

# ANOVA Comparison Simulations: Skewed, Poisson Data

*Kellie Ottoboni*

*2016-12-07*

## Poisson data

As in the other simulations that use Gaussian data, we assume the following linear data-generating process:

$$\tilde{Y}_{ij1} = \beta_0 Y_{ij0} + \beta_j + \gamma_j Z_{ij} + \varepsilon_{ij}$$

for individuals  $i = 1, \dots, n_j$ ,  $j = 1, \dots, J$ .  $\beta_0$  is the coefficient for the baseline measurement  $Y_{i0}$ ,  $\beta_j$  is the mean effect of being at site  $j$ ,  $Z_{ij}$  is the treatment level indicator,,  $\gamma_j$  is the effect of treatment at site  $j$ , and  $\varepsilon_{ij}$  takes the values 0 or 0.5 with equal probability. We will assume that  $\beta_0 = 1$ .

Now, instead of  $Y_{ij0}$  being normally distributed, we let these baseline measurements come from a truncated Poisson distribution with mean 4, with values greater than 10 censored and set to 10. This is reflective of some survey data where individuals are asked to rate their experience on a scale from 1 to 10. The data are not only discrete but also skewed: more values are in the lower half of the range.

We don't observe  $\tilde{Y}_{ij1}$  but  $Y_{ij1}$ , which is defined as

$$Y_{ij1} = \begin{cases} 1 & \text{if } \tilde{Y}_{ij1} < 1 \\ 10 & \text{if } \tilde{Y}_{ij1} > 10 \\ \lfloor \tilde{Y}_{ij1} \rfloor & \text{otherwise} \end{cases}$$

Suppose there are three sites with  $\beta_1 = 1$ ,  $\beta_2 = 1.5$ , and  $\beta_3 = 2$ . Assume that there are 16 individuals per site and treatment assignment is balanced, i.e. 8 people receive each treatment at each site.

We use two designs:

- Design 1:  $\gamma_1 = \gamma_2 = \gamma_3 = 1$ . This is the standard assumption of a constant, additive treatment effect.
- Design 2:  $\gamma_1 = \gamma > 0$ ,  $\gamma_2 = \gamma_3 = 0$ . This is a constant, additive treatment effect at site 1, but no treatment effect at sites 2 and 3. This is a simplistic case of heterogeneous treatment effects.

With these two designs, we vary the distribution of  $\varepsilon$ . In the first case, we use  $\varepsilon \sim N(0, 1)$  to mimic the usual ANOVA assumptions. In the second case,  $\varepsilon \sim t(2)$  so the errors are heavy-tailed. Thus, there are four total simulation designs.

We compare five tests:

- ANOVA: we fit a linear model of response  $Y_1$  on baseline  $Y_0$ , treatment  $Z$ , and a dummy for site.
- Stratified permutation: we permute treatment assignment within site, then take the difference in means between treated and control outcomes  $Y_1$
- Differenced permutation: we do the same permutation procedure as the stratified permutation test, except we use the difference between outcome and baseline,  $Y_1 - Y_0$
- Linear model (LM) permutation: we use the same stratified permutation procedure as above, except use the  $t$ -statistic for the coefficient on treatment in the linear regression of  $Y_1$  on  $Y_0$ ,  $Z$ , and site dummies
- Freedman-Lane test: see the other Rmd document for a full description of this procedure

Throughout our simulations, we first fix  $Y_0$  and site ID. Treatment  $Z$  and the errors  $\varepsilon$  are randomly drawn according to their respective distributions. Then,  $Y_1$  is constructed using the linear data-generating process above. We regenerate  $Z$ ,  $\varepsilon$ , and  $Y_1$  100 times for each design, then compute the empirical power of the five tests.

## Data-generation, tests, and plotting functions

```
generate_outcome <- function(baseline, tr_effect, site_effect,
  error) {
  Y1 <- baseline + tr_effect + site_effect + error
  Y1 <- floor(Y1)
  Y1 <- pmax(pmin(Y1, 10), 1)
  return(Y1)
}

generate_simulated_data <- function(gamma, effect, n = c(16,
  16, 16)) {
  # Input: gamma = the magnitude of the treatment effect effect
  # = 'same effect' or 'single site effect' - which sites have
  # a tr effect > 0? n = number of individuals at each site
  # Returns: a dataframe containing columns named Y1
  # (response), Y0 (baseline), Z (treatment), gamma_vec
  # (treatment effect per individual), SITEID (stratum),
  # site_effect (beta coefficient per individual), and epsilon
  # (errors)

  SITEID <- rep(1:3, times = n)
  N <- sum(n)
  beta <- c(1, 1.5, 2)

  # What is the treatment effect?
  if (effect == "same effect") {
    gamma_vec <- rep(gamma, N)
  } else {
    gamma_vec <- rep(c(gamma, 0, 0), times = n)
  }

  # Generate covariates
  epsilon <- rbinom(N, 1, 0.5) - 0.5
  Y0 <- rpois(N, lambda = 4)
  Y0[Y0 > 10] <- 10
  Z <- rep(0:1, length.out = N)
  site_effect <- rep(beta, times = n)
  Y1 <- generate_outcome(Y0, gamma_vec * Z, site_effect, epsilon)
  return(data.frame(Y1, Y0, Z, gamma_vec, SITEID, site_effect,
    epsilon))
}

generate_simulated_pvalues <- function(dataset, reps = 1000) {
  # Inputs: dataset = a dataframe containing columns named Y1
  # (response), Y0 (baseline), Z (treatment), and SITEID

