Unicare study power and sample size notes

May 2024

Generally we seek to detect 20% effect at 80% power, 5% significance.

We consider treatment effect

- proportion admitted;
- mean number of admissions per patient year.
- coefficient of treatment indicator variable in for logistic, Poisson, and negative binomial regression
- zero-inflated Poisson and Negative binomial models; Bayes models
- simulations of the above using values from existing Unicare and Ibis member data.
- effect of controlling for other covariates, which tend to increase efficiency; that is decrease min sample size.

All estimates assume equal number of treatment and control. Sample sizes are for the total in both arms.

Summary

Most of the analysis points to sample sizes of larger than 1000 for 20% effect, 80% power, 5% significance. The one exception is test for Poisson mean, which is 423 for 15% effect, 200 for 22% effect. But there this does not agree with the minimum sample for the coefficient for treatment in the Poisson model, which should be an almost equivalent approach. The simulations seem to give much more conservative (higher) estimates. The effect of increased efficiency seems to have been minimal. But note that these simulations were of zero inflated data, which does not share the property of the Poisson model that the variance and the mean are equal. There might be something else going on here as well. Also regarding the simulations, coefficient values were obtained by trial and error to produce the sample statistics in the Unicare data, as the raw data was not available at the time. Additionally, no other covariates were considered. In light of these, efficiency gains from including all the covariates is likely to be greater than that seen in the simulations.

Proportion admitted

Using normal approximation to binomial and assuming equal number in each group, standard sample size formulas give, for values in line with the Unicare/Ibis data:

base_rate	treat_rate	percent_reduce	odds_ratio	two_side_n	one_side_n
0.25	0.20	0.20	0.7500000	2187.478	1722.84
0.25	0.18	0.28	0.6585366	1079.023	849.7098
0.25	0.15	0.40	0.5294118	499.9639	393.5855

Poisson means- number of admissions

Treating the number of admissions per patient-year as a Poisson process with mean λ, the total number of admissions in n patient-years is approximately normal if $n\lambda > 30$. For our data λ is apparently contained in the range 0.30 - 0.45, or equivalently, 300 - 450 admissions per 1000 patient years.

base_mean	treat_mean	percent_reduce	two_side_n	one_side_n
0.45	0.40	0.1111111	850.4695	438.7239

base_mean	treat_mean	percent_reduce	two_side_n	one_side_n
0.45	0.38	0.1555556	423.7033	218.5720
0.45	0.35	0.2222222	200.1105	103.2292

These estimates seem very low and do not agree with simulation done below.

Using logistic regression A near equivalent approach is to test the coefficient of the treatment indicator variable in a logistic regression where $\text{logit}\mu = \beta_0 + \beta_1 \times \text{treat}$ where `treat` is 0 or 1 for control or treated, and the approximation with $\hat{\beta}_1/SE(\hat{\beta}_1) \sim \mathcal{N}(0,1)$. We confirmed the above results. For example, using the `WebPower` package we obtain

```
## Power for logistic regression
##
##       p0  p1     beta0      beta1        n alpha power
##     0.25 0.2 -1.098612 -0.2876821 2197.073  0.05   0.8
##
## URL: http://psychstat.org/logistic
```

Using Poisson regression Test the coefficient of the treatment indicator variable in a Poisson regression; again with the WebPower package we obtain

```
## Power for Poisson regression
##
##            n power alpha exp0 exp1      beta0      beta1
##     1576.297   0.8  0.05 0.45  0.8 -0.7985077 -0.2231436
##
## URL: http://psychstat.org/poisson
```

The values above would correspond to `base_mean` = 0.45 and `treat_mean` = 0.45*0.8 = 0.36[1].

The method above also uses the normal approximation for $\hat{\beta}_1$.

Simulations

We first do a simple simulation to check difference in Poisson means obtained above. For Poisson regression, the log of the mean is typically modeled as a linear function of the covariates.

$$\log \lambda_i = \beta_0 + \beta_{ibis} \times \text{ibis}_i$$

where

- λ_i is the mean number of admissions for patient i,
- ibis_i is the treatment indicator for patient i,
- β_0 is the mean number of admissions for the control group,
- β_{ibis} is the effect of treatment on the mean number of admissions.

It follows that the mean number of admissions is $e^{\beta_0} e^{\beta_{ibis} \times \text{ibis}_i}$ and so the effect of treatment is $e^{\beta_{ibis}}$. Thus

- $\beta_0 = \log 0.45 = -0.799$ corresponds to a base mean of 0.45 for control.
- $\beta_{ibis} = \log 0.8 = -0.223$ corresponds to 20% reduction.

