Lecture: Numerical measures + Basic probability

Hamid Raza
Associate Professor in Economics
raza@business.aau.dk

Aalborg University

Statistics - Statistik

Outline

Measures of Central Tendency

- One of the basic questions asked by researchers is that whether the data tend to center or cluster around some value
- We present numerical measures the mean, median, and mode:

Mean

The arithmetic mean (or simply mean) of a sample, is the sum of the data values divided by the number of observations:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

If the data set is the entire population of data, then the population mean, μ, is a parameter given by:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Measures of Central Tendency

Median

- The median is the middle observation of a set of observations that are arranged in increasing (or decreasing) order
- If the sample size, n, is an odd number, the median is the middle observation
- If the sample size, n, is an even number, the median is the average of the two middle observations

Mode

- The mode, if one exists, is the most frequently occurring value
- A distribution with one mode is called unimodal; with two modes, it is called bimodal; and with more than two modes, the distribution is said to be multimodal

Calculating Percentiles and Quartiles

- **Percentiles and quartiles** are measures that indicate the location, or position, of a value relative to the entire set of data
- **Example:** you are told that you scored in the 92nd percentile on your statistics exam. This means that approximately 92% of the students who took this exam scored lower than you and approximately 8% of the students who took this exam scored higher than you

- To find percentiles and quartiles, data must first be arranged in order from the smallest to the largest values
- Percentiles separate large ordered data sets into 100ths
- The 50th percentile is the median
- Quartiles are descriptive measures that separate large data sets into four quarters
- **1st Quartile** is the 25th percentile; **2nd Quartile** is the median (or the 50th percentile), and **3rd Quartile** is 75th percentile

Calculating Percentiles and Quartiles

Example:

- Let's say we've data for Elgiganten sales of iphones on black friday in 12 hours:

```
sales = c(85, 84, 80, 82, 60, 63, 65, 67, 70, 72, 75, 75)
```

We can re-arrange the data set:

```
sales = c(60, 63, 65, 67, 70, 72, 75, 75, 80, 82, 84, 85)
```

- **To calculate 1st Quartile (25th percentile):**

$$Q1 = \text{value located in } 0.25(n+1)th \text{ position}$$

$$Q1 = 0.25(12+1) = 3.25th \text{ value}$$

The value located on 3rd position is 65. 1st Quartile can be calculated as follows:

$$Q1 = 65 + 0.25(67 - 65) = 65.5 \text{ (iphones)}$$

- **To calculate 2nd Quartile (50th percentile) or median:**

$$Q2(median) = 0.5(12+1) = 6.5th \text{ value}$$

The value located on 6th position is 72. Q2 can be calculated as follows:

$$Q2 = 72 + 0.5(75 - 72) = 73.5 \text{ (iphones)}$$

Q2 can also be simply calculated: $\frac{72+75}{2}$

Calculating Percentiles and Quartiles (cont)

Example:

- **To calculate 3rd Quartile (75th percentile):**

$$Q3 = 0.75(12 + 1) = 9.75th \text{ value}$$

The value located on 9.75th position is 80. Q3 can be calculated as follows:

$$Q3 = 80 + 0.75(82 - 80) = 81.5 \text{ (iphones)}$$

Calculating Percentiles and Quartiles in R:

- R has 9 different ways of calculating quartiles. The one we discussed above can simply be calculated by typing:

```
quantile(sales, type = 6)

##   0%  25%  50%  75% 100%
## 60.0 65.5 73.5 81.5 85.0
```

- The above output contains the **5 number summary**
- **To calculate 44th percentile:**

```
quantile(sales, type = 6, prob = c(0.44))

##   44%
## 71.44
```

- Read more about different quartiles by typing ?quantile in your R

Measures of Variability

Range

Range is the difference between the largest and smallest observations

Interquartile Range

- The interquartile range (IQR) measures the spread in the middle 50% of the data
- IQR is the difference between the observation at Q3, the third quartile (or 75th percentile), and the observation at Q1, the first quartile (or 25th percentile). Thus,

$$IQR = Q3 - Q1$$

Box-and-Whisker Plot

- A box-and-whisker plot is a graph that describes the shape of a distribution in terms of the five-number summary: the minimum value, first quartile (25th percentile), the median, the third quartile (75th percentile), and the maximum value.