```

```

# (stratum) Returns: a vector of p-values first element is
# the p-value from the ANOVA second element is the p-value
# from the stratified two-sample permutation test third
# element is the p-value from the linear model test,
# permuting treatment fourth element is the p-value from the
# Freedman-Lane linear model test, permuting residuals

# ANOVA
modelfit <- lm(Y1 ~ Y0 + Z + factor(SITEID), data = dataset)
resanova <- summary(aov(modelfit))
anova_pvalue <- resanova[[1]][ "Z", "Pr(>F)"]

# Stratified permutation test of Y1
observed_diff_means <- mean(dataset$Y1[dataset$Z == 1]) -
  mean(dataset$Y1[dataset$Z == 0])
diff_means_distr <- stratified_two_sample(group = dataset$Z,
  response = dataset$Y1, stratum = dataset$SITEID, reps = reps)
perm_pvalue <- t2p(observed_diff_means, diff_means_distr,
  alternative = "two-sided")

# Diffed permutation test of Y1-Y0
dataset$diff <- dataset$Y1 - dataset$Y0
observed_diff_means2 <- mean(dataset$diff[dataset$Z == 1]) -
  mean(dataset$diff[dataset$Z == 0])
diff_means_distr2 <- stratified_two_sample(group = dataset$Z,
  response = dataset$diff, stratum = dataset$SITEID, reps = reps)
perm_pvalue2 <- t2p(observed_diff_means2, diff_means_distr2,
  alternative = "two-sided")

# Permutation of treatment in linear model
observed_t1 <- summary(modelfit)[["coefficients"]][ "Z", "t value"]
lm1_t_distr <- replicate(reps, {
  dataset$Z_perm <- permute_within_groups(dataset$Z, dataset$SITEID)
  lm1_perm <- lm(Y1 ~ Y0 + Z_perm + factor(SITEID), data = dataset)
  summary(lm1_perm)[["coefficients"]][ "Z_perm", "t value"]
})
lm_pvalue <- t2p(observed_t1, lm1_t_distr, alternative = "two-sided")

# Freedman-Lane linear model residual permutation
lm2_no_tr <- lm(Y1 ~ Y0 + factor(SITEID), data = dataset)
lm2_resid <- residuals(lm2_no_tr)
lm2_yhat <- fitted(lm2_no_tr)
lm2_t_distr <- replicate(reps, {
  lm2_resid_perm <- permute_within_groups(lm2_resid, dataset$SITEID)
  dataset$response_fl <- lm2_yhat + lm2_resid_perm
  lm2_perm <- lm(response_fl ~ Y0 + Z + factor(SITEID),
    data = dataset)
  summary(lm2_perm)[["coefficients"]][ "Z", "t value"]
})
fl_pvalue <- t2p(observed_t1, lm2_t_distr, alternative = "two-sided")

return(c(ANOVA = anova_pvalue, `Stratified Permutation` = perm_pvalue,
  `Differenced Permutation` = perm_pvalue2, `LM Permutation` = lm_pvalue,

```

```

    `Freedman-Lane` = fl_pvalue))
}

compute_power <- function(pvalues) {
  sapply((0:99)/100, function(p) mean(pvalues <= p, na.rm = TRUE))
}

plot_power_curves <- function(power_mat, title) {
  melt(power_mat) %>% mutate(pvalue = Var1/100) %>% mutate(Method = Var2) %>%
    ggplot(aes_string(x = "pvalue", y = "value", color = "Method")) +
    geom_line() + geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
    xlab("P-value") + ylab("Power") + ggtitle(title)
}

plot_pvalue_hist <- function(pvalue_mat, title) {
  melt(pvalue_mat) %>% mutate(Method = Var2) %>% ggplot(aes(x = value,
    fill = Method)) + geom_histogram() + facet_wrap(~Method) +
    ggtitle(title)
}

plot_pvalue_scatter <- function(pvalue_mat, title) {
  pvalue_mat %>% as.data.frame() %>% select(ANOVA, strat = starts_with("Stratified")) %>%
    ggplot(aes(x = ANOVA, y = strat)) + geom_point() + xlim(0,
    1) + ylim(0, 1) + ylab("Stratified Permutation") + geom_abline(intercept = 0,
    slope = 1, linetype = "dashed") + ggtitle(title)
}

```

## Test level: simulation under the null

Before testing for different kinds of effects, we begin checking that the tests have the correct level. We follow the procedure described above, using an effect size of  $\gamma = 0$  at all sites and using standard normal errors. To have the correct level means that the test rate of rejection at level  $\alpha$  is  $\alpha 100\%$ . In other words, the p-values are uniformly distributed and the power curve should coincide with the line with slope 1 through the origin. Figure ?? demonstrates that this is the case. If anything, the differenced stratified permutation test has fewer than  $\alpha 100\%$  false positives when using level  $\alpha$ .

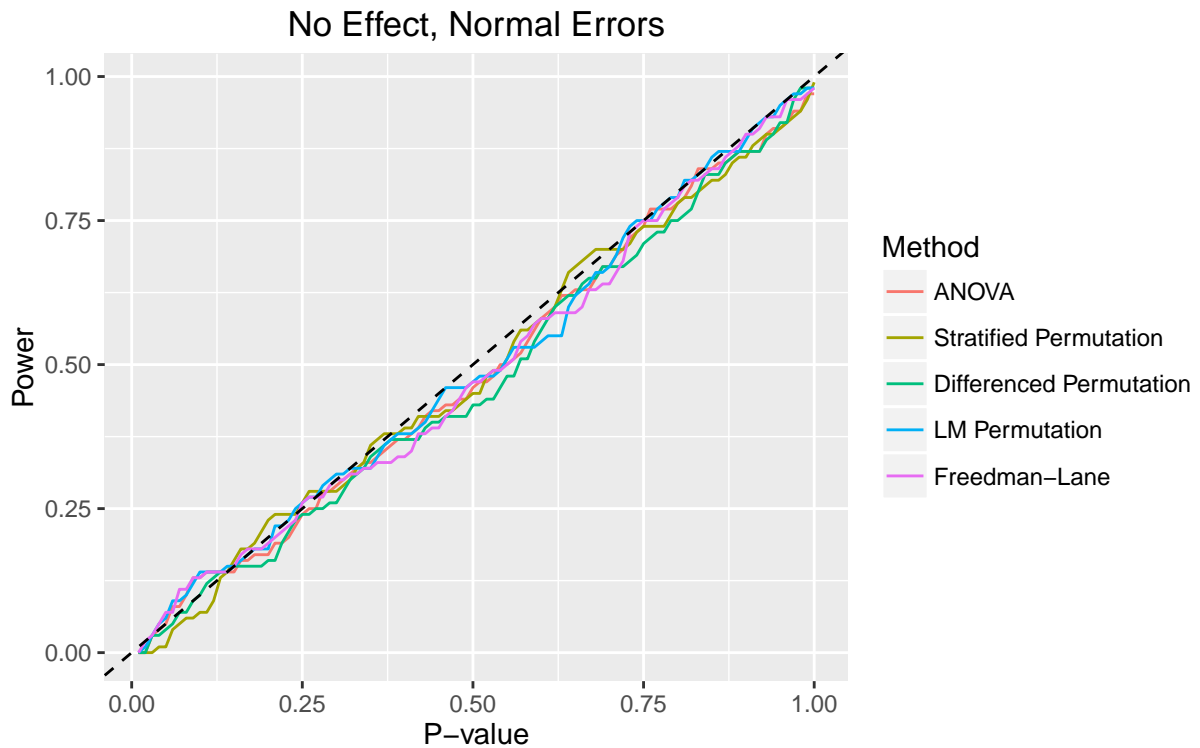
```

set.seed(760682460) # Generated from random.org Timestamp: 2016-11-14 10:21:12 UTC

tmp <- generate_simulated_data(gamma = 0, effect = "same effect")
design0_pvalues <- replicate(100, {
  tmp$epsilon <- rnorm(nrow(tmp))
  tmp$Z <- permute_within_groups(tmp$Z, tmp$SITEID)
  tmp$Y1 <- generate_outcome(tmp$Y0, rep(0, nrow(tmp)), tmp$site_effect)
  generate_simulated_pvalues(tmp)
})
design0_pvalues <- t(design0_pvalues)
colnames(design0_pvalues) <- c("ANOVA", "Stratified Permutation",
  "Differenced Permutation", "LM Permutation", "Freedman-Lane")
design0_power <- apply(design0_pvalues, 2, compute_power)

