Untitled

Background

Survey data typically come in the form of likert 5 and 7 point scales and are categorical in nature. This limits analysis to χ^2 analysis, Fischer's exact test and conversion of the categorical scales to numeric values and a t-test analysis. The former doesn't lead to understanding of pairwise differences and the latter isn't faithful to the categorical nature of the variables.

Proposal

Use Bayesian analysis and count data in order to understand the differences in the data.

Method

Generate some fake survey data, pass through a Binomial liklihood with a Beta prior. Summarise the posteriors to understand the differences.

Fake Data

```r
set.seed(336)

questions <- factor(
  c("Strongly agree", "Agree", "Neither", "Disagree", "Strongly Disagree"),
  c("Strongly Disagree", "Disagree", "Neither", "Agree", "Strongly agree"))

# Establish the vectors of probabilties
p1 <- c(.25, .25, .05, .30, .15)
p2 <-c(.25, .20, .02, .33, .20)

group_1 <- rbinom(5, 100, p1)+rbinom(5,5,.25)
group_2 <- rbinom(5, 100, p2)++rbinom(5,5,.25)

df <- data.frame(questions, group_1, group_2)

knitr::kable(df)
```

questions	group_1	group_2
Strongly agree	28	28
Agree	28	21
Neither	5	1
Disagree	30	32
Strongly Disagree	14	31

Frequentist method

For comparison:

```
chisq.test(df$group_1, df$group_2)
```

```
## Warning in chisq.test(df$group_1, df$group_2): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$group_1 and df$group_2
## X-squared = 15, df = 12, p-value = 0.2414
```

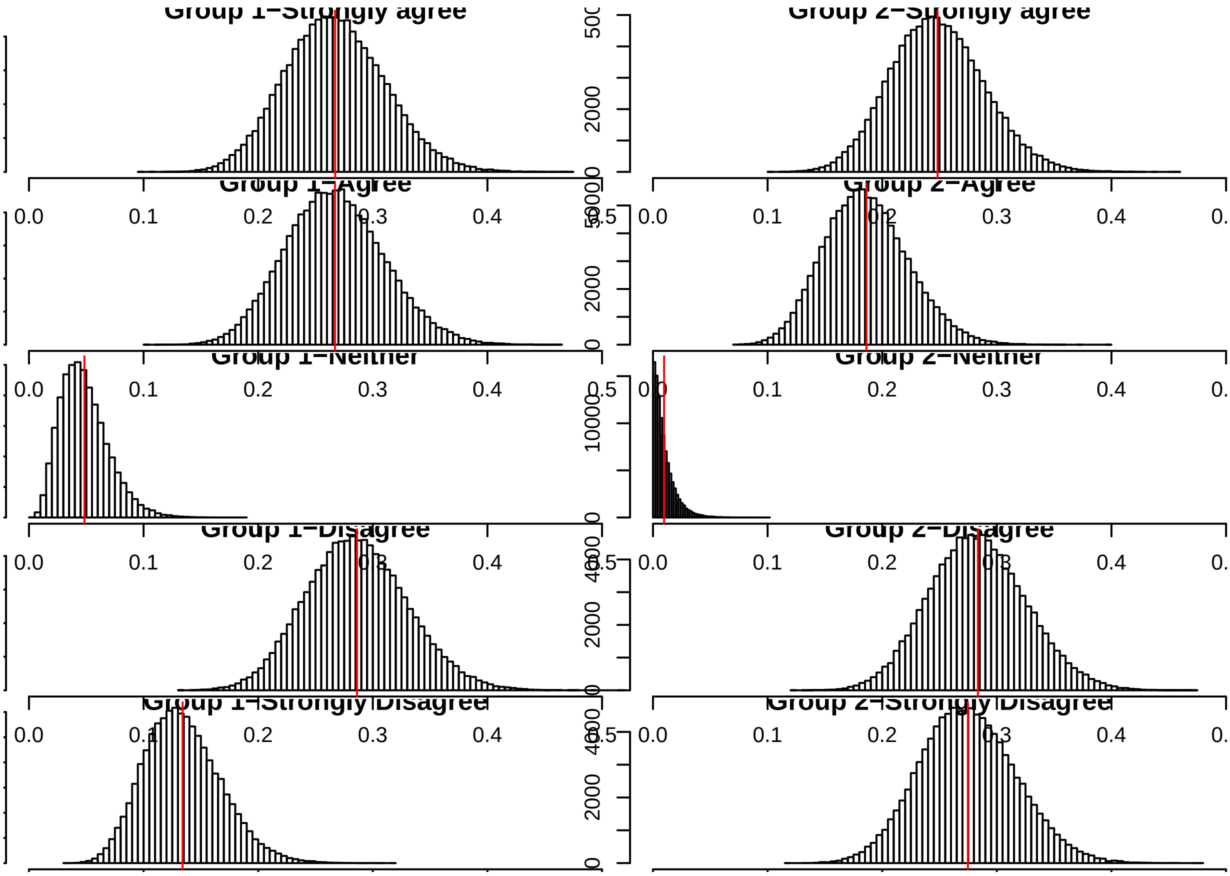
Results are not significant, these two distributions are drawn from the same distribution…so the frequentist analysis indicates.

