Binomial ANOVA Test on different group

Assumption paper

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.0     v purrr   0.3.4
v tibble  3.1.7     v dplyr   1.1.0
v tidyr   1.2.0     v stringr 1.4.0
v readr   2.1.2     v forcats 0.5.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Test Logistics:

A complete walk-through about the choose of this parameter could be found in the Chi_Review Google doc, but the idea was inspired by the paper at below link:

https://sites.ualberta.ca/~lkgray/uploads/7/3/6/2/7362679/23_-_binomial_anova.pdf

Data simulation

Here at the simulation stage, where notice that we will be needing to fit a logistic regression based on the binary response variables (0, 1 as non-response and response) , and the only variable we can use to fit such variables is to factorize into different groups and fit them as factors.

Since we are given the response rate and total number of participants in each group, the data is rather easy to simulate without any information loss (for ex. wrong assumption on distribution or similar things):

Let's simulate them here and take a look at the resulting data::

```r
Assumption_1 <- data.frame(
  Response = sample(c(rep(TRUE, 27), rep(FALSE, 37-27)), 37 ,replace = F)) |>
  mutate (Group = 'Assumption 1') #double checked

Assumption_2 <- data.frame(
  Response = sample(c(rep(TRUE, 22), rep(FALSE, 33-22)), 33 ,replace = F)) |>
  mutate (Group = 'Assumption 2') #double checked

Assumption_3 <- data.frame(
  Response = sample(c(rep(TRUE, 24), rep(FALSE, 37-24)), 37 ,replace = F)) |>
  mutate (Group = 'Assumption 3') #double checked

reminder <- data.frame(
  Response = sample(c(rep(TRUE, 20), rep(FALSE, 30-20)), 30 ,replace = F)) |>
  mutate (Group = 'Reminder') #double checked



# combining
df <- Reduce(function(x, y) merge(x, y, all=TRUE),
             list(Assumption_1,Assumption_2,Assumption_3, reminder))
# and visualize the first 15
head(df,15)
```

```
   Response        Group
1     FALSE Assumption 1
2     FALSE Assumption 1
3     FALSE Assumption 1
4     FALSE Assumption 1
5     FALSE Assumption 1
6     FALSE Assumption 1
7     FALSE Assumption 1
8     FALSE Assumption 1
9     FALSE Assumption 1
10    FALSE Assumption 1
11    FALSE Assumption 2
12    FALSE Assumption 2
13    FALSE Assumption 2
14    FALSE Assumption 2
15    FALSE Assumption 2
```

At this point of time we have simulated the data we need, next we can move on to the glm model building which is fairly easy.

GLM (Logistics Model) Fitting:

Let's fit a simple logistic regression model on the factors that we have:

```
lr_model <- glm(Response ~ as.factor(Group), family = 'binomial', data = df)

summary(lr_model)
```

```
Call:
glm(formula = Response ~ as.factor(Group), family = "binomial",
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6176  -1.4464   0.7938   0.9005   0.9304

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                    0.9933     0.3702   2.683  0.00729 **
as.factor(Group)Assumption 2  -0.3001     0.5229  -0.574  0.56600
as.factor(Group)Assumption 3  -0.3801     0.5056  -0.752  0.45212
as.factor(Group)Reminder      -0.3001     0.5358  -0.560  0.57538
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 172.00  on 136  degrees of freedom
Residual deviance: 171.35  on 133  degrees of freedom
AIC: 179.35

Number of Fisher Scoring iterations: 4
```

Notice that here Residual Deviance does not indicate a goodness of fit due to the fact that our response variable is binary. A complete proof is given below:

Under a MLE Estimation (which R uses).

↓

Maximum Likelihood Estimation;

and ℓ defined as the log (natural) of the likelihood function, we have:

Null Deviance: ℓ(saturated) − ℓ(null model).

Residual Deviance: ℓ(saturated) − ℓ(proposed model).

in proposed model: $p_i(\beta) = \text{expit}(x_i^T\beta)$

in saturated model: $\tilde{p}_i = y_i/n \sim \text{Binomial}(n_i, p_i) \rightarrow \text{Bern}(p_i)$

as $y_i = 0, 1$.

hence the saturated model:

$$L = \prod_{i=1}^{n} f_i(y_i)$$

$$= \prod_{i=1}^{n} y_i^{p_i} (1-y_i)^{1-p_i}$$

therefore:

$$\Rightarrow \ell = \sum p_i \log y_i + (1-p_i)\log(1-y_i) = 0.$$

with $y_i = 1, 0$.

$$\Rightarrow \tilde{p}_i = 1, 0.$$

∴ Residual Deviance = ℓ(proposed model) ← Not reflecting anything.

Note that there is no need to worry about the proof at the current stage, it is explanation saved for people to later referred if they encounter questions.

Therefore we will compare it and we can print out the anova table below:

```
lr.anova= anova(lr_model, test="Chisq")
lr.anova
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Response

Terms added sequentially (first to last)

                 Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                136     172.00
as.factor(Group)  3   0.6481       133     171.35   0.8853
```

Therefore, by looking at a GLM fitted Binomial ANOVA table,

with the

H0: the means of response between groups are the same i.e.

$\mu_{Assumption1} = \mu_{Assumption2} = \mu_{Assumption3} = \mu_{Reminder}$

where μ represents the average response rate within the group

H1: at least one of the means of the response rate if different from others.

Therefore on a $\chi^2(df = 3)$ test we have reached a test statistics of 0.6481, with P-value 0.8853, therefore we cannot reject the null hypothesis and hence stating:

we have no enough information to state that the group averages in response rate between Assumption 1, Assumption 2, Assumption 3, and reminder differed.