Lecture 7: Power of a Test

Caden Hewlett

2024-01-22

Power of a Hypothesis Test

The discussion here is limited to the consideration of one type of error that can occur when testing a null hypothesis H_0 against an alternative H_A.

There are two types of errors that can occur.

A **Type 1 Error** is when you reject H_0 when it is true. Consider the case of convicting an innocent person.

We let:

$$\alpha = P(\text{Type I Error}) = P(\text{ Reject } H_0 \mid H_0 \text{ is true.})$$

Futher, we let:

$$\beta = P(\text{Type II Error}) = P(\text{ Fail to Reject } H_0 \mid H_0 \text{ is false.})$$

Then, the *Power of the Test* is given by:

$$\text{Power } = 1 - \beta$$

This is all well and good, but how do we calculate this?

The Power Function

The Power can be succinctly described as "the ability to correctly detect a false null hypothesis."

In other words,

$$\text{Power } = P(\text{Reject } H_0 \mid H_0 \text{ is false.})$$

We can only calculate power against a specific alternative. In other words, for parameter Θ and hypothesized values $\theta_0, \theta_A \in \mathbb{R}$, we need hypotheses of the type:

$$H_0 : \Theta = \theta_0 \quad \text{against} \quad H_A : \Theta = \theta_A$$

The power of a test cannot be calculated in a frequentist sense when H_A proposes a range of values for the parameter of interest (i.e. $H_A : \Theta > \theta_0$)

Therefore, if we have an observed test statistic T_{obs} and a critical value from a distribution t^*, we could write the p-value in a way we already know. In the case of a one-sided test.

$$p = P(|T_{\text{obs}}| > t^* \mid \Theta = \theta_0)$$

Conversely, the power is simply:

$$\text{Power } = P(|T_{\text{obs}}| > t^* \mid \Theta = \theta_A)$$

Where, notbaly, the **critical value** t^* is the same value across both tests.

Example

From the "precognition" data, each subject's response would be out of $n = 10$ guesses, how many T did they guess correctly?

Suppose we reject H_0 when $T \geq 8$. What's the significance level?

Knowing that $T \sim \text{bin}(10, 0.5)$, we have:

```r
sum(dbinom(8:10, 10, 0.5))
```

```
## [1] 0.0546875
```

So in this case, $\alpha \approx 0.0547$.

In this study, the parameter of interest is p. The probability that the subject can correctly predict the image location.

So, if $p > 0.5$ the person is either lucky or has some "precognition."

Therefore, the power of the test is in this case probability of rejecting H_0 when in truth $p > 0.5$.

In this specific example, the power funciton is:

$$\text{Power}(p) = \pi(p) = P(T \geq 8 \mid p)$$

In this case we can actually write an `R` function to reflect the power.

```r
power <- function(p){
  return(
    sum(dbinom(8:10, 10, p))
)
}
power(0.8)
```