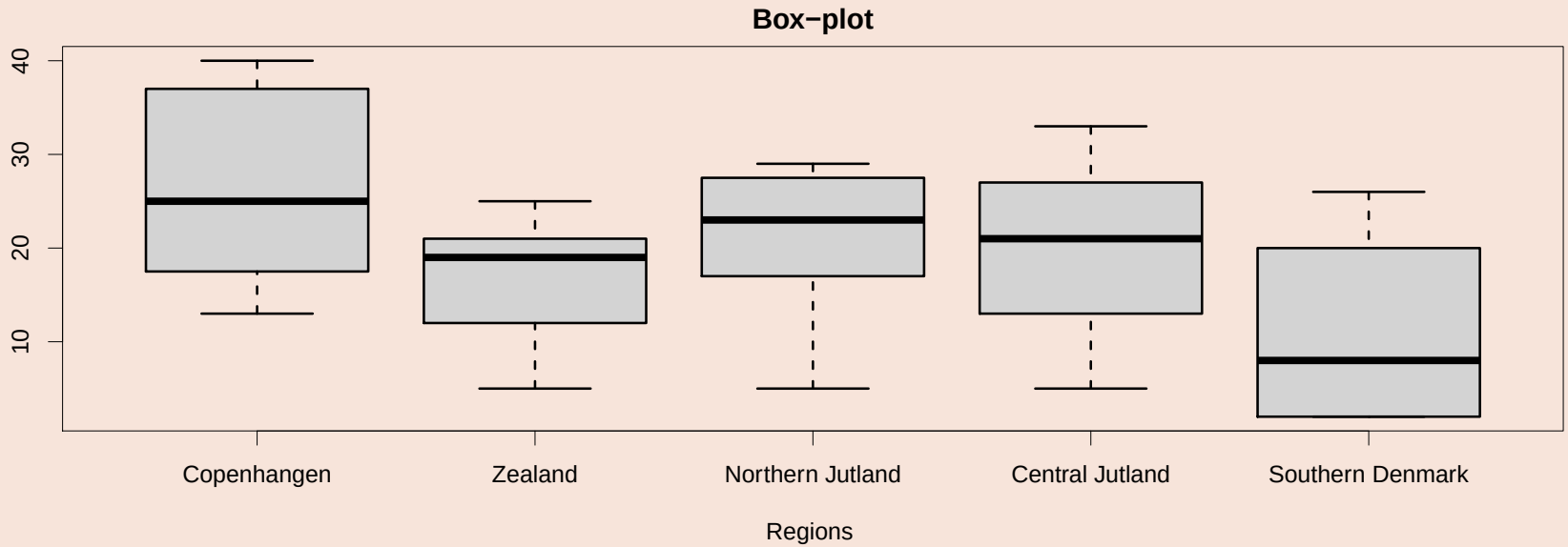
Box-and-Whisker Plot (Cont)

Example:

We look at sales of Fakta in 5 regions of Denmark over the period of 15 days:

```
set.seed(123)
copenhagen = sample(seq(from = 10, to = 40, by = 3), size = 15, replace = T)
zealand = sample(seq(from = 5, to = 25, by = 2), size = 15, replace = T)
n_jutland = sample(seq(from = 5, to = 30, by = 3), size = 15, replace = T)
central_jutland = sample(seq(from = 5, to = 35, by = 4), size = 15, replace = T)
southern_denmark = sample(seq(from = 2, to = 28, by = 6), size = 15, replace = T)
data = data.frame(copenhagen, zealand, n_jutland, central_jutland, southern_denmark)

boxplot(data, lwd=10, main="Box-plot", xlab="Regions ", ylab="Sales",
names=c("Copenhangen","Zealand","Northern Jutland","Central Jutland","Southern Denmark"))
```

