CS CM124/224 & HG CM124/224 & BIOINFO 224

DISCUSSION SECTION (APRIL11, 2014)

TA: Farhad Hormozdiari

Reminder

- Course Homepage
  - http://genetics.cs.ucla.edu/cs124/index.html
- TA Office Hours : Monday 12-2 @ Math Science (MS) 2915
- Professor Office Hours: Wednesday 4-5 @ MS 2925
- Homework 0B: Due April 15, Hard copy should be submitted to TA's office
- Lectures and Discussion section videos will be posted during the weekend.

Agenda

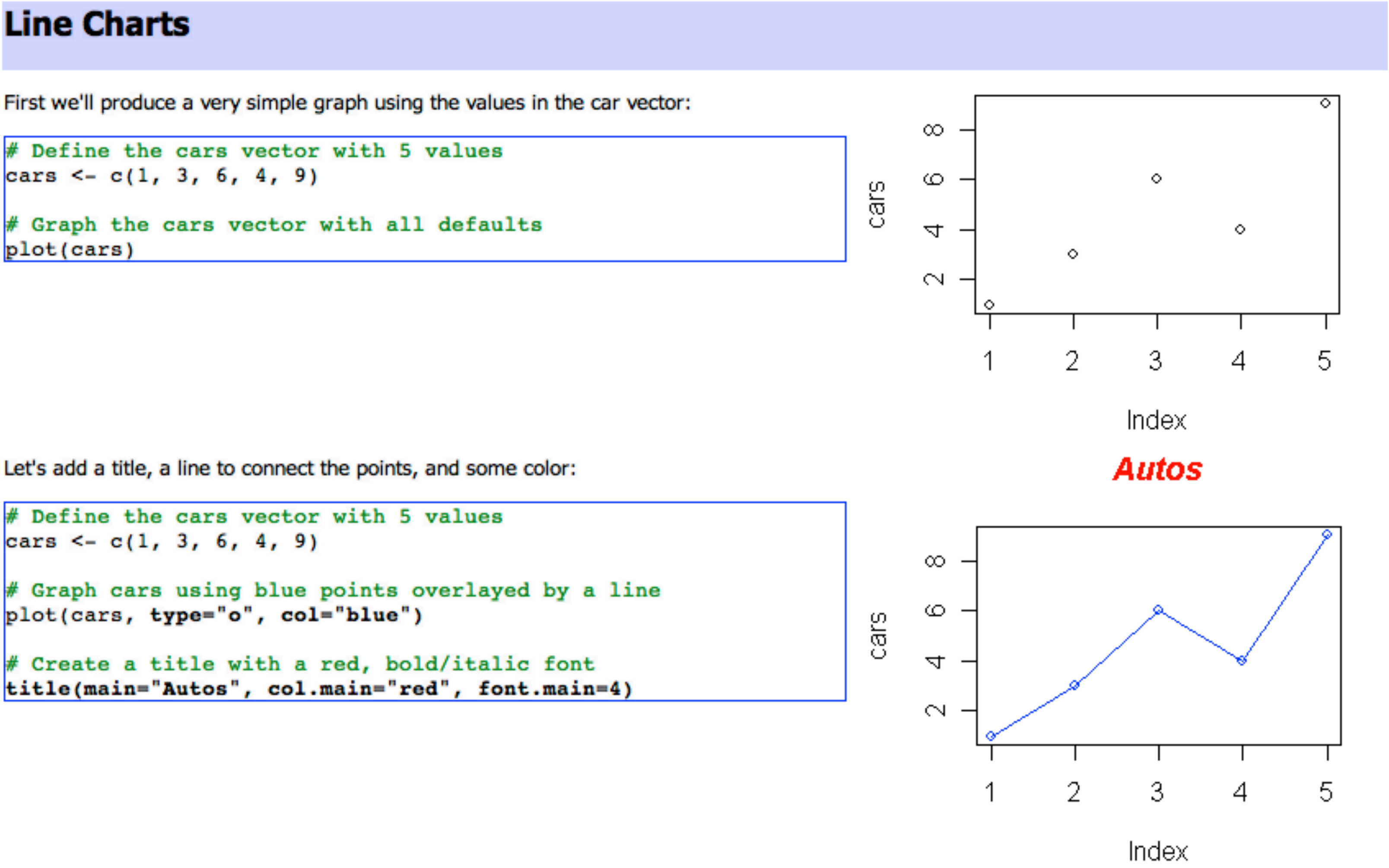
- Introduction to R
- Basic statistics
- Association Statistics
- HW0B

Introduction to R

- R is a software package for statistical and graphics
- Free to download at:
  - http://cran.r-project.org/mirrors.html
- You can use Rstudio which has nice interface
- We will learn R by going through examples posted on course webpage.
- Use ? In front of each command to get the man or help page
  - ?hist

