Statistical Data Analysis

Assist. Prof. Dr. Zeyneb KURT

(Slides have been prepared by
Prof. Dr. Nizamettin AYDIN,
updated by Zeyneb KURT)

zeyneb@yildiz.edu.tr
http://avesis.yildiz.edu.tr/zeyneb/

Data Exploration (cont'd)

Coefficient of Variation

- In general, the coefficient of variation is used to compare variables in terms of their dispersion when the means are substantially different
  – possibly as the result of having different measurement units.
- To quantify dispersion independently from units, we use the coefficient of variation,
  – which is the standard deviation divided by the sample mean
    - assuming that the mean is a positive number:

$$CV = \frac{s}{\bar{x}}$$

Scaling and Shifting Variables

- In general, when we multiply the observed values of a variable by a constant a, its mean, standard deviation, and variance are multiplied by a, $|a|$, and a^2, respectively.
  - That is, if $y = ax$, then
    - $\bar{y}=a\bar{x}$, $s_y=|a|s_x$, $s_y^2=a^2 s_x^2$
- The coefficient of variation is not affected.

$$CV_y = \frac{s_y}{\bar{y}} = \frac{as_x}{a\bar{x}} = \frac{s_x}{\bar{x}} = CV_x$$

Scaling and Shifting Variables

- If we shift the observed values by b, i.e., $y = x + b$, then

  $\bar{y}=\bar{x} + b,$ $\quad s_y=s_x,$ $\quad s_y^2=s_x^2$

- If we multiply the observed values by the constant a and then add the constant b to the result, i.e., $y = ax + b$, then

  $\bar{y}=a\bar{x} + b,$ $\quad s_y=|a|s_x,$ $\quad s_y^2=a^2 s_x^2$

- the coefficient of variation will change. If $y = ax + b$ (assuming $a > 0$ and $b = 0$), then

$$CV_y = \frac{s_y}{\bar{y}} = \frac{a s_x}{a\bar{x} + b} \neq \frac{s_x}{\bar{x}}.$$

Variable Standardization

- Variable standardization is a common *linear* transformation,
  - where we subtract the sample mean $\bar{x}$ from the observed values and divide the result by the sample standard deviation s,
    - in order to shift the mean to zero and make the standard deviation 1:

$$y_i = \frac{x_i - \bar{x}}{s}.$$

- Using such transformation is especially common in regression analysis and clustering.
- Subtracting $\bar{x}$ from the observations shifts the sample mean to zero.
  - This, however, does not change the standard deviation.
- Dividing by s, on the other hand, changes the sample standard deviation to 1

Data Exploration with R Programming

- Load *Pima.tr* data set, which is available from MASS package

```
> library(MASS)
> data(Pima.tr)
```

- The *head*() function shows only the first part of the data set.

```
> head(Pima.tr)
```

- Use the *help*() function to view description on the data available in the package

```
> help(Pima.tr)
```

- Use *table*() function to obtain the frequencies for the catagorical variable

```
> type.freq <- table(Pima.tr$type)
> type.freq
 No   Yes
 132  68
```

Note that the $ symbol is being used to access the type variable in the Pima.tr data set.

Data Exploration with R Programming

- Now, use the *type.freq* table to create the bar graph.

  > barplot(type.freq, xlab = "Type", ylab = "Frequency", main = "Frequency Bar Graph of Type")

  The first parameter to the *barplot*() function is the frequency table. The options *xlab* and *ylab* label the *x* and *y* axes, respectively. Likewise, the *main* option puts a title on the plot.

- The relative frequency can be calculated as

```
> n <- sum(type.freq)
> type.rel.freq <- type.freq/n
> round(type.rel.freq, 2)
> round(type.rel.freq, 2) * 100
```

Data Exploration with R Programming

- If the levels of a categorical variable in the data set is coded as numbers, we need to convert the type of variable to *factor* using the *factor*() function, so that R recognizes it as categorical.
- You can use the function *is.factor*() to examine whether a variable is a factor.

```
> data(birthwt)
> is.factor(birthwt$smoke)
    [1] FALSE
> birthwt$smoke <- factor(birthwt$smoke)
> is.factor(birthwt$smoke)
    [1] TRUE
> table(birthwt$smoke)
     0    1
     115  74
```

Data Exploration with R Programming

- To create a *frequency* histogram for age, use the *hist*() function with the freq option set to "TRUE" (which is the default):

  ```
  > hist(Pima.tr$age, freq = TRUE, xlab = "Age", ylab = "Frequency", col = "grey", main = "Frequency Histogram of Age")
  ```

- Then create a *density* histogram of age by setting the freq option to "FALSE":

  ```
  > hist(Pima.tr$age, freq = FALSE,xlab = "Age", ylab = "Density", col = "grey", main = "Density Histogram of Age")
  ```

Data Exploration with R Programming

- We can obtain the mean and median of numerical data with the mean() and median() functions.
- Find these statistics for numerical variables in Pima.tr:

```
> mean(Pima.tr$npreg)
    [1] 3.57
> median(Pima.tr$bmi)
    [1] 32.8
```

- The quantile() function with the probs option returns the specified quantiles:

```
> quantile(Pima.tr$bmi, probs = c(0.1, 0.25, 0.5, 0.9))
    10%     25%     50%     90%
    24.200  27.575  32.800  39.400
```

- The five-number summary along with the mean can simply be obtained with the summary() function:

```
> summary(Pima.tr$bmi)
    Min.    1st Qu.  Median  Mean    3rd Qu.  Max.
    18.20   27.58    32.80   32.31   36.50    47.90
```

Data Exploration with R Programming

- We can present the five-number summary visually with a boxplot:

```
> boxplot(Pima.tr$bmi, ylab = "BMI")
```

- While the default is to create vertical boxplots, we can also create horizontal boxplots by specifying the horizontal option to true:

```
> boxplot(Pima.tr$bmi, ylab = "BMI", horizontal = TRUE)
```

- Find the interquartile range (IQR) with the IQR() function:

```
> IQR(Pima.tr$bmi)
    [1] 8.925
```

- The smallest and largest observations can be obtained with the range() function
    - the functions min() and max() could also be applied):

```
> minMax <- range(Pima.tr$bmi)
> minMax
    [1] 18.2 47.9
```

Data Exploration with R Programming

- The variance and standard deviation are also easily calculated with var() and sd():

```
> var(Pima.tr$bmi)
    [1] 37.5795
> sd(Pima.tr$bmi)
    [1] 6.130212
```

Data Exploration with R Programming

- Creating Categories for Numerical Variables:
  - To create a categorical variable weight.status based on the *bmi* variable in *Pima.tr*, we can go through each observation one by one and assign each observation to one of the four categories:
    - "Underweight",
    - "Normal",
    - "Overweight",
    - "Obese".
  - First, we start by creating an empty vector of size 200 within the *Pima.tr* data frame:

```
> Pima.tr$weight.status <- rep(NA, 200)
```

