HW3

Ting Huang

Saturday, February 13, 2016

Problem 1:

(a)

Null hypothesis H_0 : there is no difference in average recall ability between sleep group and caffeine group. Alternative hypothesis H_a : the average recall ability of sleep group is higher than that of caffeine group.

Test statistic: $t = \frac{15.25-12.25}{s_{pooled}\sqrt{1/15+1/12}} = 2.254862$ where $s_{pooled} = \sqrt{\frac{(15-1)*3.3^2+(12-1)*3.6^2}{15+12-2}} = 3.435229$. $df = 15+12-2 = 25$.

Since $t_{1-0.5} = 1.708141$ and $t > t_{1-0.5}$, we reject the null hypothesis. Therefore, we can conclude that caffeine impairs recall ability.

(b)

Null hypothesis H_0 : there is no difference in average recall ability between sleep group and caffeine group. Alternative hypothesis H_a : the average recall ability of sleep group is higher than that of caffeine group.

Test statistic: $t = \frac{15.25-12.25}{\sqrt{3.3^2/15+3.6^2/12}} = 2.23235$, where $df = \frac{(3.3^2/15+3.6^2/12)^2}{(3.3^2/15)^2/(15-1)+(3.6^2/12)^2/(12-1)} = 22.69996 \approx 23$.

Since $t_{1-0.5} = 1.713872$ and $t > t_{1-0.5}$, we reject the null hypothesis. Therefore, we conclude that caffeine impairs recall ability.

T-test with equal variance assumption has larger degree freedom and smaller standard error.

(c)i.

For sleep group, $df = 15-1 = 14$, $t_{1-0.05/2} = 2.144787$.

$CI_1 = [15.25 - 2.144787 * 3.3/\sqrt{15}, 15.25 + 2.144787 * 3.3/\sqrt{15}] = [13.42252, 17.07748]$.

For caffeine group, $df = 12-1 = 11$, $t_{1-0.05/2} = 2.200985$.

$CI_2 = [12.25 - 2.200985 * 3.6/\sqrt{12}, 15.25 + 2.200985 * 3.6/\sqrt{12}] = [9.962669, 14.53733]$.

The two intervals have some overlap. So we couldn't conclude that there is a difference in average recall ability between the two groups.

(c)ii.

If the 95% confidence interval is an interval that contains 95% of recalled words in each group, with increasing sample size, it is highly possible that the confidence interval becomes wider and wider. But in fact as the sample size increases, the confidence interval becomes narrower. This is because a 95% confidence interval is a range of values that we can be 95% certain contains the true mean of the population. With a larger sample, we know the population mean with more precision than a smaller sample, so the confidence interval for larger sample is narrower.

(c)iii.

The df for the student distribution is $\frac{(3.3^2/15+3.6^2/12)^2}{(3.3^2/15)^2/(15-1)+(3.6^2/12)^2/(12-1)} \approx 23$. $t_{1-0.05/2} = 2.068658$.

The 95% confidence interval for the difference between the means is

$CI = [15.25 - 12.25 - 2.068658 * \sqrt{3.3^2/15 + 3.6^2/12}, 15.25 - 12.25 + 2.068658 * \sqrt{3.3^2/15 + 3.6^2/12}] = [0.2199822, 5.780018]$.

(d)

1(a) does hypothesis testing but has additional assumption that an equal population variance between two groups but it has a smaller degree freedom. 1(b) does hypothesis testing and uses no assumption about population variance.

1(c.i) calculates a separate 95% confidence interval for each group. But, the probability that both of two intervals cover their true population mean is below 95%.

1(c.iii) calculates a joint confidence interval for the difference between the means.

A hypothesis test is to test whether a value is reasonable to assume as the parameter value. P-values are clearer than confidence intervals. It can be judged whether a value is greater or less than the assumed limit. However, hypothesis testing using a p-value is not very reliable.

On the other hand, with hypothesis testing all we usually know is that the hypothesized value is not a reasonable estimate for the population parameter but you do not know what the parameter is likely to be. A confidence interval is frequently better than p-value since it gives a range of values within which we can be confident to a certain degree that the parameter lies.

Problem 2

(a)

The formula to calculate confidence interval for binomial distribution:

$\bar{y}_n \pm z^{1-\alpha/2}\sqrt{\frac{\bar{y}_n(1-\bar{y}_n)}{n}}$

Since $\bar{y}_n = 0.6$ and $n = 1024$, then 95% confidence interval is $[0.5699943, 0.6300057]$.

90% confidence interval is $[0.5748184, 0.6251816]$.

95% confidence interval is wider. Because $z^{0.975}$ is bigger than $z^{0.95}$. Confidence interval is a range that we are 95% certain contains the true mean of the population. Thus, given the same samples, if we want to increase the confidence level, we have to widen the range.

(b)

Two ways to solve this problem: (Remember that 0.55 is a fixed number and its variance is zero.)

(1) Confidence interval: We want to make $\bar{y}_n - z^{0.975}\sqrt{\frac{\bar{y}_n(1-\bar{y}_n)}{n}} > 0.55$. Then $n > 368.7801$. Thus, the minimum number of 2015 interviews that would have been required to establish that the proportion of those in favor of legalizing same-sex marriage increased since the previous year is 369.

(2) Hypothesis testing:

$$n \geq (\frac{z_{1-\beta} + z_{1-\alpha/2}}{ES})^2$$

, where $ES = \frac{\delta}{\sqrt{\bar{p}(1-\bar{p})}}$.

Let $\alpha = 0.05$ and $\beta = 0.2$, then $n \geq ((0.8416212 + 1.959964)/0.05)^2 * (0.4 * 0.6) = 753.4924$. So,the minimum number of 2015 interviews that would have been required to establish that the proportion of those in favor of legalizing same-sex marriage increased since the previous year is 754.

Problem 3

(a)

The null hypothesis is the same as the alternative hypothesis. More importantly, the null or the alternative hypotheses should be made in terms of a population parameter μ, and not in terms of a data summary such as $\bar{x}$.

(b)

The samples in first group is a subset of those in second group. So the samples in two groups are not independent. The independent two-sample t-test can't be applied for this case.

(c)

Since $0.94 > \alpha$, there is no enough evidence to reject the null hypothesis.

(d)

Since the alternative hypothesis is $\mu_1 - \mu_2 < 0$ and $t > 0$, p-value is the area under the curve from t to the lower tail. Then the p-value is 1-0.018=0.982.

