Sarah Cooper

Week 8 Q3

3. **Oneway ANOVA**
Suppose $Y_{i1}, \ldots, Y_{in_i} \overset{\text{iid}}{\sim} N(\mu_i, \sigma^2)$ for $i = 1, \ldots, k$ groups. Let Y_i be the group i average and let $\bar{Y}$ be the overall average. Consider a test of $H_o : \mu_1 = \ldots = \mu_k$ vs. H_a : "not H_o".

a) Explain how the F statistic generalizes the pooled 2-sample t statistic.

$$F - \text{stat} = \frac{\text{SSM}/(k-1)}{\text{SSE}/(n-k)} = \frac{\sum_{i=1}^k n_i (Y_i - \bar{Y})^2/(k-1)}{\sum_{i=1}^k (n_i - 1)s_i^2/(n-k)}$$

For example, show how rejecting for large values of the this is the same as rejecting for large T^2 when $k = 2$ and $N = n_1 + n_2$ (what would unusually small values of T^2 suggest?).

a) $Y_{i1}, \ldots Y_{in_i} \overset{iid}{\sim} N(\mu_i, \sigma^2)$ $i=1\ldots k$ groups $\quad$ Y_i = group i average $\quad$ $\bar{Y}$ = overall average

$H_0: \mu_1 = \ldots = \mu_k \quad H_a:$ not H_0

F-stat generalizes the pooled 2-sample t-stat:

$$F\text{-stat} = \frac{SSM/(k-1)}{SSE/(n-k)} = \frac{\sum_{i=1}^{k} n_i (Y_i - \bar{Y})^2/(k-1)}{\sum_{i=1}^{k} (n_i - 1)s_i^2/(n-k)}$$

→ variation btwn group and overall

→ variation in each group

2-sample t-test:

$$T = \frac{|\bar{Y}_1 - \bar{Y}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When $k=2$ and $N = n_1 + n_2$

$$F = \frac{SSM/(2-1)}{SSE/(n-2)} = \frac{SSM}{SSE(n-2)} = \frac{\sum n_i (Y_i - \bar{Y})^2}{\sum s_i^2 (n_i - 1)/(n-2)} = \frac{n_1(\bar{Y}_1 - \bar{Y})^2 + n_2(\bar{Y}_2 - \bar{Y})^2}{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{(n_1+n_2-2)}}$$

$$\bar{Y} = \text{overall mean} = \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2}{N}$$

numerator

$$n_1\left[\bar{Y}_1 - \left(\frac{n_1\bar{Y}_1 + n_2\bar{Y}_2}{N}\right)\right]^2 + n_2\left[\bar{Y}_2 - \left(\frac{n_1\bar{Y}_1 + n_2\bar{Y}_2}{N}\right)\right] = n_1\left[\frac{N\bar{Y}_1 - n_1\bar{Y}_1 + n_2\bar{Y}_2}{N}\right]^2 + n_2\left[\frac{N\bar{Y}_2 - n_1\bar{Y}_1 - n_2\bar{Y}_2}{N}\right]^2$$

$$= n_1\left[\frac{(N-n_1)\bar{y}_1 - n_2\bar{y}_2}{N}\right]^2 + n_2\left[\frac{(N-n_2)\bar{y}_2 - n_1\bar{y}_1}{N}\right]^2 = n_1\left[\frac{n_2\bar{y}_1 - n_2\bar{y}_2}{N}\right]^2 + n_2\left[\frac{n_1\bar{y}_2 - n_1\bar{y}_1}{N}\right]^2 = n_1\left[\frac{n_2(\bar{y}_1 - \bar{y}_2)}{N}\right]^2 + n_2\left[\frac{n_1(\bar{y}_2 - \bar{y}_1)}{N}\right]^2$$

$$= \frac{n_1 n_2^2}{N^2}(\bar{y}_1 - \bar{y}_2)^2 + \frac{n_2 n_1^2}{N^2}(\bar{y}_2 - \bar{y}_1)^2 = \left(\frac{n_1 n_2^2}{N^2} + \frac{n_2 n_1^2}{N^2}\right)(\bar{y}_1 - \bar{y}_2)^2 = \frac{n_1 n_2 (n_1 + n_2)}{N^2}(\bar{y}_1 - \bar{y}_2)^2 = \frac{n_1 n_2}{N}(\bar{y}_1 - \bar{y}_2)^2$$

$$= \frac{1}{\frac{N}{n_1 n_2}}(\bar{y}_1 - \bar{y}_2)^2 = \frac{1}{\frac{n_1 + n_2}{n_1 n_2}}(\bar{y}_1 - \bar{y}_2)^2 = \frac{1}{\frac{1}{n_1} + \frac{1}{n_2}}(\bar{y}_1 - \bar{y}_2)^2$$

$$\text{numerator} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$F = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{(n_1+n_2-2)}\right)}$$

$$T^2 = \left(\frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right) = \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

