USC Department Of Translational Genomics
Keck Medicine of USC

TRGN 527: Applied Data Science and Bioinformatics

UNIT I. Introduction and Basic Data Science

Week 3 - Lecture 1

Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S. | **Assistant Professor**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

David W. Craig, Ph.D. | **Professor and Vice Chair**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Co-Director, Institute of Translational Genomics

Topics

- Visualizing data, types of data, and **data distributions**. Data and data-types – **categorical and continuous data**. Kaplan Meier, Violin and heat maps.

Summarizing Your Data

- Problem:
  - Generate basic statistical summary of your data.
- Approach:
  - The summary function provides some useful statistics for vectors, matrices, factors, and data.

````
# Summarizing your data
```{R}
# Please load the file "trgn599.clinical.tsv" in your R-Studio.
clinical_data <- read.table("trgn599.clinical.tsv", header = TRUE, sep = "\t")
# Summarize your data
summary(clinical_data)
```

# Summarizing your data 2
```{R}
# Summarize list of vectors, apply to each list element
lapply(clinical_data, summary)
```
````

Elements of Structured Data

- Data comes from many sources:
  - Surveys
  - Sensor measurements
  - Events
  - Text
  - Images
  - Videos

- Much of this data is unstructured
  - Image: Collection of pixels
  - Texts are sequences of words
  - Clickstreams are sequences of actions

- Major challenge of Data Science is to harness this torrent of raw data into actionable information.

Key Terms for Data Types

- Continuous
  - Data that can take on any value in an interval.
    - Data that is measured.
    - Examples: height, weight, time in a race.
      - Synonyms: interval, float, numeric.

- Discrete
  - Data that can take on only integer values, such as counts.
    - Data that is counted.
    - Data that can only take certain values.
      - Examples: number of students in a class, results of rolling 2 dice: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12.
        - Synonyms: integer, count.

- Categorical
  - Data that can take on only a specific set of values representing a set of possible categories.
    - Examples: race, sex, educational level.
      - Synonyms: enums, enumerated, factors, nominal, polychotomous.

Key Terms for Data Types

- Binary
  - A special case of categorical data with just two categories of values.
    - Example: 0/1, true/false.
      - Synonyms: Dichotomous, logical, indicator, boolean.

- Ordinal
  - Categorical data that has an explicit ordering.
    - Rating happiness on the scale 1-10, positions in a race (1st , 2nd , 3rd …)
      - Synonyms: ordered factor.

Rectangular Data

- The typical frame of reference for an analysis in data science is a rectangular data object, like a spread sheet or database table.

