# STAC51 Assignment 1

Joseph Wang

2023-01-20

## Question 1

### a)

The PMF is

$$f_{\bar{Y}}(\bar{y}) = \frac{n!}{y_1! \dots y_k!} \pi_1^{y_1} \dots \pi_k^{y_k}$$

Thus is the multivariate moment generating function is

$$\begin{aligned}
M_{\bar{Y}}(\bar{t}) &= E(e^{\vec{t}'\bar{y}}) \\
&= \Sigma_{y_1}^{n} \Sigma_{y_2}^{n-y_1} \dots \Sigma_{y_k}^{n-y_1 \dots -y_{k-1}} e^{t_1 y_1 + \dots + t_k y_k} \frac{n!}{y_1! \dots y_k!} \pi_1^{y_1} \dots \pi_k^{y_k} \\
&= \Sigma_{y_1}^{n} \Sigma_{y_2}^{n-y_1} \dots \Sigma_{y_k}^{n-y_1 \dots -y_{k-1}} \frac{n!}{y_1! \dots y_k!} (e^{t_1}\pi_1)^{y_1} \dots (e^{t_k}\pi_k)^{y_k} \\
&= (\Sigma_i^k e^{t_i} \pi_i)^n
\end{aligned}$$

### b)

The moment generating function for the ith var is

$$\begin{aligned}
M_{Y_i}(t) &= M_{\bar{Y}}(0, \dots, t, \dots, 0) \\
&= (\pi_1 + \dots + e^t \pi_i + \dots \pi_k)^n
\end{aligned}$$

The mean is

$$\begin{aligned}
E(Y_i) &= \frac{\partial}{\partial t} M_{Y_i}(0) \\
&= n(\pi_1 + \dots + e^t \pi_i + \dots \pi_k)^{n-1} e^t \pi_i |_0 \\
&= n\pi_i
\end{aligned}$$

## c)

The second moment is

$$
\begin{aligned}
E(Y_i^2) &= \frac{\partial^2}{\partial t^2} M_{Y_i}(0) \\
&= \frac{\partial}{\partial t} n(\pi_1 + \dots + e^t\pi_i + \dots \pi_k)^{n-1} e^t \pi_i + e^t\pi_i * n(n-1)(\pi_1 + \dots + e^t\pi_i + \dots \pi_k)^{n-2} e^t \pi_i |_0 \\
&= n\pi_i + n^2\pi_i^2 - n\pi_i^2
\end{aligned}
$$

The variance is

$$
\begin{aligned}
Var(Y_i) &= E(Y_i^2) - E(Y_i)^2 \\
&= n\pi_i + n^2\pi_i^2 - n\pi_i^2 - n^2\pi_i^2 \\
&= n\pi_i(1 - \pi_i)
\end{aligned}
$$

## d)

The joint moment generating function for the ith and jth var is

$$
\begin{aligned}
M_{Y_iY_j}(t_i, t_j) &= M_{\bar{Y}}(0, \dots, t_i, \dots, t_j, \dots, 0) \\
&= (\pi_1 + \dots + e^{t_i}\pi_i + \dots + e^{t_j}\pi_j + \dots \pi_k)^n
\end{aligned}
$$

The expected value of $Y_iY_j$ is

$$
\begin{aligned}
E(Y_iY_j) &= \frac{\partial^2}{\partial t_i t_j} M_{Y_iY_j}(0, 0) \\
&= \frac{\partial}{\partial t_j} n(\pi_1 + \dots + e^{t_i}\pi_i + \dots + e^{t_j}\pi_j + \dots \pi_k)^{n-1} e^{t_i}\pi_i |_0 \\
&= n(n-1)(\pi_1 + \dots + e^{t_i}\pi_i + \dots + e^{t_j}\pi_j + \dots \pi_k)^{n-2} e^{t_i}\pi_i e^{t_j}\pi_j |_0 \\
&= n(n-1)\pi_i\pi_j
\end{aligned}
$$

The covariance of $Y_iY_j$ is

$$
\begin{aligned}
Cov(Y_i, Y_j) &= E(Y_iY_j) - E(Y_i)E(Y_j) \\
&= n(n-1)\pi_i\pi_j - n^2\pi_i\pi_j \\
&= -n\pi_i\pi_j
\end{aligned}
$$

### e)

Given $c = 2$, $1 = \pi_i + \pi_j$

$$
\begin{aligned}
Cor(Y_i, Y_j) &= \frac{Cov(Y_i, Y_j)}{\sqrt{Var(Y_i)Var(Y_j)}} \\
&= \frac{-n\pi_i\pi_j}{\sqrt{n\pi_i(1-\pi_i)n\pi_j(1-\pi_j)}} \\
&= \frac{-\pi_i\pi_j}{\sqrt{\pi_i\pi_j(1-\pi_i-\pi_j+\pi_i\pi_j)}} \\
&= -1
\end{aligned}
$$

This makes sense as if the number of successes one category increases, the number of successes in the other category has to decrease. The change is 1:1 and thus has a linear relationship of factor 1.

