Lecture 7: Exam Preparation

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

December 18, 2023

Plan

Plan

1) A/B testing

2) Parametric vs Non-parametric + Bootstrap

3) DiD + Matching

4) Variance reduction

A/B testing

A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

A lot of versions:

- **A/A testing** uses A/B testing to test two identical versions of a page against each other. Typically, this is done to check that the tool being used to run the experiment is statistically fair.
- **A/B/C** and
- **A/B/n testing** is an extension of A/B testing in which multiple variants of a page are compared against each other.

A/B testing

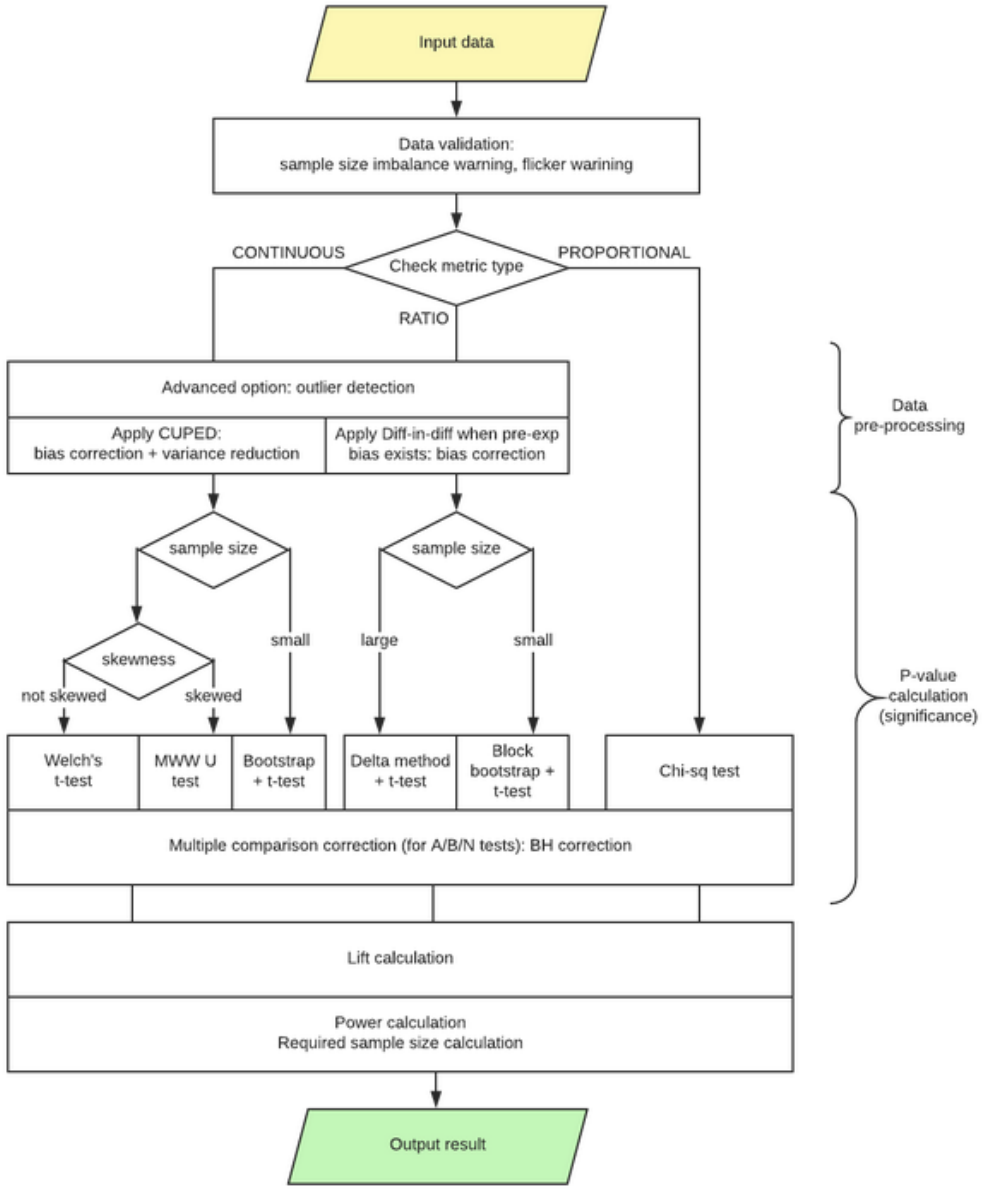


Figure: Uber's statistics engine is used for A/B/N experiments

Count metrics vs ratio metrics

Count metrics

For count metrics, the unit of analysis is the same as the unit of randomization. For example, if the unit of analysis is a user, count metrics would include *revenue per user, clicks per user, etc.*

In mathematical notation, let Y denote variable of interest. The count metric we want to estimate (i.e. the estimand) is $\widehat{M} = \frac{\sum_i Y_i}{n}$

Ratio metrics

For ratio metrics, the unit of analysis is at a more granular level than the unit of randomization. For example, when the unit of randomization is a user, ratio metrics would include *revenue per session, clicks per page view, etc.*

Let Y denote the variable of interest, and let Z denote the number of units of analysis for this randomization unit. The ratio metric we want to estimate is

$$R = \frac{\mathbb{E}[Y]}{\mathbb{E}[Z]}.$$

The most common way to estimate this is to replace both the numerator and denominator with the respective sample means:

$$\widehat{R} = \frac{\sum_i Y_i/n}{\sum_i Z_i/n} = \frac{\sum_i Y_i}{\sum_i Z_i}$$

Count metrics vs ratio metrics

1. User metrics

$$OEC_A = \text{avg}_{u \in A} X(u)$$

Example:

$$\frac{3+0+6+\ldots+5+6}{N} = \frac{3+0+6+\ldots+5+6}{1+1+1+\ldots+1+1}$$

2. Ratio metrics:

$$OEC_A = \frac{\sum_{u \in A} X(u)}{\sum_{u \in A} Y(u)}$$

Example:

$$\frac{3+0+6+\ldots+5+6}{9+8+7+\ldots+7+8}$$

Count metrics vs ratio metrics

How to conduct t-test for Ratio metrics? For T-test you need to be able to calculate,

$$R = \frac{X}{Y} \quad \mathrm{E}[R] = ? \quad \mathrm{Var}(R) = ?$$

Linearization Ratio

Taylor series up to 1st order:

$$f(a,b) + (x-a)f_x(a,b) + (y-b)f_y(a,b)$$

Expansion of $R = \frac{X}{Y}$ at point $(\mathrm{E}[X], \mathrm{E}[Y])$:

$$Z = \frac{\mathrm{E}[X]}{\mathrm{E}[Y]} + \frac{1}{\mathrm{E}[Y]}\left(X - \frac{\mathrm{E}[X]}{\mathrm{E}[Y]}Y\right)$$

$$E[Z] \approx \mathrm{E}[R] \quad \mathrm{Var}(Z) \approx \mathrm{Var}(R)$$

This formula converts two samples (numerator and denominator) into one, preserving the mean and variance (asymptotically), which allows the use of classical tests.

Count metrics vs ratio metrics

Consider $Z = \frac{X}{Y}$ with

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}\right)$$

Consider random variables R and S where S either has no mass at 0 (discrete) or has support $[0, \infty)$. Let $G = g(R, S) = R/S$.

Find approximations for EG *and* $\mathrm{Var}(G)$ *using Taylor expansions of* $g()$.

For any $f(x, y)$, the bivariate first order Taylor expansion about any $\boldsymbol{\theta} = (\theta_x, \theta_y)$ is

$$f(x, y) = f(\boldsymbol{\theta}) + f'_x(\boldsymbol{\theta})(x - \theta_x) + f'_y(\boldsymbol{\theta})(y - \theta_y) + \mathbf{R}$$

where R is a remainder of smaller order than the terms in the equation.

Count metrics vs ratio metrics

Switching to random variables with finite means $EX \equiv \mu_x$ and $EY \equiv \mu_y$, we can choose the expansion point to be $\theta = (\mu_{\mathbf{x}}, \mu_{\mathbf{y}})$. In that case the first order Taylor series approximation for $f(X, Y)$ is

$$f(X, Y) = f(\boldsymbol{\theta}) + f'_x(\boldsymbol{\theta})(X - \mu_x) + f'_y(\boldsymbol{\theta})(Y - \mu_y) + R$$

The approximation for $E(f(X, Y))$ is therefore

$$\begin{aligned} E(f(X, Y)) &= E\left[f(\boldsymbol{\theta}) + f'_x(\boldsymbol{\theta})(X - \mu_x) + f'_y(\boldsymbol{\theta})(Y - \mu_y) + R\right] \\ &\approx E[f(\boldsymbol{\theta})] + E\left[f'_x(\boldsymbol{\theta})(X - \mu_x)\right] + E\left[f'_y(\boldsymbol{\theta})(Y - \mu_y)\right] \\ &= E[f(\boldsymbol{\theta})] + f'_x(\boldsymbol{\theta})E\left[(X - \mu_x)\right] + f'_y(\boldsymbol{\theta})E\left[(Y - \mu_y)\right] \\ &= E[f(\boldsymbol{\theta})] + 0 + 0 \\ &= f(\mu_x, \mu_y) \end{aligned}$$

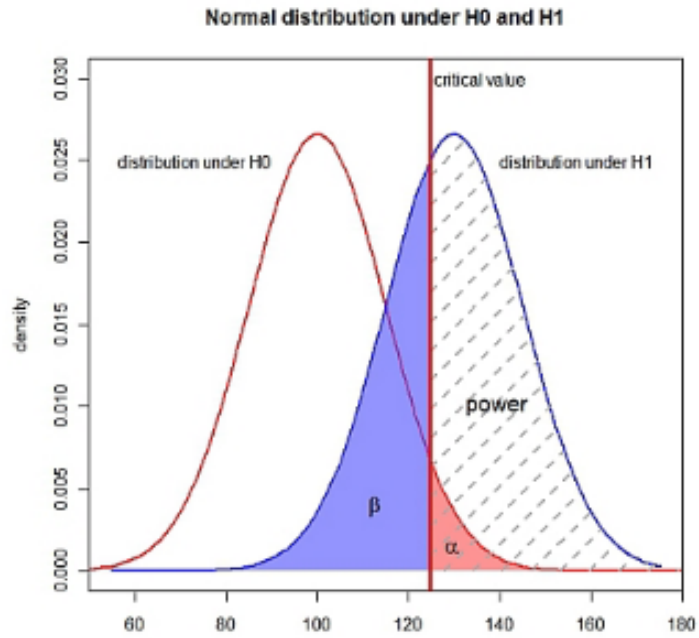
For our example where $f(x, y) = x/y$ the approximation is $E(X/Y) = E(f(X, Y)) = f(\mu_x, \mu_y) = \mu_x/\mu_y$

Count metrics vs ratio metrics

Hypothesis testing

		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	True Negative Confidence: $P\left(S^{obs} \not\subset D^{crit} \mid H_0\right) = 1 - \alpha$	False Negative Type II error: $P\left(S^{obs} \not\subset D^{crit} \mid H_1\right) = \beta$
	Reject	False Positive Type I error: $P\left(S^{obs} \subset D^{crit} \mid H_0\right) = \alpha$	True Positive Power: $P\left(S^{obs} \subset D^{crit} \mid H_1\right) = 1 - \beta$

(!) Type I error and Type II errors are inversely related



Count metrics vs ratio metrics

4 pieces of a puzzle:

- Effect size
- Sample size
- Significance
- Statistical power

Goal: Compare t-test, Welch test t-test, Mann-Whitney U-test

Count metrics vs ratio metrics

Minimal sample size to estimate a population proportion

$$n = \frac{\hat{p}(1-\hat{p})z^2}{m^2},$$

where $\hat{p}$ is the estimated proportion, m the margin of error and z the z-score corresponding to the selected confidence level (for example $z = 1.96$ for a confidence level of 95%).

Minimal sample size to estimate a population mean

$$n = \frac{\sigma^2 z^2}{m^2},$$

where σ is the (expected) standard deviation in the population, m the margin of error and z the z-score corresponding to the selected confidence level (for example $z = 1.96$ for a confidence level of 95%).