```

```
plot_power_curves(design0_power, "No Effect, Normal Errors")
```



```
plot_pvalue_hist(design0_pvalues, "No Effect, Normal Errors")
```



## Design 1: Constant additive effect

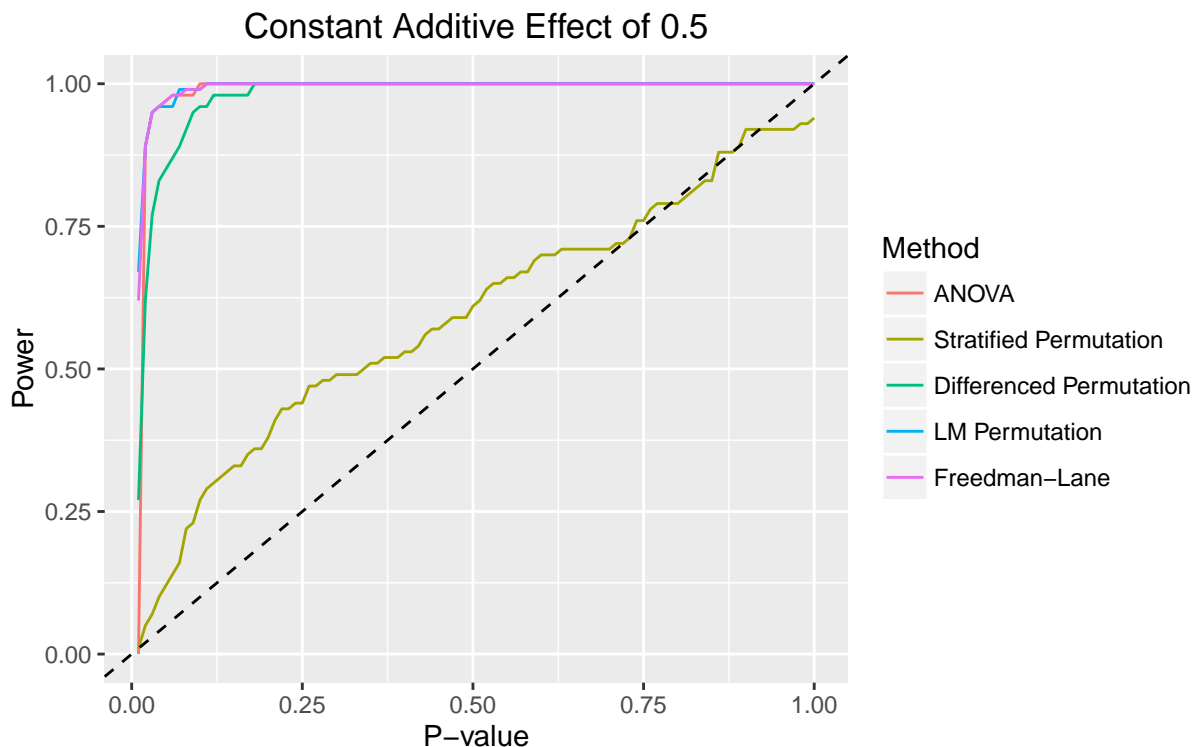
There is no discernable difference in power between the ANOVA, the differenced stratified permutation test, the LM permutations, or the Freedman-Lane test. However, the simple stratified permutation test of  $Y_1$  has substantially less power than the other four. Without controlling for the baseline values, the variance in  $Y_1$  masks the treatment effect.

```
set.seed(760682460) # Generated from random.org Timestamp: 2016-11-14 10:21:12 UTC

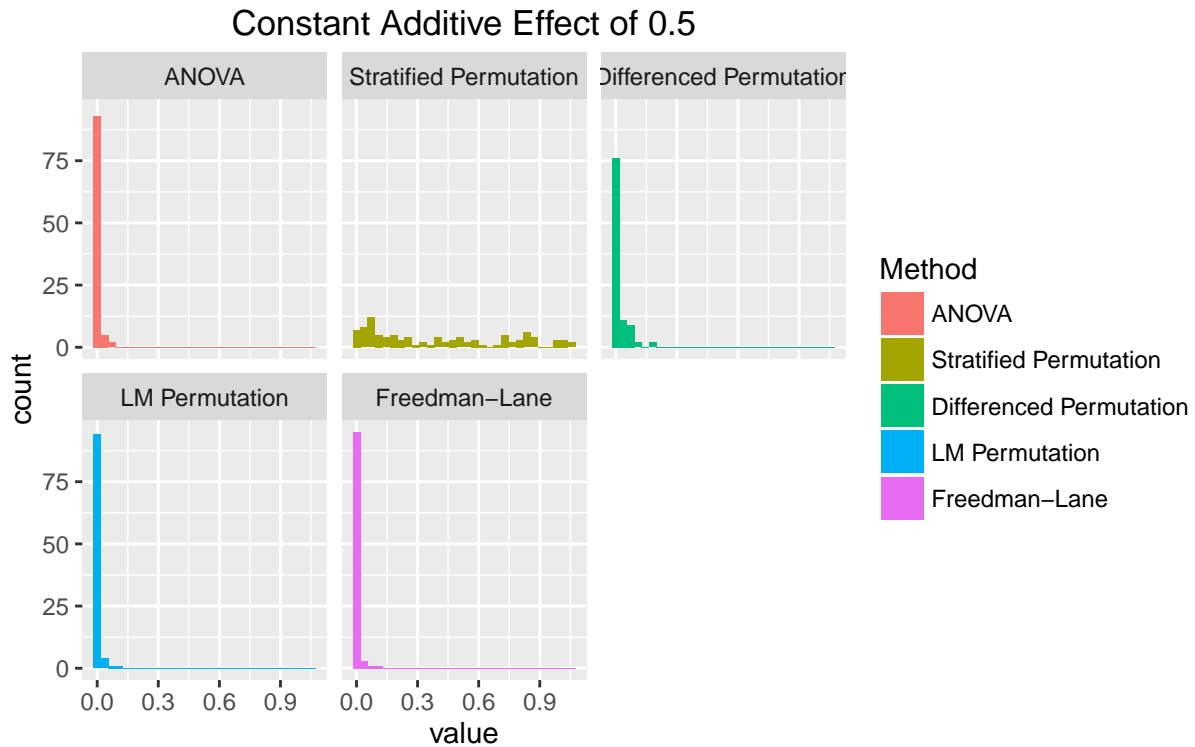
tmp <- generate_simulated_data(gamma = 0.5, effect = "same effect")
design1_pvalues <- replicate(100, {
  tmp$Z <- permute_within_groups(tmp$Z, tmp$SITEID)
  tmp$epsilon <- rbinom(nrow(tmp), 1, 0.5) - 0.5
  tmp$Y1 <- generate_outcome(tmp$Y0, tmp$Z * tmp$gamma_vec,
    tmp$site_effect, tmp$epsilon)
  generate_simulated_pvalues(tmp)
})

design1_pvalues <- t(design1_pvalues)
colnames(design1_pvalues) <- c("ANOVA", "Stratified Permutation",
  "Differenced Permutation", "LM Permutation", "Freedman-Lane")
design1_power <- apply(design1_pvalues, 2, compute_power)

p1 <- plot_power_curves(design1_power, "Constant Additive Effect of 0.5")
p1
```



```
plot_pvalue_hist(design1_pvalues, "Constant Additive Effect of 0.5")
```