Problem 4

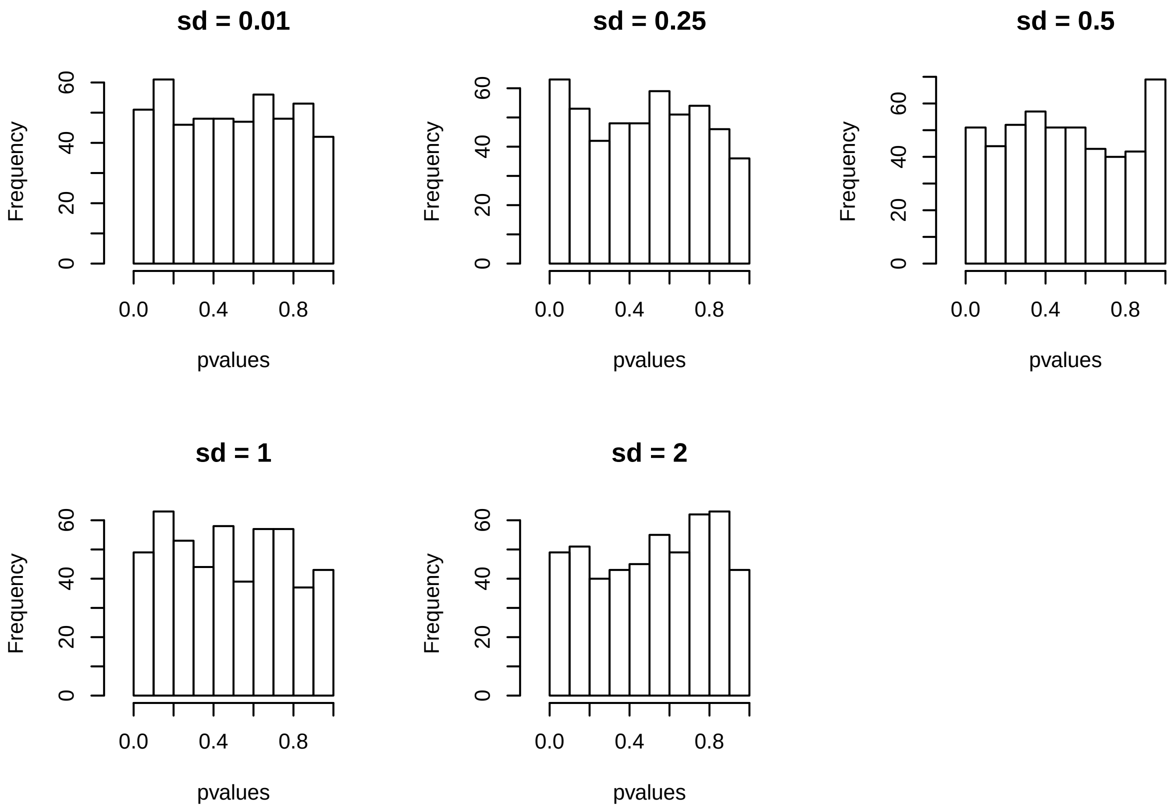
(a) and (b)

Use histograms to show the p-values.

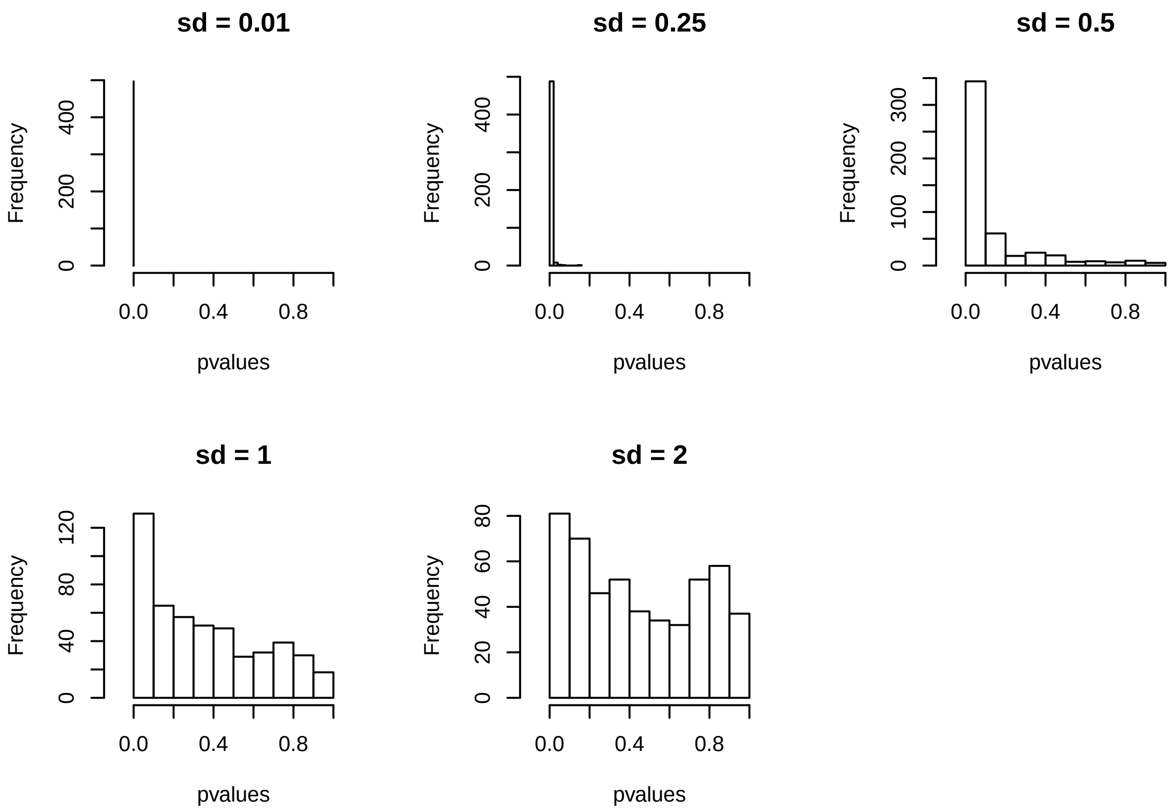
```
Sys.setenv(LANG = "en")
meandiff<-c(0,0.5,1,2,3)
sd<-c(0.01,0.25,0.5,1,2)
size <- c(10,30,300)
times<-500
for(i in size){
  cat('Sample size n=',i,"\n")
  for (mu in meandiff){
    cat('Difference between population mean is ',mu,"\n")
    op <- par(mfrow=c(2,3))
    for (s in sd){
      pvalues <- NULL
      for (r in 1:times){
        group1 <- rnorm(i, 0, s)
        group2 <- rnorm(i, mu, s)
        p <- t.test(group1, group2)$p.value
        pvalues <- c(pvalues,p)
      }
```

```
      hist(pvalues, main=paste("sd = ", s,sep=""),xlim=c(-0.1,1.1))
    }
  }
}
```