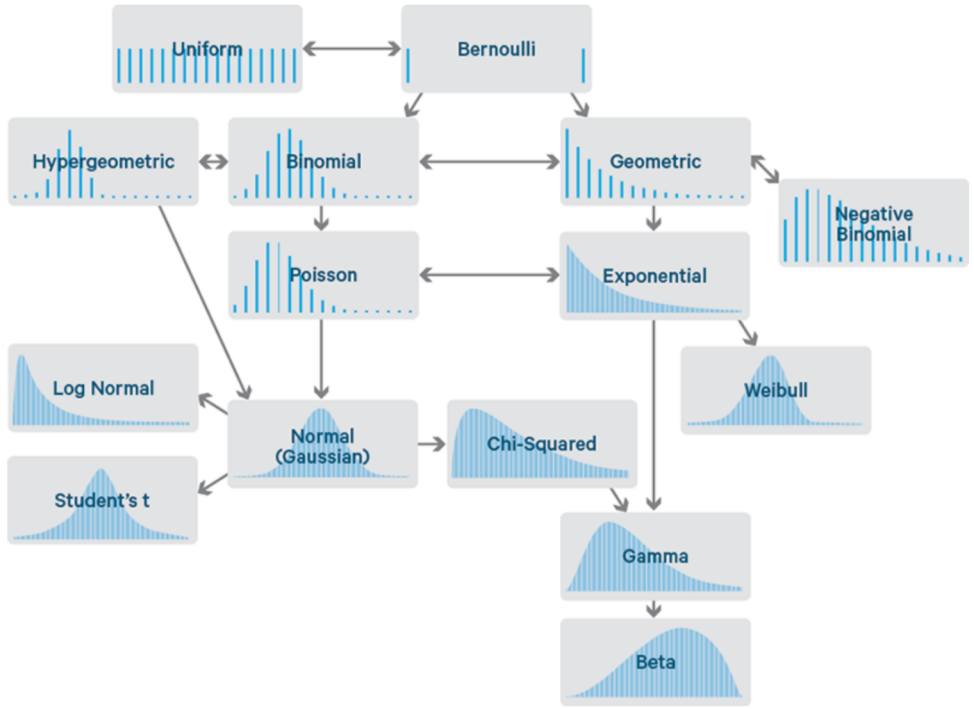
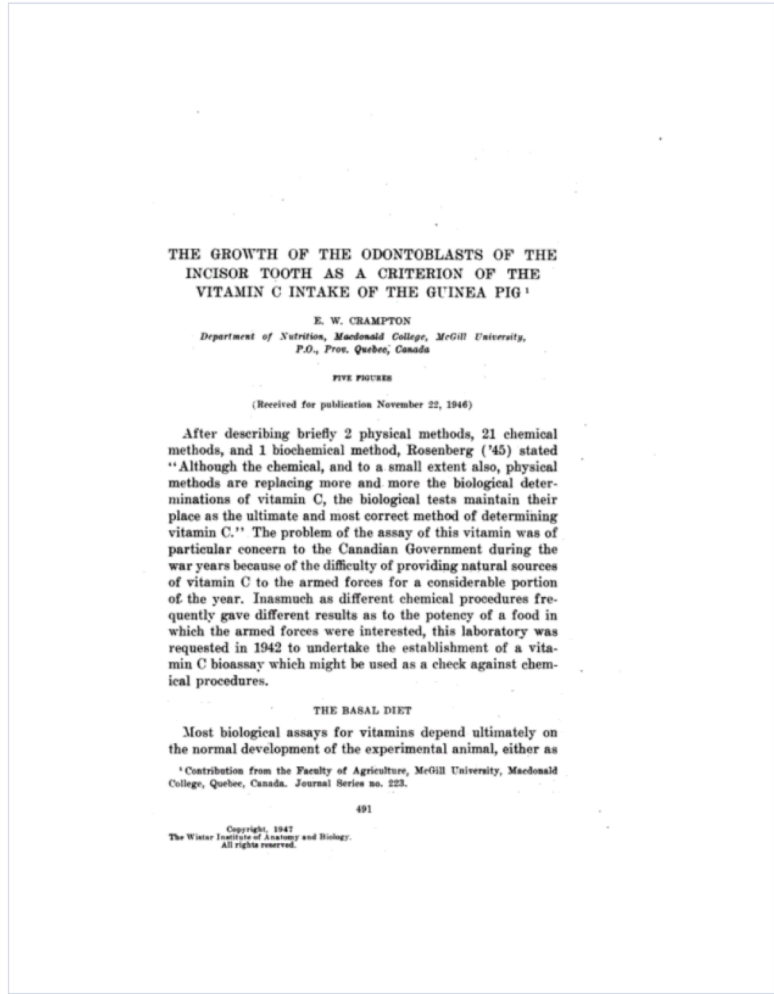
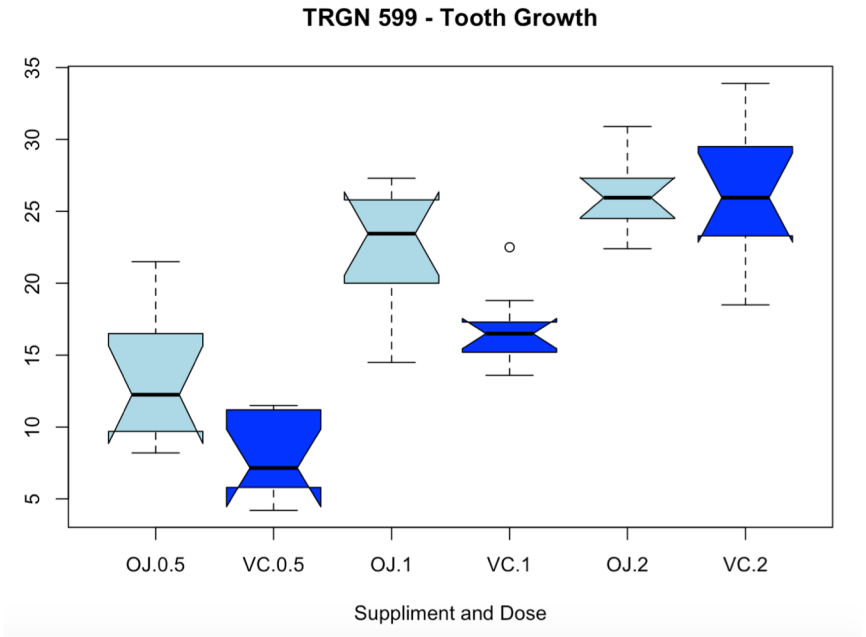
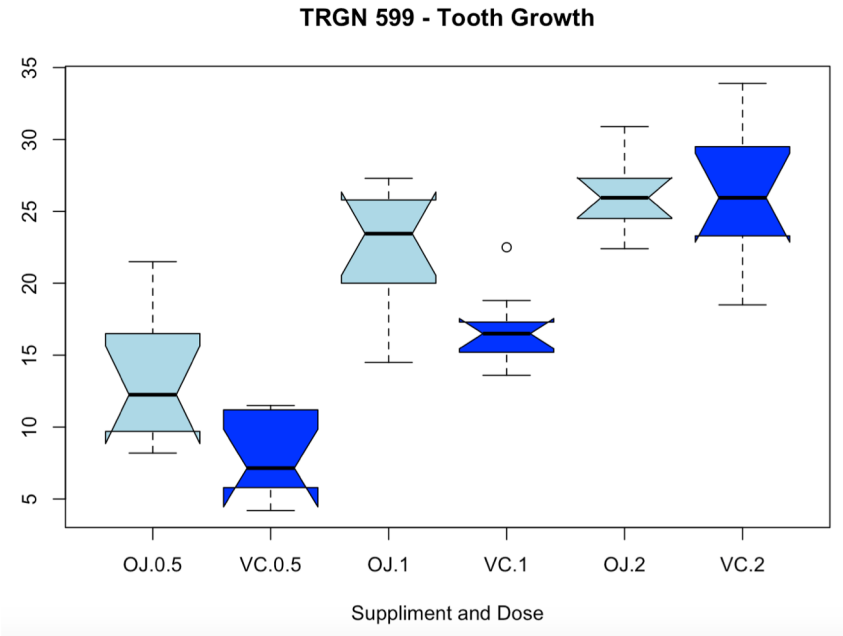
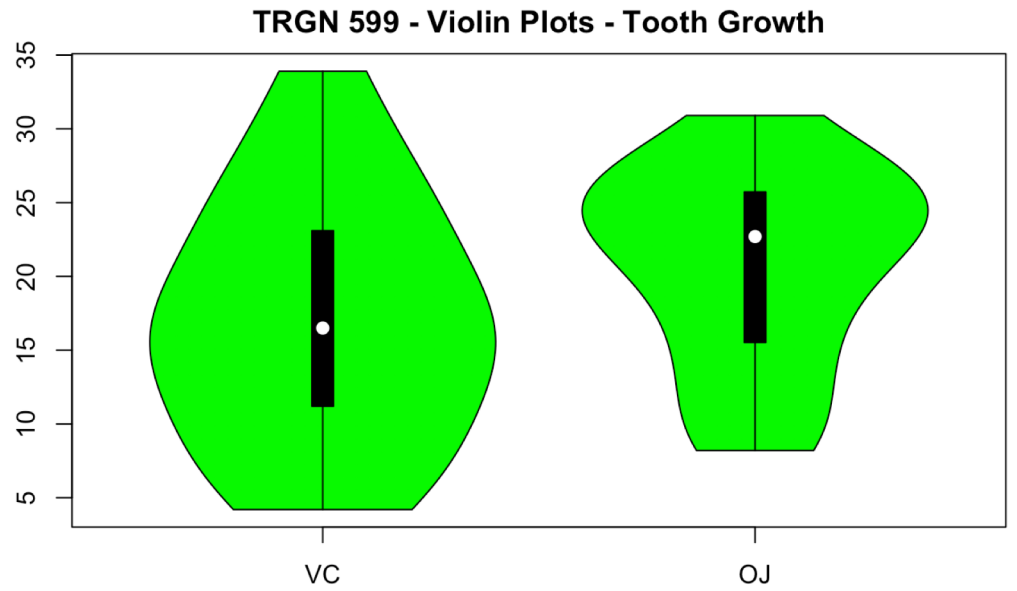
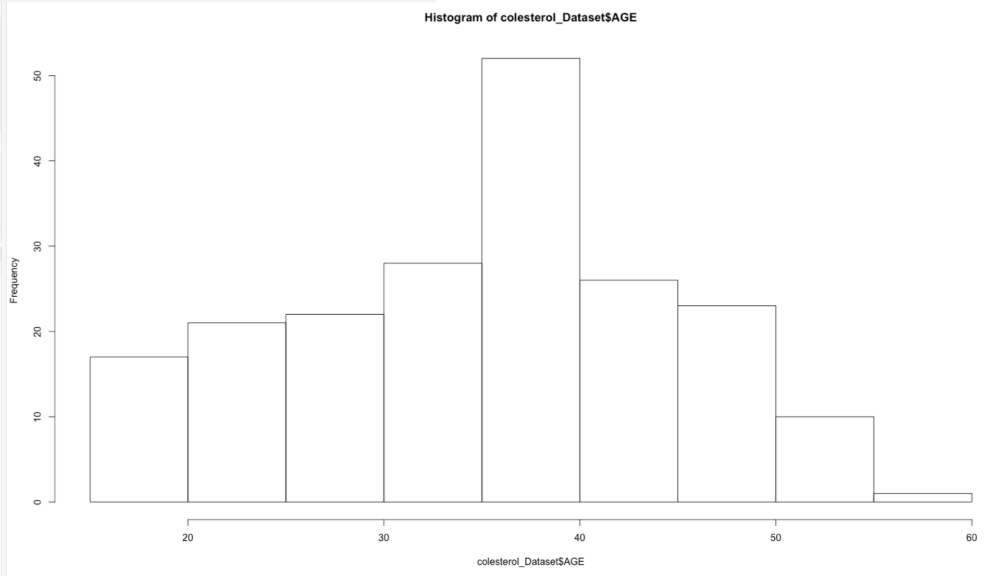
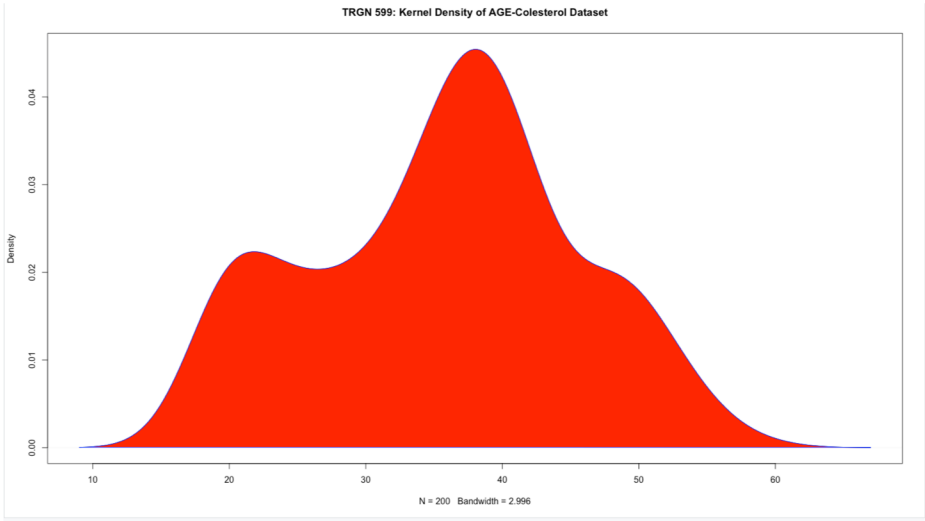
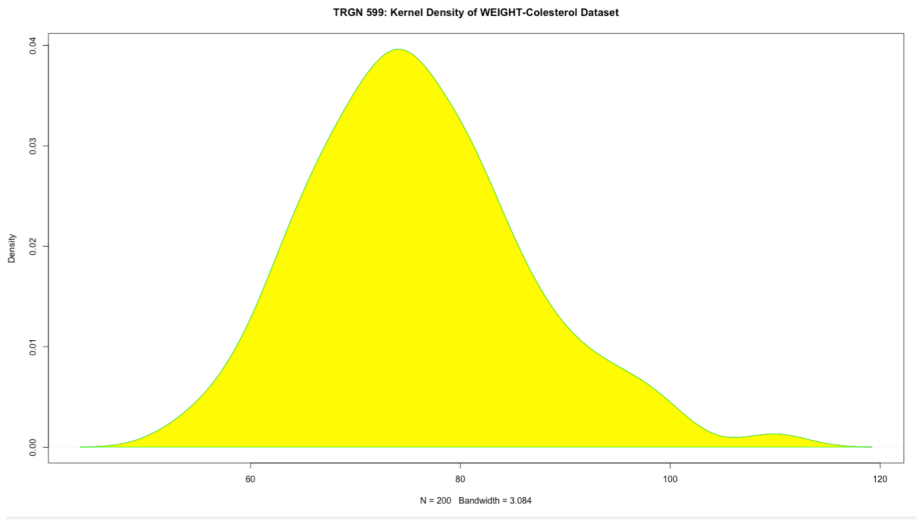
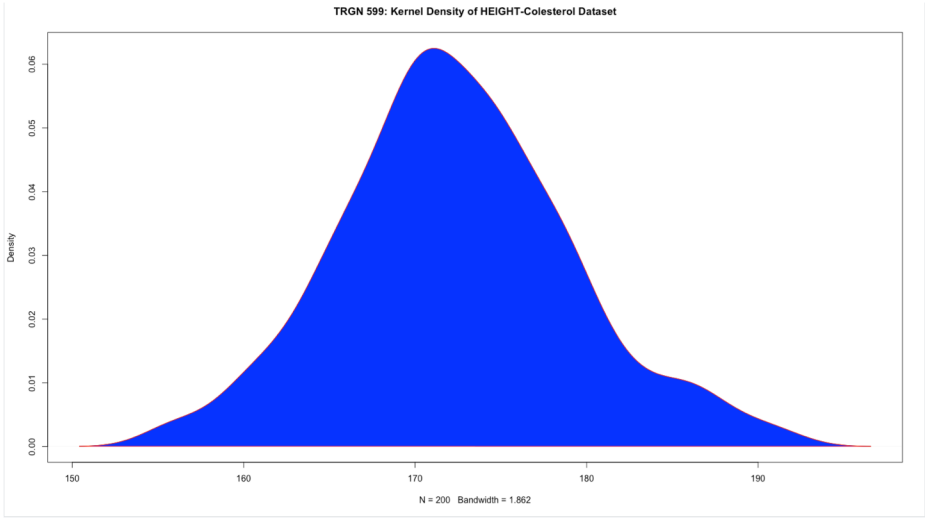
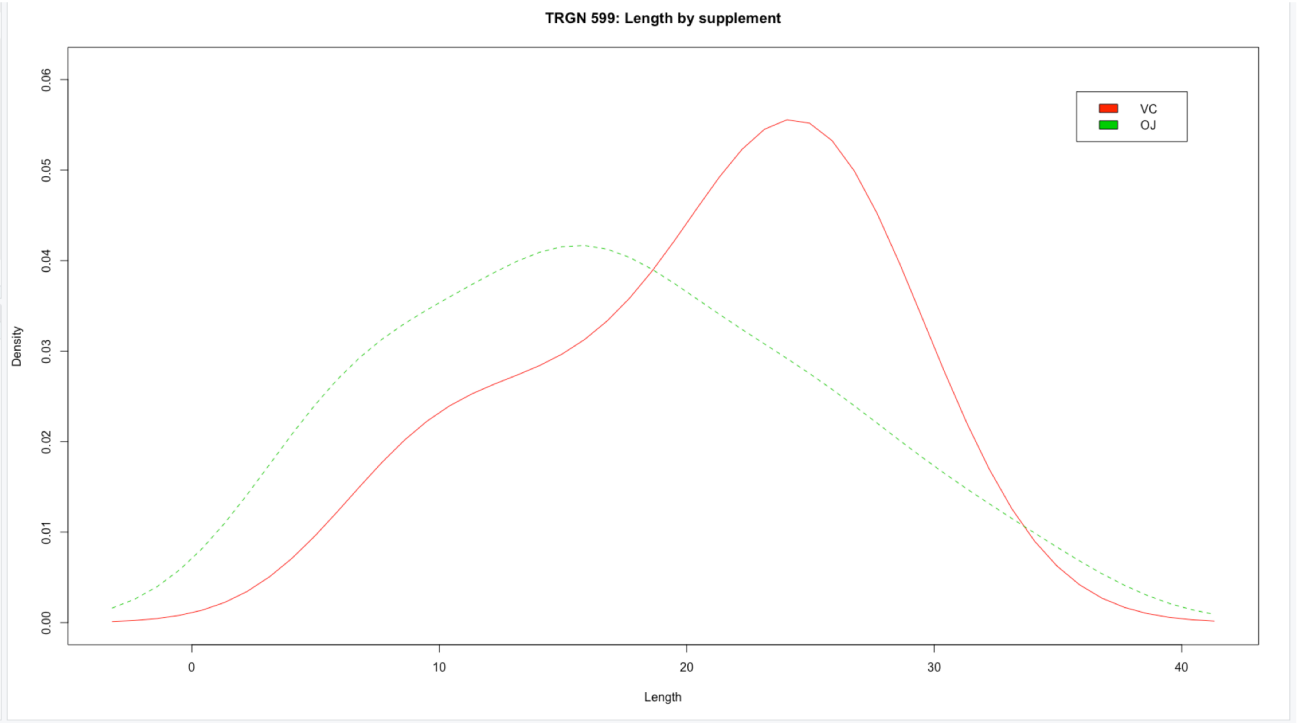
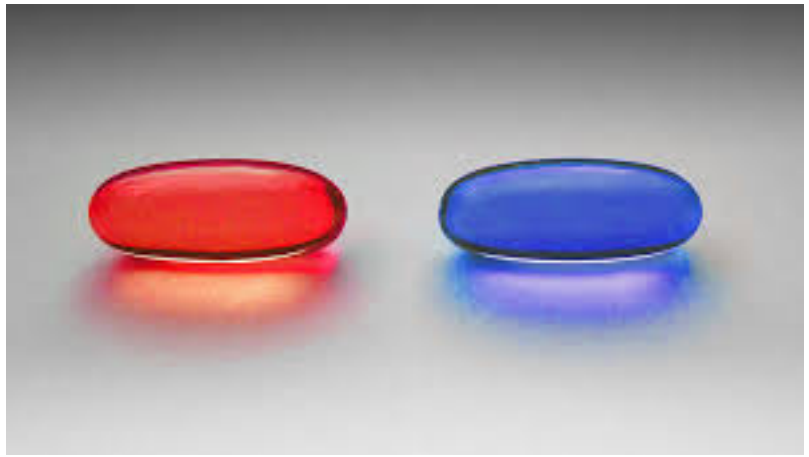
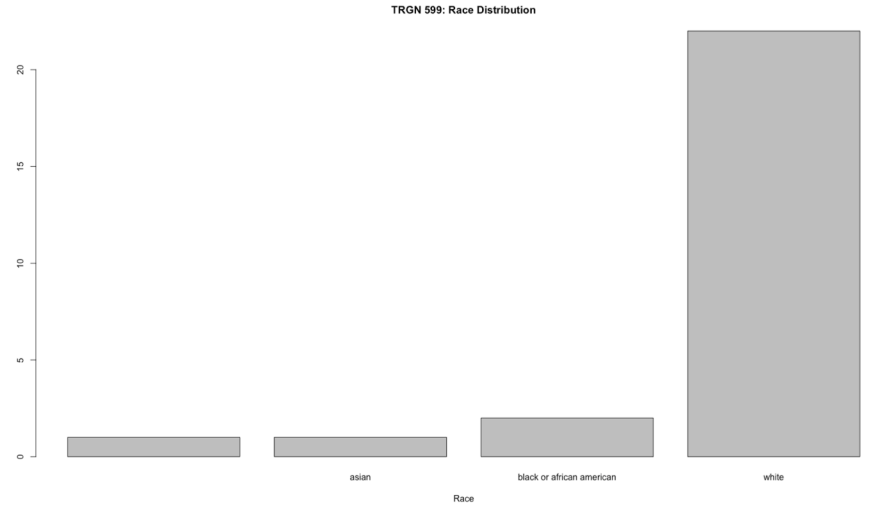
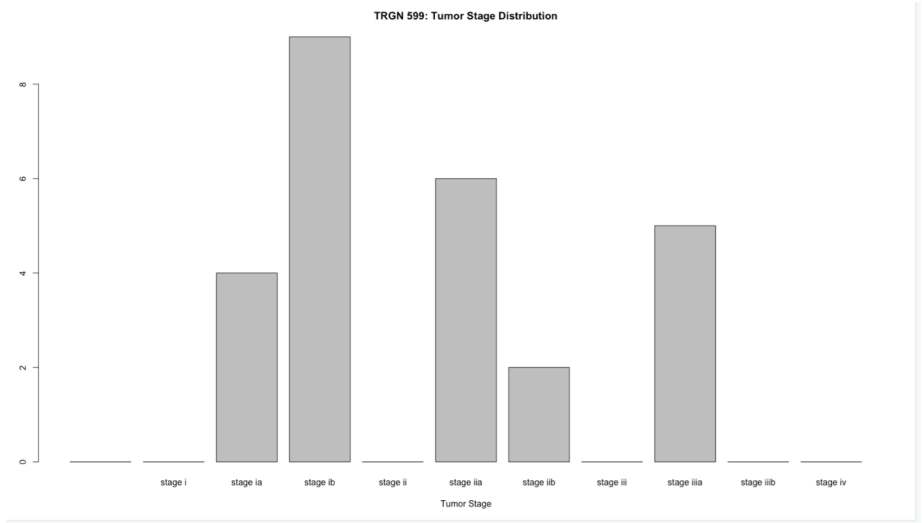
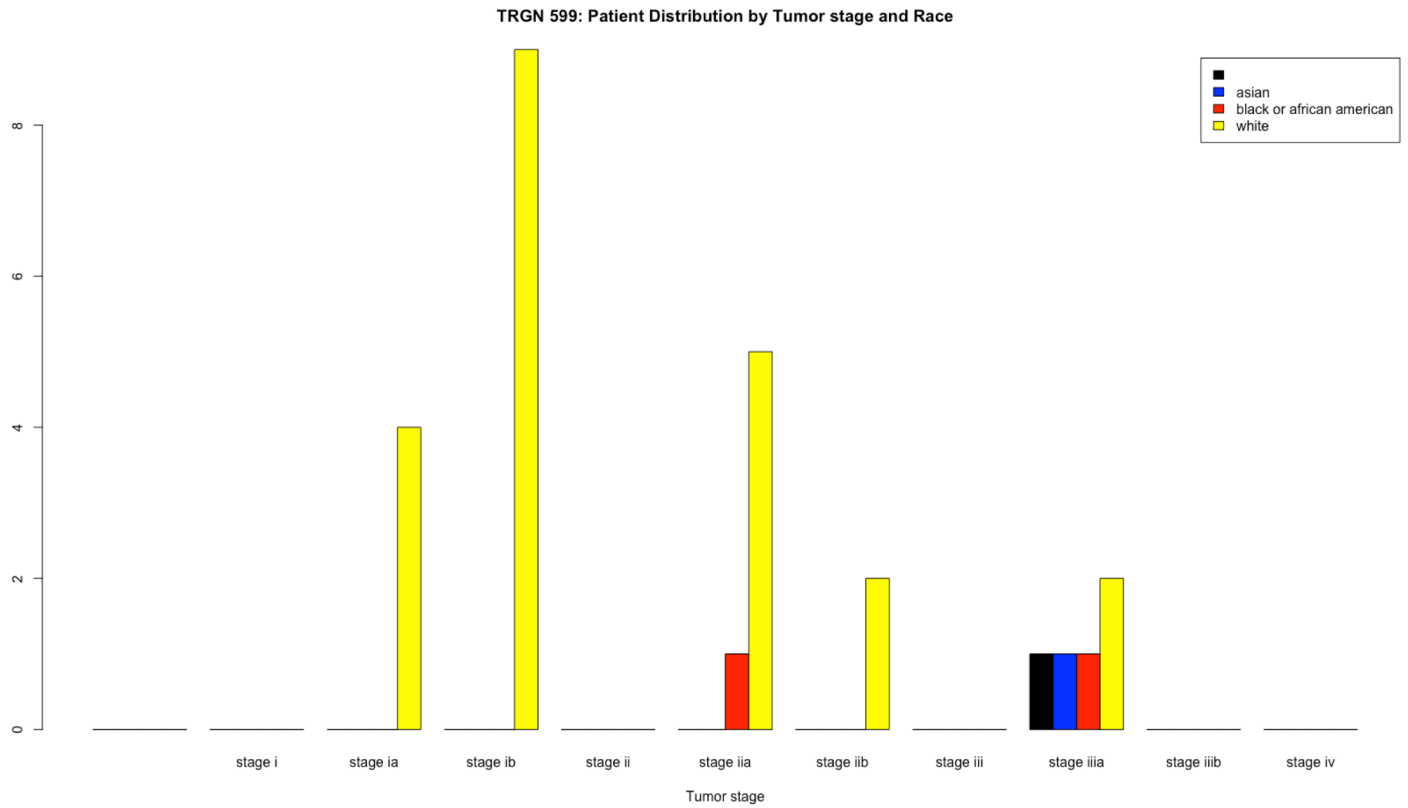
UID	Project ID	Case ID	Sample ID	Sample Type	cigarettes_per_day	weight	height	gender	year_of_birth
142bf27f-0ac8-48d3-a49d-a9ee1098bd70	TCGA-LUSC	TCGA-39-5040	TCGA-39-5040-01A	Primary Tumor	3.287671233	3.287671233	41	male	1949
d54bd3b4-1e83-4869-8b9c-76f367696f21	TCGA-LUSC	TCGA-77-8009	TCGA-77-8009-01A	Primary Tumor	5.342465753	5.342465753	39	male	1939
f66788fa-58c5-461e-aad7-2fb4023f6660	TCGA-LUSC	TCGA-77-7463	TCGA-77-7463-01A	Primary Tumor	2.301369863	2.301369863	42	male	1926