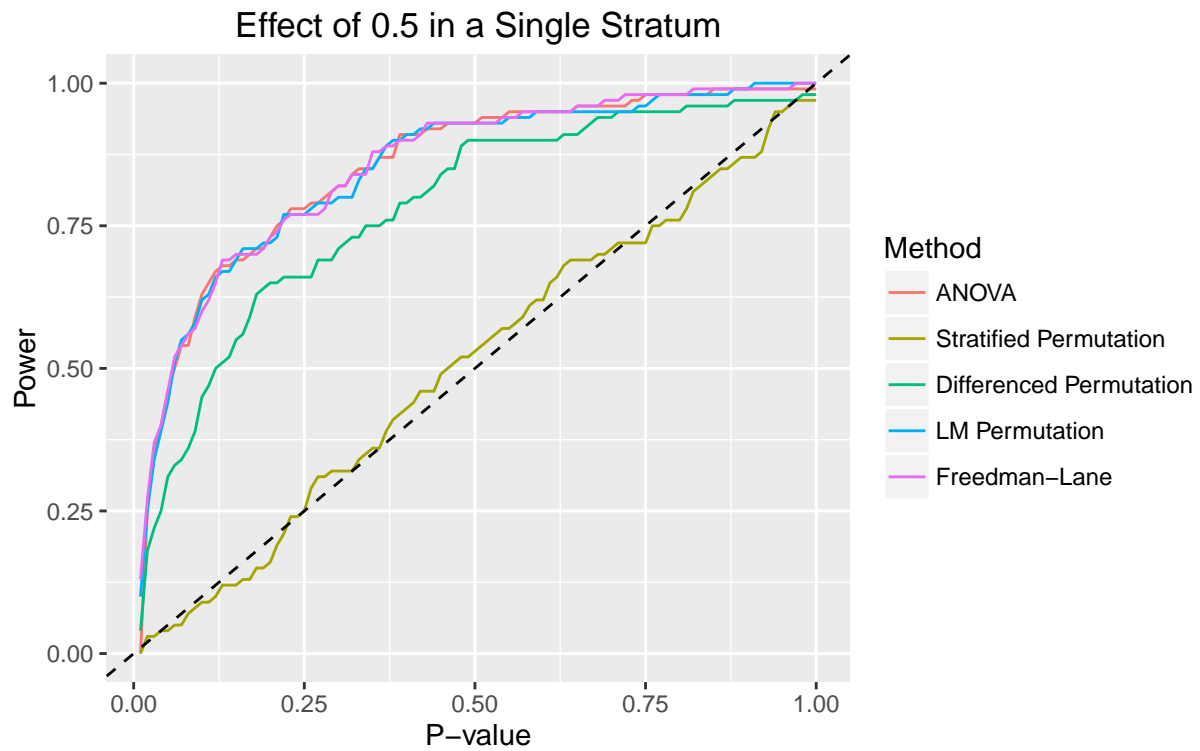
## Design 2: Single site effect

A similar pattern emerges: the simple stratified permutation test has low power, while the other four power curves roughly coincide. The Freedman-Lane test may have the highest power for small p-values, but this could also just be noise. All five power curves are closer to the line passing through the origin: since the effect is only present at one site, it is more difficult to detect.