We generate 100000 samples of random Poisson random data, with means λ_i with these parameters where `ibis` is randomly assigned 0 or 1.

Here is what a simulated data set looks like. There are a lot of zeros because of the low value of the mean.

[1]The 'exp()' is because the Poisson model would be $\log \mu = \beta_0 + \beta_1 \times \text{treat}$ where 'treat' is 0 or 1 for control or treated; resp, and so $\mu = e^{\beta_0} e^{\beta_1 \times \text{treat}}$, with e^{β_1} representing the multiplier of base mean for the treated.

y	ibis
0	0
0	1
1	0
1	1
1	0
0	0

We can see we get about a roughly 20% reduction in the mean for `ibis`.

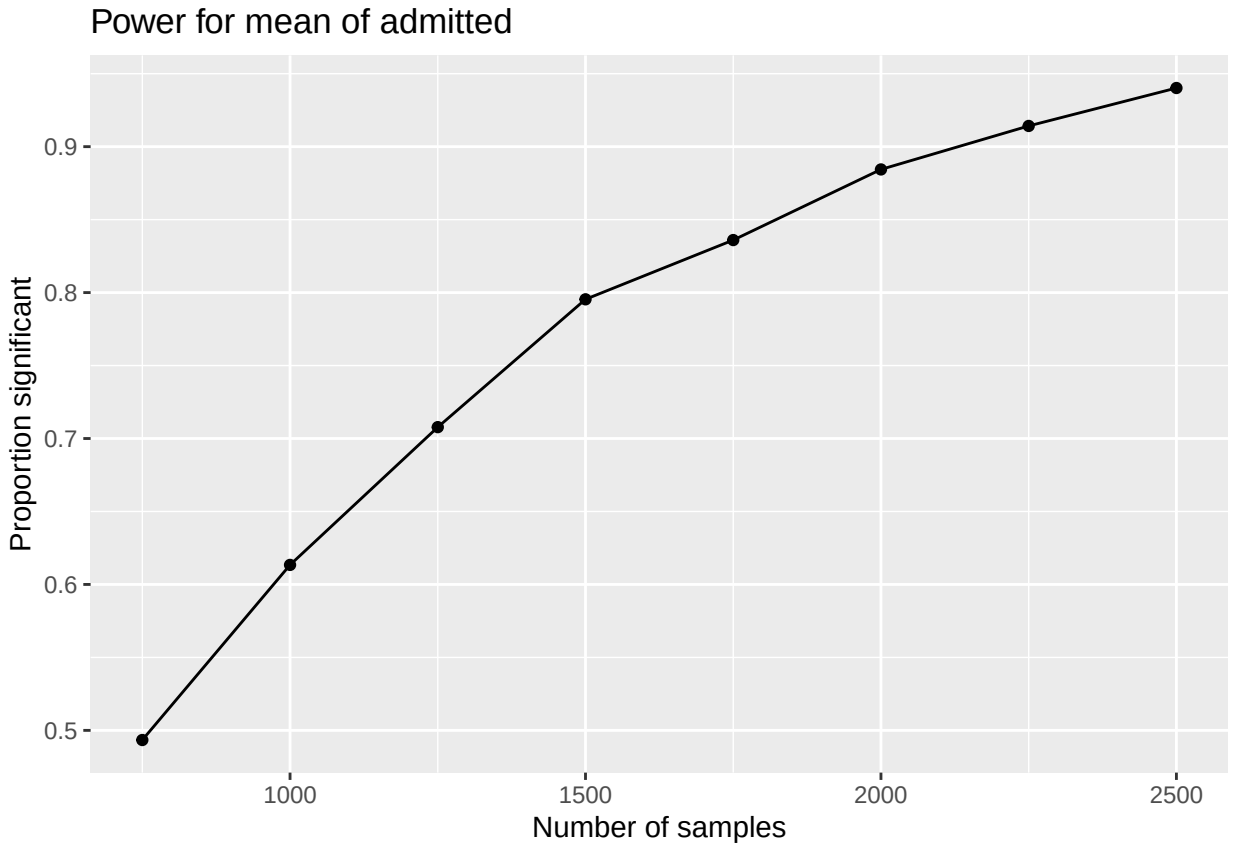
```
## ibis yes 0.3579118
```

```
## ibis no 0.4536204
```

```
## percent change 0.2109882
```

To estimate power, we simulate a given number of samples and test for significance- we use a simple t test- and repeat the process many times, and record the proportion of times the result is significant.

n	prop_sig
750	0.4934
1000	0.6134
1250	0.7078
1500	0.7954
1750	0.8360
2000	0.8844
2250	0.9142
2500	0.9402



So we see the results are not consistent with the power calculations above.