eff6361c-64b0-4aa8-ad10-184cea7edddd	TCGA-LUSC	TCGA-56-7579	TCGA-56-7579-11A	Solid Tissue Normal	0.109589041	0.109589041	17	male	1950
4bc9feca-bb58-4812-bddb-fd7f008a8d7a	TCGA-LUSC	TCGA-22-4609	TCGA-22-4609-11A	Solid Tissue Normal	4.219178082	4.219178082		male	1922
ea856e23-c5d3-4018-b98e-e2f4c63ae20c	TCGA-LUSC	TCGA-77-7138	TCGA-77-7138-01A	Primary Tumor	1.095890411	1.095890411	20	male	1932
621e56b2-5669-4dbd-b6df-b0d40c994a7d	TCGA-LUSC	TCGA-18-3414	TCGA-18-3414-01A	Primary Tumor	1.369863014	1.369863014		male	1932
b701206a-db47-4c5f-9825-8eebc831d1bc	TCGA-LUSC	TCGA-NC-A5HF	TCGA-NC-A5HF-01A	Primary Tumor				male	1934
752bf6b9-a679-4946-b390-b8760e121cb0	TCGA-LUSC	TCGA-33-4589	TCGA-33-4589-01A	Primary Tumor	2.465753425	2.465753425		female	1930
329c3d83-f0d4-4861-8297-a405777ad46a	TCGA-LUSC	TCGA-O2-A52Q	TCGA-O2-A52Q-01A	Primary Tumor	0.054794521	0.054794521		female	1961
fde5a0b0-8d5d-4468-bf31-89429c9d7837	TCGA-LUSC	TCGA-22-4591	TCGA-22-4591-01A	Primary Tumor				male	1921
