Assignment 4

Wednesday, 26 October 2022 7:51 PM

Q2

a. We have a multinomial distribution with
$K=4$, $n = X_1 + X_2 + X_3 + X_4$,
ρ_1, ρ_2, ρ_3, and $\rho_4 = 1 - \rho_1 - \rho_2 - \rho_3$

The likelihood function of the multinomial function is given by

$$\mathcal{L}(\underline{\rho} \mid \underline{X}) = \binom{n}{\underline{x}} \rho_1^{x_1} \times \dots \times \rho_K^{x_K}$$

So, for our $K=4$, we have

$$\mathcal{L}(\underline{\rho} \mid X) = \binom{n}{\underline{x}} \rho_1^{X_1} \cdot \rho_2^{X_2} \cdot \rho_3^{X_3} \cdot (1-\rho_1-\rho_2-\rho_3)^{X_4}$$

or

$$\mathcal{L}(\underline{\rho} \mid \underline{X}) = \frac{n!}{X_1! X_2! X_3! X_4!} \rho_1^{X_1} \cdot \rho_2^{X_2} \cdot \rho_3^{X_3} \cdot (1-\rho_1-\rho_2-\rho_3)^{X_4}$$

b. Under the null hypothesis, we have

$$H_0 : \rho_2 = \rho_3$$

Thus, ρ_1 is a free parameter
ρ_2 / ρ_3 count as one free parameter
and $\rho_4 = 1 - \rho_1 - \rho_2 - \rho_3$ is not a free parameter.
$= 1 - \rho_1 - 2\rho_2$
i.e. There are two free parameters under the null.

c. Under the alternative

$$H_1 : \rho_2 \neq \rho_3$$

we will have ρ_1, ρ_2, and ρ_3 as free parameters
So, we have 3 free parameters under the alternative.

d. Under the null, the likelihood becomes

$$\mathcal{L}(\underline{\rho} \mid \underline{X}) = n! \; \frac{\rho_1^{X_1}}{X_1!} \; \frac{\rho_2^{X_2+X_3}}{X_2! X_3!} \; \frac{\rho_4^{X_4}}{X_4!}$$

And the log likelihood is

$$\log \mathcal{L}(\underline{\rho}) = \log(n!) + X_1 \log \rho_1 + (X_2+X_3) \log \rho_2 + X_4 \log \rho_4 - \log(X_1! X_2! X_3! X_4!)$$

We note that $\rho_1 + \rho_2 + \rho_3 + \rho_4 = 1$, and $n = X_1 + X_2 + X_3 + X_4$
Introduce a Lagrange constraint, λ, such that

$$\sum_{i=1}^{4} \rho_i = \sum_{i=1}^{4} \frac{X_i}{\lambda}$$

$$1 = \frac{1}{\lambda} \sum_{i=1}^{4} X_i$$

$$= \frac{n}{\lambda} \quad \Rightarrow \quad \lambda = n$$

Putting this as a constraint to the log likelihood, we get that

Putting this as a constraint to the log likelihood, we get that

$$\log \mathcal{L}(p, \lambda) = \log \mathcal{L}(p \mid \underline{x}) + \lambda \left(1 - \sum_{i=1}^{4} p_i\right)$$

Next, we want to find the MLE (i.e. $\frac{\partial \log \mathcal{L}}{\partial \theta} = 0$) for p_1:

$$\frac{\partial \log \mathcal{L}(p, \lambda)}{\partial p_1} = \frac{X_1}{\hat{p}_1} - \lambda = 0$$

$$\Rightarrow \frac{X_1}{\hat{p}_1} = \lambda = n$$

$$\Rightarrow \hat{p}_1 = \frac{X_1}{n}$$

An identical result can be obtained for each $(\hat{p}_i, X_i)$ pair. However, $p_2 = p_3 \Rightarrow \hat{p}_2 = \hat{p}_3$ under the null, and so

$$\hat{p}_2 + \hat{p}_3 = \frac{X_2}{n} + \frac{X_3}{n}$$

$$2\hat{p}_2 = 2\hat{p}_3 = \frac{X_2 + X_3}{n}$$

$$\Rightarrow \hat{p}_2 = \hat{p}_3 = \frac{X_2 + X_3}{2n}$$

e. Under the null hypothesis, we'd expect that,

$$\hat{p}_i = \frac{X_i}{n} \Rightarrow X_i = \hat{p}_i \cdot n$$

for each cell. That is

$$X_1 = \hat{p}_1 \cdot n \; ; \; X_2 = X_3 = \hat{p}_2 \cdot n = \hat{p}_3 \cdot n \; ; \; X_4 = \hat{p}_4 \cdot n$$