Zero-inflated data

One way to model zero-inflated count data is as a bernouli process with probability θ to model so called "structural" zero counts, and a Poisson or negative binomial process with mean λ to model non-zero counts, and also adding additional zeros. For parameter values for generating simulated data we will use sample Unicare and Ibis data.

Zero inflated Poisson We will use a zero inflated Poisson model[2] We can express the probability of an outcome $y_i = k$ admissions as

$$P(y_i = 0) = \theta_i + (1 - \theta_i)\, P_{pois}(y_i = 0; \lambda_i)$$
$$P(y_i = k \neq 0) = (1 - \theta)\, P_{pois}(y_i = k; \lambda_i)$$

where $P_{pois}(\cdot; \lambda)$ is the Poisson probability with mean λ. A standard approach is to model θ as

$$\text{logit } \theta_i = \alpha_0 + \alpha_{age} \times \text{age}_i + \alpha_{ibis} \times \text{ibis}_i$$

where `ibis` is either 0 or 1, and logit $x = \log \frac{x}{1-x}$. We model λ as (also standard approach)

$$\log \lambda_i = \beta_0 + \beta_{age} \times \text{age}_i + \beta_{ibis} \times \text{ibis}_i$$

One thing to keep in mind is that with the model written this way, increasing logit θ increases the probability of zero admissions- a good thing- while, and increasing $\log \lambda$ increases the mean admissions- a bad thing.

[2]We will likely use a zero inflated negative binomial model for the statistical analysis, as the count data are overdispersed as well as zero inflated. Negative binomial models contain an additional dispersion parameter that can account for this.

We generate 10000 samples simulated data with

- `age` is uniform on [60, 90]
- We scale the age data so that the age coefficients for both the structural and Poisson portions of the model represent approx percent increases per ten years from baseline of 70 years, with decrease for ages less than 70. Similarly for the Poisson mean portion of the model.
- `ibis` is randomly assigned
- `ibis` confers advantage in reducing admissions; both from structural and Poisson components.

We can tune to approximate values from Unicare and Ibis data, with reductions in outcomes for `ibis`. We start with the following coefficients.

- $\alpha_0 = 0.3$, $\alpha_{age} = -0.05$, $\alpha_{ibis} = 0.2$
- $\beta_0 = -0.7$, $\beta_{age} = 0.9$, $\beta_{ibis} = -0.2$

Here is what a simulated data set looks like

y	age	ibis
0	68	0
0	85	1
3	81	1
0	76	0
1	72	1
0	67	1

Proportion of zeros:

```
## ibis yes 0.8001179
```

```
## ibis no 0.7544788
```

```
## percent change 0.06049089
```

On the other hand, the admittance rates are 1 - minus these values:

```
## ibis yes 0.1998821
```

```
## ibis no 0.2455212
```

