

Lab5_groupdata

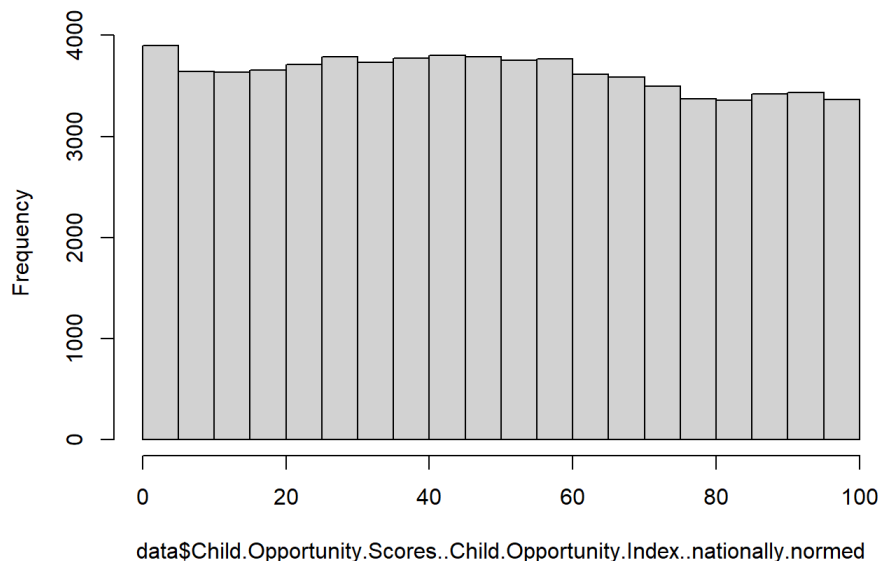
Meiheng_Liang

2024-09-26

```
data<-read.csv('final_v2_0924.csv', header = TRUE, sep = ",", stringsAsFactors = FALSE)
#View(data)
```

```
# (Min. 5), plot the distribution of one single variable and comment whether it is appearing to be Normal or Skewed which suggest Normality. If it skewed, please specify right skewed or Left skewed (Remember Skew=Tail)
hist(data$Child.Opportunity.Scores..Child.Opportunity.Index..nationally.normed)
```

gram of data\$Child.Opportunity.Scores..Child.Opportunity.Index..nationall



```
print("the distribution is slightly right skewed, indicating majority of census are ranked at medium to lower levels of environment with children opportunity")
```

```
## [1] "the distribution is slightly right skewed, indicating majority of census are ranked at medium to lower levels of environment with children opportunity"
```

```
##To start understanding the relationship of your key variables to the outcome variable, conduct individual correlation of the five variables with the outcome variable of interest
#For your kick off on correlation analysis with R please use:
#If Y is your outcome variable from your research question, and X is one of the predictors
```

```
#Obesity
Y <- data$Child.Opportunity.Scores..Child.Opportunity.Index..nationally.normed
X1 <- data$Obesity.among.adults
coi_obese <- cor(X1, Y, use = "complete.obs")
coi_obese
```

