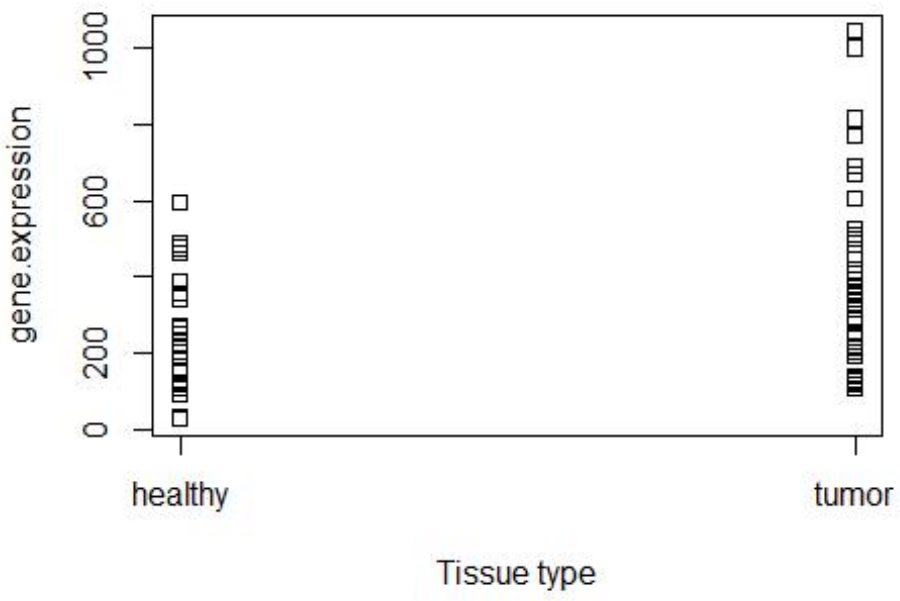
stat3701 hw4

Dongnan Liu

2024-03-17

```
#1(a)
#For the two-independent-samples model,we assume about these measurement
 are
#independent. We can assume: for the tumor samples, we have x1, . . . ,
x40
#are a realization of X1, . . . , X40, with unknown mean μ1 and standard
 deviation σ.
#For the healthy samples y1, . . . , y22 are a realization of Y1, . . . ,
 Y22,
#with unknown mean μ2 and standard deviation σ.
#If these assumptions are true, then the probability
#distribution of the random variable for which the first healthy tissue'
s
#gene expression measurement of 202.90000 is assumed to be a realization
#iid from some distribution of Y1 with unknown mean μ and standard devia
tion σ.
gene=read.table("C:/Users/DELL/Desktop/gene.txt")
#1(b)
#We can compare the response for the tumorous tissues to the healthy
#one with a stripchart.
stripchart(gene.expression ~ tissue.type,data=gene, vertical=TRUE, xlab
="Tissue type")
```




```
#We can that in the tumor tissues gene expressions, there are more varia
bility, so this is not
#a reasonable assumption for these. The distribution of the response don
't have the same
#unknown standard deviation for both levels of the categorical explanato
ry variable
#data.
#1(c)
x.list=gene$gene.expression[gene$tissue.type=="healthy"]
y.list=gene$gene.expression[gene$tissue.type=="tumor"]
xbar=mean(x.list)
n1=length(x.list)
ybar=mean(y.list)
n2=length(y.list)
sp=(sum((x.list-xbar)^2) + sum((y.list-ybar)^2))/(n1+n2-2)
t.stat=(xbar-ybar)/sqrt(sp*(1/n1 + 1/n2))
(pval=2*pt(-abs(t.stat), df=n1+n2-2))
```

```
## [1] 0.01436113
```

```
#We can see that the p-value(0.014) is greater than 0.01 so we can concl
ude that
#there is no statistical evidence at the 1% significance level
#that distribution of the expression of this gene is associated with the
 tissue type.