↳ since pooled $s_1 = s_2 = s_p$

pooled variance formula

Thus $F = T^2$ for $k=2$

You reject for large F. Indicating that variation btwn groups is larger than variation within groups. Since $F = T^2$ for $k=2$ then reject for large T^2 too. Small T^2 suggest difference btwn groups is small compared to within groups

If $n_1 = n_2 = n$ it is much more simple!

k groups, n_i individuals, Y_i = group i ave, Y_{ij} = individual j in group i

$$SSM = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{k} n_i (Y_i - \bar{Y})^2$$

when $k=2$:

$$\bar{Y} = \frac{n_1 Y_1 + n_2 Y_2}{n_1 + n_2} = \frac{n(Y_1+Y_2)}{2n} = \frac{1}{2}(Y_1+Y_2)$$

Then: $Y_1 - \bar{Y} = \frac{1}{2}(Y_1 - Y_2)$

$Y_2 - \bar{Y} = \frac{1}{2}(Y_2 - Y_1)$

So, $\sum_{i=1}^{2} n(Y_1 - \bar{Y})^2 = \frac{n}{2^2}(Y_1 - Y_2)^2 + \frac{n}{2^2}(Y_2 - Y_1)^2 = (Y_1 - Y_2)^2\left(\frac{n}{4} + \frac{n}{4}\right)$

Thus $SSM = \frac{n}{2}(Y_1 - Y_2)^2$

$$T\text{-stat} = \frac{Y_1 - Y_2}{Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{Y_1 - Y_2}{Sp\sqrt{\frac{2}{n}}}$$

$$T^2 = \frac{n(Y_1 - Y_2)^2}{2Sp^2}$$ where $Sp^2 = MSE = SSE/(N-k)$

$$F = \frac{SSM/(2-1)}{Sp^2} = \frac{\frac{n}{2}(Y_1 - Y_2)^2/(2-1)}{Sp^2} = \frac{n(Y_1 - Y_2)^2}{2Sp^2}$$

Therefore when $k=2$ F stat and T^2 are the same.

b) To simplify, suppose $n_i = n$, for $i = 1, \ldots, k$, so $N = nk$ and $\bar{Y} = \frac{1}{k}\sum_{i=1} Y_i$. Show, using facts we have already proved, that $\frac{\text{SSE}}{\sigma^2} = \sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(Y_{ij}-Y_i)^2}{\sigma^2} \sim \chi^2_{(N-k)}$ and independent of SSM, and that $\frac{\text{SST}}{\sigma^2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(Y_{ij}-\bar{Y})^2}{\sigma^2} \sim \chi^2_{(N-1)}$ if H_o is true. Also show SST = SSM + SSE, and infer that $\frac{\text{SSM}}{\sigma^2} \sim \chi^2_{(k-1)}$ if H_o is true. Conclude that the null sampling distribution of the F-statistic is $F_{(k-1,N-k)}$. Note that MSE = SSE/$(N-k)$ is an unbiased estimate for σ^2. If H_o is true, then MSM=SSM/$(k-1)$ is also unbiased for σ^2, but otherwise is biased high.

Step 1: Show that $\frac{SSE}{\sigma^2} = \sum_{i=1}^{k}\sum_{j=1}^{n} \frac{(Y_{ij}-Y_i)^2}{\sigma^2} \sim \chi^2_{(N-k)}$

Note that we have already shown that:

$Y_1, \dots Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$ $\qquad S^2 = \frac{\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_i)^2}{n-1}$

$Y_{ij}-\bar{Y}_i$ independent of $\bar{Y}$

$S^2 = \frac{\Sigma(Y_{ij}-\bar{Y}_i)^2}{n-1}$ independent of $\bar{Y}$

Also seen that: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ so $S^2 \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2}\right)$

So since $Y_{ij} \sim N(\mu, \sigma^2)$ then

$Y_{ij} - \bar{Y}_i \sim N(0, \sigma^2)$ (independently for each j)