Data Exploration with R Programming

- Then,
  - We can either use loops and conditional statements
  - Or we can simple use which() function as follows:

```
> Pima.tr$weight.status[which(Pima.tr$bmi<18.5)] <- “Underweight”
> Pima.tr$weight.status[which(Pima.tr$bmi>=18.5 & Pima.tr$bmi < 25 )] <- “Normal”
> Pima.tr$weight.status[which(Pima.tr$bmi>= 25 & Pima.tr$bmi < 30)] <- “Overweight”
> Pima.tr$weight.status[which(Pima.tr$bmi>=30)] <- “Obese”
> Pima.tr$weight.status <- factor(Pima.tr$weight.status)
> Pima.tr$weight.status <- factor(Pima.tr$weight.status, levels(Pima.tr$weight.status)[c(4,1,3,2)])
> barplot(table(Pima.tr$weight.status))
```

Exploring Relationships

Introduction

- So far, we have focused on using graphs and summary statistics to explore the distribution of individual variables.
- In this lecture we discuss using graphs and summary statistics to investigate relationships between two or more variables.
  - We want to develop a high-level understanding of the type and strength of relationships between variables.
- We start by exploring relationships between two numerical variables.
  - We then look at the relationship between two categorical variables.
- Finally, we discuss the relationships between a categorical variable and a numerical variable.

Two numerical variables

- For illustration, we use the *bodyfat* data
  - based on a study conducted by Dr. Fisher from Human Performance Research Center at Brigham Young University
    - The study involved measuring percent body fat as the target variable, along with several explanatory variables such as age, weight, height, and abdomen circumference for a sample of 252 men.
  - The collected data set *bodyfat* is available online at http://lib.stat.cmu.edu/datasets/bodyfat
  - You can also obtain this data set from the mfp package in R.
  - To install this package, enter the following command in R Console:
    - install.packages("mfp", dependencies=TRUE)

Two numerical variables

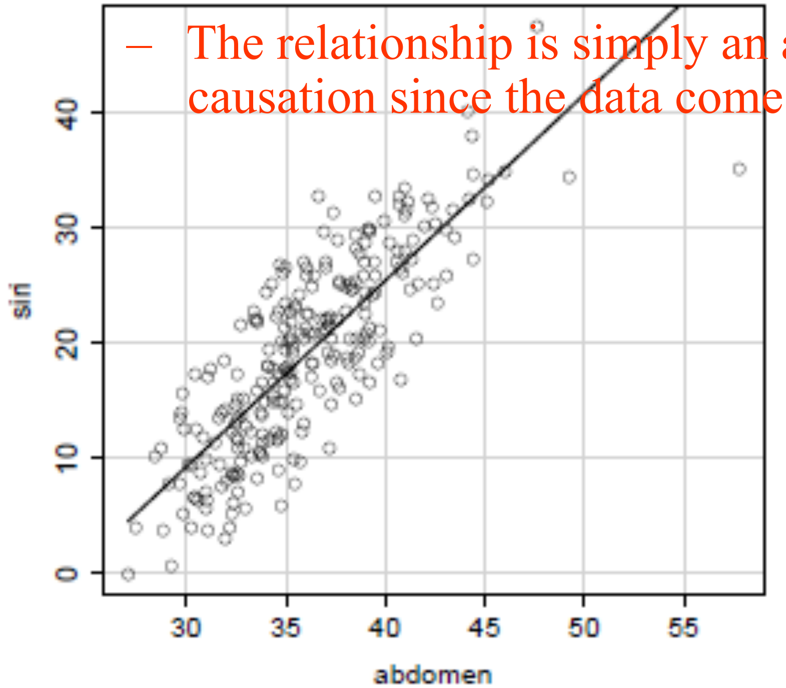
- Once the package is installed, it can be loaded into R using the following command:
  - library(mfp)
- Now you can access bodyfat by clicking
  - Data → Data in packages →Read data set from an attached package
- and selecting (doubleclicking) mfp under packages.
- You can learn more about this data set by looking at its accompanying help file.
  - In R-Commander, click
    - Data → Active data set→Help on active data set.

Two numerical variables

- Suppose that we are interested in examining the relationship between percent body fat and abdomen circumference among men.
  - Load the *bodyfat* set from the mfp package. Makesure *bodyfat* becomes the active data set and then view it.
  - For now, we are focusing on two variables, *siri* and *abdomen*.
    - The *siri* variable shows the percent body fat measurements derived based on body density using Siri's equation (percent body fat = 495/density−450).
    - The *abdomen* variable shows the abdomen circumference in centimeters.
- Both *siri* and *abdomen* are numerical variables.
  - A simple way to visualize the relationship between two numerical variables is with a scatterplot.

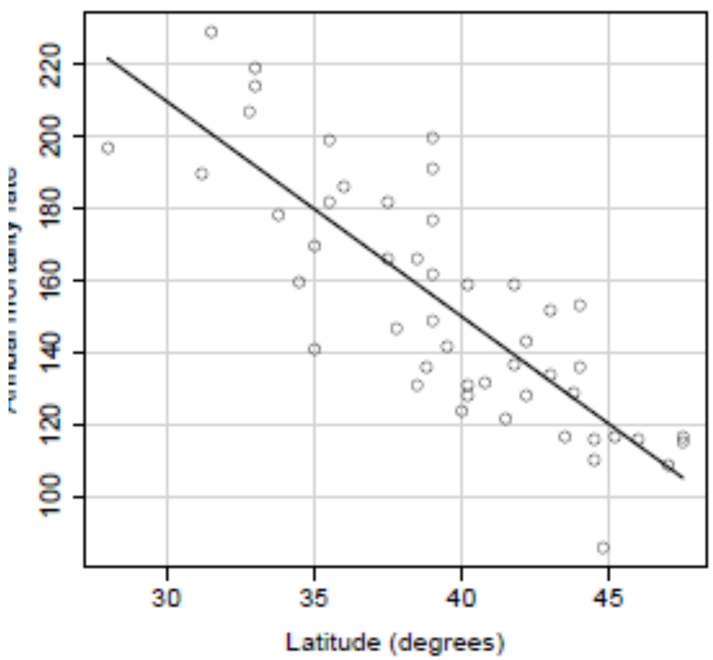
Scatterplot

- In R-Commander, click
  - Graphs → Scatterplot and select *abdomen* for the x-variable and *siri* for the y-variable.
  - Under Options, uncheck Marginal boxplots and Smooth line.
- The plot suggests that the increase in percent body fat tends to coincide with the increase in abdomen circumference.
- The two variables seem to be related with each other.
  - The relationship is simply an association and should not be regarded as causation since the data come from an observational study.