#1(d)
x.list=log(gene$gene.expression[gene$tissue.type=="healthy"])
```

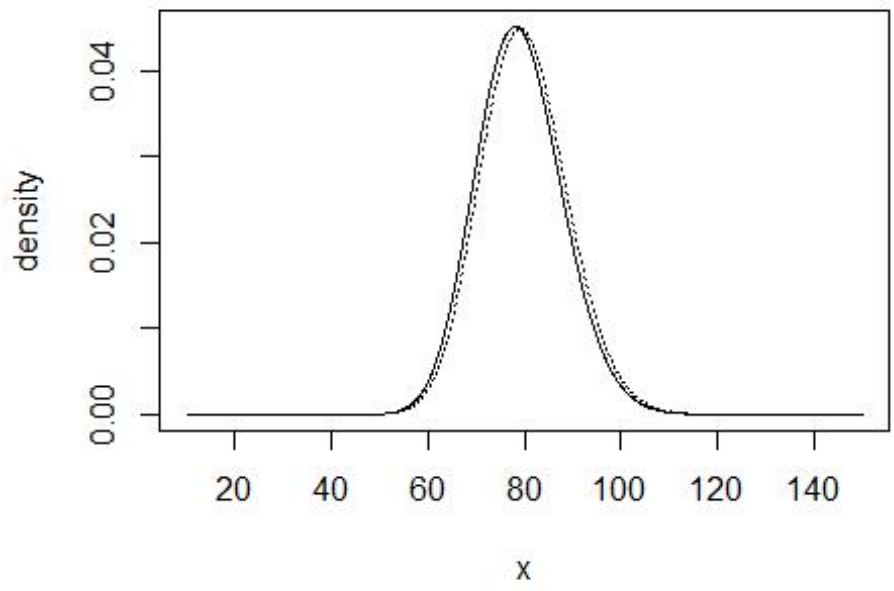
```r
y.list=log(gene$gene.expression[gene$tissue.type=="tumor"])
xbar=mean(x.list)
n1=length(x.list)
ybar=mean(y.list)
n2=length(y.list)
sp=(sum((x.list-xbar)^2) + sum((y.list-ybar)^2))/(n1+n2-2)
t.stat=(xbar-ybar)/sqrt(sp*(1/n1 + 1/n2))
(pval=2*pt(-abs(t.stat), df=n1+n2-2))
```

```
## [1] 0.003416748
```

```r
c(sd(x.list), sd(y.list))
```

```
## [1] 0.8308524 0.5842001
```

```r
#Now the p-value is 0.0034, it is smaller than 0.01 so we can conclude t
hat it is
#reasonable to sya that we have evidence at the 1% significance level th
at distribution of the natural logarithm of the expression of this gene
is associated with the tissue type. We can also see that the distributio
n of the response has the same variance after comparing the standard dea