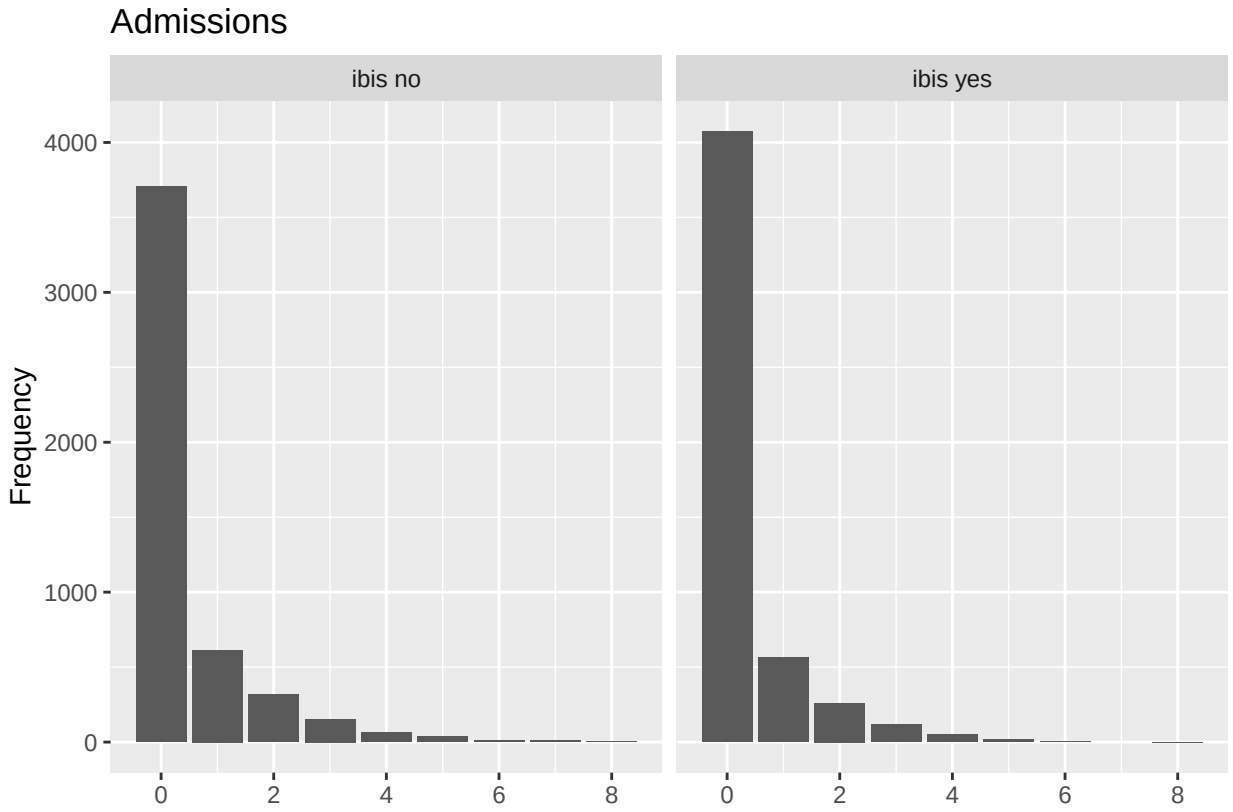
```
## percent change -0.1858866
```

Mean visits

```
## ibis yes 0.3478774
```

```
## ibis no 0.4655945
```

```
## percent change 0.2528318
```

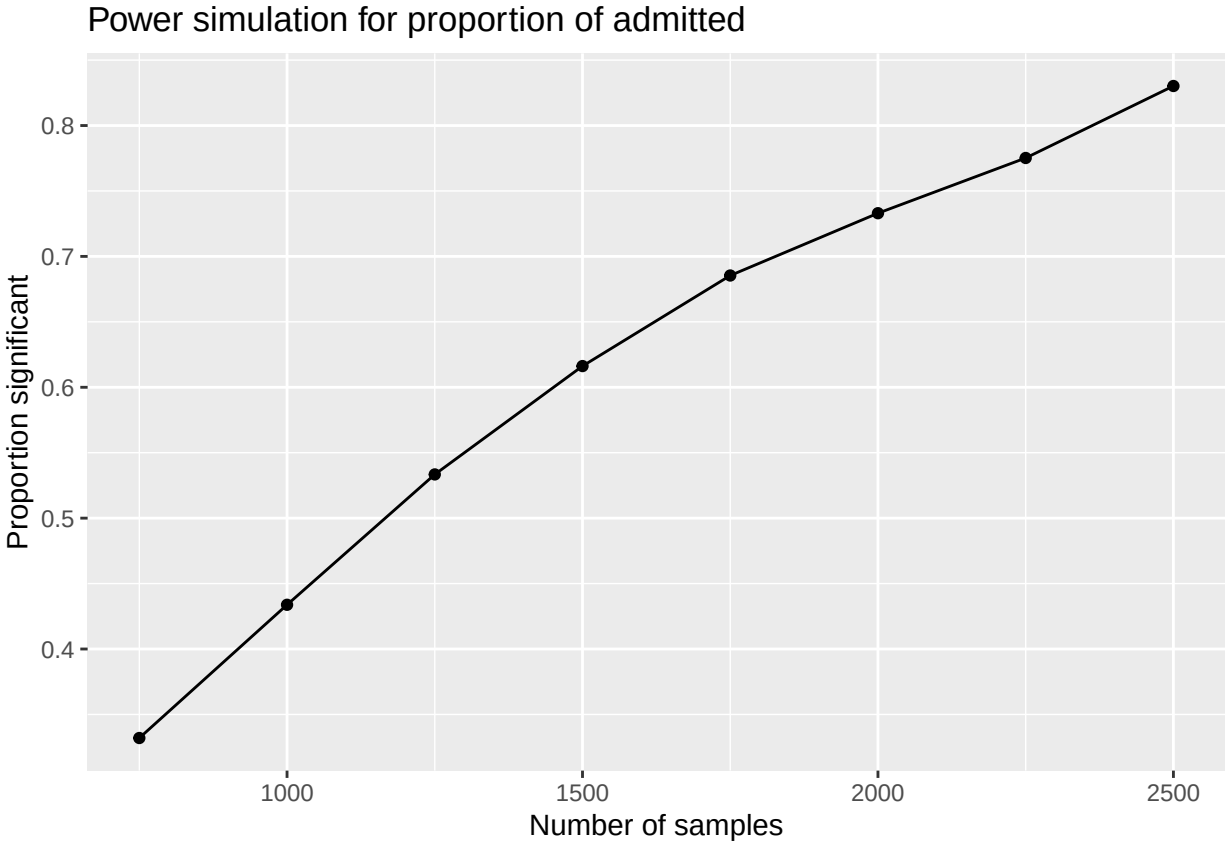


Counts

Power simulation for proportion of admitted

The results below are from 5000 simulations for each sample size.