(or $X_4 = (1 - \hat{p}_1 - 2\hat{p}_2) \cdot n$)

f. We have McNemar's test as

$$\frac{(X_2 - X_3)^2}{X_2 + X_3}$$

and Pearson's χ^2 test as

$$\chi^2 = \sum_{\text{cells}} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(X_1 - \hat{p}_1 n)^2}{X_1} + \frac{(X_2 - \hat{p}_2 n)^2}{X_2} + \frac{(X_3 - \hat{p}_3 n)^2}{X_3} + \frac{(X_4 - \hat{p}_4 n)^2}{X_4}$$

$$= \frac{(X_1 - \frac{X_1}{n} \cdot n)^2}{n} + \frac{(X_2 - \frac{X_2 + X_3}{2n} \cdot n)^2}{X_2} + \frac{(X_3 - \frac{X_2 + X_3}{2n} \cdot n)^2}{X_3} + \frac{(X_4 - \frac{X_4}{n} \cdot n)^2}{X_4}$$

$$= \frac{(X_2 - \frac{X_2}{2} - \frac{X_3}{2})^2}{X_2} + \frac{(X_3 - \frac{X_2}{2} - \frac{X_3}{2})^2}{X_3}$$

$$= \frac{(X_2 - X_3)^2}{4X_2} + \frac{(X_3 - X_2)^2}{4X_3}$$

Now, notice that

$$X_2 = \hat{p}_2 \cdot n = \frac{X_2 + X_3}{2} = \hat{p}_3 \cdot n = X_3$$

$$\Rightarrow \chi^2 = \frac{(X_2 - X_3)^2 + (X_3 - X_2)^2}{4 \cdot \frac{(X_2 + X_3)}{2}}$$

$$= \frac{X_2^2 - 2X_2X_3 + X_3^2 + X_3^2 - 2X_2X_3 + X_2^2}{\left(\frac{X_2 + X_3}{2}\right)}$$

$$= \frac{X_2^2 - 2X_2X_3 + X_3^2}{X_2 + X_3}$$

$$= \frac{X_2^2 - 2X_2X_3 + X_3^2}{X_2 + X_3}$$

$$= \frac{(X_2 - X_3)^2}{X_2 + X_3}$$

g. Since, under the null, X_2 and X_3 are described by the same MLE, they should have the same value and so there are only two free parameters, p_1 and p_2 (or p_3). Hence, $dof = n_{free} - 1 = 2 - 1 = 1$, where n_{free} is the number of free parameters.

h. Since we're dealing with $dof = 1$, and the tabulated data gives

$$\chi^2 = \frac{(X_2 - X_3)^2}{X_2 + X_3}$$

$$= \frac{(15 - 4)^2}{15 + 4} \simeq 6.368$$

our p-value is given by

$$p = P(\chi_1^2 \geq 6.368)$$

In python, we can calculate this by

```
print(1 - scipy.stats.chi2.cdf(6.639,1))
```

which yields

$$p \simeq 0.0116$$

This is strong evidence that $p_2 \neq p_3$, and that the null hypothesis doesn't explain the phenomena. i.e. that $n_{willing \to unwilling} \neq n_{unwilling \to willing}$

i. Since one cell had < 5 counts, we need to use Fisher's exact test.
Firstly, we can consider the row and column totals fixed under the null, as we essentially have only one degree of freedom (with two free parameters).
As such, we can generate multiple tables as any should be equally likely under the null.

