Power Calculations for Microbiome Data

Andy Shi

1 Research Question

How well do assumptions of power calculations based on the *t*-test and ANOVA hold for microbiome data, and how does predicted power compare to empirical power computed from simulation? How can we improve upon existing methods for calculating power in the context of the microbiome?

2 Background

2.1 The Human Microbiome

In the past 10 years, scientists have been increasingly interested in studying the human microbiome. This bacterial community is extremely large and diverse. In fact, there are 10 times more bacterial cells on the human body than human cells, and these bacteria perform various functions and have been linked to human health and disease. For example, certain gut bacteria have been linked to Crohn's disease [1] and colorectal cancer [2].

However, the microbiome is as yet poorly characterized and understood [3]. While researchers have made headway in determining which bacterial species are present in a host using new technologies, knowledge of these species' diverse interactions or functions, especially as they relate to human disease, is sorely lacking.

To learn more about the relationship between the microbiome and human disease, researchers often use case-control studies to discover which bacterial species are associated with a disease status. They will test the relative proportions of each bacterial species between two or more groups of patients with different disease statuses to look for a statistically significant difference, after doing a correction for multiple hypothesis testing. These studies can provide a shortlist of candidate bacteria or bacterial metabolic pathways associated with disease status. Further work can focus on investigating the biology behind these associations, which can possibly lead to drug targets or other insights into a disease.

For example, Morgan et al. [4] investigated the precise role of the microbiome in inflammatory bowel disease (IBD) and found associations between the bacterial species *Firmicutes* and *Enterobacteriaceae* and disease status, as well as differences in metabolic pathways between diseased and healthy patients.

Before such microbiome case-control studies are performed, researchers, funding agencies, and Institutional Review Boards (IRBs) usually try to gauge the potential risks and benefits of the study. They may want to know what kinds of differences are detectable, or learn how many patients need to be studied to detect such differences. Power calculations can be used to answer these questions.

2.2 Power Calculations

Statistical power is defined as the probability of rejecting a null hypothesis, given that it is false, and depends on sample size, variability, probability of Type I error, and effect size. In the context of the microbiome, power calculations are difficult because of the nature of microbiome data. Application of standard techniques, such as the *t*-test and ANOVA, to microbiome data makes assumptions that cannot hold.

In particular, because a microbiome sample must first be amplified by PCR, data from microbiome experiments are proportions summing to 1, instead of raw counts. If one bacteria increases in number, its proportion will increase, and the proportions of the other species will fall. Additionally, standard power calculations assume that, after applying an arcsine square-root transformation, the proportions of bacteria will follow a normal distribution. However, bacterial counts are highly skewed, with the most abundant species being several orders of magnitude more expressed compared to the least abundant ones. Additionally, many rare species of bacteria are not found in all people, thus leading to zero-inflation in the proportions. All these issues indicate that calculating power based on *t*-tests or ANOVA may be problematic.

This project aims to investigate the assumptions behind the *t*-test and ANOVA, especially normality, and assess how the empirical power differs from that calculated using *t*- and ANOVA-based methods after adjusting for multiple comparisons using a Bonferroni correction.

3 Methods

3.1 Design of Experiments

We assess both the normality assumption and empirical power through simulation.

Let $i \in \{1, \ldots, I\}$ be the number of groups, or cohorts. Typically, $I \geq 2$, corresponding to two or more cohorts of patients. Let there be N_i patients per group, for each i. For each patient, we can observe K bacterial taxa, indexed by $k \in \{1, \ldots, K\}$. Additionally, let $p_z \in [0, 1]$ indicate the probability of zero-inflation.

In the simulation, counts for bacterial species k in group i are first drawn from a LogNormal distribution with parameters $\mu_{i,k}$ and $\sigma^2_{i,k}$. Then, the counts for bacterial species k in patient n in group i are then generated independently according to

$$C_{i,n,k} \sim (1 - Z_{i,n,k}) \cdot \mathcal{LN}(\mu_{i,k}, \sigma^2_{i,k}), \tag{1}$$

where $Z_{i,n,k}$ are iid draws from $\text{Bern}(p_z)$. The counts are then normalized and arcsin square-root transformed:

$$Y_{i,n,k} = \arcsin\left(\sqrt{\frac{C_{i,n,k}}{\sum_{k=1}^{K} C_{i,n,k}}}\right). \tag{2}$$

It is assumed that the sample means of the $Y_{i,n,k}$

$$\bar{Y}_{i,k} = \frac{1}{N_i} \sum_{n=1}^{N_i} Y_{i,n,k}, \tag{3}$$

follow normal distributions.

3.2 Simulation Validation

To validate our simulation, we compared results to gut samples from data in the study by Morgan et al [4]. This dataset consisted of count data $c_{n,k}$ of number of sequences for bacterial taxon k, person $n \in \{1, 2, \ldots, N\}$. Marginal means and standard deviations for these counts are shown in Figure 13. These span several orders of magnitude, indicating the diversity of the microbiome.

Counts were normalized and transformed, according to Equation 2, yielding transformed proportions $y_{n,k}$ for the count of bacterial taxon k for person n.

To compare our simulation against this IBD dataset, we first estimated the means μ_k and standard deviations σ_k^2 of the LogNormal distribution in Equation 1 for each taxa k (note here that $I = 1$):

$$\hat{\mu}_k = \frac{\sum_n c_{n,k}}{\sum_n 1\{c_{n,k} \neq 0\}} \tag{4}$$

$$\hat{\sigma}_k^2 = \frac{\sum_n 1\{c_{n,k} \neq 0\} \cdot (c_{n,k} - \hat{\mu}_k)^2}{-1 + \sum_n 1\{c_{n,k} \neq 0\}}. \tag{5}$$

Note that for any event A,

$$1\{A\} = \begin{cases} 1, & A \text{ occurs} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Additionally, we estimated the percentage of zero inflation p_z to be the fraction of entries in our data that were zero.