Scatterplot

- As the second example, we examine the relationship between the annual mortality rate due to malignant melanoma for US states and the latitude of their geographical centers.
- The data are collected from the population of white males in the US during 1950–1969.
- You can obtain this data set, called *USmelanoma*, from the HSAUR2 package.
  - [Follow the above steps to install and load the package]



- The two variables are clearly associated since the increase in latitude tends to coincide with the decrease in mortality rate.

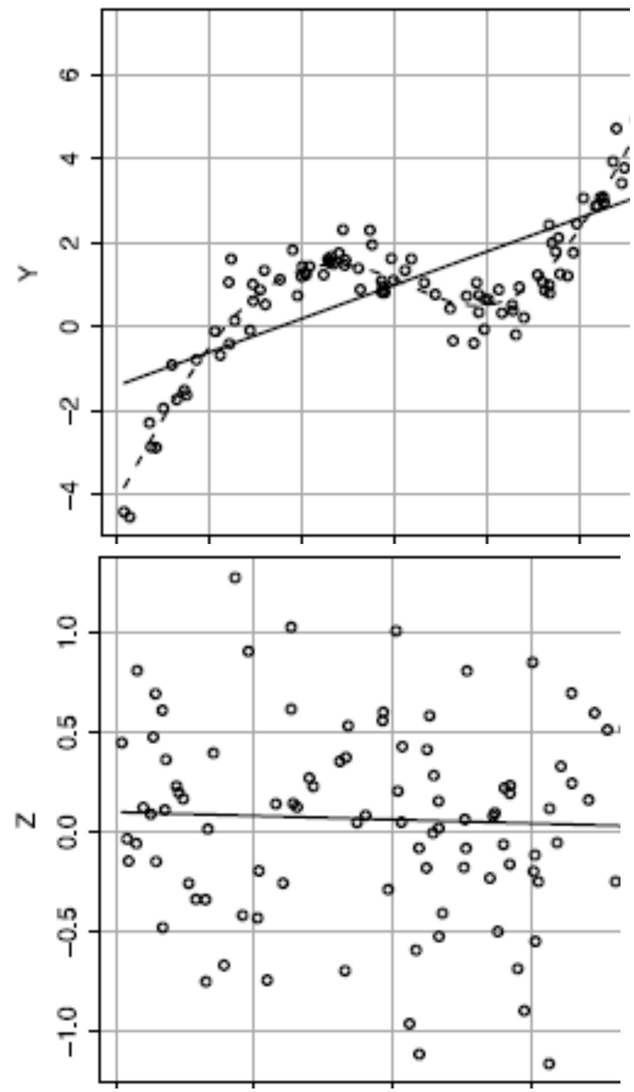
Scatterplot

- Using scatterplots, we could detect possible relationships between two numerical variables.
  - In above examples, we can see that changes in one variable coincides with substantial systematic changes (increase or decrease) in the other variable.
- Since the overall relationship can be presented by a straight line, we say that the two variables have linear relationship.
  - We say that percent body fat and abdomen circumference have positive linear relationship.
  - In contrast, we say that annual mortality rate due to malignant melanoma and latitude have negative linear relationship.

Scatterplot

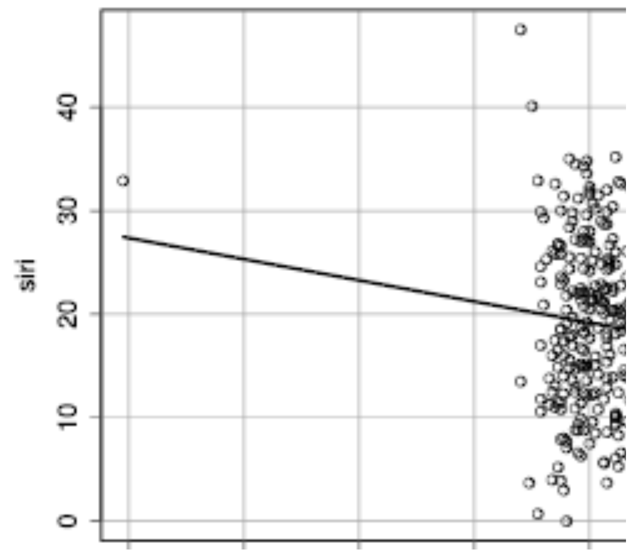
- In some cases, the two variables are related, but the relationship is not linear.

- In some cases, there is no relationship (linear or non-linear) between the two variables.

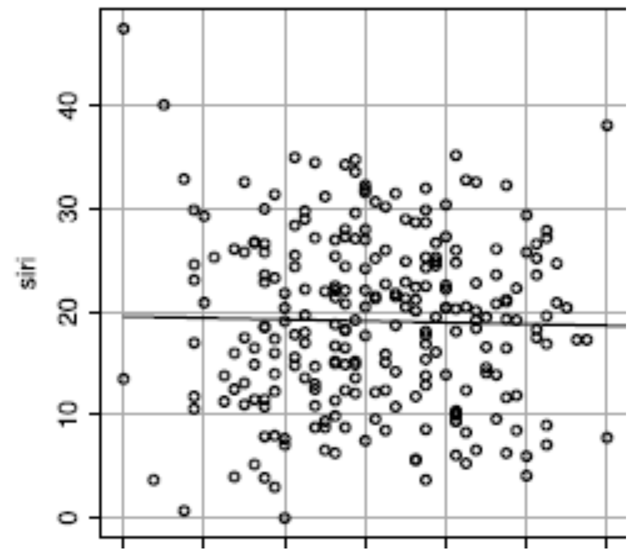


Scatterplot

- The scatterplot of percent body fat by height from the *bodyfat* data set.
  - The isolated point at the left of the graph is an outlier, which has a drastic influence on the overall pattern.



- The scatterplot of percent body fat by height after removing the outlier.
  - The two variables seem to be unrelated

