Developing Effect Sizes for Non-Normal Data in Two-Sample Comparison Studies

with an Application in E-commerce

Amin Jamalzadeh

Durham University

Apr 13, 2010

Outline

1. Introduction
   - Effect Size, Complementory for Hypothesis Tests
   - Cohen's D Effect Size
   - Common Language Effect Size (CLES)
   - Non-Overlap Effect Sizes
   - Non-parametric Effect Size

2. Effect Size for Non-Normal Data
   - Quantile Absolute Deviation
   - Divergence Effect Size
   - Simulation Study: Normal Distribution
   - Simulation Study: Weibull Distribution

3. Experimental Study: Clickstream Analysis
   - Data Source and Preparation
   - Model-Based Effect Size Computation
   - Bootstrapping

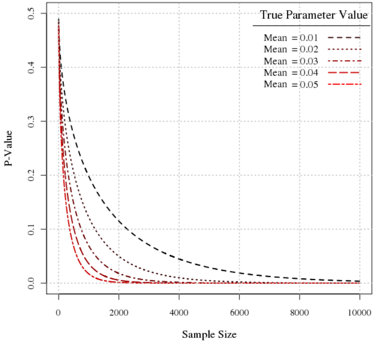
Outline

1. Introduction
   - Effect Size, Complementory for Hypothesis Tests
   - Cohen's D Effect Size
   - Common Language Effect Size (CLES)
   - Non-Overlap Effect Sizes
   - Non-parametric Effect Size
2. Effect Size for Non-Normal Data
   - Quantile Absolute Deviation
   - Divergence Effect Size
   - Simulation Study: Normal Distribution
   - Simulation Study: Weibull Distribution
3. Experimental Study: Clickstream Analysis
   - Data Source and Preparation
   - Model-Based Effect Size Computation
   - Bootstrapping

Hypothesis Test deficities

What A Statistical Hypothesis Test Provides:

The classical hypothesis testing has been widely used as a standard way of using experimental data to prove whether a phenomenon exists. BUT It does not necessarily provide information about the magnitude of the phenomenon.



- The true distribution is $N(\mu_i, 1)$
- We test $H_o : \mu_0 = 0$ versus $H_o : \mu_0 = \mu_i$.
- The true parameter is slightly bigger than 0
- For $n > 4000$ almost all tests significantly reject the null hypothesis

Effect Size

P-Value does not give information about the magnitude of an effect

We need a complementary to statistical significance to show the magintude of phenomenon

What is Effect Size?

A measures to quantify the degree to which a phenomenon exists.

Effect Size and Hypothesis tests

- Any statitical test depends on four quantities:

  P-Value – Sample size – Effect size – Power of the test.

- Test statistics are usually a function of sample size, so can not serve as ES. For example

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma/\sqrt{n}}$$

- Each hypothesis test needs a relevant ES.

Effect Size for Two sample tests

COHEN'S D EFFECT SIZE

Cohen (1977) defined the difference between the means of two populations divided by their standard deviation as a measure to represent the magnitude of mean difference between them, when two populations have the same standard deviation. This measure is one of the most widely used measures of ES for comparing the mean in two independent samples.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sigma}$$

In practice, two populations do not necessarily have an equal standard deviation. Hodges (1985) proposed to use the pooled standard deviation for computing the effect size for the non-equal standard deviation conditions.

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Effect Size for Two sample tests

COMMON LANGUAGE EFFECT SIZE (CLES)

McGraw and Wong (1992) defined Common Language Effect Size (CLES) statistic as a probability that a randomly selected individual from one group have a higher score on a variable than a randomly selected individual from another group. In other words, if X and Y follows the normal distribution with mean parameters of μ_x and μ_y respectively, and the same standard deviation σ, the CLES is:

$$\begin{aligned} \text{CLES} &= P(X < Y) \\ &= P(X - Y < 0) \\ &= \Phi\left(\frac{\mu_y - \mu_x}{\sqrt{2}\sigma}\right) \end{aligned}$$

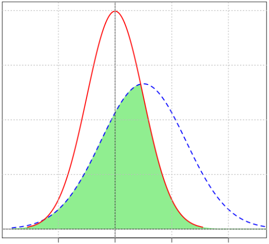
- Its easy to interpete it as it is a probability measure
- Assumes scores in both group follows the normal distribution with equally variability
- CLES$=$ 0.5 implies that two distribution entirely overlap (they are the same)
- Approching to the values of 0 and 1 implies having larger effect size

Effect Size for Two sample tests

NON-OVERLAP EFFECT SIZE

The amount of combined area under the density probability distribution function not shared by two population can serve as a measure of difference between two population (Cohen, 1992).