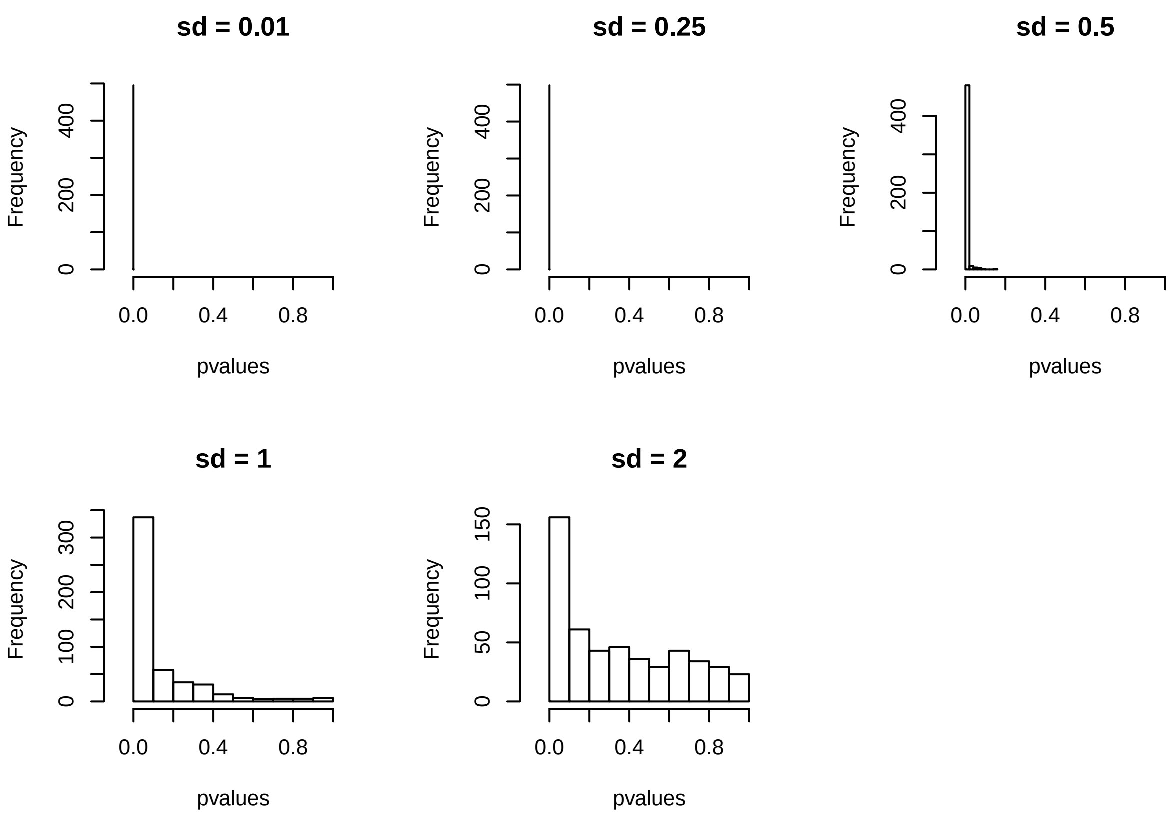
```
## Sample size n= 10
## Difference between population mean is  0
```



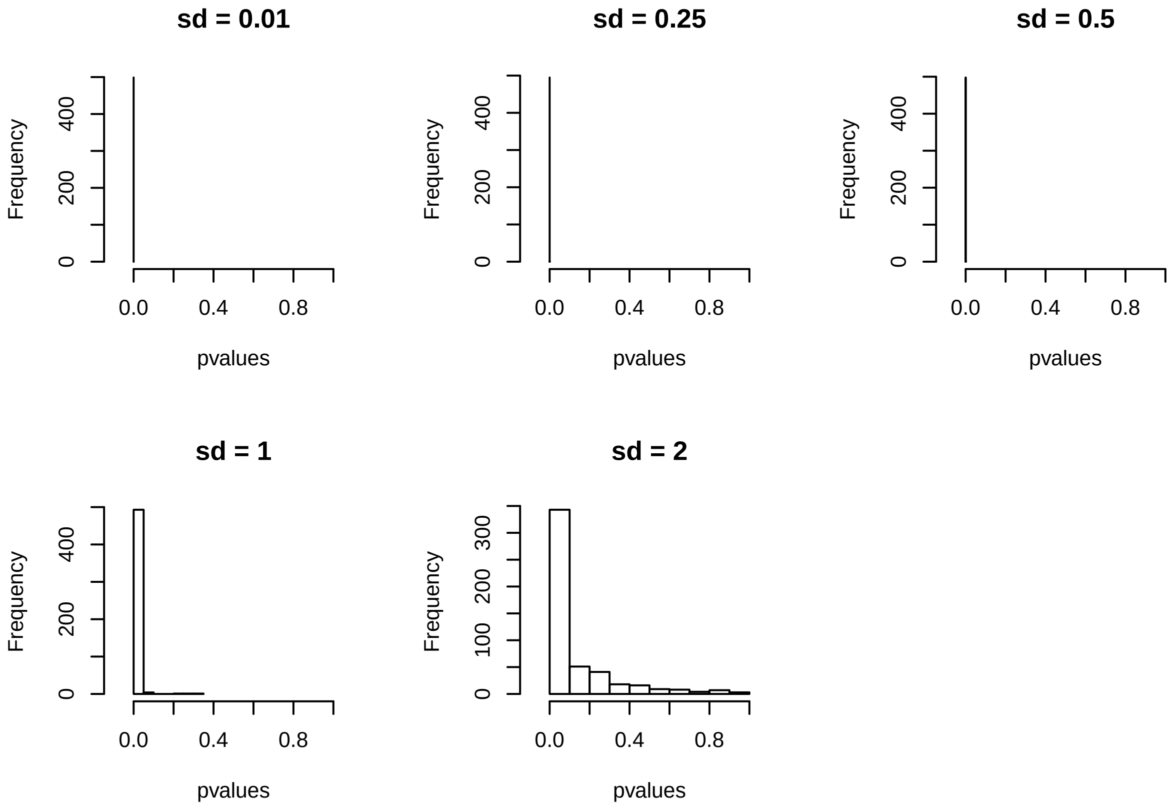

```
## Difference between population mean is  0.5
```