viation of x.list and y.list.
#deviation for both levels of the categorical explanatory variable.
#2(a)
x.seq=seq(from=10, to=150, length.out=1e3)
x.den.vals=dgamma(x=x.seq, shape=79, scale=1)
plot(x.seq, x.den.vals, type="l", xlab="x", ylab="density")
y.den.vals=dgamma(x=x.seq, shape=80, scale=1)
lines(x.seq, y.den.vals, lty=3)
```




```r
#2(b)
set.seed(3701)
reps=3e5
x.list=rgamma(n=reps, shape=79, scale=1)
t.test((x.list - 79)^2 , conf.level=0.995)$conf.int[1:2]
```

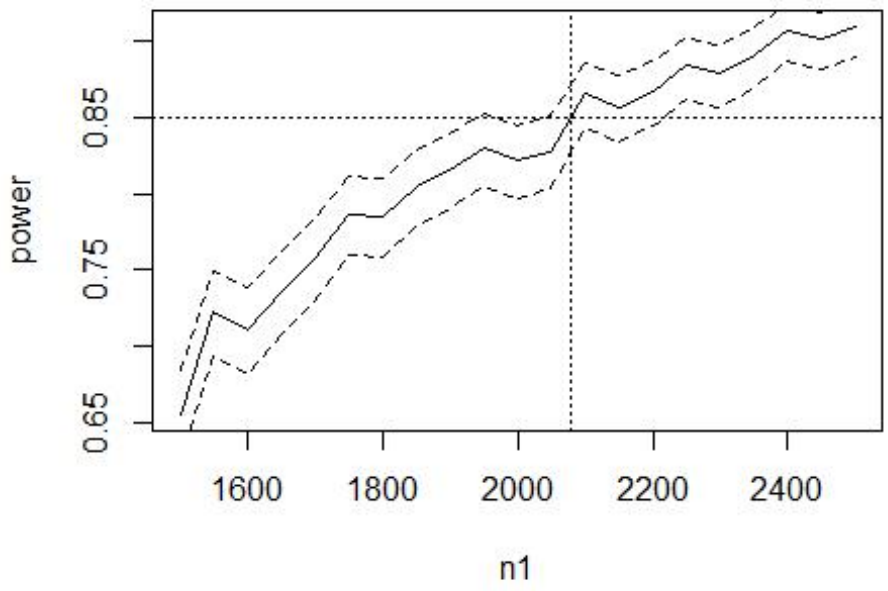
```
## [1] 78.64221 79.81225
```

```r
#The true value 79 is in this interval.
#2(c)
set.seed(3701)
reps=3e5
x.list=rgamma(n=reps, shape=80, scale=1)
t.test((x.list - 80)^2 , conf.level=0.995)$conf.int[1:2]
```

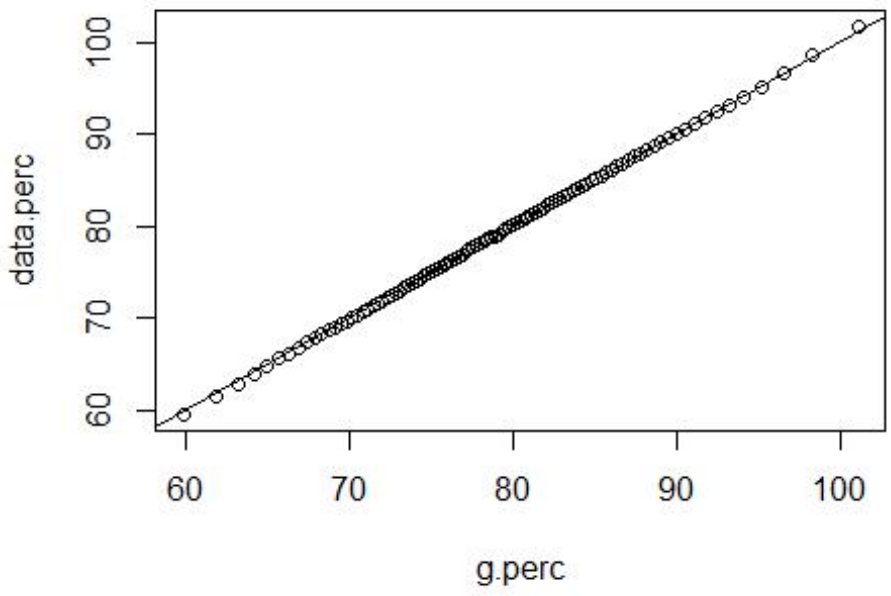
```
## [1] 79.63830 80.82303
```

```r
#The true value 80 is in this interval.