- It is easy to interpete it as it is a probability measure
- It does not need the Normality assumion
- It takes values over $[0, 1]$ for standardized case
- Approching to the value of 0 implies small effect size
- Approching to the value of 1 implies large effect size



Effect Size for Two sample tests

NON-PARAMETRIC CLIFF'S EFFECT SIZE

Cliff (1993) introduced a δ statistic is computed by enumerating the number of occurrences of an observation from one group having a higher response value than an observation from the second group, and the number of occurrences of the reverse.

$$\delta = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \operatorname{sign}(x_{i1} - x_{j2})}{n_1 \times n_2}$$

- It does not need the Normality assumion
- It serves better in the ordinal level of measurement
- It takes values over $[0, 1]$ for standardized case
- Approching to the value of 0 implies small effect size
- Approching to the value of 1 implies large effect size

Outline

1. Introduction
   - Effect Size, Complementory for Hypothesis Tests
   - Cohen's D Effect Size
   - Common Language Effect Size (CLES)
   - Non-Overlap Effect Sizes
   - Non-parametric Effect Size
2. Effect Size for Non-Normal Data
   - Quantile Absolute Deviation
   - Divergence Effect Size
   - Simulation Study: Normal Distribution
   - Simulation Study: Weibull Distribution
3. Experimental Study: Clickstream Analysis
   - Data Source and Preparation
   - Model-Based Effect Size Computation
   - Bootstrapping

Effect Size for Non-Normal Data

Quantile Absolute Deviation: QAD(F, G)

It aims to compare quantiles of the two distribution based on the the entire range of probabilities over $[0, 1]$. The quantile deviation of two populations is the average absolute distance between the quantiles of two populations. Suppose F^{-1} and G^{-1} are quantile functions for the two statistical populations corresponding to the cumulative distribution functions, F and G respectively.

$$QAD = \int_0^1 |F^{-1}(p) - G^{-1}(p)| \, dp$$

It also satisfies three properties which are referred to as the divergence properties of a criterion:

1. Self Similarity: $QAD(F, F) = 0$
2. Self Identification: $QAD(F, G) = 0$ if and only if $F = G$
3. Positivity: $QAD(F, G) \geq 0 \quad \text{for all} \quad F, G.$

The QAD is a symmetry measure, that is $QAD(F, G) = QAD(G, F)$

Effect Size for Non-Normal Data

Divergence Effect Size: $D(F,G)$

This is a probability distance for constructing effect size, as it enables researchers to compare two populations regardless to the context. Suppose that X and Y are arbitrary random variable with cumulative distribution functions F and G respectively.

$$\begin{aligned} D(F||Q) &= 2\times\int_0^1 |G\{F^{-1}(p)\} - G\{G^{-1}(p)\}|dp \\ &= 2\times\int_0^1 |G\{F^{-1}(p)\} - p|dp \end{aligned}$$

This measure satisfies divergence properties, BUT it is not a symmetric measure, as $D(F|G)\neq D(G|F)$. A symmetric measure can be defined by:

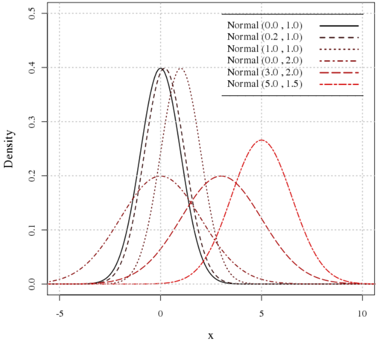
$$D(F,G) = \frac{1}{2}D(F||G) + \frac{1}{2}D(G||F)$$

The measure $D(F,G)$ is a bounded index, takes values between 0 and 1. This can be considered as an advantage of $D(F,G)$ in comparison to unbounded measures like Kullback-Leibler divergence.