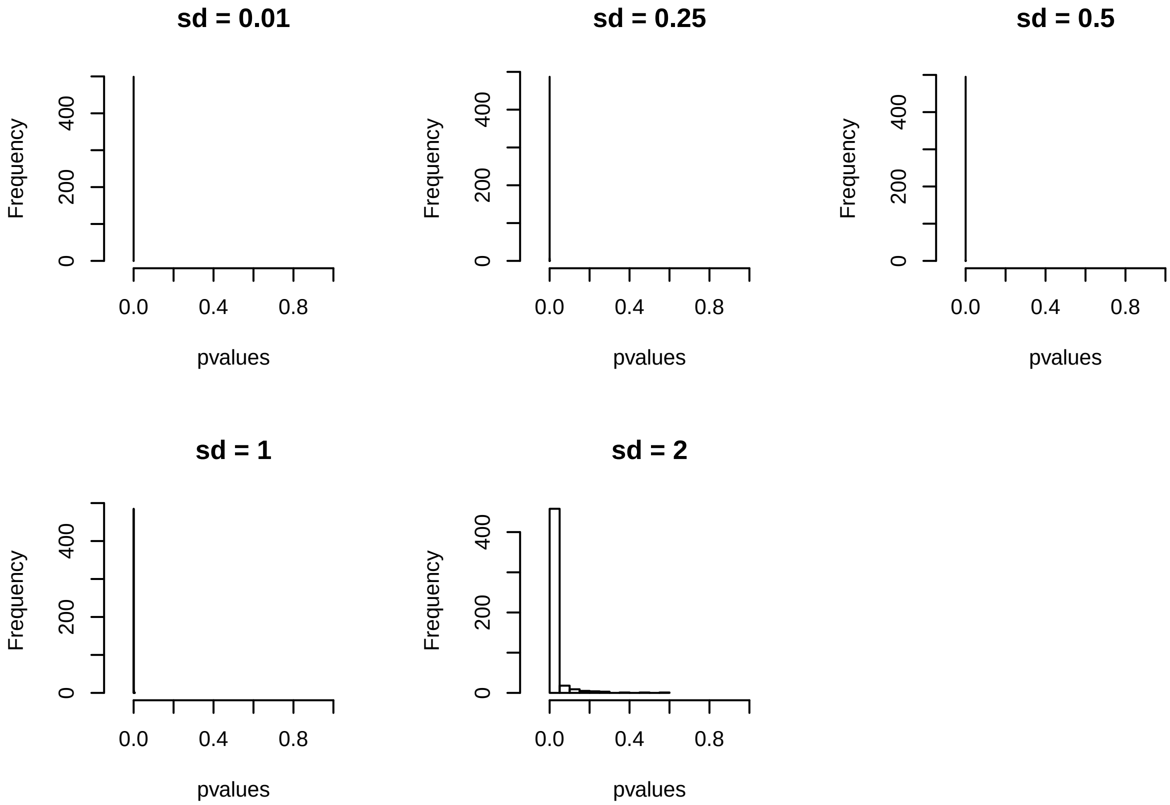

```
## Difference between population mean is  1
```



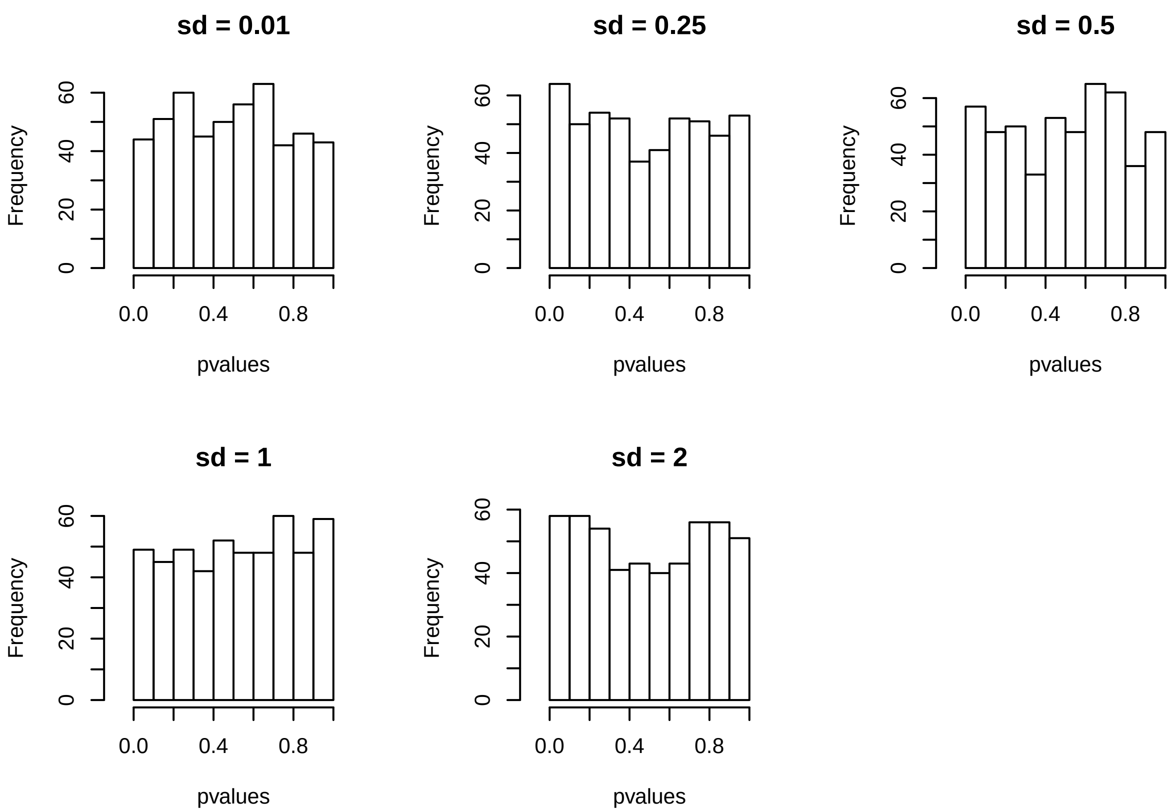
```
## Difference between population mean is  2
```