More on R plots

- Examples from http://www.harding.edu/fmccown/r/



Line Charts

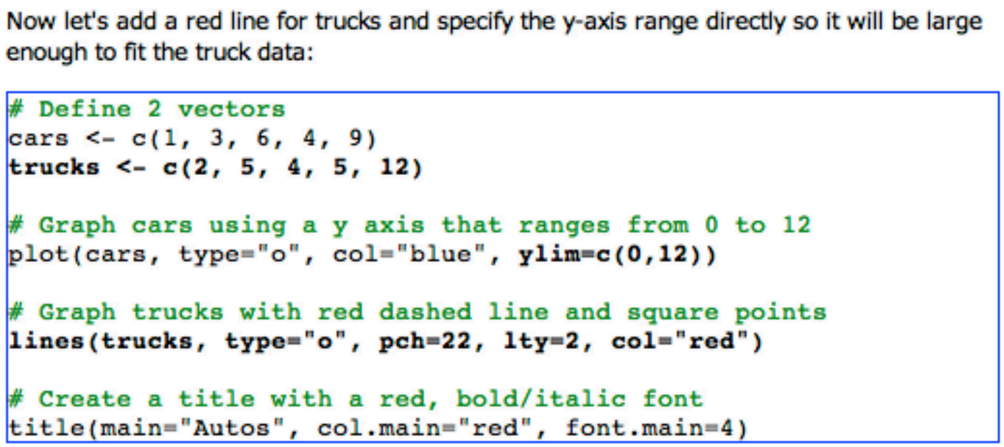
First we'll produce a very simple graph using the values in the car vector:

```
# Define the cars vector with 5 values
cars <- c(1, 3, 6, 4, 9)

# Graph the cars vector with all defaults
plot(cars)
```

Let's add a title, a line to connect the points, and some color:

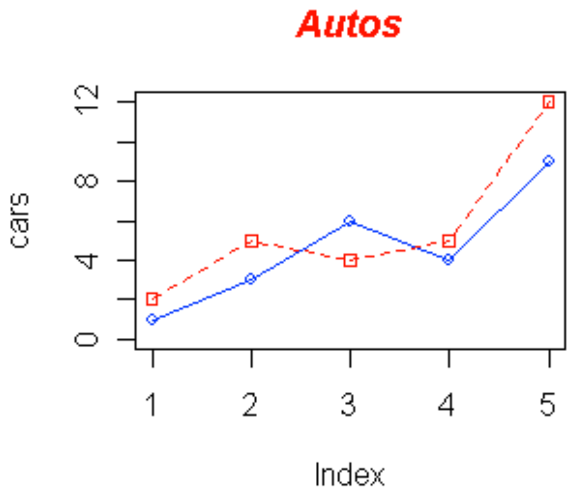
```
# Define the cars vector with 5 values
cars <- c(1, 3, 6, 4, 9)

# Graph cars using blue points overlayed by a line
plot(cars, type="o", col="blue")

# Create a title with a red, bold/italic font
title(main="Autos", col.main="red", font.main=4)
```

More on R plots



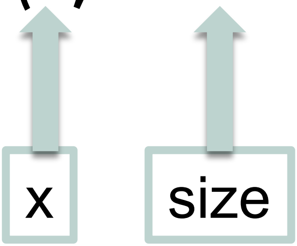
Now let's add a red line for trucks and specify the y-axis range directly so it will be large enough to fit the truck data:

```
# Define 2 vectors
cars <- c(1, 3, 6, 4, 9)
trucks <- c(2, 5, 4, 5, 12)

# Graph cars using a y axis that ranges from 0 to 12
plot(cars, type="o", col="blue", ylim=c(0,12))

# Graph trucks with red dashed line and square points
lines(trucks, type="o", pch=22, lty=2, col="red")

# Create a title with a red, bold/italic font
title(main="Autos", col.main="red", font.main=4)
```



