hw 5

Wurongyan Zhang

11/5/2019

problem 1

```
library(tidyverse)
library(ggplot2)
library(ggridges)
library(gridExtra)

set.seed(10)

iris_with_missing = iris %>%
  map_df(~replace(.x, sample(1:150, 20), NA)) %>%
  mutate(Species = as.character(Species))
```

```
colSums(is.na(iris_with_missing)) %>%
  knitr::kable()
```

	x
Sepal.Length	20
Sepal.Width	20
Petal.Length	20
Petal.Width	20
Species	20

As we can see from the summary above, the data set iris_with_missing has 20 missing values in each of the 5 variables.

```
na_func = function(x){
  if(is.character(x)){
    x=replace_na(x,"virginica")
  }
  else if(is.numeric(x)){
    x=replace_na(x, round(mean(x,na.rm=TRUE),digits=1))
  }
  x
}

iris=map_dfr(iris_with_missing,na_func)
```

```
colSums(is.na(iris)) %>%
  knitr::kable()
```

	x
Sepal.Length	0
Sepal.Width	0
Petal.Length	0
Petal.Width	0

	x
Species	0

As we can see from the second table, there is no missing values after the function of replacement.

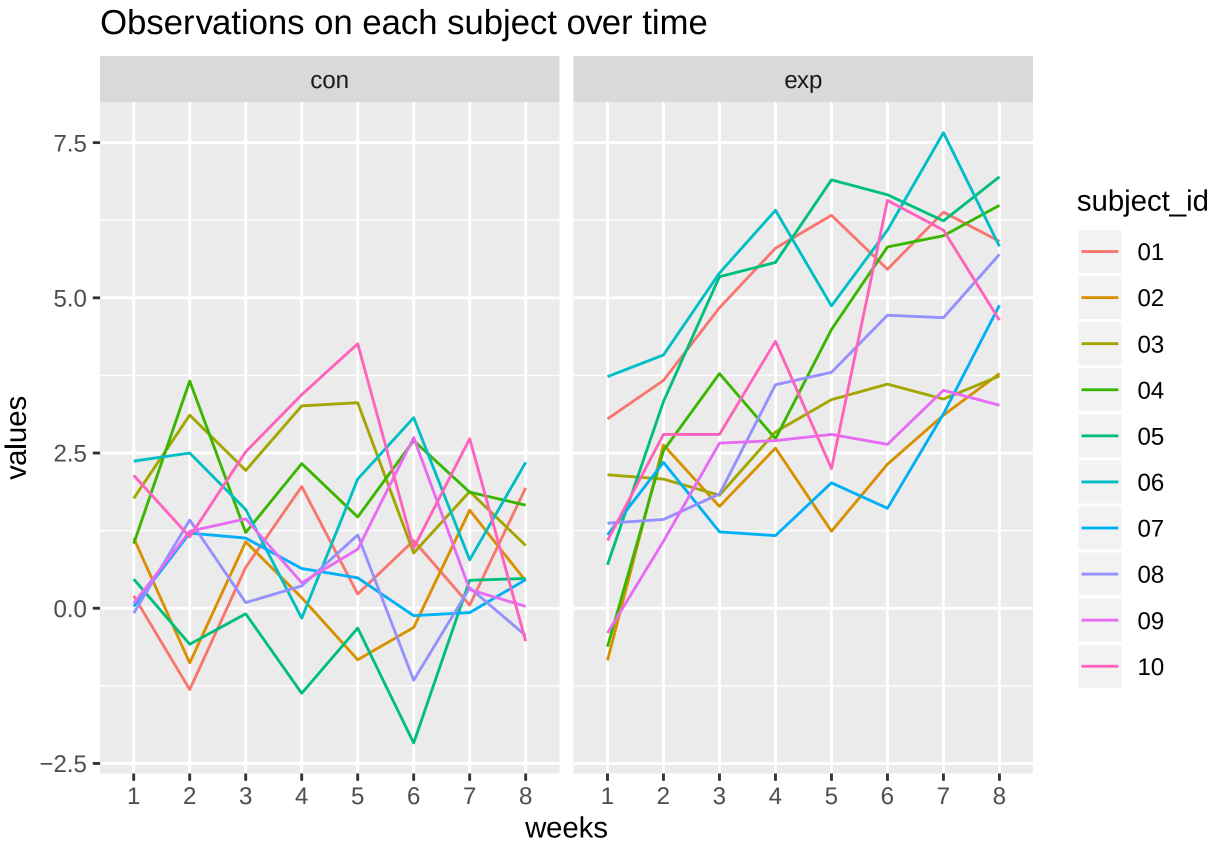
problem 2

```
file = list.files("data")

file_data = purrr::map_dfr( str_c("./data/",file), read_csv) %>%
 janitor::clean_names() %>%
  mutate(file_name=file) %>%
  mutate(file_name=str_remove(file_name,".csv")) %>%
  separate(file_name, into = c("arm","subject_id"),sep="_") %>% arrange(arm,subject_id) %>%
  select(subject_id, arm, everything())

 file_data %>%
   knitr::kable()
```

