applied_ai

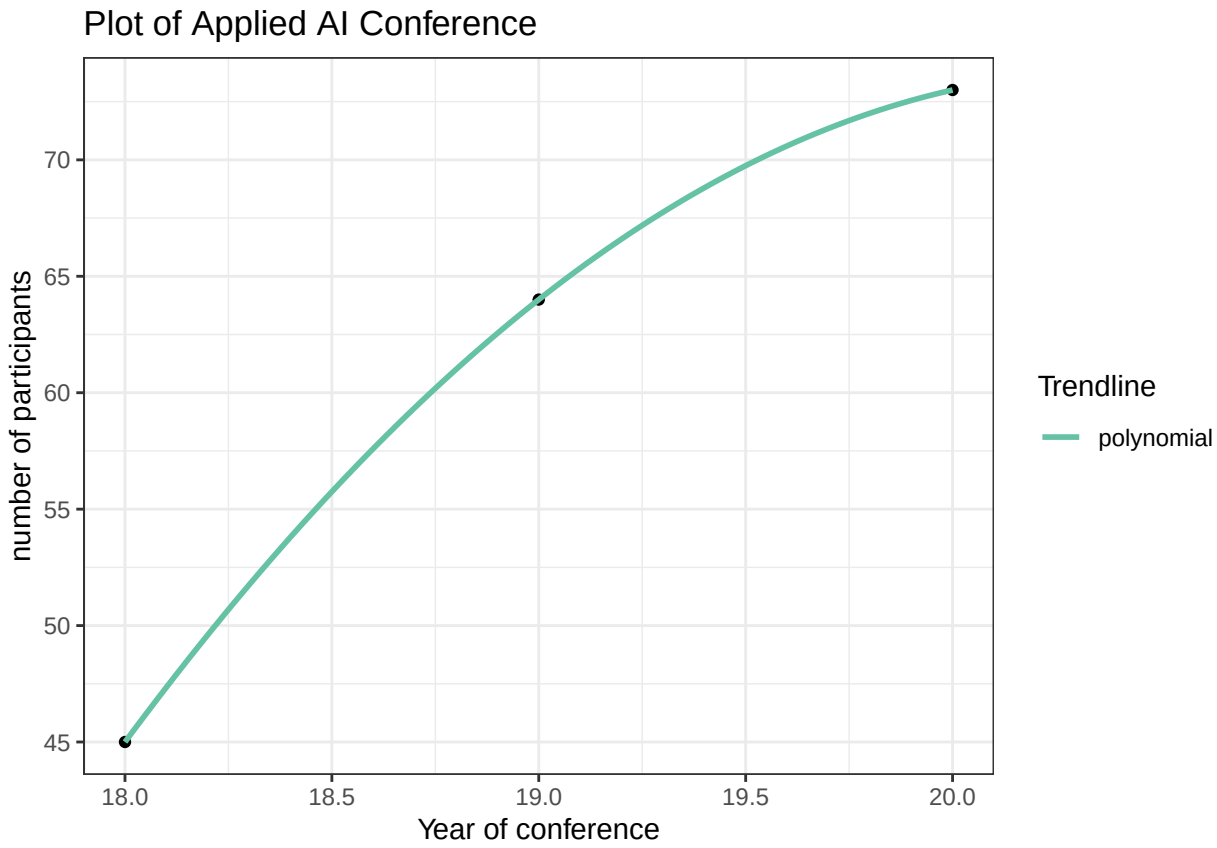
Ruijia Yang

11/07/2020

population

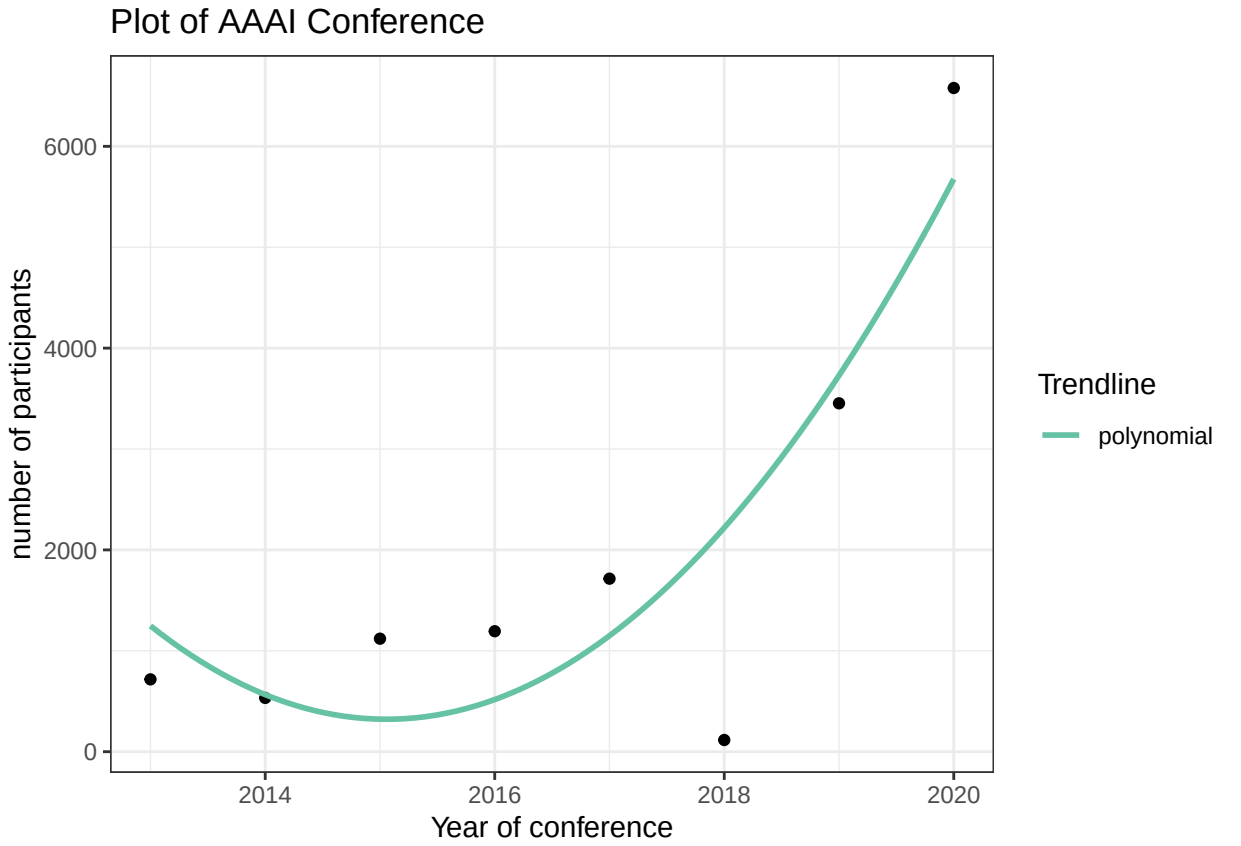
```
library("readxl")
aoolied_file <- read_excel("/Users/ricayang/Desktop/applied_race.xlsx")
applied_ai_x <- aoolied_file$year
applied_ai_y <- aoolied_file$total
library("ggplot2")
DF <- data.frame(applied_ai_x, applied_ai_y)

ggplot(DF, aes(x = applied_ai_x, y = applied_ai_y)) +
  geom_point() + ggtitle("Plot of Applied AI Conference") +
  scale_x_continuous(name="Year of conference") +
  scale_y_continuous(name="number of participants")+
  stat_smooth(method = 'lm', formula = y ~ poly(x,2), aes(colour = 'polynomial'), se= FALSE)  + theme_bu
  scale_colour_brewer(name = 'Trendline', palette = 'Set2')
```




```
aaai_file <- read_excel("/Users/ricayang/Desktop/aaai_race.xlsx")
aaai_x <- aaai_file$year
aaai_y <- aaai_file$total

DF2 <- data.frame(aaai_x, aaai_y)
ggplot(DF2, aes(x = aaai_x, y = aaai_y)) +
  geom_point() + ggtitle("Plot of AAAI Conference") +
  scale_x_continuous(name="Year of conference", limits=c(2013, 2020)) +
  scale_y_continuous(name="number of participants")+
  stat_smooth(method = 'lm', formula = y ~ poly(x,2), aes(colour = 'polynomial'), se= FALSE)  + theme_bw
  scale_colour_brewer(name = 'Trendline', palette = 'Set2')
```



For AAAI conference, there is a sudden big drop in 2018 due to the loss of data of all other participants except chairs and invited speakers, but there is still a increasing trend of attendees in the AAAI conference from 2013 to 2020. AAAI, as a long-history academica conference, there are a large group of committee every year from different places. Among participants in AAAI, in 2013, there are only about 2% participnats from Asia, (including China, Japan, Korea and India), while in 2020 there is at least 18% participants from Asia, and the other group of participants, for example, participants from North America, Europe and Oceania keeps constant. For Applied AI conference, a newly industry counference, the number of participants is increasing through 2018 to 2020. Since it is more likely that a industry conference shows regional concentration trend,the conference is helding at 1 city in 2018, 6 citites in 2020, and gping to be 11 cities in 2021. WIth the increasing number of participated city, the degree of diversity of this conference can be proved. Overall, the conferences in artificial intelligence field are having increasing attendees through these years, and more and more companies are participating, not only from different areas but different universities and companies. Therefore, taking the increasing population into account, conferences in artificial intelligence field are becoming more diverse through these years.