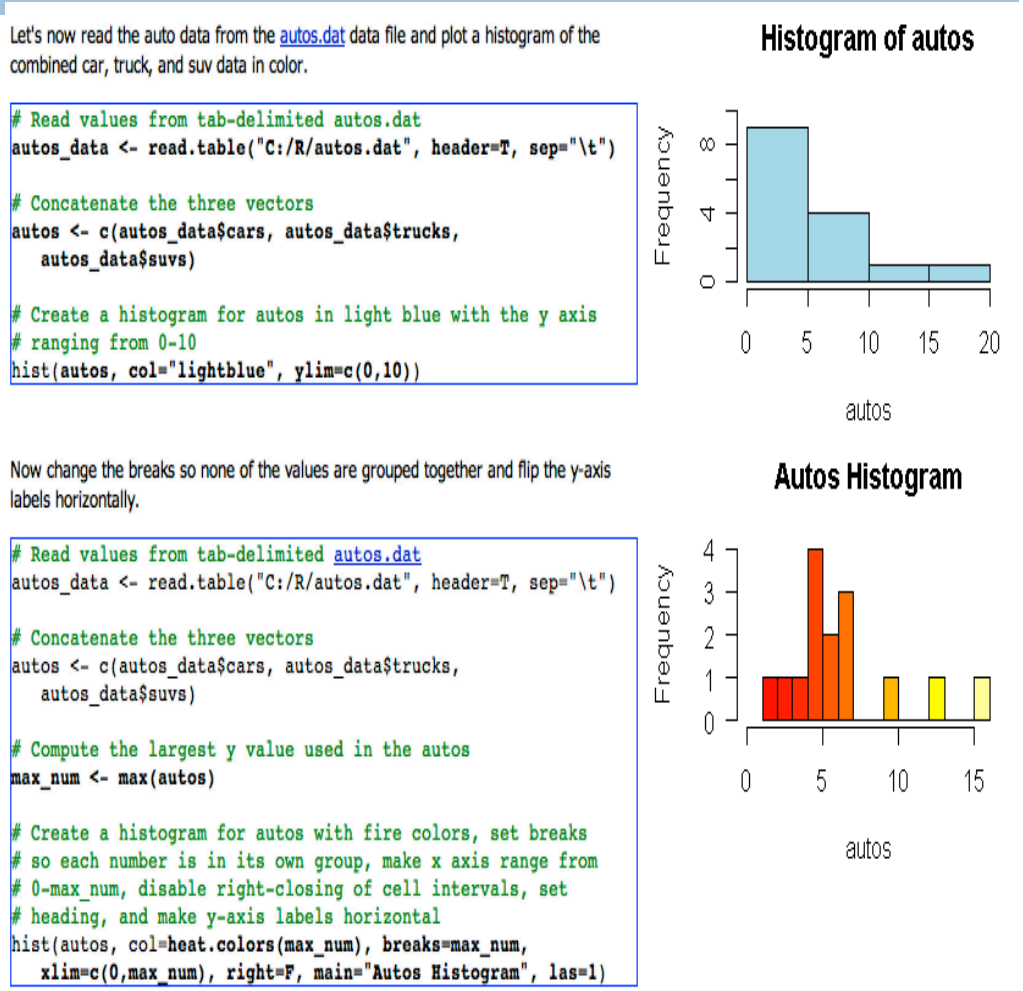
More on R plots (Histogram)

- □ "hist" command in R
- □ hist(x, breaks)



x

size



Let's now read the auto data from the autos.dat data file and plot a histogram of the combined car, truck, and suv data in color.

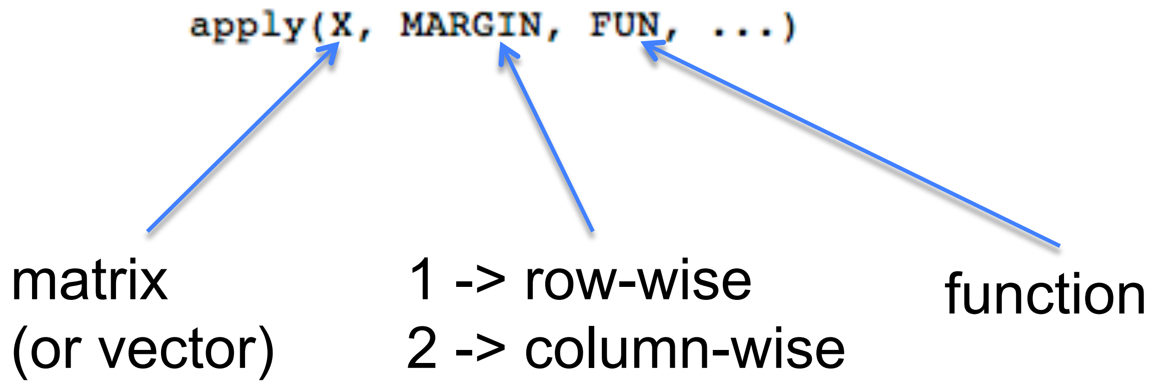
```
# Read values from tab-delimited autos.dat
autos_data <- read.table("C:/R/autos.dat", header=T, sep="\t")

# Concatenate the three vectors
autos <- c(autos_data$cars, autos_data$trucks,
   autos_data$suvs)

# Create a histogram for autos in light blue with the y axis
# ranging from 0-10
hist(autos, col="lightblue", ylim=c(0,10))
```

Now change the breaks so none of the values are grouped together and flip the y-axis labels horizontally.

```
# Read values from tab-delimited autos.dat
autos_data <- read.table("C:/R/autos.dat", header=T, sep="\t")

# Concatenate the three vectors
autos <- c(autos_data$cars, autos_data$trucks,
   autos_data$suvs)

# Compute the largest y value used in the autos
max_num <- max(autos)

# Create a histogram for autos with fire colors, set breaks
# so each number is in its own group, make x axis range from
# 0-max_num, disable right-closing of cell intervals, set
# heading, and make y-axis labels horizontal
hist(autos, col=heat.colors(max_num), breaks=max_num,
   xlim=c(0,max_num), right=F, main="Autos Histogram", las=1)
```

apply() in R

- Loops are very slow in R
  - For, while, …
- Use the built-in functions
  - colMeans, rowMeans, etc
  - colSums (x, na.rm = FALSE, dims = 1)
- If no built-in functions, use apply()



```
apply(X, MARGIN, FUN, ...)
```

matrix
(or vector)