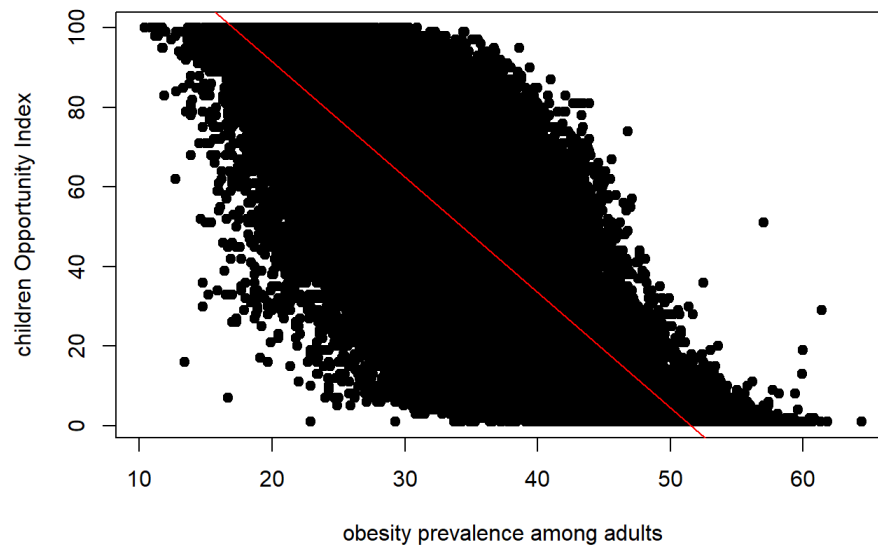
```
## [1] -0.7362566
```

```
print(paste("correlation between children opportunity and prevalence of Obesity is", coi_obese))
```

```
## [1] "correlation between children opportunity and prevalence of Obesity is -0.736256554953439"
```

```
#plot correlations:
plot(X1, Y, main = "Scatter Plot of obesity prevalence among adults vs children Opportunity Index", xlab = "obesity prevalence among adults", ylab = "children Opportunity Index", pch = 19)
abline(lm(Y ~ X1), col = "red") # Add a regression line
```

Scatter Plot of obesity prevalence among adults vs children Opportunity Ir



```
#stroke
X2 <- data$Stroke.among.adults
coi_stroke <- cor(X2, Y, use = "complete.obs")
coi_stroke
```

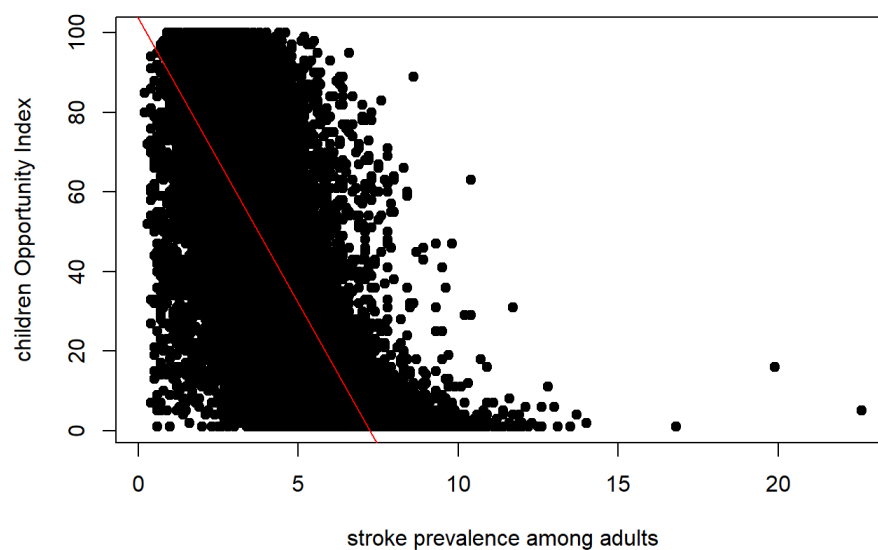
```
## [1] -0.6687204
```

```
print(paste("correlation between children opportunity and prevalence of stroke is", coi_stroke))
```

```
## [1] "correlation between children opportunity and prevalence of stroke is -0.6687203567762"
```

```
#plot
plot(X2, Y, main = "Scatter Plot of stroke prevalence among adults vs children Opportunity Index", xlab = "stroke prevalence among adults", ylab = "children Opportunity Index", pch = 19)
abline(lm(Y ~ X2), col = "red") # Add a regression Line
```

Scatter Plot of stroke prevalence among adults vs children Opportunity In



```
#Depression
X3 <- data$Depression.among.adults
coi_depress <- cor(X3, Y, use = "complete.obs")
coi_depress
```

```
## [1] -0.29544
```