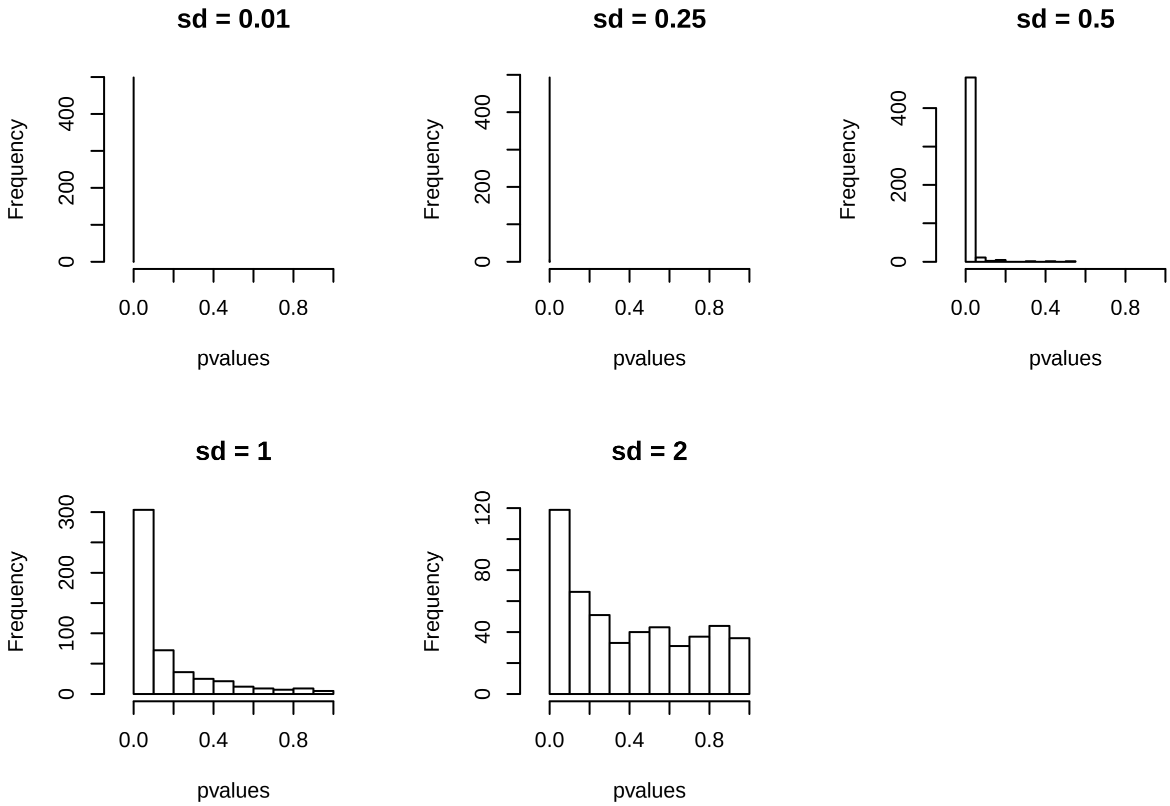

```
## Difference between population mean is  3
```



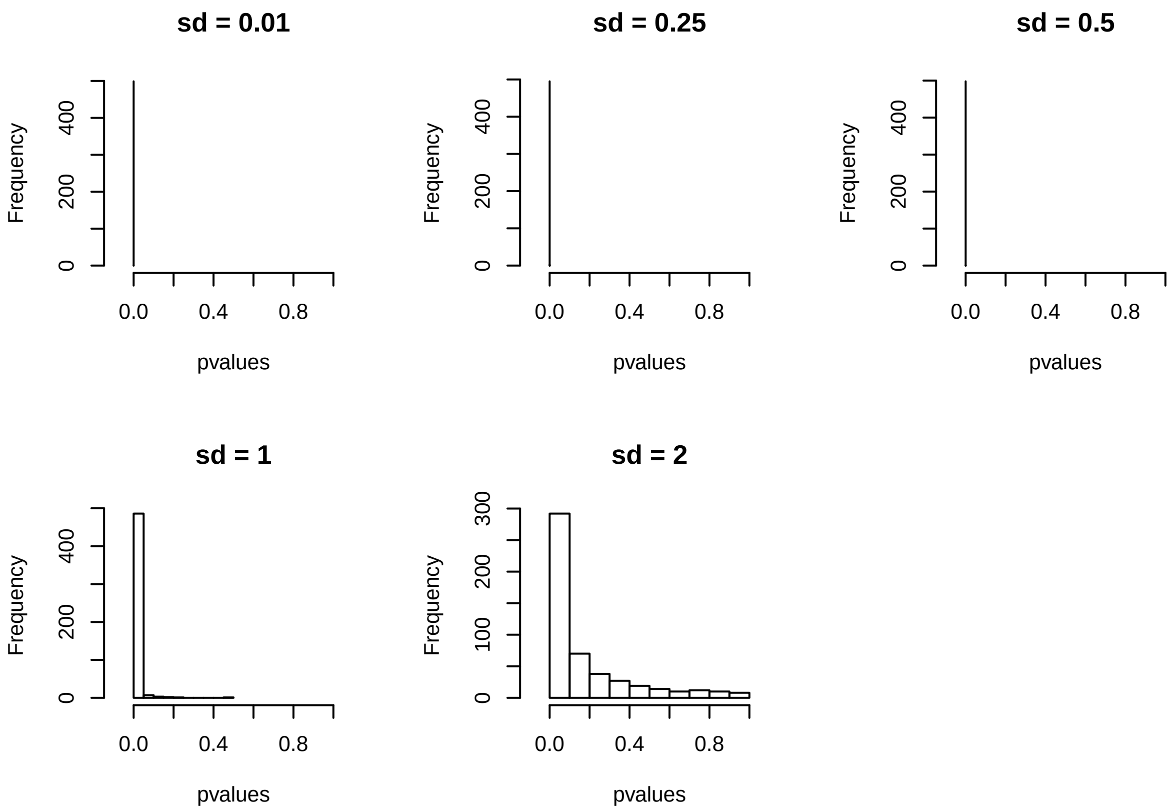

```
## Sample size n= 30
## Difference between population mean is  0
```



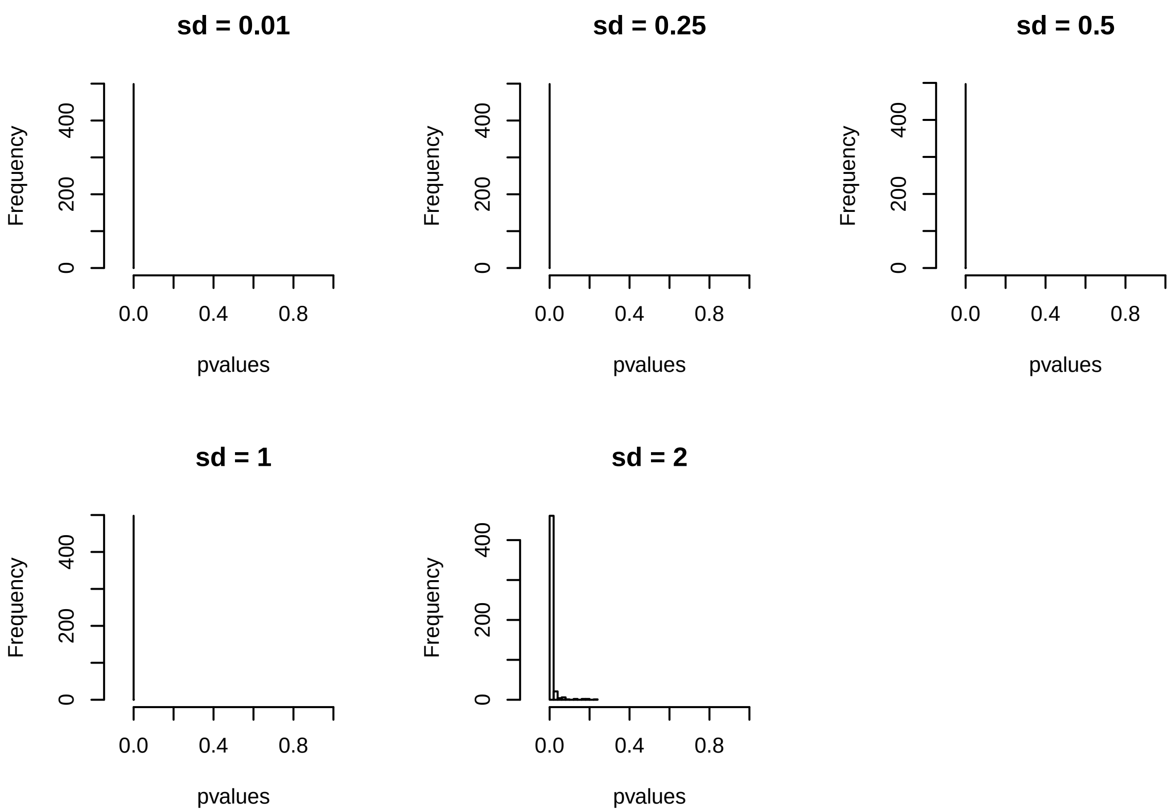
```
## Difference between population mean is  0.5
```