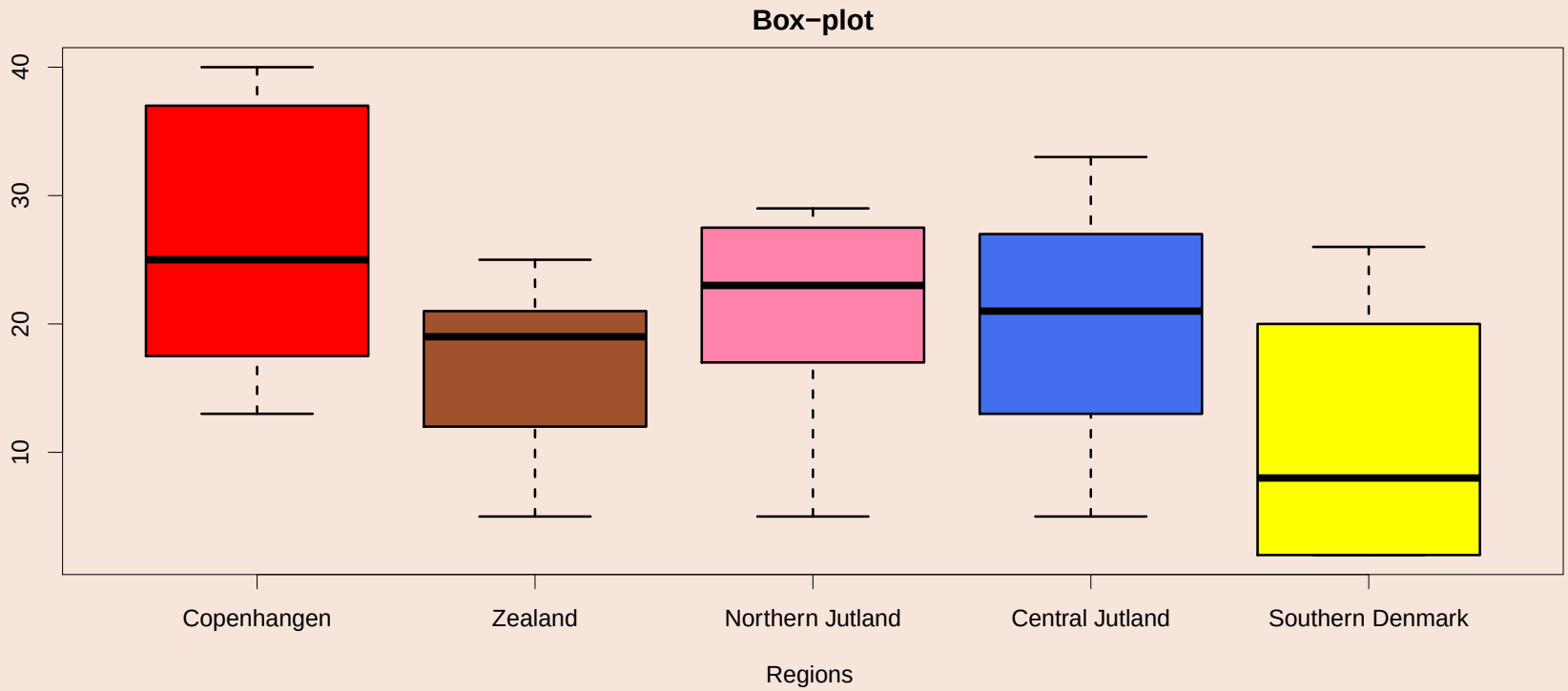


Box-and-Whisker Plot (Cont)

Example:

We can create a more fancy plots by adding colors:

```
boxplot(data, lwd=10, main="Box-plot", xlab="Regions ", ylab="Sales",
names=c("Copenhangen","Zealand","Northern Jutland","Central Jutland","Southern Denmark"),
col = c("red","sienna","palevioletred1","royalblue2","yellow"))
```



Box-and-Whisker Plot (Cont)

Description of graphs

- The inner box shows the numbers that span the range from the first to the third quartile
- A line is drawn through the box at the median. There are two "whiskers"
- One whisker is the line from the 25th percentile to the minimum value; the other whisker is the line from the 75th percentile to the maximum value.
- **TASK: Replicate the above figures and also calculate 1st quartile, second quartile (median), and third quartile of the regions which I have plotted**

Measures of Variability (Cont)

Variance

- The sample variance, s^2, is the sum of the squared differences between each observation and the sample mean divided by the sample size, n, minus 1:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- The population variance, σ^2, is the sum of the squared differences between each observation and the population mean divided by the population size, N:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

Standard deviation

- The sample standard deviation, s, is:

$$s = \sqrt{s^2}$$

- The population standard deviation, σ is:

$$\sigma = \sqrt{\sigma^2}$$

Measures of Variability (Cont)

Example

- Assume you are considering investment in a financial asset. You are presented with two assets as follows:

	Asset A	Asset B
Mean return	12.2%	12.2%
Standard deviation of rate of return	0.63	3

- Which of the above assets will you consider for investing and why?

