Hypothesis 2

Anuhya B S

2/27/2022

2. Authors can be classified based on years of experience or number of books published. Highly experienced authors can definitely be expected to have larger rating counts. From a publisher point of view, it would be useful to assess if publishing works of new authors (who typically have 0-3 works published) would look promising.

The author work experience has been divided into three categories(based on log2 (author work count) distribution):
1. Newbie : Work count of authors less than 16
2. Average : Work count of authors between 16 and 256
3. Experienced : Work count of authors more than 256

$$H_0: \mu_N = \mu_A = \mu_E$$

```
master_data = read.csv('master_dataset.csv')

colnames(master_data)

##  [1] "X"                      "book"                   "author"
##  [4] "rating_count"           "page_count"             "genre"
##  [7] "is_volume"              "author_sex"             "author_work_count"
## [10] "author_avg_rating"      "log2_author_work_count" "author_exp"
## [13] "book_size"              "genre_category"
```

Creating data subsets:

```
new_exp <- subset(master_data, author_exp == 'newbie' | author_exp ==
'experienced')
new_avg <- subset(master_data, author_exp == 'newbie' | author_exp ==
'average')
avg_exp <- subset(master_data, author_exp == 'average' | author_exp ==
'experienced')
```

2-tailed Z-test to compare mean rating count b/w New-comers and Average authors:

```
m = with(new_avg,tapply(rating_count, author_exp, mean))
s = with(new_avg,tapply(rating_count, author_exp, sd))
n = with(new_avg,tapply(rating_count, author_exp, length))
data.frame(m,s,n)
```

```
##                  m         s     n
## average 22652.46 132766.15 23612
## newbie  15532.21  93148.18  5830
```

```
z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(1-pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)
```

```
##                 z            p round_p
## average 4.762963 1.907705e-06       0
```

Results: The z-score = 4.762963 and p-value = 1.907705e-06.

Since the p-vale is less than 0.05, we have sufficient evidence to reject the null hypothesis. This means the average rating count of new comer authors is not equal to average rating counts of average work count authors.

2-tailed Z-test to compare mean rating count b/w New-comers and Experienced authors:

```
m = with(new_exp,tapply(rating_count, author_exp, mean))
s = with(new_exp,tapply(rating_count, author_exp, sd))
n = with(new_exp,tapply(rating_count, author_exp, length))
data.frame(m,s,n)
```

```
##                     m         s    n
## experienced 44184.51 189430.91 4696
## newbie      15532.21  93148.18 5830
```

```
z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(1-pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)
```

```
##                     z p round_p
## experienced 9.482697 0       0
```

Results: The z-score = 9.482697 and p-value = *0.

Since the p-vale is less than 0.05, we have sufficient evidence to reject the null hypothesis. This means the average rating count of new comer authors is not equal to average rating counts of experienced authors.

2-tailed Z-test to compare mean rating count b/w Average and Experienced authors:

```
m = with(avg_exp,tapply(rating_count, author_exp, mean))
s = with(avg_exp,tapply(rating_count, author_exp, sd))
n = with(avg_exp,tapply(rating_count, author_exp, length))
data.frame(m,s,n)
```

```
##                      m        s     n
## average     22652.46 132766.2 23612
## experienced 44184.51 189430.9  4696
```

```
z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)
```

```
##                 z            p round_p
## average -7.434607 1.048789e-13       0
```

Results: The z-score = -7.434607 and p-value = 1.048789e-13.

Since the p-vale is less than 0.05, we have sufficient evidence to reject the null hypothesis. This means the average rating count of average work count authors is not equal to average rating counts of experienced authors.

