

sixthlab

2025-11-07

```
library(MASS)
data(Boston)
X=Boston
nr=nrow(X)
mX=apply(X, 2, mean)#True mean
sdX=apply(X, 2, sd)#True sd
```

```
#Randomly selecting 40 samples, test if true mean is 6.2 (for rm column)
u0=6.2
alpha=0.05#significance level
r=X[,6]
size=40
s=sample(1:nr,size=size,replace = FALSE)
sam=r[s]
sam_mean=mean(sam)#sample Mean
sam_sd=sdX[6]/sqrt(size)#Assuming true sd is known
z=(sam_mean-u0)/(sam_sd)
#Two-tailed test; Probability of getting as extreme or more extreme values assuming null hypothesis is true
p_value=2*pnorm(abs(z), mean = 0, sd = 1, lower.tail = FALSE)
print(p_value)
```

```
##          rm
## 0.554106
```

```
if(p_value<alpha){
  sprintf("Reject Null Hypothesis:u0=%f",u0)
}else{
  sprintf("Accept Null Hypothesis:u0=%f",u0)
}
```

```
## [1] "Accept Null Hypothesis:u0=6.200000"
```

```

#Randomly selecting 40 samples and test if true proportion is 0.5 (for binary indicator of crime column)
p0=0.5#Null Hypothesis
alpha=0.05#significance level
size=40#sample size
high_crime=(X$crim>median(X$crim))
true_mean=mean(high_crime)
s=sample(1:nr,size=size,replace = FALSE)
sam=high_crime[s]
sam_p=mean(sam)
sam_sd=sqrt((p0*(1-p0))/size)#True sd assuming Null Hypothesis is True
z=(sam_p-p0)/sam_sd
#Two-tailed test; Probability of getting as extreme or more extreme values assuming null hypothesis is true
p_value=2*pnorm(abs(z), mean = 0, sd = 1, lower.tail = FALSE)
print(p_value)

```

```
## [1] 0.7518296
```

```

if(p_value<alpha){
  sprintf("Reject Null Hypothesis:p0=%f",p0)
}else{
  sprintf("Accept Null Hypothesis:p0=%f",p0)
}

```

```
## [1] "Accept Null Hypothesis:p0=0.500000"
```

```

#Randomly selecting 20 samples and test if means of two sets are same
d0=0
alpha=0.05
size=20
#Form two sets
s1=X$medv[X$chas==1]#Locations nearby charles river
s2=X$medv[X$chas==0]#Locations not nearby charles river
s1_tm=mean(s1)
s2_tm=mean(s2)
s1_sd=sd(s1)
s2_sd=sd(s2)
sprintf("Set 1:True Mean:%f,True sd:%f",s1_tm,s1_sd)

```

```
## [1] "Set 1:True Mean:28.440000,True sd:11.816643"
```

```
sprintf("Set 2:True Mean:%f,True sd:%f",s2_tm,s2_sd)
```

```
## [1] "Set 2:True Mean:22.093843,True sd:8.831362"
```

```
s=sample(1:length(s1),size=size,replace =FALSE)
sam1=s1[s]
s=sample(1:length(s2),size=size,replace=FALSE)
sam2=s2[s]
#Find mean of samples
sam1_mean=mean(sam1)
sam2_mean=mean(sam2)
sam1_sd=s1_sd/sqrt(size)
sam2_sd=s2_sd/sqrt(size)
#Total sd under H0 (m1-m2=0)
total_sd=sqrt(sam1_sd^2+sam2_sd^2)
z=(sam1_mean-sam2_mean)/total_sd#Assuming true mean of difference is zero
#Two-tailed p-value
p_value=2*pnorm(abs(z),mean=0,sd=1,lower.tail = FALSE)
print(p_value)
```

```
## [1] 0.0009215427
```

```
if(p_value<alpha){
  sprintf("Reject Null Hypothesis that u0=u1")
}else{
  sprintf("Accept Null Hypothesis that u0=u1")
}
```

```
## [1] "Reject Null Hypothesis that u0=u1"
```

```
#Randomly selecting 20 samples and test if proportions of two sets are same
dp0=0
alpha=0.05
size=20
high_tax=(X$tax>median(X$tax))#Indicator variable
#form 2 sets
s1=high_tax[X$chas==1]
s2=high_tax[X$chas==0]
s1_tm=mean(s1)
s2_tm=mean(s2)
sprintf("Set 1:True Mean:%f",s1_tm)
```

```
## [1] "Set 1:True Mean:0.428571"
```

```
sprintf("Set 2:True Mean:%f",s2_tm)
```

```
## [1] "Set 2:True Mean:0.498938"
```

```

s=sample(1:length(s1),size=size,replace=FALSE)
sam1=s1[s]
s=sample(1:length(s2),size=size,replace=FALSE)
sam2=s2[s]
sam1_mean=mean(sam1)
sam2_mean=mean(sam2)
p=mean(c(sam1,sam2))#Combined Mean
sam_sd=sqrt(p*(1-p)*(2/size))
z=(sam1_mean-sam2_mean)/sam_sd
p_value=2*pnorm(abs(z),mean=0,sd=1,lower.tail = FALSE)
print(p_value)

```

```
## [1] 0.184844
```

```

if(p_value<alpha){
  sprintf("Reject Null Hypothesis that p0=p1")
}else{
  sprintf("Accept Null Hypothesis that p0=p1")
}

```

```
## [1] "Accept Null Hypothesis that p0=p1"
```

```

#Randomly selecting 40 samples and test if two variables are uncorrelated
r=0#Null Hypotheis
alpha=0.05
size=40
X1=X$rm
X2=X$medv
true_r=cor(X1,X2)
print(true_r)

```

```
## [1] 0.6953599
```

```

s=sample(1:nr,size=40,replace=FALSE)
sam1=X1[s]
sam2=X2[s]
sam_r=cor(sam1,sam2)
z_r=0.5*log((1+sam_r)/(1-sam_r))
#print(z_r)
sd=1/sqrt(size-3)
z=z_r/sd
p_value=pnorm(abs(z),mean=0,sd=1,lower.tail = FALSE)
print(p_value)

```

```
## [1] 3.361524e-07
```

```
if(p_value<alpha){  
  sprintf("Reject Null Hypothesis that Variables are Uncorrelated")  
}  
else{  
  sprintf("Accept Null Hypothesis that Variables are Uncorrelated")  
}  
}
```

```
## [1] "Reject Null Hypothesis that Variables are Uncorrelated"
```

```
#X^2 test for Independence  
high_crime=(X$crim>median(X$crim))  
high_tax=(X$tax>median(X$tax))  
alpha=0.05  
test=chisq.test(high_crime,high_tax)#Chi square test for checking independence  
#chisquare assumes null hypothesis as two variables are independent  
p_value=test$p.value  
print(p_value)
```

```
## [1] 3.261963e-22
```

```
if(p_value<alpha){  
  sprintf("Reject Null Hypothesis that two variables are independent")  
}  
else{  
  sprintf("Accept Null Hypothesis that two variables are independent")  
}  
}
```

```
## [1] "Reject Null Hypothesis that two variables are independent"
```

```
df=as.data.frame(cbind(high_crime,high_tax))
```