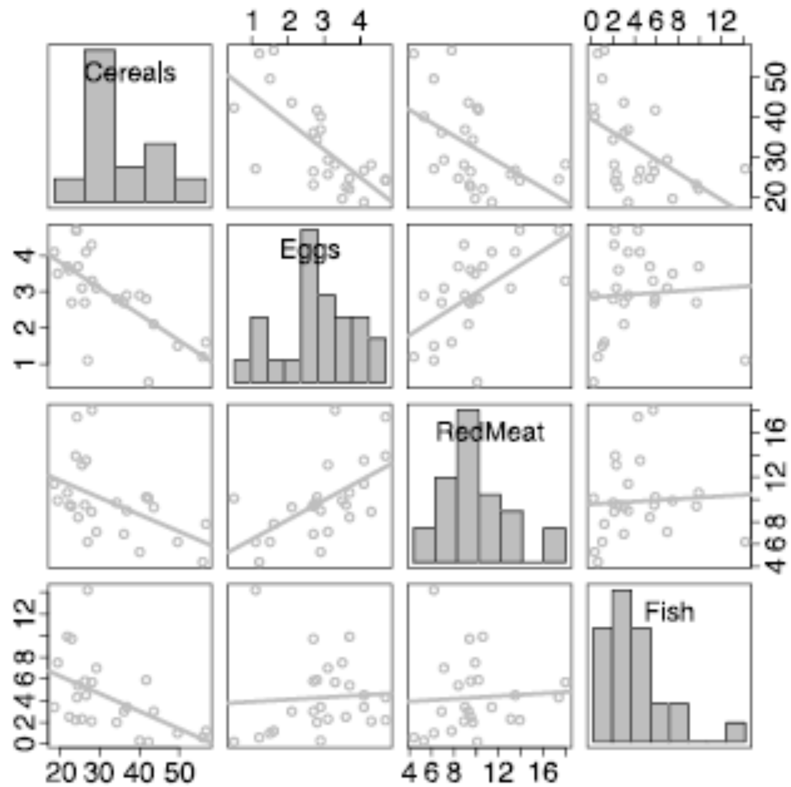


Scatterplot

- In practice, we should never remove an outlier just simply because it does not follow the overall pattern.
- Some outliers are due to rare events, which provide important information about the distribution of the corresponding variable.
- Even when we identify a data entry mistake, we should try to correct the mistake and keep the observation if possible.

Scatterplot Matrix

- Obtaining and viewing a *scatterplot matrix* in R-Commander.



  – The diagonal elements are histograms, and the off-diagonals are scatterplots with a trend line

Correlation

- To quantify the strength and direction of a linear relationship between two numerical variables,
  - we can use Pearson's correlation coefficient, r , as a summary statistic.
    - The values of r are always between -1 and +1.
    - The relationship is strong when r approaches -1 or +1.
    - The sign of r shows the direction (negative or positive) of the linear relationship.

Correlation

- Consider a set of observed pairs of values, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, for a sample of n observations.
- For these observed pairs of values, Pearson's correlation coefficient is calculated as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

  – For the two variable, s_x and s_y denote the sample standard deviations

Correlation

- Suppose that we have measured the height in inches and weight in pounds for five people.

Index	Height	Weight
1	62	160
2	71	198
3	65	173
4	73	182
5	60	143
Mean	66.2	171.2
Standard deviation	5.6	21.0

  – We denote height as X and weight as Y

Correlation

- Calculating Pearson's correlation coefficient for height and weight

Index	x	$x-\bar{x}$	y	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
1	62	-4.2	160	-11.2	47.04
2	71	4.8	198	26.8	128.64
3	65	-1.2	173	1.8	-2.16
4	73	6.8	182	10.8	73.44
5	60	-6.2	143	-28.2	174.84

$$r_{xy} = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} = \frac{1}{4}\frac{421.8}{5.6\times 21.0} = 0.89$$

Correlation

- We can use R-Commander to calculate the sample correlation coefficient.
- To calculate r for percent body fat and abdomen circumference, make sure *bodyfat* is the active data set, then click
  - *Statistics → Summaries → Correlation matrix*
- Select both *abdomen* and *siri*. (You need to hold the *control* key.)
  - The output is in the form of a symmetric matrix called the *correlation matrix*, where the value in row i and column j is the correlation coefficient between the ith and j th variables.

Correlation

- Obtaining and viewing the correlation between percent body fat and abdomen circumference in R-Commander

Output Window Submit

```
> cor(bodyfat[,c("abdomen","siri")], use="complete.obs")
          abdomen      siri
abdomen 1.0000000 0.8134323
siri    0.8134323 1.0000000
```

- Correlation matrix for most of the numerical variables in the *Protein* data set

Output Window Submit

```
> cor(Protein[,c("Cereals","Eggs","Fish","RedMeat")], use="complete.ob
           Cereals        Eggs        Fish     RedMeat
Cereals  1.0000000 -0.71243682 -0.52423080 -0.49987746
Eggs    -0.7124368  1.00000000  0.06557136  0.58560895
Fish    -0.5242308  0.06557136  1.00000000  0.06095745
RedMeat -0.4998775  0.58560895  0.06095745  1.00000000
```

Sample Covariance

- If the standard deviations are removed from the denominator in Pearson's correlation coefficient, the statistic is called the sample covariance,

$$v_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Therefore

$$r_{xy} = \frac{v_{xy}}{s_x s_y}$$

Two categorical variables

- We now discuss techniques for exploring relationships between categorical variables.
- As an example, we consider the five-year study to investigate whether regular aspirin intake reduces the risk of cardiovascular disease.
    - [“Findings from the aspirin component of the ongoing Physicians’ health study” in *New England Journal of Medicine* in 1988].
  - In this randomized experiment, 22071 physicians were randomly divided into two groups: 11037 physicians took an aspirin every other day, while 11034 physicians took a placebo. The investigators then recorded the number of people who suffered a heart attack within the five-year follow-up period.

Two categorical variables

- We usually use contingency tables to summarize such data.

	Heart attack	No heart attack	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

- Each cell shows
  - the frequency of one possible combination of disease status
    - heart attack or no heart attack
  - experiment group
    - placebo or aspirin
      - [A placebo is a substance or treatment with no active therapeutic effect. It may be given to a person in order to deceive the recipient into thinking that it is an active treatment]

Two categorical variables

- Using these frequencies, we can calculate the sample proportion of people who suffered from heart attack in each experiment group separately.
  - There were 11034 people in the placebo group, of which 189 had heart attack.
  - The proportion of people suffered from a heart attack in the placebo group is therefore

    $p_1 = 189/11034 = 0.0171.$
  - The proportion of people suffered from heart attack in the aspirin group is

    $p_2 = 104/11037 = 0.0094.$

Two categorical variables

- We refer to this as the risk (here, the sample proportion is used to measure risk) of heart attack.
- Substantial difference between the sample proportion of heart attack between the two experiment groups could lead us to believe that the treatment and disease status are related.
- One way of measuring the strength of the relationship is to calculate the difference of proportions, p_2-p_1.
  - Here, the difference of proportions is $p_2-p_1 = -0.0077$.

Two categorical variables

- The proportion of people suffered from heart attack reduces by 0.0077 in the aspirin group compared to the placebo group.
- We can present this difference as a percentage using the sample proportion (risk) in the placebo group as the baseline:

$$\frac{p_2 - p_1}{p_1} \times 100\% = \frac{-0.0077}{0.0171} \times 100\% = -45\%.$$

- This means that the risk of heart attack reduces by 45% in the aspirin group compared to the placebo group.