ANOVA TEST:

```
summary(aov(master_data$rating_count~as.factor(master_data$author_exp)))
```

```
##                                      Df    Sum Sq   Mean Sq F value Pr(>F)
## as.factor(master_data$author_exp)     2 2.369e+12 1.184e+12   63.64 <2e-16
***
## Residuals                         34135 6.352e+14 1.861e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pairwise test with unequal variances:

```
p12 =
t.test(master_data$rating_count[master_data$author_exp=='newbie'],master_data
$rating_count[master_data$author_exp=='average'],var.equal = F)
p12
```

```
##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "newbie"] and
master_data$rating_count[master_data$author_exp == "average"]
## t = -4.763, df = 12375, p-value = 1.929e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10050.525  -4189.973
## sample estimates:
## mean of x mean of y
##  15532.21  22652.46
```

```
p13 =
t.test(master_data$rating_count[master_data$author_exp=='newbie'],master_data
```

```
$rating_count[master_data$author_exp=='experienced'],var.equal = F)
p13
```

```
##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "newbie"] and
master_data$rating_count[master_data$author_exp == "experienced"]
## t = -9.4827, df = 6503.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -34575.51 -22729.10
## sample estimates:
## mean of x mean of y
##  15532.21  44184.51
```

```
p23 =
t.test(master_data$rating_count[master_data$author_exp=='average'],master_dat
a$rating_count[master_data$author_exp=='experienced'],var.equal = F)
p23
```

```
##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "average"] and
master_data$rating_count[master_data$author_exp == "experienced"]
## t = -7.4346, df = 5646.4, p-value = 1.205e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27209.71 -15854.41
## sample estimates:
## mean of x mean of y
##  22652.46  44184.51
```

Future work : Need to apply Bonferroni correction

Comparing the avg rating count for male and female authors having different work counts

Comparing average rating counts of male newbies and female newbies

```
m = with(master_data,tapply(rating_count[author_exp=='newbie'],
author_sex[author_exp=='newbie'], mean))
s = with(master_data,tapply(rating_count[author_exp=='newbie'],
author_sex[author_exp=='newbie'], sd))
n = with(master_data,tapply(rating_count[author_exp=='newbie'],
author_sex[author_exp=='newbie'], length))
data.frame(m,s,n)
```

```
##                m        s    n
## female 16665.66 105256.2 3677
## male   13596.44  67610.2 2153
```

```
z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(1-pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)
```

```
##                z         p round_p
## female 1.354282 0.1756464  0.1756
```

Results: The z-score = 1.354282 and p-value = 0.1756464. We do not have enough evidence to reject the null hypothesis. The average rating count for newbie female is equal to the average rating count of newbie male

Comparing average rating counts of male avg work count authors and female avg work count authors

```
m = with(master_data,tapply(rating_count[author_exp=='average'],
author_sex[author_exp=='average'], mean))
s = with(master_data,tapply(rating_count[author_exp=='average'],
author_sex[author_exp=='average'], sd))
n = with(master_data,tapply(rating_count[author_exp=='average'],
author_sex[author_exp=='average'], length))
data.frame(m,s,n)
```

```
##                m        s     n
## female 23773.62 154160.0 12867
## male   21309.88 101362.5 10745
```

```
z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(1-pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)
```

```
##                z         p round_p
## female 1.471525 0.1411493  0.1411
```

Results: The z-score = 1.471525 and p-value = 0.1411493. We do not have enough evidence to reject the null hypothesis. The average rating count for avg work count male authors is equal to the average rating count of avg work count female authors.

Comparing average rating counts of male experienced and female experienced

```
m = with(master_data,tapply(rating_count[author_exp=='experienced'],
author_sex[author_exp=='experienced'], mean))
s = with(master_data,tapply(rating_count[author_exp=='experienced'],
author_sex[author_exp=='experienced'], sd))
n = with(master_data,tapply(rating_count[author_exp=='experienced'],
author_sex[author_exp=='experienced'], length))
data.frame(m,s,n)
```

```
##                m        s    n
## female 40401.57 213975.2 1157
## male   45421.26 180701.0 3539
```

```
z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)
```

```
##                 z         p round_p
## female -0.7185741 0.4724034  0.4724
```

Results: The z-score = -0.7185741 and p-value = 0.4724034. We do not have enough evidence to reject the null hypothesis. The average rating count for experienced male is equal to the average rating count of experienced female.

Pairwise T-test with unequal variances:

```
p12 = t.test(master_data$rating_count[master_data$author_exp=='newbie' &
master_data$author_sex=='male' ],
master_data$rating_count[master_data$author_exp=='newbie'  &
master_data$author_sex=='female' ], var.equal = F)
p12
```