If the resulting sample size represents 10% or more of the population, the finite population correction (fpc) should be applied. [6] The required sample size should then be calculated from the formula:

$$n_e = \frac{nN}{(N+n-1)}$$

where n = sample size without fpc n_c = sample size with fpc.

Count metrics vs ratio metrics

If we assume the simple random sampling is without replacement, then the sample values are not independent, so the covariance between any two different sample values is not zero. In fact, one can show that

Covariance between two different sample values:

$$\text{cov}\left(X_i, X_j\right) = -\frac{\sigma^2}{N-1} \text{ for } i \neq j$$

This fact is used to derive these formulas for the standard deviation of the estimator and the estimated standard deviation of the estimator. The first two columns are the parameter and the statistic which is the unbiased estimator of that parameter.

		standard deviation of the estimator	estimator of the standard deviation of the estimator
μ	$\bar{X}$	$\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right)}$	$\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$
p	p	$\sqrt{\left(\frac{p(1-p)}{n}\right)\left(1-\frac{n-1}{N-1}\right)}$	$\sqrt{\frac{p(1-p)}{n}\left(\frac{n}{n-1}\right)\left(1-\frac{n}{N}\right)}$

Correction factor: $fpc = \sqrt{\frac{\mathbf{N}-\mathbf{n}}{\mathbf{N}-1}}$ where N is the population size and n is the sample size.

Count metrics vs ratio metrics

Consider the following simple regression to evaluate the impact of a randomized intervention (D):

$$Y_i = \alpha + \beta D_i + \varepsilon_i;$$

where D_i is the treatment variable and ε_i is an error term, both defined for individual i. Y_i is the outcome of interest and β is the impact of the intervention.

If we assume that the observations are independent of each other and are identically distributed, then the variance of the treatment effect can be written:

$$\text{Var}(\hat{\beta}) = \frac{1}{P(1-P)} \frac{\sigma^2}{N}$$

where σ^2 is the variance of the outcome of interest, N is the sample size and P is the fraction of the sample assigned to the treatment group.

Count metrics vs ratio metrics

The MDE is estimated as,

$$\text{MDE}(k, \alpha, N, P) = \left(t_{(1-k)} + t_{\alpha}\right) * \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}}$$

where $t_{1-k)}$ – quantile for the power (1-Type II error rate) and t_{α} is the quantile for Type I error rate.

Power can be computed as,

$$t_{(1-k)}\left(N, \alpha, \beta_E, P\right) = \frac{\beta_E}{\sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}}} - \tau_{\alpha};$$

where β_E is the effect size, or the expected change in the outcome as a consequence of the intervention.

– Notice that the MDE is also an effect size, but it is the minimum effect size for a given level of power and sample size.

– MDE is an estimate, the effect size is a parameter (an assumption made at the baseline).

The sample size can be estimated as,

$$N\left(k, \alpha, \beta_E, P\right) = \left[\frac{\sigma^* \left(t_{(1-k)} + t_{\alpha}\right) * \sqrt{\frac{1}{P(1-P)}}}{\beta_E}\right]^2;$$

Count metrics vs ratio metrics

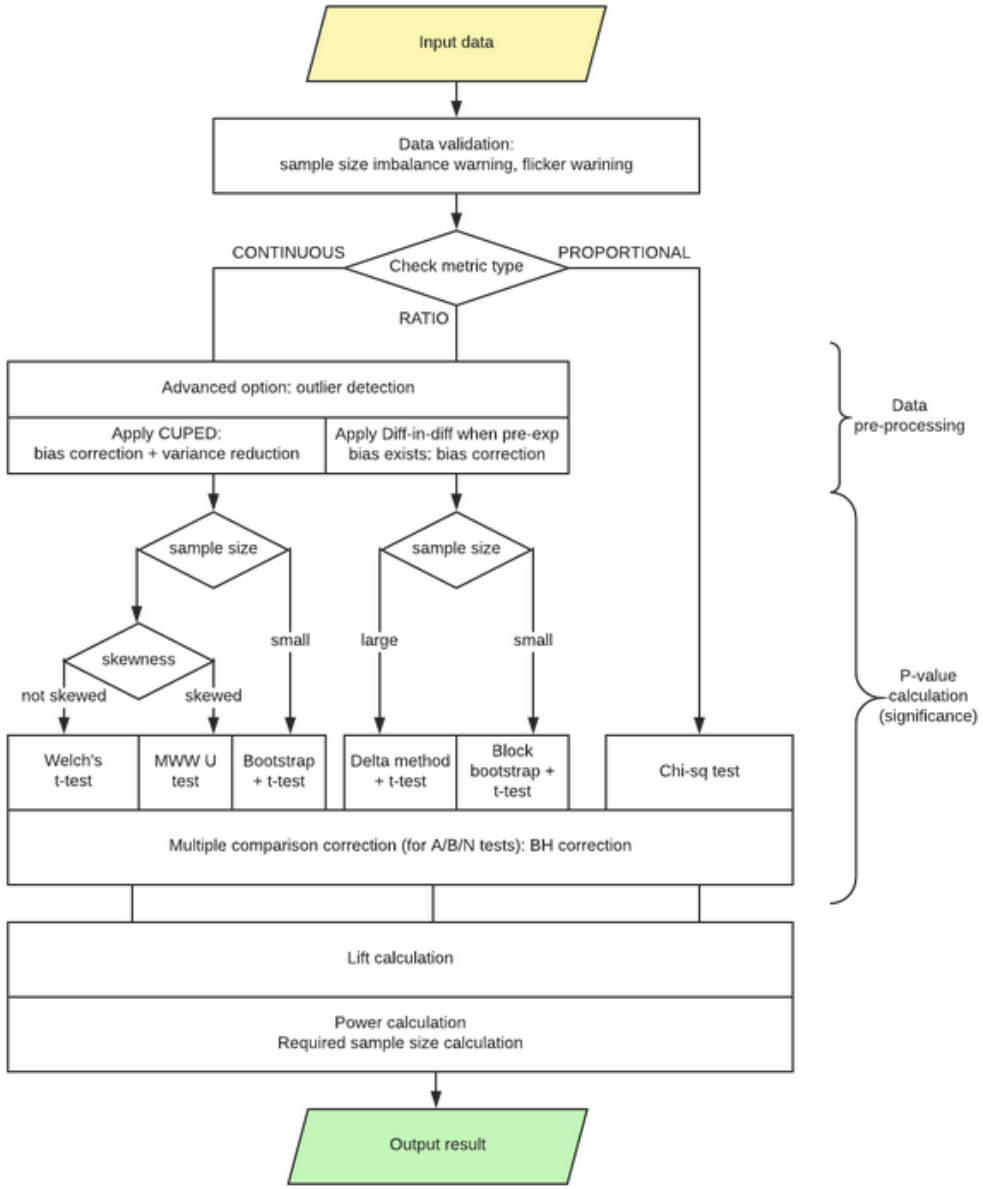


Figure: Uber's statistics engine is used for A/B/N experiments

Two-sample t-test vs Welch test

Welch test – unequal variances

The two-sample t-test for unpaired data is defined as:

$$\begin{array}{ll} \mathrm{H}_0: & \mu_1 = \mu_2 \\ \mathrm{H}_a: & \mu_1 \neq \mu_2 \\ \text{Test Statistic:} & T = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} \end{array} \tag{1}$$

where N_1 and N_2 are the sample sizes, $\bar{Y}_1$ and $\bar{Y}_2$ are the sample means, and s_1^2 and s_2^2 are the sample variances.

Critical Region: Reject the null hypothesis that the two means are equal if $|T| > t_{1-\alpha/2,\nu}$ where $t_{1-\alpha/2,\nu}$ is the critical value of the t distribution with ν degrees of freedom where

$$v = \frac{\left(s_1^2/N_1 + s_2^2/N_2\right)^2}{\left(s_1^2/N_1\right)^2/(N_1-1) + \left(s_2^2/N_2\right)^2/(N_2-1)}$$

t-test – equal variances

If equal variances are assumed, then the formula reduces to:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{1/N_1 + 1/N_2}}$$

where

$$s_p^2 = \frac{(N_1-1)\,s_1^2 + (N_2-1)\,s_2^2}{N_1 + N_2 - 2}$$

If equal variances are assumed, then $\nu = N_1 + N_2 - 2$

Two-sample t-test vs Welch test

- Welch's t-test is more robust than Student's t-test and maintains type I error rates close to nominal for unequal variances and for unequal sample sizes under normality.
- Student's t test is may be preferred if the **sample variances are not far away** from each other (e.g., requiring the variances not to differ by a factor of more than 2)
- The power of Welch's t-test comes close to that of Student's t-test, even when the population variances are equal and sample sizes are balanced.
- Neither test is exact except in the case of Student's t test where the population variances are equal. This means that the **type I error rate will deviate** somewhat from the desired significance level α.
- As the sample sizes of both groups become large, both tests are very close to the two-sample z test (Samples larger than 150-200 or so, there is a negligible difference between the results)
- It is not recommended to **pre-test for equal variances** and then choose between Student's t-test or Welch's t-test. Rather, Welch's t-test can be applied directly and without any substantial disadvantages to Student's t-test as noted above.
- Welch's t-test remains robust for skewed distributions and large sample sizes.

Non-parametric statistics

Any statistical analysis, including those outside of the scope of this course, involves the combination of:

- information that comes from assumptions, and
- information that comes from data

Assumptions that we typically have to contend with include those regarding:

- representativeness of the sample
- accuracy of the data
- underlying relationship(s) between key covariates

Non-parametric statistics

One way of classifying statistical analysis methods is on the basis of the extent of parametric assumptions regarding T, and how it depends on $\boldsymbol{X}$ or, more formally, regarding the conditional distribution of $T \mid \boldsymbol{X}$

For example, consider a linear regression analysis for a continuous response, T, based on the following components/assumptions:

1. mean model: $\mathrm{E}[T_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i^T \boldsymbol{\beta}$
2. error term: $\epsilon_i = T_i - \mathrm{E}[T_i \mid \boldsymbol{X}_i]$
3. the ϵ_i 's are independent
4. $\epsilon_i \sim \text{Normal}\left(0, \sigma^2\right)$

Implicitly, this specification corresponds to assuming:

$$T_i \mid \boldsymbol{X}_i \sim \text{Normal}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}, \sigma^2\right)$$

This is a fully parametric analysis because it characterizes the entire distribution of $T \mid \boldsymbol{X}$

Non-parametric statistics

Now suppose we relax the 4^{th} component of the previous specification as follows:

1. mean model: $\mathrm{E}[T_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i^T \boldsymbol{\beta}$
2. error term: $\epsilon_i = T_i - \mathrm{E}[T_i \mid \boldsymbol{X}_i]$
3. the ϵ_i 's are independent
4. $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$

From this, all we know about the distribution of the outcome is that:

$$\mathrm{E}[T_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i^T \boldsymbol{\beta}$$
$$\mathrm{Var}[T_i \mid \boldsymbol{X}_i] = \sigma^2$$

Thus, this model is semi-parametric $\star$ there is some structure, specifically in how the mean and variance vary (or not) across levels of $\boldsymbol{X}$ $\star$ structure does not give us the entire distribution

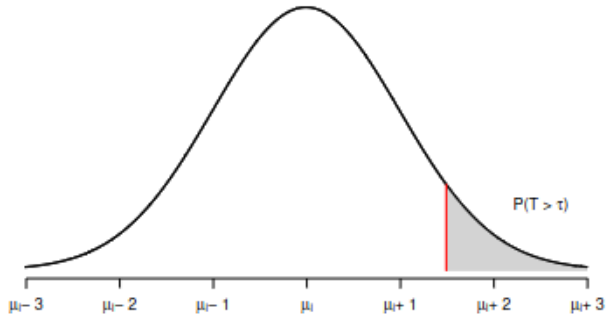
Non-parametric statistics

One benefit of specifying a fully parametric analysis is that we then can use a likelihood-based analysis such as maximum likelihood

$\star$ if the model specification is correct, such an analysis is the 'best'

A second benefit is that, following estimation, one can use the results to calculate a whole range of potentially interesting quantities

$\star$ e.g. the probability that, for a given covariate profile, the outcome is greater than some (clinically-relevant) threshold, say τ



Non-parametric statistics

If, however, it is not the case that all of the assumptions that underpin the parametric specification hold and yet we assume it to be the case, then we are not guaranteed to get valid results