Two categorical variables

- Another common summary statistic for comparing sample proportions is the relative proportion p_2/p_1.
  - Since the sample proportions in this case are related to the risk of heart attack, we refer to the relative proportion as the relative risk.
- Here, the relative risk of sufering from heart attack is

$$p_2/p_1 = 0.0094/0.0171 = 0.55$$

Two categorical variables

- This means that the risk of a heart attack in the aspirin group is 0.55 times of the risk in the placebo group.
- If the two sample proportions are equal, the relative proportion (risk) is equal to 1,
  - which is interpreted as no relationship between the two categorical variables.
- Values of the relative proportion away from 1 (either below 1 or above 1) indicate that the relationship is strong.

Two categorical variables

- It is more common to compare the sample odds,

$$o = \frac{p}{1-p}$$

  – where p is the sample proportion for the event of interest (e.g., heart attack).

- The odds of a heart attack in the placebo group, o_1, and in the aspirin group, o_2, are

$$o_1 = \frac{0.0171}{(1 - 0.0171)} = 0.0174, \quad o_2 = \frac{0.0094}{(1 - 0.0094)} = 0.0095.$$

Two categorical variables

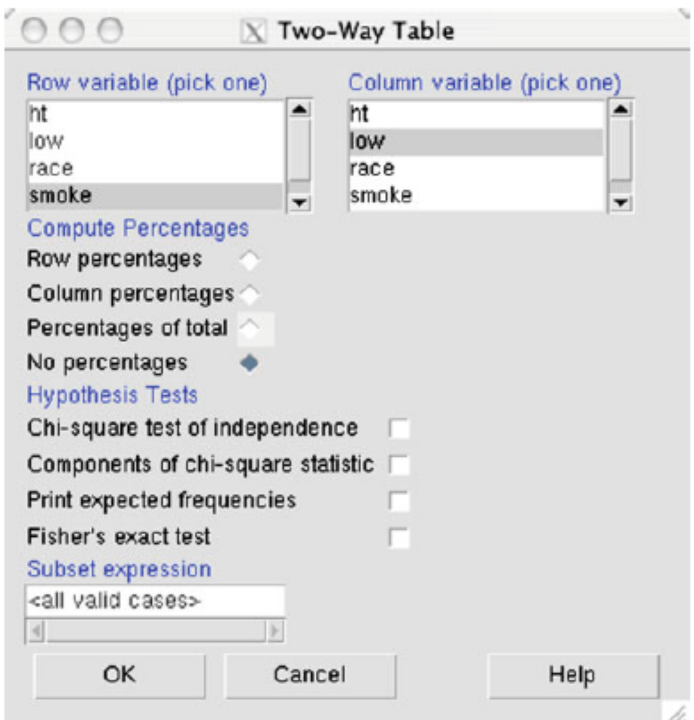
- We usually compare the sample odds using the sample odds ratio

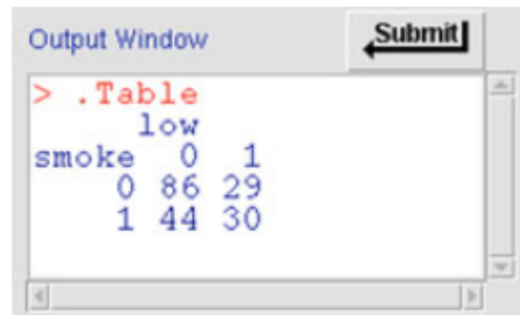
$$OR_{21} = \frac{o_2}{o_1} = \frac{0.0095}{0.0174} = 0.54.$$

  - The index “21” shows that we are dividing the odds in the second group (here, the aspirin group) by the odds in the first group (here, the placebo group).
    - An odds ratio equal to 1 means that the odds are equal in both groups and is interpreted as no relationship between the two categorical variables.
    - Values of the odds ratio away from 1 (either greater than or less than 1) indicate that the relationship is strong.
  - Note that the odds ratio cannot be negative.
    - Therefore, its smallest possible value is zero.

Two categorical variables

- Contingency table for *smoke* and *low* in *birthwt* data set
  - For creating the contingency table for smoke and low, click
    - *Statistics → Contingency tables → Two-way table.*



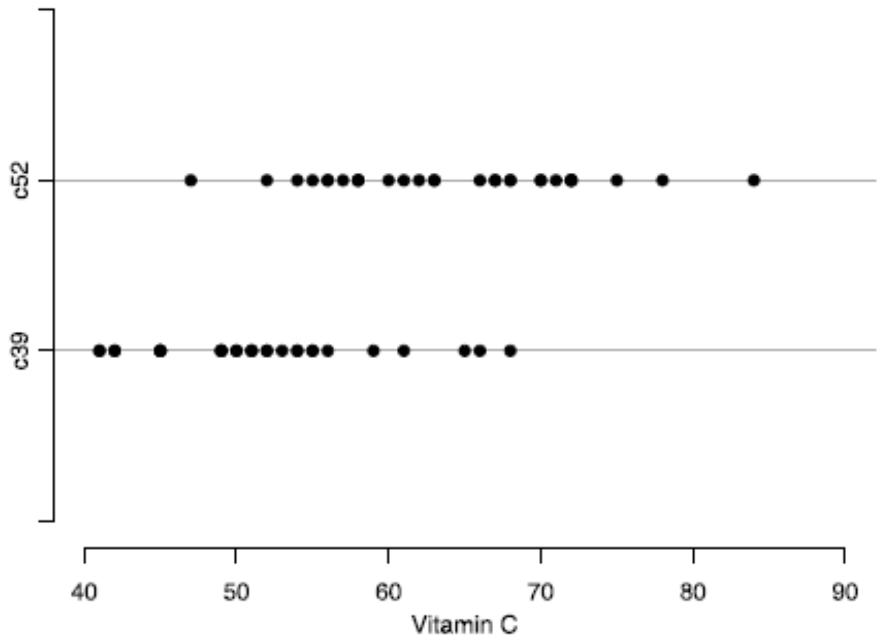


Numerical and Categorical Variables

- Very often, we are interested in the relationship between a categorical variable and a numerical random variable.
- When the sample size is small, we can visualize the relationship by simply creating dot plots of the numerical variable for different levels of the categorical variable.
- As an example, we use the *cabbages* data set available from the MASS package.

Numerical and Categorical Variables

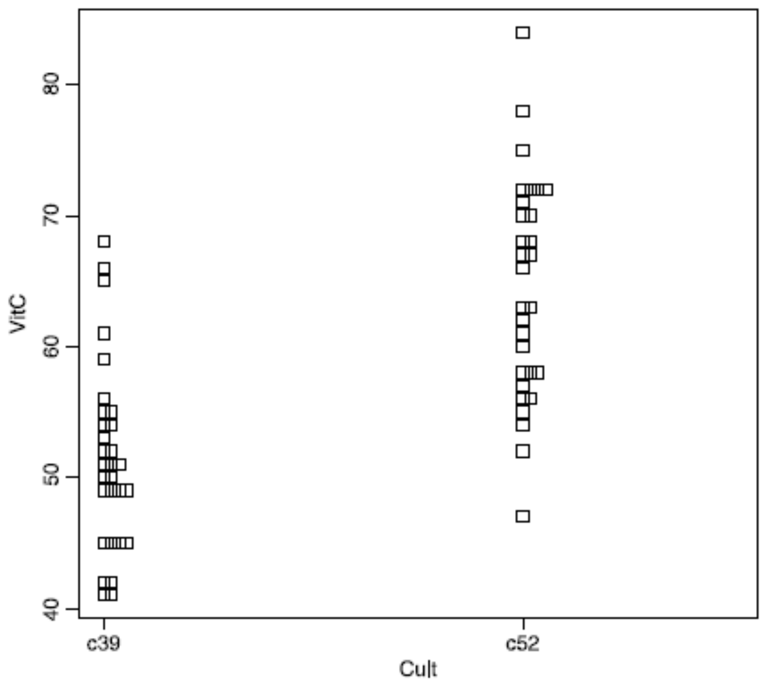
- The dot plots of *ascorbic acid* (one form of vitamin C) *content* (numerical) by *cultivar* (categorical).
- The categorical variable has two possible categories: c39 and c52.
- It shows that the distribution of *vitamin C content* is different between the two *cultivars*.



- The central tendency for the observed values in the c39 group is around 50, whereas the central tendency for the c52 group is around 65.
- **In general, we say that two variables are related if the distribution of one of them changes as the other one varies**.

Numerical and Categorical Variables

- In the above example, the two variables, *vitamin C content* and *cultivar*, seem to be related.
- We can use R-Commander to create a dot plot (a.k.a. strip chart) similar to the one presented in previous slide.

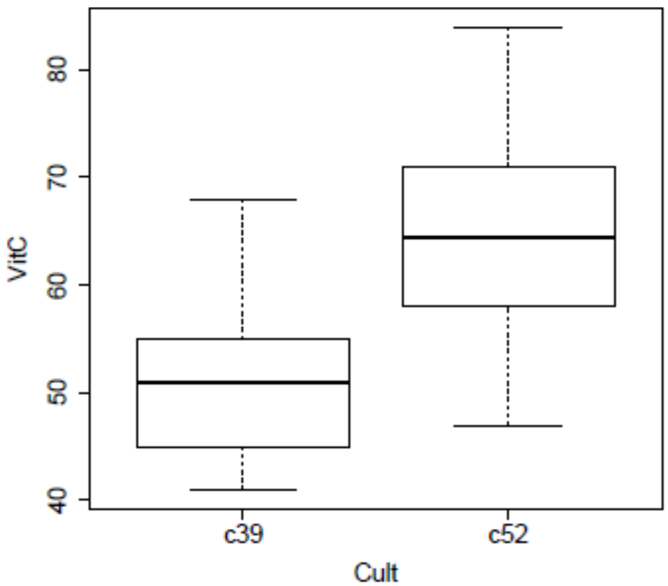


- Strip chart for *vitamin C content* (*VitC*) by *cultivar* (*Cult*) from the *cabbages* data set
- Here, multiple observations with the same value of the numerical variable are stacked toward the right.
- Overall, vitamin C content tends to be higher in the c52 group compared to the c39 group.

Numerical and Categorical Variables

- A more common way of visualizing the relationship between a numerical variable and a categorical variable is
  - to create boxplots of the numerical variable for different values of the categorical variable.
- This is especially useful when the sample size is large.
  - By focusing on some key aspects of the distributions, namely the five-number summaries, boxplots make the patterns easier to detect.
- In R-Commander, click
  - *Graphs→Boxplot*; select *VitC* as the Variable.
- Then click on
  - *Plot by groups* button and in the resulting window,
- Select
  - *Cult* as the *Groups variable*.

Numerical and Categorical Variables



– The resulting Boxplot of vitamin C content for different cultivars

– Summary statistics of *vitamin C content* by *cultivar* from the *cabbages* data set

Output Window

	mean	sd	0%	25%	50%	75%	100%	n
c39	51.5	7.123298	41	46	51.0	54.75	68	30
c52	64.4	8.455156	47	58	64.5	70.75	84	30

- This plot suggests that
  - vitamin C content tends to be higher in the c52 group compared to the c39 group.
    - This is indicative of a possible relationship between these two variables.

Numerical and Categorical Variables

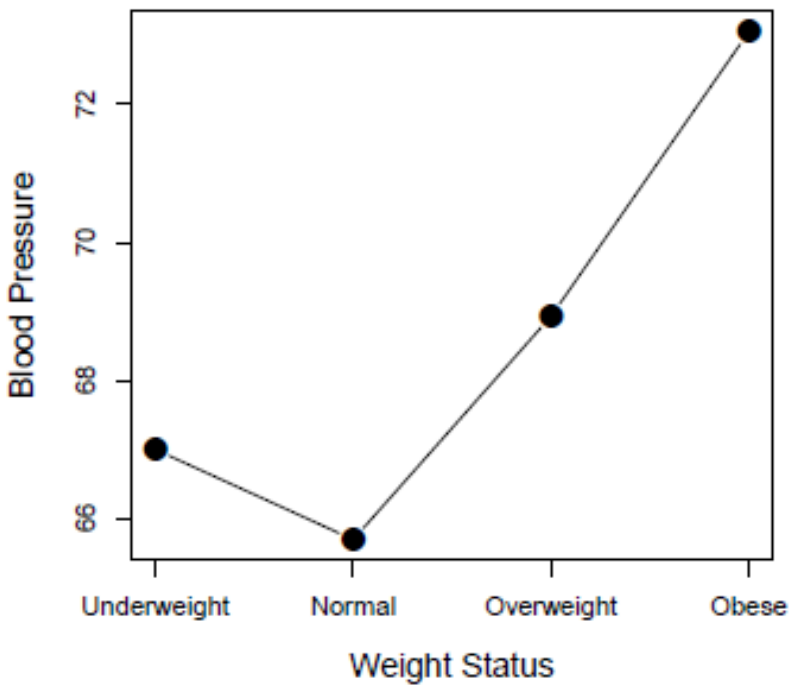
- In general, we say that two variables are related if the distribution of one of them changes as the other one varies.
- We can measure changes in the distribution of the numerical variable by obtaining its summary statistics for different levels of the categorical variable.
- It is common to use the difference of means when examining the relationship between a numerical variable and a categorical variable.
  - In the above example, the difference of means of vitamin C content is 64.4 -51.5 = 12.9 between the two cultivars.

Numerical and Categorical Variables

- When the categorical variable has multiple levels (categories), it is easier to compare the means across different levels using the plot of means.
- For example,
  - previously we created a categorical variable called *weight.status* based on *BMI* values in the *Pima.tr* data set.
  - This variable had four categories:
    - "Underweight", "Normal", "Overweight", and "Obese".
  - Here, we would like to investigate how blood pressure *bp* changes with *weight.status*, which is an *ordinal* variable

Numerical and Categorical Variables

- In R-Commander,
  - Click *Graphs* → *Plot of means* and
  - select *weight.status* as the *Factors* and *bp* as the *Response Variable*.
- For now, choose *no error bars*.