Measures of Variability (Cont)

- Sometimes, mean and standard deviations of different variables are not directly comparable, e.g., sales of fakta as compared to sales of a small shop in town.
- It is natural that both the standard deviation and mean of sales for fakta would be much larger. In this case, we cannot simply conclude that the sales of fakta are more volatile because of a higher standard deviation

Coefficient of variation

- We use **coefficient of variance (CV)**, which is a measure of relative dispersion that expresses the standard deviation as a percentage of the mean (provided the mean is positive).
- The population coefficient of variation is:

$$CV = \frac{\sigma}{\mu}$$

- The sample coefficient of variation is:

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

Lower value of CV means the standard deviation relative to the mean is lower (which means less fluctuating)

Exercise

Work in pairs (with the person next to you!)

- Assume two variables x and y in R:

```
set.seed(213)
x = sample(seq(from = 10, to = 1000, by = 3), size = 1241, replace = T)
y = sample(seq(from = 10, to = 100, by = 3), size = 1241, replace = T)
```

- What is the no. of observarion n
- Show boxer and whisker plot of the two series. What do you see?
- What is the mean of x and y, compare your results with mean(x) in R
- Calculate the variance of x and y, compare you results with var(x) in R
- Calculate standard deviation of x and y, compare your results with sd(x) in R
- Finally, calculate coefficient of variance for x and y
- Assume x and y were return on assets. Which asset would you invest in?

Outline

Probability

Basic concepts

Probability begins with the concept of a random experiment that can have two or more outcomes, but we do not know which will occur next.

- **Random experiment:** A random experiment is a process leading to two or more possible outcomes, without knowing exactly which outcome will occur, e.g.,
  1. A coin is tossed and the outcome is either a head or a tail
  2. A customer enters a store and either purchases a shirt or does not
  3. The daily change in an index of stock market prices is observed
- **Sample space:** The possible outcomes from a random experiment are called the basic outcomes, and the set of all basic outcomes is called the sample space (S), e.g., $S = \{1, 2, 3, 4, 5, 6\}$ for a die experiment.
- **Event:** An event (E) is any subset of basic outcomes from the sample space, e.g., getting odd numbers with a die experiment, $E = \{1, 3, 5\}$ is an Event.

Probability

Basic concepts (cont)

- **Union of events:** Let A and B be two events in the sample space, S. Their union, denoted by $A \cup B$, is the set of all basic outcomes in S that belong to at least one of these two events.
  - Lets say event A is odd numbers in a die experiment, $A = \{1, 3, 5\}$ and event B is even numbers, $B = \{2, 4, 6\}$. The union of A and B is $A \cup B = \{1, 3, 5, 2, 4, 6\}$

```
A=c(1,3,5);    B=c(2,4,6)
union(A,B)
## [1] 1 3 5 2 4 6
```

- **Intersection of events:** Let A and B be two events in the sample space S. Their intersection, denoted by $A \cap B$, is the set of all basic outcomes in S that belong to both A and B.
  - Lets say event A is, $A = \{1, 2, 3, 5, 6\}$ and event B, $B = \{2, 4, 6, 8, 9\}$. The intersection of A and B is $A \cap B = \{2, 6\}$

```
A=c(1,2,3,5,6);    B=c(2,4,6,8,9)
intersect(A,B)
## [1] 2 6
```

Probability

Basic concepts (cont)

- **Mutually exclusive (or disjoint):** If events A and B have no common basic outcomes, they are called mutually exclusive, and their intersection, $A \cap B$, is an empty set
  - Lets say event A is odd numbers in a die experiment, $A = \{1, 3, 5\}$ and event B is even numbers, $B = \{2, 4, 6\}$.
  - Event A and B are mutually exclusive because the intersection of A and B is, $A \cap B = \{0\}$
- **Collectively exhaustive:** Two events A and B are collectively exhaustive if their union is the entire sample space S.
  - Lets say event A is odd numbers in a die experiment, $A = \{1, 3, 5\}$ and event B is even numbers, $B = \{2, 4, 6\}$.
  - Event A and B are collectively exhaustive because the union of A and B is, $A \cup B = \{1, 2, 3, 4, 5, 6\}$, which is equal to the sample space (S), containing all basic outcomes of a die experiment

Probability

Basic concepts (cont)

- **Complement:** Let A be an event in the sample space, S. The set of basic outcomes of a random experiment belonging to S but not to A is called the complement of A and is denoted by $\bar{A}$
  - The sample space, S, in a die experiment is S=$\{1,2,3,4,5,6\}$. Lets say event A is odd numbers, $A = \{1, 3, 5\}$
  - The complement of event A is, $\bar{A} = \{2, 4, 6\}$

Probability rules

- **Complement rule:** In terms of probability, the complement of A is equal to $P(\bar{A}) = 1 - P(A)$
- Question: In the die experiment, $S = \{1, 2, 3, 4, 5, 6\}$. What is P(A) and $P(\bar{A})$, if event $A = \{1, 3\}$?