1 -> row-wise
2 -> column-wise

function

apply() in R

```
computeMean = function (x) {
return(mean(x))
}
a = matrix(c(1,2,3,4,5,6), nrow=2, ncol=3, byrow=T)

apply(a,1,computeMean)

apply(a,2,computeMean)
```

Basic Statistics

- This course requires some basic statistics.
  - We will review most topics which are needed in this course, so don't worry.
- Mean (Expectation) and Variance of a distribution.
  - Mean
  - Variance

Continues variable	Discrete variable
$E(X) = \int_{-\infty}^{\infty} xP(x)dx$	$E(X) = \sum_{x=0} xp(x)$

$$Var(X) = E(\{X - E(x)\}^2) = E(X^2) - E(X)^2$$

Basic Statistics

- Expectation is a linear function:

$$E(aX \pm bY) = aE(X) \pm bE(Y)$$

  - a and b are constant

- Variance is a bit more tricky

$$Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y)$$

Important

Basic Statistics

- □ Covariance

$$Cov(X,Y) = E\{(X - E(X))(Y - E(Y))\}$$

- □ Correlation (**Pearson's coef.**)

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

- □ Correlation seem complicated to memorize.

Basic Statistics

- Correlation (**Pearson's coef.**)

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

- Correlation seem complicated to memorize.
  - We can visualize the X and Y as vector.
  - Correlation can be seen as the angle between the two vector of X and Y

Basic Statistics (Known Distribution)

- Normal distribution (Gaussian distribution)

$$N(\mu, \sigma^2) \sim \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Bernoulli distribution.
  - An event occur with probability of p
    - Mean and variance?

- Binomial distribution
  - An event occur with probability of p, we try n times and we observe k times the event is occurred.

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Basic Statistics (Known Distribution)

- Poisson Distribution : Model rare events in nature.

$$Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

  - λ is the only parameter we need.
  - Mean and variance?
- Chi-square Distribution:
  - Sum of "k" squared normal distribution $\chi_k^2 = \sum_{i=1}^{k} Z_i^2$
  - Zi are independent.
  - k is called the degree of freedom.

Basic Statistics

- We want to obtain some information about the data
- If the data support our idea (Hypothesis)
- Question: is this coin fair or biased?
  - We flip it many times and see the outcomes.

Basic Statistics

- We flip a coin 20 times and we get 14 heads. Is the coin fair or not?
  - NULL Hypothesis the coin is fair.
  - Alternative Hypothesis is biased.
- We consider the Null Hypothesis
  - Coin pair P(Head) = P(Tail) = 0.5
  - Binomial distribution.

Basic Statistics

- We compute the p-value
  - P-value the probability of observing the same event or more significant under the Null Hypothesis.
  - Let X be the number of Heads
  - In the coin example P-value = Pr(X=14) + …P(X=20)
  - This is a Binomial trail,

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

  - P-value = P(X>=14) = 1-P(X<14)
  - If P-value is smaller than 0.05 it is significant

Basic Statistics

- P-value = P(X>=14) = 1-P(X<14)
- If P-value is smaller than 0.05 it is significant

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

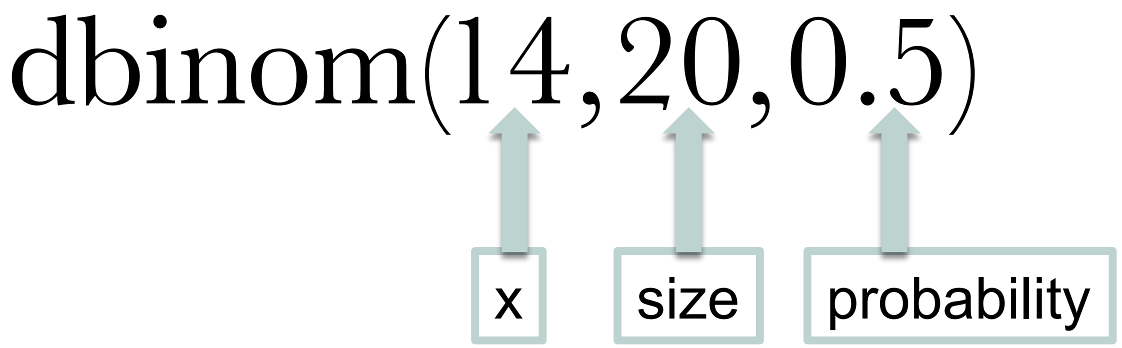
- The probability of 14 heads is 0.058
- Thus, for the significant level of 0.05, 0.058 is not consider significant

Basic Statistics

- How about we observe only 1 head?
  - P-value = $P(X=0) = (½)^{20}$
  - It is significant, so we can reject the Null (coin is fair).
  - We may believe the coin is biased.

How to use R for statistical computing

- Probability mass (or density) function (PDF):
  - Compute Pr(X = x)
  - Example: what is the probability of observing 14 heads when we toss the coin 20 times and when the probability of having heads is 0.5 (Pr(X = 14))?
  - In R, d[distribution_name] corresponds to PDF



```
dbinom(14,20,0.5)
```

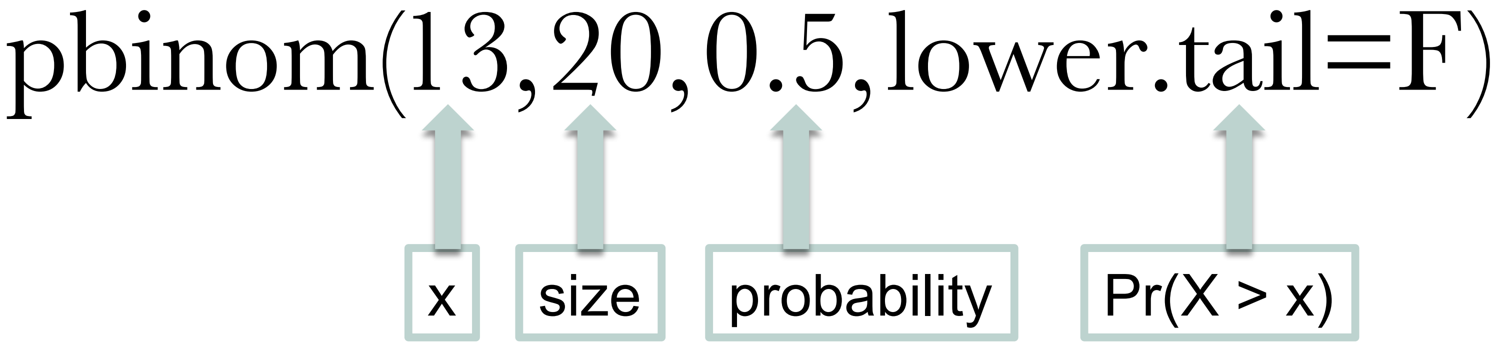
x

size

probability

How to use R for statistical computing

- Cumulative distribution function (CDF)
  - Compute the P(X<=x)
    - Easy to compute P(X>x) = 1-P(X<=x)
  - In R, p[distribution_name] correspond to CDF
    - pbinom(13,20,0.5, lower.tail = F) = P(X>x)
    - pbinom(13,20,0.5, lower.tail = T) = P(X<=x)



pbinom(13,20,0.5,lower.tail=F)

x

size

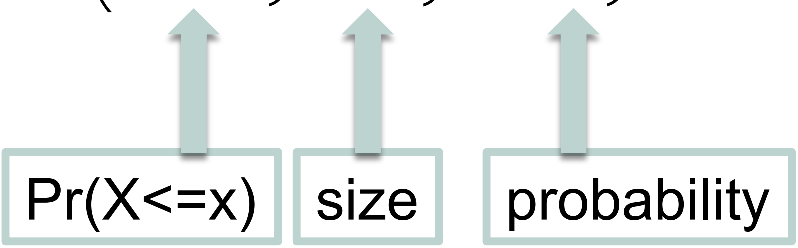
probability

Pr(X > x)

How to use R for statistical computing

- Quantile function:
  - Inverse of CDF == What is the value of x when we know P(X<=x)=0.2
  - Example: what is the number of heads (x) when P(X<=x) = 0.2
  - In R, q[distribution_name] correspond to quantile function
  - Often used to find the significant threshold

```
qbinom(0.2,20,0.5,lower.tail=T)
```



Pr(X<=x)

size

probability

How to use R for statistical computing

- For normal distribution
  - pnorm(x, mean, sd) = compute p-value
  - qnorm(x, mean, sd) = compute significance threshold
  - rnorm(n, mean, sd) = generate n random numbers from normal distribution
- Other distributions
  - [d/p/q/r]chisq = Chi-square distribution
  - [d/p/q/r]pois = Poisson distribution

Association Statistics

- Let’s go over (basic) Biology first
- You can think of human DNA as a series of 3 billion letters where each letter is either A, C, T, or G.
- Humans differ by 0.1% of their DNA.

Most of DNA look like…

```
Position X in DNA
Ind 1:  A
Ind 2:  A
Ind 3:  A
Ind 4:  A
Ind 5:  A
Ind 6:  A
Ind 7:  A
Ind 8:  A
Ind 9:  A
Ind 10: A
Ind 11: A
Ind 12: A
Ind 13: A
Ind 14: A
Ind 15: A
Ind 16: A
```

- Humans have 2 chromosomes, but for simplicity, let’s assume they have only one chromosome for now

0.1% of DNA look like…

```
Position Y in DNA
Ind 1:   A
Ind 2:   A
Ind 3:   A
Ind 4:   T
Ind 5:   A
Ind 6:   A
Ind 7:   A
Ind 8:   A
Ind 9:   T
Ind 10:  A
Ind 11:  T
Ind 12:  A
Ind 13:  A
Ind 14:  T
Ind 15:  A
Ind 16:  T
```

- This is called SNP (single nucleotide polymorphism)
- A and T (in this example) are called “alleles”
- A is one allele, T is the other allele
- A is a major allele (because it appears more often, 11/16)
- T is the minor allele (5/16)
- Only Two alleles per SNP (no three or four alleles SNP): there is no person who has C or G allele for this SNP

What is an “association”?

cases: people with a disease

```
Position X in DNA
Case 1:   A
Case 2:   A
Case 3:   A
Case 4:   A
Case 5:   A
….
Case 10^6: A
```

controls: people w/o a disease

```
Position X in DNA
Control 1:   T
Control 2:   T
Control 3:   T
Control 4:   T
Control 5:   T
….
Control 10^6: T
```

- Similar to a dice having 99 million heads on 100 million tosses
- This SNP is “associated” with a disease, simple correlation.
- If you have “A” allele, you are almost likely to have a disease

But usually we have this situation

cases: people with a disease

400 case individuals have A
600 case individuals have T

controls: people w/o a disease

300 control individuals have A
700 control individuals have T

- Is this SNP “associated” with a disease? That is, are you more likely to have a disease if you have “A” allele?
- Harder to tell than the previous example
- Let’s compute the association statistic to find it out

Association Statistics (vs. coin tossing)

	Coin Tossing	Association Statistics
1. assume no effect	Coin is fair	SNP X does not cause a disease
In terms of Math	Probability of the coin having heads is 0.5	Allele frequency in cases = Allele frequency in controls
Notation	p = 0.5	$\hat{p}_X^{+} = \hat{p}_X^{-}$

$$\hat{p}_X^{+} = \text{ Minor Allele Frequency (MAF) of SNP X in cases}$$

$$\hat{p}_X^{-} = \text{ Minor Allele Frequency (MAF) of SNP X in controls}$$

Association Statistics (vs. coin tossing)

	Coin Tossing	Association Statistics
2. Compute Quantity of interest	Count the number of heads	Compute MAF of SNP X in cases and MAF of SNP X in controls
Example	# of heads = 14	400 (out of 1000) cases have A allele 300 (out of 1000) controls have A allele *A is the minor allele
Notation	NA	$\hat{p}_X^+ = \frac{400}{1000 * 2} = 0.2$ $\hat{p}_X^- = \frac{300}{1000 * 2} = 0.15$

- Now, we consider that humans have two chromosomes (that's why we divide by 1000***2**, not 1000)

Association Statistics (vs. coin tossing)

	Coin Tossing	**Association Statistics**
3. Transform data to distribution	NA	Compute Z-score of SNP from MAF of cases and controls
Equation	NA	$z_X = \dfrac{\hat{p}_X^+ - \hat{p}_X^-}{\sqrt{2/N}\sqrt{\hat{p}_X(1-\hat{p}_X)}}, \hat{p}_X = \dfrac{\hat{p}_X^+ + \hat{p}_X^-}{2}$
Example	NA	$z_X = \dfrac{0.2-0.15}{\sqrt{2/2000}\sqrt{0.175(1-0.175)}} = 4.161252$

Association Statistics (vs. coin tossing)

	Coin Tossing	**Association Statistics**
4. Use distribution to measure significance	What is the probability of observing 14 heads or more up to 20 when p = 0.5	What is the probability of observing Zx under the null hypothesis? $\hat{p}_X^+ = \hat{p}_X^- \longrightarrow Z_X \sim \mathcal{N}(0,1)$ Under null, Zx follows the standard normal distribution (mean = 0, variance = 1)
Equation	2*pbinom(13,20,0.5,lower.tail=F)	2*pnorm(4.161252, lower.tail=F) (no need to specify mean and variance because default values are standard normal distribution)
P-value	0.115	3.165076e−05
Is significant? (significance level = 0.05)	No	Yes

Association Statistics (conclusion)

- Since p-value we got (3.165076e−05) is less than the significance level (0.05), we reject the null hypothesis
  - Null hypothesis says MAF in cases = MAF in controls
- In other words, this SNP is associated with a disease

HW0B Pr 1 & 2

1. Imagine you toss a fair coin 20 times. Let A be the event that at least one coin toss comes up heads. What is the probability that A occurs? (Hint: Think about the complement of A, A^c) Lets do this calculation using R. Write a R code to calculate the probability.

- A^c is none of coin tosses comes up heads, and $P(A) = 1 - P(A^c)$.
- $P(A^c) = (0.5) ^ 20$, so $P(A) = 1 - (0.5)^20$
- pbinom(0,20,0.5,lower.tail=FALSE) => Pr(X > 0)

2. It is known any one screw produced by a certain company will be defective with a probability of 0.01. The company sells the screws in packages of 10 and offers a money-back guarantee if at least 2 of 10 screws are defective. What proportion of packages sold must the company replace?

- Compute the probability that at least 2 of 10 screws are defective
- pbinom(1,10,0.01,lower.tail=FALSE) => Pr(X > 1)

HW0B Pr 3

3. A long question, but essentially, this questions asks you to compute p-values given statistics.

- CDF of chi-square distribution in R is pchisq
- R-code for when LRT statistics is 0.05
  - pchisq(0.05,1,lower.tail=FALSE) = 0.8230
  - The first parameter is a statistic
  - The second parameter is a degree of freedom
  - This is not significant (or not causal) since p-value is greater than significance threshold (0.001)

HW0B Pr 4

Part A. Imagine that we have divided the genome up into 611 bins of 20 kb each. Assuming that linkages are uniformly distributed across the genome (any given linkage has an equal probability of being in one of the bins), what's the probability that a linkage occurs in a bin?

- 1/611

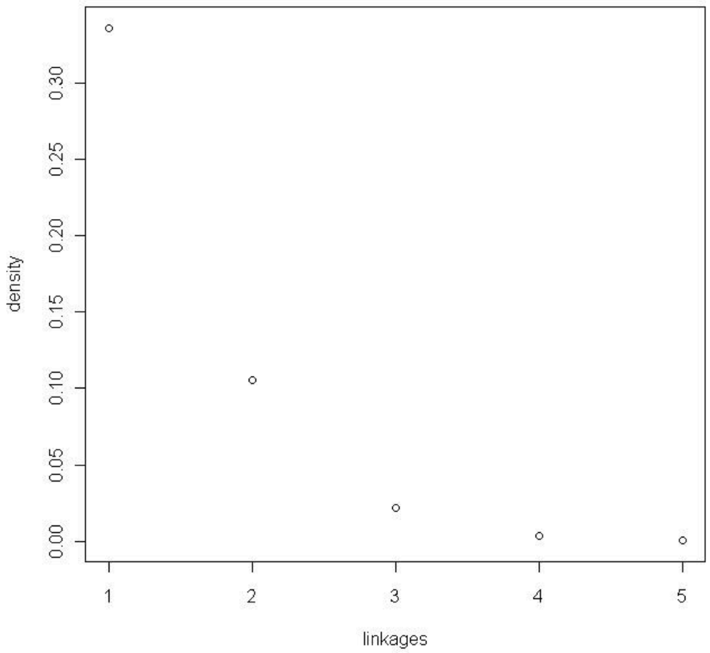
Part B. Let X be the random variable counting the number of significant linkages a bin would contain. What's the probability that a bin would contain X > 1 significant linkages. (Hint: Assume a binomial distribution with n = 385 and p equal to what was computed in Part A and use pbinom. If you couldn't compute Part A, use p = 0.001.)

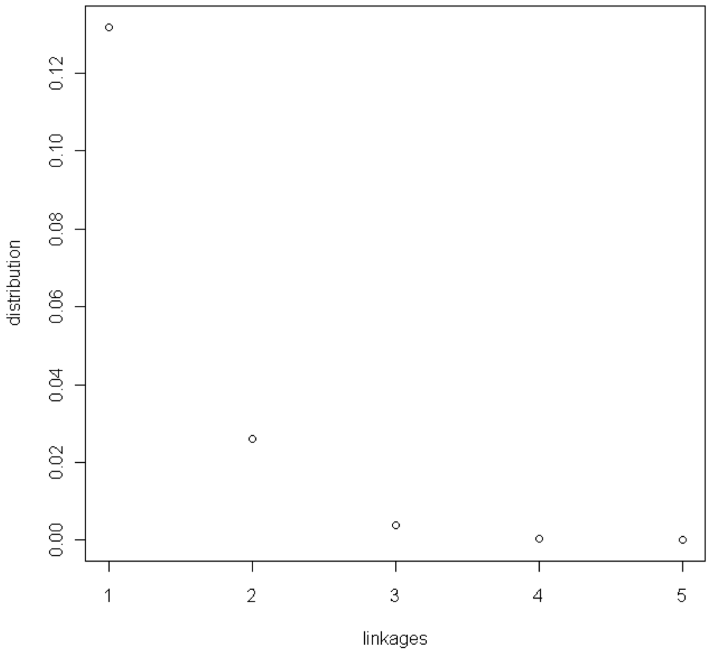
- pbinom(1, 385, (1/611), lower.tail = FALSE)

Part C. Again, let X be the number of significant linkages a bin would contain. Compute and plot the probability density functions for X = x and the cumulative distribution functions for X > x where x ∈ {1, 2, 3, 4, 5}