- potential for bias
- potential for incorrect inference

For example, if we assume $\epsilon_i \sim \text{Normal}\left(0, \sigma^2\right)$ in a linear regression analysis but this is not the case then standard errors, and hence confidence intervals and p-values, will, in general, be biased

If we are uncertain, it may be 'safest' to proceed with the semi-parametric model

$\star$ although you still have to contend with whether $\mathrm{E}\left[\epsilon_i\right] = 0$ and $\text{Var}\left[\epsilon_i\right] = \sigma^2$ truly hold

If this is the path that is pursued, it's important to note that there are some potential drawbacks

Non-parametric statistics

Moreover, because the specification does not tell us everything about the distribution we are limited in what we can (easily) estimate

This may not be a big deal, however, if we only care about β

Another consequence, however, is that we cannot use a likelihood-based analysis

$\star$ we need other statistical tools

$\star$ e.g. in the case of the linear regression analysis we use least squares estimation

Non-parametric analyses

A non-parametric statistical analysis procedure places no structure on the distribution of T

$\star$ in terms of the shape as well as how differences across levels of $\boldsymbol{X}$ manifest

In general, a non-parametric statistical analysis procedure in the context of this course places no structure on the distribution of the response variable

$\star$ i.e. on $T_i \mid \boldsymbol{X}_i$

Non-parametric statistics

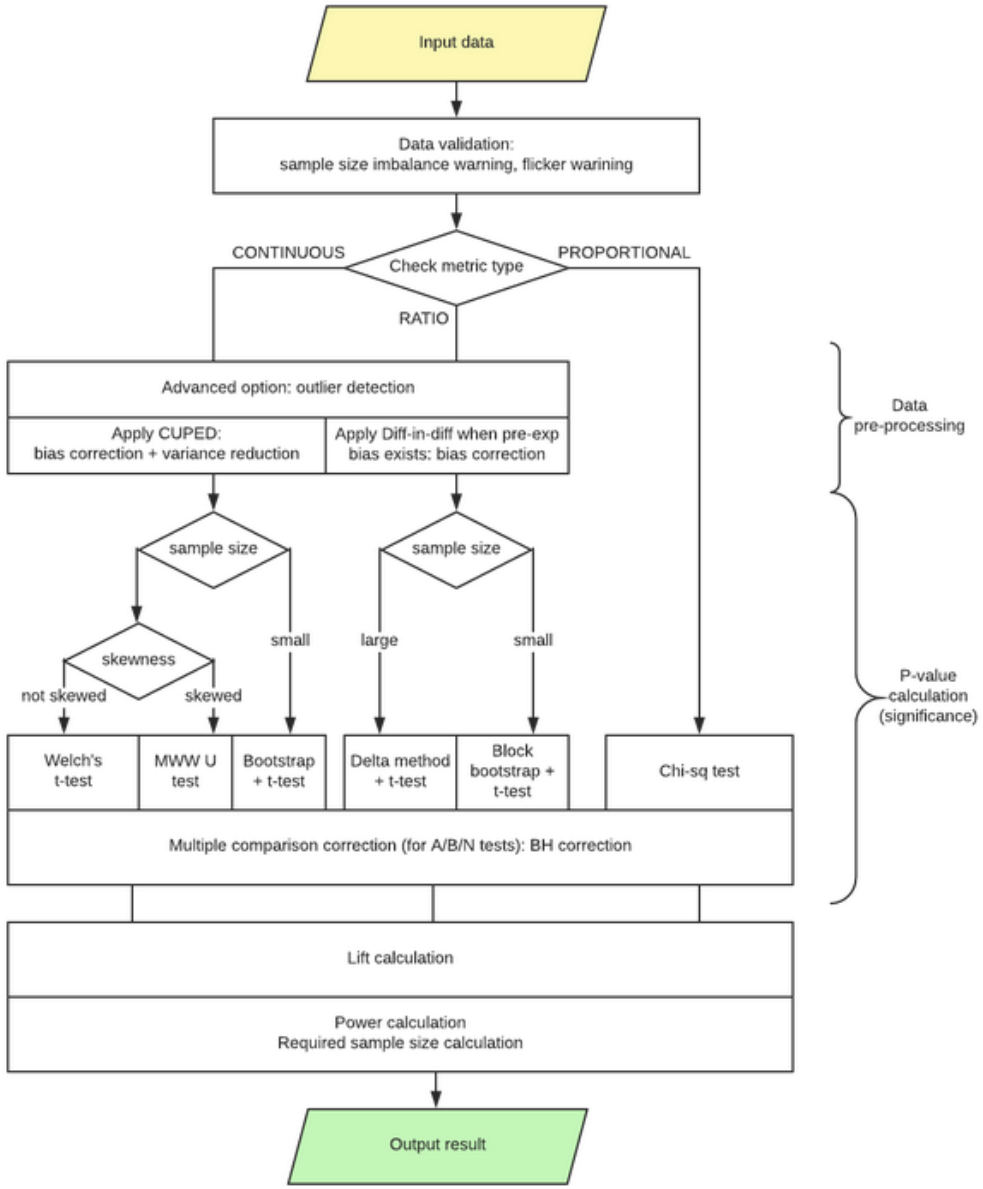


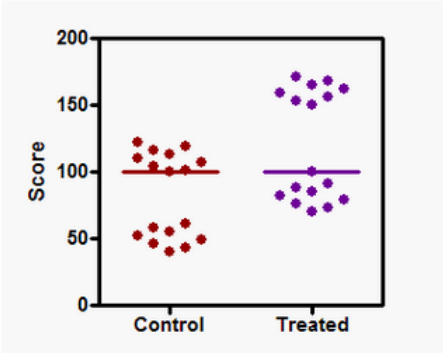
Figure: Uber's statistics engine is used for A/B/N experiments

Mann-Whitney U test

Suppose we have a sample of n_x observations $\{x_1, x_2, \dots x_n\}$ in one group (i.e. from one population) and a sample of n_y observations $\{y_1, y_2, \dots y_n\}$ in another group (i.e. from another population).

The Mann-Whitney test is based on a comparison of every observation x_i in the first sample with every observation y_j in the other sample. The total number of pairwise comparisons that can be made is $n_x n_y$.

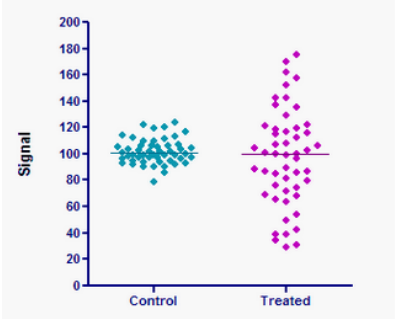
You'll sometimes read that the Mann-Whitney test compares the medians of two groups. But this is not exactly true, as this example demonstrates.



The graph shows each value obtained from control and treated subjects. The two-tail P value from the Mann-Whitney test is 0.0288, so you conclude that there is a statistically significant difference between the groups. But the two medians, shown by the horizontal lines, are identical. The Mann-Whitney test ranked all the values from low to high, and then compared the mean ranks. The mean of the ranks of the control values is much lower than the mean of the ranks of the treated values, so the P value is small, even though the medians of the two groups are identical.

Mann-Whitney U test

It is also not entirely correct to say that the Mann-Whitney test asks whether the two groups come from populations with different distributions. The two groups in the graph below clearly come from different distributions, but the P value from the Mann-Whitney test is high (0.46). The standard deviation of the two groups is obviously very different. But since the Mann-Whitney test analyzes only the ranks, it does not see a substantial difference between the groups.



The Mann-Whitney test **compares the mean ranks** – it does not compare medians and does not compare distributions. More generally, the P value answers this question: What is the chance that a randomly selected value from the population with the larger mean rank is greater than a randomly selected value from the other population?

If you make an additional assumption – that the distributions of the two populations have the **same shape**, even if they are shifted (have different medians) – then the Mann-Whiteny test can be considered a test of medians. If you accept the assumption of identically shaped distributions, then a small P value from a Mann-Whitney test leads you to conclude that the difference between medians is statistically significant.

However, if the groups have the same distribution, then a shift in location will move medians and means by the same amount and so the difference in medians is the same as the difference in means. Thus the Mann-Whitney test is also a test for the difference in means.

Mann-Whitney U test

Hence, when we apply Mann-Whitney test under assumption that the two populations have the **same shape**, we can set our hypotheses as following:

$$H_0 : f_x(x) = f_y(y)$$

$$H_1 : f_x(x) = f_y(y + a)$$

We count the number of times an x_i from sample 1 is greater than a y_j from sample 2 . This number is denoted by U_x. Similarly, the number of times an x_i from sample 1 is smaller than a y_j from sample 2 is denoted by U_y. Under the null hypothesis we would expect U_x and U_y to be approximately equal.

Procedure for carrying out the test:

1. Arrange all the observations in order of magnitude.
2. Under each observation, write down X or Y (or some other relevant symbol) to indicate which sample they are from.
3. Under each x write down the number of y s which are to the left of it (i.e. smaller than it); this indicates $x_i > y_j$. Under each y write down the number of x s which are to the left of it (i.e. smaller than it); this indicates $y_j > x_i$
4. Add up the total number of times $x_i > y_j$ - denote by U_x. Add up the total number of times $y_j > x_i$ - denote by U_y. Check that $U_x + U_y = n_x n_y$.
5. Calculate $U = \min(U_x, U_y)$
6. Use statistical tables for the Mann-Whitney U test to find the probability of observing a value of U or lower. If the test is one-sided, this is your p-value; if the test is a two-sided test, double this probabililty to obtain the p-value.

NOTE: If the number of observations is such that $n_x n_y$ is large enough (¿20), a normal approximation can be used with $\mu_U = \frac{n_x n_y}{2}, \sigma_U = \sqrt{\frac{n_x n_y (N+1)}{12}}$, where $N = n_x + n_y$.

Mann-Whitney U test

Example: The following data shows the age at diagnosis of type II diabetes in young adults. Is the age at diagnosis different for males and females?

Males: 19 22 16 29 24; Females: 20 11 17 12

Solution:

1. 1. Arrange in order of magnitude

Age	11	12	16	17	19	20	22	24	29
M/F	F	F	M	F	M	F	M	M	M
$M > F$			2		3		4	4	4
$F > M$	0	0		1		2			

2. 2. Affix M or F to each observation (see above).
3. 3. Under each M write the number of F s to the left of it; under each F write the number of Ms to the left of it (see above).
4. 4. $U_M = 2 + 3 + 4 + 4 + 4 = 17 \quad U_F = 0 + 0 + 1 + 2 = 3$
5. 5. $U = \min(U_M, U_F) = 3$
6. 6. Using tables for the Mann-Whitney U test we get a two-sided p-value of $p = 0.11$
7. 7. If we use a normal approximation we get:

$$z = \frac{U - \frac{n_x n_y}{2}}{\sqrt{\frac{n_x n_y (N+1)}{12}}} = \frac{3 - 10}{\sqrt{50/3}} = -1.715$$ This gives a two-sided p-value of $p = 0.09$.

The exact test and the normal approximation give similar results. We would conclude that there is no real evidence that the age at diagnosis is different for males and females, although the results are borderline and the lack of statistical significance in this case may just be due to the very small sample. The actual median age at diagnosis is 14.5 years for females and 22 for males, which is quite a substantial difference. In this case it would be advisable to conduct a larger study.

Welch vs Mann-Whitney

Welch T-test	Mann-Whitney U-test
+ easy to interpret	- difficult to interpret
- sensitive to outliers	+ not sensitive to outliers

It would seem prudent to use non-parametric tests in all cases, which would save one the bother of testing for Normality. Parametric tests are preferred, however, for the following reasons:

1. We are rarely interested in a significance test alone; we would like to say something about the population from which the samples came, and this is best done with estimates of parameters and confidence intervals.
2. It is difficult to do flexible modelling with non-parametric tests, for example allowing for confounding factors using multiple regression.
3. Parametric tests usually have more statistical power than their non-parametric equivalents. In other words, one is more likely to detect significant differences when they truly exist.

