"Statistical Spidey knows the score." by Matthew B. Wall
https://computingforpsychologists.wordpress.com/2013/04/11/comment-on-the-button-et-al-2013-neuroscience-power-failure-article-in-nrn/

Introduction to Biostatistics
Lecture 10A

Henrik Renlund

April 14, 2016

What shall we learn today?

This:

- statistical power

and its relation to

- sample size calculations.

(Bootstrap example in notes outside the scope of this course.)

How to make 'null' results meaningful

Generally, it 'H_0 not rejected' is uninformative *unless* we know that the study had a good chance of detecting an effect that is interesting.

E.g: "Two methods of pain relief were compared. The difference was not statistically significant."

This would be enhanced by;

"The study was designed to have a 90% chance of detecting a clinically significant difference of 9 (on the VAS scale)".

Attempt to visualize the power of a test

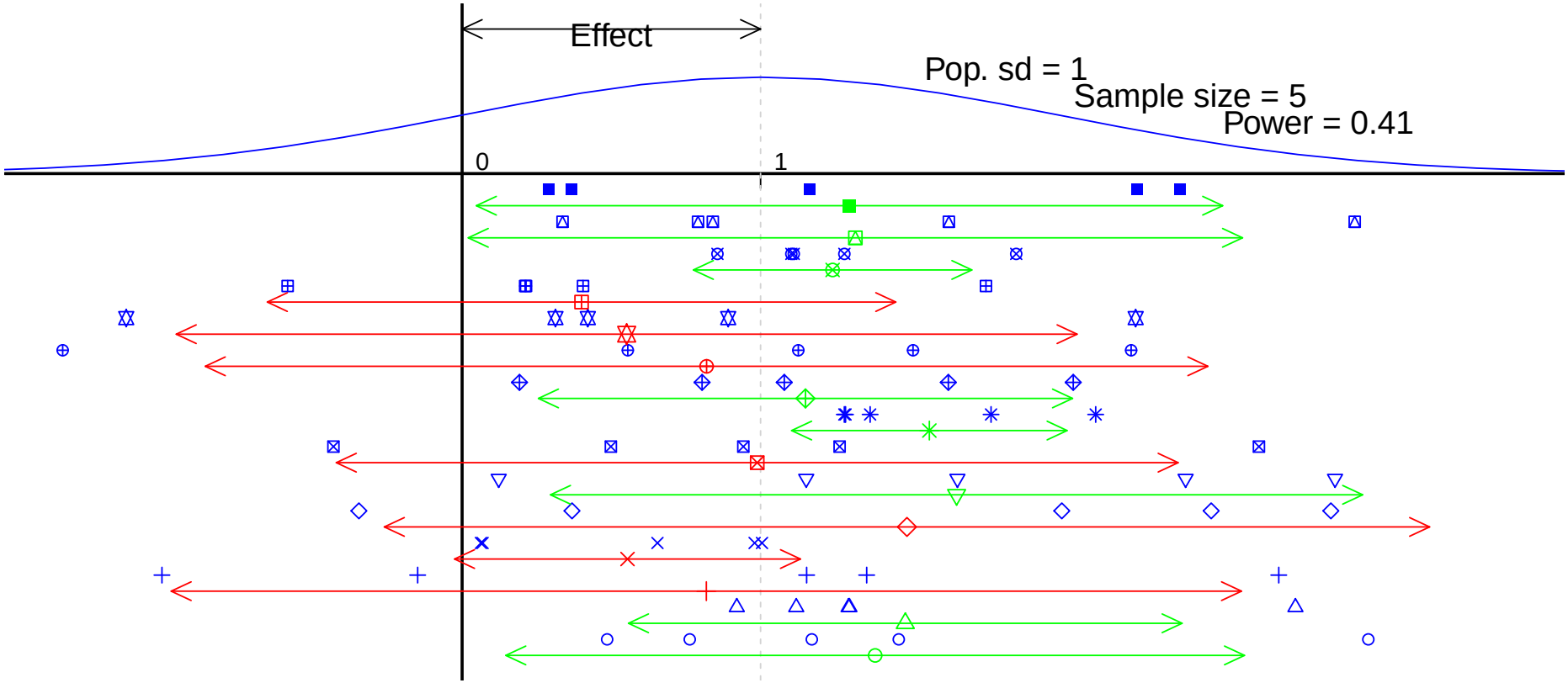
Suppose we measure the effect of a drug that on average does decrease the blood pressure by a clinically significant amount (defined as 1 unit).