proportion

I used the two-proportions z-test to compare two observed proportions. Null hypothesis H_0: $p_{2018} = p_{2020}$ Alternative hypothesis H_a: $p_{2018} > p_{2020}$ where p_{2018} is the proportion of female in the 2018 Applied AI conference,
p_{2020} is the proportion of female in the 2020 Applied AI conference,

```
library(pwr)
X18_file <- read_excel("/Users/ricayang/Desktop/applied/18_file.xlsx")
```

```
## New names:
```

```
## * `` -> ...1
```

```
X20_file <- read_excel("/Users/ricayang/Desktop/applied/20_file.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
pwr.2p2n.test(h = 0.2, n1 = 45, n2 = 74, sig.level = NULL, power = 0.8)
```

```
##
##      difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.2
##              n1 = 45
##              n2 = 74
##       sig.level = 0.6618683
##           power = 0.8
##     alternative = two.sided
##
## NOTE: different sample sizes
```

```
x <- factor(X18_file$gender,levels=c("male","female"), labels=c(0,1))
y <- factor(X20_file$gender,levels=c("male","female"), labels=c(0,1))
prop.test(x = c(sum(x == 1),sum(y == 1)), n = c(45,74), alternative = "greater",conf.level = 0.34)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(x == 1), sum(y == 1)) out of c(45, 74)
## X-squared = 0.88282, df = 1, p-value = 0.1737
## alternative hypothesis: greater
## 34 percent confidence interval:
##  0.1170556 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.3555556 0.2567568
```

Conducting power analysis on the two sample, using significance level 0.66 can be enough to detect some difference, Since p-value $< \alpha$, there is enough evidance to reject null hypothesis with a 0.66 significance level, which suggests that the proportion of female in this conference is statistically different from 2018 to 2020, which means I am 90% confident that the true value of $p_{2018} - p_{2020}$ is in between 0.12 and 1.

aaai

Null hypothesis H_0: $p_{2013} = p_{2020}$ Alternative hypothesis H_a: $p_{2013} < p_{2020}$ where p_{2013} is the proportion of female in the 2013 AAAI conference,
p_{2020} is the proportion of female in the 2020 AAAI conference,
and my significance level is 0.1.

```
X13_aaai_file <- read_excel("/Users/ricayang/Desktop/aaai/13_aaai_file.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
X20_aaai_file <- read_excel("/Users/ricayang/Desktop/aaai/20_aaai_file.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
x <- factor(X13_aaai_file$gender,levels=c("male","female"), labels=c(0,1))
y <- factor(X20_aaai_file$gender,levels=c("male","female"), labels=c(0,1))
aaai_trend <- prop.test(x = c(sum(x == 1),sum(y == 1)), n = c(720,6581), alternative = "less",conf.level
aaai_trend
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(x == 1), sum(y == 1)) out of c(720, 6581)
## X-squared = 9.6894, df = 1, p-value = 0.0009267
## alternative hypothesis: less
## 90 percent confidence interval:
##  -1.00000000 -0.03191394
## sample estimates:
##    prop 1    prop 2
## 0.1861111 0.2385656
```

Since p-value $< \alpha$, there is enough evidance to reject null hypothesis, which suggests that the proportion of female in this conference is statistically different from 2013 to 2020, and the proportion of female is increasing through the years. And I am 90% confident that the true value of $p_{2020} - p_{2013}$ is in between 0.031 and 1.

same year comparsion

```
X18_aaai_file <- read_excel("/Users/ricayang/Desktop/aaai/18_aaai_file.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
X18_file <- read_excel("/Users/ricayang/Desktop/applied/18_file.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
x <- factor(X18_aaai_file$gender,levels=c("male","female"), labels=c(0,1))
y <- factor(X18_file$gender,levels=c("male","female"), labels=c(0,1))
year18_trend <- prop.test(x = c(sum(x == 1),sum(y == 1)), n = c(113,45), alternative = "less",conf.level
year18_trend
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(x == 1), sum(y == 1)) out of c(113, 45)
## X-squared = 1.9929, df = 1, p-value = 0.07902
## alternative hypothesis: less
## 90 percent confidence interval:
##  -1.000000000 -0.005348504
## sample estimates:
##    prop 1    prop 2
## 0.2300885 0.3555556
```

```
X19_aaai_file <- read_excel("/Users/ricayang/Desktop/aaai/19_aaai_file.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
X19_file <- read_excel("/Users/ricayang/Desktop/applied/19_file.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
x <- factor(X19_aaai_file$gender,levels=c("male","female"), labels=c(0,1))
y <- factor(X19_file$gender,levels=c("male","female"), labels=c(0,1))
year19_trend <- prop.test(x = c(sum(x == 1),sum(y == 1)), n = c(3457,64), alternative = "less", conf.lev
year19_trend
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(x == 1), sum(y == 1)) out of c(3457, 64)
## X-squared = 2.7329, df = 1, p-value = 0.04915
## alternative hypothesis: less
## 90 percent confidence interval:
##  -1.00000000 -0.01419927
## sample estimates:
##    prop 1    prop 2
## 0.2597628 0.3593750
```

```
x <- factor(X20_aaai_file$gender,levels=c("male","female"), labels=c(0,1))
y <- factor(X20_file$gender,levels=c("male","female"), labels=c(0,1))
year20_trend <- prop.test(x = c(sum(x == 1),sum(y == 1)), n = c(6581,74), alternative = "less", conf.lev
year20_trend
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(x == 1), sum(y == 1)) out of c(6581, 74)
## X-squared = 0.051942, df = 1, p-value = 0.4099
## alternative hypothesis: less
## 90 percent confidence interval:
```

```
##  -1.00000000  0.05406878
## sample estimates:
##    prop 1    prop 2
## 0.2385656 0.2567568
```