- The resulting graph shows that
  - compared to the Normal group, the average blood pressure increases for both Underweight and Overweight group.
  - The Obese group has the highest blood pressure average.
- Also, note that
  - as we move toward higher levels of weight group, average blood pressure first decreases and then increases.

Probability

Probability as a Measure of Uncertainty

- Plots and summary statistics are used to learn about the distribution of variables and to investigate their relationships.
  - However, we always remain uncertain about the true distributions and relationships in the population since we almost never have access to all of its members.
  - Furthermore, our findings based on the observed sample can change if different samples from the population were obtained.
- Therefore, when we generalize our findings from a sample to the whole population, we should explicitly specify the extent of our uncertainty.
  - We use probability as a measure of uncertainty.

Some Commonly Used Genetic Terms

- Gene
  - a segment of double-stranded DNA, which itself is made of a sequence of four different nucleotides:
    - adenine (A), guanine (G), thymine (T), or cytosine (C).
- Single Nucleotide Polymorphisms (SNPs)
  - Genetic variation is caused by changes in the DNA sequence of a gene.
  - SNPs are the most common type of genetic variation.
  - SNPs occur when a single nucleotide is replaced by another one.
    - An example of a SNP would be replacing "G" in the sequence {TAGCAAT} by "T" to create {TATCAAT}.
- Alleles
  - alternate forms of a gene
  - responsible for variation in phenotypes.
    - Phenotypes, in general, are observable traits, such as eye color, disease status, and blood pressure, due to genetic factors and/or environmental factors
  - In the above example, the alleles could be denoted as T and G.
    - We denote the genes with bold face letters (e.g., **A**) and the two different alleles as capital and small letters (e.g., *A* and *a*).

Some Commonly Used Genetic Terms

- Genotype
  - Genetic materials are stored on chromosomes.
  - Human somatic cells have two copies of each chromosome
    - one inherited from each parent; hence, they are called diploid.
  - Each pair of similar chromosomes are called homologous chromosomes.
  - The genotype (i.e., genetic makeup) of an individual for the bi-allelic gene **A** can take one of the three possible forms:
    - AA, aa, or Aa.
- Homozygous vs. heterozygous
  - The first two genotypes, AA and aa, are called homozygous,
    - which means the same version of the allele was inherited from both parents.
      - That is, both homologous chromosomes have the same allele.
  - The last genotype, Aa, is called heterozygous,
    - which means different alleles were inherited.

Some Commonly Used Genetic Terms

- Phenotype
  - the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment
- Recessive vs. dominant
  - The presence of a specific allele does not always result in its corresponding trait (a characteristic such as eye color).
  - Some alleles are recessive,
    - producing their trait only when both homologous chromosomes carry that specific variant.
  - On the other hand, some alleles are dominant,
    - producing their traits when they appear on at least one of the homologous chromosomes.
      - {For example, suppose that the allele a for gene A is responsible for a specific disease.
      - Furthermore, assume that a is a recessive allele.
      - Then, only a person with genotype aa will be affected by the disease.
      - Individuals with genotype AA or Aa will not have the disease.}

Random phenomena and their sample space

- A phenomenon is called random if its outcome (value) cannot be determined with certainty before it occurs.
  - For example, coin tossing and genotypes are random phenomena.
- The collection of all possible outcomes S is called the sample space.

$$
\begin{array}{lll}
\text{Coin tossing} & : & S = \{H, T\}, \\
\text{Die rolling} & : & S = \{1, 2, 3, 4, 5, 6\}, \\
\text{Bi-allelic gene} & : & S = \{A, a\}, \\
\text{Genotype} & : & S = \{AA, Aa, aa\}.
\end{array}
$$

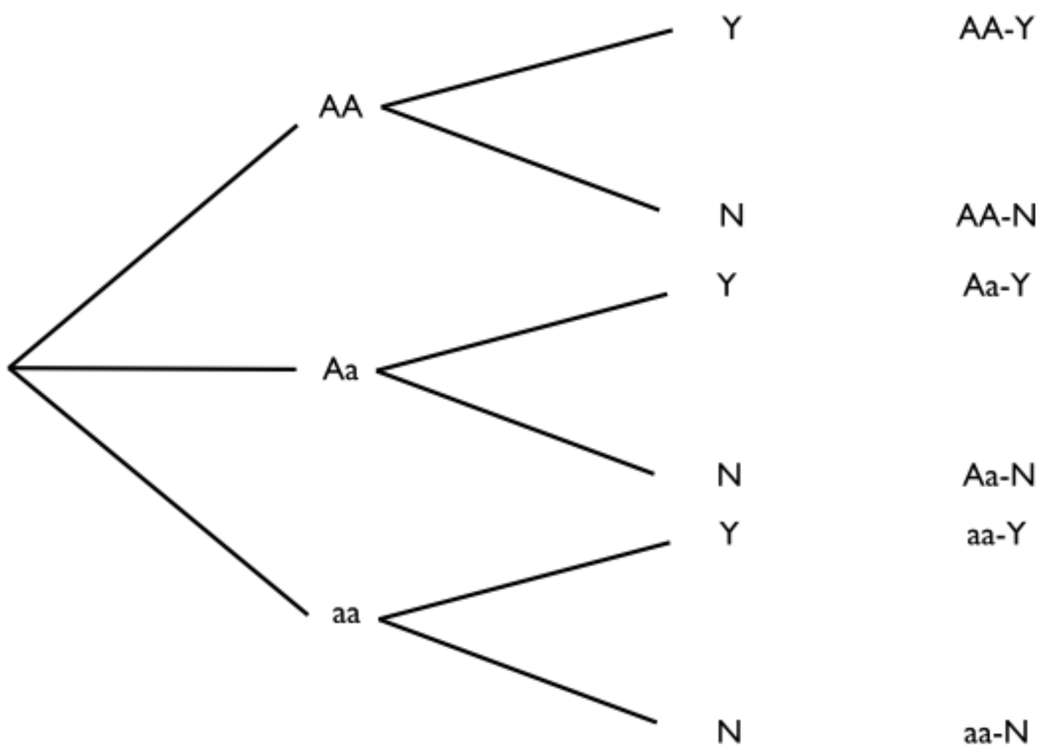
Random phenomena and their sample space

- The sample space might include an infinite number of possible outcomes.
  - For example, the value of blood pressure is random since it cannot be determined with certainty before measuring it.
    - The corresponding sample space for blood pressure values is (theoretically) the set of positive real numbers, which is infinite.
- For a complex random phenomenon that is a combination of two or more other random phenomena, it might be easier to view the sample space with tree diagrams.