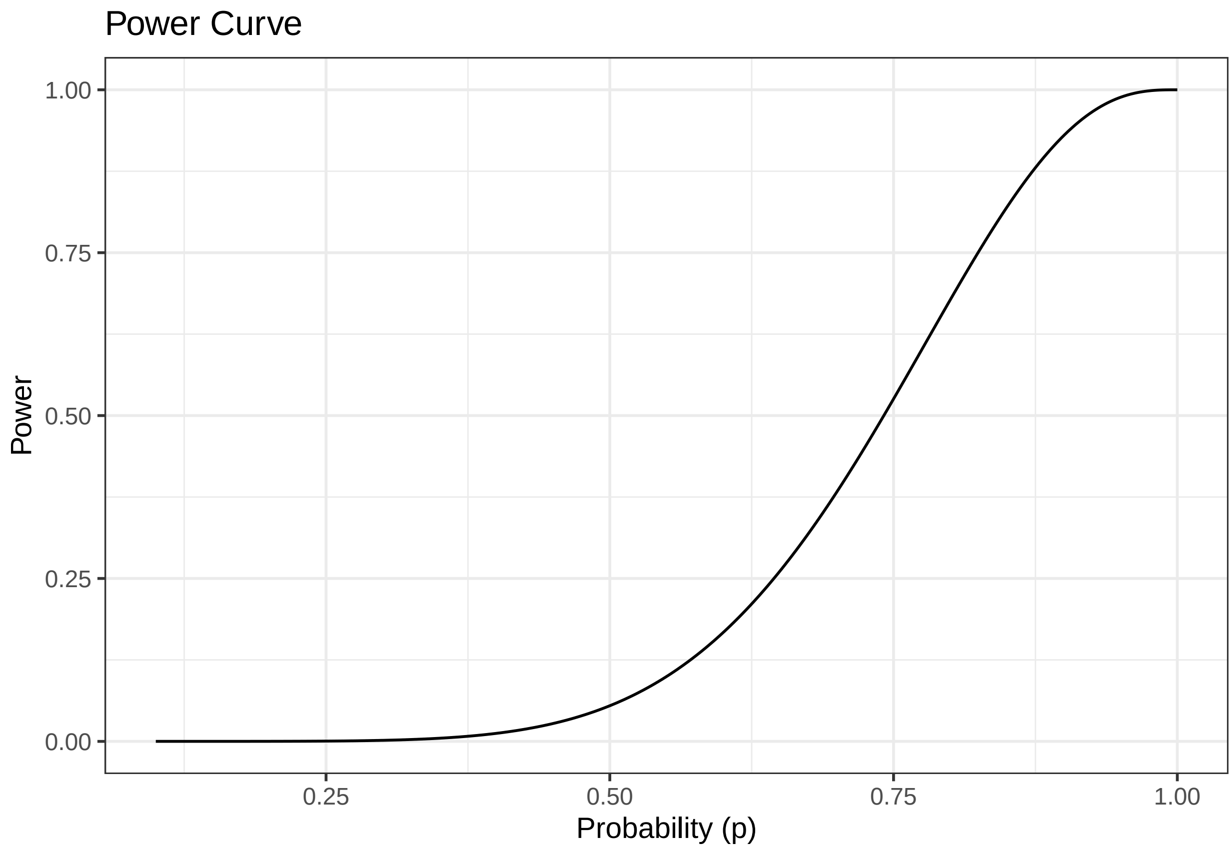
```
## [1] 0.6777995
```

You can also use the equation for the binomial distribution to find the discrete sum of observing values in $\{8, 9, 10\}$.

Then, if we want to find the power for all "valid" values of p, in this case $p \in [0, 1]$, we can plot it in a **Power Curve.**

```r
p_values <- seq(0.1, 1, length.out = 1000)
power_values <- sapply(p_values, power)
power_data <- data.frame(p_values, power_values)

p<-ggplot(power_data, aes(x = p_values, y = power_values)) +
  geom_line() +
  labs(title = "Power Curve", x = "Probability (p)", y = "Power") +
  theme_bw()
print(p)
```