With a 0.1 confedence level α, comparing conference in the same 3 year, the proportion of female in AAAI conference is always smaller than the proportion of female in Applied AI conference in 2018 and 2019. However, in 2020, p-value is larger than α, which means there is not a statistically difference between the proportion of female of two conferences in 2020. Gender diversity is equitable or fair representation of people of different genders. It most commonly refers to an equitable ratio of men and women, the result of the same year comparison tells that the proportion of female is increasing these years, who contributes to the gender diversity of conference in artificial intelligence field.

Race

Categorizing participants' likely race/ethnicity according to US Census taxonomy: W_NL (white, non latino), HL (hispano latino), A (asian, non latino), B_NL (black, non latino).

In order to know is there difference race distribution between years of different conference, Using Chi-Square test of independence, assuming 0.1 significance level.

Ho: participants in each year and race are independent. Ha: participants in each year and race are not independent.

```
# aaai_race_dif
tulip2 <- c(210,63,111,333)
res2 <- chisq.test(tulip2, p = c(4907/6578,645/6578,451/6578,575/6578))
res2
```

```
##
##  Chi-squared test for given probabilities
##
## data:  tulip2
## X-squared = 1441.8, df = 3, p-value < 2.2e-16
```

```
# applied_race_dif
tulip <- c(22, 3,1, 19)
ind_res <- chisq.test(tulip, p = c(36/73, 1/73, 8/73,28/73))
```

```
## Warning in chisq.test(tulip, p = c(36/73, 1/73, 8/73, 28/73)): Chi-squared
## approximation may be incorrect
```

```
ind_res
```

```
##
##  Chi-squared test for given probabilities
##
## data:  tulip
## X-squared = 12.528, df = 3, p-value = 0.005778
```

Because of p-value $< \alpha$, race distribution and conference in different year are statistically significantly associated.

Using Chi-Square goodness of fit test, the goal is to compare multiple observed proportions to expected probabilities. If these race were equally distributed, the expected proportion would be 1/4 for each of the race.
H0: There is no significant difference between the observed and the expected value. Ha: There is a significant difference between the observed and the expected value.

```
# applied_ai_dif
tulip <- c(22, 3,1, 19)
equal_res <- chisq.test(tulip, p = c(1/4,1/4,1/4,1/4))
equal_res
```

```
##
##  Chi-squared test for given probabilities
##
## data:  tulip
## X-squared = 31, df = 3, p-value = 8.5e-07
```

```
# aaai_race_dif
tulip2 <- c(210,63,111,333)
eq_res2 <- chisq.test(tulip2, p = c(1/4,1/4,1/4,1/4))
eq_res2
```

```
##
##  Chi-squared test for given probabilities
##
## data:  tulip2
## X-squared = 238.53, df = 3, p-value < 2.2e-16
```

The p-value of both tests are less than the significance level alpha = 0.1. We can conclude that the races are significantly not commonly distributed and there is a significant difference between the observed and the expected value.

Below showing the trend line of race ratio in different conference. For AAAI conference, the number of Asian is increasing dramatically from 2013 to 2020, and the number of black, non latino also showing a increasing trend. While the percentage of White, non latino is decreasing significantly, the percentage of hispano latino is slightly decreasing. Adding a 0.25 reference line, assume that if each line approach 0.25 more, the participant coposition of race for the conference is more equal. The ratio of Asian and White nearly composite the main part of particiants (each above 0.25) For Applied AI conference, the number of hispano in increasing while the black, non hispano in decreasing and Asian and White,non latino dominate the participants through 2018 to 2020. We can see that there might be some issues in the race distribution of the two conference through these year.