n	prop_sig
750	0.3320
1000	0.4338
1250	0.5334
1500	0.6162
1750	0.6854
2000	0.7330
2250	0.7752
2500	0.8302



If we wanted to cheat we could aim for 20% increase in zero count rate. Tuning the data generation and we get, for 10000 samples

Proportion of zeros:

```
## ibis yes 0.8758625
```

```
## ibis no 0.7220362
```

```
## percent change 0.2130452
```

But note that the corresponding change in the admittance rate is now

```
## [1] 0.5534041
```

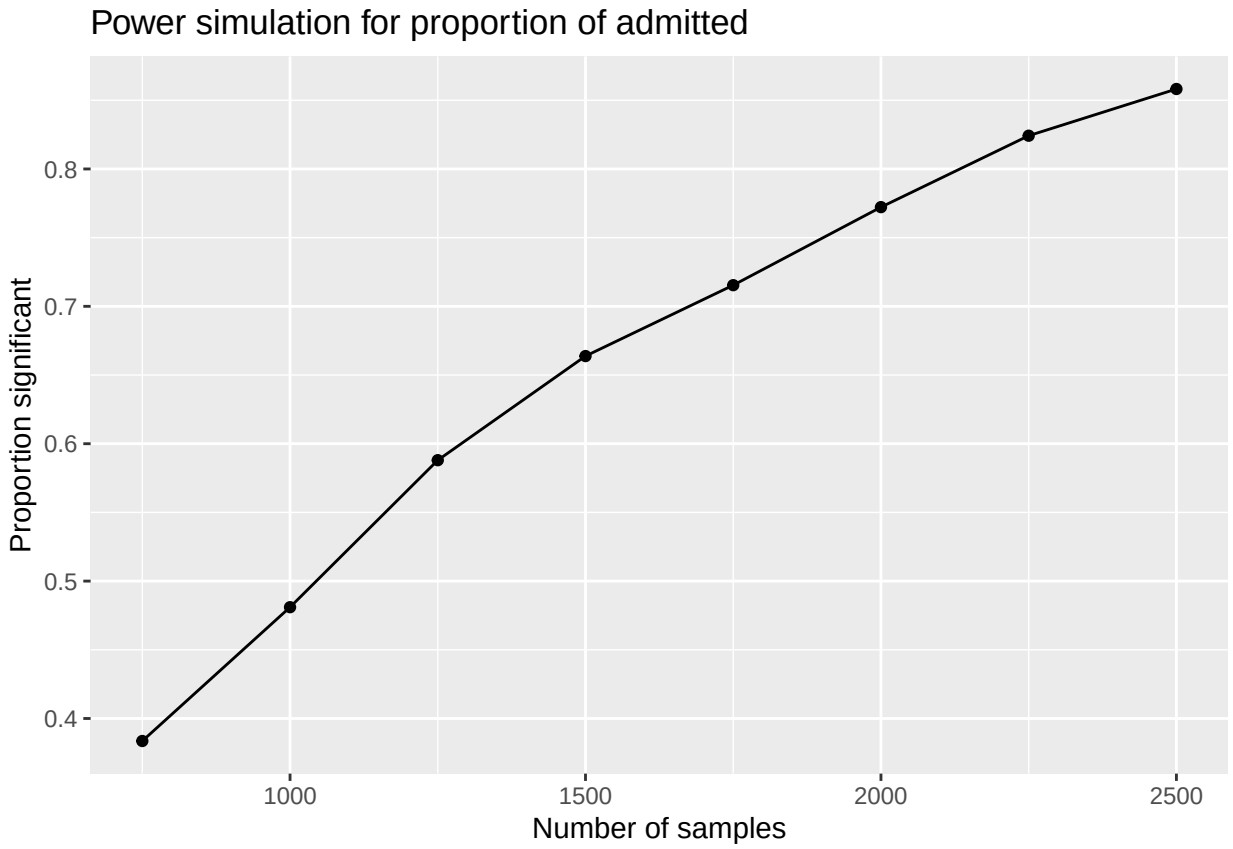
But in any case if we do this and simulate we obtain

n	prop_sig
200	0.7286
300	0.8906
500	0.9908

Test on regression coefficient; include age as covariate The above should be equivalent to testing the coefficient of the treatment indicator variable in a logistic regression with just the `ibis` covariate. But if include covariates such as `age` that are predictive of the outcome, we should see an increase in efficiency; that is, less variance in estimates, which translates to greater power.