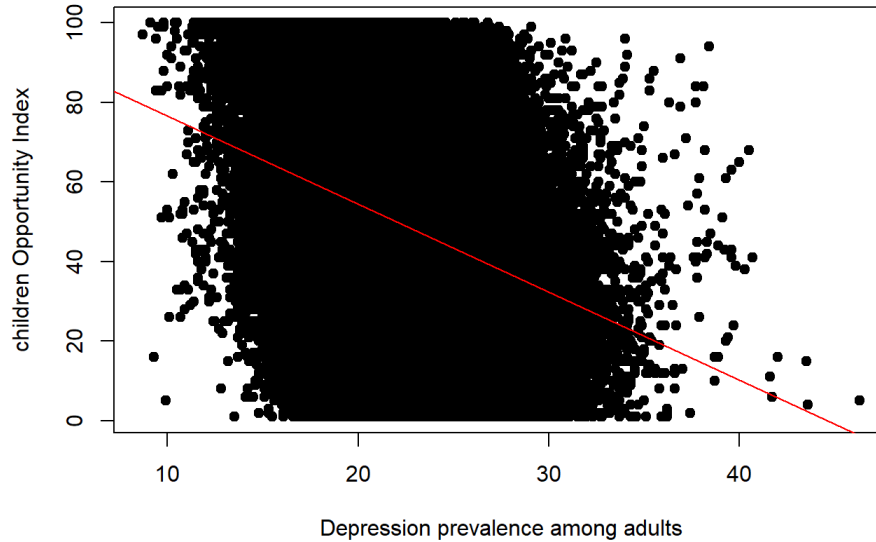
```
print(paste("correlation between children opportunity and prevalence of depression is", coi_depress))
```

```
## [1] "correlation between children opportunity and prevalence of depression is -0.295440026825368"
```

```
#plot correlations:
```

```
plot(X3, Y, main = "Scatter Plot of Depression prevalence among adults vs children Opportunity Index", xlab = "Depression prevalence among adults", ylab = "children Opportunity Index", pch = 19)  
abline(lm(Y ~ X3), col = "red") # Add a regression line
```

Scatter Plot of Depression prevalence among adults vs children Opportunity



```
#Diabetes
```

```
X4 <- data$Diagnosed.diabetes.among.adults  
coi_diabetes <- cor(X4, Y, use = "complete.obs")  
coi_diabetes
```

```
## [1] -0.7141614
```

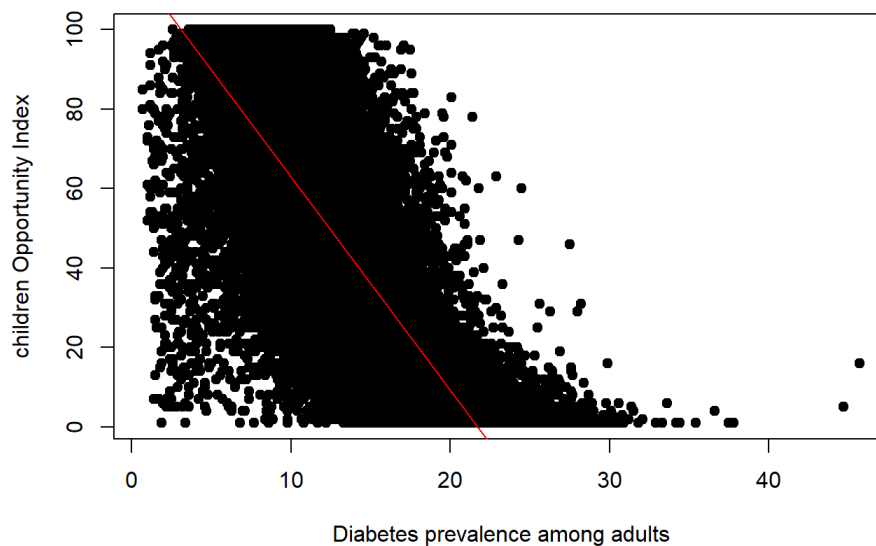
```
print(paste("correlation between children opportunity and prevalence of diabetes is", coi_diabetes))
```

```
## [1] "correlation between children opportunity and prevalence of diabetes is -0.714161382545456"
```

```
#plot correlations:
```

```
plot(X4, Y, main = "Scatter Plot of Diabetes prevalence among adults vs children Opportunity Index", xlab = "Diabetes prevalence among adults", ylab = "children Opportunity Index", pch = 19)  
abline(lm(Y ~ X4), col = "red") # Add a regression line
```