- PDF: plot(1:5, dbinom(1:5, 385, (1/611)), xlab="linkages", ylab="density")
- CDF: plot(1:5, pbinom(1:5, 385, (1/611), lower.tail=FALSE), xlab="linkages", ylab="distribution")

HW0B Pr 4


density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1
2
3
4
5
linkages


distribution
0.00
0.02
0.04
0.06
0.08
0.10
0.12
1
2
3
4
5
linkages

HW0B Pr 4

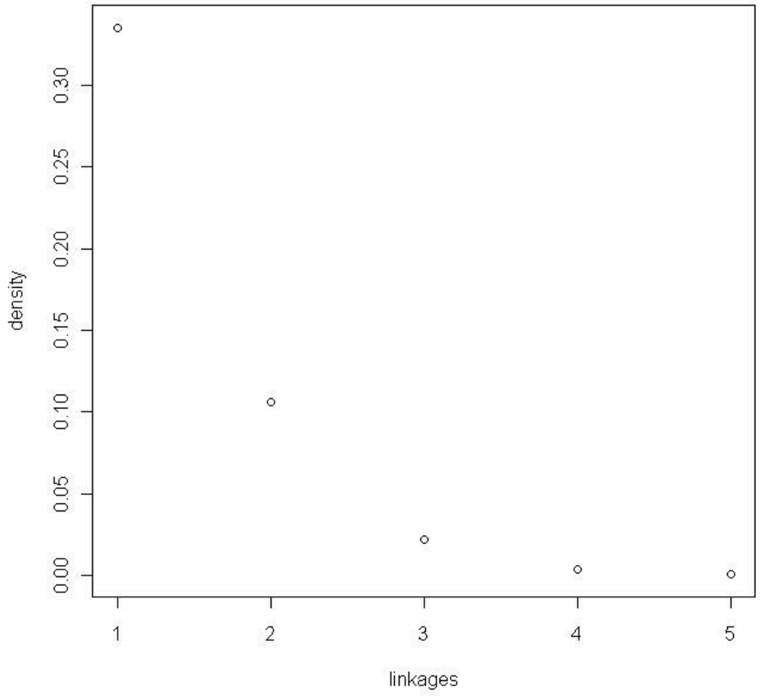
Part D. We often model rare events using the Poisson distribution. In this case, we can model the rare event of a significant linkage occuring in a bin. At what rate do significant linkages fall into a bin? (Hint: Rate is the ratio between the number of significant linkages and the number of bins.)

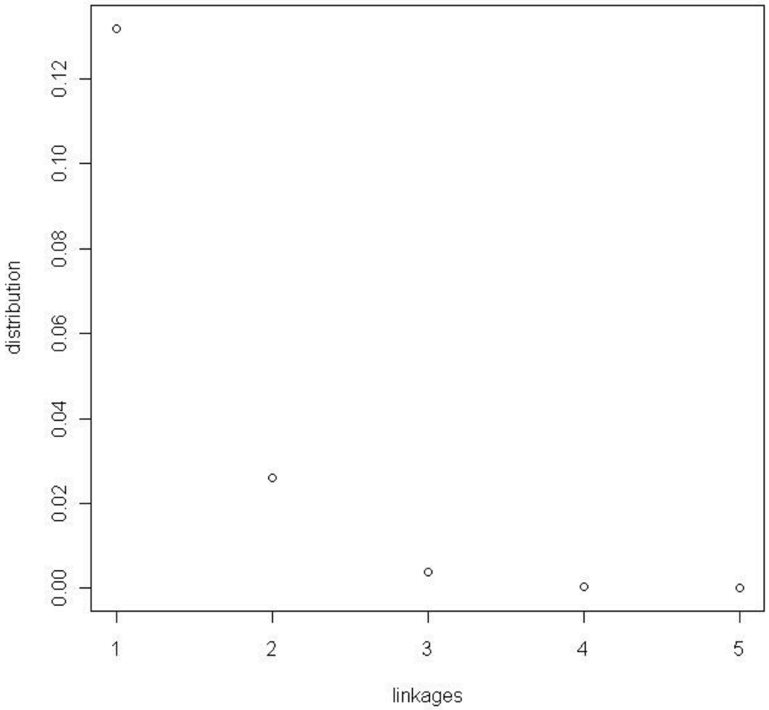
- 385/611

Part E. The rate is the only parameter you need for the Poisson distribution. Compute and plot the probability density functions for X = x and the cumulative distribution functions for the Poisson distribution for X > x where x ∈ {1, 2, 3, 4, 5}. Does it look similar to Part B?

- PDF and CDF for the Poisson distribution in R is dpois, ppois
- PDF: plot(1:5, dpois(1:5, (385/611)), xlab="linkages", ylab="density”)
- CDF: plot(1:5, ppois(1:5, (385/611), lower.tail=FALSE), xlab="linkages", ylab="distribution")

HW0B Pr 4


density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1
2
3
4
5
linkages


distribution
0.00
0.02
0.04
0.06
0.08
0.10
0.12
1
2
3
4
5
linkages