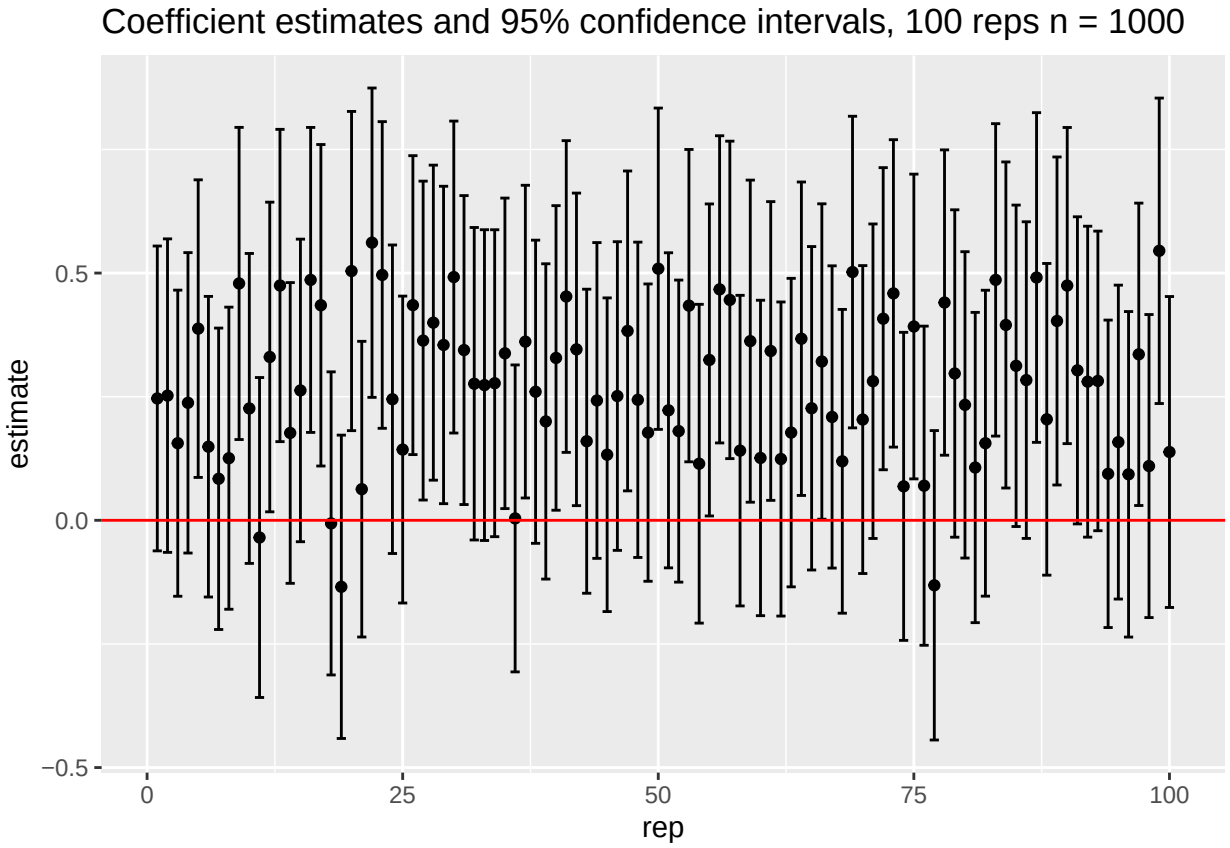
We do the simulations with the original parameter values above. We see only a marginal increase in power with the age covariate.

n	prop_sig
750	0.3836
1000	0.4810
1250	0.5880
1500	0.6638
1750	0.7154
2000	0.7722
2250	0.8242
2500	0.8582



We can see there is a lot of variability in the coefficient estimates, even with n = 1000. For significance we want the confidence interval to not contain zero[3].

[3]Note that the coefficient is positive, corresponding to higher probability of zero



Power simulation for mean number of admissions

We adjust the generating process for effect size approx 20% reduction in mean number of admissions, by decreasing the reduction to the mean due to `ibis`. A sample of 10000 gives the following.