Probability rules (cont)

- **Addition rules of probabilites:** Let A and B be two events. Using the addition rules of probabilities, the probability of their union is as follows:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- If we have two mutually exclusive (or disjoint) events, then the probability of their union is simply:

$$P(A \cup B) = P(A) + P(B)$$

  - **Example:** A cell phone company found that 75% of all customers want text messaging on their phones, 80% want photo capability, and 65% want both. What is the probability that a customer will want at least one of these?
  - **Solution:** Let A be the event "customer wants text messaging" and B be the event "customer wants photo capability." Thus:

$$P(A) = 0.75 \quad P(B) = 0.80 \quad \text{and} \quad P(A \cap B) = 0.65$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.75 + 0.80 - 0.65 = 0.90$$

Probability

Probability rules (cont)

- **Joint probability:** We use the term joint probability of A and B to denote the probability of the intersection of A and B, i.e., $P(A \cap B)$ is the joint probability of A and B
- **Relative Frequency:** It shows how often something happens relative to total outcomes:
- Example 1: 50 people were surveyed to answer "What are the top 2 reasons you are late to work?". The responses were following:

```
counts <- matrix(c(6, 14, 9, 2, 17, 2), ncol = 1, byrow = T)
rownames(counts) <- c("Weather", "Overslept", "Alarm Failure", "Time Change", "Traffic", "Other")
colnames(counts) <- c("Reason for being late")
counts
```

```
##               Reason for being late
## Weather                           6
## Overslept                        14
## Alarm Failure                     9
## Time Change                       2
## Traffic                          17
## Other                             2
```

Probability

Probability rules (cont)

- **Relative Frequency (cont):** We can calculate the relative frequency of the above example:

```
100 * prop.table(counts, 2)
```

```
##                Reason for being late
## Weather                           12
## Overslept                         28
## Alarm Failure                     18
## Time Change                        4
## Traffic                           34
## Other                              4
```

- Example 2: For example, employees of a company are asked their membership status of a volleyball club in Aalborg as well as their gender:

```
volley1 <- matrix(c(30, 20, 60, 30, 20, 20), ncol = 2, byrow = TRUE)
colnames(volley1) <- c("Male", "Female")
rownames(volley1) <- c("Current", "Former", "Never")
volley1 <- as.table(volley1)
100 * prop.table(volley1, 2)  # This calculates column-wise relative frequencies
```

```
##          Male Female
## Current 27.27  28.57
## Former  54.55  42.86
## Never   18.18  28.57
```

27.27% of the total male employees are current members of the club

Probability

Probability rules (cont)

```
100 * prop.table(volley1, 1)  # This calculates row-wise relative frequencies

##          Male Female
## Current 60.00  40.00
## Former  66.67  33.33
## Never   50.00  50.00
```

Current members of the volley ball club consists of 60% males and 40% females

Discrete distribution:

```
100 * prop.table(volley1)

##          Male Female
## Current 16.67  11.11
## Former  33.33  16.67
## Never   11.11  11.11
```

Here the sample is divided into 6 mutually exclusive events corresponding to volleyball membership and gender, such that the sum of all 6 probabilites equal 1 (or 100%).

Exercise

- Use the following survey data:

```
library(tidyverse)  # Data wrangling and plotting tools
library(stargazer)  # Package made to export regression and summary tables
library(haven)  # And a package for loading different data types fast(er)
#### Load some data that we want to test out Data comes from an experiment
#### run by Bertrand and Mullainathan (2004) on the effect of having a
#### non-black sounding name.
dat <- read_dta("http://masteringmetrics.com/wp-content/uploads/2015/02/bm.dta")
```

- Calculate a contigency table between education and interview call
  - "education" consists of 4 categories (0, not reported; 1, primary school; 2, high school; 3, college; 4, university)
  - interview call consits of two categories (0, no interview call; 1, interview call)
  - Note: when you define labels for each category, make sure that your labels correctly reflect the data, i.e., obs no. 4 in your sample must correspond to university
- Calculate row wise probabilities
- Calculate column wise probabilities
- Calculate discrete distribution.