Two-Tailed Power

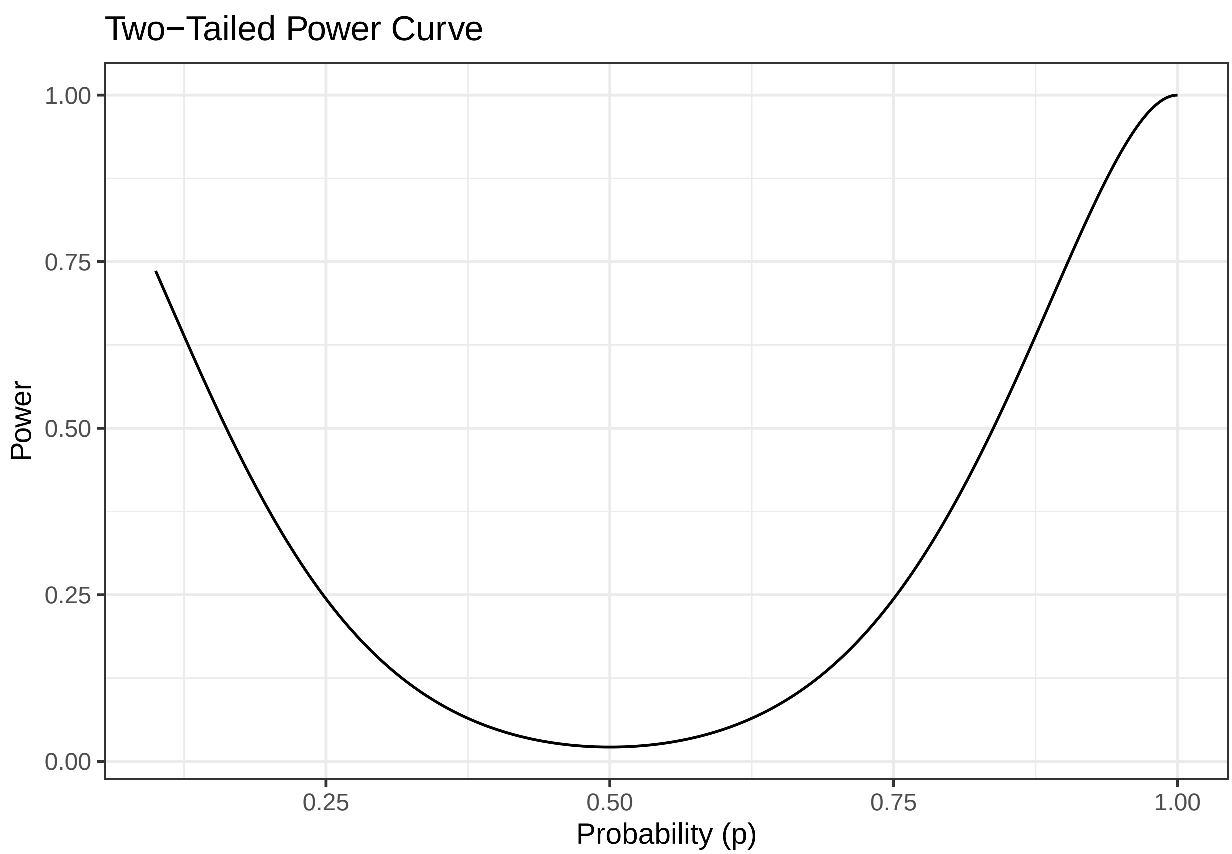
In this case, we consider probabilities on either tail. To get an α near-ish to 0.05, we would reject if we observe $T_{\text{obs}} \in \{0, 1, 9, 10\}$

The power curve would then look a bit different, since there is high probability density associated with lower values and higher values, but there would be a dip in the center.

```r
power_new <- function(p){
  return(
    sum(dbinom(9:10, 10, p), dbinom(0:1, 10, p))
)}

power_values <- sapply(p_values, power_new)
new_power_data <- data.frame(p_values, power_values)

p<-ggplot(new_power_data, aes(x = p_values, y = power_values)) +
  geom_line() +
  labs(title = "Two-Tailed Power Curve", x = "Probability (p)", y = "Power") +
  theme_bw()
print(p)
```



Example

In the premonition study, we observed $p = 0.531$, and the researcher chose to reject the null hypothesis with a p-value of 0.01. We cannot tell from these results of premonitions are a real phenomenon. These results are still theoretically possible by chance. We may question the power of this test since the value was so close to the hypothesized value, therefore it would be wise to replicate this study to verify these results.

Further notes

It is also important to know that the T required to reject H_0 will change with sample size. In our ten people, we'd need to observe $p = 0.8$ to reasonably reject H_0. If we had 100 individuals, we would only need to observe $p = 0.58$ to reject. The reasonable conclusion is that an increase in sample size allows us to observe smaller differences. We can compare the power curves if we wish.

Summary

The general approach (via the Neymann-Pearson Lemma) is to fix α, then attempt to maximize power with respect to this fixed significance.

Power is affected by four main things. Firstly, the **sample size** the value of the **specific alternative**, the **significance level** and the **effect size** (approximate differnece between hypothesized and true.)

Power is often used in sample size calculations e.g. how many people do we need to observe a given power for our test?