you are going to run the test on? (unit of diversion)
2. Choose population: What population are you going to use (US only?)
3. Size
4. Duration

test how to change, compute evaluation metrics) on equivalent population

Typically you want to assign people and not events since the same user may see different changes. If you use a person, you typically use a cookie which may change by platform. The alternative then is to use a user id .

3.1 Unit of Diversion

Commonly used units of diversion are:

1. User identifier (id): Typically the username or email address used on the website. It is typically stable and unchanging. If user id is used as a unit of diversion, then it is either in the test group or the control group. User ID is personally identifiable
2. Anonymous id: This is usually an anonymous identifier such as a cookie. It changes with browser or device. People may often refresh their cookies every time they log in. It is difficult to refresh a cookie on an app or a phone compared to the computer.
3. Event: An event is a page load that can change for each user. This is used typically for changes that is not user facing.

Lesser used units of diversion are

4. Device id: Typically available for mobile devices. It is tied to a specific device and cannot be changed by the user.
5. IP address: The ip address is location specific, but may change as the user changes location (e.g. testing on infrastructure change to test impact on latency)

3 main considerations in selecting an appropriate unit of diversion

1. Consistency
2. Ethical
3. Variability

Variability is higher when it is calculated empirically than when calculated analytically. This is because the unit of analysis (i.e. the denominator in the metric) is different from the unit of variability.

E.g. If unit of diversion is a query, then coverage (= #queries with ads/ # queries) will have lower variability compared to using a cookie as a unit of diversion. This is because when a query is used, the unit of diversion matches the unit of analysis (which is the denominator of the metric i.e. query)

In the medical industry, users are paired with each other based on location, demographics. However, given how little information there is on users on the internet, this is not widely practiced.

e.g. Consider an experiment where you are analyzing data for a particular region (NZ), and for the rest of the world. For the global data that includez NZ and the rest of the world, what is the pooled standard error?

For the other, you have

```
N_cont = 50000 + 6021
X_cont = 2500 + 302
N_exp = 50000 + 5979
X_exp = 2500 + 374

p_cont = X_cont/N_cont
p_cont
```

```
## [1] 0.05001696
```

```
p_exp = X_exp/N_exp
p_exp
```

```
## [1] 0.05134068
```

```
p_pool = (X_cont + X_exp)/(N_cont + N_exp)
se_pool = sqrt(p_pool*(1-p_pool)*(1/N_cont + 1/N_exp))
se_pool
```

```
## [1] 0.00131081
```

Since $abs(p_{cont} - p_{exp}) < 1.96 * se_pool$, the global difference is not statistically significant

3.2 Population vs Cohort

A cohort is like an entering class for an analysis. A cohort may make more sense to look at a population when:

1. Looking for leaning effects
2. Examining user retention
3. Want to increase user activity
4. Anything that requires the user to be established

Practical considerations in experimental design

1. Duration
2. When to run the experiment
3. Fraction of the traffic to send to the experiment

Two different types of learning effects

1. Change aversion
2. Knowledge effect

When users first encounter a change they will react, but will eventually adapt to a change.

4 Analyzing Results

4.1 Sanity Tests

One of the first things to do once you finish collecting experimental data is to analyze the invariants. This is done by calculating the values for one or more invariants on the test and control group, and check if the difference is statistically significant. For e.g. if the values for an invariant (say total # of cookies) are x and y, then calculate the se as $\sqrt{\frac{0.5*0.5}{x+y}}$, since one would expect the same number of cookies in both groups. Then calculate the margin as $1.96 * se$. If the margin is greater than $x/(x+y) - y/(x+y)$, then the difference of the invariant is insignificant. However if the difference is greater than the margin, then the difference is insignifiant and needs to be investigated further

An example is provided below:

```
control_event_ct = c(2451,2475,2394,2482,2374,1704,1468)
test_event_ct = c(2404,2507,2376,2444,2504,1612,1465)
control_total = sum(control_event_ct)
test_total= sum(test_event_ct)
p_cont = control_total/(control_total+test_total)
p_test = test_total/(control_total + test_total)
p_cont
```

```
## [1] 0.5005871
```

```
p_test
```

```
## [1] 0.4994129
```

```
se = sqrt(0.5*0.5/(control_total+test_total))
margin = 1.96*se
p_cf_min = 0.5 - margin
p_cf_max = 0.5 + margin
p_cf_min
```