Welch vs Mann-Whitney

Power analysis Lab:

– Comparison of type I error rates and powers

– Fair comparison of powers

– $\sigma_x < \sigma_y$ and $N_x > N_y$ vs $\sigma_x > \sigma_y$ and $N_x > N_y$

– Test for equality of variances + Welch/t-test

Bootstrap

Principle of obtaining **the sampling distribution**

1. Draw samples from the *population*
2. Compute the statistic of interest for each sample (such as the mean, median, etc.)
3. The distribution of the statistics is the *sampling distribution*

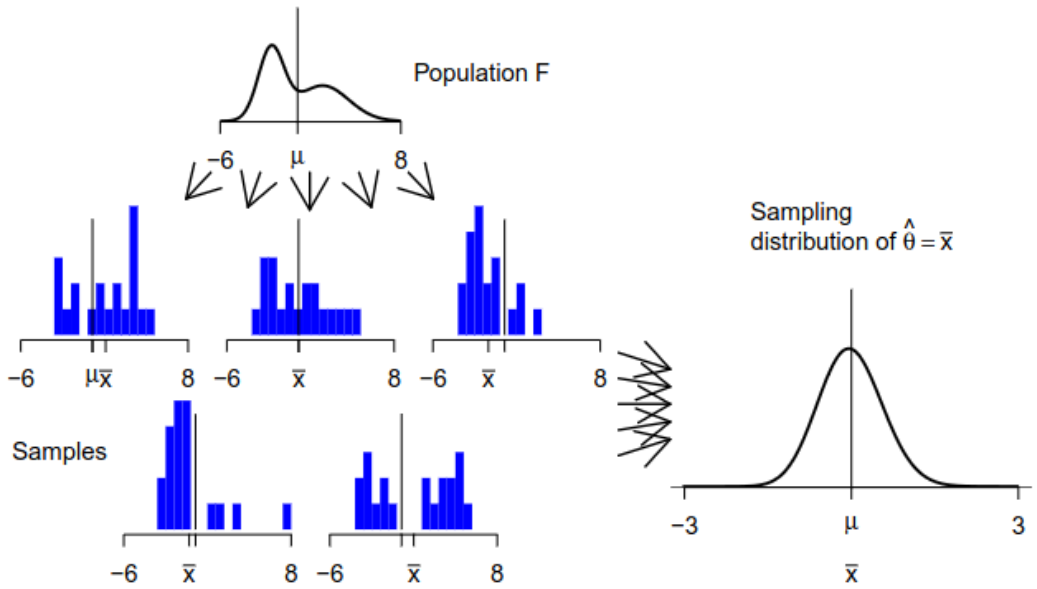


Figure: Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution.

Bootstrap principle

1. Draw samples from an estimate of the population.
2. Compute the statistic of interest for each sample.
3. The distribution of the statistics is the bootstrap distribution.

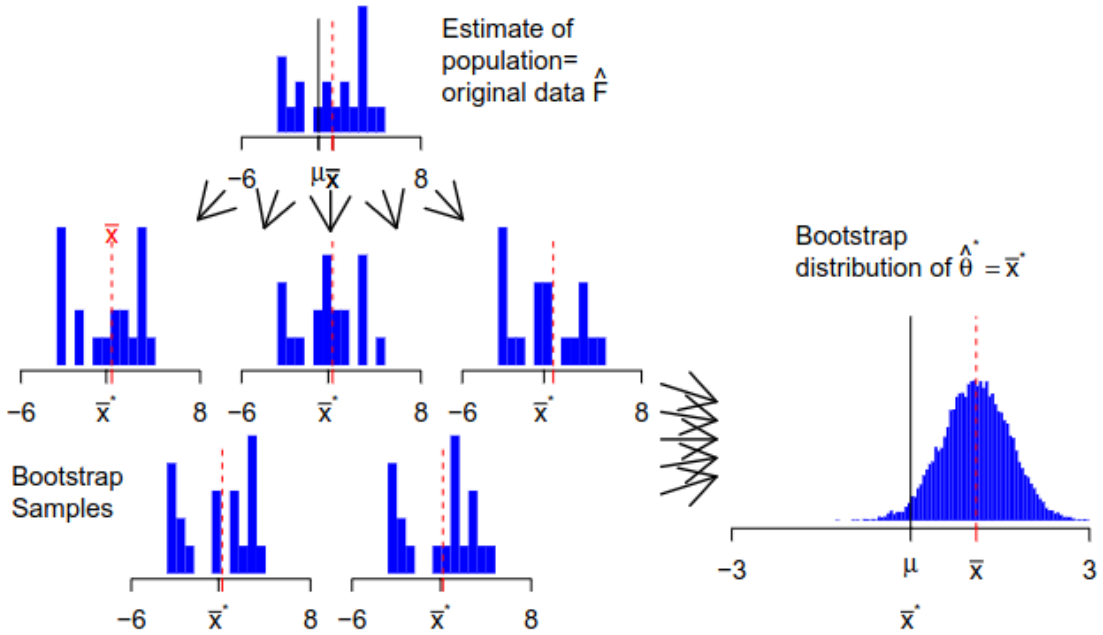


Figure: The bootstrap distribution is obtained by drawing repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics. The distribution is centered at the observed statistic ($\bar{x}$), not the parameter (μ)

Bootstrap

The set-up

1. $x_1, x_2, \ldots, x_n$ is a data sample drawn from a distribution F.
2. u is a statistic computed from the sample.
3. F^* is the empirical distribution of the data (the resampling distribution).
4. $x_1^*, x_2^*, \ldots, x_n^*$ is a resample of the data of the same size as the original sample
5. u^* is the statistic computed from the resample.

Bootstrap principles:

- $F^* \approx F$ and the variation of u is well-approximated by the variation of u^*.
- Sampling with replacement from the data is used: resampling with replacement maintains data structure but reshuffles values, extrapolating to the population

Useful when:

1. data are not normal
2. have unknown statistical properties (e.g., PCA results)
3. lack a standard calculation (e.g., R^2 or coefficient of variation)

Bootstrap

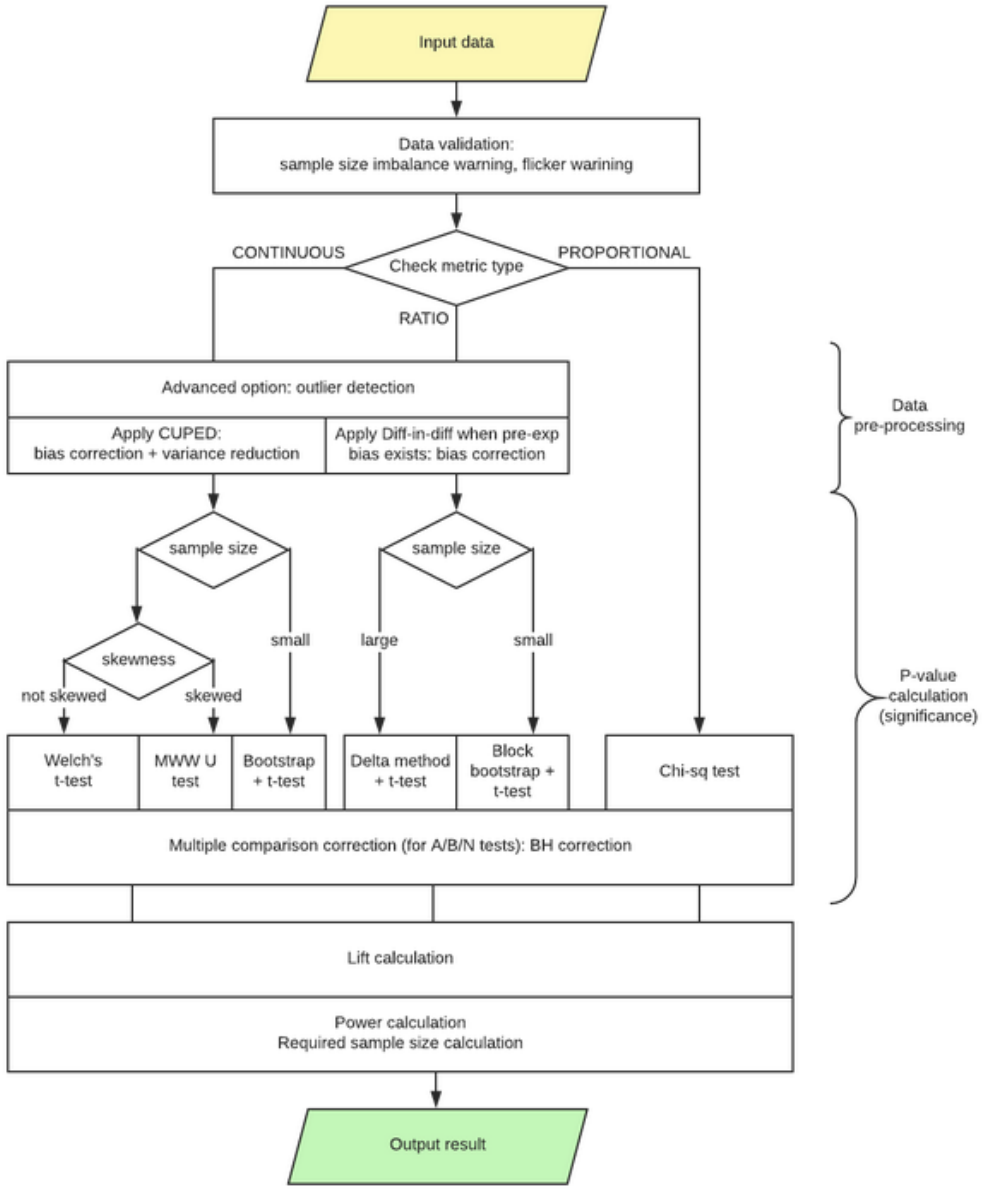


Figure: Uber's statistics engine is used for A/B/N experiments

Causes of Bias

There are three common causes of bias.

- Bias caused by **nonlinear transformations** for complex statistics – Bias correction would be harmful

Example: estimating the relative risk; $E(\hat{p}_1/\hat{p}_2) = E(\hat{p}_1)\,E(1/\hat{p}_2) \neq E(\hat{p}_1)/E(\hat{p}_2)$.

In this case the median bias is near zero, but the mean bias estimate $\hat{\theta}^* - \hat{\theta}$ can be large and have high variability, and is strongly dependent on how close the denominator is to zero.

Example: s^2 is unbiased but s is not; $E(s) \neq \sqrt{E(s^2)} = \sigma$.

- Inherently biased parameters + Bias due by **optimization**, when one or more parameters are chosen to optimize some measure, then the estimate of that measure is biased. – Bias correction can be helpful.

Example: R^2, sample variance.

- Bias due **lack of model fit** – Bias correction would not be apparent to the bootstrap

Here the bootstrap may not even show that there is bias. It can only quantify the performance of the procedure you actually used, not what you should have used.

(!) Inference, Not Better Estimates

The bootstrap distribution is centered at the observed statistic, not the population parameter, e.g. at $\bar{x}$, not μ.

- We do not use the bootstrap to get better estimates. For example, we cannot use the bootstrap to improve on $\bar{x}$; no matter how many bootstrap samples we take, they are always centered at $\bar{x}$, not μ. We'd just be adding random noise to $\bar{x}$. Instead we use the bootstrap to tell how accurate the original estimate is $=>$ A different approach to estimating the standard error
- We do not use quantiles of the bootstrap distribution of $\hat{\theta}^*$ to estimate quantiles of the sampling distribution of $\hat{\theta}$. Instead, we use the bootstrap distribution to estimate the standard deviation of the sampling distribution, or the expected value of $\hat{\theta} - \theta$.

Instead we use the bootstrap to tell how accurate the original estimate is

– Can be used for complex statistics

– Non-parametric bootstrap does not help for small samples

Bootstrap

The bootstrap distribution reflects the original sample.

Typically for large samples the data represent the population well; for small samples they may not. **Bootstrapping does not overcome the weakness of small samples as a basis for inference.**

Indeed, for the very smallest samples, you may not want to bootstrap; it may be better to make **additional assumptions** such as smoothness or a parametric family.

When there is a lot of data (sampled randomly from a population) we can trust the data to represent the shape and spread of the population; when there is little data we cannot.

Classification of sampling techniques

Features of **non-probability sampling** include the following.

– Some units in the population have a zero probability of selection.

– Individual units in populations have an unknown probability of being selected.