Simulation for Perception about Effect Sizes

DIFFERENT PARAMETERS FOR NORMAL DISTRIBUTION

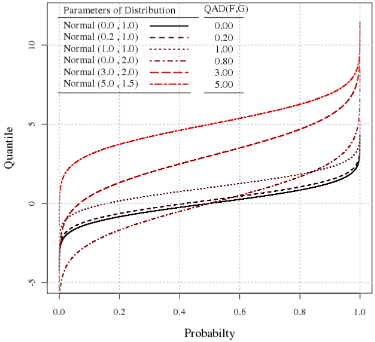
We compare five different normal distribution of different mean and standard deviation parameters with the standard normal distribution, as a control group.

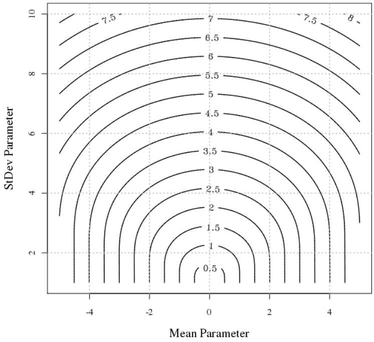


Simulation for Perception about Effect Sizes

DIFFERENT PARAMETERS FOR NORMAL DISTRIBUTION

We compare five different Normal distribution of different mean and standard deviation parameters to the standard normal distribution, as a control group.

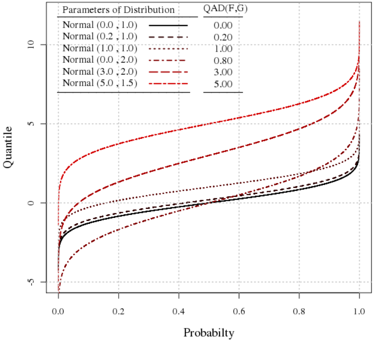




Simulation for Perception about Effect Sizes

DIFFERENT PARAMETERS FOR NORMAL DISTRIBUTION

We compare five different Normal distribution of different mean and standard deviation parameters to the standard normal distribution, as a control group.



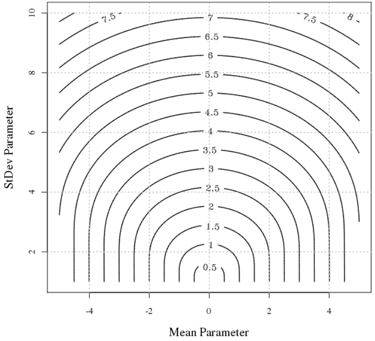
- The area between $N(0, 1)$ and the quantiles depicts the $QAD(P, G)$.
- The slight change of parameter make a slight difference in quantiles and small QAD.
- The mean of larger values cause the N(0,1) line to shift up
- larger standard deviation produce steeper curves
- The line $N(0, 2)$ intersect the $N(0, 1)$, as it has longer tails (the same mean, larger Std parameter)
- The non-intersects happend when the distribution dominates the $N(0, 1)$
-

Simulation for Perception about Effect Sizes

DIFFERENT PARAMETERS FOR NORMAL DISTRIBUTION

We compare five different Normal distribution of different mean and standard deviation parameters to the standard normal distribution, as a control group.

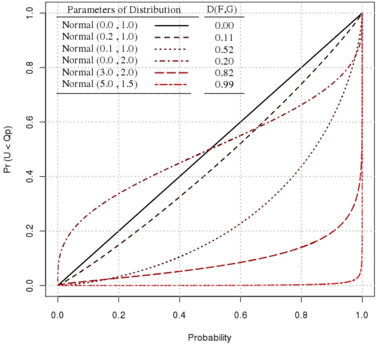
- Getting far from the mean point 0 will result in having larger effect size.
- As the standard deviations values increases, the steeper curves will be
- The effect size is more sensitive to the change of mean parameter rather than scale parameter

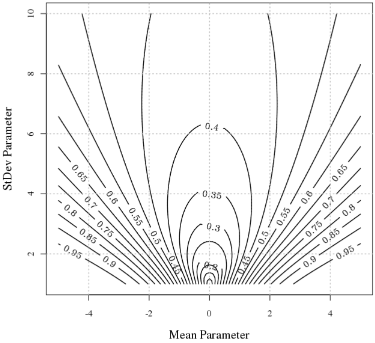


Simulation for Perception Divergence Effect Size

DIFFERENT PARAMETERS FOR NORMAL DISTRIBUTION

We compare five different Normal distribution of different mean and standard deviation parameters to the standard normal distribution, as a control group.