d428b125-149d-44fc-b3ac-513a3e0a2a66	TCGA-LUSC	TCGA-66-2773	TCGA-66-2773-01A	Primary Tumor	3.178082192	3.178082192	23	male	1938
43ed092c-c187-410f-a404-12bea37963a0	TCGA-LUSC	TCGA-60-2707	TCGA-60-2707-01A	Primary Tumor	2.684931507	2.684931507	33	male	1933
6af47543-4d7a-4e1b-9803-eae5195f74d1	TCGA-LUSC	TCGA-22-5477	TCGA-22-5477-01A	Primary Tumor	5.479452055	5.479452055		male	1939
382cbf8f-9811-49f4-acc1-a7b79d302be3	TCGA-LUSC	TCGA-85-8664	TCGA-85-8664-01A	Primary Tumor	3.287671233	3.287671233		male	1938
e799ab87-6ec2-410c-9921-a97d8a87ab4a	TCGA-LUSC	TCGA-77-8008	TCGA-77-8008-11A	Solid Tissue Normal	3.369863014	3.369863014	41	male	1933
8b06e5a0-0668-41b1-b59f-9a70096c3b26	TCGA-LUSC	TCGA-6A-AB49	TCGA-6A-AB49-01A	Primary Tumor	2.739726027	2.739726027		female	1936
5d581b48-477b-488c-800c-3e7c2a2c7998	TCGA-LUSC	TCGA-34-7107	TCGA-34-7107-01A	Primary Tumor				male	1941
15ca6bac-19f2-4a56-a803-0b1d05b8ae47	TCGA-LUSC	TCGA-33-4586	TCGA-33-4586-01A	Primary Tumor	5.917808219	5.917808219	43	male	1950