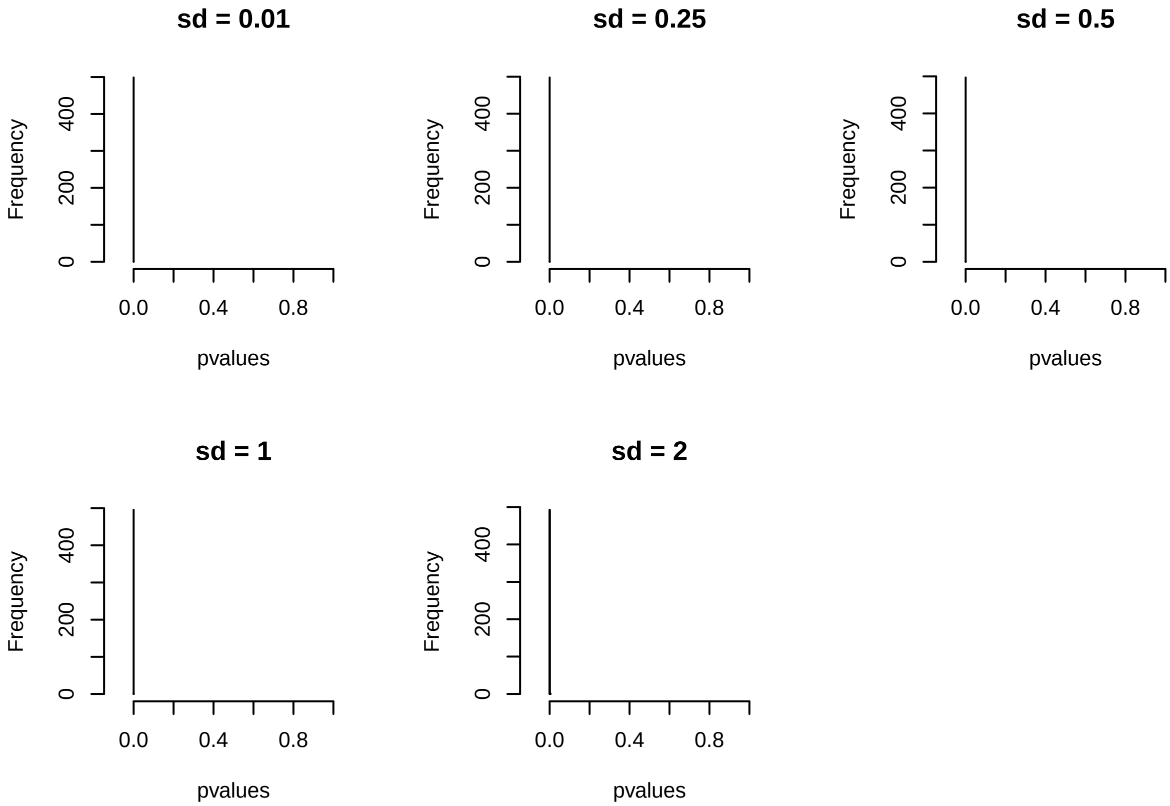

```
## Difference between population mean is  1
```



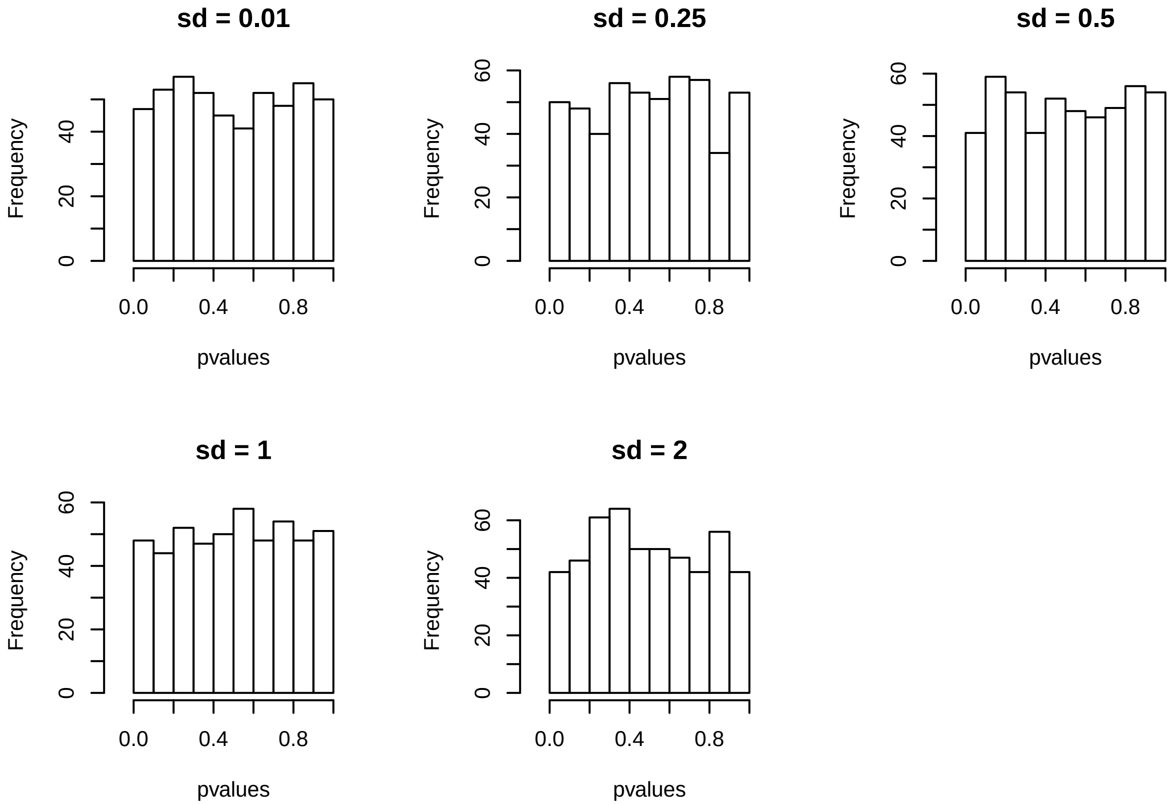

```
## Difference between population mean is  2
```