subject_id	arm	week_1	week_2	week_3	week_4	week_5	week_6	week_7	week_8
01	con	0.20	-1.31	0.66	1.96	0.23	1.09	0.05	1.94
02	con	1.13	-0.88	1.07	0.17	-0.83	-0.31	1.58	0.44
03	con	1.77	3.11	2.22	3.26	3.31	0.89	1.88	1.01
04	con	1.04	3.66	1.22	2.33	1.47	2.70	1.87	1.66
05	con	0.47	-0.58	-0.09	-1.37	-0.32	-2.17	0.45	0.48
06	con	2.37	2.50	1.59	-0.16	2.08	3.07	0.78	2.35
07	con	0.03	1.21	1.13	0.64	0.49	-0.12	-0.07	0.46
08	con	-0.08	1.42	0.09	0.36	1.18	-1.16	0.33	-0.44
09	con	0.08	1.24	1.44	0.41	0.95	2.75	0.30	0.03
10	con	2.14	1.15	2.52	3.44	4.26	0.97	2.73	-0.53
01	exp	3.05	3.67	4.84	5.80	6.33	5.46	6.38	5.91
02	exp	-0.84	2.63	1.64	2.58	1.24	2.32	3.11	3.78
03	exp	2.15	2.08	1.82	2.84	3.36	3.61	3.37	3.74
04	exp	-0.62	2.54	3.78	2.73	4.49	5.82	6.00	6.49
05	exp	0.70	3.33	5.34	5.57	6.90	6.66	6.24	6.95
06	exp	3.73	4.08	5.40	6.41	4.87	6.09	7.66	5.83
07	exp	1.18	2.35	1.23	1.17	2.02	1.61	3.13	4.88
08	exp	1.37	1.43	1.84	3.60	3.80	4.72	4.68	5.70
09	exp	-0.40	1.08	2.66	2.70	2.80	2.64	3.51	3.27
10	exp	1.09	2.80	2.80	4.30	2.25	6.57	6.09	4.64

The data frame after cleaning is shown above.

```
file_data_week=file_data %>%
  pivot_longer(week_1:week_8,
               names_to="weeks",
               values_to = "values") %>%
  separate(weeks, into = c("week","weeks"),sep = "_")

file_data_week %>%
  ggplot(aes(x=weeks, y=values, group=subject_id, color=subject_id))+
  geom_line()+facet_grid(~arm) +
  labs(title = "Observations on each subject over time")
```



We can see from the plot that the observation values for experimental group are higher than control group on average for each person in each week on average. The values of experimental and control groups were similar at week 1 but the experimental group increased later on. Moreover, the experimental group shows increasing trend on values but the control group only fluctuate without an increasing or decreasing trend.