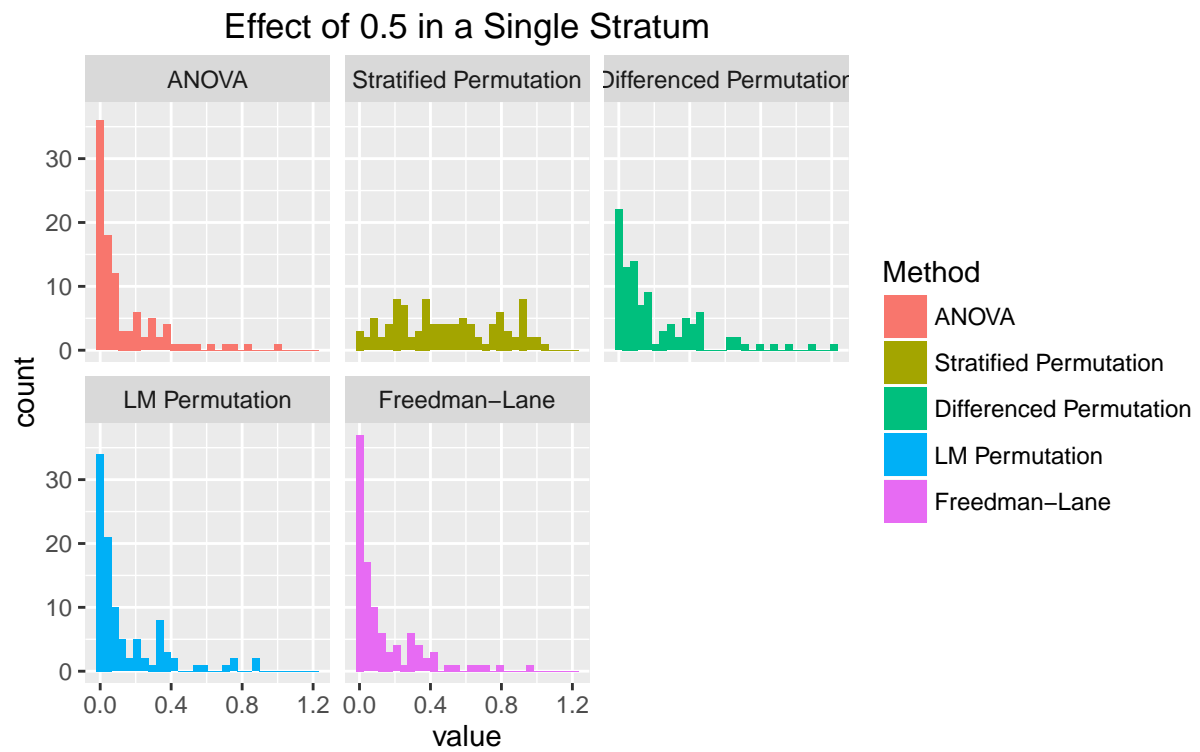
```
set.seed(760682460) # Generated from random.org Timestamp: 2016-11-14 10:21:12 UTC

tmp <- generate_simulated_data(gamma = 0.5, effect = "single site effect")
design2_pvalues <- replicate(100, {
  tmp$Z <- permute_within_groups(tmp$Z, tmp$SITEID)
  tmp$epsilon <- rbinom(nrow(tmp), 1, 0.5) - 0.5
  tmp$Y1 <- generate_outcome(tmp$Y0, tmp$Z * tmp$gamma_vec,
    tmp$site_effect, tmp$epsilon)
  generate_simulated_pvalues(tmp)
})
design2_pvalues <- t(design2_pvalues)
colnames(design2_pvalues) <- c("ANOVA", "Stratified Permutation",
  "Differenced Permutation", "LM Permutation", "Freedman-Lane")
design2_power <- apply(design2_pvalues, 2, compute_power)

p2 <- plot_power_curves(design2_power, "Effect of 0.5 in a Single Stratum")
p2
```



```
plot_pvalue_hist(design2_pvalues, "Effect of 0.5 in a Single Stratum")
```





### Design 3: Unpredictive baseline

Suppose that the baseline measure is not strongly predictive of the outcome. Then, we'd expect that controlling for baseline will not improve the power of the test, and may even make it worse. Suppose that in the linear data-generating process above, we have  $\beta_0 = 0.25$ , meaning that  $Y_1$  and  $Y_0$  have a correlation of 0.25.

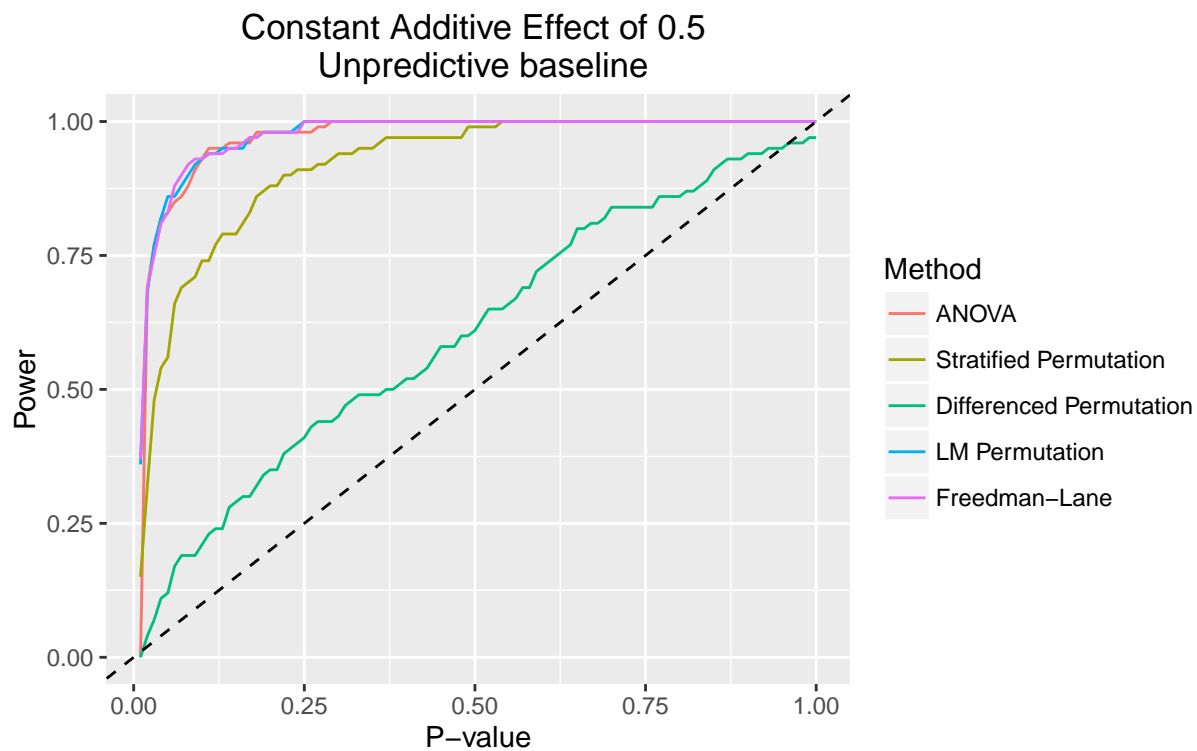