Key Terms for Rectangular Data

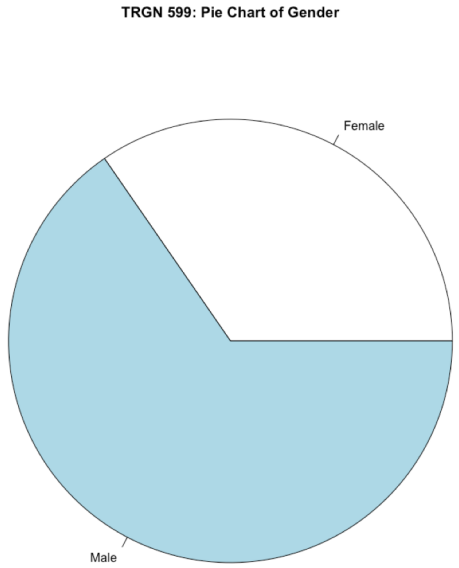
- Data frame
  - Rectangular data (like a spreadsheet) is the basic data structure for statistical and machine learning models.

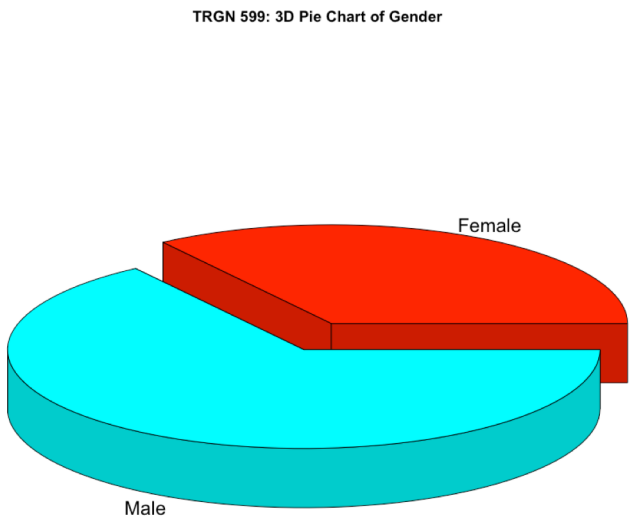
```
## Uploading the file "trgn599.clinical.tsv":
```{R}
clinical_data <- read.table("trgn599.clinical.tsv", header = TRUE, sep = "\t")
```


#Create a Data frame by typing the data
```{R}
a = c(3, 7, 9)
b = c("CG", "AT", "GC")
c = c(TRUE, FALSE, TRUE)
df = data.frame(a, b, c)       # df is a data frame
df
```


#Create a Data frame by selecting data from a table
```{R}
df_clinical_data = data.frame(clinical_data$UID, clinical_data$weight, clinical_data$gender)
df_clinical_data[1:5,1:3]
```
```

```
> df
  a  b     c
1 3 CG  TRUE
2 7 AT FALSE
3 9 GC  TRUE
> df_clinical_data[1:5,1:3]
                     clinical_data.UID clinical_data.weight clinical_data.gender
1 142bf27f-0ac8-48d3-a49d-a9ee1098bd70             3.287671                 male
2 d54bd3b4-1e83-4869-8b9c-76f367696f21             5.342466                 male
3 f66788fa-58c5-461e-aad7-2fb4023f6660             2.301370                 male
4 eff6361c-64b0-4aa8-ad10-184cea7edddd             0.109589                 male
5 4bc9feca-bb58-4812-bddb-fd7f008a8d7a             4.219178                 male
```