problem 3

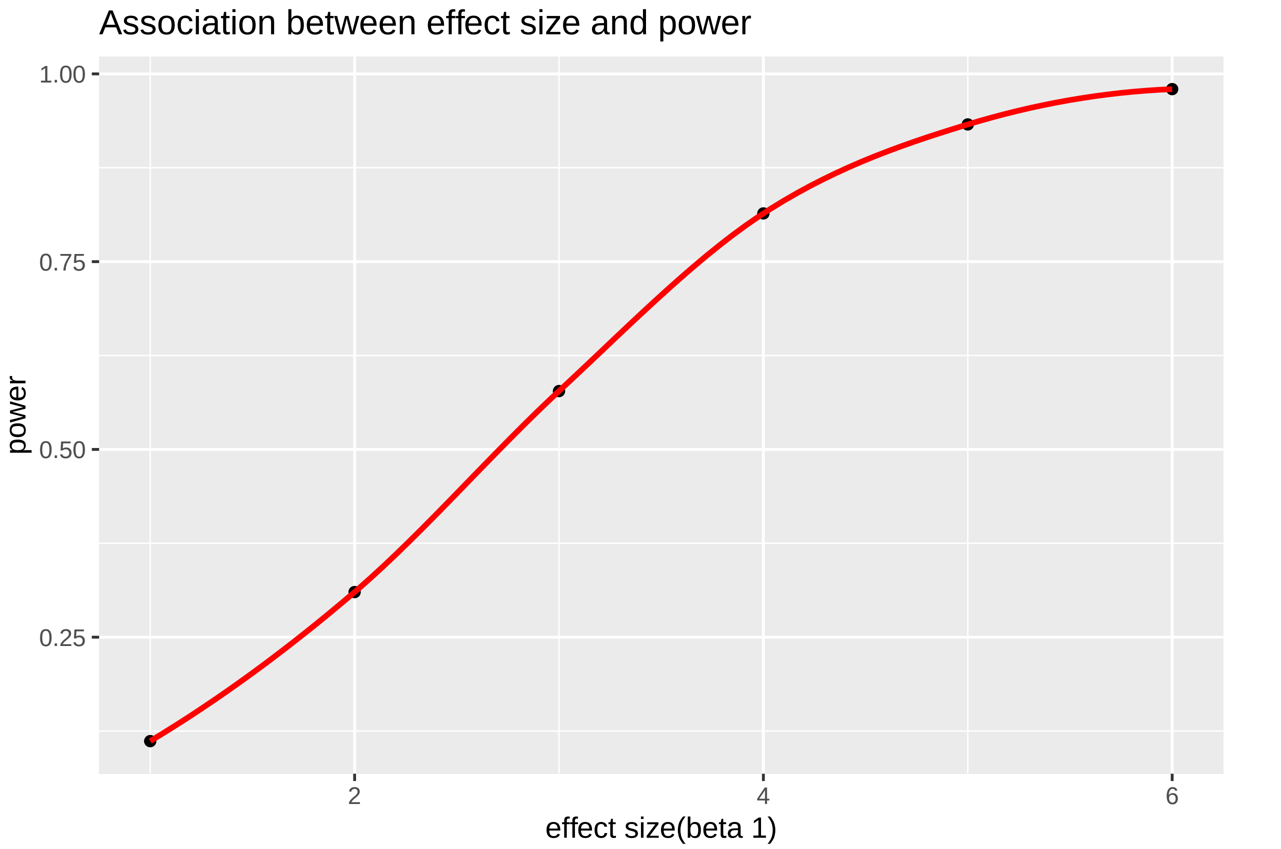
```
set.seed(100)

sim_regression= function(beta1,n=30, beta0=2,sigma_squared=50){
  sim_data= tibble(
    x=rnorm(n,mean=0, sd=1),
    y=beta0+beta1*x+rnorm(n,mean=0,sd=sqrt(sigma_squared))
  )

  ls_fit= lm(y~x, data=sim_data) %>%
    broom::tidy() %>%
    select(term, estimate, p.value) %>%
    mutate(term=recode(term, "x"="beta1_hat")) %>%
    filter(term=="beta1_hat")

}
```

```r
#generate 10000 datasets from the model
sim_results=
  rerun(10000, sim_regression(beta1=0)) %>%
  bind_rows()
```

```r
#repeat above for beta1=1,2,3,4,5,6
sim_results16=
  tibble(beta1=c(1:6)) %>%
  mutate(model= map(beta1,~rerun(10000, sim_regression(beta1=.x)))) %>%
  unnest() %>%
  unnest
```

```r
sim_results16 %>%
  group_by(beta1) %>%
  summarise(total=n(),
            alpha=sum(p.value<0.05)/total) %>% ggplot(aes(y=alpha, x=beta1)) +geom_point()+
  geom_smooth(color="red")+
  labs( title = "Association between effect size and power",
              x= "effect size(beta 1)", y= "power")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The relationship between effect size and power is positive and at a certain point the rate of increasing will decrease. Thus increase β_1 would increase power but the increase would not be very significant when β_1 reaches certain value.