We measure the blood pressure on n indivuals before and after taking the drug.

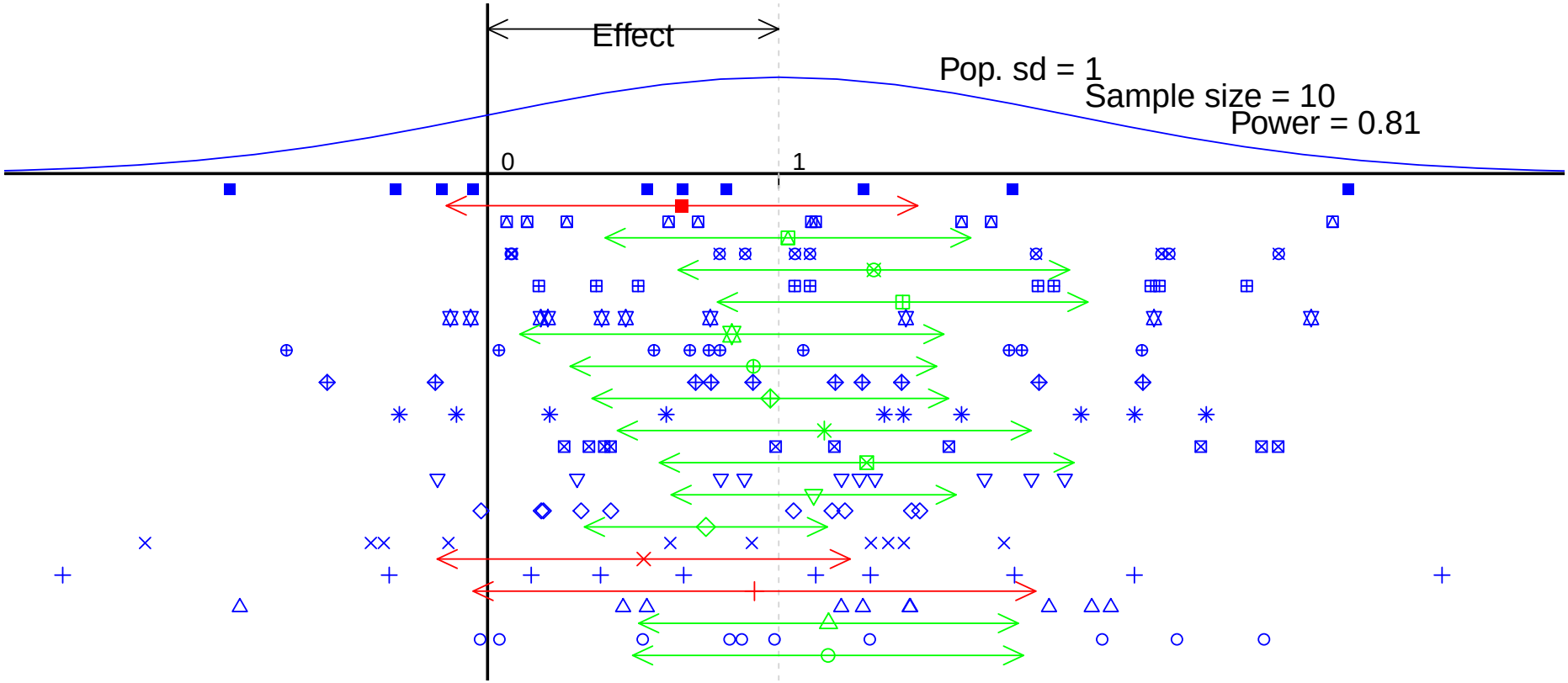
Our data consist of n 'indivual effects' (before - after) which are positive if the drug works. We assume these are Normal with a standard deviation of 1 unit.