We used the python code below to analyse this

```
#q2i
row1 = 52; row2 = 30
col1 = 41; col2 = 41
iters = 10**6
counts = 0

for i in range(iters):
    x1 = np.random.randint(0, min(row1 + 1, col1))
    x2 = row1 - x1
    x3 = col1 - x1
    x4 = col2 - x2
    if (x2 - x3)**2 / (x2 + x3) >= 6.368:
        counts += 1

print(f"p-value from {iters} trials is ", counts / iters)
```

This produced a p-value of

$$p \simeq 0.098 \quad \text{(on average)}$$

from 10^6 table variations.

$p \simeq 0.098$ (on average)
from 10^6 table variations.
This says that there is no significant evidence against the null hypothesis via Fisher's exact test. (i.e. our conclusion from h) was likely wrong).

Q3
We have the numerical data,

maths terms	incorrect solutions	correct solutions	total
μ	298	165	463
$\overline{X}$	146	264	410
S or S^2	119	16	135
$\leq$ or $\geq$ or $<$ or $>$	106	328	434
1.96 or -1.96	29	203	232
$\sqrt{n}$	133	165	298
∞	29	30	59
Total	860	1171	2031

a. In this situation, our null and alternative hypotheses are

H_0: The relative frequency of each math term is the same across correct/incorrect solutions

H_1: The relative frequency of at least one math term differs between correct/incorrect solutions.

We are interested here in the counts of each term (for correct and incorrect solutions), as well as each row/column total.
Notice that we'll have $(n_{rows}-1)\times(n_{cols}-1)$ degrees of freedom. That is,

$$dof = (7-1)\times(2-1) = 6$$

So we can calculate an approximate p-value via

$$p\text{-value} = \mathbb{P}(\chi^2_6 \geq \chi^2)$$

Where χ^2 is Pearson's χ^2 statistic, given by

$$\chi^2 = \sum_{cells} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed value in the $(i,j)^{th}$ cell, and

$$E_{ij} = \frac{(\Sigma O_{:j})\times(\Sigma O_{i:})}{N}$$

where N is the sum of the cell counts.
Using the following python code,

```
col1 = np.array([298, 146, 119, 106, 29, 133, 29])
col2 = np.array([165, 264, 16, 328, 203, 165, 30])
total = sum(col1) + sum(col2)
chi2 = 0
for i in range(len(col1)):
    for col in [col1, col2]:
        Eij = (sum(col) * sum([col1[i], col2[i]])) / total
        chi2 += (col[i] - Eij)**2 / Eij
pval = 1 - stats.chi2.cdf(chi2, 6)
print(f"We get a value of {chi2 = }, and a corresponding {pval = }")
```

we obtain a Pearson's χ^2 of

$$\chi^2 \simeq 359.06$$

and a p-value $< 1e-16$ (almost certain significance)
(python displays a pval = 0.0, so the p-value is

$\chi^2 \simeq 331.06$

and a p-value $< 1e-16$ (almost certain significance)
(python displays a pval = 0.0, so the p-value is below floating point precision).
So, we have incredibly strong evidence that the usage of at least one mathematical term differs between correct and incorrect solutions.

b. For the 8 solutions that got 0/4 marks, we can employ a similar process to part a)

	incorrect solution								
maths terms	1	2	3	4	5	6	7	8	
μ	12	11	12	12	13	11	13	12	96
$\overline{X}$	6	6	7	6	5	8	7	5	50
S or S^2	5	5	5	4	5	3	4	4	35
$\leq$ or $\geq$ or $<$ or $>$	4	4	6	2	4	2	6	6	34
1.96 or -1.96	2	0	0	0	2	0	2	2	8
$\sqrt{n}$	5	6	4	5	4	4	6	5	39
∞	2	0	2	2	0	0	0	2	8
	36	32	36	31	33	28	38	36	270

In this case, our null and alternative hypotheses will be
H_0: The relative frequency of math terms is consistent across the 8 answers
H_1: There is at least one solution that uses a different relative frequency of math terms.