```r
average =
  sim_results16 %>%
  pivot_wider(names_from = term,
              values_from = estimate) %>%
  group_by(beta1) %>%
  summarise(avg_beta=mean(beta1_hat))
```

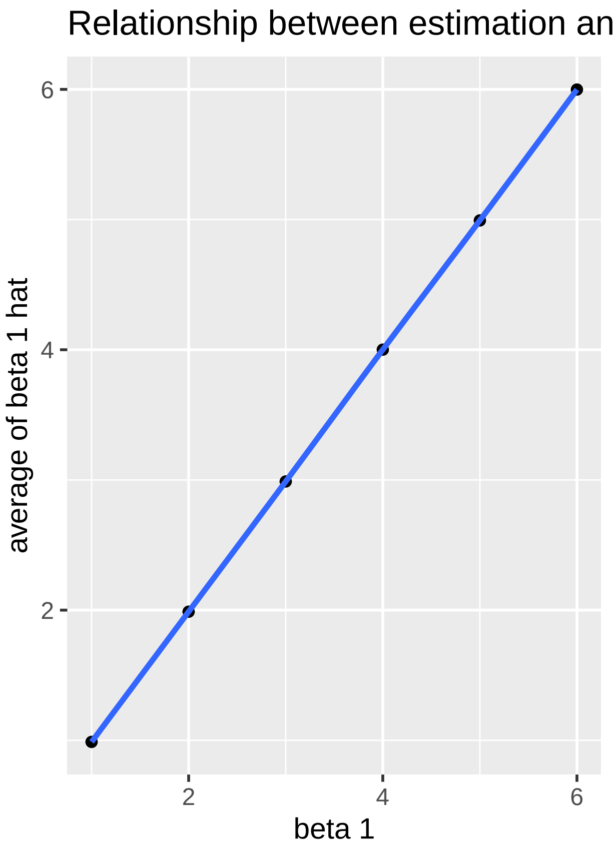
```
null_reject =
  sim_results16 %>%
  pivot_wider(names_from = term,
              values_from = estimate) %>% filter(p.value<0.05) %>% group_by(beta1) %>% summarise(avg_be

# average estimation
average %>%
  knitr::kable()
```

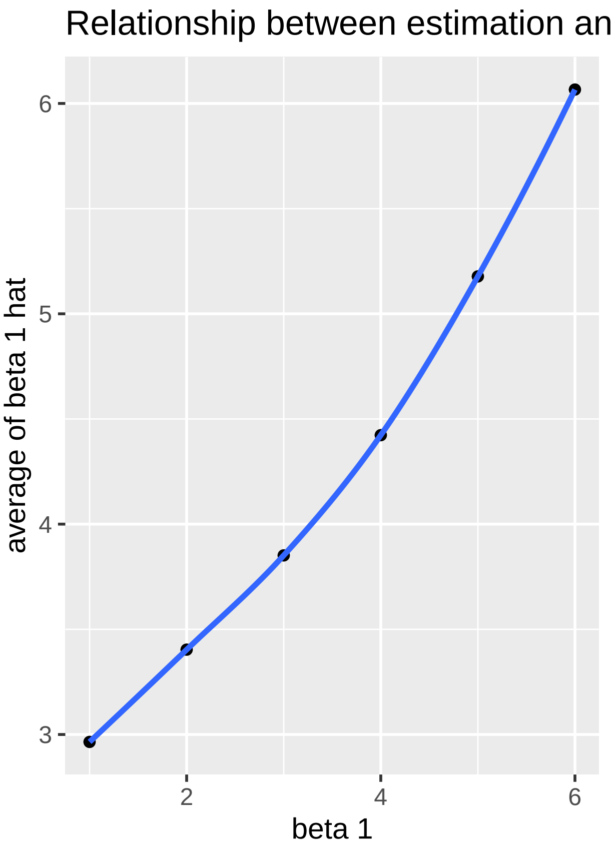
beta1	avg_beta
1	0.9879664
2	1.9881544
3	2.9882483
4	4.0007868
5	4.9933934
6	5.9985428

```
# average estimation of rejected data
null_reject %>%
  knitr::kable()
```

beta1	avg_beta_null
1	2.964758
2	3.403373
3	3.851496
4	4.423086
5	5.177751
6	6.066117

```
plot1=average %>% ggplot(aes(x=beta1, y=avg_beta))+ geom_point()+geom_smooth()+labs(title="Relationship

plot2=null_reject %>% ggplot(aes(x=beta1, y=avg_beta_null))+ geom_point()+geom_smooth()+labs(title="Rel

grid.arrange(plot1,plot2, nrow = 1)
```





As we can see from those two plots, sample average of $\hat{\beta}_1$ for which the null is rejected is not equal to the true value of β_1, and the sample average is always higher than the true value of β_1. However, at certain point, in our example approximately when β_1 equals to 6, the sample average of $\hat{\beta}_1$ is approximately equals to the true value of β_1. This can be explained by the power and the increase of effective size that as true value of β_1 increases, the probability of the sample to reject the null hypothesis($\beta_1 = 0$) given that the null hypothesis is false increases. Since the estimates follow normal distribution, as sample size increases, the sample mean became a good estimation, then the average of estimation would be approximately equals to the true value.