Mean number of admissions:

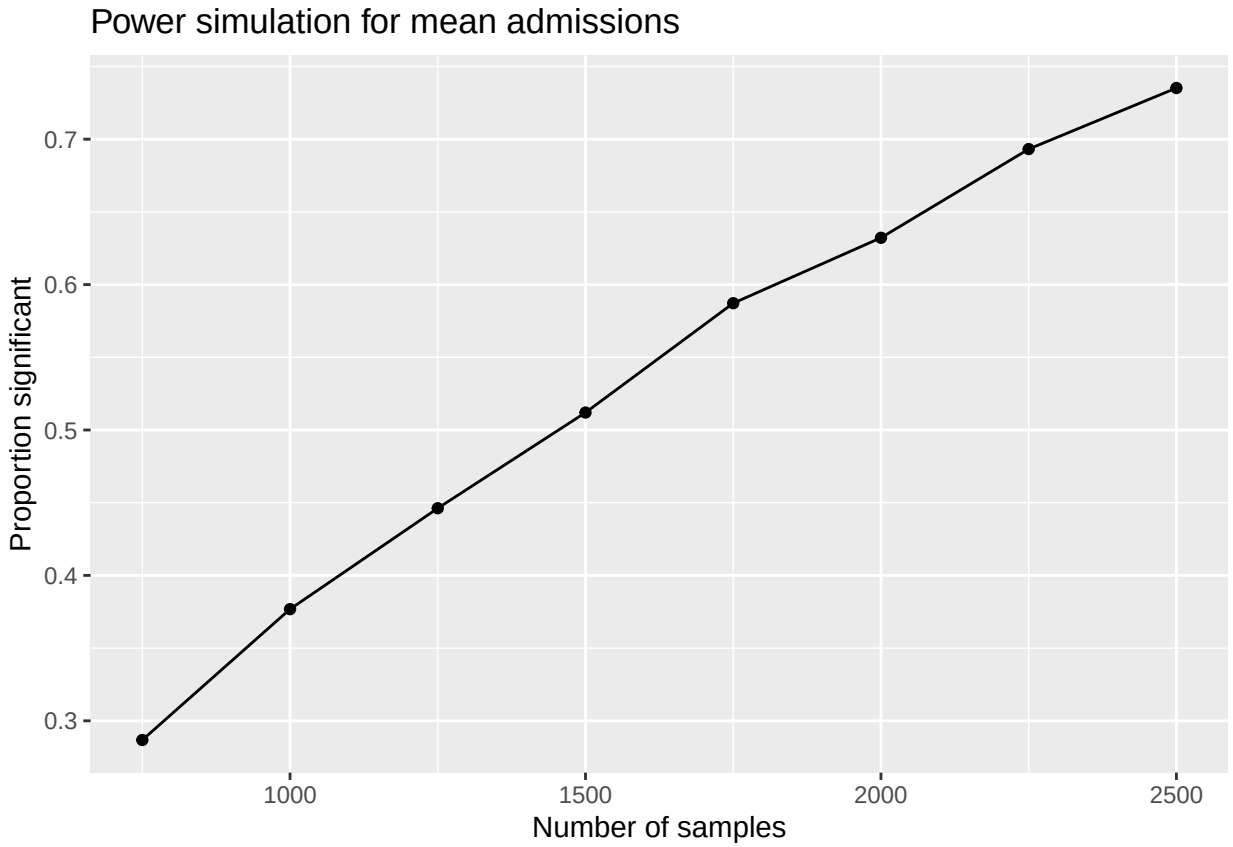
```
## ibis yes 0.3273857
```

```
## ibis no 0.4553073
```

```
## percent change 0.2809565
```

Now run the simulations, using t test to test significance of difference in means.

n	prop_sig
750	0.2868
1000	0.3768
1250	0.4462
1500	0.5120
1750	0.5872
2000	0.6322
2250	0.6932
2500	0.7352