We can use Pearson's χ^2 test again, with
$dof = (n_{rows} - 1) \times (n_{cols} - 1) = (7-1) \times (8-1) = 42$

and p-value $= P(\chi^2_{42} \geq \chi^2)$
Using the following python code,

```
col1 = np.array([12, 6, 5, 4, 2, 5, 2])
col2 = np.array([11, 6, 5, 4, 0, 6, 0])
col3 = np.array([12, 7, 5, 6, 0, 4, 2])
col4 = np.array([12, 6, 4, 2, 0, 5, 2])
col5 = np.array([13, 5, 5, 4, 2, 4, 0])
col6 = np.array([11, 8, 3, 2, 0, 4, 0])
col7 = np.array([13, 7, 4, 6, 2, 6, 0])
col8 = np.array([12, 5, 4, 6, 2, 5, 2])
total = np.sum([[col1, col2, col3, col4, col5, col6, col7, col8]])
chi2 = 0
for i in range(len(col1)):
    for col in [col1, col2, col3, col4, col5, col6, col7, col8]:
        Eij = (sum(col) * sum([col1[i], col2[i], col3[i], col4[i],
                               col5[i], col6[i], col7[i], col8[i]])) / total
        chi2 += (col[i] - Eij)**2 / Eij
pval = 1 - stats.chi2.cdf(chi2, 42)
print(f"We get a value of {chi2 = }, and a corresponding {pval = }")
```

We obtain $\chi^2 \simeq 22.08$ and p-val $\simeq 0.995$
This is incredibly weak evidence against the null hypothesis, and so we conclude that all 8 incorrect solutions use the same relative frequency of math terms.

c. The p-value in part b) is definitely suspicious, as it is very large, and we would not expect that from independent answers.
The cause that first comes to mind (in the context of an assignment solution) is plagiarism amongst at least these 8 students (and possibly more when looking at other incorrect solutions).

d. Since so many of the cells in the table have less than 5 counts, we don't meet the required conditions to definitively use Pearson's χ^2 test. As such, we would need to perform

have less than ~5 counts, we don't meet the required conditions to definitively use Pearson's χ^2 test. As such, we would need to perform Fisher's exact test on the observed data.

Q1

a. Using a short python script with numpy, the data can be summarised by

$\mu \simeq 3.096$ $\quad \sigma \simeq 1.59$

median = 3 $\quad$ mode = 5 (52 counts)

b. For our hypothesis test, we propose

H_0: the proportion of each opinion is consistant between regular people and influencers

H_1: there is at least one difference in opinion proportions between regular people and influencers.

To test these hypotheses, we can employ Pearson's χ^2 statistic. Using the following python script,

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

# q1a
data = pd.read_csv("likes.csv", delimiter=',')

print('mean = ', np.mean(data['rating']), ' sd = ', np.std(data['rating']),
      ' median = ', np.median(data['rating']),
      ' mode = ', stats.mode(data['rating']))