Using these $\hat{\mu}_k$, $\hat{\sigma}_k^2$, and p_z, data was simulated (with $I = 1$) according to Equations 1 and 2. The means of

$$\bar{Y}_k = \frac{1}{N} \sum_{n=1}^{N} Y_{n,k} \tag{7}$$

for each taxon k across 10000 simulations was plotted against the empirical means

$$\bar{y}_k = \frac{1}{N} \sum_{n=1}^{N} y_{n,k} \tag{8}$$

in a QQ plot, as shown in Figure 1. The near-linear relationship between the observed and simulated microbiome data supports the validity of our simulation. Additional data from other members of the Huttenhower lab has shown that a LogNormal distribution performs reasonably for modeling underlying counts for microbiome data.

3.3 Assessing Normality

We would like to assess, for a given taxa k and group i, the distribution of the marginal mean of the relative abundance for each taxa, $\bar{Y}_{i,k}$.

The distribution of $\bar{Y}_{i,k}$ for a particular i and k are assessed using a quantile-quantile plot, compared against a normal distribution. We desire $\bar{Y}_{i,k}$ to be normally distributed because t-based tools are used to calculate the power of testing $\bar{Y}_{i,k}$ vs. $\bar{Y}_{j,k}$.

We assessed normality under several conditions. For each of the following conditions, we used $I = 1$ groups and assessed normality for $n_1 \in \{10, 50, 100, 500, 1000\}$. $K = 100$ taxa were used. For each n_1, 1000 simulations were run. Two sets of means and variances were used, and the levels of zero-inflation were varied also. The experimental conditions are described below:

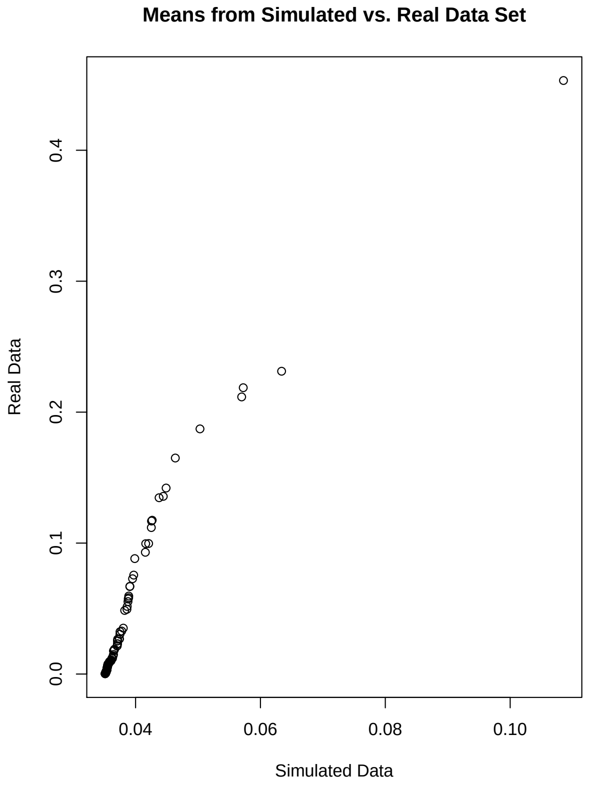


Figure 1: QQ Plot of taxa means from observed microbiome data (y-axis) vs. means from average over 10000 iterations of simulated data (x-axis).

1. All the $\mu_{1,k} = 0$ and all the variances $\sigma^2_{1,k} = 1$ for all k, and $p_z = 0$. Results shown in Figure 2.
2. Means were changed according to Equation 9, variances were kept the same as the previous configuration, and p_z set to 0. Results shown in Figure 6.

$$\mu_{1,k} = \begin{cases} 1 & : k \bmod 4 = 1 \\ 2 & : k \bmod 4 = 2 \\ 3 & : k \bmod 4 = 3 \\ 4 & : k \bmod 4 = 0 \end{cases} \tag{9}$$

3. $\mu_{1,k} = 0$ for all k, variances changed according to Equation 10, and p_z set to 0. Results shown in Figure 7.

$$\sigma^2_{1,k} = \begin{cases} 1 & : k \bmod 4 = 1 \\ 2 & : k \bmod 4 = 2 \\ 3 & : k \bmod 4 = 3 \\ 4 & : k \bmod 4 = 0 \end{cases} \tag{10}$$

4. Means were set according to Equation 9, variances according to Equation 10, and p_z set to 0. Results shown in Figure 3.
5. $p_z = 0.1$, and for all k, $\mu_{1,k} = 0$ and $\sigma^2_{1,k} = 1$. Results shown in Figure 8.

6. $p_z = 0.1$, and means and variances were set using Equations 9 and 10, respectively. Results shown in Figure 10.

7. $p_z = 0.5$, and for all k, $\mu_{1,k} = 0$ and $\sigma^2_{1,k} = 1$. Results shown in Figure 11.

8. $p_z = 0.5$, and means and variances were set using Equations 9 and 10, respectively. Results shown in Figure 12.

3.4 Assessing Empirical Power

We would like to determine the power when comparing transformed proportions of a particular taxa across groups, e.g. $\bar{Y}_{i,k}$ vs. $\bar{Y}_{j,k}$. To evaluate empirical power, we simulated 1000 synthetic microbiome datasets with 2 groups ($I = 2$) and varying parameters described below. In each simulation, we fixed a difference between bacteria of a specific taxa k. We conduct a t-test between the transformed proportions $\bar{Y}_{1,k}$ vs. $\bar{Y}_{2,k}$ and calculated the power of a t-test to detect this difference by examining the proportion of empirical p-values less than Bonferroni-adjusted cutoff of 0.05 / (total number of taxa).