And we know that $\frac{(n_i-1)S_i^2}{\sigma^2} \overset{iid}{\sim} \chi^2_{(n_i-1)}$

Thus $\frac{SSE}{\sigma^2} = \frac{\sum_{i=1}^{k}(n_i-1)S_i^2}{\sigma^2} = \frac{\Sigma\Sigma(Y_{ij}-\bar{Y}_i)^2}{\sigma^2} \sim \chi^2_{(N-k)}$ were $N = \sum_{i=1}^{k} n_i$

summing k independent groups so add degrees of freedom $n_1+n_2+\dots -1-1-1\dots = N-k$

Show that $MSE = SSE/N-k$ is unbiased for σ^2

$MSE = \frac{SSE}{N-k}$

$E(MSE) = E\left(\frac{SSE}{N-k}\right) = \frac{(N-\cancel{k})\sigma^2}{(N-\cancel{k})} = \sigma^2$

Also note that $MSE \sim \text{Gamma}\left(\frac{N-k}{2}, \frac{N-k}{2\sigma^2}\right)$ and $E(MSE) = \frac{\cancel{N-k}}{\cancel{2}}\left(\frac{2\sigma^2}{\cancel{N-k}}\right) = \sigma^2$

So unbiased estimate.

SSM and SSE are independent.

- we know that $Y_{ij} \sim N(\mu, \sigma^2)$ thus $(Y_{ij}-Y_i)$ or S_i^2 is independent of $\bar{Y}_i$
- $\bar{Y}_i$ are calculated from Y_{ij}'s but $(Y_{ij}-\bar{Y}_i)$ is independent of $\bar{Y}_i$.

So, SSM depends on group means $\bar{Y}_i$, but SSE depends on deviations from the group means

we already know that $(Y_{ij}-\bar{Y}_i)$ is independent of $\bar{Y}_i$ so SSE and SSM are indep.

Show that $\frac{SST}{\sigma^2} = \sum_{i=1}^{k} \frac{(Y_{ij}-\bar{Y})^2}{\sigma^2} \sim \chi^2_{(N-1)}$ under H_0

Note: Under H_0: $\mu_1 = \mu_2 = \mu_3 = \dots \mu_k = \mu$ so $Y_{ij} \overset{iid}{\sim} N(\mu, \sigma^2)$

$SST = \Sigma\Sigma(Y_{ij}-\bar{Y})^2$

Each $(Y_{ij}-\mu)^2 \sim N(0, \sigma^2)$

Summing standard normals is chi-square (proved at earlier date)

So $\frac{SST}{\sigma^2} \sim \chi^2_{(N-1)}$

It is $N-1$ because $\bar{Y}$ is estimated

Since SST = SSM + SSE and SSE and SSM are independent.

$$\frac{SST}{\sigma^2} = \frac{SSM}{\sigma^2} + \frac{SSE}{\sigma^2}$$

$$\chi^2_{(N-1)} = ? + \chi^2_{(N-k)}$$ - DF add

so $\frac{SSM}{\sigma^2} = \chi^2(k-1)$ only if Ho is true

Conclude null sampling of f-stat is $F(k-1, N-k)$

$$\text{F-stat} = \frac{SSM/(k-1)}{SSE/(N-k)} = \frac{MSM}{MSE}$$

under null $\frac{SSE}{\sigma^2} \sim \chi^2_{N-k}$ $\quad \frac{SSM}{\sigma^2} \sim \chi^2_{k-1}$

The ratio of two independent χ^2 variables follows an F distribution

Thus $F(k-1, N-k)$

Note: $X_1 \sim \text{Gamma}(a, \lambda)$ independent $X_2 \sim \text{Gamma}(b, \lambda)$

Here $X_1 = SSM \sim \text{Gamma}\left(\frac{k-1}{2}, \frac{1}{2\sigma^2}\right)$ and $X_2 = SSE \sim \text{Gamma}\left(\frac{N-k}{2}, \frac{1}{2\sigma^2}\right)$

$$\frac{X_1}{X_2} \sim F^*(a, b, c)$$

$$\frac{X_1/a}{X_2/b} \sim F(2a, 2b) \quad \text{So} \quad \frac{SSM/k-1}{SSE/N-k} \sim F(k-1, N-k)$$

MSM Biases:

$$MSM = SSM/(k-1)$$

Under Ho: $E(MSM) = E\left(\frac{SSM}{k-1}\right) = \frac{E(SSM)}{k-1} = \frac{\sigma^2(k-1)}{k-1} = \sigma^2$ because $SSM \sim \chi^2_{(k-1)}$ under Ho

Otherwise: $E(SSM)$ is greater than $\sigma^2(k-1)$ because variation between groups will no longer be due to pure random sampling.