The power curves for this design look similar to those from Design 1 with one important difference: the stratified permutation test has power very close to ANOVA and the two linear model tests, while the differenced permutation test has lower power. This suggests that one should be careful when incorporating control variables; naively taking the difference  $Y_1 - Y_0$  does not capture the correct relationship between baseline and outcome.

```
set.seed(760682460) # Generated from random.org Timestamp: 2016-11-14 10:21:12 UTC

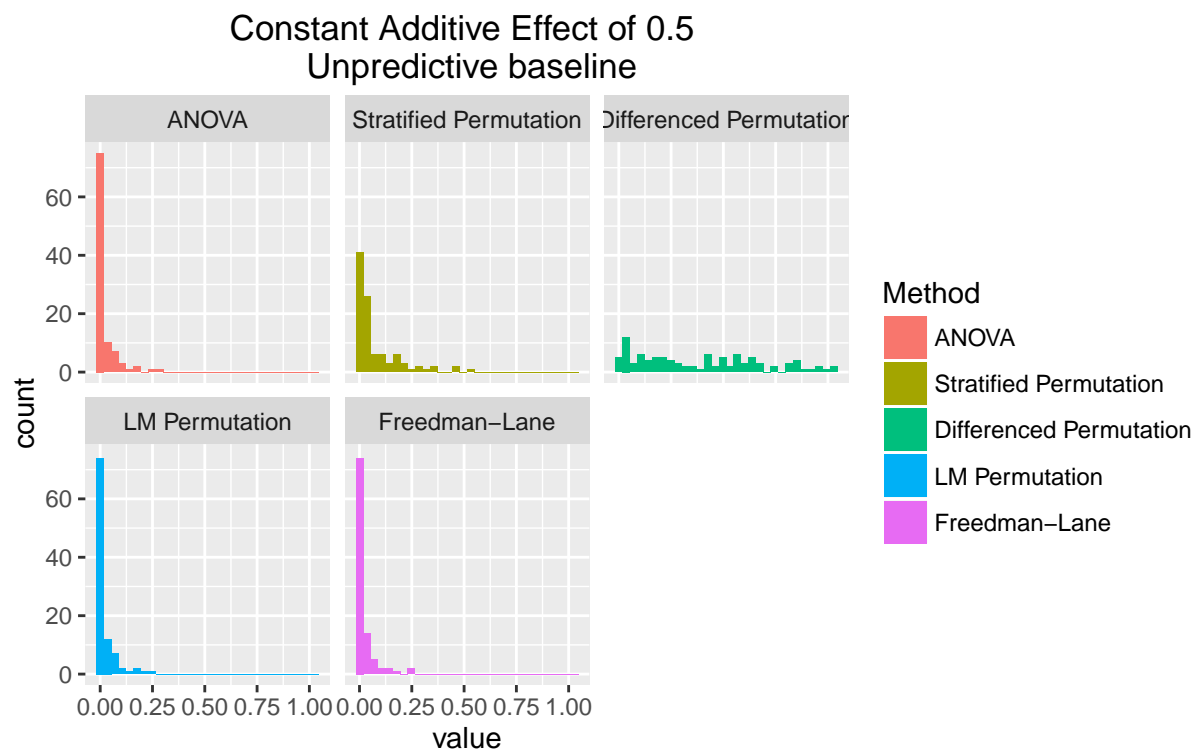
tmp <- generate_simulated_data(gamma = 0.5, effect = "same effect")
design3_pvalues <- replicate(100, {
  tmp$Z <- permute_within_groups(tmp$Z, tmp$SITEID)
  tmp$epsilon <- rbinom(nrow(tmp), 1, 0.5) - 0.5
  tmp$Y1 <- generate_outcome(0.1 * tmp$Y0, tmp$Z * tmp$gamma_vec,
    tmp$site_effect, tmp$epsilon)
  generate_simulated_pvalues(tmp)
})

design3_pvalues <- t(design3_pvalues)
colnames(design3_pvalues) <- c("ANOVA", "Stratified Permutation",
  "Differenced Permutation", "LM Permutation", "Freedman-Lane")
design3_power <- apply(design3_pvalues, 2, compute_power)

p3 <- plot_power_curves(design3_power, "Constant Additive Effect of 0.5 \n Unpredictive baseline")
p3
```



```
plot_pvalue_hist(design3_pvalues, "Constant Additive Effect of 0.5 \n Unpredictive baseline")
```