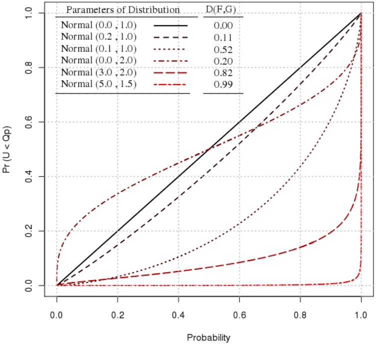




Simulation for Perception Divergence Effect Size

DIFFERENT PARAMETERS FOR NORMAL DISTRIBUTION

We compare five different Normal distribution of different mean and standard deviation parameters to the standard normal distribution, as a control group.



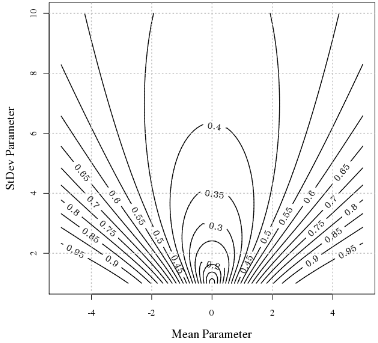
- The 45-degree straight line represents the vertical compasion function for control group $N(0, 1)$.
- The area between straight line and the curves depicts the $\mathrm{D}(P, G)$.
- The slight change of parameter make a slight difference in quantiles and small $\mathrm{D}(F, G)$.
- The mean of larger values cause the $N(0, 1)$ line to shift up.
- The line of $N(0, 2)$ intersects the $N(0, 1)$, as it has longer tails (the same mean, larger Std parameter).
- The non-intersect shape occures when the distribution dominates the $N(0, 1)$.

Simulation for Perception Divergence Effect Size

DIFFERENT PARAMETERS FOR NORMAL DISTRIBUTION

We compare five different Normal distribution of different mean and standard deviation parameters to the standard normal distribution, as a control group.

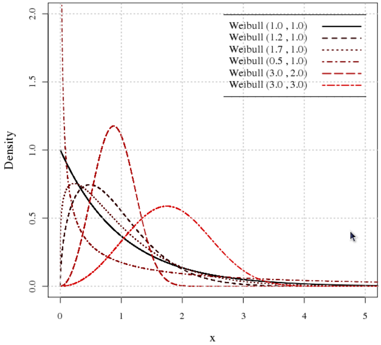
- Getting far from the mean point 0 will result in having larger effect size.
- For a fixed mean values close to the 0, increase of Std will increase the $\mathrm{D}(F, G)$.
- For a fixed mean values far from the 0, increase of Std will decrease the $\mathrm{D}(F, G)$.



Simulation for Perception about Effect Size

DIFFERENT PARAMETERS FOR WEIBULL DISTRIBUTION

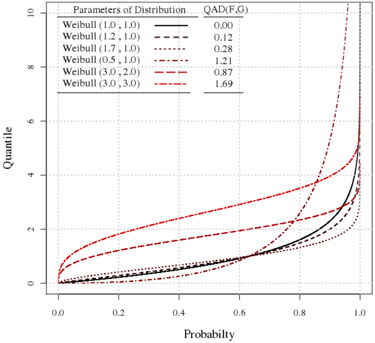
We compare the Weibull distribution of different shape and scale parameters to the Weibull distribution where both shape and scale parameters take $\alpha = \lambda = 1$ which is equivalent to the exponential distribution of the rate parameter 1.

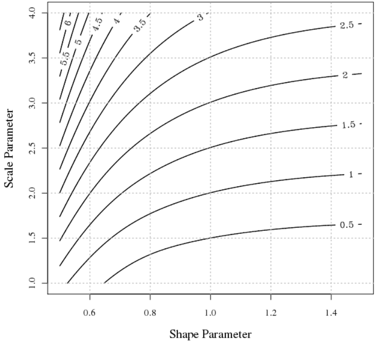


Simulation for Perception about Effect Sizes

DIFFERENT PARAMETERS FOR WEIBULL DISTRIBUTION

We compare the Weibull distribution of different shape and scale parameters to the Weibull distribution where both shape and scale parameters take $\alpha = \lambda = 1$ which is equivalent to the exponential distribution of the rate parameter 1.

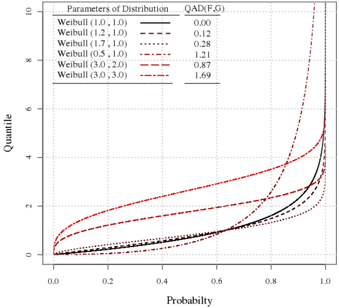




Simulation for Perception about Effect Sizes

DIFFERENT PARAMETERS FOR WEIBULL DISTRIBUTION

We compare the Weibull distribution of different shape and scale parameters to the Weibull distribution where both shape and scale parameters take $\alpha = \lambda = 1$, which is equivalent to the exponential distribution of the rate parameter 1.



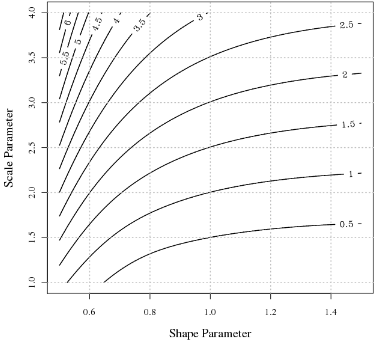
- The control group follows the *Weibull*$(1, 1)$ distribution (Solid line).
- The *Weibull*$(0.5, 1)$ line intersects the *Weibull*$(1, 1)$ as it has longer tails.
- The non-intersects distributions dominate the *Weibull*$(1, 1)$.
- The area between *Weibull*$(1, 1)$ curve and each of the other curves depicts the corresponding $D(P, G_i)$

Simulation for Perception about Effect Sizes

DIFFERENT PARAMETERS FOR WEIBULL DISTRIBUTION

We compare the Weibull distribution of different shape and scale parameters to the Weibull distribution where both shape and scale parameters take $\alpha = \lambda = 1$, which is equivalent to the exponential distribution of the rate parameter 1.

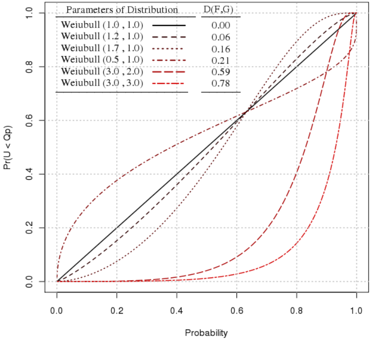
- For small values of shape parameter, change of the scale parameter make a big change for QAD.
- The QAD is not sensitive to the change of the shape parameters greater than 2.

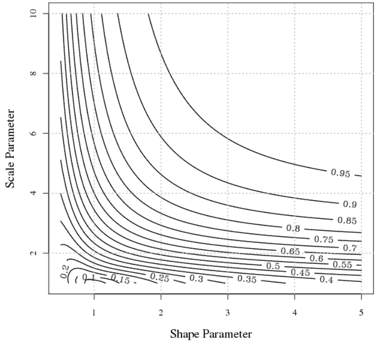


Simulation for Perception Divergence Effect Size

DIFFERENT PARAMETERS FOR WEIBULL DISTRIBUTION

We compare the Weibull distribution of different shape and scale parameters to the Weibull distribution where both shape and scale parameters take $\alpha = \lambda = 1$, which is equivalent to the exponential distribution of the rate parameter 1.





Outline

1. Introduction
   - Effect Size, Complementory for Hypothesis Tests
   - Cohen's D Effect Size
   - Common Language Effect Size (CLES)
   - Non-Overlap Effect Sizes
   - Non-parametric Effect Size
2. Effect Size for Non-Normal Data
   - Quantile Absolute Deviation
   - Divergence Effect Size
   - Simulation Study: Normal Distribution
   - Simulation Study: Weibull Distribution
3. Experimental Study: Clickstream Analysis
   - Data Source and Preparation
   - Model-Based Effect Size Computation
   - Bootstrapping

Data Source and Data Preparation

DATA SOURCE

- Server log files from commercial websites, selling products and services on the internet, belonging to clients of a local web management company.
- Conversion data files: purchase information for the visitors to the website. In the E-commerce cotext, the conversion is an action of online purchase whilst a user surf the website. A common statistic for web owner is the conversion rate hat show how many visitors to a website actually buy something.
- Registered users web data, demographical information of the visitor is not available.
- This data was collected from May 25th to June 2nd 2008,

DATA PREPARATION STEP

- The Main preprocessing tasks on the raw web log files executed by the company
  - Despidering: eliminating bots from the web log files
  - Filtering: eliminating irrelevant elements in the log files
  - User identification: generating the IDs to enable us to identify users
- Sessionization: splitting the sessions where the time stamp between two consequent page requests lasts more than 30 minutes
- Data extraction: Using original data to extract new variables (e.g. visit on week day, holidays, day time of the session, etc.).
- Filtering: Records of the sessions of only one page request (single-page visits) were filtered out

Data Source and Data Preparation

DATA SOURCE

- Server log files from commercial websites, selling products and services on the internet, belonging to clients of a local web management company.
- Conversion data files: purchase information for the visitors to the website. In the E-commerce cotext, the conversion is an action of online purchase whilst a user surf the website. A common statistic for web owner is the conversion rate hat show how many visitors to a website actually buy something.
- Registered users web data, demographical information of the visitor is not available.
- This data was collected from May 25th to June 2nd 2008,

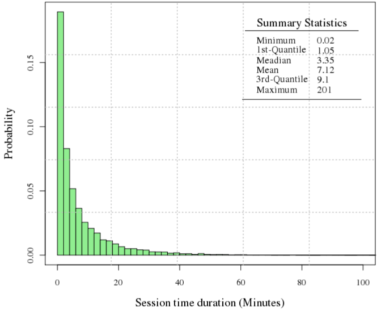
DATA PREPARATION STEP

- The Main preprocessing tasks on the raw web log files executed by the company
  - Despidering: eliminating bots from the web log files
  - Filtering: eliminating irrelevant elements in the log files
  - User identification: generating the IDs to enable us to identify users
- Sessionization: splitting the sessions where the time stamp between two consequent page requests lasts more than 30 minutes
- Data extraction: Using original data to extract new variables (e.g. visit on week day, holidays, day time of the session, etc.).
- Filtering: Records of the sessions of only one page request (single-page visits) were filtered out

Preprocessing and Data Preparation

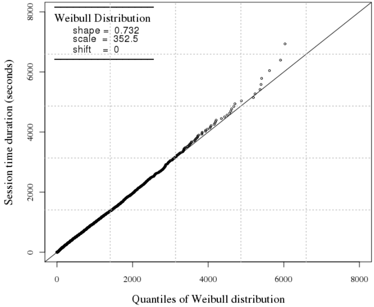
RESEARCH QUESTION

A comparison between customers who purchase goods online versus those who do not, with respect to the browsing time spent on the website. In other words, we want to investigate whether the time spent in a website is affected by the purchase decision of a visitor on that website?



Summary Statistics

Minimum	0.02
1st-Quantile	1.05
Meadian	3.35
Mean	7.12
3rd-Quantile	9.1
Maximum	201

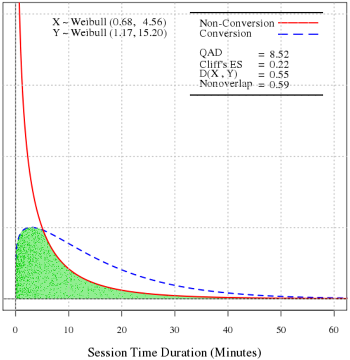


Weibull Distribution
shape = 0.732
scale = 352.5
shift = 0

Model-Based Effect Size

WEIBULL DISTRIBUTION FITTED ON EACH GROUP

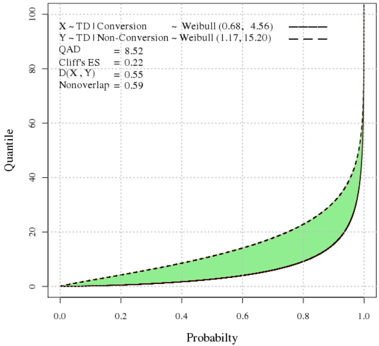
- Fit a Weibull distribution for the session time duration for the conversion group.
- Fit a Weibull distribution for the session time duration for the non-conversion group.
- Compute the effect sizes using the estimated parameters for each group

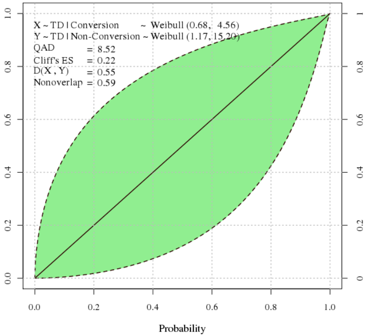


Model-Based Effect Size

WEIBULL DISTRIBUTION FITTED ON EACH GROUP

- Fit a Weibull distribution for the session time duration for the conversion group.
- Fit a Weibull distribution for the session time duration for the non-conversion group.
- Compute the effect sizes using the estimated parameters for each group

