Problem 6: Differing Distributions

Rakesh Vishwabrahmana

Question 1

Solution:

The given chunk below will implement a simulation of the Shapiro-Wilk normality test, which has a false positive rate of $\alpha = 0.05$ when testing with a *p-value* threshold of 0.05. We will use 1000 repetitions to find the rejection percentage of the null hypothesis that the random sample comes from a normal distribution.

```
n_sim <- 1000
pvalue_vec <- vector(length = n_sim)
for(i in 1 : n_sim){
  x <- rnorm(100)
  pvalue_vec[i] <- shapiro.test(x)$p.value
}
fp <- round(sum(pvalue_vec < 0.05) / n_sim, 2)
fp

## [1] 0.07
```

Therefore, when repeatedly running a Shapiro-Wilk test on data randomly generated from a normal distribution, the null hypothesis of normality is rejected $\sim 5\%$ of the time.

Question 2

Solution:

The given chunk below will estimate the statistical power $(1 - \beta)$ of the Shapiro-Wilk normality test for different alternative distribution scenarios when testing at $\alpha = 0.05$.

```
sw_test <- function(n_sam, null_dist = "unif") { # use "t" for t-distribution
if(null_dist != "unif") {
  x <- rt(n_sam, df = 3)
} else x <- runif(n_sam)
 p_value <- shapiro.test(x)$p.value
 return(p_value)
}
sw_power <- function(n_sim, n_sam, null_dist = "unif") {
p_values <- replicate(n_sim, sw_test(n_sam = n_sam, null_dist = null_dist))
est_power <- sum(p_values<0.05)/n_sim
return(est_power)
}
unif_10 <- sw_power(n_sim = 1000, n_sam = 10, null_dist = "unif")
unif_50 <- sw_power(n_sim = 1000, n_sam = 50, null_dist = "unif")
```

```r
unif_200 <- sw_power(n_sim = 1000, n_sam = 200, null_dist = "unif")
t_10 <- sw_power(n_sim = 1000, n_sam = 10, null_dist = "t")
t_50 <- sw_power(n_sim = 1000, n_sam = 50, null_dist = "t")
t_200 <- sw_power(n_sim = 1000, n_sam = 200, null_dist = "t")
power_vec <- c(unif_10,unif_50,unif_200,t_10,t_50,t_200)
n <- rep(c(10,50,200),times =2)
distribution <- c(rep("uniform",3),rep("t(dof=3)",3))
power_df <- data.frame(distribution, n, power_vec)
names(power_df) <- c("Distribution", "n", "$1-\\beta_{SW}$")
#Table 1 Summary: Statistical power of the Shapiro-Wilk normality test
power_df
```

```
##   Distribution   n $1-\\beta_{SW}$
## 1      uniform  10           0.085
## 2      uniform  50           0.742
## 3      uniform 200           1.000
## 4     t(dof=3)  10           0.184
## 5     t(dof=3)  50           0.642
## 6     t(dof=3) 200           0.991
```

Question 3

Solution:

The given chunk below will estimate the statistical power $(1 - \beta)$ of the Kolmogorov-Smirnov two-sample test for distinguishing data sampled from a uniform distribution from data sampled from a standard normal distribution when testing at $\alpha = 0.05$.

```r
ks_test <- function(n1, n2) {
 x1 <- runif(n1, -1.75, 1.75)
 x2 <- rnorm(n2)
 p_value <- ks.test(x1, x2)$p.value
 return(p_value)
}
ks_power <- function(n_sim, n1, n2) {
p_values <- replicate(n_sim, ks_test(n1 = n1, n2 = n2))
est_power <- sum(p_values<0.05)/n_sim
return(est_power)
}
ks_20 <- ks_power(n_sim = 1000, n1 = 20, n2 = 20)
ks_50 <- ks_power(n_sim = 1000, n1 = 50, n2 = 50)
ks_200 <- ks_power(n_sim = 1000, n1 = 200, n2 = 200)
ks_500 <- ks_power(n_sim = 1000, n1 = 500, n2 = 500)

power_vec <- c(ks_20, ks_50, ks_200, ks_500)
n1 <- c(20, 50, 200, 500)
n2 <- c(20, 50, 200, 500)
power_df <- data.frame(n1, n2, power_vec)
names(power_df) <- c("$n_1$", "$n_2$", "$1-\\beta_{KS}$")
```

```
#Table 2 Summary: Statistical power of the Kolmogorov-Smirnov two-sample test
power_df
```

```
##   $n_1$ $n_2$ $1-\\beta_{KS}$
## 1    20    20           0.041
## 2    50    50           0.082
## 3   200   200           0.230
## 4   500   500           0.633
```

Question 4

Solution:

The given chunk below will estimate the statistical power $(1-\beta)$ of the Anderson-Darling two-sample test for distinguishing data sampled from a uniform distribution from data sampled from a standard normal distribution when testing at $\alpha = 0.05$.

```
ad_test <- function(n1, n2) {
 x1 <- runif(n1, -1.75, 1.75)
 x2 <- rnorm(n2)
 p_value <- kSamples::ad.test(x1, x2)$ad[2,3]
 return(p_value)
}
ad_power <- function(n_sim, n1, n2) {
p_values <- replicate(n_sim, ad_test(n1 = n1, n2 = n2))
est_power <- sum(p_values<0.05)/n_sim
return(est_power)
}
ad_20 <- ad_power(n_sim = 1000, n1 = 20, n2 = 20)
ad_50 <- ad_power(n_sim = 1000, n1 = 50, n2 = 50)
ad_200 <- ad_power(n_sim = 1000, n1 = 200, n2 = 200)
ad_500 <- ad_power(n_sim = 1000, n1 = 500, n2 = 500)

power_vec <- c(ad_20, ad_50, ad_200, ad_500)
n1 <- c(20, 50, 200, 500)
n2 <- c(20, 50, 200, 500)
power_df <- data.frame(n1, n2, power_vec)
names(power_df) <- c("$n_1$", "$n_2$", "$1-\\beta_{AD}$")
#Table 3 Summary: Statistical power of the Anderson-Darling two-sample test
power_df
```

```
##   $n_1$ $n_2$ $1-\\beta_{AD}$
## 1    20    20           0.053
## 2    50    50           0.091
## 3   200   200           0.318
## 4   500   500           0.904
```

Question 5

Solution:

From **Table 2** and **Table 3**, we observed that the Anderson-Darling test produced better power than the Kolmogorov-Smirnov test. For example, the Kolmogorov-Smirnov and Anderson-Darling test power for (large) sample size 500 is 0.633 and 0.904, respectively. Although the Anderson-Darling test might slow, it is clear that the Anderson-Darling test is superior to the Kolmogorov-Smirnov test. Therefore, based on these results, we recommend the Anderson-Darling test.