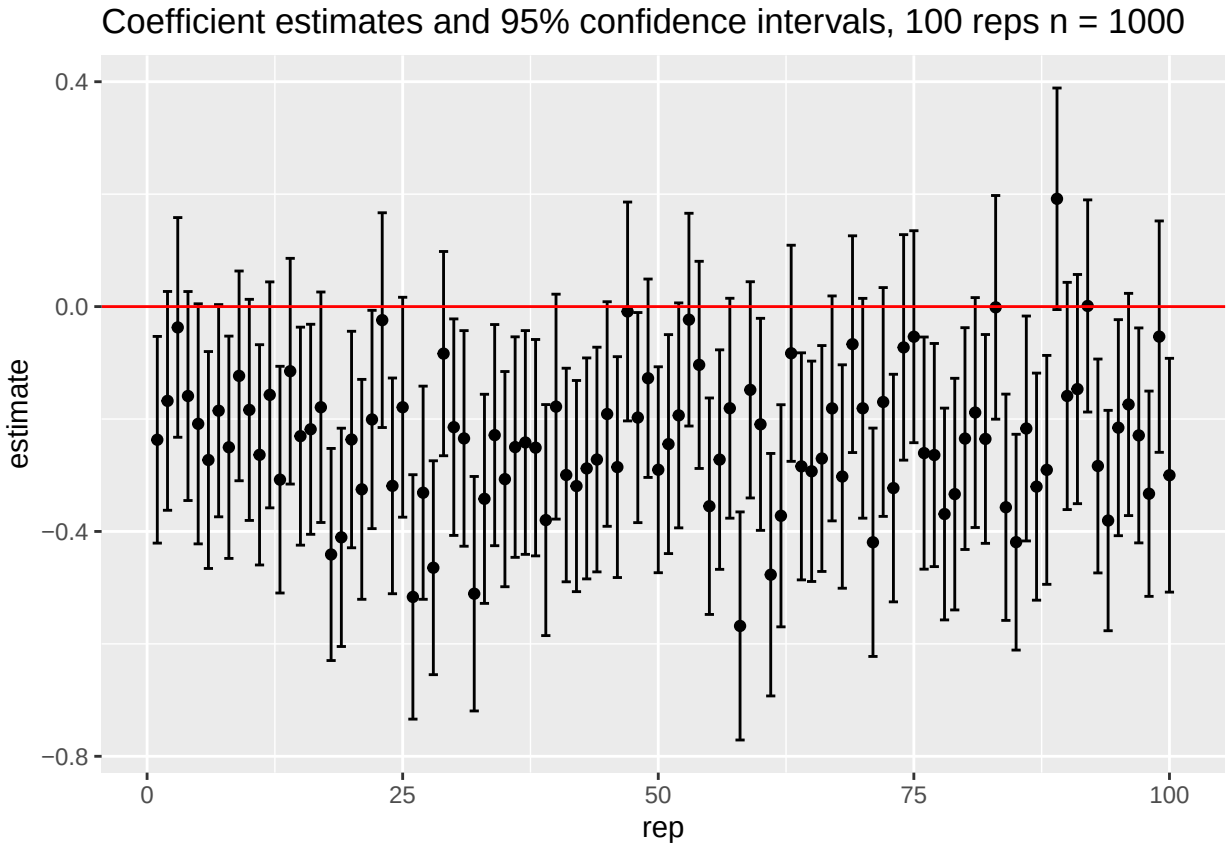


Test on regression coefficient; include age as covariate As before, we test the coefficient of the treatment indicator variable in a Poisson regression, including the age covariate.

n	prop_sig
750	0.5238
1000	0.6216
1250	0.6894
1500	0.7508
1750	0.8086
2000	0.8542
2250	0.8788

We again look at the variability in the coefficient estimates with n = 1000[4].

[4]Here the coefficient is negative, reflecting a reduction in the mean



Math notes

Most of our power calculations use a normal approximation for an estimate $\hat{\theta}$ of parameter θ, which in our case is either a mean or a proportion, and we look for differences in treated and untreated, θ_t, and θ_c; resp. When $\theta_c - \theta_t = \delta$, $\hat{\theta}_c - \hat{\theta}_t - \delta$ is normally distributed with zero mean. If the θ are proportions, with equal sample size equal n for treatment and control, the standard error is $SE = \sqrt{\frac{2\hat{p}(1-\hat{p})}{n}}$ where $\hat{p}$ is the pooled mean proportion. We get an analogous expression when the θ are means; in particular, the standard error is again proportional to $1/\sqrt{n}$.

For a two sided test, at significance α, the probability of rejecting the null hypothesis $H_0 : \theta_t = \theta_c$ is

$$\begin{aligned}
\beta &= P\left(\frac{\hat{\theta}_c - \hat{\theta}_t}{SE} > z_{\alpha/2}\right) \\
&= P\left(\frac{\hat{\theta}_c - \hat{\theta}_t - \delta}{SE} > z_{\alpha/2} - \frac{\delta}{SE}\right) \\
&= 1 - \Phi\left(z_{\alpha/2} - \frac{\delta}{SE}\right) \\
&= \Phi\left(\frac{\delta}{SE} - z_{\alpha/2}\right)
\end{aligned}$$

where $\Phi(x)$ is the "erf" function giving the probability $P(Z < x)$ where $Z \sim \mathcal{N}(0, 1)$, and "critical values" z_γ are defined by $1 - \Phi(z_\gamma) = \gamma$. It follows that

$$z_\beta = \frac{\delta}{SE} - z_{\alpha/2}$$

For given β, α, and δ, we can solve for n, which is inside the expression for SE. The expression for SE depends on what is being estimated, but all are proportional to $1/\sqrt{n}$. In the case of means, if standard deviations are not known, the Z standard normal distribution is replaced by t distribution but the arguments are similar.