Probability

Conditional probability:

Consider a pair of events, A and B. Suppose that we are concerned about the probability of A, given that B has occurred denoted by P(A|B). This problem can be approached using the concept of conditional probability.

- Let A and B be two events. The conditional probability of event A, given that event B has occurred, is found to be as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{provided that } P(B) > 0$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad \text{provided that } P(A) > 0$$

Conditional probability (cont):

- **Example:** Consider that 75% of the customers of mobile company want text messaging, 80% want photo capability, and 65% want both. What are the probabilities that a person who wants text messaging also wants photo capability and that a person who wants photo capability also wants text messaging?
- **Solution:** Event A is text messaging and B is photo capability. We know that $P(A) = 0.75$, $P(B) = 0.80$, and $P(A \cap B) = 0.65$.

  The probability that a person who wants photo capability also wants text messaging is the conditionl probability of event A, given event B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.65}{0.80} = 0.8125$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.65}{0.75} = 0.8667$$

Probability

Statistical independence:

- Statistical independence is a special case for which the conditional probability of A, given B, is the same as the unconditional probability of A — that is, $P(A|B) = P(A)$
- We see that knowing that event B has occurred does not change the probability of event A.

Exercise

- A company gathers data on the gender of its employees and their volleyball membership status

```
100 * prop.table(volley1)

##           Male Female
## Current 16.67  11.11
## Former  33.33  16.67
## Never   11.11  11.11
```

The above information represents probabilities. e.g., 16.7 is the joint probability of male and current membership, i.e., $P(male \cap current) = 0.16$, which means that 16.7% of the employees are male as well as current members of the club.

What is the probability that an employee is a male or current member of the volleyball club, i.e., find $P(male \cup current)$?

What is the probability that an employee is a female or a former member of the volleyball club, i.e., find $P(female \cup former)$?

Hint: First, find probabilites of all events separately.

Exercise

Question:

- **What is the probability that an employee who is a male is also a former club member?**

- **What is the probability that an employee who is a female is also a current club member?**

Hint: Conditional probability

Outline

Bayes' Theorem:

- Let A1 and B1 be two events. Then Bayes' theorem states that:

$$P(A_1|B_1) = \frac{P(B_1|A_1).P(A_1)}{P(B_1)}$$

Sometimes the denominator $P(B_1)$ is not known, the Bayes' theorem can also be written as:

$$P(A_1|B_1) = \frac{P(B_1|A_1).P(A_1)}{P(B_1|A_1).P(A_1) + P(B_1|A_2).P(A_2) + ... + P(B_1|A_k).P(A_k)}$$

$P(A_2)$ is the complement of $P(A_1)$. In other words, $P(A_2)$ simply means something not included in $P(A)$, e.g., if we say $P(A_1)$ is the probability that 80% of the people at AAU are students. Then $P(A_2)$ in this case will be 100 - 80 = 40% who are not students at AAU.

Bayes' Theorem (cont):

- **Example:** Based on an examination of past records of a corporation's account balances, an auditor finds that 15% have contained errors. Of those balances in error, 60% were regarded as unusual values based on historical figures. Of all the account balances, 20% were unusual values. If the figure for a particular balance appears unusual on this basis, what is the probability that it is in error?
- **Solution:** Let A_1 be "error in account balance" and B_1 be "unusual value based on historical figures." Then, from the available information:

$$P(A_1) = 0.15 \quad P(B_1) = 0.20 \quad P(B_1|A_1) = 0.60$$

Using Bayes' theorem:

$$P(A_1|B_1) = \frac{P(B_1|A_1).P(A_1)}{P(B_1)} = \frac{(0.60)(0.15)}{0.20} = 0.45$$

Outline

Exercise

Question: 1

- Given $P(A_1) = 0.4$, $P(B_1|A_1) = 0.6$, and $P(B_1|A_2) = 0.7$, what is the probability of $P(A_1|B_1)$?
  Hint: calculate the complements of the given probabilites

- Given $P(A_1) = 0.4$, $P(B_1|A_1) = 0.6$, and $P(B_1|A_2) = 0.7$, what is the probability of $P(A_2|B_2)$?

Exercise

Question: 2

A publisher sends advertising materials for an accounting text to 80% of all professors teaching the appropriate accounting course. 30% of the professors who received this material adopted the book, as did 10% of the professors who did not receive the material. What is the probability that a professor who adopts the book has received the advertising material?