```
##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "newbie" &
master_data$author_sex == "male"] and
master_data$rating_count[master_data$author_exp == "newbie" &
master_data$author_sex == "female"]
## t = -1.3543, df = 5779.7, p-value = 0.1757
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7512.044  1373.594
## sample estimates:
## mean of x mean of y
##  13596.44  16665.66
```

```
p13 = t.test(master_data$rating_count[master_data$author_exp=='average' &
master_data$author_sex=='male' ],
master_data$rating_count[master_data$author_exp=='average'  &
master_data$author_sex=='female' ], var.equal = F)
p13
```

```
##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "average" &
master_data$author_sex == "male"] and
master_data$rating_count[master_data$author_exp == "average" &
```

```
master_data$author_sex == "female"]
## t = -1.4715, df = 22435, p-value = 0.1412
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5745.4312   817.9585
## sample estimates:
## mean of x mean of y 
##  21309.88  23773.62

p23 = t.test(master_data$rating_count[master_data$author_exp=='experienced' &
master_data$author_sex=='male' ],
master_data$rating_count[master_data$author_exp=='experienced'  &
master_data$author_sex=='female' ], var.equal = F)
p23

## 
##  Welch Two Sample t-test
## 
## data:  master_data$rating_count[master_data$author_exp == "experienced" &
master_data$author_sex == "male"] and
master_data$rating_count[master_data$author_exp == "experienced" &
master_data$author_sex == "female"]
## t = 0.71857, df = 1727.2, p-value = 0.4725
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8681.496 18720.886
## sample estimates:
## mean of x mean of y 
##  45421.26  40401.57
```

Comparing average rating counts of male newbie and male avg work count

```
m = with(new_avg,tapply(rating_count[author_sex=='male'],
author_exp[author_sex=='male'], mean))
s = with(new_avg,tapply(rating_count[author_sex=='male'],
author_exp[author_sex=='male'], sd))
n = with(new_avg,tapply(rating_count[author_sex=='male'],
author_exp[author_sex=='male'], length))
data.frame(m,s,n)

##                m        s     n
## average 21309.88 101362.5 10745
## newbie  13596.44  67610.2  2153

z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(1-pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)

##                z            p round_p
## average 4.395607 1.104634e-05       0
```

Results: The z-score = 4.395607 and p-value = 0. We have sufficient evidence to reject the null hypothesis. The average rating count for newbie male authors is not equal to the average rating count of avg work count male authors.

Comparing average rating counts of male newbie and male experienced author

```
m = with(new_exp,tapply(rating_count[author_sex=='male'],
author_exp[author_sex=='male'], mean))
s = with(new_exp,tapply(rating_count[author_sex=='male'],
author_exp[author_sex=='male'], sd))
n = with(new_exp,tapply(rating_count[author_sex=='male'],
author_exp[author_sex=='male'], length))
data.frame(m,s,n)

##                    m        s    n
## experienced 45421.26 180701.0 3539
## newbie      13596.44  67610.2 2153

z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(1-pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)

##                    z p round_p
## experienced 9.446552 0       0
```

Results: The z-score = 9.446552 and p-value = 0. We have sufficient evidence to reject the null hypothesis. The average rating count for newbie male authors is not equal to the average rating count of experienced male authors.

Comparing average rating counts of male experienced and male avg work count

```
m = with(avg_exp,tapply(rating_count[author_sex=='male'],
author_exp[author_sex=='male'], mean))
s = with(avg_exp,tapply(rating_count[author_sex=='male'],
author_exp[author_sex=='male'], sd))
n = with(avg_exp,tapply(rating_count[author_sex=='male'],
author_exp[author_sex=='male'], length))
data.frame(m,s,n)

##                    m        s     n
## average     21309.88 101362.5 10745
## experienced 45421.26 180701.0  3539

z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)

##                 z            p round_p
## average -7.555948 4.158188e-14       0
```

Results: The z-score = -7.555948 and p-value = 0. We have sufficient evidence to reject the null hypothesis. The average rating count for experienced male authors is not equal to the average rating count of avg work count male authors.

ANOVA Test:

