# PennSUIP Biostatistics Workshop 2023

## Introduction to Biostatistics and R

*Nina Alfaro, MS (alfarom@pennmedicine.upenn.edu)*

### Installing packages

```
install.packages("PACKAGE")
library(PACKAGE)

install.packages("ggplot2")
library(ggplot2)
```

### Retrieve the working directory

```
getwd()
```

### Re-define the working directory

```
setwd()
```

### Reading a .csv file for Mac and PC

The initial portion is retrived using getwd()!

```
mydata <- read.csv("/Users/mkca/Downloads/mydata.csv", header = TRUE)
mydata <- read.csv("C:\\Users\\mkca\\Downloads\\mydata.csv", header = TRUE)
```

### Common arithmetic functions

```
sum()
sqrt()
round()
log()
exp()
abs()
```

## Create a variable

### Enclose characters in quotations

```
x <- 1
y <- "PREP"
```

## Create a vector

```
odd_num <- c(1, 3, 5)
pets <- c("cat", "dog")
```

## Retrieving a specific field entry

```
mydata$Age[1]
mydata[1, 2]
mydata[1, "Age"]
```

## Retrieving specified row(s) with all available variables

```
dataset[1, ]
dataset[c(1, 2, 3), ]
```

## Retrieve the data type for a given variable

```
class(mydata$Age)
```

## Retrieve a specific cohort within your dataset

```
subset(mydata, Age > 30)
subset(mydata, Gender == "Female")
```

## Descriptive Statistics

To ignore nulls, add the following: na.rm = TRUE

summary()
mean()
median()
quantile(mydata$Age, c(0.25, 0.50, 0.75))
range()
min()
max()
IQR()
sd()
var()

## Graphical methods in Base R

To add labels, add the following: xlab = "X-AXIS", ylab = "Y-AXIS", main = "Title"

hist()
plot()
barplot()
boxplot()
pie()

## Retrieve table of counts for a given variable

```
table(mydata$Gender)
```

## Retrieve table of proportions for a given variable

```
prop.table(table(mydata$Gender))
```

## Retrieve table of counts for 2 variables

```
table(mydata$Gender, mydata$HighCholesterol)
```

## Retrieve table of counts for 2 variables under a particular condition

```
table(mydata$Gender, HighCholes = mydata$HighCholesterol, Death = mydata$Death)
```

## Generate bar charts for counts and proportions

```
barplot(table(mydata$Gender))
barplot(prop.table(table(mydata$Gender)))
```

## Correlation between two variables, numerically and visually

```
cor(mydata$Age, mydata$BMI)
plot(mydata$Age, mydata$BMI, xlab = "Age", ylab = "BMI", main = "Age vs. BMI")
```

## Graphical methods in ggplot2

### hist(), plot(), barplot(), boxplot() equivalencies, respectively

```
ggplot(data = mydata, mapping = aes(x = BMI)) + geom_histogram()
ggplot(data = mydata, mapping = aes(x = Age, y = BMI)) + geom_point()
ggplot(data = mydata, mapping = aes(x = Gender)) + geom_bar()
ggplot(data = mydata, mapping = aes(x = Gender, y = BMI)) + geomboxplot()
```

## Common statistical tests for continuous data

```
t.test()
res.aov()
wilcox.test()
kruskal.test()
```

## Common statistical tests for binary data

```
prop.test()
chisq.test()
mcnemar.test()
binom.test()
fisher.test()
```

## Linear regression, simple and multiple, for a continuous outcome

```
lm()
```