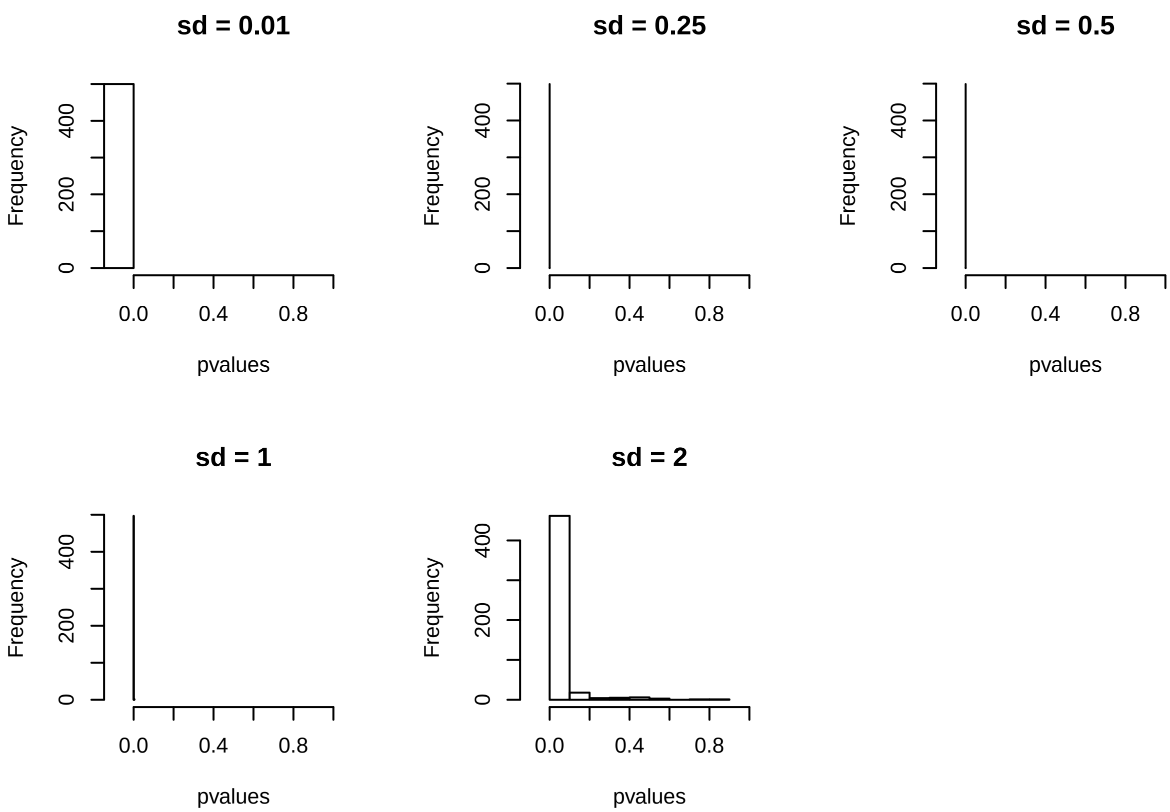
```
## Difference between population mean is  3
```



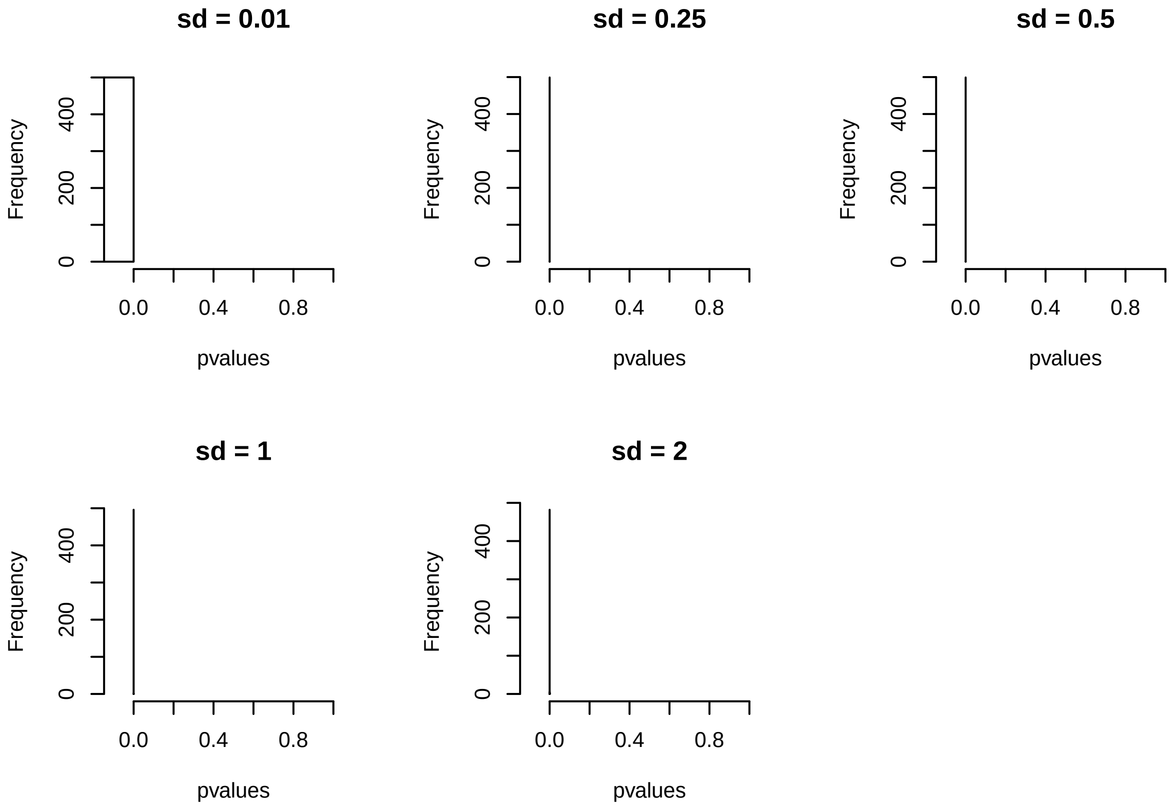

```
## Sample size n= 300
## Difference between population mean is  0
```