#2(d)
pval.dist=function(n1, n2, reps=1e4) {
  p.list=numeric(reps)
  for(r in 1:reps)
  {
    x.list=rgamma(n1, shape=79, scale=1)
    y.list=rgamma(n2, shape=80, scale=1)
    xbar=mean(x.list)
    ybar=mean(y.list)
```

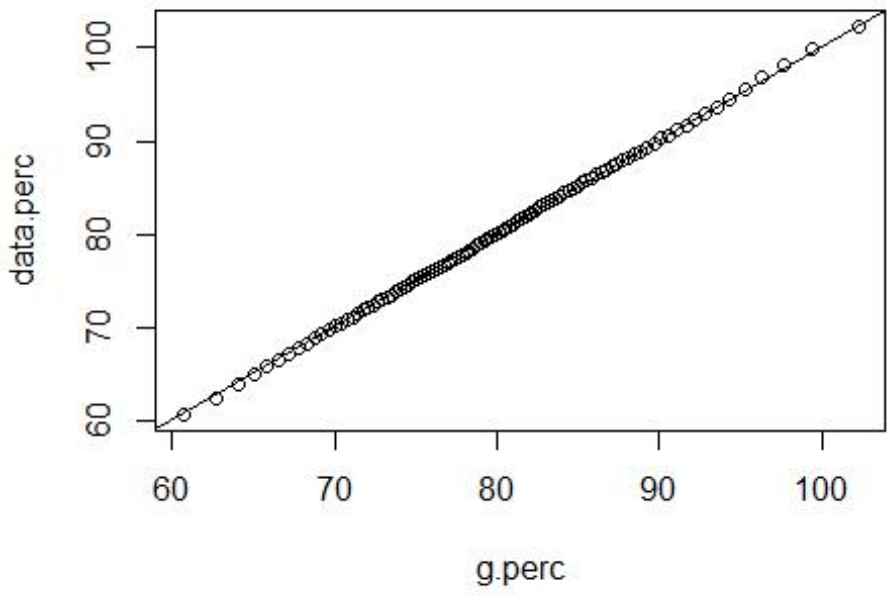
```
    residuals=c(x.list-xbar, y.list-ybar)
    sp.sq=sum(residuals^2)/(n1+n2-2)
    t=(xbar-ybar)/sqrt(sp.sq* (1/n1+1/n2))
    p.list[r] = 2*pt(-abs(t), n1+n2-2) }
    return(p.list)
}
set.seed(3701)
reps=1e3
n.list=seq(from=1500, to=2500, by=50)
est.power=numeric(length(n.list))
LB.list=numeric(length(n.list))
UB.list=numeric(length(n.list))
for(j in 1:length(n.list))
{
  pvals=pval.dist(n1=n.list[j], n2=n.list[j], reps=reps)
  est.power[j]=mean(pvals < 0.01)
  bounds=binom.test(x=sum(pvals<0.01), n=reps, conf.level=0.95)$conf.in
t[1:2]
  LB.list[j]=bounds[1]
  UB.list[j]=bounds[2]
}
plot(n.list, est.power, t="l",
     xlab="n1", ylab="power")
lines(n.list, LB.list, lty=2)
lines(n.list, UB.list, lty=2)
abline(h=0.85, lty=3)
abline(v=2080, lty=3)
```




```
#We can see that when power is 85%, the n1 and n2 need to be approximate
ly equal to 2080.
#3(a)
set.seed(3701)
reps=1e4
t.list=rgamma(reps, shape=60, scale=1)
u.list=rgamma(reps, shape=19, scale=1)
x.list=t.list+u.list
probs=seq(from=0.01, to=0.99, by=0.01)
data.perc=quantile(x.list, probs)
g.perc=qgamma(probs, shape=79, scale=1)
plot(g.perc, data.perc)
abline(0,1)
```




```r
#We can that it follows y=x, so the sample percentiles of these realizat
ions
#aligns with the percentiles of the Gamma(79, 1) distribution.