## Design 4: Constant additive effect, imbalanced design

Suppose that instead of having 16 individuals per site, we distribute them unequally across sites: site 1 has 8 patients, site 2 has 16 patients, and site 3 has 24 patients. This imbalance does not violate any assumptions of the ANOVA. However, it may reduce power if the effect is concentrated in sites with fewer patients.

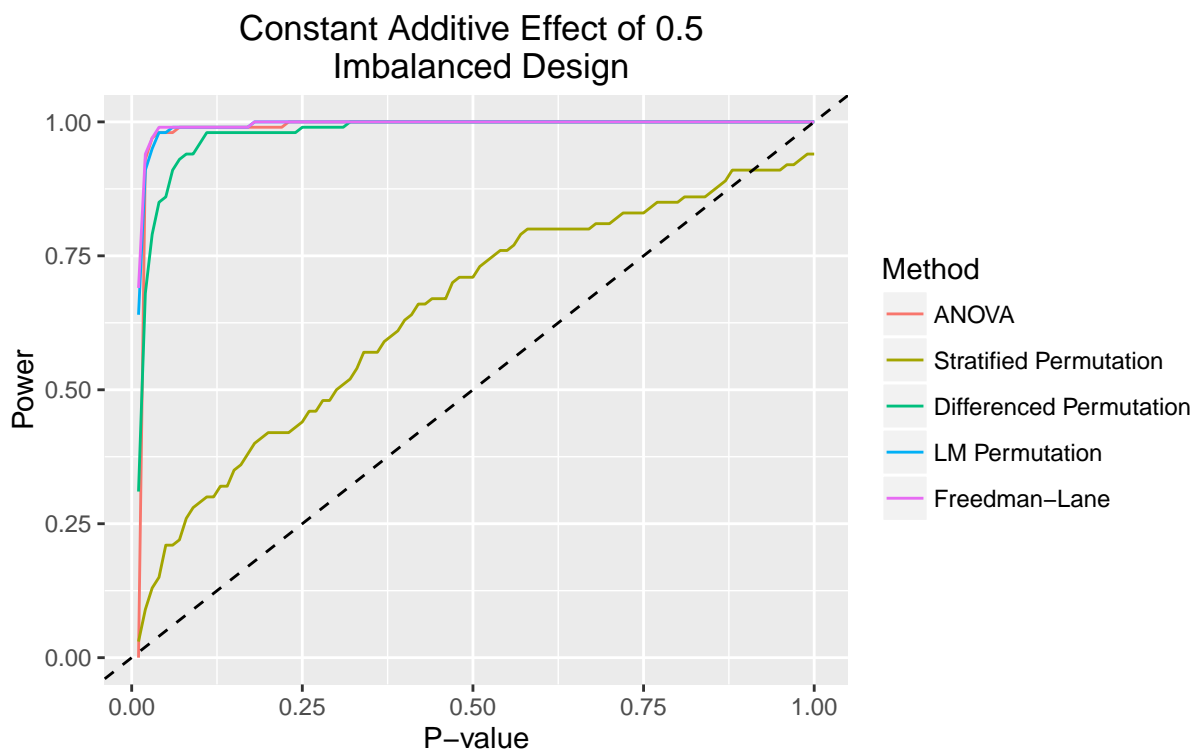
The power curves here are very similar to those in Design 1. This result is expected, since the effect is still present among all patients who received treatment.

```
set.seed(760682460) # Generated from random.org Timestamp: 2016-11-14 10:21:12 UTC

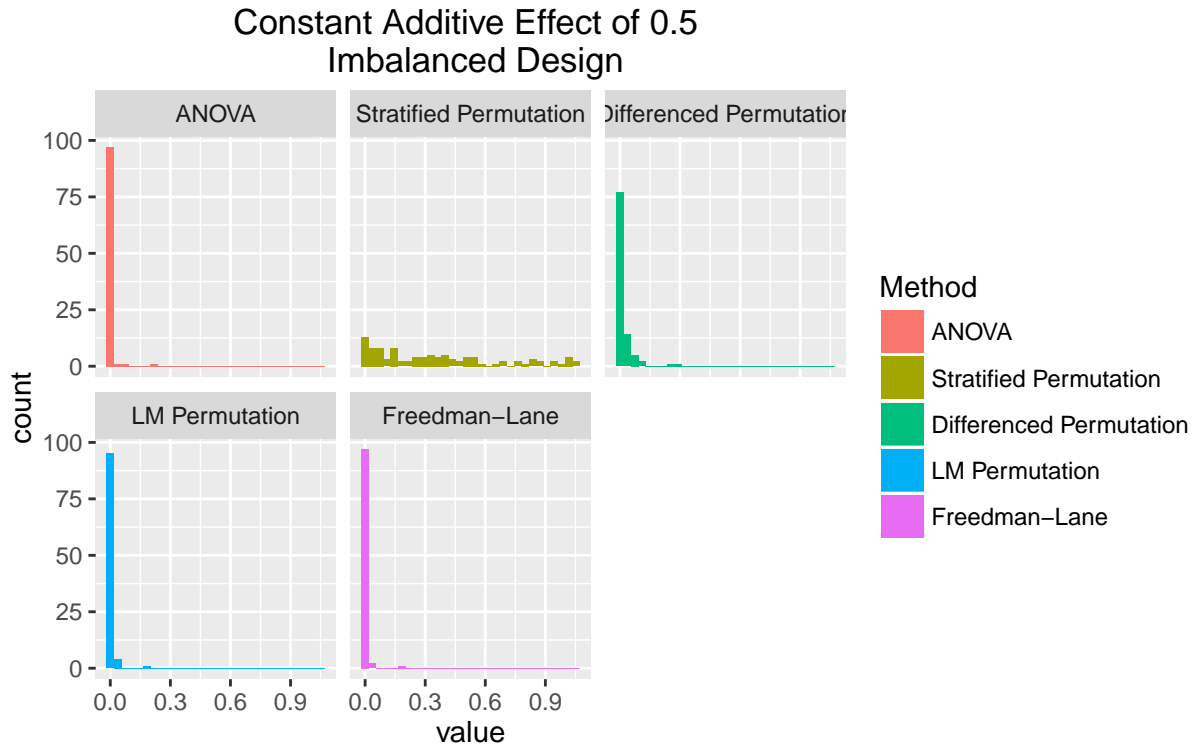
tmp <- generate_simulated_data(gamma = 0.5, effect = "same effect",
  n = c(8, 16, 24))
design4_pvalues <- replicate(100, {
  tmp$Z <- permute_within_groups(tmp$Z, tmp$SITEID)
  tmp$epsilon <- rbinom(nrow(tmp), 1, 0.5) - 0.5
  tmp$Y1 <- generate_outcome(tmp$Y0, tmp$Z * tmp$gamma_vec,
    tmp$site_effect, tmp$epsilon)
  generate_simulated_pvalues(tmp)
})

design4_pvalues <- t(design4_pvalues)
colnames(design4_pvalues) <- c("ANOVA", "Stratified Permutation",
  "Differenced Permutation", "LM Permutation", "Freedman-Lane")
design4_power <- apply(design4_pvalues, 2, compute_power)

p4 <- plot_power_curves(design4_power, "Constant Additive Effect of 0.5 \n Imbalanced Design")
p4
```



```
plot_pvalue_hist(design4_pvalues, "Constant Additive Effect of 0.5 \n Imbalanced Design")
```



## Design 5: Single site effect, imbalanced design

There is very low power to detect the effect here, since in the simulated data, the site with only 8 patients was the only site with a nonzero effect of treatment. The treatment effect is on the same order of magnitude as the error variance.

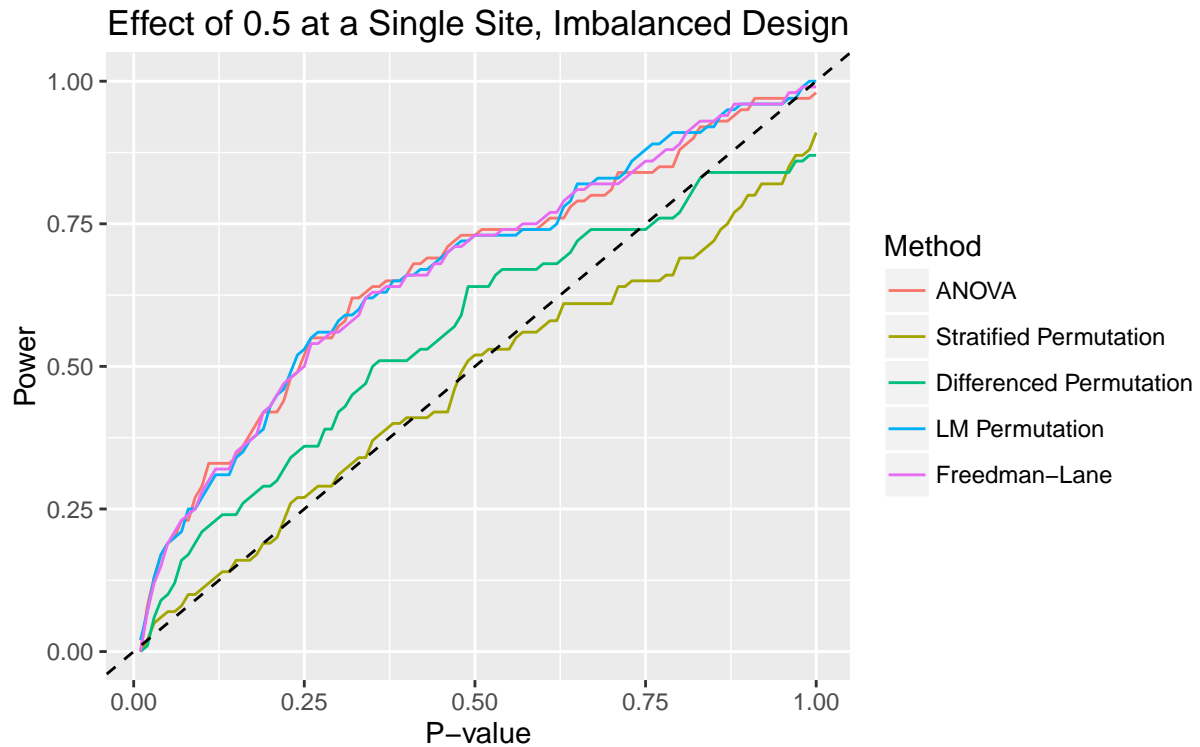
```
set.seed(760682460) # Generated from random.org Timestamp: 2016-11-14 10:21:12 UTC
```

```
tmp <- generate_simulated_data(gamma = 0.5, effect = "single site effect",
  n = c(8, 16, 24))
```

