Statistical Power

The Power of a Hypothesis test

Statistical hypothesis tests have two possible errors that can occur, referred to as Type I and Type II errors. A Type I error occurs when the null hypothesis, H_0, is rejected when it is true: $\mathbb{P}\big(\text{reject } H_0|H_0 \text{ is true}\big)$. Type I errors are associated with the confidence level of the test: For a 95% confidence level, the probability of a type I error is 5%, i.e. you will reject the null hypothesis approximately 5% of the time by chance alone.

A Type II error occurs when the null hypothesis is not rejected when in fact the alternative is true: $\mathbb{P}\big(\text{ do not reject } H_0|H_1 \text{ is true}\big)$. The *Power* of a test is defined as $1 - \mathbb{P}\big(\text{Type II error}\big)$, i.e. the probability the test will correctly reject the null hypothesis given that an alternative is true.

For a statistical test of the form $H_0 : p_1 = p_2$ vs. $H_1 : p_1 \neq p_2$, or equivalently, $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 > 0$, we can simulate the power of the test for various values of the parameters. Sample proportions are simulated from given sample sizes and assumed success probabilities. (Success probability here is the probability the customer will pay). The following code simulates, using 10,000 simulation runs, and calculates the power for a range of alternative hypotheses:

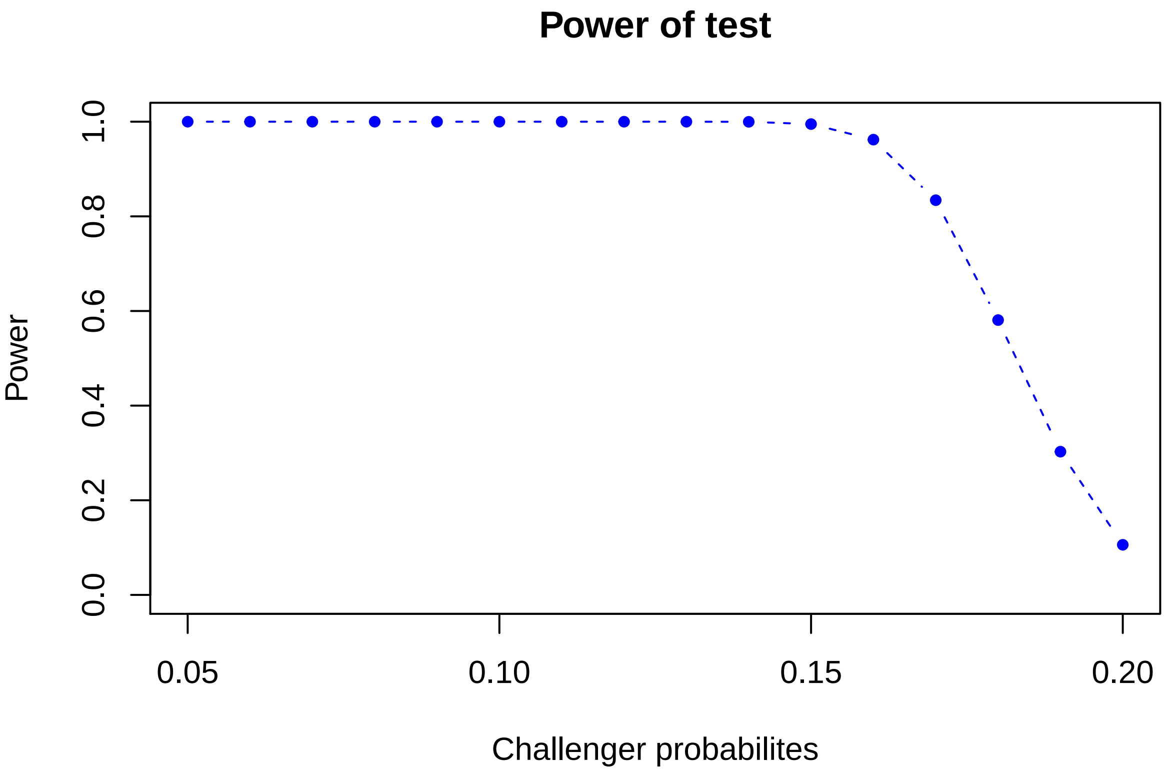
```
n1 <- 5000  ## champion sample size
n2 <- 1000  ## challenger sample size
p1 <- 0.2   ## champion probability
p2 <- 0.15  ## challenger probability

test_sim <- function(n1,p1,n2,p2,sims = 10000,conf.level = .9) {
test_thresh <- qnorm(conf.level) ## threshold to test significance against z-score
results <- rep(NA,sims) ## store results of test for each sim
for (i in 1:sims) {
  pchamp <- sum(rbinom(n1,1,p1))/n1  ## generate n1 samples, calculate est of p1
  pchall <- sum(rbinom(n2,1,p2))/n2  ## same as above for challenger
  var1 <- pchamp*(1-pchamp)/n1  ## champ variance
  var2 <- pchall*(1-pchall)/n2  ## chall variance
  stanDev <- sqrt(var1+var2) ##  stand deviation of (pchamp - pchall)
  z <- (pchamp-pchall)/stanDev  ## z score
  if (z > test_thresh) {
    results[i] <- 1 ## assign 1 if test shows significance
  } else {
    results[i] <- 0
  }

 }
return(sum(results)/sims)
}

p_alt <- seq(0.05,p1, by = 0.01)  ## alternative probs for challenger
power <- rep(NA,length(p_alt))  ## vector to hold power of tests

for (i in 1:length(p_alt)) {
  power[i] <- test_sim(n1,p1,n2,p_alt[i])
  }
```

Results

```
data.frame(cbind(p_alt,power))
```

```
##    p_alt  power
## 1   0.05 1.0000
## 2   0.06 1.0000
## 3   0.07 1.0000
## 4   0.08 1.0000
## 5   0.09 1.0000
## 6   0.10 1.0000
## 7   0.11 1.0000
## 8   0.12 1.0000
## 9   0.13 1.0000
## 10  0.14 0.9998
## 11  0.15 0.9950
## 12  0.16 0.9621
## 13  0.17 0.8341
## 14  0.18 0.5808
## 15  0.19 0.3027
## 16  0.20 0.1059
```

```
 plot(p_alt,power, pch = 20, col = 'blue',type='b',lty=2,ylim=c(0,1),main = "Power of test", xlab = "Cha
```



The simulation fixes the champion probability at 0.2. As expected, when the challenger probability is 0.2 as well, the power of the test is 0.1 since the confidence level is 0.9. The power decreases as we approach 0.2 since it becomes harder to detect significant differences when the probabilities are close.