3.4.1 Experimental Setup

We used $I = 2$ groups of patients, with equal number of patients in each group. The sample sizes for each group are $n \in \{10, 25, 50, 100, 150, 200, 250, 500\}$. We tested three different configurations for $p_z \in \{0, 0.1, 0.5\}$. Means and variances of the underlying LogNormal distribution were set according to four configurations. Note that in each configuration, the difference we are interested in testing is between bacterial species 1.

1. $\mu_{1,1} = 0.5$, and all other μ set to 0; σ^2 all set to 1.

2. μ set to the formula in Equation 11; σ^2 all set to 1.

$$\mu_{i,k} = \begin{cases} 0.5, & k \bmod 5 = 1 \text{ and } i = 1 \\ 0.5, & k \bmod 5 = 2 \text{ and } i = 2 \\ 0, & k \bmod 5 = 1 \text{ and } i = 2 \\ 0, & k \bmod 5 = 2 \text{ and } i = 1 \\ 1, & k \bmod 5 = 3 \\ 2, & k \bmod 5 = 4 \\ 3, & k \bmod 5 = 0 \end{cases} \tag{11}$$

3. $\mu_{1,1} = 0.5$, and all other μ set to 0; σ^2 set to Equation 12.

$$\sigma^2_{i,k} = \begin{cases} 0.5, & k \bmod 5 = 1 \text{ and } i = 1 \\ 0.5, & k \bmod 5 = 2 \text{ and } i = 2 \\ 1, & k \bmod 5 = 1 \text{ and } i = 2 \\ 1, & k \bmod 5 = 2 \text{ and } i = 1 \\ 2, & k \bmod 5 = 3 \\ 3, & k \bmod 5 = 4 \\ 4, & k \bmod 5 = 0 \end{cases} \tag{12}$$

4. μ set according to Equation 13; σ^2 set to Equation 12.

$$\mu_{i,k} = \begin{cases} 0.5, & k \bmod 2 = 1 \text{ and } i = 1 \\ 0.5, & k \bmod 2 = 0 \text{ and } i = 2 \\ 0, & k \bmod 2 = 0 \text{ and } i = 1 \\ 0, & k \bmod 2 = 1 \text{ and } i = 2 \end{cases} \tag{13}$$

3.4.2 Calculating Theoretical Power

The theoretical power was calculated using t tools. Because we assume normality for the transformed proportions $Y_{i,n,k}$ instead of the counts $C_{i,n,k}$, we have no direct way of computing the means and variances for $Y_{i,n,k}$ from those for the LogNormal distributions underlying $C_{i,n,k}$. To compute these summary statistics for $Y_{i,n,k}$. Instead, we simulated 10000 data sets with the same settings as those used to calculate the empirical power, and we used the mean (across 10000 simulations) sample mean and sample variance of the nonzero entries to estimate the mean and variance of $Y_{i,n,k}$:

$$\hat{\mu}_{i,k} = \frac{\sum_{n=1}^{N_i} y_{i,n,k}}{\sum_{n=1}^{N_i} 1\{y_{i,n,k} \neq 0\}} \tag{14}$$

$$\hat{\sigma}^2_{i,k} = \frac{\sum_{n=1}^{N_i} 1\{y_{i,n,k} \neq 0\} \cdot (y_{i,n,k} - \hat{\mu}_{i,k})^2}{-1 + \sum_{n=1}^{N_i} 1\{y_{i,n,k} \neq 0\}} \tag{15}$$

We calculate the power of the unpooled two-sample t-test to test $\bar{Y}_{i,k}$ vs.$\bar{Y}_{j,k}$ as follows: Let the degrees of freedom ν be given by

$$\nu = \frac{\left(\frac{\hat{\sigma}^2_{i,k}}{N_i} + \frac{\hat{\sigma}^2_{j,k}}{N_j}\right)^2}{\frac{\hat{\sigma}^4_{i,k}}{N_i^2(N_i-1)} + \frac{\hat{\sigma}^4_{j,k}}{N_j^2(N_j-1)}}, \tag{16}$$

and let the significance level $\alpha = 0.05/K$, where K is the total number of taxa. Then, the critical value is $t_{\nu,1-\frac{\alpha}{2}}$, the $1 - \frac{\alpha}{2}$ quantile from a standard t distribution with ν degrees of freedom. We use a noncentral t distribution to calculate the power. The noncentrality parameter is calculated as

$$ncp = \frac{\hat{\mu}_{i,k} - \hat{\mu}_{j,k}}{\sqrt{\frac{\hat{\sigma}^2_{i,k}}{N_i} + \frac{\hat{\sigma}^2_{j,k}}{N_j}}}, \tag{17}$$

so our power is

$$P(T_\nu > t_{v,1-\frac{\alpha}{2}}) + P(T_\nu < -t_{v,1-\frac{\alpha}{2}}), \tag{18}$$

where T_ν follows a noncentral t distribution with ν degrees of freedom and noncentrality parameter ncp.

4 Results

4.1 Normality Assumption

The main results are shown in Figure 2 and 3. We see deviations from normality, especially in smaller sample sizes, as we change the variances or increase p_z, the probability of zero inflation. As sample size increases, normality seems to improve, but there are still problems at the tails. The lack of normality, especially for smaller sample sizes, indicates a potential problem for calculating power based on t-tests or ANOVA.

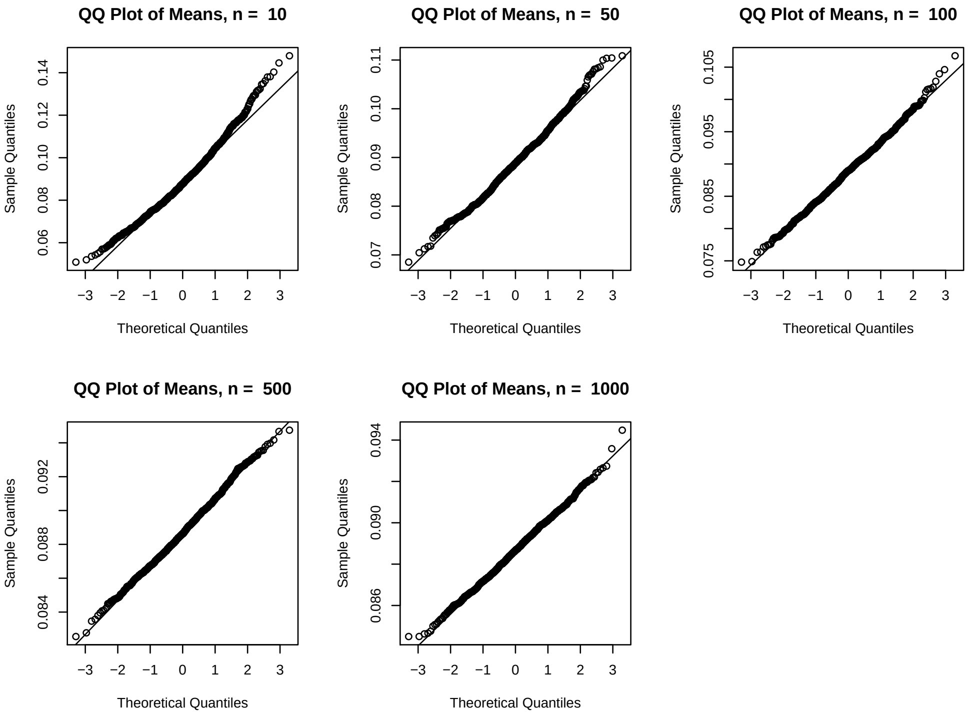


Figure 2: QQ Plots of $\bar{Y}_{i,k}$ for a randomly chosen k, under equal means and variances.

We see QQ plots of transformed and untransformed proportions for a single bacterial taxa across 1000 patients in Figure 4. We see a high degree of skewing, with a few extremely large values caused by simulating from a LogNormal distribution. This skewing drives the deviations from normality of the sample means $\bar{Y}_{i,k}$.

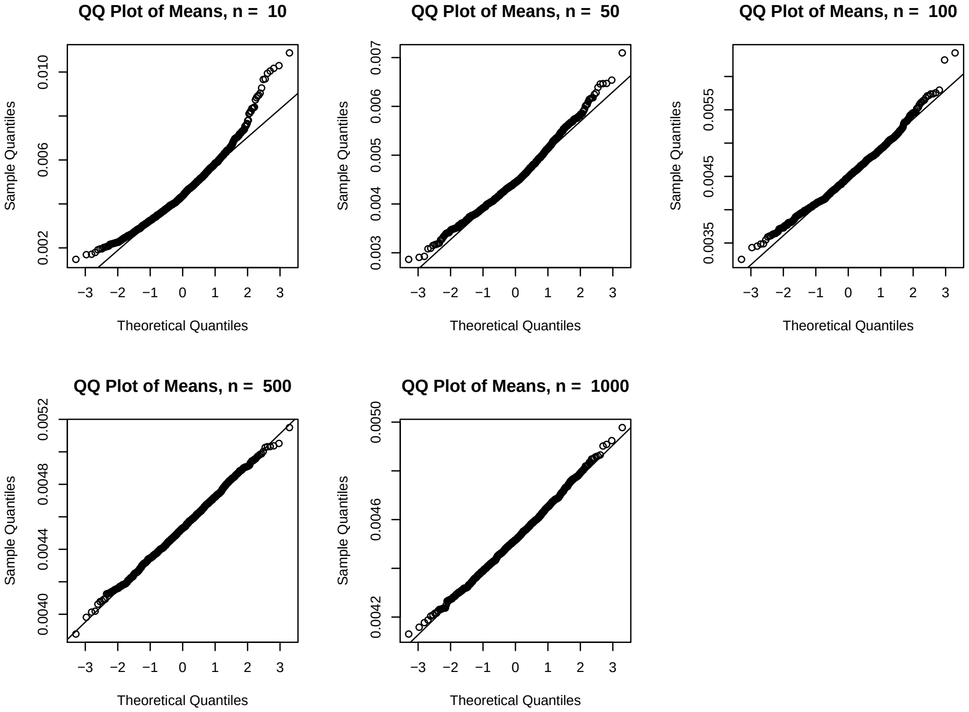


Figure 3: QQ Plots of $\bar{Y}_{i,k}$ for a randomly chosen k, under unequal means and unequal variances.

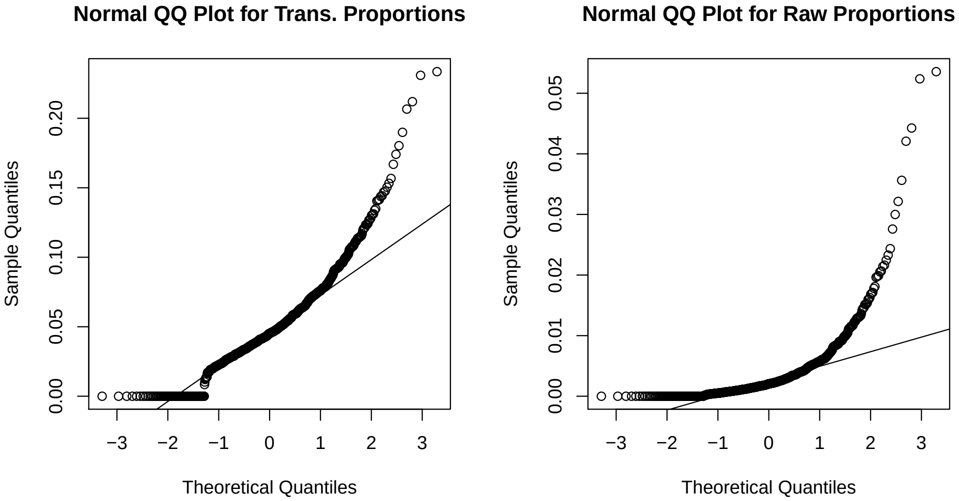


Figure 4: QQ Plots of transformed proportions $Y_{i,n,k}$ (left) and untransformed proportion $(\sin(Y_{i,n,k}))^2$ (right) for one particular bacterial species.

4.2 Empirical Power

4.2.1 Marginal *t*-test

Plots of empirical and theoretical power across different testing conditions are shown in Figure 5. As the amount of zero inflation becomes worse (moving from left to right in the figure), the difference between the theoretical power and the empirical power becomes greater. Overall, the *t*-tools tend to overestimate the actual amount of power.

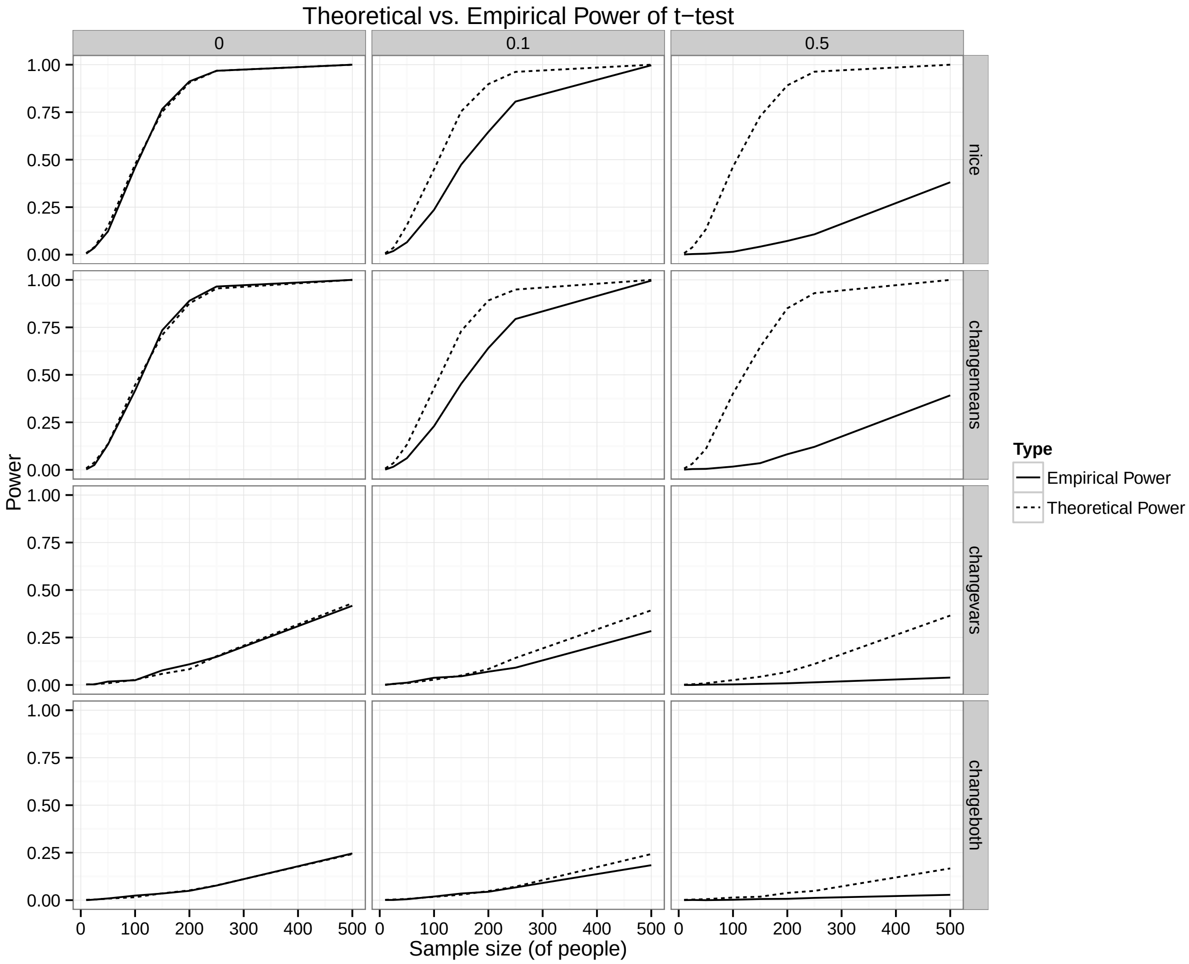


Figure 5: Plot of Empirical and Theoretical power, as a function of sample size. Solid line: Empirical power; Dotted line: theoretical power. The columns represent different levels of zero inflation, and the rows correspond to each of the simulation schemes described earlier.

5 Discussion

Our simulations show deviations from normality of the sample means of transformed proportions, which led us to believe calculating power using parametric tools like the *t*-tools might be problematic. Indeed, we observed that theoretical power computed via *t*-tools generally overestimated the empirical power of the *t*-test via simulation, and that this overestimation increased as the amount of zero inflation increased. Our result suggests that using *t*-based tools is not sufficient to calculate power in the context of the microbiome. The next step for this project is to develop an improved method for computing these power calculations.

Acknowledgment

The author would like to thank the staff of Stat 98 and Huttenhower lab, especially Emma Schwager, for their helpful comments on this project, and Boyu Ren, for providing insights into simulation procedures and data to validate the simulations.

References

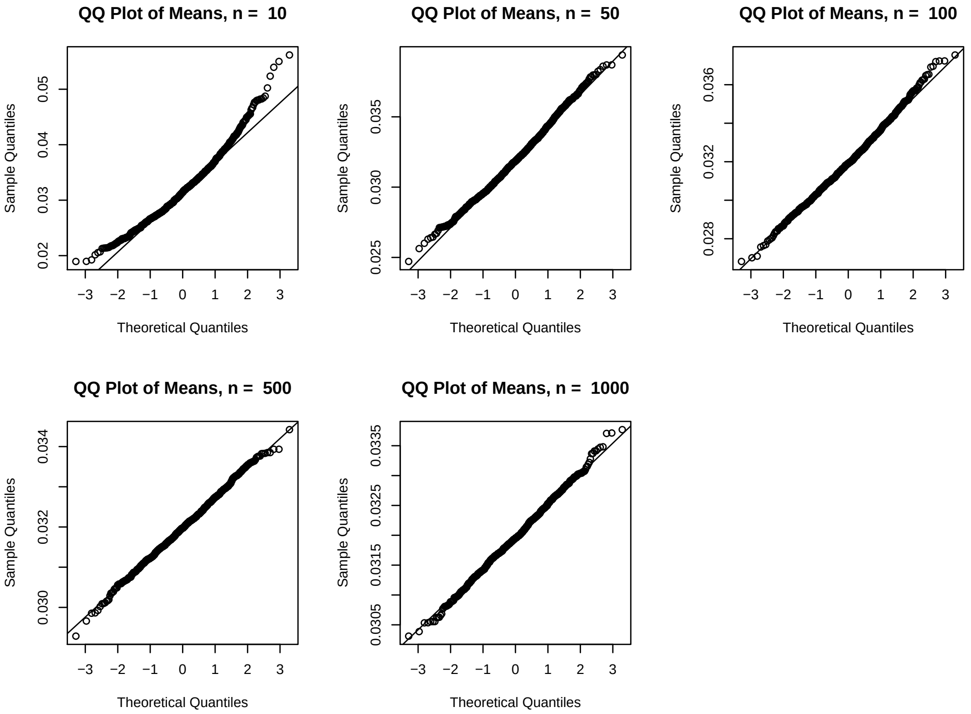
[1] Dirk Gevers et al. "The treatment-naive microbiome in new-onset Crohn's disease." In: *Cell host & microbe* 15.3 (Mar. 12, 2014), pp. 382–392. ISSN: 1934-6069. URL: `http://view.ncbi.nlm.nih.gov/pubmed/24629344`.

[2] Aleksandar D. Kostic et al. "Genomic analysis identifies association of Fusobacterium with colorectal carcinoma". In: *Genome Research* 22.2 (Feb. 1, 2012), pp. 292–298. ISSN: 1549-5469. URL: `http://dx.doi.org/10.1101/gr.126573.111`.

[3] Junjie Qin et al. "A human gut microbial gene catalogue established by metagenomic sequencing". In: *Nature* 464.7285 (Mar. 4, 2010), pp. 59–65. ISSN: 0028-0836. URL: `http://dx.doi.org/10.1038/nature08821`.

[4] Xochitl C. Morgan et al. "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment". In: *Genome Biology* 13.9 (Sept. 26, 2012), R79+. ISSN: 1465-6906. URL: `http://dx.doi.org/10.1186/gb-2012-13-9-r79`.

A Supplementary Figures


QQ Plot of Means, n = 10

QQ Plot of Means, n = 50

QQ Plot of Means, n = 100

QQ Plot of Means, n = 500

QQ Plot of Means, n = 1000

Figure 6: QQ Plots of $\bar{Y}_{i,k}$ for a randomly chosen k, under different means but equal variances.

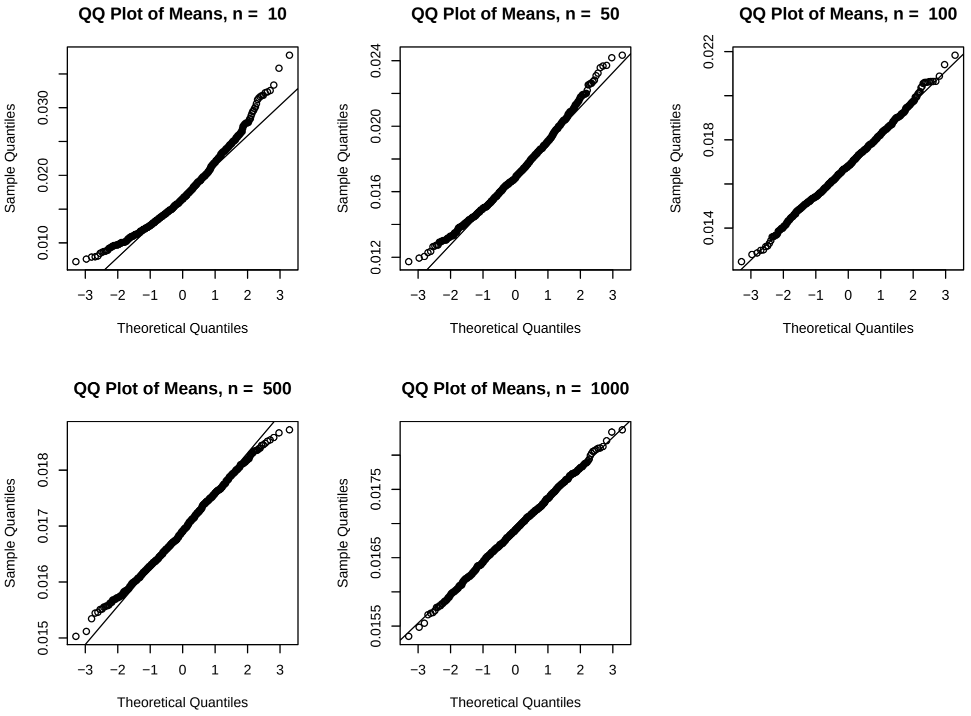


Figure 7: QQ Plots of $\bar{Y}_{i,k}$ for a randomly chosen k, under equal means but different variances.

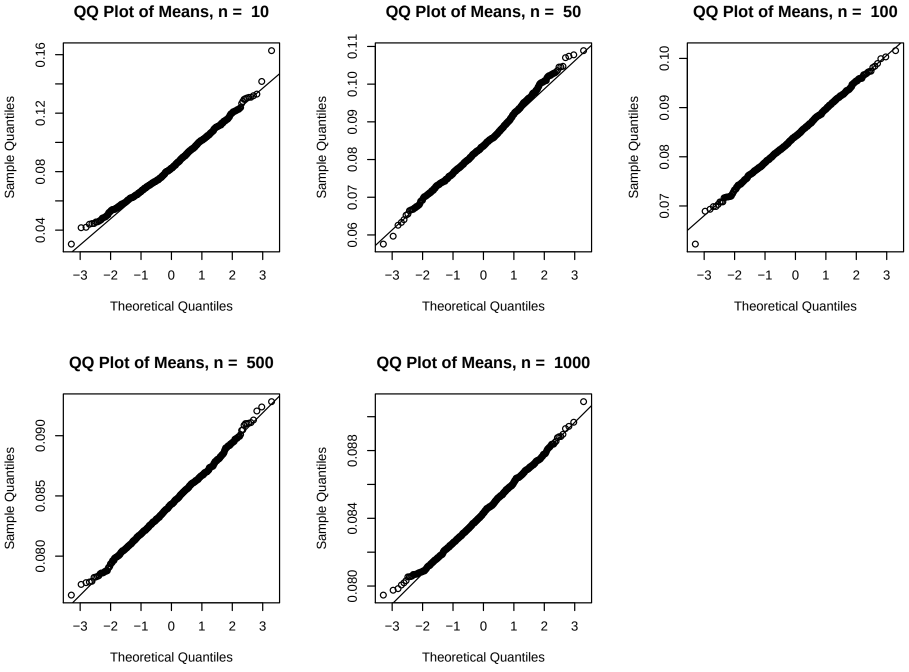


Figure 8: QQ Plots of $\bar{Y}_{i,k}$ for a randomly chosen k, under equal means and variances, with 10% zero-inflation probability.

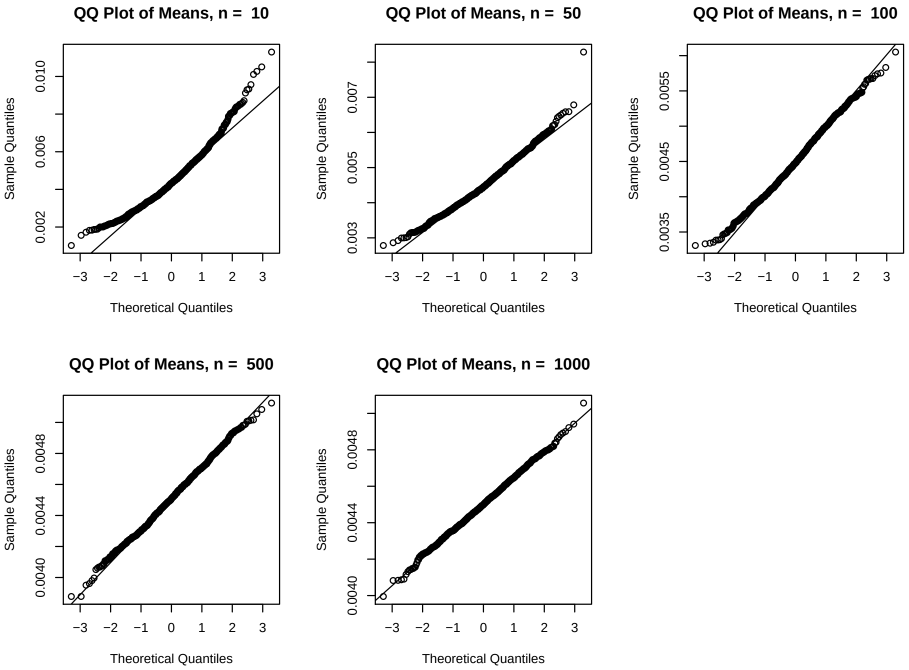


Figure 9: c

Figure 10: QQ Plots of $\bar{Y}_{i,k}$ for a randomly chosen k, under unequal means and unequal variances, with 10% zero-inflation probability.

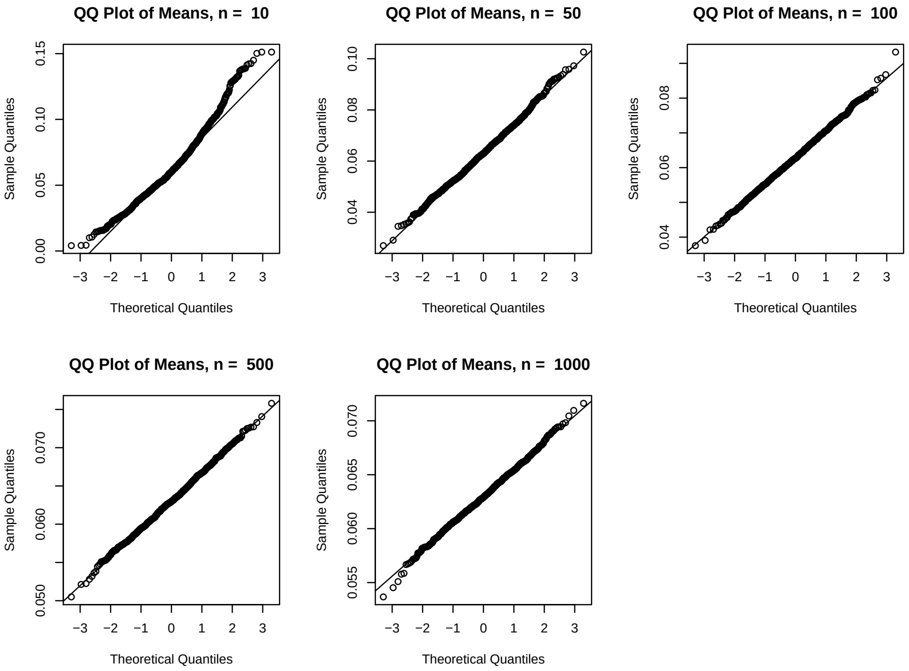


Figure 11: QQ Plots of $\bar{Y}_{i,k}$ for a randomly chosen k, under equal means and variances, with 50% zero-inflation probability.

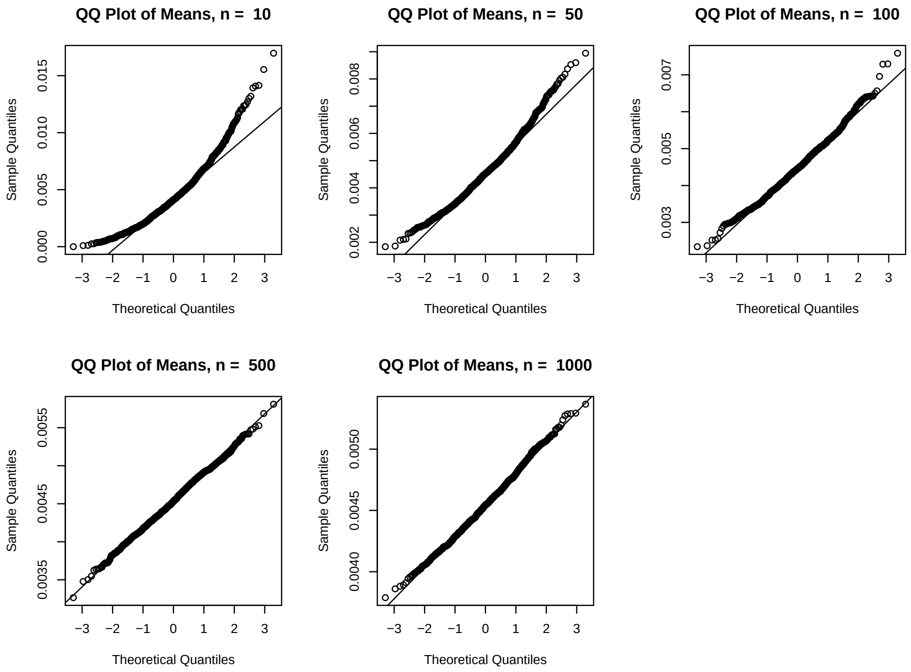


Figure 12: QQ Plots of $\bar{Y}_{i,k}$ for a randomly chosen k, under unequal means and unequal variances, with 50% zero-inflation probability.

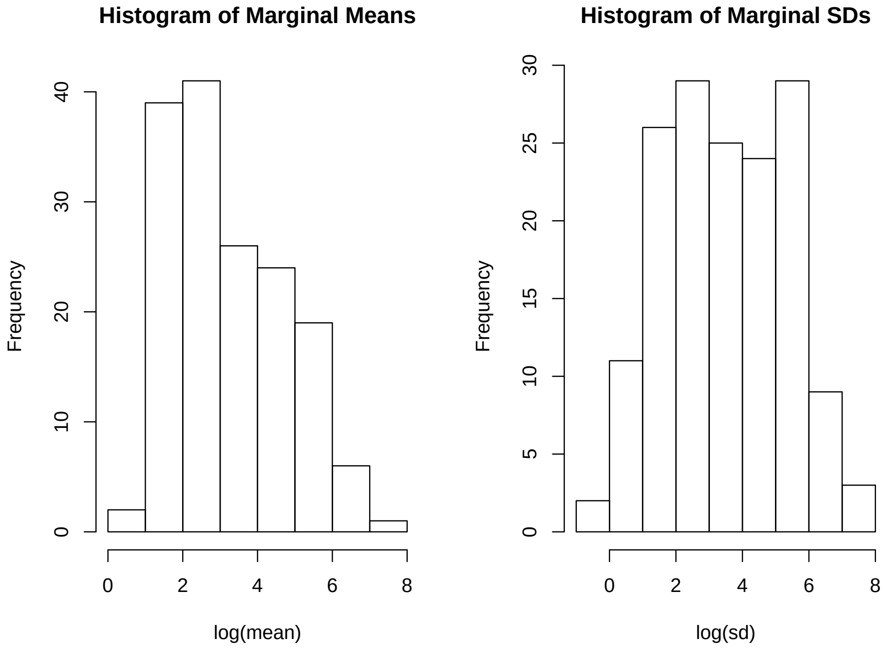


Figure 13: Histogram of marginal means (left) and standard deviations (right) for counts in the IBD study.