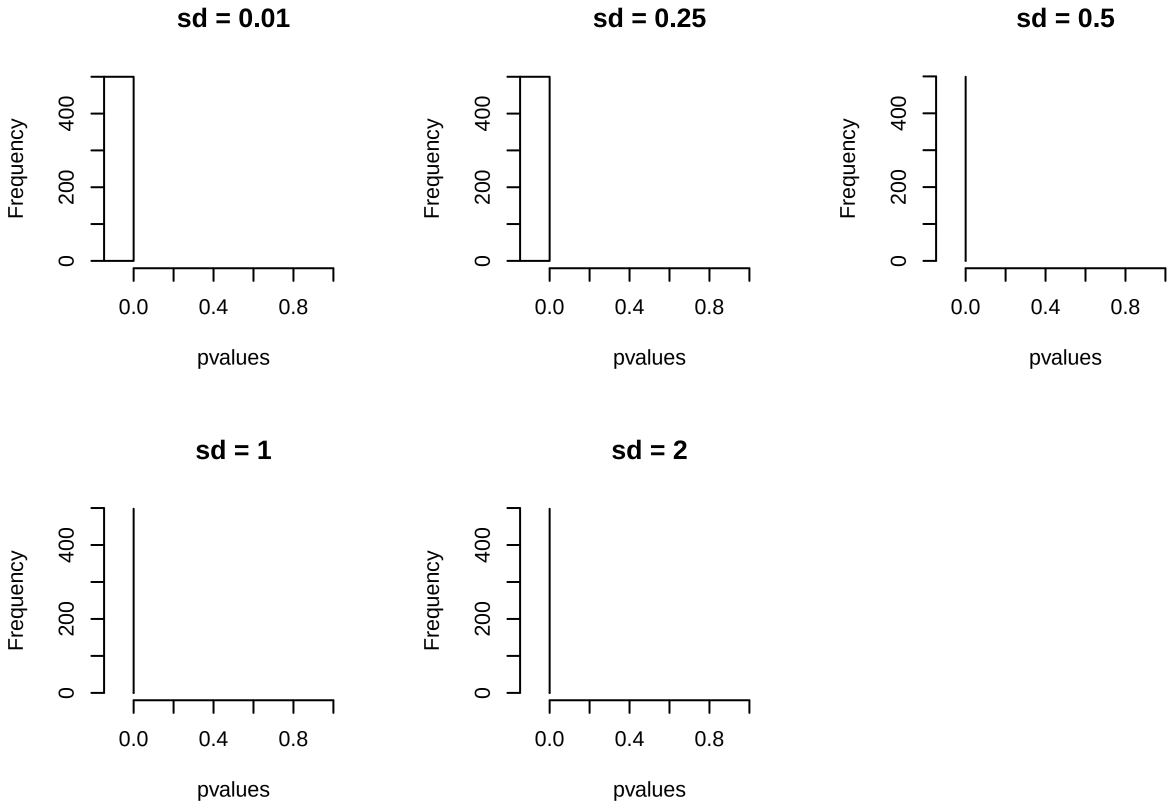

```
## Difference between population mean is  0.5
```



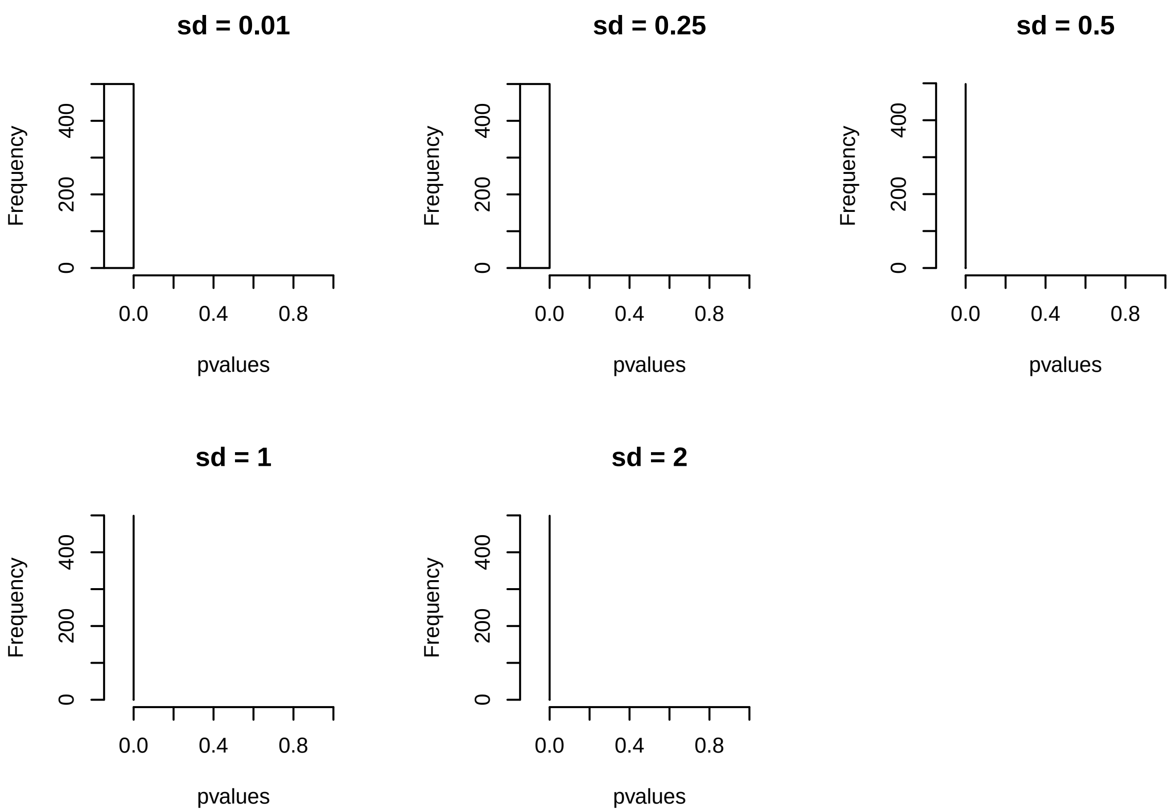
```
## Difference between population mean is  1
```




```
## Difference between population mean is  2
```




```
## Difference between population mean is  3
```



(c)

i. Given the same sample size and same true difference between the population means, as the standard deviation in the population increases, more p-values move toward the right (larger p-values) and it becomes less possible to detect the differencies.

ii. Given the same sample size and same standard deviation, as the true difference between the population means increases, more p-values move toward the left (smaller p-values) and it becomes more possible to detect the differencies.

iii. Given the same true difference between the population means and same standard deviation, as the sample size increases, more p-values move toward the left (smaller p-values) and it becomes more possible to detect the differencies.