Scatter Plot of Diabetes prevalence among adults vs children Opportunity I



```
summary(data$Obesity.among.adults)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's  
##    10.40   29.60   34.60   34.43   39.20   64.40  12450
```

```
obes_sd <- sd(data$Obesity.among.adults, na.rm=TRUE)  
#probability X is less than (or equal to 39.20)  
pnorm(39.20, 34.43, obes_sd)
```

```
## [1] 0.7472924
```

```
#probability X is greater than 39.20  
pnorm(39.20, 34.43, obes_sd, lower.tail = FALSE)
```

```
## [1] 0.2527076
```

```
#probability Z is less than (or equal to) 1  
pnorm(1)
```

```
## [1] 0.8413447
```

```
#probability Z is greater than 1  
pnorm(1, lower.tail = FALSE)
```

```
## [1] 0.1586553
```

```
#identify X value  
qnorm(0.841, 34.43, obes_sd)
```

```
## [1] 41.58203
```

```
qnorm(0.159, 34.43, obes_sd, lower.tail = FALSE)
```

```
## [1] 41.58203
```

```
#identify Z value  
qnorm(0.841)
```

```
## [1] 0.9985763
```

```
qnorm(0.159, lower.tail = FALSE)
```

```
## [1] 0.9985763
```

```
summary(data$Child.Opportunity.Scores..Child.Opportunity.Index..nationally.normed)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's  
##      1.00   25.00   49.00   49.49   74.00   100.00  23388
```

```
coi_sd <- sd(data$Child.Opportunity.Scores..Child.Opportunity.Index..nationally.normed, na.rm=TRUE)  
coi_sd
```

```
## [1] 28.63313
```

```
# a) What number of census tract in the upper 2.5% of the children opportunity distribution? In other words, what is the 97.5 percentile of this index distribution?  
coi_high_quatile<-qnorm(0.975, mean = 49.49, sd = coi_sd)  
  
print(paste(coi_high_quatile, " is the 97.5 percentile of this index distribution "))
```

```
## [1] "105.609899927513 is the 97.5 percentile of this index distribution "
```

```
#b) What proportion of census who give birth are 21 years of age or older?  
coi_quatile_75n_higher <-pnorm(74.00, mean = 49.49, sd = coi_sd, lower.tail = FALSE)  
  
print(paste(coi_quatile_75n_higher, "of census tract region are ranked at 75 or higher quatile "))
```

```
## [1] "0.195998481067549 of census tract region are ranked at 75 or higher quatile "
```

```
#2. hypertensive  
summary(data$High.blood.pressure.among.adults)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's  
##      4.10   28.90   33.40   33.79   38.30   80.20  17527
```

```
hyper_sd <- sd(data$High.blood.pressure.among.adults, na.rm=TRUE)  
  
high_hypertension_prevalence <- pnorm(50, mean = 33.79, sd = hyper_sd) - pnorm(40, mean = 33.79, sd = hyper_sd)  
print(paste(high_hypertension_prevalence, "of census tract area are likely to have population with high prevalence of hypertension"))
```

```
## [1] "0.186238567306772 of census tract area are likely to have population with high prevalence of hypertension"
```

```
#3. census tract classified as low education domain has low ranking index.  
summary(data$High.blood.pressure.among.adults)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's  
##      4.10   28.90   33.40   33.79   38.30   80.20  17527
```

```
hyper_sd <- sd(data$High.blood.pressure.among.adults, na.rm=TRUE)  
census<-length(data$High.blood.pressure.among.adults)  
  
#a) # Probability of prevalence > 40%  
pnorm(40, mean = 33.79, sd = hyper_sd, lower.tail = FALSE)
```

```
## [1] 0.2003796
```

```
#b) How many census tracts have a prevalence higher than 50%?  
high_prevalence_count <- sum(data$High.blood.pressure.among.adults > 38.30, na.rm = TRUE)  
high_prevalence_count