– Inability to measure sampling error.

Features of **probability sampling** include the following.

– Every population element has a known, non-zero probability of being selected in the sample.

– Probability sampling makes it possible to estimate the margins of sampling error.

Classification of sampling techniques

Technique	*Strengths*	*Weaknesses*
Non-probability sampling		
Convenience sampling	Least expensive, least time-consuming, most convenient	Selection bias, sample not representative, not recommended for descriptive or causal research
Judgemental sampling	Low cost, convenient, not time-consuming ideal for exploratory research designs	Does not allow generalisation, subjective
Quota sampling	Sample can be controlled for certain characteristics	Selection bias, no assurance of representativeness
Snowball sampling	Can estimate rare characteristics	Time-consuming
Probability sampling		
Simple random sampling (SRS)	Easily understood, results projectable	Difficult to construct sampling frame, expensive, lower precision, no assurance of representativeness
Systematic sampling	Can increase representativeness, easier to implement than SRS, sampling frame not always necessary	Can decrease representativeness depending upon 'order' in the sampling frame
Stratified sampling	Includes all important subpopulations, precision	Difficult to select relevant stratification variables, not feasible to stratify on many variables, expensive
Cluster sampling	Easy to implement, cost effective	Imprecise, difficult to compute and interpret results

Stratified sampling vs Cluster sampling

Factor	**Stratified sampling**	**Cluster sampling (one-stage)**
Objective	Increase precision	Decrease cost
Subpopulations	All strata are included	A sample of clusters is chosen
Within subpopulations	East stratum should be homogeneous	Each cluster should be heterogeneous
Across subpopulations	Strata should be heterogeneous	Clusters should be homogeneous
Sampling frame	Needed for the entire population	Needed only for the selected clusters
Selection of elements	Elements selected from each stratum randomly	All elements from each selected cluster are included

Imbalanced samples

Problem:

– Data imbalance is predominant and inherent in the real world.

– Data often demonstrates skewed distributions with a long tail.

– However, most of algorithms are designed around the assumption of a uniform distribution over each target category (classification).

One way the imbalance may affect the algorithm is when algorithm completely ignores the **minority class**.

The reason this is an issue is because the minority class is often the class that we are most interested in.

E.g.

1) a classifier to classify fraudulent and non-fraudulent transactions from various observations

2) probability of bankruptcies withing the industry

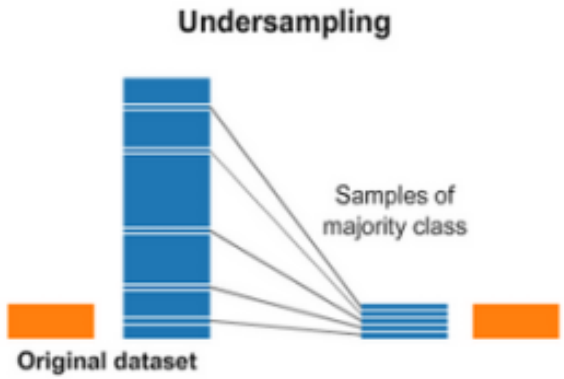
Imbalanced samples

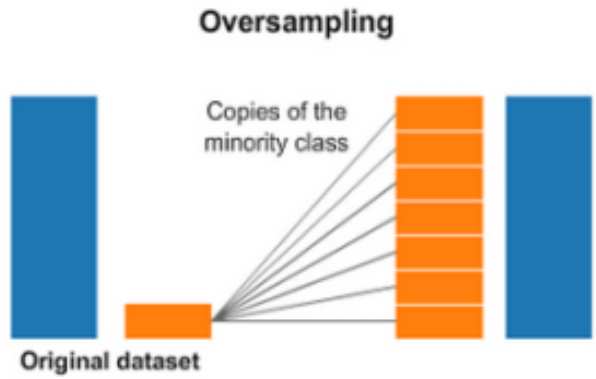
There are two main approaches to random resampling for imbalanced classification; they are oversampling and undersampling.

– Random Oversampling: Randomly duplicate examples in the minority class.

– Random Undersampling: Randomly delete examples in the majority class.

Random oversampling involves randomly selecting examples from the minority class, with replacement, and adding them to the training dataset. Random undersampling involves randomly selecting examples from the majority class and deleting them from the training dataset.





Undersampling

Advantages

– Reduce the risk of their analysis or machine learning algorithm skewing toward the majority.

Without resampling, scientists might come up with *the accuracy paradox* where they run a classification model with 90% accuracy. On closer inspection, though, they will find the results are heavily within the majority class.

– Fewer storage requirements and better run times for analyses. Less data means you or your business needs less storage and time to gain valuable insights.

Disadvantages

– Removing enough majority examples to make the majority class the same or similar size to the minority class results in a significant loss of data.

– The sample of the majority class chosen could be biased, meaning, it might not accurately represent the real world, and the result of the analysis may be inaccurate. Therefore, it can cause the classifier to perform poorly on real unseen data.

Oversampling

Advantages

– It improves the overfitting caused by random oversampling as synthetic examples are generated rather than a copy of existing examples.

–No loss of information.

Disadvantages

– While generating synthetic examples, SMOTE does not take into consideration neighboring examples that can be from other classes. This can increase the overlapping of classes and can introduce additional noise.

– SMOTE is not very practical for high-dimensional data.

Estimating treatment effect

The problem of counterfactual

The main challenge of an impact evaluation is to determine what would have happened to the beneficiaries if the program had not existed.

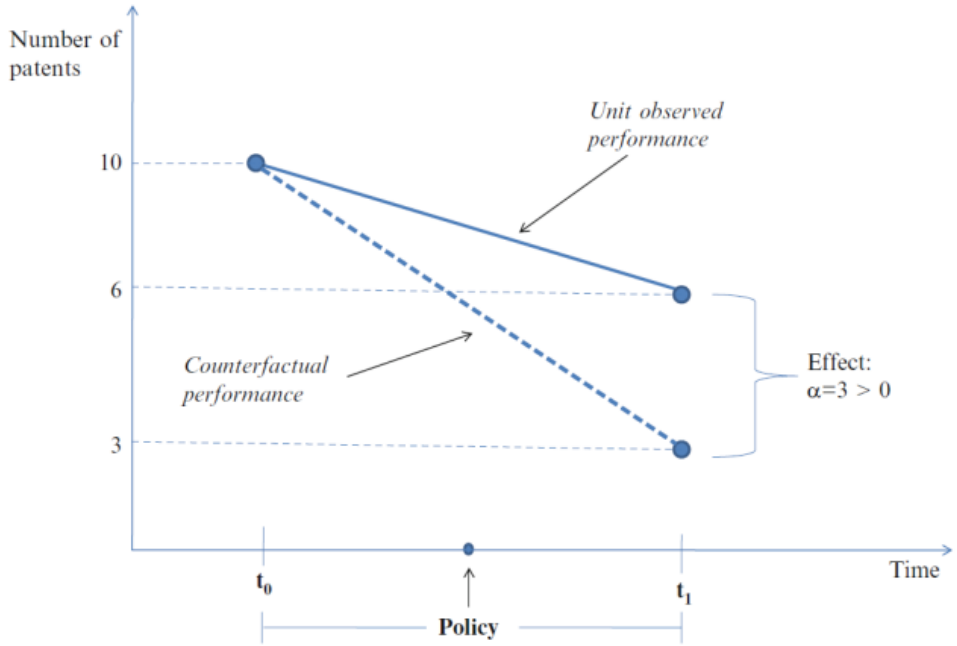


Figure:

Without information on the counterfactual, the next best alternative is to compare outcomes of treated individuals with those of a comparison group that has not been treated.

Estimating treatment effect

The level of health that we actually observe is the value of the dependent variable available to us in the data for a particular individual (observed outcome):

$$Y_i = \begin{cases} Y_i(1), \text{ if } D_i = 1; \\ Y_i(0), \text{ if } D_i = 0. \end{cases}$$

Sometimes it is convenient to write Y_i as follows:

$$Y_i = Y_i(0) + D_i \cdot (Y_i(1) - Y_i(0)).$$

Since we cannot directly calculate the impact effect for an individual object $Y_i(1) - Y_i(0)$, we cannot calculate its mathematical expectation $E(Y_i(1) - Y_i(0))$, i.e. ATE, as is.

Instead, we can try to estimate this effect using observed data.

Let's consider comparing the expected health levels of those who were hospitalized with the expected health levels of everyone else:

$$E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0)$$

where $E(Y_i \mid D_i = 1)$ is the expected value of the dependent variable for objects that were exposed; $E(Y_i \mid D_i = 0)$ is the expected value of the dependent variable for objects that were not exposed.

Estimating treatment effect

To find out how this difference in mathematical expectations relates to the value ATE of interest to us, let us carry out the following transformations:

$$E\left(Y_i \mid D_i = 1\right) - E\left(Y_i \mid D_i = 0\right) = E\left(Y_i(1) \mid D_i = 1\right) - E\left(Y_i(0) \mid D_i = 0\right) =$$
$$= E\left(Y_i(1) \mid D_i = 1\right) - E\left(Y_i(0) \mid D_i = 1\right) + E\left(Y_i(0) \mid D_i = 1\right) - E\left(Y_i(0) \mid D_i = 0\right) =$$
$$= \underbrace{E\left(Y_i(1) - Y_i(0) \mid D_i = 1\right)}_{\text{ATET}} + \underbrace{E\left(Y_i(0) \mid D_i = 1\right) - E\left(Y_i(0) \mid D_i = 0\right)}_{\text{selection bias}}.$$

The last expression consists of two terms:

- $E\left(Y_i(1) - Y_i(0) \mid D_i = 1\right)$ is the average treatment effect on the treated, ATET

- $E\left(Y_i(0) \mid D_i = 1\right) - E\left(Y_i(0) \mid D_i = 0\right)$ - this expression is called selection bias

  – the expected level of health of hospitalized people $(D = 1)$ if they had not gone for treatment $(Y_i(0))$.

  – **minus** the expected level of health of people who did not go for treatment.

Thus, the initial difference in mathematical expectations can be written as follows:

$$E\left(Y_i \mid D_i = 1\right) - E\left(Y_i \mid D_i = 0\right) = ATET + \text{ selection bias}.$$

Example: people with low levels of health are more likely to enter the experiment $=>$ $Bias < 0$

Estimating treatment effect

Under conditions of random distribution into groups (**random assignment**), whether an object falls into one group or another will not depend on its characteristics.

In terms of mathematical expectations, this would mean that:

$$E\left(Y_i(0) \mid D_i = 1\right) = E\left(Y_i(0) \mid D_i = 0\right) = E\left(Y_i(0)\right).$$

In this situation, there is no self-selection bias:

$$\text{selection bias } = E\left(Y_i(0) \mid D_i = 1\right) - E\left(Y_i(0) \mid D_i = 0\right) = 0.$$

Consequently, the difference in conditional mathematical expectations is equal to the average impact effect of interest to us:

$$E\left(Y_i \mid D_i = 1\right) - E\left(Y_i \mid D_i = 0\right) = ATET.$$

Estimating treatment effect

In case of no bias, mathematical expectations can be consistently estimated by averages, a consistent estimate of the average impact effect can be calculated as follows:

$$\bar{Y}_1 - \bar{Y}_0 = \widehat{ATET}$$

where $\bar{Y}_1$ is the sample average value of Y_i for objects included in the test group;

$\bar{Y}_0$ – sample average value of the dependent variable for objects included in the control group.

Estimating treatment effect

An estimate of the treatment effect can also be obtained using ordinary pairwise regression. To do this, you need to estimate the model parameters:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 \cdot D_i,$$

where $\hat{\beta}_2 = \bar{Y}_1 - \bar{Y}_0$.

If the experiment is constructed correctly, then in ordinary pairwise regression the explanatory variable is exogenous.

Therefore, ordinary pairwise regression provides an unbiased and consistent estimate of the average treatment effect. Therefore, it is not necessary to use **control variables** in the regression.

However, there are two reasons why their use can still be useful:

1. **Increased estimation precision:** Including control variables allows us to better describe the dependent variable, reduce the standard error of the regression, and obtain more accurate coefficient estimates.

2. **Checking the quality of randomization:** if the experiment is constructed correctly and the binary variable D is truly exogenous, then the estimates of the coefficient for this variable in paired and multiple regression should not differ much (since both estimates are consistent).