Posterior Draws

Set up for the Bayesian analysis

```
#Prep for MC
group_1_n <- sum(group_1)
group_2_n <- sum(group_2)

#Priors
a <- .1
b <- .1

#Iterations
sampz <- 100000

#Draw from posterior
group_1_posterior <-list()
for(i in 1:length(group_1)){
  group_1_posterior[[i]] <- rbeta(sampz,group_1[i]+a,group_1_n-group_1[i]+b)

}

group_2_posterior <-list()
for(i in 1:length(group_2)){
  group_2_posterior[[i]] <- rbeta(sampz,group_2[i]+a,group_2_n-group_2[i]+b)

}
```

Examine Posterior Distribution



```
# library(tidyverse)
# #name my lists
# try_1<-list()
# for(i in 1:5){
#   catz <- rep(as.character(questions[i]),sampz)
#   try_1[[i]] <- group_1_posterior[[i]] %>%
#     as_data_frame() %>%
#     add_column(., catz)
# }
#
# try_2<-list()
# for(i in 1:5){
#   catz <- rep(as.character(questions[i]),sampz)
#   try_2[[i]] <- group_2_posterior[[i]] %>%
#     as_data_frame() %>%
#     add_column(., catz)
# }
```

```
# library(ggplot2)
# library(dplyr)
# try_1 %>%
#   bind_rows() %>%
#   mutate(group = "group_1") %>%
#   union(., try_2 %>%
#   bind_rows() %>%
#   mutate(group = "group_2")) %>%
#   ggplot(aes(value, color = group, group = group))+
#   geom_density()+
#   facet_wrap(~catz)+theme_minimal()
```

Summarise Results

First, I need a helper function to summarise the two different groups.

```
#Helper Function
library(purrr)

summarise_posterior <- function(x,y){
  mu_1 <- map_dbl(x, mean)
  mu_2 <- map_dbl(y, mean)

  sd_1 <- map_dbl(x, sd)
  sd_2 <- map_dbl(y, sd)

  delta_<-list()
  pooled_sd <- list()
  pro_difference <-vector()
  for(i in 1:length(x)){
    delta_[[i]] <- x[[i]]-y[[i]]
    delta_mu <- map_dbl(delta_, mean)
    pooled_sd[[i]] <-sqrt(var(x[[i]]) + var(y[[i]]))
    cohens <-delta_mu/ unlist(pooled_sd)
    pro_difference[i] <- max(mean(x[[i]] > y[[i]]),
                             1-mean(x[[i]] > y[[i]]))
  }

  out <- cbind(mu_1, mu_2, sd_1, sd_2,delta_mu, cohens,pro_difference)

  out
}
```

Now look at the results in table form.

```
statz <- summarise_posterior(group_1_posterior, group_2_posterior)

knitr::kable(cbind(statz, p1, p2 ), digits = 3)
```

mu_1	mu_2	sd_1	sd_2	delta_mu	cohens	pro_difference	p1	p2
0.267	0.248	0.043	0.040	0.019	0.319	0.625	0.25	0.25
0.267	0.186	0.043	0.036	0.081	1.439	0.923	0.25	0.20
0.048	0.010	0.021	0.009	0.039	1.704	0.973	0.05	0.02
0.286	0.283	0.044	0.042	0.003	0.045	0.517	0.30	0.33
0.134	0.275	0.033	0.042	-0.141	-2.643	0.995	0.15	0.20

Summary

Now we can look at the individual differences as well as the individual propabilties of difference.