```

```
## [1] 19355
```

```
#c)What is the probability that exactly 100 census tracts have a high blood pressure prevalence greater than 40%?  
prob_greater_40 <- pnorm(40, mean = 33.79, sd = hyper_sd, lower.tail = FALSE)  
print(paste("likelihood to randomly select a census with hypertension prevalence higher than 40% is", prob_greater_40))
```

```
## [1] "likelihood to randomly select a census with hypertension prevalence higher than 40% is 0.200379591938601"
```

```
prob_100 <- pbinom(100, size = census, prob = prob_greater_40)
prob_100
```

```
## [1] 0
```

```
print(paste("the probability of picking exact 100 census tracts with hypertension prevalence higher than 40% is", prob_100))
```

```
## [1] "the probability of picking exact 100 census tracts with hypertension prevalence higher than 40% is 0"
```

```
#d) What is the probability that between 100 and 200 census tracts have a prevalence above 40%?
prob_200 <- pbinom(200, size = census, prob = prob_greater_40)
prob_200
```

```
## [1] 0
```

```
prob_100 <- pbinom(100, size = census, prob = prob_greater_40)
prob_100
```

```
## [1] 0
```

```
prob_between <- prob_200 - prob_100
prob_between
```

```
## [1] 0
```

```
table(data$Child.Opportunity.Levels..Child.Opportunity.Index..nationally.normed)
```

```
##
##           High           Low Moderate Very High  Very Low
##    23388    14067    15001    15110    13578    14828
```

```
#1. What is the probability that a census tract has exactly 5 occurrences of a high child opportunity index?
lambda_high <- 14067
high<-dpois(5, lambda_high)
high
```

```
## [1] 0
```

```
#2. What is the probability that a census tract has 5 or fewer occurrences of a low child opportunity index?
lambda_low <- 15001
low<-ppois(5, lambda_low)
low
```

```
## [1] 0
```

```
#3. What is the probability that the occurrence of moderate child opportunity index is greater than 5 in a census tract?
Moderate <- 15110
med<-ppois(5, Moderate, lower.tail = FALSE)
med
```

```
## [1] 1
```

```
#1. Probability of observing 10,000 or fewer census tracts with high COI
```

```
lambda_high <- 14067
prob_10000_or_fewer_high <- ppois(10000, lambda = lambda_high)
print(paste("The probability of observing 10,000 or fewer tracts with high COI is", prob_10000_or_fewer_high))
```

```
## [1] "The probability of observing 10,000 or fewer tracts with high COI is 7.56464245640121e-287"
```

```
#2. Expected number of census tracts at a moderate COI level
lambda_moderate <- 15110
expected_moderate <- lambda_moderate
print(paste(expected_moderate, "tracts are expected to have a moderate COI level next year"))
```

```
## [1] "15110 tracts are expected to have a moderate COI level next year"
```

```
#3. Probability of observing 15,000 or fewer tracts with moderate COI
prob_moderate_less_15000 <- ppois(15000, lambda = lambda_moderate)
print(paste("The probability of observing 15,000 or fewer tracts with moderate child opportunity index is", prob_moderate_less_15000))
```

```
## [1] "The probability of observing 15,000 or fewer tracts with moderate child opportunity index is 0.18659199662587"
```

```
# Expected number of census tracts with high hypertension prevalence (e.g., >40%)
# prob_greater_40 previously calculated in Q3
# Total number of census tracts
n <- 95972
expected_high_hypertension <- n * prob_greater_40
print(paste(expected_high_hypertension, "census tracts are expected to have a hypertension prevalence greater than 40%"))
```

```