- Feature
  - A column in the table is commonly referred to as a feature.
    - Synonyms: attribute, input, predictor, variable

Key Terms for Rectangular Data

- Outcome
  - Many data science projects involve predicting an outcome-often a yes/no outcome. The features are sometimes used to predict the outcome in an experiment or study.
    - Synonyms: dependent variable, response, target, output

- Records
  - A row in the table is commonly referred to as a record
    - Synonyms: case, example, instance, observation, pattern, sample

UID	Project ID	Case ID	Sample ID	Sample Type	cigarettes_per_day	weight	height	gender	year_of_birth
142bf27f-0ac8-48d3-a49d-a9ee1098bd70	TCGA-LUSC	TCGA-39-5040	TCGA-39-5040-01A	Primary Tumor	3.287671233	3.287671233	41	male	1949
d54bd3b4-1e83-4869-8b9c-76f367696f21	TCGA-LUSC	TCGA-77-8009	TCGA-77-8009-01A	Primary Tumor	5.342465753	5.342465753	39	male	1939
f66788fa-58c5-461e-aad7-2fb4023f6660	TCGA-LUSC	TCGA-77-7463	TCGA-77-7463-01A	Primary Tumor	2.301369863	2.301369863	42	male	1926
eff6361c-64b0-4aa8-ad10-184cea7edddd	TCGA-LUSC	TCGA-56-7579	TCGA-56-7579-11A	Solid Tissue Normal	0.109589041	0.109589041	17	male	1950

Exploring the Data Distribution

- Each of the estimates we've covered sums up the data in a single number to describe the location or variability of the data.
- It is also useful to explore how the data is distributed overall.



R Datasets

ToothGrowth {datasets} R Documentation

The Effect of Vitamin C on Tooth Growth in Guinea Pigs

Description

The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as vc).

Usage

```
ToothGrowth
```

Format

A data frame with 60 observations on 3 variables.

[,1] len numeric Tooth length
[,2] supp factor Supplement type (VC or OJ).
[,3] dose numeric Dose in milligrams/day

Source

C. I. Bliss (1952). *The Statistics of Bioassay*. Academic Press.

References

McNeil, D. R. (1977). *Interactive Data Analysis*. New York: Wiley.

Crampton, E. W. (1947). The growth of the odontoblast of the incisor teeth as a criterion of vitamin C intake of the guinea pig. *The Journal of Nutrition*, **33**(5), 491–504. doi: 10.1093/jn/33.5.491.

Examples

```
require(graphics)
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```

The Growth of the Odontoblasts of the Incisor Tooth as a Criterion of the Vitamin C Intake of the Guinea Pig: Five Figures

E. W. Crampton

The Journal of Nutrition, Volume 33, Issue 5, 1 May 1947, Pages 491–504,
https://doi.org/10.1093/jn/33.5.491
Published: 01 May 1947 **Article history ▾**

Article PDF first page preview



THE GROWTH OF THE ODONTOBLASTS OF THE INCISOR TOOTH AS A CRITERION OF THE VITAMIN C INTAKE OF THE GUINEA PIG [1]

E. W. CRAMPTON
Department of Nutrition, Macdonald College, McGill University, P.O., Prov. Quebec, Canada

FIVE FIGURES

(Received for publication November 22, 1946)

After describing briefly 2 physical methods, 21 chemical methods, and 1 biochemical method, Rosenberg ('45) stated "Although the chemical, and to a small extent also, physical methods are replacing more and more the biological determinations of vitamin C, the biological tests maintain their place as the ultimate and most correct method of determining vitamin C." The problem of the assay of this vitamin was of particular concern to the Canadian Government during the war years because of the difficulty of providing natural sources of vitamin C to the armed forces for a considerable portion of the year. Inasmuch as different chemical procedures frequently gave different results as to the potency of a food in which the armed forces were interested, this laboratory was requested in 1942 to undertake the establishment of a vitamin C bioassay which might be used as a check against chemical procedures.

THE BASAL DIET

Most biological assays for vitamins depend ultimately on the normal development of the experimental animal, either as

[1] Contribution from the Faculty of Agriculture, McGill University, Macdonald College, Quebec, Canada. Journal Series no. 223.

491

THE GROWTH OF THE ODONTOBLASTS OF THE INCISOR TOOTH AS A CRITERION OF THE VITAMIN C INTAKE OF THE GUINEA PIG [1]

E. W. CRAMPTON

Department of Nutrition, Macdonald College, McGill University, P.O., Prov. Quebec, Canada

FIVE FIGURES

(Received for publication November 22, 1946)

After describing briefly 2 physical methods, 21 chemical methods, and 1 biochemical method, Rosenberg ('45) stated "Although the chemical, and to a small extent also, physical methods are replacing more and more the biological determinations of vitamin C, the biological tests maintain their place as the ultimate and most correct method of determining vitamin C." The problem of the assay of this vitamin was of particular concern to the Canadian Government during the war years because of the difficulty of providing natural sources of vitamin C to the armed forces for a considerable portion of the year. Inasmuch as different chemical procedures frequently gave different results as to the potency of a food in which the armed forces were interested, this laboratory was requested in 1942 to undertake the establishment of a vitamin C bioassay which might be used as a check against chemical procedures.

THE BASAL DIET

Most biological assays for vitamins depend ultimately on the normal development of the experimental animal, either as

[1] Contribution from the Faculty of Agriculture, McGill University, Macdonald College, Quebec, Canada. Journal Series no. 223.

Key Terms for Exploring the Distribution

- Boxplot
  - A plot introduced by Tukey as a quick way to visualize the distribution of data
    - They can be created using the boxplot() function
      - Synonyms: Box and whiskers plot

```
# Creating a Box Plot using public R datasets
```{R}
# Boxplot of tooth growth (dataset) against two factors
boxplot(len~supp*dose, data=ToothGrowth, notch=TRUE,
  col=(c("lightblue","blue")),
  main="TRGN 599 - Tooth Growth", xlab="Suppliment and Dose")

```
```



Key Terms for Exploring the Distribution

- Violin Plot
  - A violin plot is a combination of a boxplot and a kernel density plot.
  - Is an enhancement to the boxplot and plots the density estimate with the density on the y-axis.
  - The density is mirrored and flipped over and the resulting shape is filled in, creating an image resembling a violin.
  - The advantage of a violin plot is that it can show nuances in the distribution that aren't perceptible in a boxplot.
    - R vioplot package
      - vioplot( ) function

```
29 # Creating a Box Plot using public R datasets
30 ```{R}
31 # Boxplot of tooth growth (dataset) against two factors
32 boxplot(len~supp*dose, data=ToothGrowth, notch=TRUE,
33   col=(c("lightblue","blue")),
34   main="TRGN 599 - Tooth Growth", xlab="Suppliment and Dose")
35
36 ```
```



Key Terms for Exploring the Distribution

- Frequency table
  - A tally of the count of numeric data values that fall into a set of intervals (bins).