Estimating treatment effect

The average treatment effect on the treated (ATET) and the average treatment effect on the untreated (ATENT), defined respectively as:

$$\begin{aligned} \text{ATET} &= \text{E}\left(Y_1 - Y_0 \mid D = 1\right) \\ \text{ATENT} &= \text{E}\left(Y_1 - Y_0 \mid D = 0\right) \end{aligned}$$

The ATET is the average treatment effect calculated within the subsample of treated units.

The ATENT is the average treatment effect calculated within the subsample of untreated units.

$$\text{ATE} = \text{ATET} \cdot p(D = 1) + \text{ATENT} \cdot p(D = 0)$$

We can also define the previous parameters as conditional on x as "individual-specific average treatment effects".

Estimating treatment effect

Regression-Adjustment

$$\text{ATE}(\mathbf{x}) = \mathrm{E}(Y \mid \mathbf{x}, D = 1) - \mathrm{E}(Y \mid \mathbf{x}, D = 0)$$

that can be interpreted as a conditional DIM estimator. By simply denoting:

$$m_1(\mathbf{x}) = \mathrm{E}(Y \mid \mathbf{x}, D = 1)$$

and

$$m_0(\mathbf{x}) = \mathrm{E}(Y \mid \mathbf{x}, D = 0)$$

we have that:

$$\text{ATE}(\mathbf{x}) = m_1(\mathbf{x}) - m_0(\mathbf{x})$$

If consistent estimators of $\mathrm{m}_1(\mathrm{x})$ and $\mathrm{m}_0(\mathrm{x})$ are available, causal parameters ATEs can be estimated through the sample equivalents:

Sample equivalents of ATEs in RA

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \left[\widehat{m}_1(\mathbf{x}_i) - \widehat{m}_0(\mathbf{x}_i)\right]$$
$$\widehat{\text{ATET}} = \frac{1}{N_1} \sum_{i=1}^{N} D_i \cdot \left[\widehat{m}_1(\mathbf{x}_i) - \widehat{m}_0(\mathbf{x}_i)\right]$$
$$\widehat{\text{ATENT}} = \frac{1}{N_0} \sum_{i=1}^{N} (1 - D_i) \cdot \left[\widehat{m}_1(\mathbf{x}_i) - \widehat{m}_0(\mathbf{x}_i)\right]$$

Both $m_1(x)$ and $m_0(x)$ can be estimated either parametrically, semi-parametrically, or non-parametrically.

Difference-in-differences method

The main factors that can influence occupancy in a typical restaurant:

- specific features of the state in which the restaurant is located (state effect);

- features of different periods of time, say, changes in economic conditions (temporary effect);

- the effect of changing the minimum wage (the same effect that we are trying to estimate).

Formally we can write it like this:

$$Y_{ist} = \alpha_s + \mu_t + \delta \cdot D_{ist} + \varepsilon_{ist},$$

where index i – restaurant number; Y_{ist} – the number of workers employed in this restaurant; the variable $D = 1$, for New Jersey after wages changed, and $D = 0$ otherwise;

α_s – state effect: α_{control} , for Pennsylvania; $\alpha_{\text{treatment}}$ for New Jersey;

μ_t – time effect: μ_{before} before the wage change and μ_{after} after the change;

δ **is the effect of a wage increase on employment**;

ε_{ist} – random errors of the model.

Difference-in-differences method

New Jersey before the wage change:

$$E\left(Y_{\text{ist}} \mid s = \text{ treatment }, t = \text{ before }\right) = \mu_{\text{before}} + \alpha_{\text{treatment}}.$$

New Jersey after the change:

$$E\left(Y_{\text{ist}} \mid s = \text{ treatment }, t = \text{ after }\right) = \mu_{\text{after}} + \alpha_{\text{treatment}} + \delta.$$

The expected change in employment in New Jersey:

$$\Delta_{\text{treatment}} = \mu_{\text{after}} - \mu_{\text{before}} + \delta$$

Pennsylvania before the wage change:

$$E\left(Y_{\text{ist}} \mid s = \text{ control}, t = \text{ before }\right) = \mu_{\text{before}} + \alpha_{\text{control}}.$$

Pennsylvania after the change:

$$E\left(Y_{\text{ist}} \mid s = \text{ control }, t = \text{ after }\right) = \mu_{\text{after}} + \alpha_{\text{control}}.$$

The expected change in employment in Pennsylvania:

$$\Delta_{\text{control}} = \mu_{\text{after}} - \mu_{\text{before}}.$$

Difference-in-differences method

The expected change in employment in New Jersey:

$$\Delta_{\text{treatment}} = \mu_{\text{after}} - \mu_{\text{before}} + \delta$$

The expected change in employment in Pennsylvania:

$$\Delta_{\text{control}} = \mu_{\text{after}} - \mu_{\text{before}} .$$

Finally,

$$\Delta_{\text{treatment}} - \Delta_{\text{control}} = \delta$$

Thus, the treatment effect can be represented as the difference between the differences of conditional mathematical expectations:

$$\delta = \Delta_{\text{treatment}} - \Delta_{\text{control}} =$$
$$= [E(Y_{\text{ist}} \mid s = \text{ treatment }, t = \text{ after }) - E(Y_{\text{ist}} \; mids = \text{ treatment }, t = \text{ before })] -$$
$$- [E(Y_{\text{ist}} \mid s = \text{ control }, t = \text{ after }) - E(Y_{\text{ist}} \; mids = \text{ control }, t = \text{ before })] .$$

And by LLN, a consistent estimate of each of these mathematical expectations is the corresponding average value. Therefore:

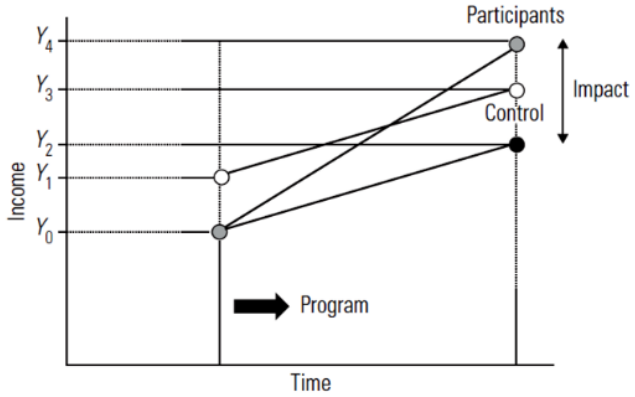
$$\hat{\delta} = \left[\bar{Y}_{\text{treatment, after}} - \bar{Y}_{\text{treatment, before}}\right] - \left[\bar{Y}_{\text{control, after}} - \bar{Y}_{\text{control, before}}\right] ,$$

Difference-in-differences method

Graphical illustration:

$$\mathrm{DD} = E(Y_1^T - Y_0^T \mid T_1 = 1) - E(Y_1^C - Y_0^C \mid T_1 = 0) \cdot \qquad (5.1)$$

Figure 5.1 An Example of DD



$$\mathrm{DD} = (Y_4 - Y_0) - (Y_3 - Y_1).$$

$$\mathrm{DD} = (Y_4 - Y_2)$$

Figure:

Difference-in-differences method

The difference-in-differences method is directly related to the estimation of models using regressions.

$$Y_{it} = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot z_t + \delta \cdot x_i \cdot z_t + \varepsilon_{it},$$

where x_i is a binary variable that equals one if the ith restaurant is located in New Jersey (i.e., belongs to the test group);

$z_t -$ is a binary variable that is equal to one for all observations related to the second period (the period after the minimum wage increase).

Difference-in-differences method

In this case, applying the least squares method, we obtain the following equation:

$$\hat{Y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i + \hat{\beta}_2 \cdot z_t + \hat{\delta} \cdot x_i \cdot z_t.$$

where FE-estimator:

$$\begin{gathered}
\hat{\beta}_0 = \bar{Y}_{\text{control, before}} ; \\
\hat{\beta}_1 = \bar{Y}_{\text{treatment,before}} - \bar{Y}_{\text{control, before}} ; \\
\hat{\beta}_2 = \bar{Y}_{\text{control, after}} - \bar{Y}_{\text{control, before}} ; \\
\hat{\delta} = \left[\bar{Y}_{\text{treatment, after}} - \bar{Y}_{\text{treatment,before}}\right] - \left[\bar{Y}_{\text{control, after}} - \bar{Y}_{\text{control, before}}\right].
\end{gathered}$$

Thus, the coefficient of the product $x_i \cdot z_t$ is equal to the same estimate of the impact effect that we derived above.

An equivalent estimation method is to apply OLS to the following pairwise regression:

$$\Delta Y_i = \beta_2 + \delta \cdot x_i + u_i,$$

where x_i is still a binary variable that is equal to one if the i-th object belongs to the test group;

hence, DD-estimator: $\Delta Y_i = Y_{i1} - Y_{i0}$.

Difference-in-differences method

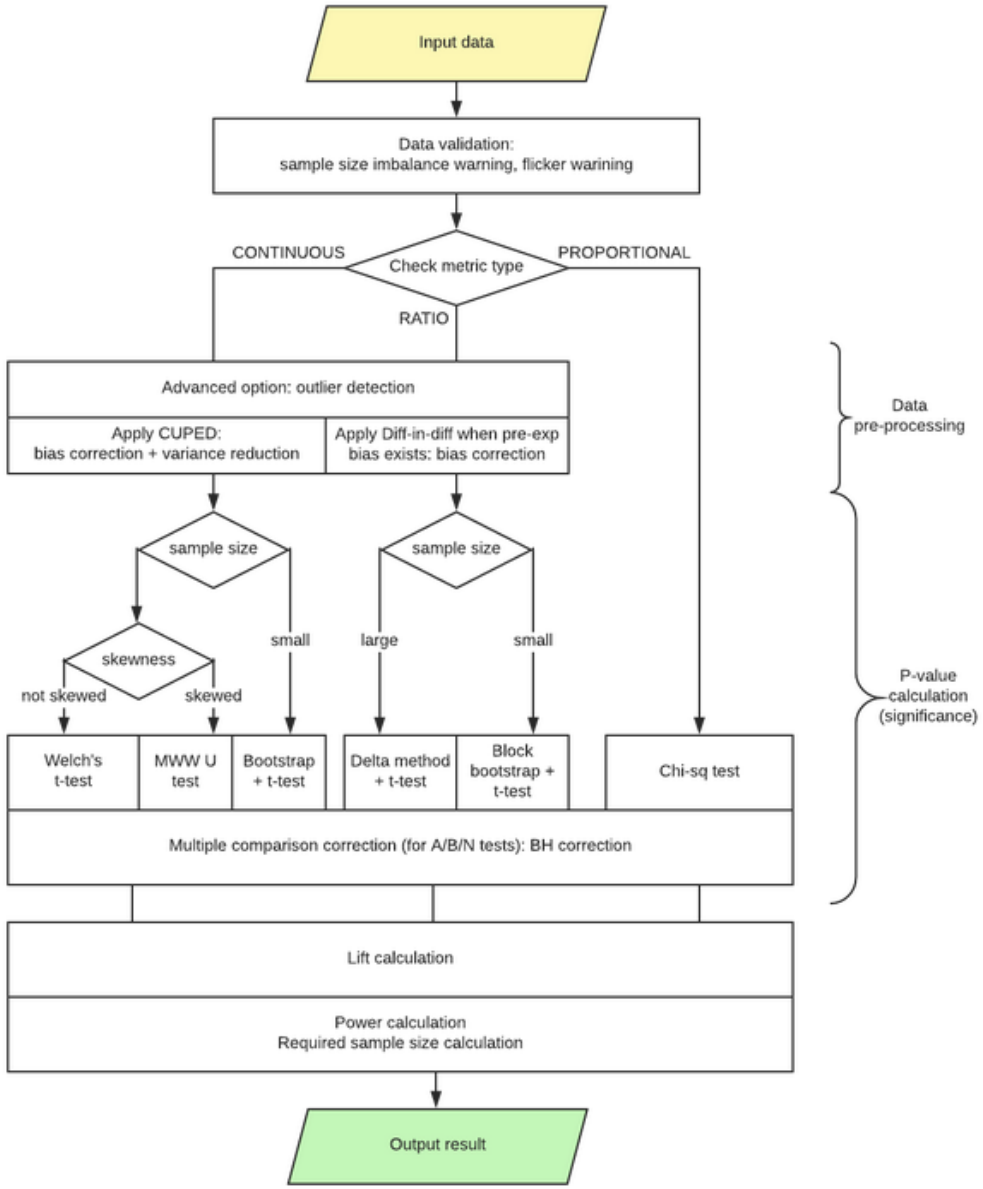


Figure: Uber's statistics engine is used for A/B/N experiments

PSM with DD (Doubly-robust estimand)

DD can be combined with PSM to better match control and project units on pre-program characteristics.