# q1b
InfAnswers = data[data['user'] == 'influencer']
RegAnswers = data[data['user'] == 'regular']
InfCounts, RegCounts = np.zeros(5), np.zeros(5)
for i in range(len(InfAnswers['rating'])):
    InfCounts[InfAnswers['rating'].to_numpy()[i] - 1] += 1
for i in range(len(RegAnswers['rating'])):
    RegCounts[RegAnswers['rating'].to_numpy()[i] - 1] += 1
print([RegCounts, InfCounts])
total = sum(RegCounts) + sum(InfCounts)
chi2 = 0
for i in range(len(RegCounts)):
    for col in [RegCounts, InfCounts]:
        Eij = (sum(col) * sum([RegCounts[i], InfCounts[i]])) / total
        chi2 += (col[i] - Eij)**2 / Eij
pval = 1 - stats.chi2.cdf(chi2, 4)
print(f"We get a value of {chi2 = }, and a corresponding {pval = }")
```

we obtain the table,

Response	Respondent Type		
	Regular	Influencer	
Str Disagree	19	27	
Disagree	18	15	
Neutral	21	6	
Agree	22	18	
Str Agree	20	32	
Totals	100	98	198

We have

$$dof = (n_{rows} - 1) \times (n_{cols} - 1) = (2-1) \times (5-1)$$
$$= 4$$

degrees of freedom here, and we see that >80% of cells in the table have ≥10 counts, and all cells have ≥5 counts, so Pearson's χ^2 is

of cells in the table have ≥ 10 counts, and all cells have ≥ 5 counts, so Pearson's χ^2 is a suitable statistic.

From the python code, we get

$$\chi^2 \simeq 13.148 \Rightarrow p\text{-val} = \mathbb{P}(\chi^2_4 \geq \chi^2)$$
$$= \mathbb{P}(\chi^2_4 \geq 13.148)$$
$$\simeq 0.0106$$

So we obtain significant evidence that there is at least one differing opinion between a population of regular people and influencers.

(i.e. that regular people and influencers somewhat disagree on the removal of 'likes' from this particular social media platform.)

c. Check audio file. Tl;dr:

- We assume that the data points were independent of each other, i.e. that no one person "influenced" (haha) another person's answer.
- People with 10000+ followers are representative of a sample of "influencers"
- Pearson's χ^2 is an accurate test statistic for this data set. We determined that the counts satisfied the minimum requirements to use Pearson's χ^2, but ideally all cells would have a count of 10 or more.

d. Firstly, the two-sample t-test relies on the assumption that the data is described by an underlying standard normal distribution.
We don't know if the data can be described by this.
Secondly, our domain of discrete measurements is simply too small to accurately approximate our discrete data with a continuous normal distribution.

Q4

We have a two-way ANOVA, with

$$Y_{jki} = \mu + \alpha_j + \beta_k + \delta_{jk} + \epsilon_{jki}$$

with constraints

$$\sum_j \alpha_j = 0 \; ; \; \sum_k \beta_k = 0 \; ; \; \sum_j \delta_{jk} = \sum_k \delta_{jk} = 0$$

a. Under the sum constraints, the joint likelihood function is given by

$$\mathcal{L}(\mu, \underline{\alpha}, \underline{\beta}, \underline{\delta}, \underline{\epsilon}, \sigma^2 / \underline{Y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{} \sum_{k=1}^{} \sum_{i=1}^{} \left(Y_{jki} - \mu - \alpha_j - \beta_k - \delta_{jki}\right)^2\right)$$

We can then maximise the log likelihood:

$$\log \mathcal{L} = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{} \sum_{k=1}^{} \sum_{i=1}^{} \left(Y_{jki} - \mu - \alpha_j - \beta_k - \delta_{jki}\right)^2 \quad (1)$$

Differentiating (1) w.r.t. μ and setting to 0 gives the MLE for μ, $\hat{\mu}$

$$\Rightarrow \frac{\partial \log \mathcal{L}}{\partial \mu} = 0 = -\frac{1}{\sigma^2} \sum_{j=1}^{} \sum_{k=1}^{} \sum_{i=1}^{} \left(Y_{jki} - \hat{\mu} - \alpha_j - \beta_k - \delta_{jki}\right)$$

$$\Rightarrow N\hat{\mu} = \sum\sum\sum Y_{jki} - \mu \sum\sum\sum 1 - \sum_j \alpha_j \sum\sum 1 - \sum_k \beta_k \sum\sum 1 - \sum_j \sum_k \delta_{jk} \sum 1$$

$$\Rightarrow \frac{\partial \log L}{\partial \mu} = 0 = -\frac{1}{\sigma^2}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{r}(Y_{jki} - \hat{\mu} - \alpha_j - \beta_k - \delta_{jki})$$

$$\Rightarrow N\hat{\mu} = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{r} Y_{jki} - Kr\sum_{j=1}^{J}\alpha_j - Jr\sum_{k=1}^{K}\beta_k - \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{r}\delta_{jki}$$

with the sum constraints, this is then

$$N\hat{\mu} = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{r} Y_{jki} - 0 - 0 - 0$$

$$\Rightarrow \hat{\mu} = \frac{1}{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{r} Y_{jki} = \overline{Y}_{...}$$

where $N = J \cdot K \cdot r$

Similarly, differentiating (1) w.r.t. α_j and setting to 0 gives

$$\frac{\partial \log L}{\partial \alpha_j} = 0 = -\frac{1}{\sigma^2}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{r} Y_{jki} - \mu - \hat{\alpha}_j - \beta_k - \delta_{jki}$$

$$\Rightarrow Kr\hat{\alpha}_j = \sum_{k=1}^{K}\sum_{i=1}^{r} Y_{jki} - Kr\mu - r\sum_{k=1}^{K}\beta_k - \sum_{k=1}^{K}\sum_{i=1}^{r}\delta_{jki}$$

Applying the sum constraints gives

$$Kr\hat{\alpha}_j = \sum_{k=1}^{K}\sum_{i=1}^{r} Y_{jki} - Kr\mu - r \cdot 0 - 0$$

$$\Rightarrow \hat{\alpha}_j = \frac{1}{Kr}\sum_{k=1}^{K}\sum_{i=1}^{r} Y_{jki} - \mu$$

$$= \overline{Y}_{j..} - \overline{Y}_{...}$$

for β_k, we get

$$\frac{\partial \log L}{\partial \beta_k} = 0 = -\frac{1}{\sigma^2}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{r} Y_{jki} - \mu - \alpha_j - \hat{\beta}_k - \delta_{jki}$$

$$\Rightarrow Jr\hat{\beta}_k = \sum_{j=1}^{J}\sum_{i=1}^{r} Y_{jki} - Jr\mu - r\sum_{j=1}^{J}\alpha_j - \sum_{j=1}^{J}\sum_{i=1}^{r}\delta_{jki}$$

Applying sum constraints gives

$$\hat{\beta}_k = \frac{1}{Jr}\sum_{j=1}^{J}\sum_{i=1}^{r} Y_{jki} - \mu - 0 - 0$$

$$= \overline{Y}_{.k.} - \overline{Y}_{...}$$

And finally for δ_{jk},

$$\frac{\partial \log L}{\partial \delta_{jk}} = 0 = -\frac{1}{\sigma^2}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{r} Y_{jki} - \mu - \alpha_j - \beta_k - \hat{\delta}_{jki}$$

$$\Rightarrow Jr\hat{\delta}_{jk} = \sum_{i=1}^{r} Y_{jki} - Jr\mu - Jr\alpha_j - Jr\beta_k$$

$$\Rightarrow \hat{\delta}_{jk} = \frac{1}{Jr}\sum_{i=1}^{r} Y_{jki} - \mu - \alpha_j - \beta_k$$

$$= \overline{Y}_{jk.} - \overline{Y}_{...} - \overline{Y}_{j..} - \overline{Y}_{.k.}$$

as required.