```
summary(aov(master_data$rating_count[master_data$author_sex=='male']~as.facto
r(master_data$author_exp[master_data$author_sex=='male'])))
```

```
##                                                                         Df
## as.factor(master_data$author_exp[master_data$author_sex == "male"])     2
## Residuals                                                           16434
##                                                                        Sum
Sq
## as.factor(master_data$author_exp[master_data$author_sex == "male"])
1.898e+12
## Residuals
2.358e+14
##                                                                       Mean
Sq
## as.factor(master_data$author_exp[master_data$author_sex == "male"])
9.491e+11
## Residuals
1.435e+10
##                                                                     F
value
## as.factor(master_data$author_exp[master_data$author_sex == "male"])
66.16
## Residuals
##                                                                     Pr(>F)
## as.factor(master_data$author_exp[master_data$author_sex == "male"]) <2e-16
***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pairwise T-test with unequal variances:

```
p12 = t.test(master_data$rating_count[master_data$author_exp=='newbie' &
master_data$author_sex=='male'],
master_data$rating_count[master_data$author_exp=='average' &
master_data$author_sex=='male'],var.equal = F)
p12
```

```
##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "newbie" &
master_data$author_sex == "male"] and
master_data$rating_count[master_data$author_exp == "average" &
master_data$author_sex == "male"]
```

```
## t = -4.3956, df = 4350.1, p-value = 1.131e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11153.760  -4273.127
## sample estimates:
## mean of x mean of y
##  13596.44  21309.88
```

```
p13 = t.test(master_data$rating_count[master_data$author_exp=='newbie' &
master_data$author_sex=='male'],
master_data$rating_count[master_data$author_exp=='experienced' &
master_data$author_sex=='male'],var.equal = F)
p13
```

```
##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "newbie" &
master_data$author_sex == "male"] and
master_data$rating_count[master_data$author_exp == "experienced" &
master_data$author_sex == "male"]
## t = -9.4466, df = 4924.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -38429.44 -25220.21
## sample estimates:
## mean of x mean of y
##  13596.44  45421.26
```

```
p23 = t.test(master_data$rating_count[master_data$author_exp=='average' &
master_data$author_sex=='male'],
master_data$rating_count[master_data$author_exp=='experienced' &
master_data$author_sex=='male'],var.equal = F)
p23
```

```
##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "average" &
master_data$author_sex == "male"] and
master_data$rating_count[master_data$author_exp == "experienced" &
master_data$author_sex == "male"]
## t = -7.5559, df = 4294.1, p-value = 5.052e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -30367.48 -17855.28
## sample estimates:
## mean of x mean of y
##  21309.88  45421.26
```

Comparing average rating counts of female newbie and female avg work count

```
m = with(new_avg,tapply(rating_count[author_sex=='female'],
author_exp[author_sex=='female'], mean))
s = with(new_avg,tapply(rating_count[author_sex=='female'],
author_exp[author_sex=='female'], sd))
n = with(new_avg,tapply(rating_count[author_sex=='female'],
author_exp[author_sex=='female'], length))
data.frame(m,s,n)

##                m        s     n
## average 23773.62 154160.0 12867
## newbie  16665.66 105256.2  3677

z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(1-pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)

##                z           p round_p
## average 3.224229 0.001263123  0.0013
```

Results: The z-score = 3.224229 and p-value = 0.0013 We have sufficient evidence to reject the null hypothesis. The average rating count for newbie female authors is not equal to the average rating count of avg work count female authors.

Comparing average rating counts of female experienced and female newbie authors

```
m = with(new_exp,tapply(rating_count[author_sex=='female'],
author_exp[author_sex=='female'], mean))
s = with(new_exp,tapply(rating_count[author_sex=='female'],
author_exp[author_sex=='female'], sd))
n = with(new_exp,tapply(rating_count[author_sex=='female'],
author_exp[author_sex=='female'], length))
data.frame(m,s,n)

##                    m        s    n
## experienced 40401.57 213975.2 1157
## newbie      16665.66 105256.2 3677

z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(1-pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)

##                    z           p round_p
## experienced 3.637264 0.000275549   3e-04
```

Results: The z-score = 3.637264 and p-value = 0.000275549 We have sufficient evidence to reject the null hypothesis. The average rating count for newbie female authors is not equal to the average rating count of experienced female authors.