Thus $E(MSM) > \sigma^2$

Also under Ho: $\bar{Y}_i \overset{iid}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$

$$E\left(\frac{\Sigma(\bar{Y}_i - \bar{Y})^2}{k-1}\right) = \frac{\sigma^2}{n}$$

c) Define $R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$ as the proportion of variability explained by the groups. Show that the null sampling distribution of R^2 is Beta$(k-1, N-k)$. With $k = 5$ and $n = 10$, what values of the F-statistic would lead you to reject H_o? What values of R^2 would lead you to reject? What conclusion could you make?

$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$ proportion of variability explained by the groups.

Under Ho:

All group means are equal, so the model does not explain any variability beyond what is expected by random

$\frac{SSM}{\sigma^2} \sim \chi^2_{k-1}$ and $\frac{SSE}{\sigma^2} \sim \chi^2_{N-k}$ also $\frac{SST}{\sigma^2} \sim \chi^2_{N-1}$ (part b)

$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$

$SSM = \sigma^2 X$ where $X \sim \chi^2_{k-1}$ $\sim \text{Gamma}\left(\frac{k-1}{2}, \frac{1}{2\sigma^2}\right)$ - under Ho

So $SST \sim \text{Gamma}\left(\frac{N-1}{2}, \frac{1}{2\sigma^2}\right)$

$SSE = \sigma^2 Y$ where $Y \sim \chi^2_{N-k}$ $\sim \text{Gamma}\left(\frac{N-k}{2}, \frac{1}{2\sigma^2}\right)$ - always

$SST = \sigma^2 W$ where $Z \sim \chi^2_{N-1}$

Also could be shown that

thus $R^2 = \frac{\sigma^2 X}{\sigma^2 Z} = \frac{X}{W} = \frac{X}{X+Y}$ Let $\frac{X}{X+Y} = Z$

$b = SST = SSM + SSE$

$\frac{X}{X+Y} \sim \text{Beta}(a, b)$

so $\frac{SSE}{SSM} \sim \text{Beta}\left(\frac{k-1}{2}, \frac{N-k}{2}\right)$

$Z = \frac{X}{X+Y}$ $Y = M$

Inverses: $X = \frac{ZW}{1-Z}$ $Y = M$

$J = \begin{vmatrix} \frac{dx}{dz} & \frac{dx}{dm} \\ \frac{dy}{dz} & \frac{dy}{dm} \end{vmatrix} = \begin{bmatrix} \frac{M}{(1-Z)^2} & \frac{Z}{(1-Z)} \\ 0 & 1 \end{bmatrix}$ $|J| = \frac{M}{(1-Z)^2}$

X and Y are independent (in part b SSM and SSE indep.)

so, $f_{X,Y}(x,y) = f_X(x) f_Y(y)$

Then, $f_{Z,M}(z,m) = f_{X,Y}(x,y) \cdot |J|$

$f_{Z,M}(Z,M) = f_X\left(\frac{ZM}{1-Z}\right) f_Y(M) \cdot |J|$

$= \frac{1}{\Gamma(\frac{k-1}{2}) \cdot 2^{k-1/2}} \left(\frac{ZM}{1-Z}\right)^{\frac{k-1}{2}-1} e^{-\frac{1}{2}\left(\frac{ZM}{1-Z}\right)} \cdot \frac{1}{\Gamma(\frac{N-k}{2}) 2^{\frac{N-k}{2}}} e^{-\frac{M}{2}} \cdot \frac{M}{(1-Z)^2}$

$= \frac{1}{\Gamma(\frac{k-1}{2}) \Gamma(\frac{N-k}{2}) 2^{\frac{(k-1+N-k)}{2}}} Z^{\frac{k-1}{2}-1} (1-Z)^{-\frac{k-1}{2}-1} M^{\frac{k-1+N-k}{2}-1} e^{-\frac{1}{2}\left(\frac{M}{1-Z}\right)}$

$f_Z(z) = \int_0^\infty f_{Z,M}(z,m)\, dm$

$= \frac{1}{\Gamma(\frac{k-1}{2}) \Gamma(\frac{N-k}{2}) 2^{\frac{(k-1+N-k)}{2}}} Z^{\frac{k-1}{2}-1} (1-Z)^{-\frac{k-1}{2}-1} \cdot \int_0^\infty M^{\frac{N-1}{2}} e^{-\frac{1}{2}\left(\frac{M}{1-Z}\right)} dm$