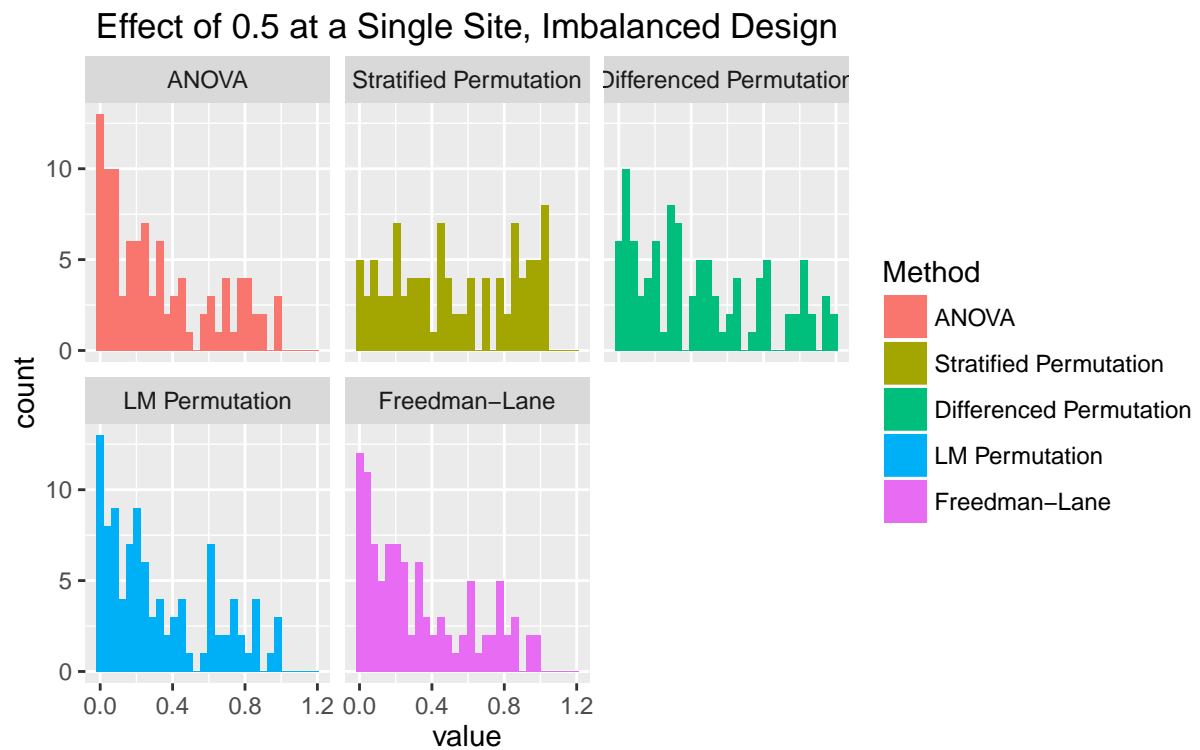
```
design5_pvalues <- replicate(100, {
  tmp$Z <- permute_within_groups(tmp$Z, tmp$SITEID)
  tmp$epsilon <- rbinom(nrow(tmp), 1, 0.5) - 0.5
  tmp$Y1 <- generate_outcome(tmp$Y0, tmp$Z * tmp$gamma_vec,
    tmp$site_effect, tmp$epsilon)
  generate_simulated_pvalues(tmp)
})
```

```
design5_pvalues <- t(design5_pvalues)
colnames(design5_pvalues) <- c("ANOVA", "Stratified Permutation",
  "Differenced Permutation", "LM Permutation", "Freedman-Lane")
design5_power <- apply(design5_pvalues, 2, compute_power)
```

```
p5 <- plot_power_curves(design5_power, "Effect of 0.5 at a Single Site, Imbalanced Design")
p5
```



```
plot_pvalue_hist(design5_pvalues, "Effect of 0.5 at a Single Site, Imbalanced Design")
```



```
# plots4 <- lapply(list(p1, p2, p4, p5), function(pl) pl +
# theme(legend.position = 'none') + ggtitle(''))
# do.call(grid.arrange, plots4)
```

```

powers <- list(design1_power %>% as.data.frame() %>% mutate(Treatment = rep("Constant Additive Effect",
  nrow(design1_power)), Design = rep("Balanced", nrow(design1_power))),
  design2_power %>% as.data.frame() %>% mutate(Treatment = rep("Single Site Effect",
  nrow(design2_power)), Design = rep("Balanced", nrow(design2_power))),
  design4_power %>% as.data.frame() %>% mutate(Treatment = rep("Constant Additive Effect",
  nrow(design4_power)), Design = rep("Imbalanced", nrow(design4_power))),
  design5_power %>% as.data.frame() %>% mutate(Treatment = rep("Single Site Effect",
  nrow(design5_power)), Design = rep("Imbalanced", nrow(design5_power))))

all_power_curves <- do.call(rbind, powers)
twobytwo <- melt(all_power_curves, id.vars = c("Treatment", "Design")) %>%
  mutate(pvalue = rep((1:100)/100, 5 * 4)) %>% mutate(Method = variable) %>%
  ggplot(aes_string(x = "pvalue", y = "value", color = "Method")) +
  geom_line() + geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
  xlab("P-value") + ylab("Power") + facet_wrap(Treatment ~
  Design) + ggtitle("Discrete, Skewed Data")

twobytwo

```