Combining different methods may sometimes lead to an estimation of the treatment effects having **better properties in terms of robustness**.

The robustness of this approach lies in the fact that either the **conditional mean** or the **propensity-score** needs to be correctly specified but not both.

The propensity score can be used to match participant and control units, and the treatment effect is calculated across participant and matched control units **within the common support**.

Matching

Causal inference "balancing technique"

Often used with "Diff-in-diff" to minimize sampling bias.

The idea is to find, from a large group of nonparticipants, individuals who are observationally similar to participants in terms of characteristics not affected by the program.

- When a treatment cannot be randomized, the next best thing to do is to try to mimic randomization—that is, try to have an observational analogue of a randomized experiment.

- Matching is not the only way to eliminate bias (e.g. regressions with control variables and/or instruments).

- Matching is a non-parametric method. You do not need to assume any functional form for the causal relationship being investigated.

Matching

1) **Simple matching** compares groups of observations corresponding to the same discrete values $\boldsymbol{x}$.

$$\Delta^M = \sum_k w_k * (\bar{y}_{1,k} - \bar{y}_{0,k})$$

$\bar{y}_{1,k}$ — average result for the treatment group

$\bar{y}_{0,k}$ — average result for control group

w_k is the proportion of observations belonging to the k-th group among the entire sample

2) Nearest-neighbor matching

The i-th observation from the experimental group is compared with the set of closest observations from the control group.

Euclidean distance is used as a measure of proximity.

$$A_i = \left\{ j \mid \min_j \left\| x_i - x_j \right\| \right\}$$

A_i — set of objects from the control group that are compared with the i-th observation and the experimental group

Matching

3) Propensity score matching

If the vector of explanatory variables has a large dimension, or if there are continuous variables among the variables, then an exact comparison is not entirely convenient.

In this case, a propensity score is used - the conditional probability that an object will be affected given the given values of the regressors.

The propensity score is usually estimated using a logit or probit model.

$$P\left(D_i = 1 \mid x_i^{(1)}, x_i^{(2)} \ldots, x_i^{(k)}\right) = \\ = p\left(x_i^{(1)}, x_i^{(2)} \ldots, x_i^{(k)}\right)$$

Thus, matching is carried out in two stages:

1. For each observation, the value of the propensity measure is estimated (for example, based on a logit model)

2. Then a comparison of objects with similar values of the propensity measure is carried out using

– Nearest neighbor matching

– Comparison with stratification

– Radial matching

Matching

Assumptions for PSM

– PSM is a useful approach when only **observed characteristics** are believed to affect program participation.

Whether this belief is actually the case **depends on the unique features** of the program itself, in terms of targeting as well as individual take-up of the program.

– Assuming **selection on observed characteristics** is **sufficiently strong** to determine program participation, baseline data on a wide range of pre-program characteristics will allow the probability of participation based on observed characteristics to be specified more **precisely**.

– Some tests can be conducted to **assess the degree of selection bias** or participation on unobserved characteristics.

Variance reduction

Main factors when estimating any measure:

1. sample size is finite
2. the estimate should be as accurate as possible.

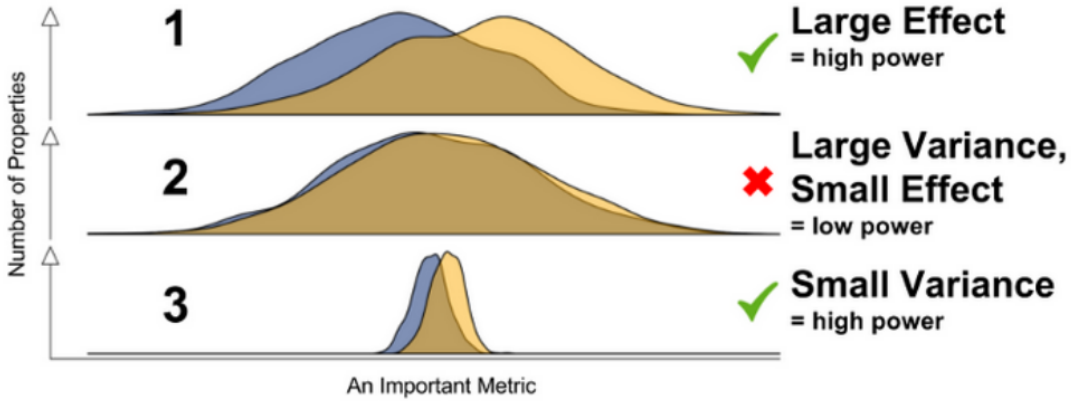


Figure:

Variance reduction: $2 => 3$

Variance reduction

In mathematics, more specifically in the theory of **Monte Carlo methods**, variance reduction is a procedure used to increase the precision of the estimates obtained for a given simulation, i.e. reducing the variance which limits the precision of the simulation results.

General idea: taking advantage of the covariance between the measures in an attempt to reduce the variance of the measure of interest.

1) Simplest example – Paired t-test

No reduction: compare the average of the individuals outcomes before and after the treatment.

Paired t-test variance reduction:

If we want the causal effect of the treatment, then we can reduce variance by **controlling for each individual's pre-treatment measure.**

The covariance between each individual's pre-treatment and post-treatment measure is utilized to reduce the variance of our estimation of the treatment effect.

CUPED (Controlled-experiment Using Pre-Experiment Data)

2) CUPED (Controlled-experiment Using Pre-Experiment Data)

Developed by the Experiment Platform team at Microsoft (Deng, Xu, Kohavi,& Walker, 2013) that tries to remove variance in a metric that can be accounted for by pre-experiment information.

– Probably the most used variance reduction technique in A/B testing in the tech industry.

– Special case of applying the technique of **control variates** to the A/B testing set-up.

CUPED (Controlled-experiment Using Pre-Experiment Data)

Assume we are in the A/B testing setting and we want to evaluate the impact of some treatment on a response metric.

For individual i, let:

– $Y_i(T)$ denote the value of the metric we would see if the individual was given the treatment,

– $Y_i(C)$ denote the value of the metric we would see if the individual was not given the treatment (i.e. was in control),

– Y_i denote the observed value (i.e. $Y_i = Y_i(T)$ or $Y_i = Y_i(C)$, depending on whether i was in treatment or control).

We want to estimate the average treatment effect (ATE) across individuals, $\Delta = \mathbb{E}[Y_i(T) - Y_i(C)]$.

The most commonly used estimator for this is the **difference-in-means estimator**

$$\hat{\Delta} = \left(\frac{\sum_{i \text{ in treatment}} Y_i}{\#\{i \text{ in treatment }\}} \right) - \left(\frac{\sum_{i \text{ in control}} Y_i}{\#\{i \text{ in control }\}} \right) =: \bar{Y}_T - \bar{Y}_C.$$

The difference-in-means estimator is unbiased for the ATE and has a certain variance.

CUPED is another estimator for the ATE that is (approximately) unbiased and usually has smaller variance than the difference-in-means estimator.

CUPED (Controlled-experiment Using Pre-Experiment Data)

Key idea

Imagine that on top of collecting metric values $Y_1, Y_2, \ldots, Y_{nt}$ in the treatment group, we also collected pre-experiment values on another (real-valued) variable $X_1, X_2, \ldots, X_{nt}$.

Let's also assume that we know the mean of X (which denote by $\mathbb{E}[X]$).

For any fixed parameter θ, we have

$$\begin{aligned}\mathbb{E}\left[Y_i(T)\right] &= \mathbb{E}\left[\bar{Y}_T\right] \\ &= \mathbb{E}\left[\bar{Y}_T - \theta X\right] + \theta\mathbb{E}[X] \\ &= \mathbb{E}\left[\bar{Y}_T - \theta\bar{X}_T\right] + \theta\mathbb{E}[X]\end{aligned}$$

Hence,

$$\widetilde{Y}_T = \bar{Y}_T - \theta\bar{X}_T + \theta\mathbb{E}[X]$$

is an unbiased estimator for $\mathbb{E}\left[Y_i(T)\right]$.

Additionally, the variance of $\widetilde{Y}_T$ is minimized when $\theta = \mathrm{Cov}(Y, X)/\mathrm{Var}(X)$, and at this value of θ, we have

$$\mathrm{Var}\left(\tilde{Y}_T\right) = \left(1 - \rho^2\right)\mathrm{Var}\left(\bar{Y}_T\right) \leq \mathrm{Var}\left(\bar{Y}_T\right)$$

where ρ is the correlation between Y and X.

CUPED (Controlled-experiment Using Pre-Experiment Data)

1. Take a random variable X independent of Y
2. Let's imagine the new metric as the difference between Y and θX.

$$Y_{CUPED} = Y - \theta X$$

3. Variance is calculated using the formula:

$$\left(\text{var}(Y) + \theta^2 \,\text{var}(X) - 2\theta \,\text{cov}(Y, X)\right)$$

4. Variance is minimized when:

$$\theta = \text{cov}(Y, X) / \,\text{var}(X)$$

Final variance

$$\text{var}_{srs}\left(Y_{CUPED}\right)_{\text{min}} = \text{var}_{srs}(Y)\left(1 - \rho^2\right)$$

Stratification

3) Stratification

Idea: (1) divide the sampling region into strata, (2) sample within each stratum separately and (3) combine results from individual strata together to give an overall estimate.

Mathematically, we want to estimate $\mathbb{E}(Y)$, where Y is the variable of interest.

Assume we can divide the sampling region of Y into K subregions (strata) with w_k the probability that Y falls into the k th stratum, $k = 1, \ldots, K$.

If we fix the number of points sampled from the k th stratum to be $n_k = n \cdot w_k$, we can define a stratified average to be

$$\widehat{Y}_{\text{strat}} = \sum_{k=1}^{K} p_k \bar{Y}_k,$$

where $\bar{Y}_k$ is the average within the k th stratum.

The stratified average $\widehat{Y}_{\text{strat}}$ and the standard average $\bar{Y}$ have the same expected value but the former gives a smaller variance when the means are different across the strata.

Stratification

The **intuition** is that the variance of $\bar{Y}$ can be decomposed into the within-strata variance and the between-strata variance, and the latter is removed through stratification.

Example: the variance of children's heights in general is large. However, if we stratify them by their age, we can get a much smaller variance within each age group.

More formally,

$$\begin{aligned}\operatorname{var}(\bar{Y}) &= \sum_{k=1}^{K} \frac{p_k}{n}\sigma_k^2 + \sum_{k=1}^{K} \frac{p_k}{n}\left(\mu_k - \mu\right)^2 \\ &\geq \sum_{k=1}^{K} \frac{p_k}{n}\sigma_k^2 = \operatorname{var}\left(\widehat{Y}_{\text{strat}}\right)\end{aligned}$$

where $\left(\mu_k, \sigma_k^2\right)$ denote the mean and variance for users in the k th stratum.

A good stratification is the one that aligns well with the underlying clusters in the data. By explicitly identifying these clusters as strata, we essentially remove the extra variance introduced by them.

Summary of comparison

– Both estimates are unbiased.

– The variance of the estimate in stratified sampling is smaller than that in simple random sampling

$$\begin{aligned}\operatorname{var}(\bar{Y}_{srs}) &= \sum_{k=1}^{K} \frac{p_k}{n}\sigma_k^2 + \sum_{k=1}^{K} \frac{p_k}{n}(\mu_k - \mu)^2 \\ &\geq \sum_{k=1}^{K} \frac{p_k}{n}\sigma_k^2 = \operatorname{var}\left(\widehat{Y}_{strat}\right)\end{aligned}$$

The intuition:

Variance of SRS estimate can be decomposed into within-strata variance and between-strata variance.