```
36  # Violin Plots
37  #install.packages("sm") - type "no" when asking for compilation
38  #install.packages("vioplot")
39  library("vioplot")
40  a1 <- ToothGrowth$len[ToothGrowth$supp=="VC"]
41  a2 <- ToothGrowth$len[ToothGrowth$supp=="OJ"]
42  vioplot(a1, a2, names=c("VC", "OJ"),
43     col="green")
44  title("TRGN 599 - Violin Plots - Tooth Growth")
45  ```
```



Key Terms for Exploring the Distribution

- Histogram
  - A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y-axis.

```
50 # Creating a Histogram using biology related public R datasets
51 ```{R}
52 colesterol_Dataset <- read.table("col.txt", header = TRUE)
53 colesterol_Dataset[1:2,1:5]
54 hist(colesterol_Dataset$AGE)
55 hist(colesterol_Dataset$HEIGHT)
56 hist(colesterol_Dataset$WEIGHT)
```



Key Terms for Exploring the Distribution

- Density plot
  - A smoothed version of the histogram often based on a kernel density estimate.

```
# Creating a Density Plot using biology related public R datasets
```{R}
d1 <- density(colesterol_Dataset$AGE)
plot(d1, main="TRGN 599: Kernel Density of AGE-Colesterol Dataset")
polygon(d1, col="red", border="blue")

d1 <- density(colesterol_Dataset$HEIGHT)
plot(d1, main="TRGN 599: Kernel Density of HEIGHT-Colesterol Dataset")
polygon(d1, col="blue", border="red")

d1 <- density(colesterol_Dataset$WEIGHT)
plot(d1, main="TRGN 599: Kernel Density of WEIGHT-Colesterol Dataset")
polygon(d1, col="yellow", border="green")
```
```







Key Terms for Exploring the Distribution

- Density plot
  - A smoothed version of the histogram often based on a kernel density estimate.

```
# Comparing Groups VIA Kernal Density
```{R}
# Compare length distributions for GP with VC, OJ
library(sm)
attach(ToothGrowth)

# create value labels
supp.f <- factor(supp, levels= c("VC", "OJ"),
  labels = c("VC", "OJ"))

# plot densities
sm.density.compare(len, supp, xlab="Length")
title(main="TRGN 599: Length by supplement")

# add legend via mouse click
colfill<-c(2:(2+length(levels(supp.f))))
legend(locator(1), levels(supp.f), fill=colfill)
```
```



Exploring Binary and Categorical Data

- For categorical data, simple proportions or percentages tell the story of the data.



Figure 2: Supervised principal component analysis plots for the two tissues and technologies showing the separation of the phenotypes using the maximum set classifiers. AR, acute rejection; PC, principal component; RNA-seq, RNA sequencing; subAR, subclinical acute rejection; TX, transplant with stable function.

Creating a new dataset without missing data (NA)

- The function **na.omit()** returns the object with listwise deletion of missing values.

```
> new_clinical_data <- na.omit(clinical_data)
> new_clinical_data$weight
 [1] 1.6438356 2.1917808 3.2876712 2.1917808 3.2876712 2.7397260 1.3698630 1.0958904 2.7397260 0.5479452 0.4383562 5.4794521 0.4931507 2.1917808 4.6575342 4.6027397 2.7397260 1.6438356 3.5068493 6.2465753 2.6301370 1.6438356 0.8219178 1.6438356 2.3013699 0.6575342
> clinical_data$weight
  [1]  3.28767123  5.34246575  2.30136986  0.10958904  4.21917808  1.09589041  1.36986301          NA  2.46575342  0.05479452          NA  3.17808219  2.68493151  5.47945206  3.28767123  3.36986301  2.73972603          NA  5.91780822  2.90410959          NA  3.28767123  1.36986301
 [24]  3.28767123  2.73972603  4.27397260  1.64383562  2.19178082  2.19178082  2.08219178  2.73972603          NA          NA          NA          NA  2.41095890  1.09589041  1.09589041          NA  0.82191781          NA  2.73972603  4.71232877  4.93150685  4.38356164          NA
 [47]          NA  4.05479452  3.28767123  2.46575342  4.38356164  2.46575342  2.19178082  1.36986301  2.46575342          NA  3.28767123          NA          NA  2.52054795  1.86301370  2.02739726  2.73972603  1.36986301  2.19178082  3.28767123          NA          NA  2.73972603
 [70]  1.64383562          NA  3.28767123  1.91780822          NA  1.58904110  2.46575342  6.02739726  2.19178082          NA  1.64383562          NA  2.73972603  3.28767123  2.63013699          NA          NA  2.19178082          NA          NA  1.36986301          NA  3.01369863
 [93]  1.64383562          NA          NA          NA  6.57534247  1.04109589  0.27397260  5.17808219  0.82191781  2.19178082  6.02739726  1.36986301  2.19178082  6.84931507  6.57534247          NA  2.57534247  1.64383562          NA          NA          NA  1.64383562          NA
[116]  0.54794520  0.27397260  1.91780822          NA  2.46575342          NA  2.24657534          NA  1.04109589  8.43835616  2.73972603  3.50684932  4.05479452  0.54794520  2.19178082  3.28767123 10.95890411  3.83561644  4.93150685  2.73972603  1.64383562  1.36986301          NA
[139]  2.19178082  2.73972603  4.38356164          NA  1.91780822  0.10958904  1.09589041  1.91780822          NA          NA  4.38356164  0.65753425  3.36986301  1.26027397  3.06849315  2.26027397  1.64383562  2.73972603  4.10958904          NA  2.19178082  1.91780822          NA
[162]  2.79452055  0.54794520          NA  4.21917808  3.56164384  0.22602740  0.43835616  4.60273973  1.09589041  2.73972603  1.64383562  1.91780822  3.28767123  3.56164384  3.28767123  2.84931507  5.47945206  4.10958904  4.38356164          NA  4.60273973  1.31506849  5.47945206
[185]  0.95890411          NA  3.50684932  2.08219178  1.36986301          NA          NA  2.73972603  1.91780822          NA          NA          NA  1.09589041  4.93150685          NA          NA          NA  1.91780822          NA  4.27397260  0.60273973          NA          NA
[208]  3.28767123  2.19178082  0.76712329          NA  0.16438356  2.19178082  6.41095890  3.28767123          NA  2.52054795  4.38356164  5.47945206  4.35616438  0.49315068          NA  0.65753425  2.73972603  2.73972603  1.36986301  2.73972603  5.47945206  2.73972603  4.35616438
[231]  2.19178082  0.13698630  1.64383562  3.50684932  2.19178082  2.73972603  1.53424658  2.46575342  2.73972603          NA  6.13698630  2.73972603  3.56164384  2.46575342          NA  2.63013699  5.31506849  0.54794520  0.60273973          NA          NA  4.65753425          NA
[254]  3.28767123  1.36986301  5.47945206          NA          NA  0.13698630  4.60273973  1.36986301  4.27397260  1.36986301  2.73972603  2.19178082  2.73972603  1.91780822  2.63013699  5.47945206  2.19178082  1.97260274  2.46575342  2.19178082  3.28767123  1.64383562  1.04109589
[277]          NA  2.40000000  0.54794520          NA  1.36986301  1.53424658          NA          NA  5.69863014          NA  2.19178082  1.09589041  3.83561644  3.94520548  1.64383562  1.20547945  2.46575342  1.36986301  2.19178082  3.12328767  1.64383562  5.47945206  3.28767123
[300]  2.73972603          NA  2.73972603  0.16438356          NA  2.95890411  2.73972603  3.56164384          NA  1.09589041  3.28767123  2.73972603          NA          NA  1.64383562  2.19178082  6.16438356  2.73972603  1.36986301  1.64383562  2.63013699          NA          NA
[323]  2.73972603          NA          NA          NA          NA  2.19178082  1.86301370          NA          NA          NA          NA  0.82191781  3.04109589  0.82191781  0.10958904  2.19178082  1.91780822          NA  2.46575342  3.50684932  1.64383562 13.15068493  2.19178082
[346]  7.72602740  2.02739726  2.19178082  0.82191781  6.19178082  0.54794520          NA  3.78082192  3.12328767  6.24657534  2.19178082  1.64383562          NA  3.28767123  2.73972603  3.28767123          NA  0.82191781          NA  2.19178082  2.19178082          NA  0.82191781
[369]  5.15068493  2.19178082  2.73972603  0.32876712          NA  0.82191781  0.54794520          NA          NA  0.27397260  2.63013699  3.28767123  3.12328767  3.01369863  2.19178082  1.40273973  1.64383562  4.71232877  4.10958904  1.09589041  0.05479452 10.52054795  1.80821918
[392]  0.82191781  1.36986301  2.79452055  2.73972603  3.06849315  2.02739726  2.19178082  1.09589041          NA  1.26027397  2.19178082          NA  2.84931507  3.78082192  1.23287671          NA  2.84931507          NA          NA  0.82191781  1.64383562  1.64383562  2.73972603
[415]  7.39726027          NA  1.36986301  3.28767123  4.54794521  2.19178082  3.28767123  4.05479452  2.26027397  1.09589041  6.35616438  1.09589041  2.30136986  3.28767123          NA  2.73972603          NA  0.65753425  1.09589041          NA          NA          NA  1.36986301
[438]  0.82191781  4.38356164  0.54794520          NA  0.65753425  2.73972603          NA  2.19178082          NA          NA  3.12328767  2.46575342          NA  5.17808219          NA  1.20547945          NA  2.35616438  2.73972603  2.19178082          NA  1.09589041          NA
[461]  3.28767123  1.52328767  3.28767123  3.50684932          NA
```

Key terms for Exploring Binary and Categorical Data

- Bar charts
  - The frequency or proportion for each category plotted as bars