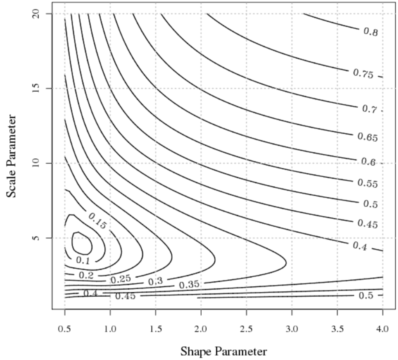




Model-Based Effect Size

MAGNITUDE OF THE EFFECT SIZE BASED ON PARAMETERS

Using a contour plot we will find the set of all parameters settings (α, λ) which produce the same numerical value for effect size as the observed effect size. This provides information about the magnitude of the effect size with respect to the Weibull distribution parameters.



Inference based on Bootstrapping

BOOTSTRAP FOR A PARAMETER

Given independent SRSs of size *n* from a population:

- Draw a resample of size *n* with replacement from the original sample
- Compute a statistic of interest (such as mean, median, quantiles, etc.)
- Repeat this resampling process hundreds of times
- Construct the bootstrap distribution of the statistic. Inspect its shape, bias, and bootstrap standard error in the usual way.

BOOTSTRAP FOR COMPARING TWO POPULATIONS

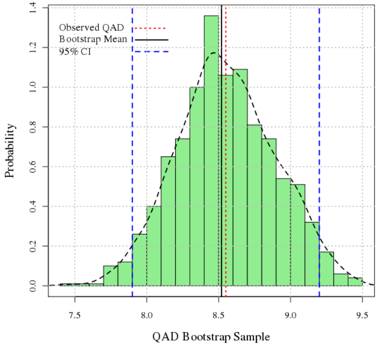
Given independent SRSs of sizes *n* and *m* from two populations:

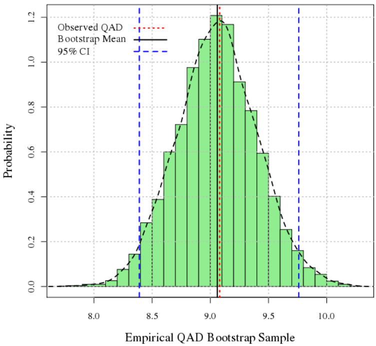
- Draw a resample of size *n* with replacement from the first sample and a separate resample of size *m* from the second sample.
- Compute a statistic that compares the two groups, such as the difference between the two sample means.
- Repeat this resampling process hundreds of times
- Construct the bootstrap distribution of the statistic. Inspect its shape, bias, and bootstrap standard error in the usual way.

Inference based on Bootstrapping

BOOTSTRAP: MODEL-BASED *QAD* AND EMPIRICAL *QAD* EFFECT SIZES

The bootstrap distribution for 10000 resample model-based *QAD* and empirical *QAD* effect sizes for time spent on the website for conversion and non-conversion groups.

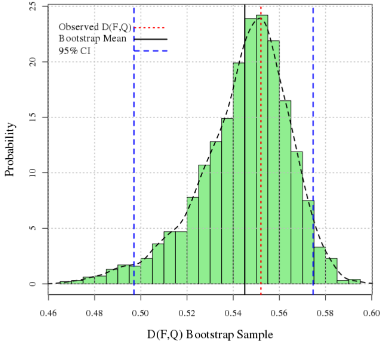


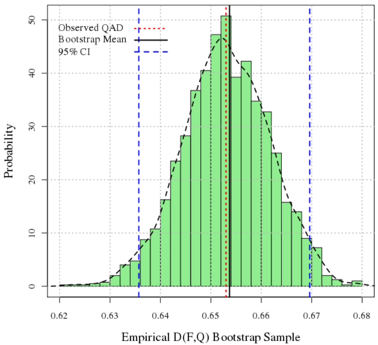


Inference based on Bootstrapping

BOOTSTRAP: MODEL-BASED $D(F,G)$ AND EMPIRICAL $D(F,G)$ EFFECT SIZES

The bootstrap distribution for 10000 resample model-based $D(F,G)$ and empirical $D(F,G)$ effect sizes for time spent on the website for conversion and non-conversion groups.





THANKS FOR YOUR ATTENTION!

THANKS FOR YOUR PATIENCE!