Comparing average rating counts of female experienced and female avg work count

```
m = with(avg_exp,tapply(rating_count[author_sex=='female'],
author_exp[author_sex=='female'], mean))
s = with(avg_exp,tapply(rating_count[author_sex=='female'],
author_exp[author_sex=='female'], sd))
n = with(avg_exp,tapply(rating_count[author_sex=='female'],
author_exp[author_sex=='female'], length))
data.frame(m,s,n)

##                    m        s     n
## average     23773.62 154160.0 12867
## experienced 40401.57 213975.2  1157

z = (m[1]-m[2])/sqrt(sum(s^2/n))
p = 2*(pnorm(z))
round_p = round(p,4)
data.frame(z,p,round_p)

##                 z           p round_p
## average -2.583666 0.009775639  0.0098
```

Results: The z-score = -2.583666 and p-value = 0.0098 We have sufficient evidence to reject the null hypothesis. The average rating count for experienced female authors is not equal to the average rating count of avg work count female authors.

ANOVA Test:

```
summary(aov(master_data$rating_count[master_data$author_sex=='female']~as.fac
tor(master_data$author_exp[master_data$author_sex=='female'])))

##
Df
## as.factor(master_data$author_exp[master_data$author_sex == "female"])
2
## Residuals
17698
##
Sum Sq
## as.factor(master_data$author_exp[master_data$author_sex == "female"])
5.030e+11
## Residuals
3.994e+14
##
Mean Sq
## as.factor(master_data$author_exp[master_data$author_sex == "female"])
2.515e+11
## Residuals
2.257e+10
##                                                                      F
value
## as.factor(master_data$author_exp[master_data$author_sex == "female"])
11.14
```

```
## Residuals
##
Pr(>F)
## as.factor(master_data$author_exp[master_data$author_sex == "female"])
1.46e-05
## Residuals
##
## as.factor(master_data$author_exp[master_data$author_sex == "female"]) ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pairwise T-test with unequal variances:

```
p12 = t.test(master_data$rating_count[master_data$author_exp=='newbie' &
master_data$author_sex=='female'],
master_data$rating_count[master_data$author_exp=='average' &
master_data$author_sex=='female'],var.equal = F)
p12

##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "newbie" &
master_data$author_sex == "female"] and
master_data$rating_count[master_data$author_exp == "average" &
master_data$author_sex == "female"]
## t = -3.2242, df = 8636.9, p-value = 0.001268
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11429.387  -2786.522
## sample estimates:
## mean of x mean of y
##  16665.66  23773.62

p13 = t.test(master_data$rating_count[master_data$author_exp=='newbie' &
master_data$author_sex=='female'],
master_data$rating_count[master_data$author_exp=='experienced' &
master_data$author_sex=='female'],var.equal = F)
p13

##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "newbie" &
master_data$author_sex == "female"] and
master_data$rating_count[master_data$author_exp == "experienced" &
master_data$author_sex == "female"]
## t = -3.6373, df = 1336.3, p-value = 0.000286
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -36537.75 -10934.06
## sample estimates:
## mean of x mean of y
##  16665.66  40401.57
```

```
p23 = t.test(master_data$rating_count[master_data$author_exp=='average' &
master_data$author_sex=='female'],
master_data$rating_count[master_data$author_exp=='experienced' &
master_data$author_sex=='female'],var.equal = F)
p23
```

```
##
##  Welch Two Sample t-test
##
## data:  master_data$rating_count[master_data$author_exp == "average" &
master_data$author_sex == "female"] and
master_data$rating_count[master_data$author_exp == "experienced" &
master_data$author_sex == "female"]
## t = -2.5837, df = 1266.2, p-value = 0.009887
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -29253.953  -4001.951
## sample estimates:
## mean of x mean of y
##  23773.62  40401.57
```