b. WTS: $SS_{total} = SS_A + SS_B + SS_{AB} + SS_{resid}$

We have that $SS_{total} = \sum_{jki}(Y_{jki} - \overline{Y}_{...})^2$, and

$$Y_{jki} - \overline{Y}_{...} = (Y_{jki} - \overline{Y}_{jk.}) + (\overline{Y}_{j..} - \overline{Y}_{...}) + (\overline{Y}_{.k.} - \overline{Y}_{...}) + (\overline{Y}_{jk.} - \overline{Y}_{j..} - \overline{Y}_{.k.} - \overline{Y}_{...})$$

$$\Rightarrow \sum_{jki}(Y_{jki} - \overline{Y}_{...})^2 = \sum_{jki}\Big([Y_{jki} - \overline{Y}_{jk.}]^2 + [\overline{Y}_{j..} - \overline{Y}_{...}]^2 + [\overline{Y}_{.k.} - \overline{Y}_{...}]^2$$
$$+ [\overline{Y}_{jk.} - \overline{Y}_{j..} - \overline{Y}_{.k.} - \overline{Y}_{...}]^2$$
$$+ [Y_{jki} - \overline{Y}_{jk.}]\left(\overline{Y}_{j..} - \overline{Y}_{...} + \overline{Y}_{.k.} - \overline{Y}_{...} + \overline{Y}_{jk.} - \overline{Y}_{j..} - \overline{Y}_{.k.} - \overline{Y}_{...}\right)$$
$$+ [\overline{Y}_{j..} - \overline{Y}_{...}]\left(Y_{jki} - \overline{Y}_{jk.} + \overline{Y}_{.k.} - \overline{Y}_{...} + \overline{Y}_{jk.} - \overline{Y}_{j..} - \overline{Y}_{.k.} - \overline{Y}_{...}\right)$$

$$+[Y_{jki}-\bar{Y}_{jk\cdot}]\left(\bar{Y}_{j\cdot\cdot}-\bar{Y}_{\cdots}+\bar{Y}_{\cdot k\cdot}-\bar{Y}_{\cdots}+\bar{Y}_{jk\cdot}-\bar{Y}_{j\cdot\cdot}-\bar{Y}_{\cdot k\cdot}-\bar{Y}_{\cdots}\right)$$
$$+\left[\bar{Y}_{j\cdot\cdot}-\bar{Y}_{\cdots}\right]\left(Y_{jki}-\bar{Y}_{jk\cdot}+\bar{Y}_{\cdot k\cdot}-\bar{Y}_{\cdots}+\bar{Y}_{jk\cdot}-\bar{Y}_{j\cdot\cdot}-\bar{Y}_{\cdot k\cdot}-\bar{Y}_{\cdots}\right)$$
$$+\left[\bar{Y}_{\cdot k\cdot}-\bar{Y}_{\cdots}\right]\left(Y_{jki}-\bar{Y}_{jk\cdot}+\bar{Y}_{j\cdot\cdot}-\bar{Y}_{\cdots}+\bar{Y}_{jk\cdot}-\bar{Y}_{j\cdot\cdot}-\bar{Y}_{\cdot k\cdot}-\bar{Y}_{\cdots}\right)$$
$$+\left[\bar{Y}_{jk\cdot}-\bar{Y}_{j\cdot\cdot}-\bar{Y}_{\cdot k\cdot}-\bar{Y}_{\cdots}\right]\left(Y_{jki}-\bar{Y}_{jk\cdot}+\bar{Y}_{j\cdot\cdot}-\bar{Y}_{\cdots}+\bar{Y}_{\cdot k\cdot}-\bar{Y}_{\cdots}\right)$$

$$= SS_{resid}+SS_A+SS_B+SS_{AB}$$
$$+\sum_{jki}\Big[(Y_{jki}-\bar{Y}_{jk\cdot})(\bar{Y}_{jk\cdot}-3\bar{Y}_{\cdots})+(Y_{j\cdot\cdot}-Y_{\cdot\cdot})(Y_{jki}-2\bar{Y}_{\cdot\cdot})$$
$$+(\bar{Y}_{\cdot k\cdot}-\bar{Y}_{\cdots})(Y_{jki}-2\bar{Y}_{\cdots})+(\bar{Y}_{jk\cdot}-\bar{Y}_{j\cdot\cdot}-\bar{Y}_{\cdot k\cdot}-\bar{Y}_{\cdots})(Y_{jki}-\bar{Y}_{jk\cdot}+\bar{Y}_{j\cdot\cdot}+\bar{Y}_{\cdots}-2\bar{Y}_{\cdot\cdot})\Big]$$

$$= SS_A+SS_B+SS_{AB}+SS_{resid}$$
$$+\sum_{jki}0$$

$$= SS_A+SS_B+SS_{AB}+SS_{resid}.$$

c. We'd expect that the residual sum of squares has a distribution given by

$$\frac{SS_{resid}}{\sigma^2}\sim\chi^2_{df_{resid}}$$

as the residuals over the variance should have a standard normal distribution, and so the square of the residuals should have a χ^2 distribution. The sum of the SS_{resid}/σ^2 should be $\sim\chi^2_{df_{resid}}$, as we would expect the degrees of freedom to be the total df of all data, minus the number of means we are estimating. That is

$$df_{resid} = df_{tot} - \text{no. } \mu_i$$
$$= JK(r)-JK$$
$$= JK(r-1)$$

Which, for $J=K=2$ and $r=3$, we get

$$df_{resid} = 8.$$