## Question 2

```
y = 5
n = 30
pihat = y/n
alpha = 0.1
```

### a)

Test P-values

```
pi0 = 0.1
yfit = pi0*n

sewald= sqrt(pihat*(1-pihat)/n)
sescore = sqrt(pi0*(1-pi0)/n)

score = (pihat - pi0)/sescore

wald = (pihat - pi0)/sewald

lrt = 2*(y*log(y/yfit) + (n-y)*log((n-y)/(n-yfit)))


2*pnorm(score, lower.tail = FALSE)
```

```
## [1] 0.2235429
```

```
2*pnorm(wald, lower.tail = FALSE)
```

```
## [1] 0.3271869
```

```r
pchisq(lrt,1,lower.tail = FALSE)
```

```
## [1] 0.2616124
```

## b)

Wald CI

```r
z = qnorm(1-alpha/2)

c(pihat - z*sewald, pihat + z*sewald)
```

```
## [1] 0.05474855 0.27858478
```

## c)

Score CI

```r
c(((n*pihat+z^2/2)-z*sqrt(n*pihat*(1-pihat)+z^2/4))/(n+z^2),((n*pihat+z^2/2)+z*sqrt(n*pihat*(1-pihat)+z
```

```
## [1] 0.08356237 0.30492050
```

## d)

Agresti-Coull

```r
pis = (y+z^2/2)/(n+z^2)
ns = n+z^2

c(pis-z*sqrt(pis*(1-pis)/ns), pis+z*sqrt(pis*(1-pis)/ns))
```

```
## [1] 0.08045514 0.30802774
```

## e)

```r
prop.test(y,n,p=0.1, correct=F)
```

```
## Warning in prop.test(y, n, p = 0.1, correct = F): Chi-squared approximation may
## be incorrect
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  y out of n, null probability 0.1
## X-squared = 1.4815, df = 1, p-value = 0.2235
## alternative hypothesis: true p is not equal to 0.1
```

```
## 95 percent confidence interval:
##  0.07336542 0.33564351
## sample estimates:
##         p
## 0.1666667
```

### f)

```r
library(binom)

binom.confint(x=5,n=30,conf.level=0.90, methods="all")
```

```
##           method x  n      mean      lower     upper
## 1  agresti-coull 5 30 0.1666667 0.08045514 0.3080277
## 2     asymptotic 5 30 0.1666667 0.05474855 0.2785848
## 3          bayes 5 30 0.1774194 0.06782250 0.2829193
## 4        cloglog 5 30 0.1666667 0.07379893 0.2917865
## 5          exact 5 30 0.1666667 0.06805557 0.3189712
## 6          logit 5 30 0.1666667 0.08201730 0.3092501
## 7         probit 5 30 0.1666667 0.07848132 0.3017123
## 8        profile 5 30 0.1666667 0.07582948 0.2964459
## 9            lrt 5 30 0.1666667 0.07579682 0.2964405
## 10     prop.test 5 30 0.1666667 0.06303555 0.3545108
## 11        wilson 5 30 0.1666667 0.08356237 0.3049205
```

## Question 3

The likelihood function is $l(\theta|y) = \binom{n}{y}(\theta)^y(1-\theta)^{n-y}$. With n = 30 and y = 5.

### a)

Thus the likelihood under $\pi_0 = 0.1$ is

$$l(0.1|5) = \binom{30}{5}(0.1)^5(0.9)^{25}$$

### b)

The MLE is the maximized likelihood over all possible values. $\hat{\pi} = \frac{y}{n} = \frac{5}{30} = \frac{1}{6}$.

$$l(\frac{1}{6}|5) = \binom{30}{5}(\frac{1}{6})^5(\frac{1}{6})^{25}$$

### d)

```
qchisq(0.9, df=1)
```

```
## [1] 2.705543
```

### e)

```
binom.confint(x=5,n=30,conf.level=0.9, methods="lrt")
```

```
##   method x  n      mean      lower     upper
## 1    lrt 5 30 0.1666667 0.07579682 0.2964405
```

# Question 4

```
N = 100000
n = 25
pi = 0.06
count = 0
for(i in 1:N){
  z = rbinom(n,1,pi)
  y = sum(z)
  pi_hat = y/n

  l = pi_hat - 1.96*sqrt(pi_hat*(1-pi_hat)/n)
  t = pi_hat + 1.96*sqrt(pi_hat*(1-pi_hat)/n)

  if(l<= pi & pi <= t){
    count = count + 1
  }
}

count/N
```

```
## [1] 0.78522
```

# Question 6

```
freq = c(229, 221, 93, 35, 7, 1)
sum(freq)
```

```
## [1] 586
```

```r
f = function(x){
  return(x+1)
}


(lapply(freq, f))
```

```
## [[1]]
## [1] 230
##
## [[2]]
## [1] 222
##
## [[3]]
## [1] 94
##
## [[4]]
## [1] 36
##
## [[5]]
## [1] 8
##
## [[6]]
## [1] 2
```