Random phenomena and their sample space

- For example, suppose that we suspect that gene **A** is related to a specific disease, but genetic variation alone does not determine the disease status.
  - Rather, it affects the risk of the disease.
  - Further, we suspect that smoking (an environmental factor) is also related to the disease.
- In this case, the random phenomenon we are interested in is the combination of genotype and smoking status
- All possible combinations (i.e., sample space) are identified using the following tree diagram.

Random phenomena and their sample space



- Genotypes (*AA*, *Aa*, and *aa*) are represented by the first set of branches
- Smoking status (Y for smokers and N for nonsmokers) is represented by the second set of branches
- Following each branch from root to tip, we obtain the following sample space;
  - $S = \{AA\text{-}Y, AA\text{-}N, Aa\text{-}Y, Aa\text{-}N, aa\text{-}Y, aa\text{-}N\}$.
    - For example, Aa - Y represents the outcome of having heterozygous genotype and smoking.

Probability Measure

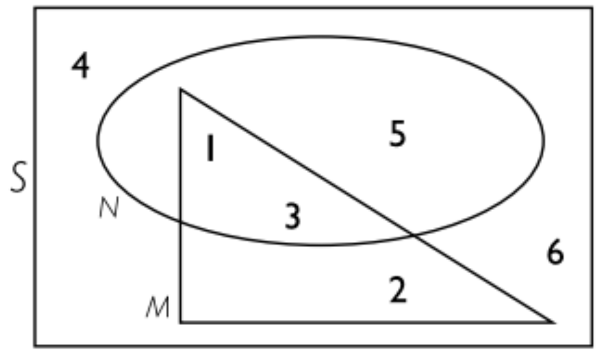
- To each possible outcome in the sample space, we assign a probability P,
  - which represents how certain we are about the occurrence of the corresponding outcome.
    - For an outcome o, we denote the probability as $P(o)$,
      - where $0 \leq P(o) \leq 1$.
- The total probability of all outcomes in the sample space is always 1.
  - Coin tossing : $P(H) + P(T) = 1$
  - Die rolling : $P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1$
- Therefore, if the outcomes are equally probable,
  - the probability of each outcome is $1/n_S$,
    - where n_S is the number of possible outcomes.

Random events

- An event is a subset of the sample space S.
  - A possible event for die rolling is
    - $E = \{1,3,5\}$.
      - This is the event of rolling an odd number.
  - For the genotype example,
    - $E = \{AA, aa\}$
      - This is the event that a person is homozygous.
- An event occurs when any outcome within that event occurs.
- We denote the probability of event E as $P(E)$.
- The probability of an event is the sum of the probabilities for all individual outcomes included in that event.

Random events – Example 1

- Consider the die rolling example presented in the form of a Venn diagram below.

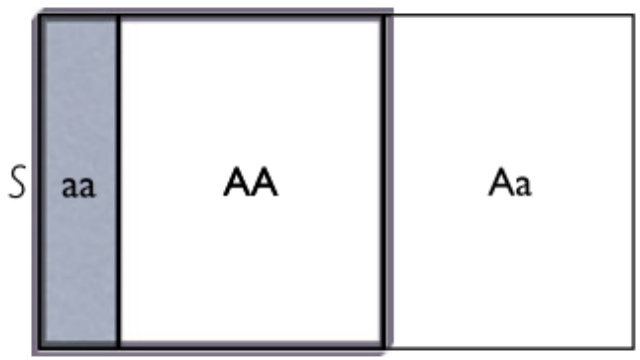


- All the possible outcomes are contained inside the sample space *S*, which is represented by the rectangle.

- We define two events.
  - The event *M* (shown as a triangle) occurs when the outcome is less than 4.
  - The event *N* (shown as an oval) occurs when the outcome is an odd number.
- In this example, $P(M) = 1/2$ and $P(N) = 1/2$

Random events – Example 2

- As a running example, we consider a bi-allelic gene **A** with two alleles *A* and *a*.
- We assume that allele *a* is recessive and causes a specific disease.
  - Then only people with the genotype *aa* have the disease.
    - A schematic representation for a bi-allelic gene with a recessive allele *a* that causes a specific disease.



- The *shaded area* shows the disease event (*D*).
- The *unshaded area* shows the no-disease event (*ND*).
- The *area with shaded border lines* shows the homozygous event (*HM*).
- The *remaining part* of the sample space, which includes the outcome *Aa* only, corresponds to the heterozygous event

Random events - Example

- We can define four events as follows:
  - The homozygous event : $HM = \{AA, aa\}$;
  - The heterozygous event : $HT = \{Aa\}$;
  - The no-disease event : $ND = \{AA, Aa\}$;
  - The disease event : $D = \{aa\}$:
- Assume that the probabilities for different genotypes are
  - $P(AA) = 0.49$, $P(Aa) = 0.42$, and $P(aa) = 0.09$.
- Then,
  - $P(HM) = 0.49 + 0.09 = 0.58$;
  - $P(HT) = 0.42$;
  - $P(ND) = 0.49 + 0.42 = 0.91$;
  - $P(D) = 0.09$.

Complement

- For any event E, we define its complement, E^c, as the set of all outcomes that are in the sample space S but not in E.
  - For the gene-disease example, the complement of the homozygous event $HM = \{AA, aa\}$ is the heterozygous event $\{Aa\}$;
    - we show this as $HM^c = HT$.
  - Likewise, the complement of the disease event, $D = \{aa\}$, is the no-disease event, $ND = \{AA, Aa\}$;
    - we show this as $D^c = ND$.
- The probability of the complement event is
  - 1 minus the probability of the event:

$$P(E^c) = 1 - P(E)$$

Complement - example

- For the event that the outcome is an odd number, we have
  - $P(N^c) = 1 - P(N) = 1 - (1/2) = 1/2$
    - equal to the probability that the outcome is an even number.
- In the gene disease example, the probability of the complement of the homozygous event is
  - $P(HM^c) = 1 - P(HM) = 1 - 0.58 = 0.42.$
    - equal to the probability of the heterozygous event $P(HT) = 0.42$.
- Likewise, the probability of the complement of the disease event is
  - $P(D^c) = 1 - P(D) = 1 - 0.09 = 0.91$
    - equal to the probability of the no-disease event, $P(ND) = 0.91$.

Complement

- The odds of an event shows how much more certain we are that the event occurs than we are that it does not occur.
- For event E, we calculate the odds as follows:

$$\frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

- For the gene-disease example, the odds for ND (i.e., not having the disease) are

$$\frac{P(ND)}{P(ND^c)} = \frac{P(ND)}{1 - P(ND)} = \frac{0.91}{1 - 0.91} = 10.11$$

- Therefore, it is almost 10 times more likely that a person is not affected by the disease than it is for having the disease.
  - In this case, we say that the odds for not having the disease are 10 to 1.