H_0 :"average effect $= 0$" is determined by creating a 95% confidence interval for the mean effect.

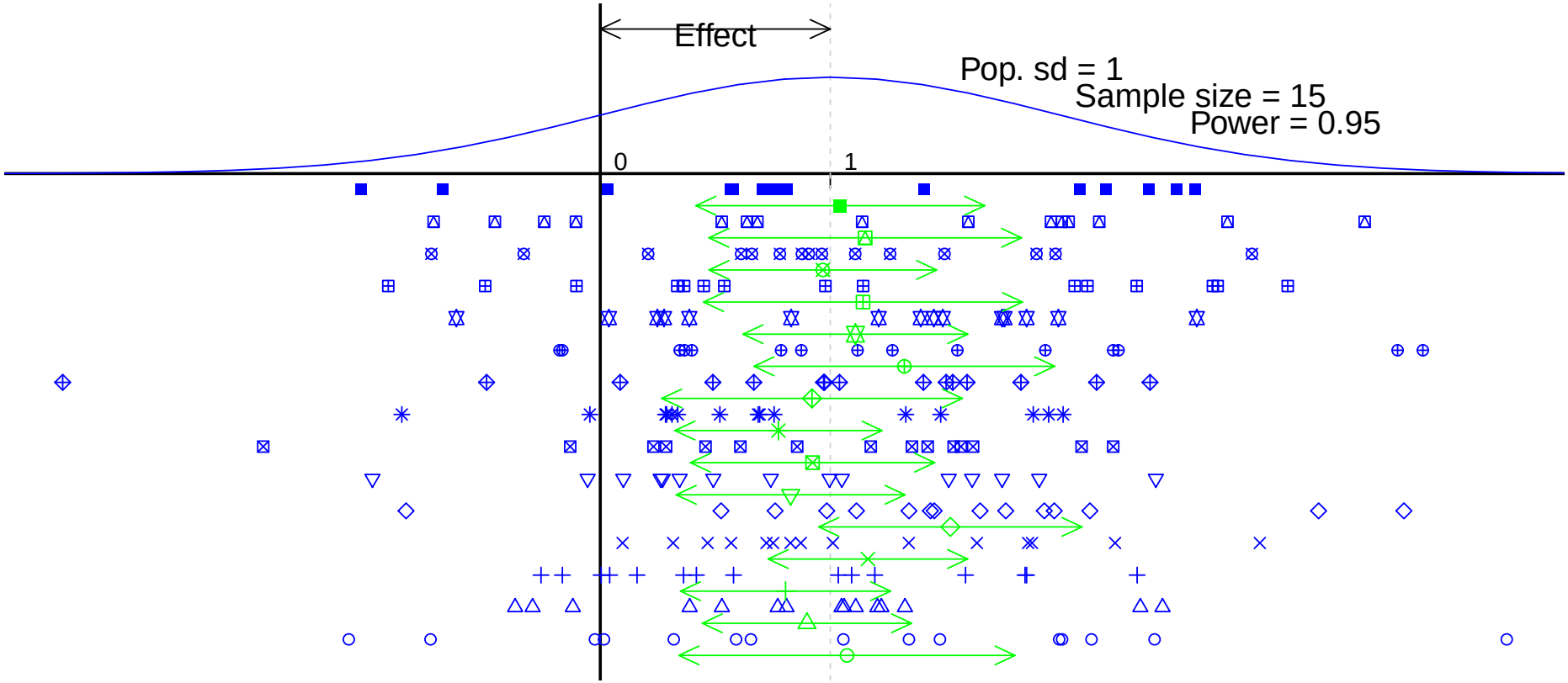
15 samples of size 5



15 samples of size 10



15 samples of size 15



Power

The power of a test is the probability of rejecting the null hypothesis.

The power depends on

- effect size
- sample size (maybe under your control)
- the spread of the data
- (the statistical test, significance level, etc.)

We want a test to have high power as soon as the effect is *interesting*.

The power is often thought of as a function of sample size (n) and effect. Set

- the power wanted,
- the effect to be *at least interesting*,

and figure out what n needs to be.

Sample size calculation

Recall the breastfeeding example from lecture 4.

pair	1	2	3	4	5	6	7	8	9	10	11	12	13
bottle	8	8	31	8	13	21	26	39.0	77.0	29	35	182	186
breast	7	10	28	12	19	14	15	51.0	65.0	12	11	24	17
diff	1	−2	3	−4	−6	7	11	−12.0	12.0	17	24	158	169
rank	1	2	3	4	5	6	7	8.5	8.5	10	11	12	13

part	rank	sum
negative	2, 4, 5, 8.5	19.5
positive	1, 3, 6, 7, 8.5, 10, 11, 12, 13	71.5

The null hypothesis of no difference was tested with the Wilcoxon signed rank test. (This tests for a shift in median.)
The p-value was 0.07

Think of this as being a pilot study and assume the data is representative of some target population. If the observed effect, a median shift of 9 days, is considered interesting (and thought to be true) what sample size would we need in order to show this?

Two possible approaches:

- Make a model (Normal? via transformation?) for the data and solve analytically or with software
- "Pull yourself up by your own bootstraps" - resample from the pilot study data.

In either case we will use the given sample:

$$z = \{1, -2, 3, -4, -6, 7, 11, -12, 12, 17, 24, 158, 169\}$$

Approximation with the R software

Simple sample-size software exists on-line and possibly in your statistical software. In (base) R we can use function for t-test to get a ballpark figure.

```
power.t.test(power = 0.8, delta = mean(z), sd = sd(z),
             type = "one.sample")
```

```
     One-sample t test power calculation

              n = 35.94785
          delta = 29.07692
             sd = 60.49995
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

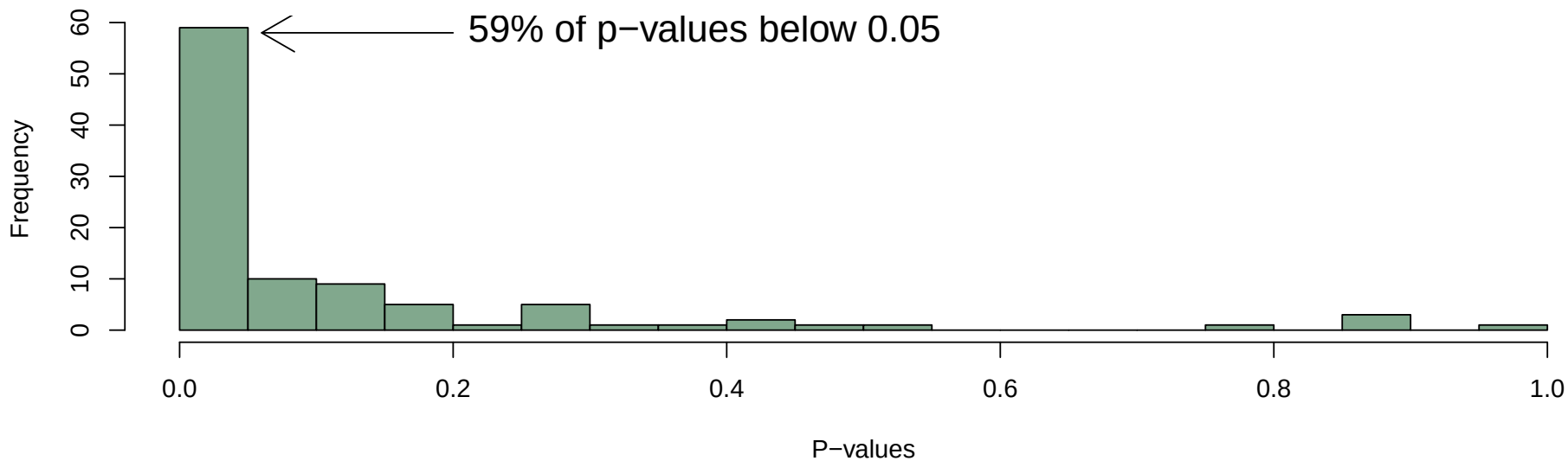
Since the distribution is so skewed it is not unreasonable to think that the Wilcoxon-test will be more powerful.

Bootstrap method

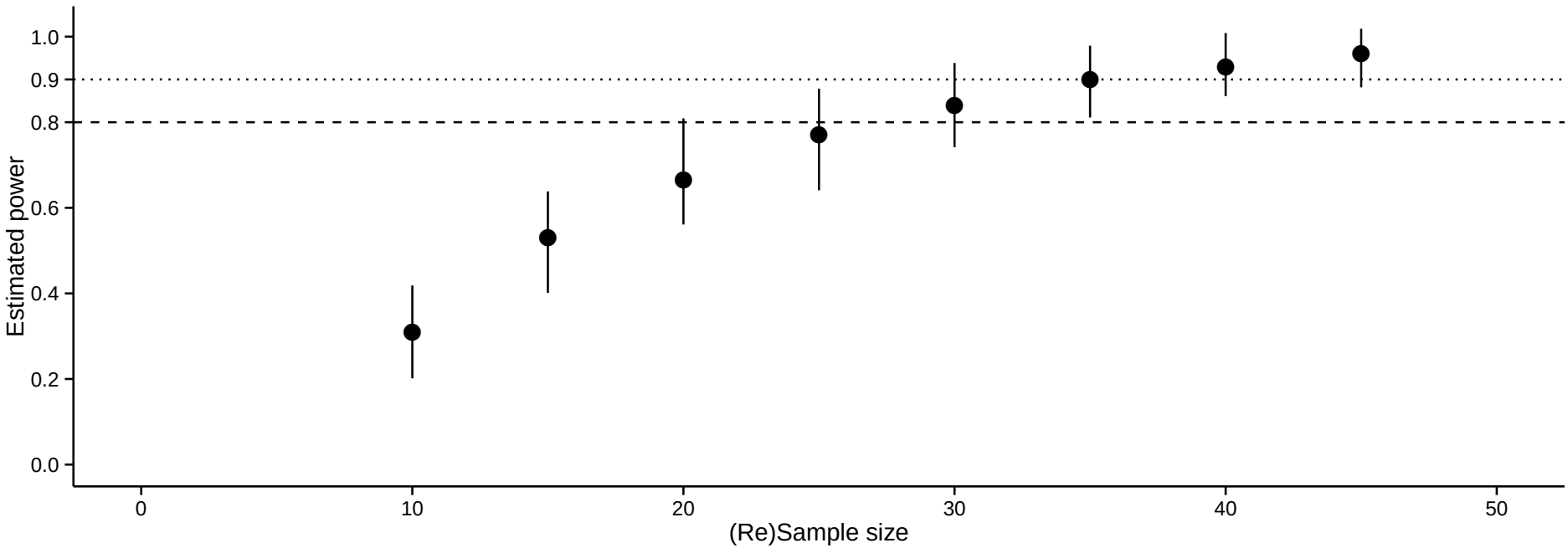
Resampling using the same same sample size (13).

$$3, 1, -6, 24, 158, 17, 1, -2, 7, -12, 1, -2, 24, 1$$

We are interested in the power of the test, so we want to see how often the null is rejected. For that purpose we collect the p-value. In this case it is 0.15.
Now lets repeat this process 100 times. What kind of p-values do we get?



These 100 'regenerated' p-values, suggest that the power was 0.59.
This is an estimate! (The error will depend on both the resample *and* inital sample.)
The plot shows a bootstrap estimate for varying sample sizes:



So you should probably use at least 35 individuals. (There the power is estimated to be at least 80%.)

References

- Chapters 23-25: Petrie & Sabin. *Medical Statistics at a Glance*, Wiley-Blackwell (2009).