Final Project

Miles Keppler, Phillip Thomas

March 22, 2018

Abstract

Power and statistical significance are assessed with hypothesis testing for equality of means for two samples with different binomial link transformations. Issues of traditional p-value assessment are addressed and refined. The generalized test statistic for the two-sample t-test is explained and its limiting distribution is derived using the delta method, central limit theorem and weak law of large numbers. Simulate data 10,000 iterations were ran for each combination of sample sizes of 5, 15, 50 and 120 form each of the two groups and probabilities of 0.01, 0.05, 0.2, 0.5, 0.8, 0.95 and 0.99 for Bernoulli random variables. We conclude that effective size transformations are more robust under samples with differing sample sizes.

1 Research Question

We're considering the powers and significance levels error rates of transformations of Bernoulli random variables. The transformations we'll be looking at (and their derivatives) are

identity	$g_1(p) = p$	$g_1'(p) = 1$
log	$g_2(p) = \log(p)$	$g_2'(p) = \frac{1}{p}$
logit	$g_3(p) = \log\left(\frac{p}{1-p}\right)$	$g_3'(p) = \frac{1}{p - p^2}$
probit	$g_4(p) = \mathrm{qnorm(p)}$	$g_4'(p) = \frac{1}{\mathrm{dnorm}(\mathrm{qnorm}(p))}$
cloglog	$g_5(p) = \log(-\log(1-p))$	$g_5'(p) = -\frac{1}{(1-p)\log(1-p)}$.

The common theme here are that these are common binomial link transformations used in generalized linear regression. The identity and log transformations are used mainly for comparison.

2 Background

We use hypothesis testing in many fields of science as a statistically viable way to analyze and support quantitative reasoning. However, hypothesis tests by themselves do not provide adequate evidence to claim research findings. For example, p-values generated by implementing these tests can take on arbitrarily low values based on high sample sizes or even random chance from sampling. This causes trouble when it comes to replicating research claims and is referred to as the replication crisis. Issues with reproducibility are common in disciplines where categorical data analysis is required.

Benjamin et al. (2017) proposes a solution suggesting that a new cut-off a = 0.005 would make findings more reproducible. There is also a connection between this stricter cut-off and the Bayes factor, which relates the ratio of probabilities for H_1 and H_0. A p-value of 0.005 gives substantial to strong evidence according to Bayes factor classifications and also drastically decreases false positive rates for different power levels. The American Statistical Association (ASA) (Wasserstein and Lazar, 2016) recently discussed this issue and the misuse of p-values in hypothesis testing. It is mentioned that proper statistical inference requires full reporting, especially with binary decision making and just a p-value does not entirely illustrate what data tells us.

To avoid any matter of unreasonable conclusions, it is necessary to observe multiple measures of statistical significance when considering research findings; one of these being effect size. Effect size measures the practical significance, with a lower impact of sample sizes compared to p-values. When considering the two-sample case, some common effect sizes are standardized sample mean difference, log ratio and log odds of the sample mean, all of which give us evidence to enhance the strength of our conclusion. We will consider the aforementioned effect size measures along with the probit and cloglog transformations. Given that $\bar{X}_1$ and $\bar{X}_2$ are the sample means of the first and second sample, we can generalize these transformations as $g(\bar{X}_1) - g(\bar{X}_2)$.

3 Asymptotic Theory

Showing limiting distribution of T_g.

$$T_g = \frac{[g(\bar{X}_1) - g(\bar{X}_2)] - [g(\mu_1) - g(\mu_2)]}{\sqrt{\frac{[g'(\bar{X}_1)]^2 S_1^2}{n_1} + \frac{[g'(\bar{X}_2)]^2 S_2^2}{n_2}}}$$

Take $\lambda_1 = n_1/N, \lambda_2 = n_2/N$. By CLT

$$\sqrt{N}\left(\begin{pmatrix}\bar{X}_1\\ \bar{X}_2\end{pmatrix} - \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right) \xrightarrow{d} \mathrm{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\frac{\sigma_1^2}{\lambda_1} & 0\\ 0 & \frac{\sigma_2^2}{\lambda_2}\end{pmatrix}\right).$$

Using Δ-method,

$$\sqrt{N}\left(\begin{pmatrix}g(\bar{X}_1)\\ g(\bar{X}_2)\end{pmatrix} - \begin{pmatrix}g(\mu_1)\\ g(\mu_2)\end{pmatrix}\right) \xrightarrow{d} \mathrm{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}[g'(\mu_1)]^2\frac{\sigma_1^2}{\lambda_1} & 0\\ 0 & [g'(\mu_2)]^2\frac{\sigma_2^2}{\lambda_2}\end{pmatrix}\right).$$

Using bivariate Δ-method,

$$\sqrt{N}\left(f\begin{pmatrix}g(\bar{X}_1)\\ g(\bar{X}_2)\end{pmatrix} - f\begin{pmatrix}g(\mu_1)\\ g(\mu_2)\end{pmatrix}\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma_f^2\right).$$

Taking $f\begin{pmatrix}x_1\\ x_2\end{pmatrix} = f(x_1, x_2) = x_1 - x_2$,

$$\begin{aligned}
\sigma_f^2 &= \begin{bmatrix}\frac{\partial f(x_1,x_2)}{\partial x_1} & \frac{\partial f(x_1,x_2)}{\partial x_2}\end{bmatrix}\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}\begin{bmatrix}\frac{\partial f(x_1,x_2)}{\partial x_1}\\ \frac{\partial f(x_1,x_2)}{\partial x_2}\end{bmatrix}\\
&= \begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}[g'(\mu_1)]^2\frac{\sigma_1^2}{\lambda_1} & 0\\ 0 & [g'(\mu_2)]^2\frac{\sigma_2^2}{\lambda_2}\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix}\\
&= [g'(\mu_1)]^2\frac{\sigma_1^2}{\lambda_1} + [g'(\mu_2)]^2\frac{\sigma_2^2}{\lambda_2}
\end{aligned}$$

it can be shown

$$\sqrt{N}([g(\bar{X}_1) - g(\bar{X}_2)] - [g(\mu_1) - g(\mu_2)]) \xrightarrow{d} \mathrm{N}\left(0, \frac{[g'(\mu_1)]^2\sigma_1^2}{\lambda_1} + \frac{[g'(\mu_2)]^2\sigma_2^2}{\lambda_2}\right).$$

By WLLN,

$$\sqrt{\frac{\frac{[g'(\mu_1)]^2\sigma_1^2}{\lambda_1} + \frac{[g'(\mu_2)]^2\sigma_2^2}{\lambda_2}}{\frac{[g'(\bar{X}_1)]^2 S_1^2}{\lambda_1} + \frac{[g'(\bar{X}_2)]^2 S_2^2}{\lambda_2}}} \xrightarrow{p} 1.$$

By Slutsky's Theorem,

$$\sqrt{\frac{\frac{[g'(\mu_1)]^2\sigma_1^2}{\lambda_1} + \frac{[g'(\mu_2)]^2\sigma_2^2}{\lambda_2}}{\frac{[g'(\bar{X}_1)]^2 S_1^2}{\lambda_1} + \frac{[g'(\bar{X}_2)]^2 S_2^2}{\lambda_2}}} \frac{\sqrt{N}([g(\bar{X}_1) - g(\bar{X}_2)] - [g(\mu_1) - g(\mu_2)])}{\sqrt{\frac{[g'(\mu_1)]^2\sigma_1^2}{\lambda_1} + \frac{[g'(\mu_2)]^2\sigma_2^2}{\lambda_2}}} \xrightarrow{d} \mathrm{N}(0,1)$$

$$\implies \frac{\sqrt{N}([g(\bar{X}_1) - g(\bar{X}_2)] - [g(\mu_1) - g(\mu_2)])}{\sqrt{\frac{[g'(\bar{X}_1)]^2 S_1^2}{\lambda_1} + \frac{[g'(\bar{X}_2)]^2 S_2^2}{\lambda_2}}} \xrightarrow{d} \mathrm{N}(0,1).$$

Therefore,

$$\begin{aligned} T_g &= \frac{[g(\bar{X}_1) - g(\bar{X}_2)] - [g(\mu_1) - g(\mu_2)]}{\sqrt{\frac{[g'(\bar{X}_1)]^2 S_1^2}{n_1} + \frac{[g'(\bar{X}_2)]^2 S_2^2}{n_2}}} \\ &= \sqrt{N}\frac{[g(\bar{X}_1) - g(\bar{X}_2)] - [g(\mu_1) - g(\mu_2)]}{\sqrt{\frac{[g'(\bar{X}_1)]^2 S_1^2}{n_1/N} + \frac{[g'(\bar{X}_2)]^2 S_2^2}{n_2/N}}} \\ &= \frac{\sqrt{N}([g(\bar{X}_1) - g(\bar{X}_2)] - [g(\mu_1) - g(\mu_2)])}{\sqrt{\frac{[g'(\bar{X}_1)]^2 S_1^2}{\lambda_1} + \frac{[g'(\bar{X}_2)]^2 S_2^2}{\lambda_2}}} \\ &\xrightarrow{d} \mathrm{N}(0,1). \end{aligned}$$

□

4 Methods

To find empirical power and robustness, generated 10 000 simulations of each test, comparing samples of Bernoulli random variables. Each test was ran with significance level $\alpha = .05$. Parameters were adjusted to compare all combinations of samples of size $n_1 \in \{5, 15, 50, 120\}$ from $\mathrm{Bernoulli}(p_1)$ with $p_1 \in \{.05, .2, .5, .8, .95\}$ to samples of size $n_2 \in \{5, 15, 50, 120\}$ from $\mathrm{Bernoulli}(p_2)$ with $p_2 \in \{.01, .05, .2, .5, .8, .95, .99\}$. Examined upper-tailed, lower-tailed, and two-sided tests for each cases, resulting in a total of 1680 different cases. The test statistic we used was the one discussed above,

$$T_g = \frac{[g(\bar{X}_1) - g(\bar{X}_2)] - [g(\mu_1) - g(\mu_2)]}{\sqrt{\frac{[g'(\bar{X}_1)]^2 S_1^2}{n_1} + \frac{[g'(\bar{X}_2)]^2 S_2^2}{n_2}}}.$$

T_g is approximately $t(v_g)$ distributed, where the effective degrees of freedom are

$$v_g = \frac{\left(\frac{[g'(\bar{X}_1]^2 S_1^2}{n_1} + \frac{[g'(\bar{X}_2]^2 S_2^2}{n_2}\right)^2}{\frac{[g'(\bar{X}_1]^4 S_1^4}{n_1^2(n_1-1)} + \frac{[g'(\bar{X}_2]^4 S_2^4}{n_2^2(n_2-1)}}.$$

5 Results

Figures for results are provided at the end of the paper. The first noticeable difference was the poor empirical significance level in low sample sizes for all transformations of the mean. Going up in sample size provides much better results, and actually shows an improvement in power over the identity and log transformations. Some cases (as seen in in figure 1) show extremely poor results from the log transformation, while the identity function provides consistent results in all cases. The log and cloglog transformations provide similar results to the probit and logit transformations under lower values of p, while probit and logit are nearly the same in all cases. This would be expected from comparing the transformations themselves. Link transformations provide strong powers under unequal sample sizes, much better than the identity and more robust that log. The

cloglog transformation generally provides worse results in all aspects when compared to the probit and logit models. Differing significance levels only exemplified errors and strengthened the results already discussed.

6 Conclusion & Further Research

The probit and logit transformations provide robust sizes and powers under extreme parameter values, however can perform poorly under low sample sizes. Unequal sample sizes seem to provide low power without transformation, but power goes up with any of the link transformations. Use these transformations with sample sizes 50 and above, and with extreme parameter values or with unequal sample sizes to improve empirical results. With an appropriate transformation, generalized linear regression on proportions should retain normality in errors, and comparing proportions in a reasonable task. We would be interested in future research on the use of the transformations beta random variables and perhaps other distributions that represent proportions. Bernoulli random variables are somewhat boring, as the only have one parameter. Would like to look at inverses of these transformations as well, as that is important in interpreting the results.

References

[1] Daniel J Benjamin et al. "Redefine statistical significance". In: *Nature Human Behaviour* 2.1 (2018), p. 6.

[2] Kimihiro Noguchi and Fernando Marmolejo-Ramos. "Assessing Equality of Means Using the Overlap of Range-Preserving Confidence Intervals". In: *The American Statistician* 70.4 (2016), pp. 325–334. DOI: 10.1080/00031305.2016.1200487. eprint: https://doi.org/10.1080/00031305.2016.1200487. URL: https://doi.org/10.1080/00031305.2016.1200487.

[3] Ronald L. Wasserstein and Nicole A. Lazar. "The ASA's Statement on p-Values: Context, Process, and Purpose". In: *The American Statistician* 70.2 (2016), pp. 129–133. DOI: 10.1080/00031305.2016.1154108. eprint: https://doi.org/10.1080/00031305.2016.1154108. URL: https://doi.org/10.1080/00031305.2016.1154108.

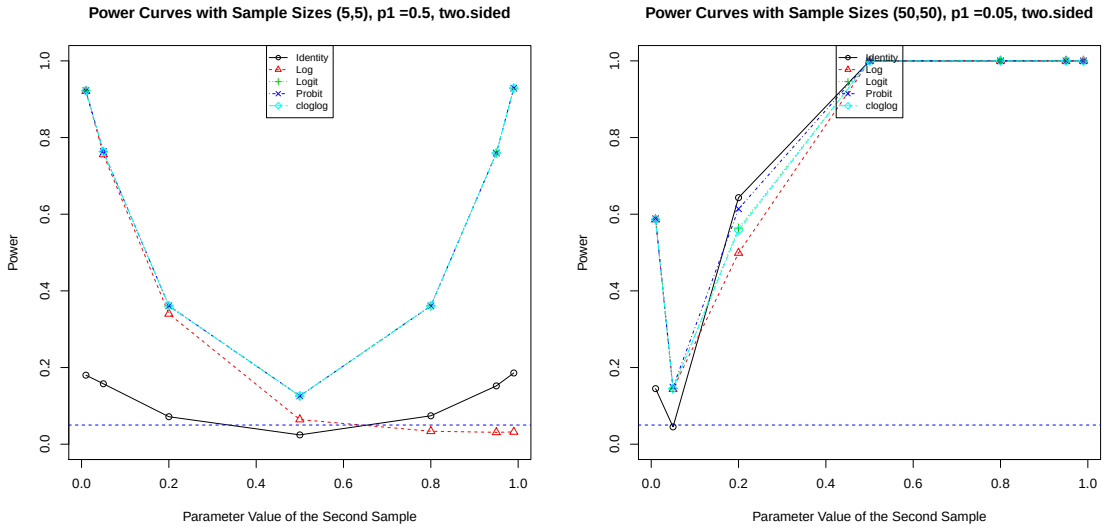

Figure 1: Transformations are from theoretical significance value of .05

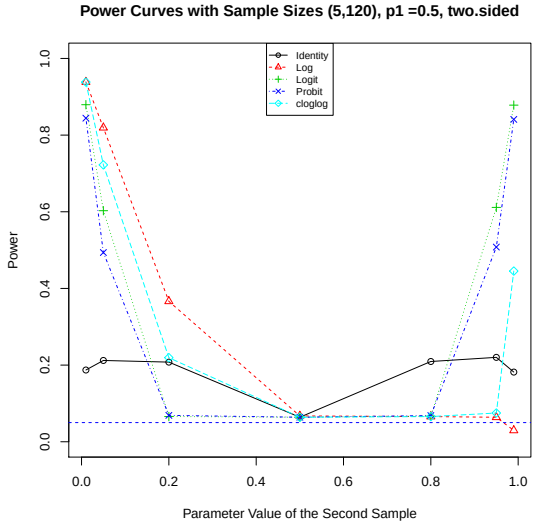

Figure 2: Transformations have high power with extreme parameter values, but less robust under middle values.

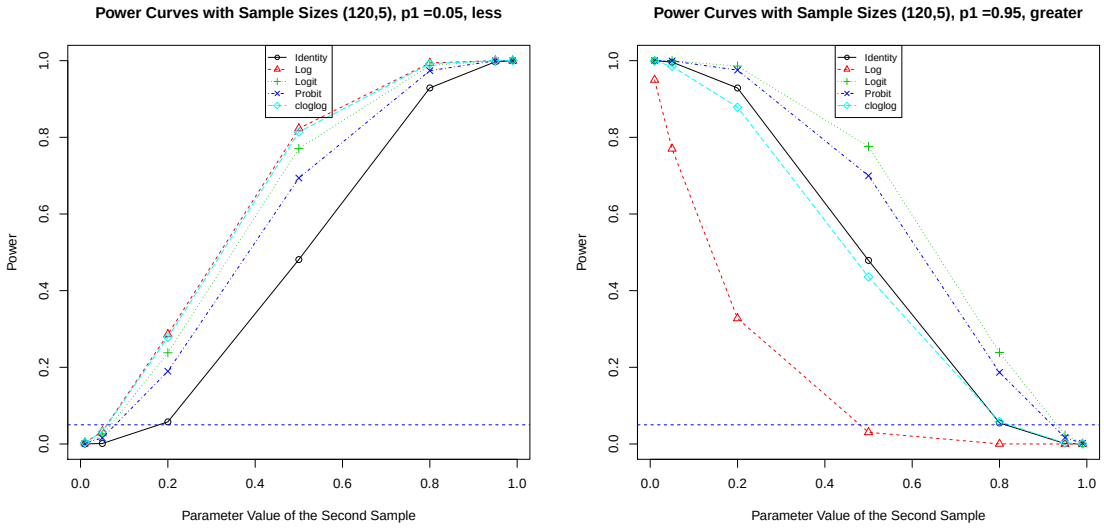

Figure 3: Transformations provide high power under unequal sample sizes and extreme parameters, and have robust empirical significance levels.