#3(b)
set.seed(3701)
reps=1e4
t.list=rgamma(reps, shape=60, scale=1)
v.list=rgamma(reps, shape=20, scale=1)
x.list=t.list+v.list
probs=seq(from=0.01, to=0.99, by=0.01)
data.perc=quantile(x.list, probs)
g.perc=qgamma(probs, shape=80, scale=1)
plot(g.perc, data.perc)
abline(0,1)
```

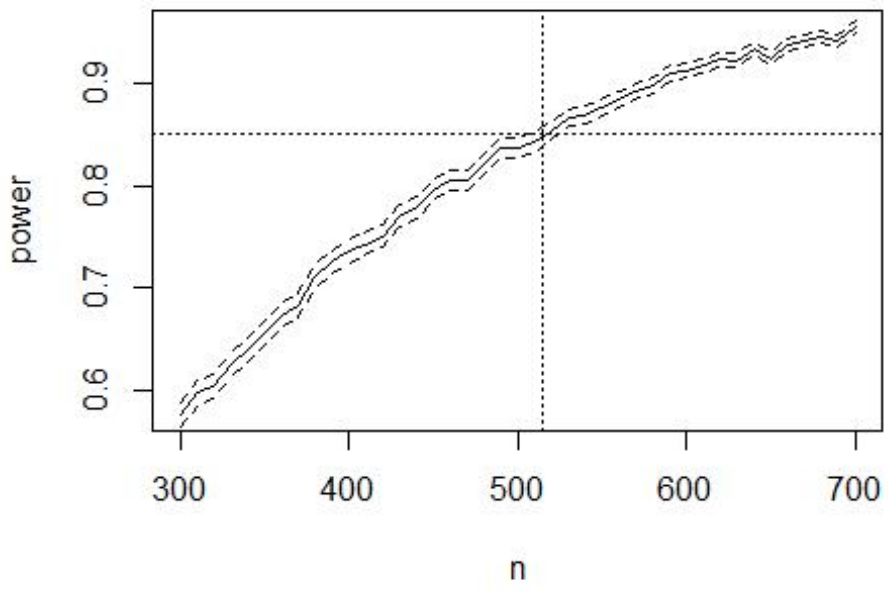



```
#We can that it follows y=x, so the sample percentiles of these realizat
ions
#aligns with the percentiles of the Gamma(80, 1) distribution.
#3(c)
gen.paired.data=function(n, shape1, shape2,shape3, scale)
{
  t.list=rgamma(n, shape=shape1, scale=scale)
  u.list=rgamma(n, shape=shape2, scale=scale)
  v.list=rgamma(n, shape=shape3, scale=scale)
  x.list=t.list+u.list
  y.list=t.list+v.list
  return(list(x.list=x.list, y.list=y.list))
}
set.seed(3701)
reps=1e4
shape1=60; shape2=19; shape3=20; scale=1
pairdata=gen.paired.data(n=reps, shape1=shape1, shape2=shape2,shape3=sh
ape3, scale=scale)
t.test(((pairdata$x.list-79)*(pairdata$y.list-80))/sqrt(79*80), conf.le
vel=0.99)$conf.int[1:2]

## [1] 0.7333425 0.8001369

#3(d)
pval.dist=function(n, shape1, shape2,shape3, scale) {
  p.list=numeric(reps)
  for(r in 1:reps)
```

```
  {
    pdat=gen.paired.data(n=n, shape1=60, shape2=19,shape3=20, scale=scal
e)
    z.list=pdat$x.list-pdat$y.list
    t=(mean(z.list)- 0)/(sd(z.list)/sqrt(n))
    p.list[r] = 2*pt(-abs(t), n-1)
    }
    return(p.list)
}
set.seed(3701)
reps=6e3
n.list=seq(from=300, to=700, by=10)
est.power=numeric(length(n.list))
LB.list=numeric(length(n.list))
UB.list=numeric(length(n.list))
for(j in 1:length(n.list))
{
  pvals=pval.dist(n=n.list[j], shape1=60, shape2=19,shape3=20,scale=1)
  est.power[j]=mean(pvals < 0.01)
  bounds=binom.test(x=sum(pvals<0.01), n=reps, conf.level=0.95)$conf.in
t[1:2]
  LB.list[j]=bounds[1]
  UB.list[j]=bounds[2]
}
plot(n.list, est.power, t="l",
     xlab="n", ylab="power")
lines(n.list, LB.list, lty=2)
lines(n.list, UB.list, lty=2)
abline(h=0.85, lty=3)
abline(v=515, lty=3)
```




```
#We can see in the plot that when the power=0.85, the n is approxiamtely
 equal to 515.
```