## [1] "19230.8301975315 census tracts are expected to have a hypertension prevalence greater than 40%"
```

```
#b) What is the expected number of census with hypertension prevalence <40%?
prob_less_than_40 <- pnorm(40, mean = 33.79, sd = hyper_sd)
expected_low_hypertension_Prev <- n * prob_less_than_40
print(paste(expected_low_hypertension_Prev, "census tracts are expected to have a hypertension prevalence less than 40%"))
```

```
## [1] "76741.1698024685 census tracts are expected to have a hypertension prevalence less than 40%"
```

```
#c) Probability of observing 500 or more tracts with high hypertension prevalence
prob_more_than_500_high_hypertension <- ppois(499, lambda = expected_high_hypertension, lower.tail = FALSE)
print(paste("The probability of observing 500 or more census tracts with hypertension prevalence > 40% is", prob_more_than_500_high_hypertension))
```

```
## [1] "The probability of observing 500 or more census tracts with hypertension prevalence > 40% is 1"
```

```
#d) Probability of observing less than 500 tracts with hypertension prevalence < 40%
prob_fewer_than_500 <- ppois(499, lambda = expected_low_hypertension_Prev)
print(paste("The probability of observing fewer than 500 census tracts with hypertension prevalence less than 40% is", prob_fewer_than_500))
```

```
## [1] "The probability of observing fewer than 500 census tracts with hypertension prevalence less than 40% is 0"
```

```
summary(data$Taking.medicine.to.control.high.blood.pressure.among.adults.with.high.blood.pressure)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.   Max.    NA's
##  12.90   73.80   77.80   76.51   80.70   95.10  17527
```

```
#a) What is the probability of observing 85% or more adults with high blood pressure taking medication in a specific county, assuming the national rate is the same as the mean rate?
```

```
pop_bp <- 12850 #assume Population of adults with high blood pressure in the county
```

```
rate_meds <- 76.51 / 100 # National rate (mean from the dataset), no normalization
```

```
lambda_meds <- pop_bp * rate_meds
```

```
ppois_85_or_more <- ppois(0.85 * pop_bp, lambda = lambda_meds, lower.tail = FALSE)
```

```
print(paste("The probability of observing 85% or more adults taking medication to control high blood pressure is", ppois_85_or_more))
```

```
## [1] "The probability of observing 85% or more adults taking medication to control high blood pressure is 1.5474865061131e-27"
```

```
#b) What is the observed rate of adults taking medication in this county in 2022 if 8,500 adults out of 12,850 are reported to take medication?
```

```
observed_meds <- 8500
```

```
# Observed rate in percentage
```

```
observed_rate <- (observed_meds / pop_bp) * 100
```

```
print(paste("The observed rate of adults taking medication is", observed_rate, "%"))
```

```
## [1] "The observed rate of adults taking medication is 66.147859922179 %"
```

c) In 2022, 8500 adults with high blood pressure were reported to take medication in the same county. Using the rate from part b), calculate the probability of 450,000 or more such cases.

```
observed_rate_b <- observed_meds / pop_bp
lambda_observed <- pop_bp * observed_rate_b
ppois_850_more <- ppois(8500 - 1, lambda = lambda_observed, lower.tail = FALSE)

print(paste("The probability of observing 8500 or more adults taking medication is", ppois_850_more))
```

```
## [1] "The probability of observing 8500 or more adults taking medication is 0.501442379801329"
```