$= \frac{1}{\Gamma(\frac{k-1}{2}) \Gamma(\frac{N-k}{2}) 2^{\frac{(k-1+N-k)}{2}}} Z^{\frac{k-1}{2}-1} (1-Z)^{-\frac{k-1}{2}-1} \left(\frac{\Gamma(\frac{N-1}{2})}{\left(\frac{1}{2(1-Z)}\right)^{\frac{N-1}{2}}}\right) \cdot \int_0^\infty \frac{\left(\frac{1}{2(1-Z)}\right)^{\frac{N-1}{2}}}{\Gamma(\frac{N-1}{2})} \cdot M^{\frac{N-1}{2}-1} e^{\frac{-1}{2(1-Z)}M} dw$

Gamma $a = \frac{N-1}{2}$ $b = \frac{1}{2(1-Z)}$

know pdf integrates to 1 so

$f_Z(z) = \frac{1}{\Gamma(\frac{k-1}{2}) \Gamma(\frac{N-k}{2}) 2^{\frac{N-1}{2}}} Z^{\frac{k-1}{2}-1} (1-Z)^{-\frac{k-1}{2}-1} \cdot \frac{\Gamma(\frac{N-1}{2})}{\left(\frac{1}{2(1-Z)}\right)^{\frac{N-1}{2}}}$

$= \frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{k-1}{2}) \Gamma(\frac{N-k}{2})} Z^{\frac{k-1}{2}-1} (1-Z)^{\frac{N-k}{2}-1}$

Beta$(k-1/2, N-k/2)$

$k=5$ and $n=10$ $N=50$

$k-1=4$ and $N-k=45$ $R^2 = \text{Beta}(2, \frac{50-5}{2} = 22.5)$

$F = \frac{MSM}{MSE} = \frac{SSM/(k-1)}{SSE/(N-k)}$ $\alpha = .05$

$F(k-1, N-k) \sim F(4,45)$

at $\alpha = .05$ F critical value $= 2.578$ qf(.95,4,45, lower.tail=TRUE)

Reject for fstat > 2.58

$R^2 \sim Beta(2, 22.5)$

$R^2_{.95} = 0.1752$ qbeta(.95,2,22.5, lower.tail=TRUE)

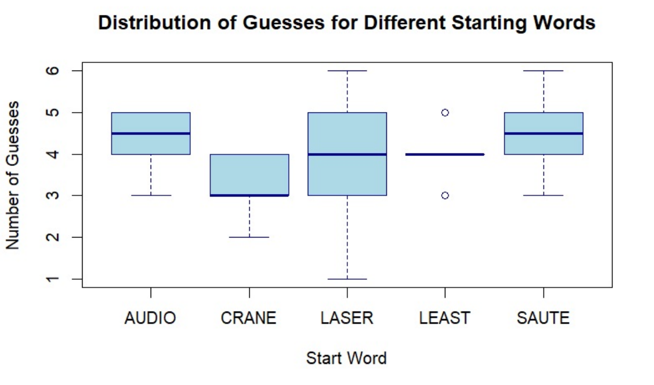
Reject for $R^2 > 0.18647$

A higher R^2 is equivalent to rejecting for a large f-stat. However R^2 is more intuitive because it explains that a large R^2 means a large variability btwn groups. Meaning Ho seems likely untrue.

Link to $R^2 \sim Beta(\frac{m}{2}, \frac{n}{2})$ proof: https://statproofbook.github.io/P/beta-chi2.html

√ d) Use my Wordle data to demonstrate, with $k = 5$ different start words and $n = 10$ games played with each start word. The responses are the numbers of attempts to guess the word (assume these are approximately Normal with a constant variance but possibly different means). Show a graph of the data, fill in the ANOVA table and compute the F-statistic and R^2.

Graph



ANOVA Table

Source of variation	Df	Sum of Squares	Mean Square	F-Stat	Pr(>F)
Between Groups	4	9.92	2.480	2.79	.0374
Within Groups	45	40.0	0.8889		
Total	49	49.92			

$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST} = \frac{9.92}{49.92} = 0.1987$

Both F-stat and R^2 allow us to reject Ho. Meaning each individual word does not yield the same average guesses.