Stratified sampling achieves variance reduction by removing the between-strata variance.

Post-stratification

Post stratification assumes simple random sampling but uses the estimate in

$$\hat{Y}_{\text{strat}} = \sum_{k=1}^{K} p_k \bar{Y}_k$$

$$\bar{Y} = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}.$$

Note that, when simple random sampling is used, these estimates are different.

This is because the sample size n_k from the k th stratum is not necessarily equal to np_k under simple random sampling.

In fact, $n_1, \ldots, n_K$ are all random under simple random sampling.

The intuition behind post stratification is very simple.

The weighted average gives more weights to observations from the strata that are under-represented in the sample.

Thus if a sample is *badly balanced* for some covariate such as signup country, the weighted average estimate automatically corrects for it.

Post-stratification

For large enough n, post stratification leads to variance reduction for large enough sample size:

$$\begin{aligned}
&\text{var}_{\text{strat}}\left(\hat{Y}_{\text{strat}}\right) = \text{var}_{\text{srs}}\left(\hat{Y}_{\text{strat}}\right) + O\left(\frac{1}{n^2}\right) = \text{var}_{\text{srs}}\left(\bar{Y}\right) + O\left(\frac{1}{n}\right), \\
&\text{var}_{\text{strat}}\left(\hat{Y}_{\text{strat}}\right) \leq \text{var}_{\text{srs}}\left(\hat{Y}_{\text{strat}}\right) \leq \text{var}_{\text{srs}}\left(\bar{Y}\right).
\end{aligned}$$

Post-stratification

SAMPLE SIZES REQUIRED FOR EACH METHOD

	Sample size to estimate a proportion	Sample size to estimate an average
Simple random sampling	$\frac{Z^2 p(1-p)}{e^2}$	$\frac{Z^2 \sigma^2}{e^2}$
Proportional stratified sampling	$\frac{Z^2 \sum_{h=1}^{L} W_h p_h (1-p_h)}{e^2}$	$\frac{Z^2 \sum_{h=1}^{L} W_h \sigma_h^2}{e^2}$
Best stratified sampling	$\frac{Z^2 \left(\sum_{h=1}^{L} W_h \sqrt{p_h (1-p_h)}\right)^2}{e^2}$	$\frac{Z^2 \left(\sum_{h=1}^{L} W_h \sigma_h\right)^2}{e^2}$

where Z – quantiles of the Gaussian distribution, L is the number of strata, e is the accepted margin of error; σ^2 is the variance of the data within the total population. σ_h^2 is the variance within every stratum.

- p is the proportion of the total population that we are trying to determine (e.g. the percent of the Mexican population that smokes). p_h represents that proportion within each stratum.

- W_h is the stratum's weight within the sample (the size of the stratum with respect to the whole sample).

Proportional – W_h is equal to the proportion represented by that stratum in the population. *Optimal* – W_h is calculated based on the dispersion within each stratum.

CUPED vs Stratification

When the covariates are categorical, stratification and control variates produce identical results.

For clarity and simplicity, we assume X is binary with values 1 and 0 . Let $w = \mathbb{E}(X)$. The two estimates are

$$\widehat{Y}_{\text{strat}} = w\bar{Y}_1 + (1-w)\bar{Y}_0$$
$$\widehat{Y}_{cv} = \bar{Y} - \hat{\theta}\bar{X} + \hat{\theta}w$$

where $\bar{Y}_1$ denotes the average of Y in the $\{X = 1\}$ stratum and $\hat{\theta} = \widehat{\text{cov}}(Y, X)/\widehat{\text{var}}(X) = \bar{Y}_1 - \bar{Y}_0$. Plugging in the expression for $\hat{\theta}$, we have

$$\begin{aligned}\widehat{Y}_{cv} &= \bar{Y} - \left(\bar{Y}_1 - \bar{Y}_0\right)\bar{X} + \left(\bar{Y}_1 - \bar{Y}_0\right)w \\ &= (1-\bar{X})\bar{Y}_0 + \bar{Y}_0\bar{X} + \left(\bar{Y}_1 - \bar{Y}_0\right)w \\ &= w\bar{Y}_1 + (1-w)\bar{Y}_0 = \widehat{Y}_{\text{strat}}\end{aligned}$$

where the second equality follows from the fact that $\bar{Y} = \overline{XY}_1 + (1-\bar{X})\bar{Y}_0$.

To prove for the case with $K > 2$, we construct $K-1$ indicator variables as control variates. With the observation that the coefficients $\hat{\theta}_k = \bar{Y}_k - \bar{Y}_0$, the proof follows the same steps as the binary case outlined above.

CUPED and CTR

Click-through rate (CTR) is the ratio of clicks on a specific link to the number of times a page, email, or advertisement is shown.

To achieve variance reduction for non-user level metrics, we need to incorporate **delta method**.

We use CTR as an example and derive for the control variates formulation since it's more general.

Let n be the number of users (non-random). Denote $Y_{i,j}$ the number of clicks on user i 's j th page-view during the experiment and $X_{i,k}$ the number of clicks on user i 's k th page-view during the pre-experiment period. Let N_i and M_i be the numbers of page-views from user i during the experiment and pre-experiment respectively. The estimate for CTR using $X_{i,j}$ as the control variate becomes

$$\begin{aligned}\widehat{Y}_{cv} &= \frac{\sum_{i,j} Y_{i,j}}{\sum_{i,j} 1} - \theta \frac{\sum_{i,k} X_{i,k}}{\sum_{i,k} 1} + \theta \mathbb{E}\left(X_{i,k}\right) \\ &= \frac{\sum_i Y_{i,+}}{\sum_i N_i} - \theta \frac{\sum_i X_{i,+}}{\sum_i M_i} + \theta \mathbb{E}\left(X_{i,j}\right),\end{aligned}$$

where $Y_{i,+} = \sum_j Y_{i,j}$ is the total number of clicks from user i. Similar notation applies to $X_{i,+}$.

CUPED and CTR

Following the same derivation as for count metrics, we know var $\left(\widehat{Y}_{cv}\right)$ is minimized at

$$\begin{aligned}\theta &= \text{cov}\left(\frac{\sum_i Y_{i,+}}{\sum_i N_i}, \frac{\sum_i X_{i,+}}{\sum_i M_i}\right) / \text{var}\left(\frac{\sum_i X_{i,+}}{\sum_i M_i}\right) \\ &\doteq \text{cov}\left(\frac{\bar{Y}}{\mu_N} - \frac{\mu_Y \bar{N}}{\mu_N^2} - \frac{\mu_Y}{\mu_N}, \frac{\bar{X}}{\mu_M} - \frac{\mu_X \bar{M}}{\mu_M^2} - \frac{\mu_X}{\mu_M}\right) / \text{var}\left(\frac{\bar{X}}{\mu_M} - \frac{\mu_X \bar{M}}{\mu_M^2} - \frac{\mu_X}{\mu_M}\right) \\ &= \text{cov}\left(\frac{\bar{Y}}{\mu_N} - \frac{\mu_Y \bar{N}}{\mu_N^2}, \frac{\bar{X}}{\mu_M} - \frac{\mu_X \bar{M}}{\mu_M^2}\right) / \text{var}\left(\frac{\bar{X}}{\mu_M} - \frac{\mu_X \bar{M}}{\mu_M^2}\right)\end{aligned}$$

where the second equality follows from using Taylor expansion to linearize the ratios and $\bar{Y} = \frac{1}{n}\sum_i Y_{i,+}$ with $\mu_Y = \mathbb{E}(\bar{Y})$ (similarly for μ_X, μ_N and μ_M).

Because the user is the randomization unit and user level observations are i.i.d., we have

$$\sqrt{n}(\bar{Y}, \bar{N}, \bar{X}, \bar{M}) \Rightarrow N(\mu, \Sigma),$$

following a multivariate normal distribution with mean vector μ and covariance matrix Σ easily estimated from the i.i.d. samples. It is now straight forward to estimate θ using

$$\theta = \left(\beta_1^T \Sigma \beta_2\right) / \left(\beta_2^T \Sigma \beta_2\right),$$

where $\beta_1 = \left(1/\mu_N, -\mu_Y/\mu_N^2, 0, 0\right)^T$ and $\beta_2 = \left(0, 0, 1/\mu_M, -\mu_X/\mu_M^2\right)^T$.

We can easily see that the derivation works generally for various combinations. The metric can be at user level while the covariate can be at page-view level, etc.

CUPED and CTR

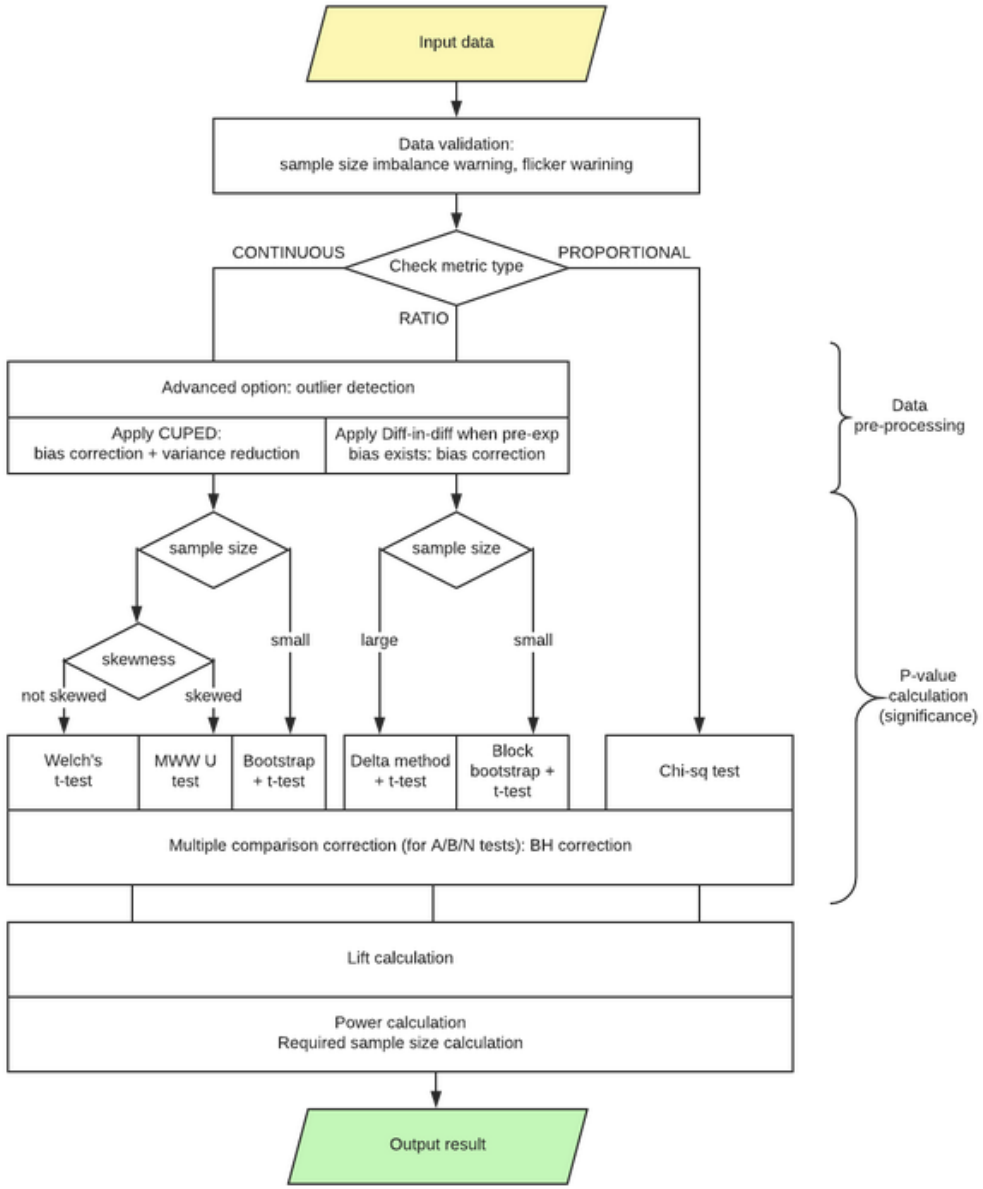


Figure: Uber's statistics engine is used for A/B/N experiments