```
99  # Bar plots
100 ```{R}
101 # Simple Bar Plot
102 counts <- table(new_clinical_data$race)
103 barplot(counts, main="TRGN 599: Race Distribution",
104    xlab="Race")
105
106 # Simple Bar Plot
107 counts <- table(new_clinical_data$tumor_stage)
108 barplot(counts, main="TRGN 599: Tumor Stage Distribution",
109    xlab="Tumor Stage")
110
```





Key terms for Exploring Binary and Categorical Data

- Grouped Bar Plot

```
# Creating grouped Bar Plot
```{R}
# Simple Bar Plot
counts <- table(new_clinical_data$race, new_clinical_data$tumor_stage)
barplot(counts, main="TRGN 599: Patient Distribution by Tumor stage and Race",
   xlab="Tumor stage", col=c("black", "blue", "red", "yellow"),
  legend = rownames(counts), beside=TRUE)
```
```



Key terms for Exploring Binary and Categorical Data

- Pie charts
  - The frequency or proportion for each category plotted as wedges in a pie

```
# Creating a Simple Pie Chart
```{R}
#
slices <- table(new_clinical_data$gender)
lbls <- c("Female", "Male")
pie(slices, labels = lbls, main="TRGN 599: Pie Chart of Gender")


# 3D Exploded Pie Chart
# install.packages("plotrix")
library(plotrix)
slices <- table(new_clinical_data$gender)
lbls <- c("Female", "Male")
pie3D(slices,labels=lbls,explode=0.1,
   main="TRGN 599: 3D Pie Chart of Gender")
```



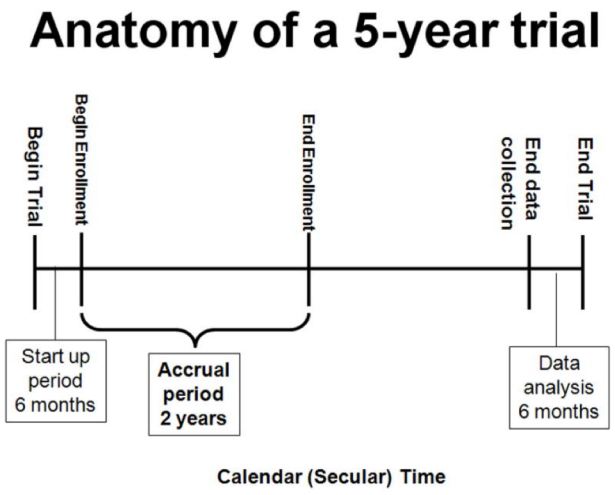

```

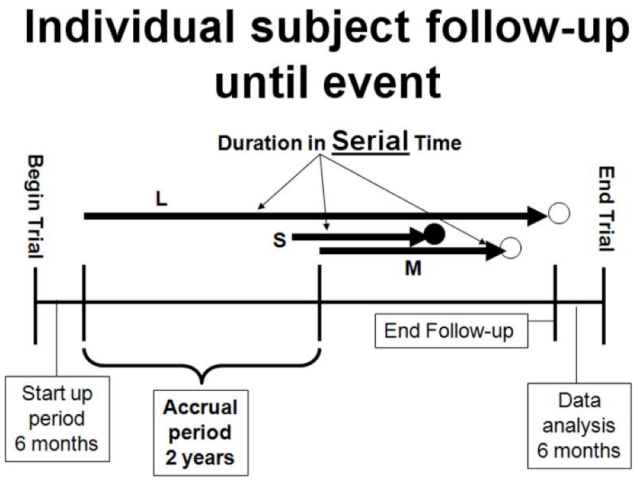
Key terms for Exploring Binary and Categorical Data

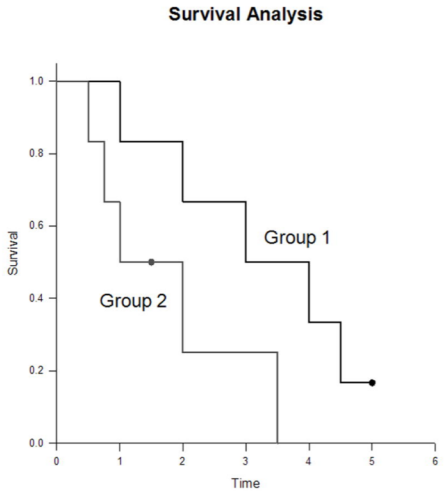
- Mode
  - The most commonly occurring category or value in a data set.
  - **3**, **3**, 2, 1, 4, 5, **3**, **3**, 4, 9,7
- Expected value
  - When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.
    - In R, the function "mean" return the expected value.

Kaplan Meier

- Kaplan Meier
  - Kaplan-Meier curves and estimates of survival data have become a familiar way of dealing with differing survival times (times-to-event), especially when not all the subjects continue in the study.
  - “Survival” times need not relate to actual survival with death being the event; the “event” may be any event of interest.
  - Kaplan-Meier analyses are also used in non-medical disciplines.







https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3932959/

Heat map

- Heat map
  - A heat map (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors.