MTH113TC
Introduction to Probability and Statistics
(Statistics Component)

2020/21 Academic Year

Semester 1

Jionglong Su

2020-09-20

Xi’an Jiaotong-Liverpool University

西交利物浦大学

MTH113TC: Introduction to Probability and Statistics

- Lecturer: Dr Jionglong Su (苏炯龙)
  - Email: Jionglong.Su@xjtlu.edu.cn
  - Phone Number: 8188 4848
  - Office: SC527
- Office Hour
  - Wed 7-9 pm
- Teaching Materials
  - Most will be uploaded to ICE in due time, so please check the following website regularly:
  - https://ice.xjtlu.edu.cn
  - ***N.B. This file itself will be updated frequently but with the same file name; please check the date on the first page to ensure you get the latest version.***

Chapter 4: Introduction and Review

- Reading task: Moore-McCabe-Craig Chapters ***1, 4, 5***

4.1.1 Data and its Representation

4.1.2 Four scales of measurement

4.1.3 Describing Distributions

- Shape
- Location
- Spread
- Outliers

4.1.4 Box-and-whisker plot

4.1.5 Matlab Implementation

4.2.1 Random Variable and its Properties

4.2.2 Sampling Distributions

4.2.3 What is the Sampling Distribution of $\bar{X}$?

Chapter 4: Introduction and Review

4.2.4 What is the CLT used for?

4.2.5 Sampling Distribution of the Difference between Two Means, $\bar{X}_1 - \bar{X}_2$

4.3.1 χ^2-Distribution

4.3.2 How does the χ^2-Distribution look like?

4.3.3 What is the χ^2-Distribution Used For?

4.4.1 Student t-Distribution

4.4.2 How does the t-Distribution look like?

4.4.3 What is the t-distribution Used For?

4.5.1 F-Distribution

4.5.2 How does the F-Distribution look like?

4.5.3 What is the F-Distribution Used For?

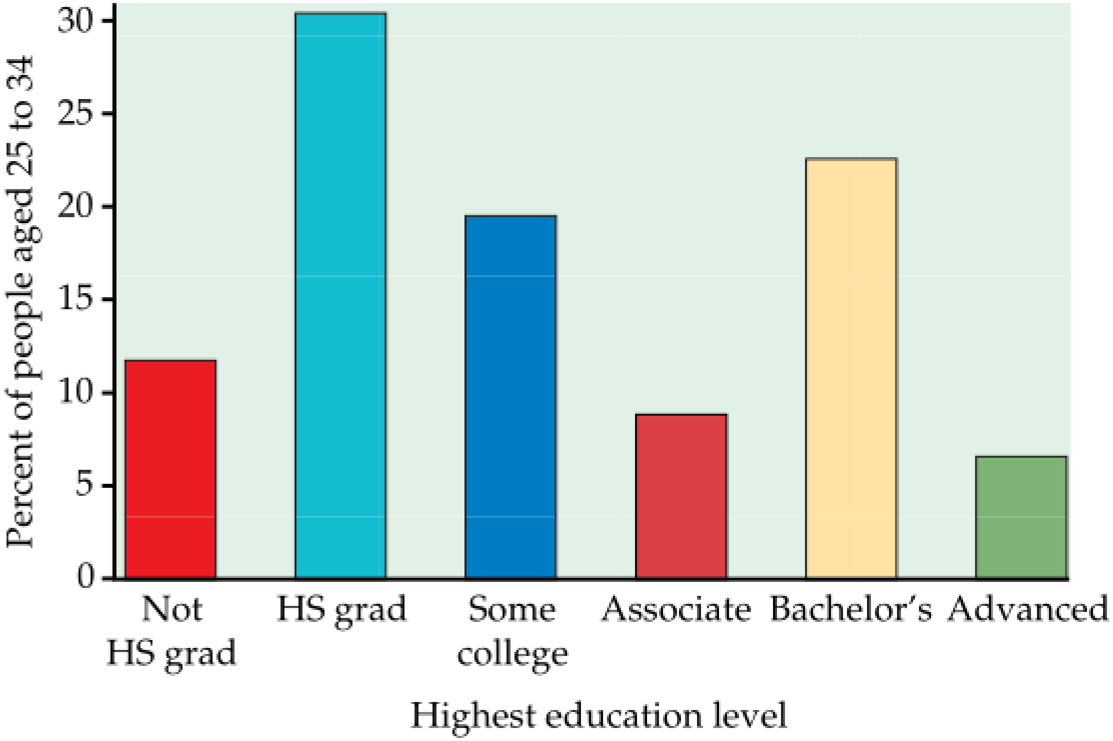
4.6 Matlab Implementation

4.7 Summary

Xi'an Jiaotong-Liverpool University 西交利物浦大學

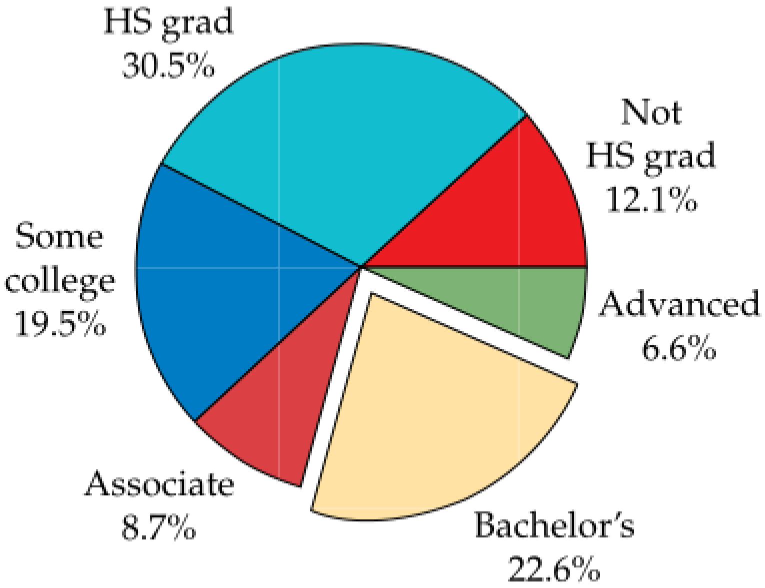
4.1.1 Data and its Representation

Bar graph



4.1.1 Data and its Representation

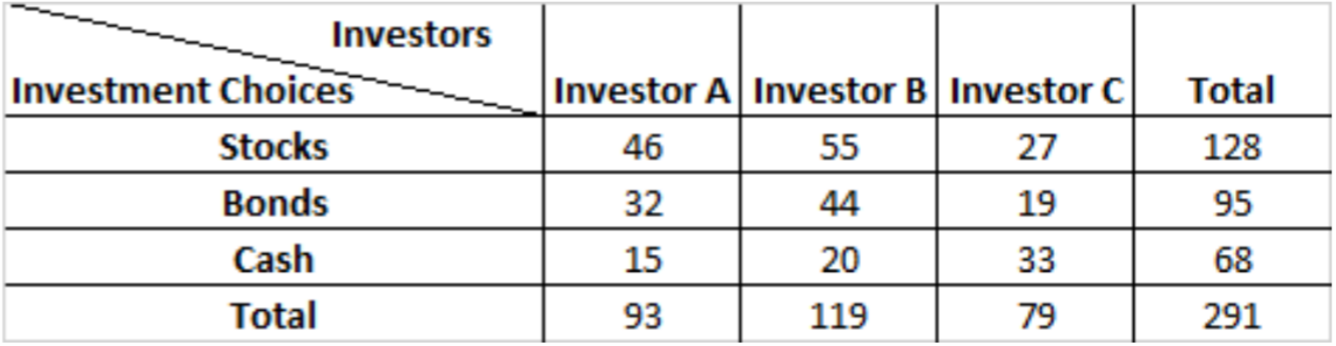
Pie Chart



4.1.1 Data and its Representation

Cross Table

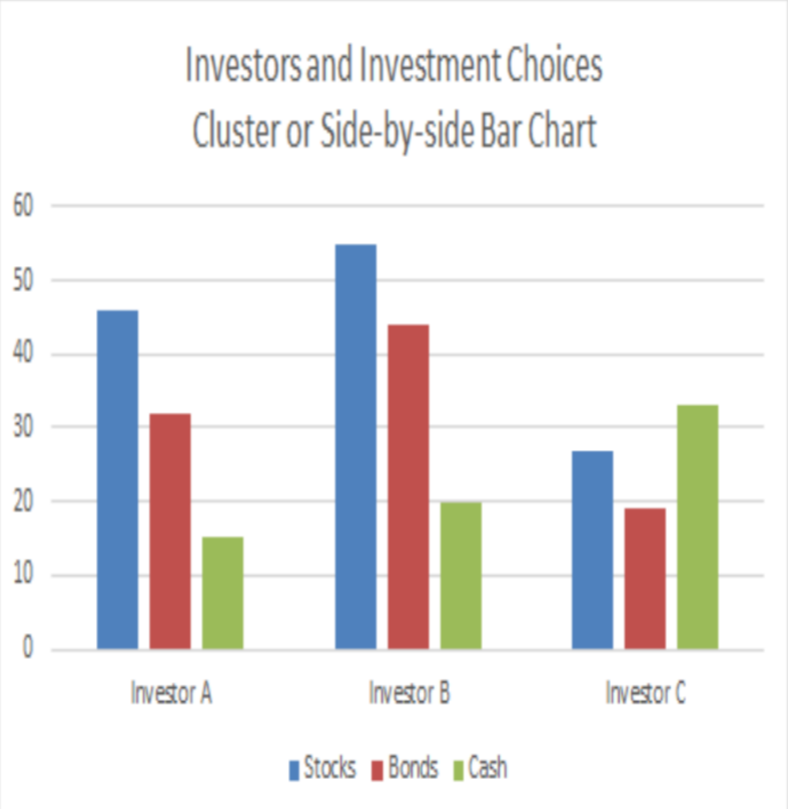
A ***cross table*** lists the number of observations for every combination of values for **two categorical variables**. If there are *r* categories for the first variable and *c* categories for the second variable (*r* rows and *c* columns), the table is called an *r* × *c* cross table. The following is an example of 3 × 3 cross table

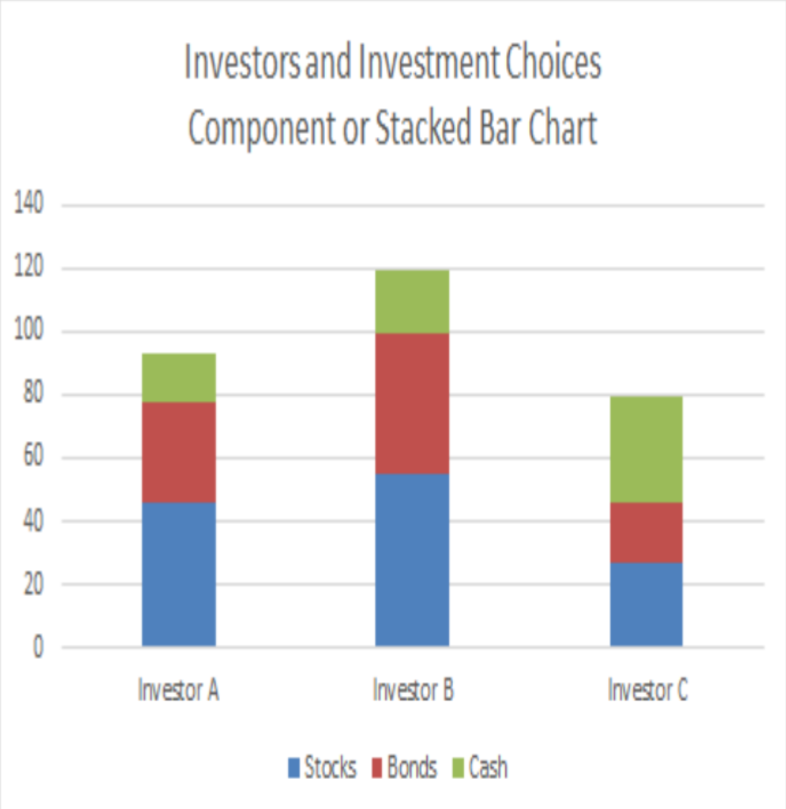


Investors Investment Choices	Investor A	Investor B	Investor C	Total
Stocks	46	55	27	128
Bonds	32	44	19	95
Cash	15	20	33	68
Total	93	119	79	291

4.1.1 Data and its Representation

We can graph this 3 × 3 cross table using either a cluster (side-by-side) or a component (stacked) bar chart.

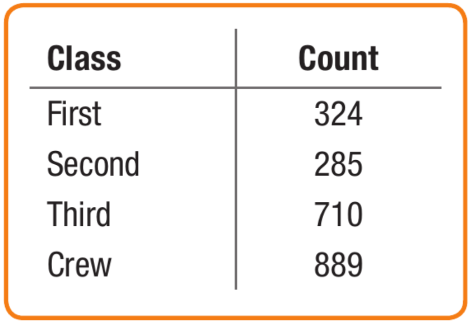




4.1.1 Data and its Representation

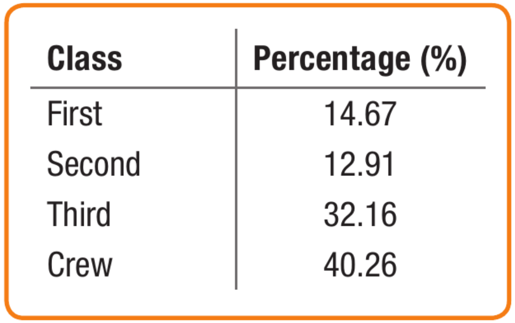
(Relative) Frequency Table for categorial data

A frequency table that lists the number of cases (i.e., how many times it appears) in each category along with its name. A relative frequency table displays percentages (or proportions) rather than the counts in each category.



Class	Count
First	324
Second	285
Third	710
Crew	889

Table 2.2
A frequency table of the *Titanic* passengers.



Class	Percentage (%)
First	14.67
Second	12.91
Third	32.16
Crew	40.26

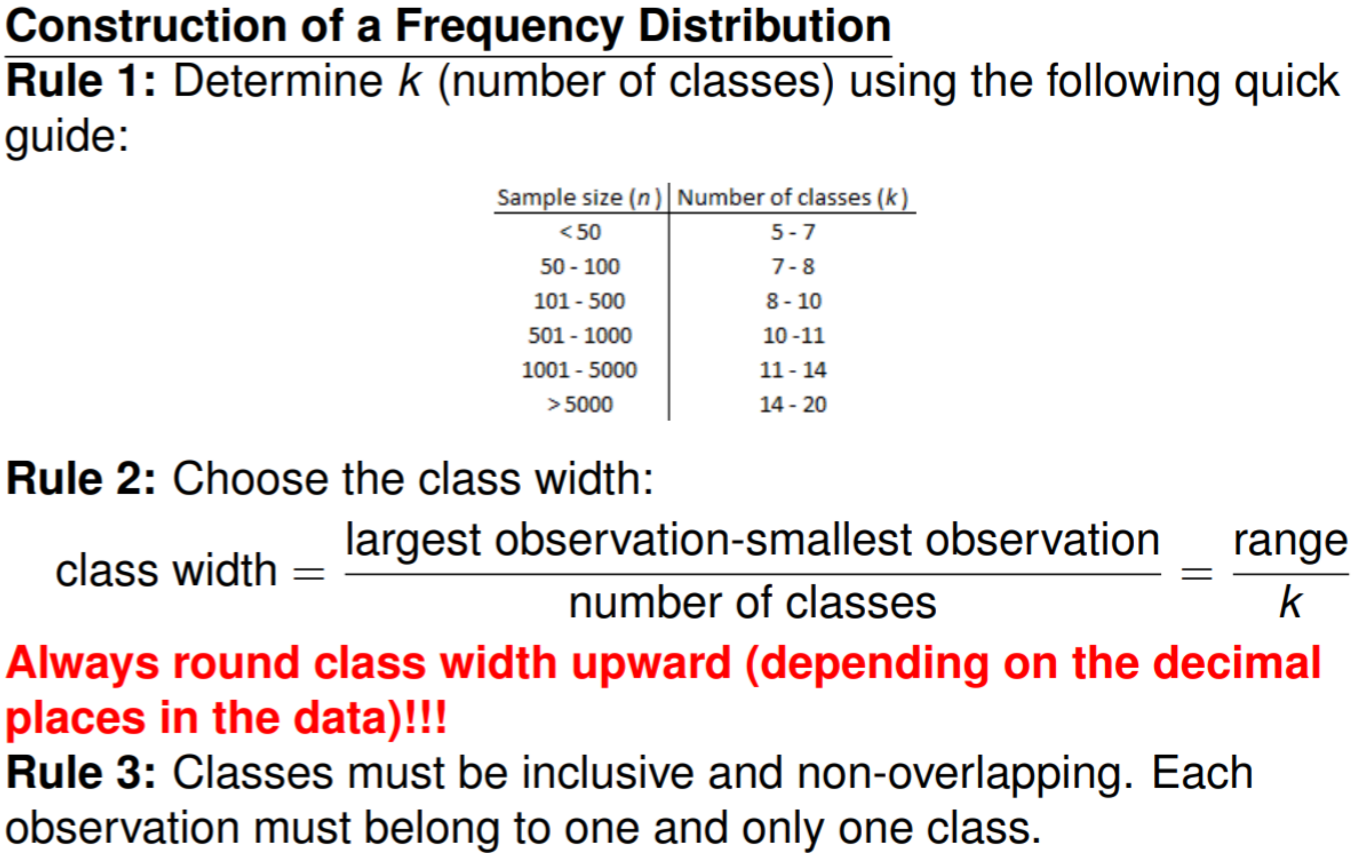
Table 2.3
A relative frequency table for the same data.

Frequency table is sometimes called frequency distribution.

4.1.1 Data and its Representation

(Relative) Frequency Table for numerical data

A frequency table (distribution) for numerical summarizes data by listing the classes in the left column and the frequencies in the right column.



Construction of a Frequency Distribution

Rule 1: Determine *k* (number of classes) using the following quick guide:

Sample size (n)	Number of classes (k)
< 50	5 - 7
50 - 100	7 - 8
101 - 500	8 - 10
501 - 1000	10 -11
1001 - 5000	11 - 14
> 5000	14 - 20

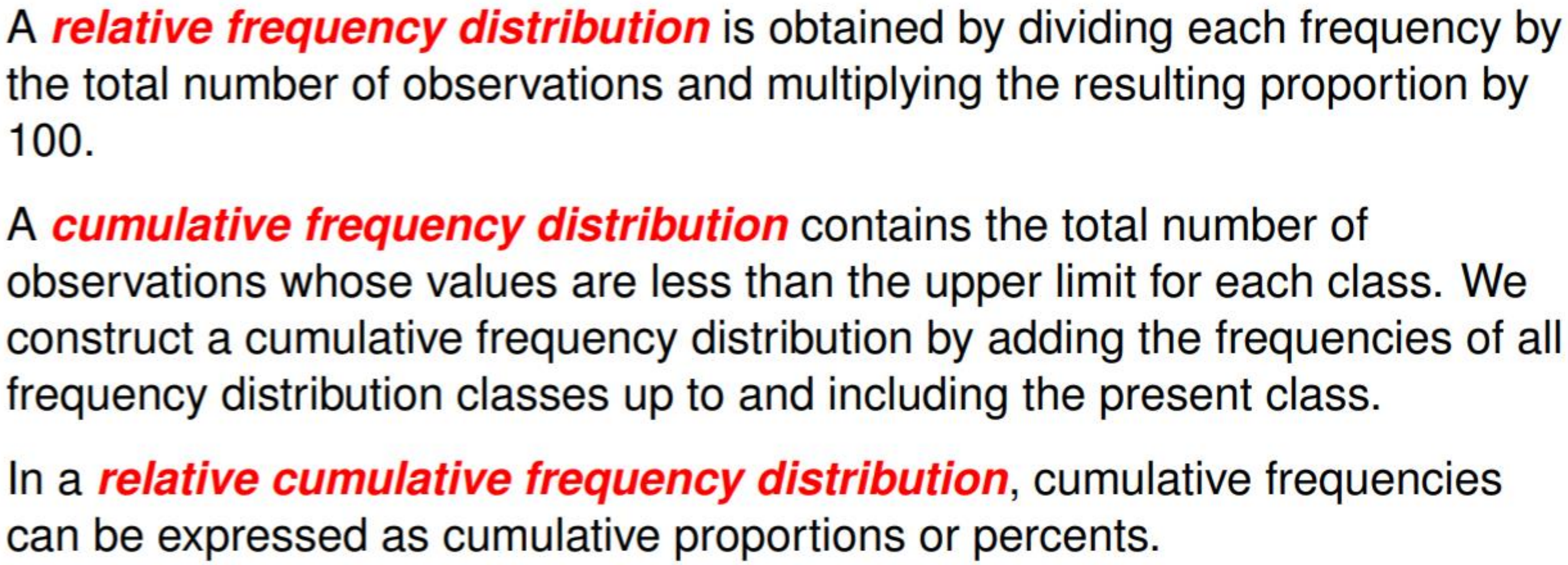
Rule 2: Choose the class width:

$$\text{class width} = \frac{\text{largest observation-smallest observation}}{\text{number of classes}} = \frac{\text{range}}{k}$$

Always round class width upward (depending on the decimal places in the data)!!!

Rule 3: Classes must be inclusive and non-overlapping. Each observation must belong to one and only one class.

4.1.1 Data and its Representation

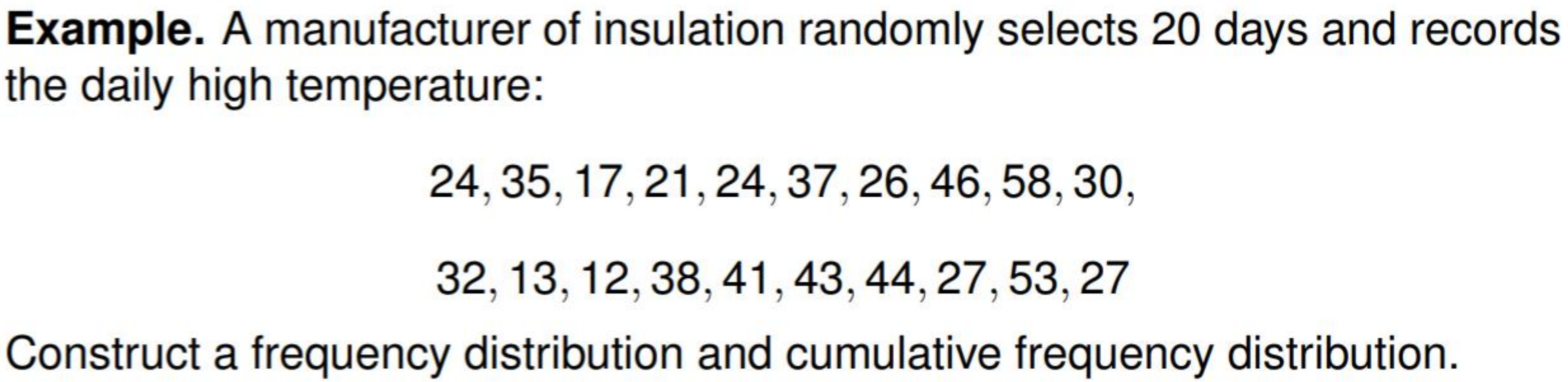


A ***relative frequency distribution*** is obtained by dividing each frequency by the total number of observations and multiplying the resulting proportion by 100.

A ***cumulative frequency distribution*** contains the total number of observations whose values are less than the upper limit for each class. We construct a cumulative frequency distribution by adding the frequencies of all frequency distribution classes up to and including the present class.

In a ***relative cumulative frequency distribution***, cumulative frequencies can be expressed as cumulative proportions or percents.

4.1.1 Data and its Representation



Example. A manufacturer of insulation randomly selects 20 days and records the daily high temperature:

24, 35, 17, 21, 24, 37, 26, 46, 58, 30,

32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Construct a frequency distribution and cumulative frequency distribution.

4.1.1 Data and its Representation

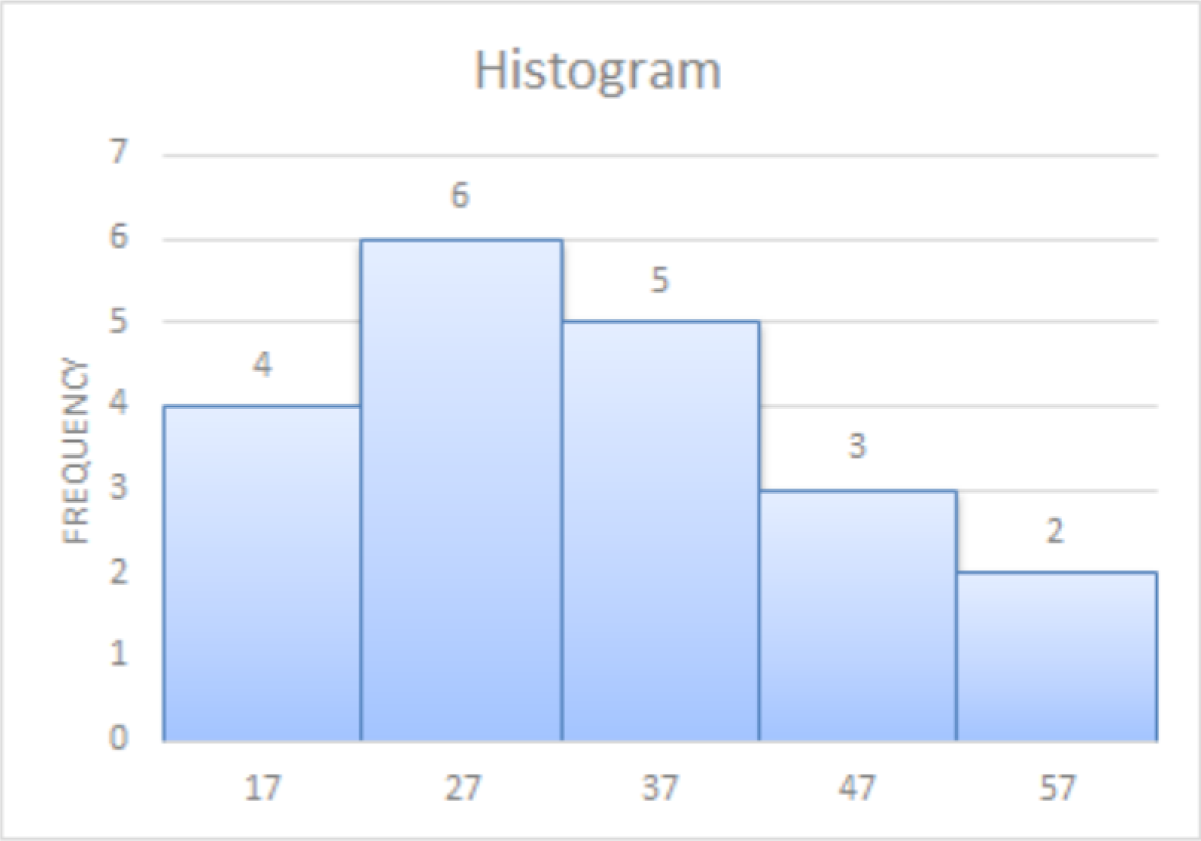
Solution. Since the sample size is $20 < 50$, the number of classes should be between 5 and 7, say 5. We need to sort the data given in an increasing order. Since $\frac{58-12}{5} = 9.2$, we round up this number and have class *width* $= 10$. Hence we have

$$12, 13, 17, 21 | 24, 24, 26, 27, 27, 30 | 32, 35, 37, 38, 41 | 43, 44, 46 | 53, 58$$

Class	Frequency	Relative Frequency(%)	Cumulative Frequency	Relative Cumulative Frequency(%)
$[12, 22)$	4	$20(=\frac{4}{20}100\%)$	4	$20(=\frac{4}{20}100\%)$
$[22, 32)$	6	$30(=\frac{6}{20}100\%)$	10	$50(=\frac{10}{20}100\%)$
$[32, 42)$	5	25	15	75
$[42, 52)$	3	15	18	90
$[52, 62)$	2	10	20	100
total	20	100		

4.1.1 Data and its Representation

Histogram



4.1.1 Data and its Representation

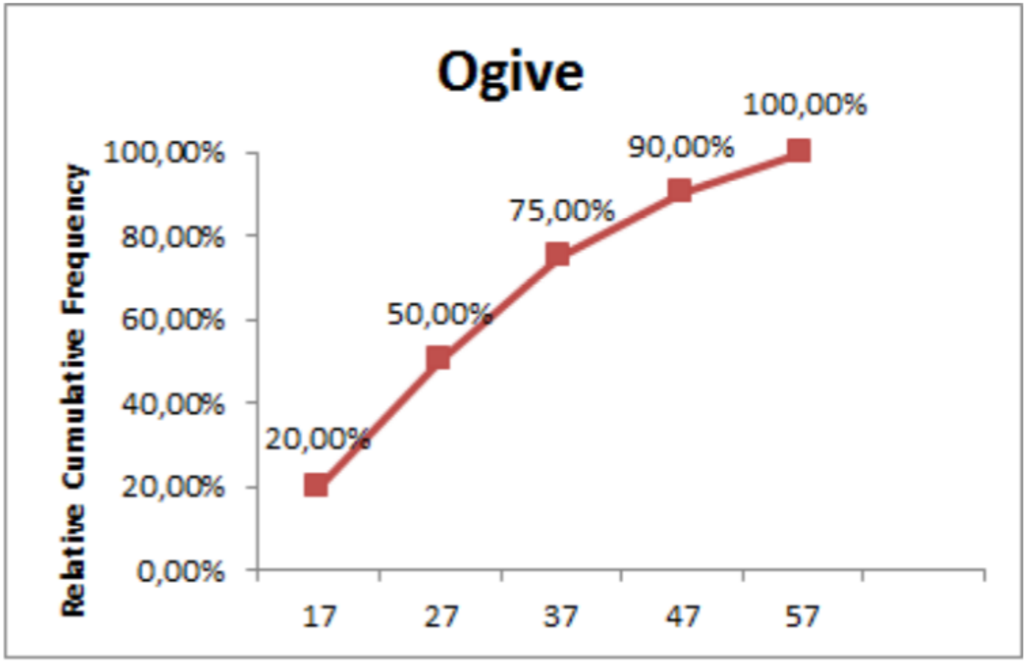
In a histogram

- the intervals correspond to the classes in a frequency distribution,
- the widths of the bars are equal,
- the height of each bar is proportional to the number of observations in that interval,
- the number of observations can be displayed above the bars,
- there are no gaps between the bars.

4.1.1 Data and its Representation

Ogive

An ***ogive*** is a line that connects points that are the cumulative percent of observations below the upper limit of each interval in a cumulative frequency distribution. The ogive related with the previous frequency distribution is



4.1.1 Data and its Representation

Stem and Leaf Plot

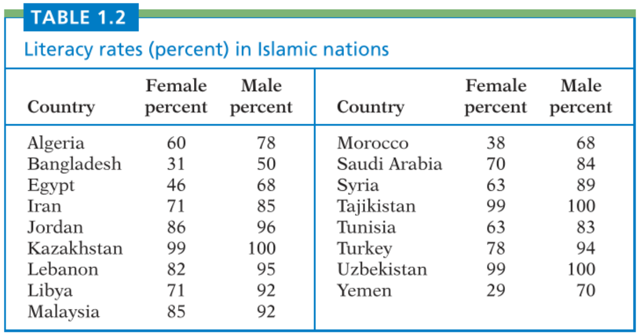
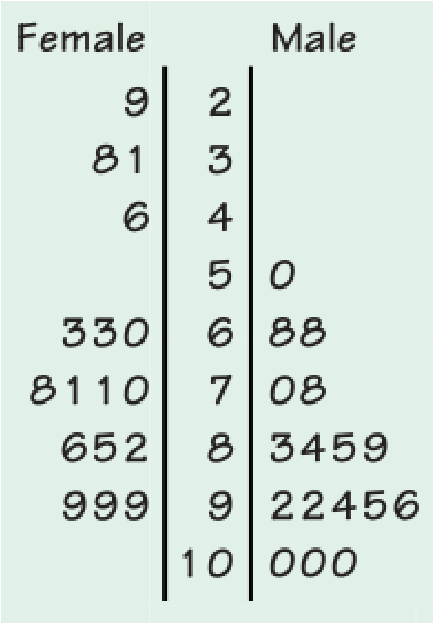


TABLE 1.2

Literacy rates (percent) in Islamic nations

Country	Female percent	Male percent	Country	Female percent	Male percent
Algeria	60	78	Morocco	38	68
Bangladesh	31	50	Saudi Arabia	70	84
Egypt	46	68	Syria	63	89
Iran	71	85	Tajikistan	99	100
Jordan	86	96	Tunisia	63	83
Kazakhstan	99	100	Turkey	78	94
Lebanon	82	95	Uzbekistan	99	100
Libya	71	92	Yemen	29	70
Malaysia	85	92			

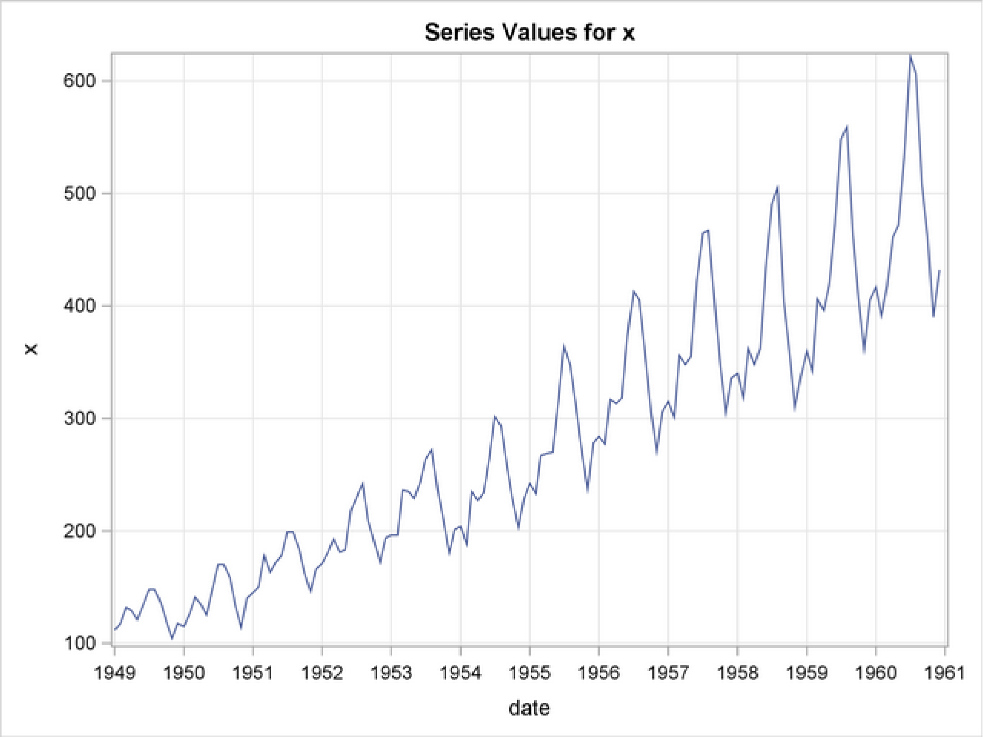


Female		Male
9	2	
81	3	
6	4	
	5	0
330	6	88
8110	7	08
652	8	3459
999	9	22456
	10	000

How do you describe the data?

4.1.1 Data and its Representation

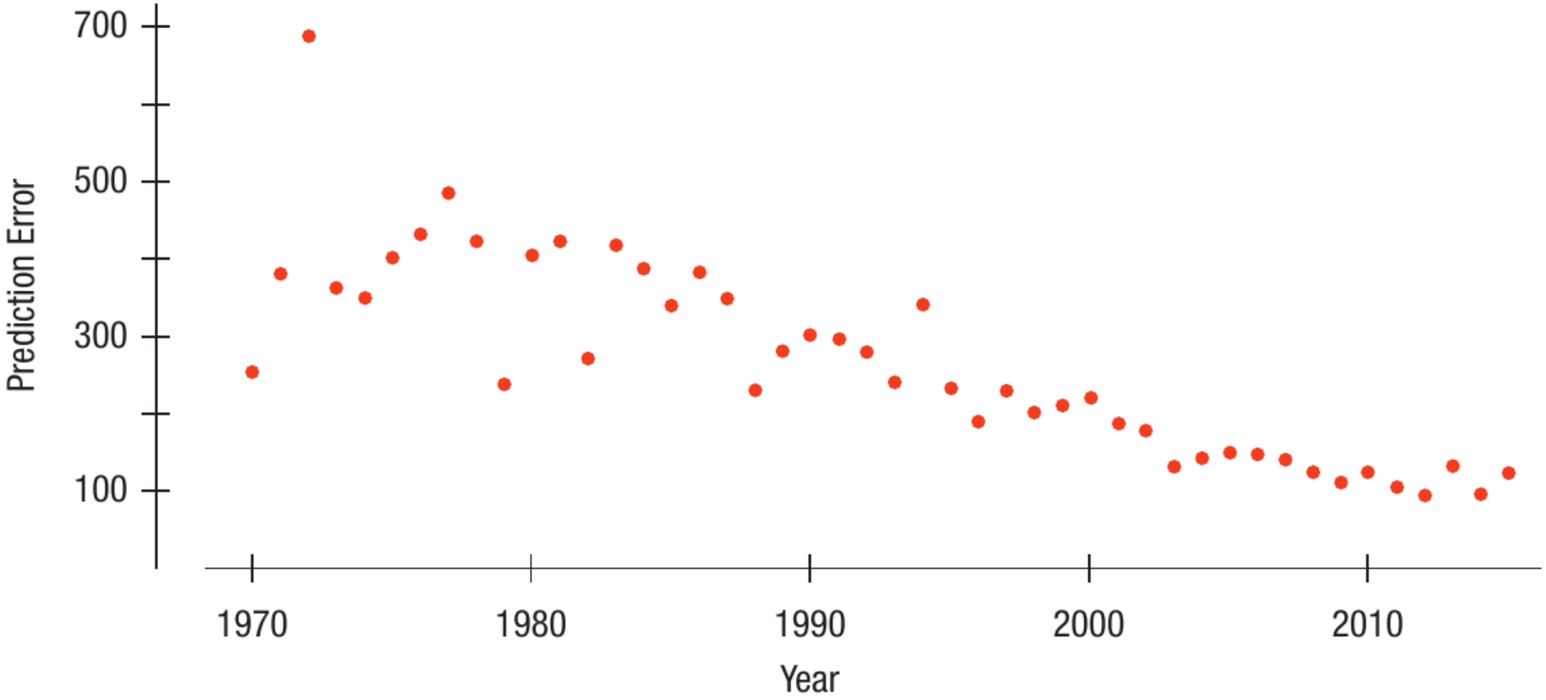
Time Series Plot



How do you describe the data?

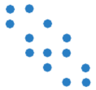
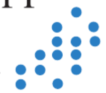
4.1.1 Data and its Representation

Scatterplot

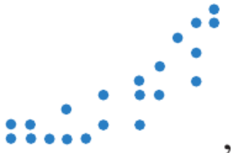
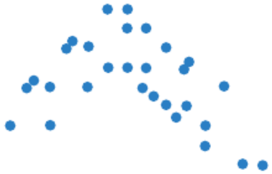


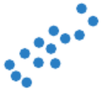
Scatterplots are the best way to start observing the relationship (i.e., association) between two *quantitative* variables.

4.1.1 Data and its Representation

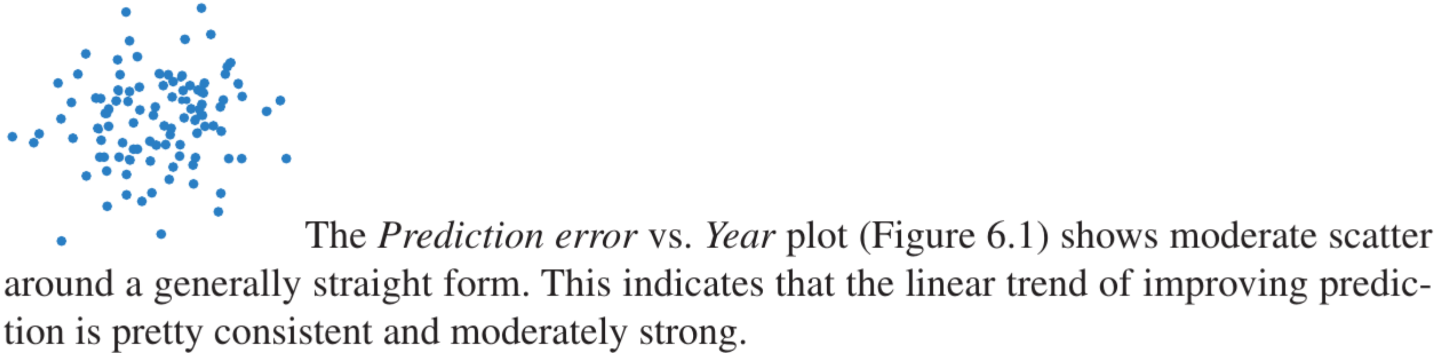
You might say that the **direction** of the association is important. Over time, the NHC's prediction errors have decreased. A pattern like this that runs from the upper left to the lower right  is said to be **negative**. A pattern running the other way  is called **positive**.

The second thing to look for in a scatterplot is its **form**. A plot that appears as a cloud or swarm of points stretched out in a generally consistent, straight form is called linear. For example, the scatterplot of *Prediction Error* vs. *Year* has such an underlying linear form, although some points stray away from it.

If the relationship isn't straight, but curves gently while still increasing or decreasing steadily , we can often find ways to make it more nearly straight. But if it curves sharply up and down, for example like this: , there is much less we can say about it with the methods of this book.

The third feature to look for in a scatterplot is the **strength** of the relationship. At one extreme, do the points appear tightly clustered in a single stream  (whether straight, curved, or bending all over the place)? Or, at the other extreme, does the swarm of points seem to form a vague cloud through which we can barely discern any trend or pattern?

4.1.1 Data and its Representation



The *Prediction error* vs. *Year* plot (Figure 6.1) shows moderate scatter around a generally straight form. This indicates that the linear trend of improving prediction is pretty consistent and moderately strong.

To mathematically measure the linear strength, we look linear correlation.

Definition Let (x_i, y_i) for $i = 1,2,\ldots,n$ be the corresponding values of two quantitative variables then correlation r is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the mean of x and y respectively.

4.1.1 Data and its Representation

Assumptions and conditions for correlation:

1. To use r, there must be a underlying linear relationship between the two variables. That is, you should first identify a linear association possibly through investigating scatterplots, then calculate r to quantify such linear relationship.
2. The variables must be quantitative
3. Outliers can strongly affect the correlation.

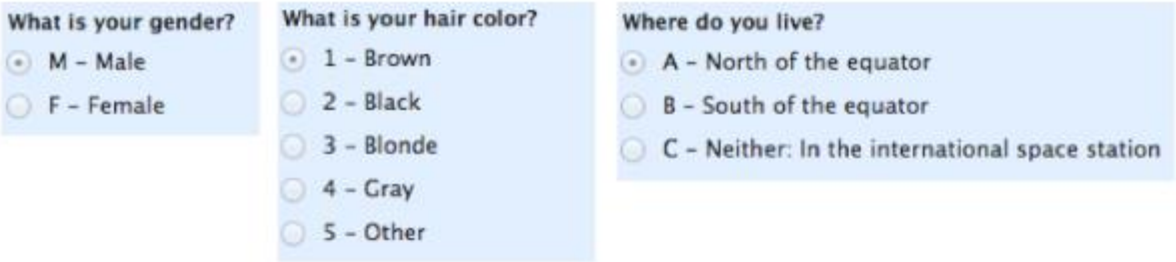
Remarks:

1. Correlation r has no units, so changing the units of x and y does not affect r.
2. Note that $|r| \leq 1$. When $r = \pm 1$ then the points (x_i, y_i) lie exactly on a straight line

4.1.2 Four Scales of Measurements

1) Nominal (Name/Label)

- Eg. Countries in Europe, the number pinned on the T-shirt of sportsperson
- Not quantitative value



What is your gender?

● M – Male
○ F – Female

What is your hair color?

● 1 – Brown
○ 2 – Black
○ 3 – Blonde
○ 4 – Gray
○ 5 – Other

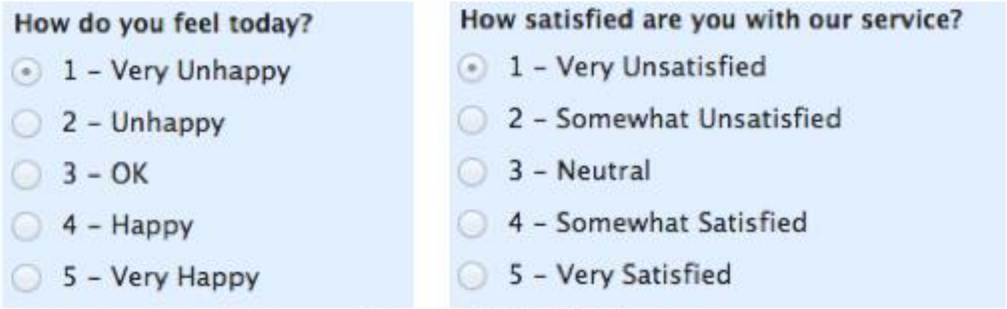
Where do you live?

● A – North of the equator
○ B – South of the equator
○ C – Neither: In the international space station

Examples of Nominal Scales

2) Ordinal (Order)

- E.g. Happiness level on a scale of 1 to 10
- Not quantitative value



How do you feel today?

● 1 – Very Unhappy
○ 2 – Unhappy
○ 3 – OK
○ 4 – Happy
○ 5 – Very Happy

How satisfied are you with our service?

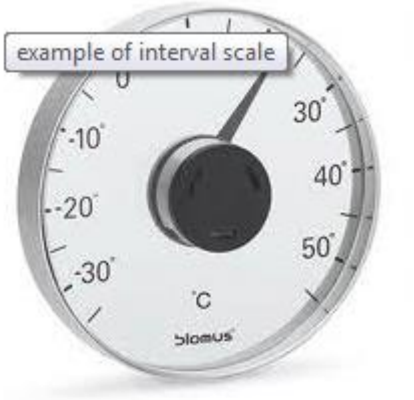
● 1 – Very Unsatisfied
○ 2 – Somewhat Unsatisfied
○ 3 – Neutral
○ 4 – Somewhat Satisfied
○ 5 – Very Satisfied

Example of Ordinal Scales

4.1.2 Four Scales of Measurements

3) Interval

- quantitative; order known; differences between values known
- E.g. Celsius scale
- “Difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees”
- No “true zero”; impossible to obtain ratio



Example of Interval Scale

4.1.2 Four Scales of Measurements

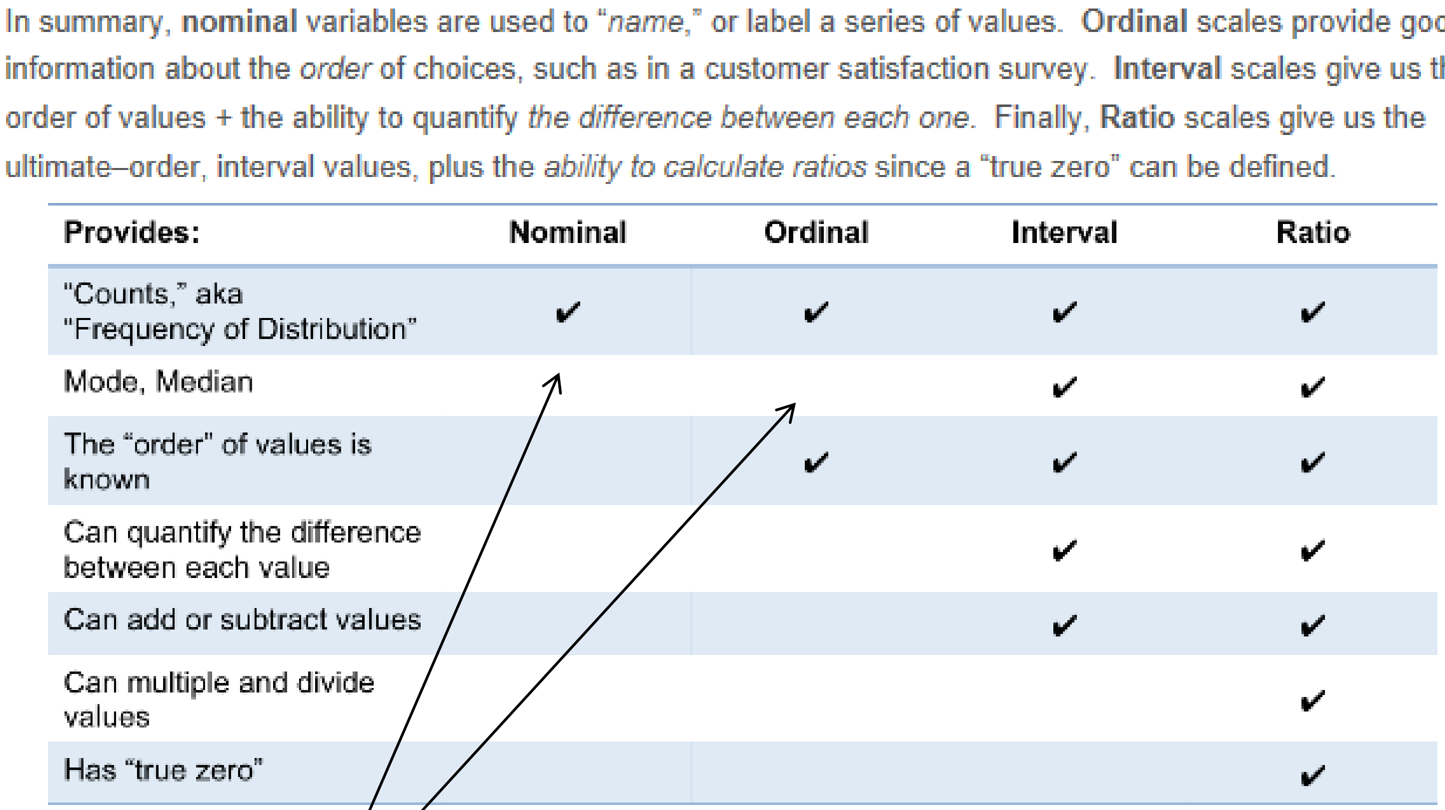
4) Ratio

- quantitative; order known; differences between values known, “true zero”
- E.g. Weight, height
- Possible to obtain ratio; “20 kg is twice as heavy as 10 kg”

This Device Provides Two Examples of
Ratio Scales (height and weight)

4.1.2 Four Scales of Measurements

Summary



In summary, **nominal** variables are used to "*name*," or label a series of values. **Ordinal** scales provide good information about the *order* of choices, such as in a customer satisfaction survey. **Interval** scales give us the order of values + the ability to quantify *the difference between each one*. Finally, **Ratio** scales give us the ultimate–order, interval values, plus the *ability to calculate ratios* since a "true zero" can be defined.

Provides:	**Nominal**	**Ordinal**	**Interval**	**Ratio**
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode, Median			✔	✔
The "order" of values is known		✔	✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

summary of data types and scale measures

Mode ✓

Median ×

4.1.3 Describing Distributions

1) Measures of Central Tendency: the single numerical value that is considered to be the most *typical* of data.

i. **Mean $\overline{\boldsymbol{x}}$**: arithmetic average

E.g. the mean of the observations 2, 3, 3, and 4, is equal to 3.

ii. **Median m**: the centre point in a set of ordered numbers; it is also the fiftieth percentile.

$$m := \begin{cases} x_{\frac{n+1}{2}}, \text{when } n \text{ is odd.} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, \text{when } n \text{ is even.} \end{cases}$$

E.g. the median of the observations 5, 1, 2, 4, and 3 is equal to 3

iii. **Mode**: the most frequently occurring number

4.1.3 Describing Distributions

2) Measures of Spread: information regarding the variability of the data.

i. **Range**: difference between the biggest and smallest numbers

ii. **Interquartile range $\boldsymbol{IQR}$**: difference between upper quartile Q_3 and lower quartile Q_1

Given a set of data $\{x_1, x_2, \cdots, x_n\}$ in ascending order,

- If n is odd: Q_1 is the median of $\{x_1, x_2, \cdots, x_{\frac{n+1}{2}-1}\}$ and Q_3 is the median of $\{x_{\frac{n+1}{2}+1}, \cdots, x_n\}$.
- If n is even: Q_1 is the median of $\{x_1, x_2, \cdots, x_{\frac{n}{2}}\}$ and Q_3 is the median of $\{x_{\frac{n}{2}+1}, \cdots, x_n\}$.

$$\boxed{IQR = Q_3 - Q_1}$$

4.1.3 Describing Distributions

iii. **Variance s^2 and Standard deviation s**: The standard deviation tells you, on average, how far the numbers vary from the mean.

For n observations $x_1, x_2, \cdots, x_n$ taken from a population

- Variance $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$
- Standard deviation $s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$

s^2 and s are large if the observations are widely spread about $\bar{x}$, and small if the observations are all close to $\bar{x}$.

Note

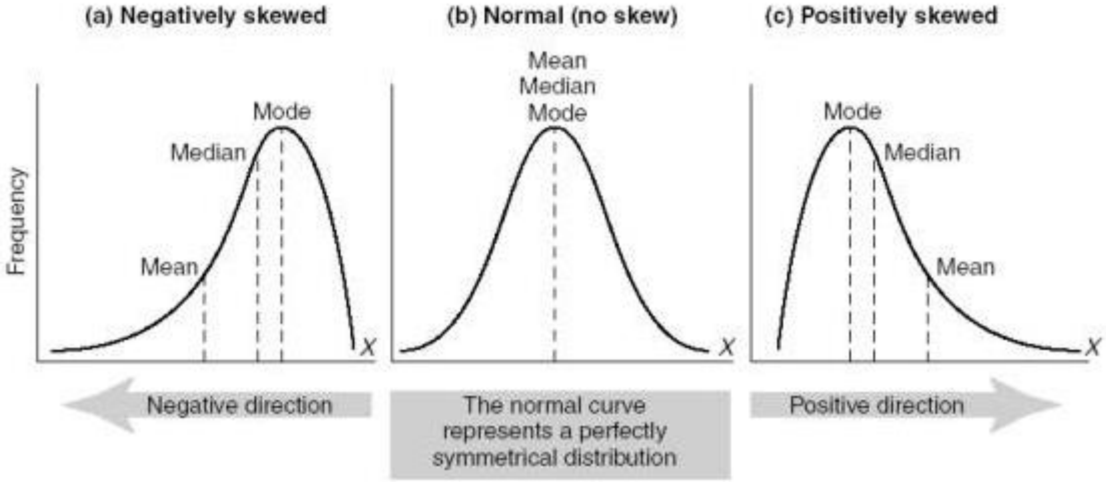
In the variance formula, $n - 1$ and not n, is the divisor. This is because the sum of the deviations $\sum(x_i - \bar{x})$ is always zero. So the last deviation can be found once we know the other $n - 1$ deviations. So we are not averaging n unrelated numbers.

Only $n - 1$ of the squared deviations can vary freely. So we average by dividing the total by $n - 1$.

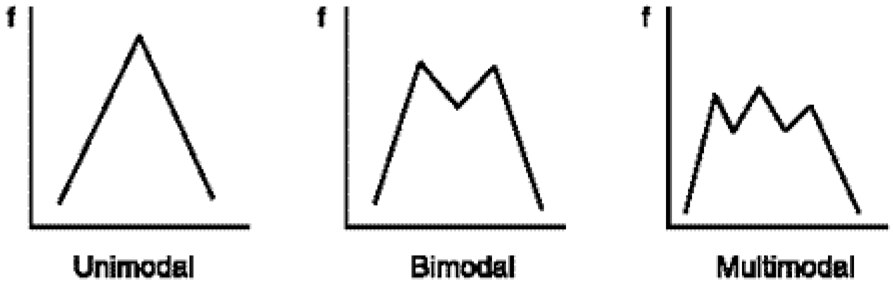
4.1.3 Describing Distributions

3) Shape:

i. **Skewness**: Lack of symmetry



ii. **Number of modes**



4.1.3 Describing Distributions

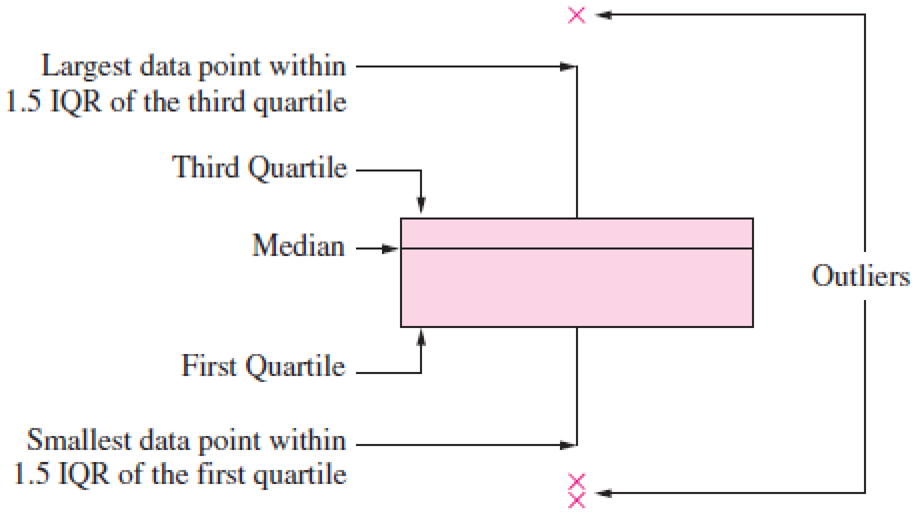
iii. **Outlier**: An observation that is considered to be unusually far from the bulk of the data. It may be due to experimental error; the point is sometimes excluded from the data set

- an observation maybe an outlier if it falls more than $1.5IQR$ above Q_3 or below Q_1.

4.1.4 Box-and-whisker plot

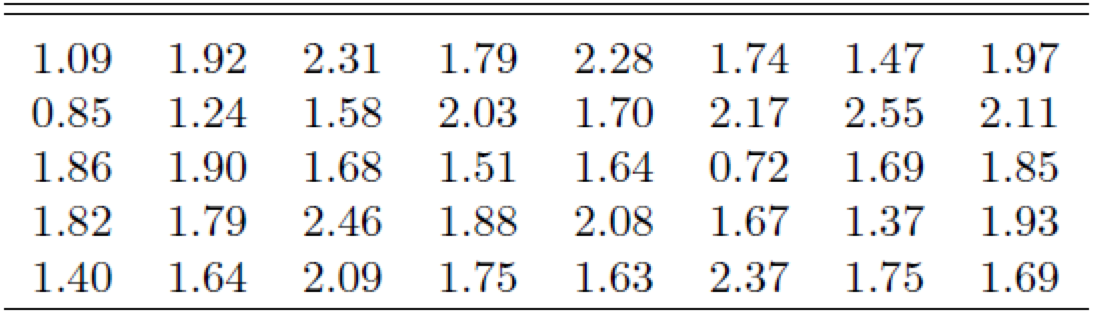
A Box-and-whisker plot is a graphic that presents the median, the Q_1 and Q_3, and any (potential) outliers that are present in a sample.

The visual information in the box-and-whisker plot or box plot is **not intended** to be a formal test for outliers. Rather, it is viewed as a diagnostic tool.


Largest data point within
1.5 IQR of the third quartile
Third Quartile
Median
First Quartile
Smallest data point within
1.5 IQR of the first quartile
Outliers

4.1.5 Matlab Implementation

Nicotine content was measured in a random sample of 40 cigarettes. The data are displayed below.



1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

Use Matlab to give a Box-and-whisker plot.

Solution

```
>>  A=[1.09 1.92 2.31 1.79 2.28 1.74 1.47 1.97 0.85 1.24 1.58 2.03 1.70
    2.17 2.55 2.11 1.86 1.90 1.68 1.51 1.64 0.72 1.69 1.85 1.82 1.79 2.46
    1.88 2.08 1.67 1.37 1.93 1.40 1.64 2.09 1.75 1.63 2.37 1.75 1.69]'
```

4.1.5 Matlab Implementation

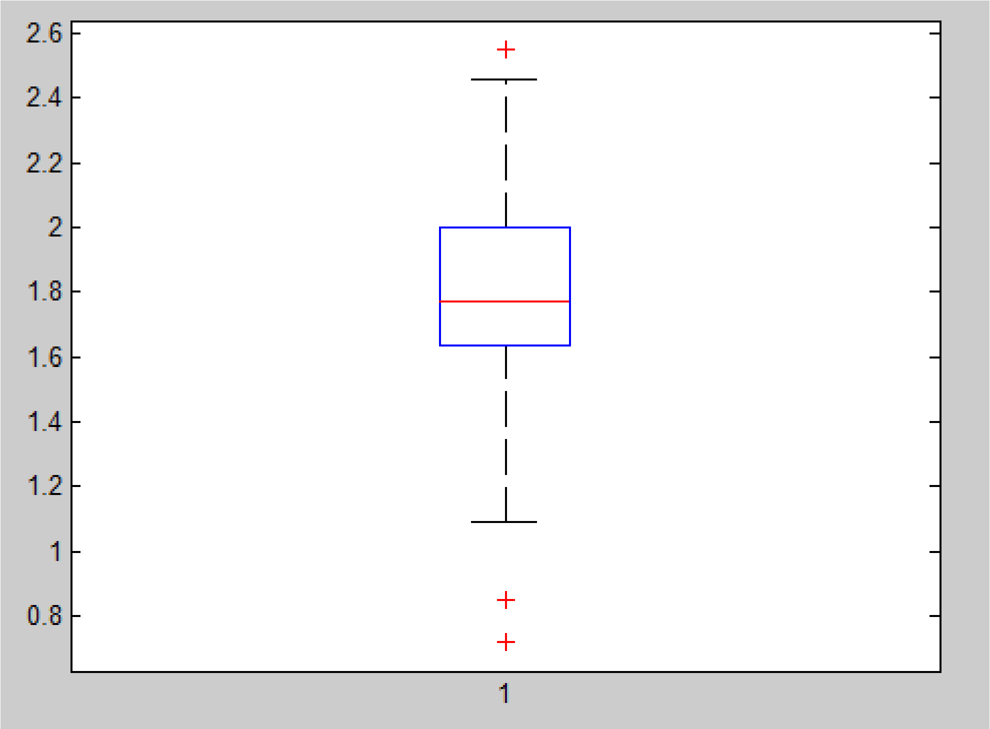
```
A =

    1.0900
    1.9200
    2.3100
    1.7900
    2.2800
    1.7400
    1.4700
    1.9700
    0.8500
    1.2400
    1.5800
    2.0300
    1.7000
    2.1700
    2.5500
    2.1100
    1.8600
    1.9000
    1.6800
    1.5100
    1.6400
    0.7200
    1.6900
    1.8500
    1.8200
    1.7900
    2.4600
    1.8800
    2.0800
    1.6700
    1.3700
    1.9300
    1.4000
    1.6400
    2.0900
    1.7500
    1.6300
    2.3700
    1.7500
    1.6900
```

4.1.5 Matlab Implementation

```
>> median(A)

ans =

    1.7700

>> quantile(A,[0.25 0.5 0.75])

ans =

    1.6350    1.7700    2.0000

>> iqr(A)

ans =

    0.3650
```

4.1.5 Matlab Implementation

```
>> boxplot(A)
```



4.2.1 Random Variable and its Properties

1) A random variable X is a *rule* that assigns a real number to each outcome (sample point) of a random experiment.

Example In an experiment in which a coin is tossed ten times.
What is the sample space of this experiment? Define a random variable for this sample space and then find the space(range) of the random variable.

Solution
The sample space $S = \{s | s$ is a sequence of 10 heads or tails$\}$. The random variable can be the function $X: S \to \mathbb{R}$ defined as follows:

$X(s) =$ number of heads in sequence $s \in S$.

This random variable, for example, maps the sequence $HHTTTHTTHH$ to the real number 5, i.e. $X(HHTTTHTTHH) = 5$.
The space(range) of this random variable is $R_X = \{0,1,2,\cdots,10\}$. ■

4.2.1 Random Variable and its Properties

2) A random variable X may be discrete or continuous.

3) Given random variable X, its mean, median, mode, variance are as defined:

i) **Mean**

- A number μ such that

$$\mu = \begin{cases} \int xf(x)\,dx & \text{for continuous variable} \\ \sum xp(x) & \text{for discrete variable} \end{cases}$$

where $f(x)$ is probability density function and $p(x)$ is probability mass function.

ii) **Median**

- A number m such that $P(X \leq m) \geq \frac{1}{2}$ and $P(X \geq m) \geq \frac{1}{2}$ for discrete r.v. X
- The area under the probability density function $f(x)$ is divided into two parts by the vertical line $x = m$, with equal area $\frac{1}{2}$.

iii) **Mode**

- Local maxima of the probability density function of X
- Could be more than one

4.2.1 Random Variable and its Properties

iv) **Variance**

A number σ^2 such that

$$\sigma^2 = \begin{cases} \int (x-\mu)^2 f(x)\,dx & \text{for continuous variable} \\ \sum (x-\mu)^2 p(x) & \text{for discrete variable} \end{cases}$$

where $f(x)$ is probability density function, $p(x)$ is probability mass function and μ is the expected value of the random variable X.

4) The **cumulative distribution function** for a random variable X is

$$F(x) = P(X \le x) = \begin{cases} \int_{-\infty}^{x} f(t)\,dt & \text{for continuous variable} \\ \sum_{t \le x} p(t) & \text{for discrete variable} \end{cases}$$

where $f(t)$ is probability density function and $p(t)$ is probability mass function.

4.2.2 Sampling Distributions

- Previously, given a sample $x_1, x_2, \cdots, x_n$ we were interested in statistics such as sample mean $\bar{x} = \frac{\sum x_i}{n}$ and sample variance $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$.
- This sample $x_1, x_2, \cdots, x_n$ is one of the many possible outcomes of a random set of variables $X_1, X_2, \cdots, X_n$ where each of them is independent and has same probability density function.

Definition

Let $X_1, X_2, \cdots, X_n$ be n independent random variables, each having the probability density function $f(x)$. Define $X_1, X_2, \cdots, X_n$ to be a **random sample** of size n from the population. Its joint probability pdf is

$f(x_1, x_2, \cdots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$.

E.g. We know for a given (i.e. observed) sample $x_1, x_2, \cdots, x_n$ its sample mean is $\bar{x} = \frac{\sum x_i}{n}$ which is a number. For a **random sample** $X_1, X_2, \cdots, X_n$ its sample mean is $\bar{X} = \frac{\sum X_i}{n}$ which is a function of random variables.

4.2.2 Sampling Distributions

E.g. We know for a given (i.e. observed) sample $x_1, x_2, \cdots, x_n$ its sample variance is $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ which is a number. For a **random sample** $X_1, X_2, \cdots, X_n$ its sample variance mean is $S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$ which is a function of random variables.

- For a random sample, its statistics e.g. $\bar{X}$ and S^2 are functions of random variables. Therefore a statistic is a random variable too.

Definition

The probability distribution of a statistic is called a **sampling distribution**.

- *Why are we so interested in them?* They allow us to make inferences on the unknown population parameters, e.g. μ and σ^2.

4.2.3 What is the Sampling Distribution of $\bar{X}$?

The sampling distribution of $\bar{X}$ with sample size n is the distribution that results when an experiment is **conducted over and over** (always with sample size n) and **many values of $\bar{X}$ result**.

The sampling distribution of $\bar{X}$ describes the variability of sample averages around population mean μ.

Case 1) $X \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Proof $E[\bar{X}] = \frac{\sum_{i=1}^{n} EX_i}{n} = \frac{n\mu}{n} = \mu$.

$$Var[\bar{X}] = E[(\bar{X} - E[\bar{X}])^2] = E[(\bar{X} - \mu)^2] = E\left[\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right)^2\right] = \frac{1}{n^2} E\left[\left(\sum_{i=1}^{n}(X_i - \mu)\right)^2\right]$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{n}(X_i - \mu)(X_j - \mu)\right] = \frac{1}{n^2} n\sigma^2 + \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} E\left(X_i X_j - \mu X_j - \mu X_i + \mu^2\right)$$

$$= \frac{\sigma^2}{n} + \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \left(E[X_i]E[X_j] - \mu E[X_j] - \mu E[X_i] + \mu^2\right) = \frac{\sigma^2}{n} + \frac{1}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} (\mu^2 - \mu^2 - \mu^2 + \mu^2) = \frac{\sigma^2}{n} \quad \blacksquare$$

4.2.3 What is the Sampling Distribution of $\bar{X}$?

Case 2) Population with unknown distribution – **Use Central Limit Theorem**!

Theorem (Central Limit Theorem)

If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then if n is sufficiently large,

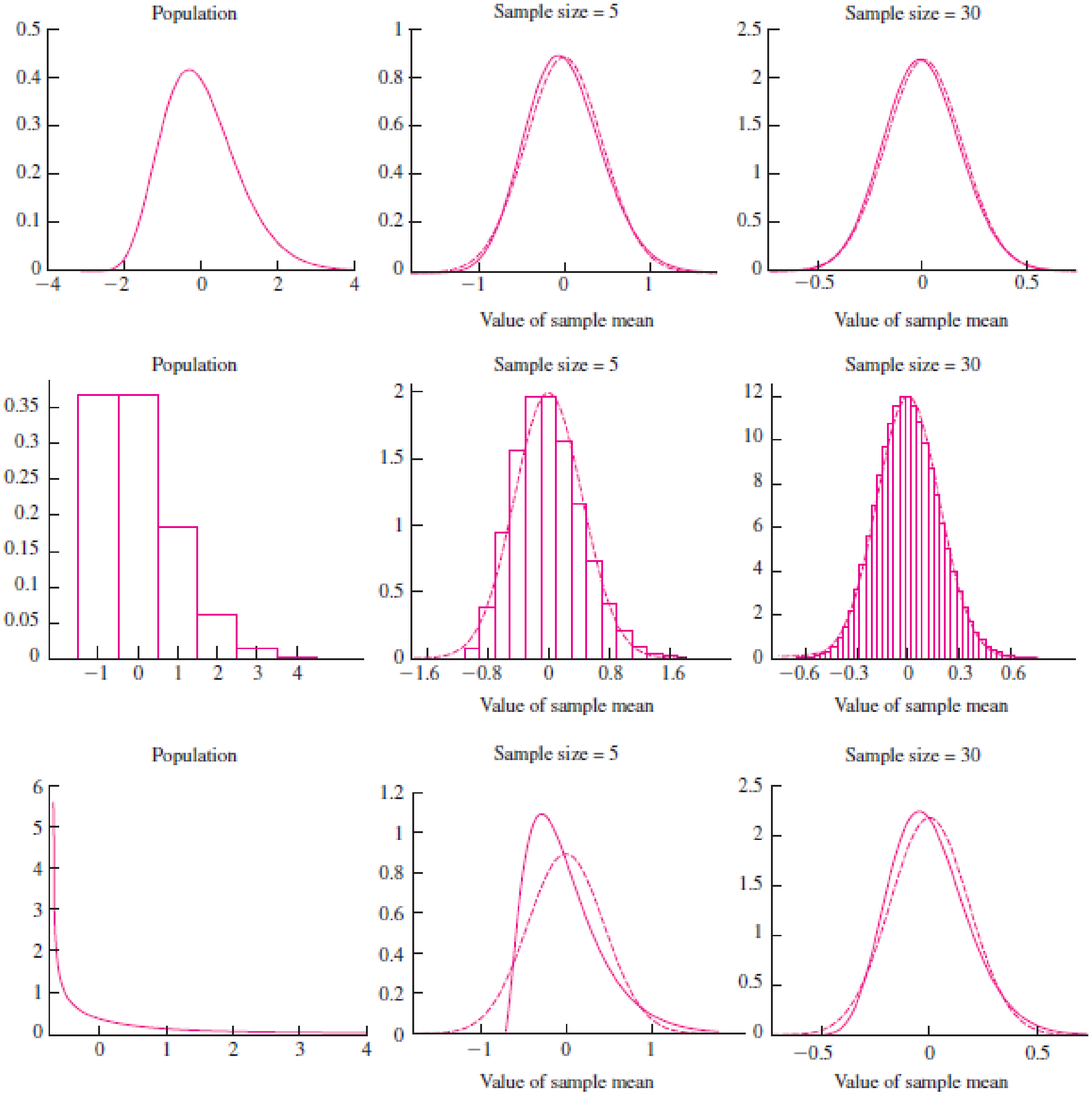
$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ approximately.

Note

1. For most population, if the sample size $n \geq 30$, the CLT approximation is good.
2. If $n < 30$, approximation is good only if the population is not too different from a normal population.
3. If population is normal (i.e. Case 1) then $\bar{X}$ is normally distributed, no matter how small n is.

4.2.3 What is the Sampling Distribution of $\bar{X}$?

The diagram below shows the sampling distribution of $\bar{X}$, for different populations.



Observations

1) The distribution of $\bar{X}$ becomes closer to normal as n increases.

2) The mean of $\bar{X}$ remains μ for any sample size and the variance of $\bar{X}$ gets smaller as n increases.

Example 1

At a large university, the mean age of the students is 22.3 years, and the standard deviation is 4 years. A random sample of 64 students is drawn. What is the probability that the average age of these students is greater than 23 years?

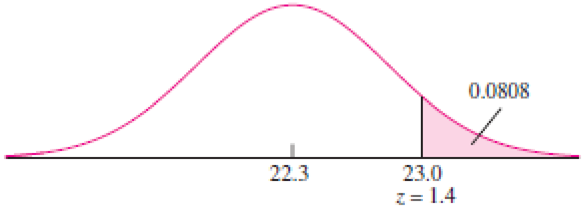
Solution

Let $X_1, X_2, \cdots, X_{64}$ be the ages of the 64 students in the random sample.

Given: $\mu = 22.3$, $\sigma^2 = 16$ and $n = 64$. We want to find $P(\bar{X} > 23)$.

Using CLT, $\bar{X} \sim N(22.3, 0.25)$. Therefore

$$P(\bar{X} > 23) = P\left(\frac{\bar{X} - 22.3}{\sqrt{0.25}} > \frac{23 - 22.3}{\sqrt{0.25}}\right) = P(Z > 1.40)$$



From the z table, the area to the right of $z = 1.40$ is 0.0808, thus the required probability is 0.0808 . ■

Example 2

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with $\mu = 800$ hours and $\sigma = 40$ hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

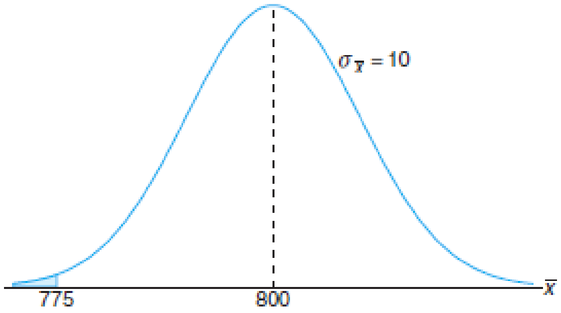
Solution

Let $X_1, X_2, \cdots, X_{16}$ be the length of life of 16 bulbs in the random sample.

Given: $\mu = 800$, $\sigma^2 = 40^2, n = 16$ and population is **approximately normal**.

We want to find $P(\bar{X} < 775)$.

Using CLT, $\bar{X} \sim N(800, 10^2)$. Therefore



$$P(\bar{X} < 775) = P\left(\frac{\bar{X} - 800}{10} < \frac{775 - 800}{10}\right) = P(Z < -2.5) = 0.0062 \text{ from } z \text{ table.}$$

Hence the required probability is 0.0062 . ■

4.2.4 What is the CLT used for?

1) One application of CLT is to obtain the sampling distribution of $\bar{X}$ (based on population parameters μ, σ^2 and sample size n).

2) Another application of CLT is to determine reasonable values of unknown μ (given σ^2 and n).
    - Hypothesis Testing
    - Estimation
    - Quality Control

For example, in hypothesis testing, we decide if the given data (in the form of $\bar{x}$) supports our belief about μ taking a particular value. This will be taught later.

4.2.5 Sampling Distribution of the Difference between Two Means, $\bar{X}_1 - \bar{X}_2$

The CLT can be extended to the two-sample, two-population case.

Theorem

If independent samples of size n_1 and n_2 are randomly drawn from two populations, discrete or continuous, with means μ_1 and μ_2 and finite variances σ_1^2 and σ_2^2 respectively. Then if n_1, n_2 are sufficiently large,

$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ approximately.

Note

1. For most population, if the sample size $n_1, n_2 \geq 30$, the CLT approximation is good.
2. If $n_1, n_2 < 30$, approximation is good only if the populations are not too different from normal populations.
3. If **both** populations are normal then $\bar{X}_1 - \bar{X}_2$ is normally distributed, no matter how small n_1, n_2 are.

Example 3

Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0 .

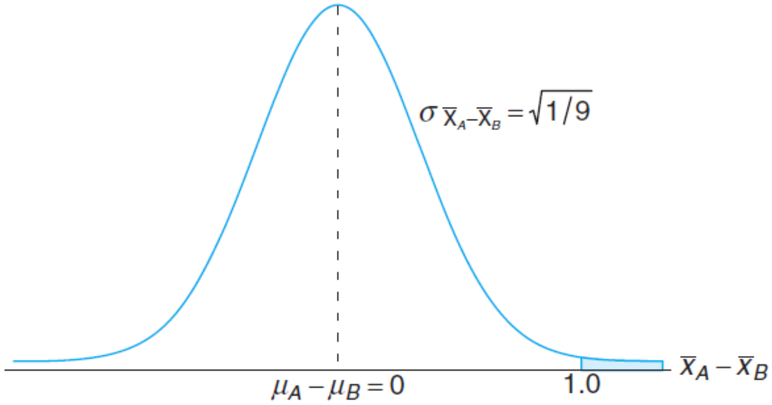
Assuming the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A$ and $\bar{X}_B$ are respectively the average drying times for samples of size $n_A = n_B = 18$.

You may assume that the two populations are approximately normal.

Solution

By CLT, $\bar{X}_A - \bar{X}_B \sim N\left(\mu_A - \mu_B, \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right)$. Therefore in this case $\bar{X}_A - \bar{X}_B \sim N\left(0, \frac{1}{9}\right)$.

The required probability is given by the shaded region.



Therefore $P(\bar{X}_A - \bar{X}_B > 1.0) = P\left(Z > \frac{1.0-0}{\sqrt{\frac{1}{9}}}\right) = P(Z > 3.0) = 1 - P(Z \le 3.0)$

$= 1 - 0.9987 = 0.0013$ ■

4.3.1 χ^2-Distribution

We now consider the sampling distribution of $\frac{(n-1)S^2}{\sigma^2}$.

If a random sample of size n is drawn from a normal population with mean μ and variance σ^2, and the sample variance is computed, we obtain a value of the statistic S^2. We shall proceed to obtain the distribution of the statistic $\frac{(n-1)S^2}{\sigma^2}$.

Considering $\sum_{i=1}^{n}(X_i - \mu)^2$,

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}[(X_i - \bar{X}) + (\bar{X} - \mu)]^2$$
$$= \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - \mu)^2 + 2\,(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \bar{X})$$
$$= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

Dividing by σ^2 and substituting $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$, we have

$$\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}. \qquad (*)$$

We know from statistical theory that

1) The random variable $Y = \frac{(X-\mu)^2}{\sigma^2}$ has a χ^2-distribution with 1 degree of freedom when X has a normal distribution with mean μ and variance σ^2.

So LHS of $(*)$ becomes sum of n (independent) χ^2 random variable each with 1 (since $X_1, X_2, \cdots, X_n$ are i.i.d.) degree of freedom.

2) If $W_1, W_2, \cdots, W_n$ are independent χ^2 random variables with, respectively, $v_1, v_2, \cdots, v_n$ degrees of freedom, then the random variable $Y = W_1 + W_2 + \cdots + W_n$ has a χ^2-distribution with $(v_1 + v_2 + \cdots + v_n)$ degrees of freedom.

So LHS of $(*)$ becomes a chi-squared random variable with n degrees of freedom.

3) The random variable $Y = \frac{(\bar{X}-\mu)^2}{\sigma^2/n}$ has a χ^2-distribution with 1 degree of freedom when $\bar{X}$ has a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

So 2nd term of RHS of $(*)$ becomes a chi-squared random variable with 1 degree of freedom.

Therefore $\frac{(n-1)S^2}{\sigma^2}$ is a χ^2 random variable with $n-1$ degrees of freedom. We formalize this in the following theorem.

Xi'an Jiaotong-Liverpool University 西交利物浦大学

Theorem

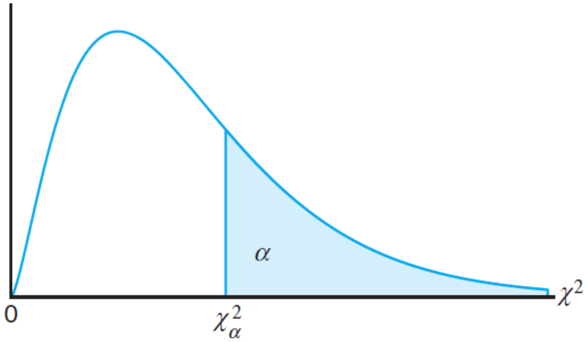
If S^2 is the variance of a random sample of size n taken from a normal population having variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

has a chi-squared distribution with $v = n - 1$ degrees of freedom.

4.3.2 How does the χ^2-Distribution look like?

- The general shape of the χ^2-distribution is given here.



- Note $\chi^2 \in [0, \infty)$.
- We denote χ^2_α to be the value of χ^2 above which we can find area of α.
- From statistical table, for 7 degrees of freedom, we have $\chi^2_{0.05} = 14.067$ and $\chi^2_{0.95} = 2.167$.

- Exactly 95% of a χ^2-distribution lies between $\chi^2_{0.975}$ and $\chi^2_{0.025}$. A χ^2 value falling to the right of $\chi^2_{0.025}$ is not likely to occur unless our assumed value of σ^2 is too small. A χ^2 value falling to the left of $\chi^2_{0.975}$ is not likely to occur unless our assumed value of σ^2 is too large.

Example 4

A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5 and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.

Solution

We first find the sample variance s^2.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{(1.9-3)^2+(2.4-3)^2+(3.0-3)^2+(3.5-3)^2+(4.2-3)^2}{4} = \frac{3.26}{4} = 0.815$$

Then $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(4)(0.815)}{1} = 3.26$

is a value from a chi-squared distribution with 4 degrees of freedom. Since 95% of the χ^2 values with 4 degrees of freedom fall between 0.484 and 11.143, the computed value with $\sigma^2 = 1$ is reasonable. The manufacturer has no reason to suspect that the standard deviation is not 1 year. ■

4.3.3 What is the χ^2-Distribution Used For?

The χ^2-distribution is commonly used for statistical inferences. For example, it is used in tests to compare actual data with data we would expect to obtain according to a certain assumption.

As statisticians, we want to know:

1) Were the deviations (differences between observed and expected) the result of chance, or were they due to other factors?
2) How much deviation can occur before we must conclude that something other than chance is at work, causing the observed to differ from the expected?

The actual use of χ^2-tests shall be covered in future topics.

4.4.1 Student t-Distribution

Theorem

Let Z be a standard normal random variable and V a chi-squared random variable. If Z and V are independent, then the distribution of the random variable $T = \frac{Z}{\sqrt{V/v}}$ is given by the density function

$$h(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{\pi v}}\left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}} \quad ; \quad -\infty < t < \infty.$$

This is know as the t-distribution with v degrees of freedom. ■

Corollary

Let $X_1, X_2, \cdots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Then the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $v = n - 1$ degrees of freedom. ■

4.4.2 How does the t-Distribution look like?

- Distribution of T very similar to the distribution of Z
- bell-shaped; mean zero; variance depends on sample size n
- As $n \to \infty$, distribution of T becomes distribution of Z

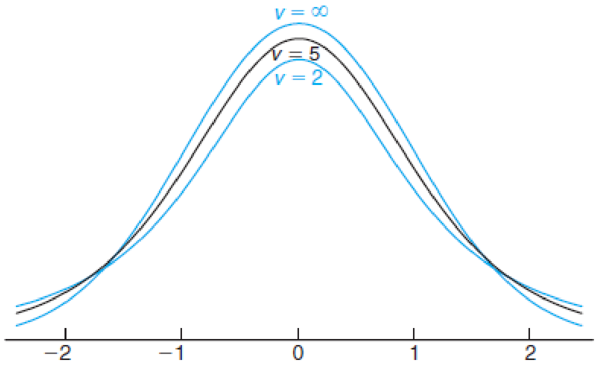


Figure 8.8: The t-distribution curves for $v = 2, 5,$ and ∞.

4.4.2 How does the t-Distribution look like?

- We let t_α denote the t-value above which we find an area equal to α
- By symmetry about zero, $t_{1-\alpha} = -t_\alpha$

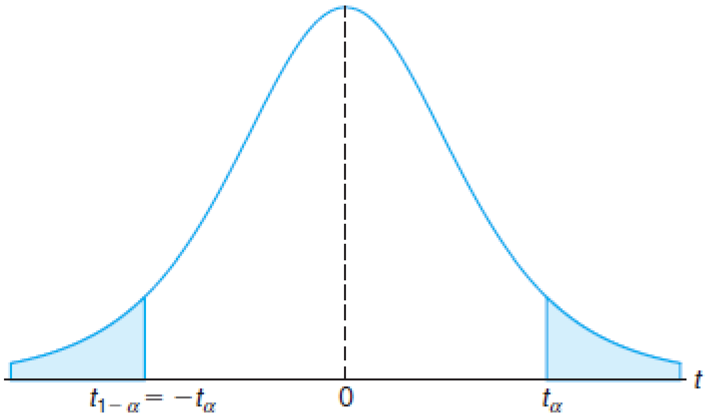


Figure 8.9: Symmetry property (about 0) of the t-distribution.

Example 5

Given the random variable T has a t-distribution, find $P(-t_{0.025} < T < t_{0.05})$.

Solution

Since $t_{0.05}$ leaves an area of 0.05 to the right, and $-t_{0.025}$ leaves an area of 0.025 to the left, we find an area of

$$1 - 0.05 - 0.025 = 0.925$$

between $-t_{0.025}$ and $t_{0.05}$. Hence $P(-t_{0.025} < T < t_{0.05}) = 0.925$ ■

4.4.3 What is the t-distribution Used For?

Used extensively in

1) problems that deal with inference about population mean μ (when population variance σ^2 is unknown)
2) problems where one is trying to determine if means from two samples are significantly different (when population variances σ_1^2 and σ_2^2 are unknown)

Note

1. The use of the t-distribution for the statistic $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ requires that $X_1, X_2, \cdots, X_n$ are **normal**.

2. The use of the t-distribution and the sample size consideration is **not related** to the Central Limit Theorem.

Example 6

A chemical engineer claims that the population mean yield of a certain batch process is 500 g/ml of raw material. To check this claim he samples 25 batches each month. If the computed t-value falls between $-t_{0.05}$ and $t_{0.05}$, he is satisfied with this claim.

1) What conclusion show he draw from a sample that has a mean $\bar{x} = 518$ g/ml and a sample standard deviation $s = 40$ g? Assume the distribution of yields to be normal.

2) What if the distribution of yields is not normal?

Solution

1) From statistical table, $t_{0.05} = 1.711$ with 24 degrees of freedom. Therefore, engineer is satisfied if a sample of 25 batches yields a t-value between -1.711 and 1.711.

If $\mu = 500$, then $t = \frac{518-500}{40/\sqrt{25}} = 2.25$, so he should **not** be satisfied with his claim.

In fact, the probability of obtaining a t-value, with $v = 24$, equal to or greater than 2.25 is approximate 0.02 . If $\mu > 500$, the value of t computed would be smaller, and therefore more reasonable (since corresponding probability would be bigger).

Hence the engineer is likely to conclude that the process produces a better product than he thought.

2) The statistic $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ **cannot be used** since it requires that $X_1, X_2, \cdots, X_n$ to be normal. Furthermore, the Central Limit Theorem also **cannot be used**. This is because the statistic $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ requires σ to be **explicitly known**. ■

4.5.1 F-Distribution

Theorem

Let U and V be two independent random variables having chi-squared distributions with v_1 and v_2 degrees of freedom, respectively. Then the distribution of the random variable $F = \frac{U/v_1}{V/v_2}$ is given by the density function

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}} & ; \quad f > 0 \\ 0 \quad ; \quad f \leq 0 \end{cases}$$

Corollary

If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2 respectively, then

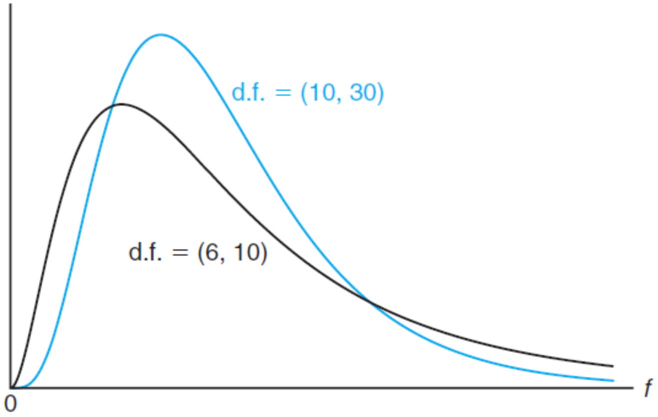
$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

has an F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

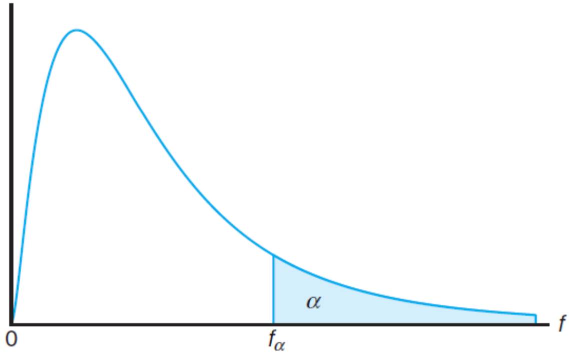
■

4.5.2 How does the F-Distribution look like?

- Density function of F depends on v_1, v_2 but also on the order that we state them
- Typical F-distributions are given below



- Let f_α be the f-value above which we find an area equal to α.



4.5.2 How does the F-Distribution look like?

- Usually statistical tables give values of f_α for $\alpha = 0.05$ or 0.01.
- E.g. $f_\alpha = 3.22$ for $\mathrm{d.f.} = (6{,}10)$
- The following theorem allows us to obtain values of $f_{0.95}$ or $f_{0.99}$

Theorem

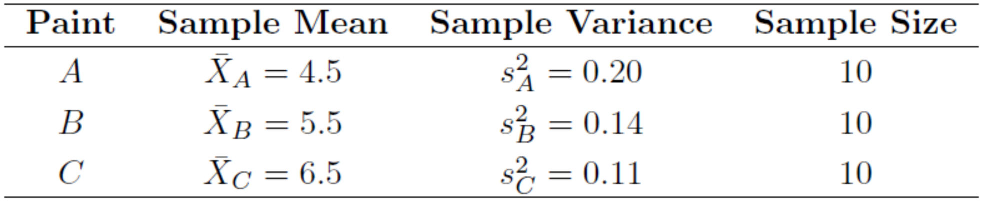
Writing $f_\alpha(v_1, v_2)$ for f_α with v_1 and v_2 degrees of freedom, we have

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2,v_1)}. \quad ■$$

- E.g. $f_{0.95}(6{,}10) = \frac{1}{f_{0.05}(10{,}6)} = \frac{1}{4.06} = 0.246$

4.5.3 What is the F-Distribution Used For?

- The F-distribution is called the variance ratio distribution. It is generally used for the **analysis of variance**.
- E.g. Suppose we sample from three types of paints A, B, C. We wish to find out if $\mu_A = \mu_B = \mu_C$. The information obtained from sampling:



Paint	Sample Mean	Sample Variance	Sample Size
A	$\bar{X}_A = 4.5$	$s_A^2 = 0.20$	10
B	$\bar{X}_B = 5.5$	$s_B^2 = 0.14$	10
C	$\bar{X}_C = 6.5$	$s_C^2 = 0.11$	10

- For $\mu_A = \mu_B = \mu_C$, we will need the sample means $\bar{x}_A, \bar{x}_B, \bar{x}_C$ to be close enough. It would seem reasonable if the variability between $\bar{x}_A, \bar{x}_B, \bar{x}_C$ is smaller than what we would expect <u>*by chance*</u>. So how to find out?

- Analysis involves:

1. Variability within samples (i.e. s_A^2, s_B^2, s_C^2)
2. Variability between samples (i.e. between $\bar{x}_A, \bar{x}_B, \bar{x}_C$)

4.5.3 What is the F-Distribution Used For?

- It is unlikely that data from populations with $\mu_A = \mu_B = \mu_C$ could have variability between sample averages **considerably larger** than variability within samples.
- Therefore it is highly possible that $\mu_A = \mu_B = \mu_C$ in this case:

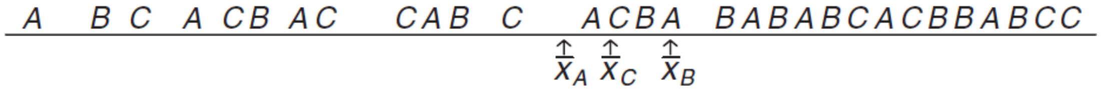


Figure 8.14: Data that easily could have come from the same population.

but not in this case:

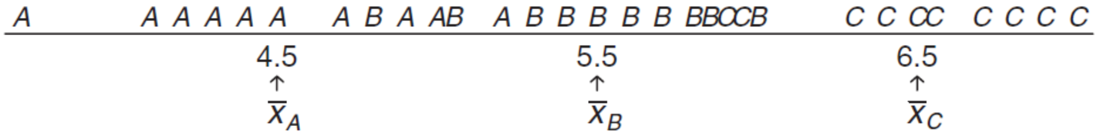


Figure 8.13: Data from three distinct samples.

- The two variabilities generate important ratios of variances, whose ratio are used together with the F-distribution. ■

4.6 Matlab Implementation

To find area under the distribution curve, use the Matlab cdf functions.

```
>>  chi2cdf(X,V)
```

computes the *chi-square* cdf at the value of X using the corresponding degrees of freedom in V.

```
>>  fcdf(X,V1,V2)
```

computes the *F* cdf at each of the value of X using the corresponding degrees of freedom $V1$ and $V2$.

```
>> normcdf(X,mu,sigma)
```

computes the *normal* cdf at the value of X using the corresponding mean *mu* and standard deviation *sigma*.

```
>> normcdf(X)
```

computes the *standard normal* cdf at the value of X using the corresponding mean 0 and standard deviation *1*.

4.6 Matlab Implementation

```
>>   tcdf(X,V)
```

computes *Student's t* cdf at the value of X using the corresponding degrees of freedom in V.

4.7 Summary

- Know Data and its representations
- Measures of central tendency and spread
- How to identify observations that are possible outliers
- Understand what is meant by random sample and sampling distribution
- The sampling distribution of $\bar{X}$, $\bar{X}_1 - \bar{X}_2$, $\frac{(n-1)S^2}{\sigma^2}$, $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ and F-distribution
- The use of the statistical table.

Chapter 5: Estimation of Population Parameters

Previously, we looked at the distribution of $\bar{X}$ and S^2 of a random sample. The sampling distributions allow us to conclude about population parameters such as μ and σ^2.

Statistical inferences can be divided into

1. Estimation:
   - Point estimation
   - Confidence interval
2. Hypothesis Testing
   - Parametric test
   - Nonparametric test

We look at point estimation and confidence interval in this chapter,

Chapter 5: Estimation of Population Parameters

- Reading task: Moore-McCabe-Craig Chapter ***6.1***

5.1 Point Estimation

5.1.1 Unbiased Estimator

5.1.2 Variance of an Estimator

5.2 Interval Estimation

5.2.1 Single Sample: Estimating μ

5.2.2 Two Samples: Estimating $\mu_1 - \mu_2$

5.2.4 Single Sample: Estimating σ^2

5.2.5 Two Samples: Estimating σ_1^2/σ_2^2

5.3 Summary

5.1 Point Estimation

- Reading task: Moore-McCabe-Craig Chapters ***6.1***

Definition

An *estimator* $\widehat{\Theta}$ is a statistic (i.e. a function of the random sample $X_1, X_2, \cdots, X_n$) that is used to infer the value of an unknown population parameter θ. The value $\hat{\theta}$ taken by an estimator is called a (point) *estimate* of the unknown parameter.

E.g. The statistic $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ is an *estimator* of the population mean μ. The value $\bar{x}$ based on observations $x_1, x_2, \cdots x_n$ is an *estimate* of μ.

We can have different estimators for an unknown population parameter. To decide which one to use, we look at

1. Unbiasedness
2. Variance

of an estimator.

5.1.1 Unbiased Estimator

Definition

A statistic $\widehat{\Theta}$ is said to be an **unbiased estimator** of the parameter θ if $E\left[\widehat{\Theta}\right] = \theta$.

E.g. S^2 is an unbiased estimator of σ^2.

Certainly we would like $\widehat{\Theta}$ to be an unbiased estimator of θ.

5.1.2 Variance of an Estimator

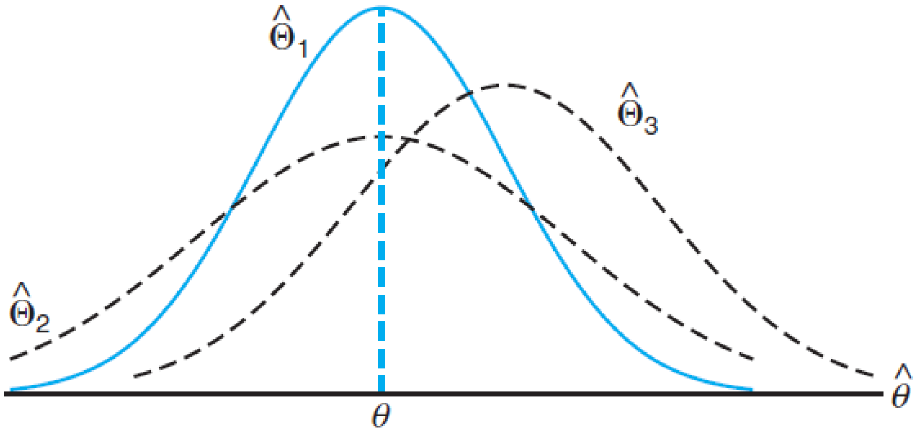
If $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$ are two unbiased estimates of the same population parameter θ, the estimator whose sampling distribution has the smallest variance should be chosen.

Definition

If we consider all possible unbiased estimators of some parameter θ, the one with the smallest variance is called the most efficient estimator of θ.

Example

The sampling distributions of three different estimators, $\widehat{\Theta}_1, \widehat{\Theta}_2$ and $\widehat{\Theta}_3$, are given.



We observe that

1. Only $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$ are unbiased
2. The estimator $\widehat{\Theta}_1$ has a smaller variance than $\widehat{\Theta}_2$, so is more efficient.

Our choice of estimator would be $\widehat{\Theta}_1$.

5.2 Interval Estimation

- Even the most efficient unbiased estimator is unlikely to estimate population parameter θ exactly, i.e. its point estimate from any sample is not exactly equal to the value of θ.
- It is preferable to determine an interval within which we would expect to find the value of the parameter. This is called **interval estimate**.

Definition

An interval estimate of a parameter θ is an interval of the form $\hat{\theta}_L < \theta < \hat{\theta}_U$, where $\hat{\theta}_L$ and $\hat{\theta}_U$ depends on the value of $\widehat{\Theta}$ for a particular sample and also on the sampling distribution of $\widehat{\Theta}$.

Note

Different samples will generate different values of $\widehat{\Theta}$, so $\hat{\theta}_L$ and $\hat{\theta}_U$ are values of random variables which we denote $\widehat{\Theta}_L$ and $\widehat{\Theta}_U$.

Definition

A $100(1-\alpha)\%$ confidence interval is an interval of the form $\hat{\theta}_L < \theta < \hat{\theta}_U$, where $\hat{\theta}_L$ and $\hat{\theta}_U$ are respectively values of $\widehat{\Theta}_L$ and $\widehat{\Theta}_U$ obtained for a particular sample, based on

$$P\left(\widehat{\Theta}_L < \theta < \widehat{\Theta}_U\right) = 1-\alpha \quad ; \quad 0 < \alpha < 1$$

in the estimation of population parameter θ.

Note

1. The point estimate is a single number derived from a set of observations, while the confidence interval estimate is an interval that is reasonable for the parameter based on a set of observations. These estimates are **not random**.

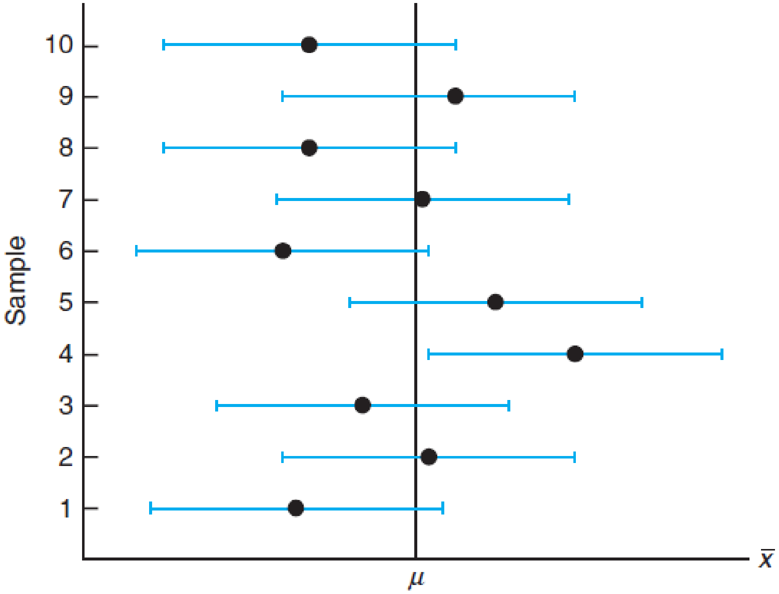
2. Ideally, we prefer a shorter interval with a high degree of confidence. E.g. It is better to be 95% confident that $\mu \in [6,7]$ than to be 99% that $\mu \in [3,10]$.

3. A 95% confidence interval **does not mean** that

(i) there is a 95% probability that this interval covers the population parameter, nor that

(ii) there is a 95% probability that the population parameter lies within this interval.

Once an experiment is done and an interval calculated, this interval either covers the parameter value or it does not. It is no longer a matter of probability. The 95% probability relates to the reliability of the estimation procedure, **not** to a specific calculated interval, i.e. 95% of the confidence intervals computed this way will contain the unknown parameter.

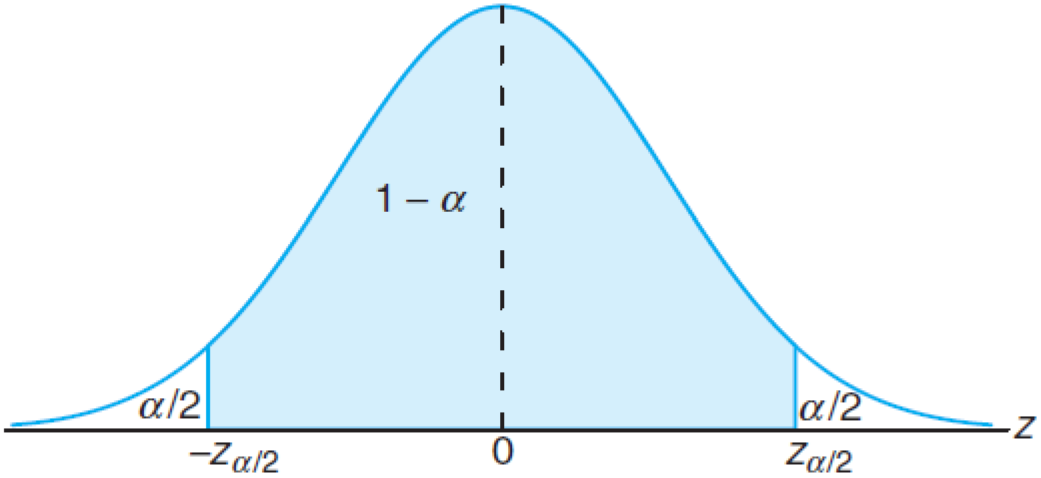


5.2.1 Single Sample: Estimating μ

There are several cases. We shall look at each case separately.

Case 1: Confidence Interval on μ, σ^2 known

According to CLT, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ for sufficiently large sample. Writing $z_{\alpha/2}$ for the z-value above which we find an area of $\alpha/2$ under the standard normal curve, we see from the diagram below that



$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$$

where $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.

Hence,

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

Rearranging, we have

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Based on this, the $100(1-\alpha)\%$ confidence interval can be calculated.

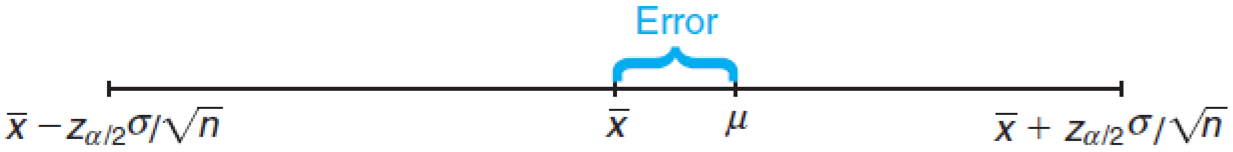
If $\bar{x}$ is the mean of a random sample of size n from a population with known variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

Note

1. The confidence interval is based on data that is already observed; there is **nothing random** about it.
2. For samples of size $n \geq 30$, sampling theory guarantees good results.
3. Different samples will yield different values of $\bar{x}$, giving different confidence intervals. But in general $100(1-\alpha)\%$ of these intervals will cover μ.
4. The confidence interval reflects of the accuracy of our point estimate $\bar{x}$. When $\bar{X}$ is an estimator of μ, we are $100(1-\alpha)\%$ confident the size of error $|\bar{X}-\mu|$ will not exceed $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.


Error
$\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}$
$\bar{x}$
μ
$\bar{x} + z_{\alpha/2}\sigma/\sqrt{n}$

Example 1

To estimate the unknown population mean μ, the sample mean $\bar{x}$ is used. Given that the population standard deviation $\sigma = 0.3$, what is the minimum sample size required if we want to be at least 95% confident that our estimate is off by less than 0.05?

Solution

We have $P(|\bar{X} - \mu| < 0.05) \geq 0.95$. Standardizing, we obtain

$P\left(\frac{|\bar{X}-\mu|}{\frac{0.3}{\sqrt{n}}} < \frac{0.05}{\frac{0.3}{\sqrt{n}}}\right) \geq 0.95$. That is, $P\left(|Z| < \frac{0.05\sqrt{n}}{0.3}\right) \geq 0.95$

From statistical table, $z_{\alpha/2} = \frac{0.05\sqrt{n}}{0.3} \geq 1.96$ to be at least 95% confident.

Solving for n, we have $n = \left(\frac{(1.96)(03)}{0.05}\right)^2 \geq 138.3$

To be at least 95% confident, a random sample of minimum size 139 will provide an estimate $\bar{x}$ differing from μ by an amount less than 0.05 . ■

Example 2

The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 g/ml. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 g/ml.

Solution

The point estimate of μ is $\bar{x} = 2.6$.

For 95% C.I., we have $z_{0.025} = 1.96$, so the interval is

$$2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right)$$

$$\Rightarrow 2.50 < \mu < 2.70$$

For 99% C.I., we have $z_{0.005} = 2.575$, so the interval is

$$2.6 - (2.575)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (2.575)\left(\frac{0.3}{\sqrt{36}}\right)$$

$$\Rightarrow 2.47 < \mu < 2.73$$

We see that a longer interval is required to estimate μ with a higher degree of confidence. ■

Case 2: One-sided confidence bounds on μ, σ^2 known

The confidence interval in Case 1 is *two*-sided, i.e. upper and lower bounds are given. However, there are many applications where only *one* bound is required.

This is especially true if we are interested in the worse-case scenario. E.g. tensile strength of steel bars in a factory (we are interested in knowing the lower bound), mercury level in river (we are interested in upper bound).

For $100(1-\alpha)\%$ confidence bound, we consider

$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1-\alpha$$

for lower one-sided bound. Rearranging, we obtain

$$P\left(\mu > \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Similarly, we consider

$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > -z_\alpha\right) = 1-\alpha$$

for upper one-sided bound. Rearranging, we obtain

$$P\left(\mu < \bar{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Summarizing, we have the following results.

If $\bar{x}$ is the mean of a random sample of size n from a population with variance σ^2, the one-sided $100(1-\alpha)\%$ confidence bounds for μ are given by

upper one-sided bound: $\mu < \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}$

lower one-sided bound: $\mu > \bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}}$

Example 3

In an experiment, 25 subjects are selected randomly and their reaction time, in seconds, to a particular chemical is measured. Past experience suggests that the variance in reaction time to the chemical is $\sigma^2 = 4$ and the distribution of reaction time is approximately normal. The average time for the subjects is 6.2 seconds. Give an upper 95% bound for the mean reaction time μ.

Solution

The upper 95% bound is given by

$$\bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} = 6.2 + (1.645)\left(\frac{2}{\sqrt{25}}\right) = 6.2 + 0.658 = 6.858$$

Hence required bound is $\mu < 6.858$. ■

Case 3: Confidence Interval on μ, σ^2 unknown

Often we need to estimate the population mean μ with unknown variance σ^2.

We recall that for $n < 30, X_i$'s are i.i.d. such that $X_i \sim N(\mu, \sigma^2) \quad ; \quad i = 1,2,\cdots,n$

then $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

In this case, with σ unknown, T can be used to construct a $100(1-\alpha)\%$ confidence interval on μ. Consider

$$P\left(-t_{\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha$$

Rearranging,

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

We summarize our result below.

If $\bar{x}$ and s are respectively the mean and standard deviation of a small random sample $(n < 30)$ from a normal population with unknown variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the t-value with $n-1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

Note

1. The T statistic is used when sample size $n < 30$, population is normal **and** population variance σ^2 is unknown.

2. When sample size n is large, i.e. $n \geq 30$, the distribution of T is very close to normal, so the normal distribution can be used instead of the Student's t.

Example 4

The contents of seven similar containers of sulphuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2 and 9.6 litres. Find a 95% confidence interval for the mean content of all such containers in the warehouse, assuming a normal distribution.

Solution

From given data, $\bar{x} = 10.0$ and $s = 0.283$. From statistical table, $t_{0.025} = 2.447$ with 6 degrees of freedom. Hence 95% C.I. for μ is

$$10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right)$$

$$\Rightarrow 9.74 < \mu < 10.26$$

■

5.2.2 Two Samples: Estimating $\mu_1 - \mu_2$

Given two populations with means and variances respectively

1. μ_1 and σ_1^2
2. μ_2 and σ_2^2,

a point estimator of the difference between μ_1 and μ_2 is given by the statistic $\bar{X}_1 - \bar{X}_2$.

Clearly, to construct $100(1-\alpha)\%$ C.I. about $\mu_1 - \mu_2$, the sampling distribution of $\bar{X}_1 - \bar{X}_2$ must be known. We consider a few cases.

Case 1: Confidence Interval for $\boldsymbol{\mu_1 - \mu_2}$, both $\boldsymbol{\sigma_1^2}$ and $\boldsymbol{\sigma_2^2}$ known

If sample size $n_1, n_2 \geq 30$, then by CLT, $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ approximately. Letting $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$, then $Z \sim N(0, 1^2)$.

Xi'an Jiaotong-Liverpool University 西交利物浦大学

Writing $z_{\alpha/2}$ for the z-value above which we find an area of $\alpha/2$ under the standard normal curve, we consider

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$$

where $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

Rearranging, we have

$$P\left((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha.$$

Based on this, the $100(1 - \alpha)\%$ confidence interval can be calculated.

If $\bar{x}_1$ and $\bar{x}_2$ are means of independent random samples of sizes $n_1 \geq 30$ and $n_2 \geq 30$ from populations with variances σ_1^2 and σ_2^2, respectively, a $100(1 - \alpha)\%$ C.I. for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

Example 5

A study was conducted in which two types of engines, A and B, were compared. Gas mileage, in miles per gallon, was measured. Fifty experiments were conducted using engine type A and 75 experiments were done with engine type B. The gasoline used and other conditions were held constant. The average gas mileage was 36 miles per gallon for engine A and 42 miles per gallon for engine B. Find a 96% confidence interval on $\mu_B - \mu_A$, where μ_A and μ_B are population mean gas mileage for engines A and B respectively. Assume the population standard deviations are 6 and 8 for engines A and B respectively.

Solution

The point estimate for $\mu_B - \mu_A$ is $\bar{x}_B - \bar{x}_A = 42 - 36 = 6$. The sample sizes are large enough for CLT to be used. From statistical table, $z_{0.02} = 2.05$. Hence, the 96% C.I. is

$$6 - 2.05\sqrt{\frac{64}{75} + \frac{36}{50}} < \mu_B - \mu_A < 6 + 2.05\sqrt{\frac{64}{75} + \frac{36}{50}}$$

$\Rightarrow 3.43 < \mu_B - \mu_A < 8.57$. ■

Case 2: Confidence Interval for $\mu_1 - \mu_2$, $\sigma_1^2 = \sigma_2^2$ but both unknown

Case 1 cannot be used when sample sizes n_1 and n_2 are small, i.e. less than 30. This is because we cannot invoke CLT to obtain the sampling distribution of $\bar{X}_1 - \bar{X}_2$. However, if both populations are normal, the Student's t-distribution can be used to compute a C.I. for $\mu_1 - \mu_2$.

To show this, we consider both populations to be normal and $\sigma_1^2 = \sigma_2^2$ but both unknown. Let $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Then

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}} \sim N(0,1^2)\ .$$

Furthermore, the two random variables

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_1-1}$$

and

$$\frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

Since these random variables are independent, then the sum

$$V = \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

Since Z and V are independent, then the statistic

$$T = \frac{Z}{\sqrt{V/(n_1+n_2-2)}} = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma^2[(1/n_1)+(1/n_2)]}} \Big/ \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}} \sim t_{n_1+n_2-2}$$

This looks messy. So we denote

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

to be the pooled estimator. Simplifying the expression for T, we obtain

$$T = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{(1/n_1)+(1/n_2)}} \sim t_{n_1+n_2-2}.$$

Using this T statistic, we have

$$P\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) = 1-\alpha,$$

where $t_{\alpha/2}$ is the t-value with n_1+n_2-2 degrees of freedom, above which we find an area of $\alpha/2$. Substituting for T in the inequality and rearranging, the $100(1-\alpha)\%$ C.I. for $\mu_1-\mu_2$ is obtained.

If $\bar{x}_1$ and $\bar{x}_2$ are respectively the means of small independent random samples of sizes n_1 and n_2 (where $n_1, n_2 < 30$), from normal populations with unknown by equal variances, a $100(1-\alpha)\%$ C.I. for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where s_p is the pooled estimate of the population standard deviation and $t_{\alpha/2}$ is the t-value with $n_1 + n_2 - 2$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

Note

1. The value of s_p^2 can be seen as a weighted average of the two sample variances s_1^2 and s_2^2, where the weights are the degrees of freedom.
2. The T statistic is used when sample sizes $n_1, n_2 < 30$, populations are normal **and** population variances σ_1^2, σ_2^2 are unknown.
3. How to interpret the C.I.? If we can conclude $\mu_1 - \mu_2 > 0$ with high confidence (based on positive values for both bounds), then it is reasonable to infer that $\mu_1 > \mu_2$ with little risk of being in error, based on the observations.

Case 3: Confidence Interval for $\boldsymbol{\mu_1 - \mu_2}$, $\boldsymbol{\sigma_1^2 \neq \sigma_2^2}$ and both unknown

We consider the situation when the unknown population variances are not likely to be equal.

The statistic used is

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}},$$

which has a t-distribution with v degrees of freedom, where

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}.$$

Since v is seldom an integer, we **round it down** to the nearest whole number.

Using the statistic T', we write $P\left(-t_{\alpha/2} < T' < t_{\alpha/2}\right) \approx 1 - \alpha$, where $t_{\alpha/2}$ is the value of the t-distribution with v degrees of freedom, above which we find an area of $\alpha/2$. Substituting T' and rearranging, the $100(1 - \alpha)\%$ C.I. can be found.

If $\bar{x}_1, s_1^2$ and $\bar{x}_2, s_2^2$ are respectively the means and variances of small independent random samples of sizes n_1 and n_2 (where $n_1, n_2 < 30$) from normal populations with unknown and unequal variances, a $100(1-\alpha)\%$ C.I. for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

where $t_{\alpha/2}$ is the t-value with

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{[(s_1^2/n_1)^2/(n_1-1)] + [(s_2^2/n_2)^2/(n_2-1)]}$$

degrees of freedom (rounded down), leaving an area of $\alpha/2$ to the right.

Example 7

A study on pollution was conducted to estimate the difference in the amounts of a chemical at two different stations on a river. Fifteen samples were collected from station 1, and 12 samples were obtained from station 2. The 15 samples from station 1 has an average chemical content of 3.84 mg/l and a standard deviation of 3.07 mg/l, while the 12 samples from station 2 had an average content of 1.49 mg/l and a standard deviation of 0.80 mg/l.

Find a 95% C.I. for the difference in the average chemical content at these two stations, assuming that the observations came from normal populations with different variances.

Solution

Station 1: $\bar{x}_1 = 3.84, s_1 = 3.07, n_1 = 15$

Station 2: $\bar{x}_2 = 1.49, s_2 = 0.80, n_2 = 12$

Since the population variances are unequal, we find an approximate 95% C.I. based on t-distribution with degrees of freedom

$$v = \frac{(3.07^2/15 + 0.80^2/12)^2}{[(3.07^2/15)^2/14] + [(0.80^2/12)^2/11]} = 16.3 \approx 16$$

Our point estimate of $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2 = 3.84 - 1.49 = 2.35$.

From statistical table, $t_{0.025} = 2.120$ for $v = 16$ degrees of freedom.

Therefore, the 95% C.I. for $\mu_1 - \mu_2$ is

$$2.35 - 2.120\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}} < \mu_1 - \mu_2 < 2.35 + 2.120\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}}$$

$$\Rightarrow 0.60 < \mu_1 - \mu_2 < 4.10$$

■

5.2.4 Single Sample: Estimating σ^2

The statistic S^2 (of a random sample of size n) is an unbiased estimator of population variance σ^2. The value of S^2 is used as a point estimate of σ^2.

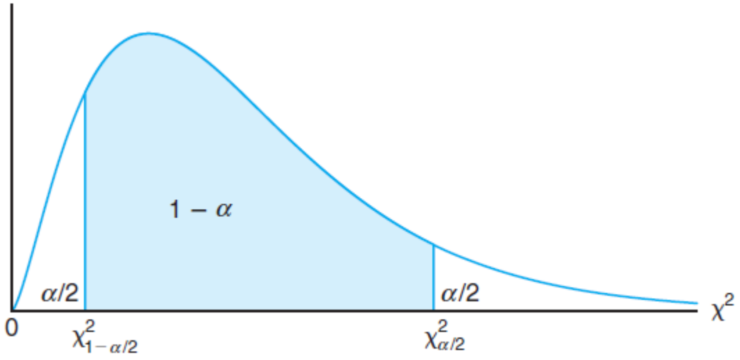
An interval estimate of σ^2 is established by using the statistic

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

which has chi-squared distribution with $n-1$ degrees of freedom when the samples are chosen from a normal population. We would like to have

$$P\left(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}\right) = 1-\alpha,$$

where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are values of chi-squared distribution with $n-1$ degrees of freedom, leaving areas of $1-\alpha/2$ and $\alpha/2$, respectively, to the right.



Substituting for X^2 and rearranging, we have

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha.$$

Based on this expression, the $100(1-\alpha)\%$ C.I. for σ^2 is obtained below.

If s^2 is the variance of a random sample of size n from a normal population, a $100(1-\alpha)\%$ C.I. for σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are χ^2-values with $n-1$ degrees of freedom, leaving areas of $\alpha/2$ and $1-\alpha/2$, respectively, to the right.

Note

An approximate $100(1-\alpha)\%$ C.I. for σ is obtained by taking the square root of the lower and upper bounds of the interval for σ^2.

Example 9

The following are the weights, in grams, or 10 samples of chemicals sold by a company: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2 and 46.0 . Find a 95% C.I. for the variance of the weights of all such chemical products, assuming that they are normally distributed.

Solution

We find

$$s^2 = \frac{n\sum x_i^2 - (\sum x_i)^2}{n(n-1)} = \frac{(10)(21273.12) - (461.2)^2}{(10)(9)} = 0.286$$

From statistical table, we find $\chi^2_{0.025} = 19.023$ and $\chi^2_{0.975} = 2.700$ with 9 degrees of freedom. Therefore the 95% C.I. for σ^2 is

$$\frac{(9)(0.286)}{19.023} < \sigma^2 < \frac{(9)(0.286)}{2.700}$$

$$\Rightarrow 0.135 < \sigma^2 < 0.953$$

■

5.2.5 Two Samples: Estimating σ_1^2/σ_2^2

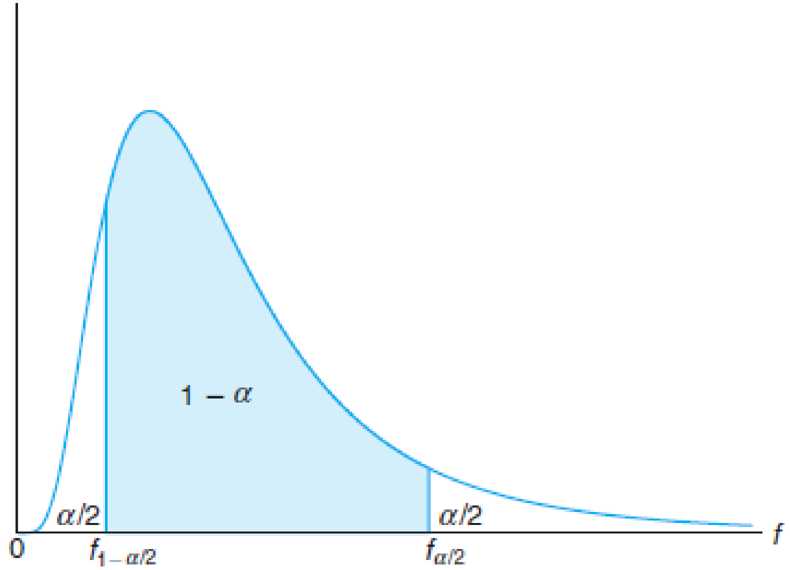
A point estimate of σ_1^2/σ_2^2 is given by the ratio s_1^2/s_2^2 of the sample variances. If σ_1^2 and σ_2^2 are the variances of normal populations, then we have

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \sim F_{n_1-1,n_2-1}.$$

Therefore we write

$$P\left[f_{1-\alpha/2}(n_1-1,n_2-1) < F < f_{\alpha/2}(n_1-1,n_2-1)\right] = 1-\alpha,$$

where $f_{1-\alpha/2}(n_1-1,n_2-1)$ and $f_{\alpha/2}(v_1,v_2)$ are the values of the F-distribution with degrees of freedom n_1-1 and n_2-1, leaving areas of $1-\alpha/2$ and $\alpha/2$, respectively, to the right.



Substituting for F and rearranging, we have

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(n_1-1,n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{1-\alpha/2}(n_1-1,n_2-1)}\right] = 1-\alpha$$

Furthermore, we have learnt that

$$f_{1-\alpha/2}(n_1-1,n_2-1) = \frac{1}{f_{\alpha/2}(n_2-1,n_1-1)},$$

therefore

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(n_1-1,n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}f_{\alpha/2}(n_2-1,n_1-1)\right] = 1-\alpha.$$

Hence for any two independent random samples of sizes n_1 and n_2 selected from two normal populations, the ratio of the sample variances $\frac{s_1^2}{s_2^2}$ is computed, and the following $100(1-\alpha)\%$ C.I. for $\frac{\sigma_1^2}{\sigma_2^2}$ is obtained.

If s_1^2 and s_2^2 are the variances of independent samples of sizes n_1 and n_2, respectively, from normal populations, then a $100(1-\alpha)\%$ C.I. for σ_1^2/σ_2^2 is

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(n_1-1,n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(n_2-1,n_1-1)$$

where $f_{\alpha/2}(v_1,v_2)$ is an f-value with v_1 and v_2 degrees of freedom, leaving an area of $\alpha/2$ to the right, and $f_{\alpha/2}(v_2,v_1)$ is a similar f-value with v_2 and v_1 degrees of freedom.

Note

A $100(1-\alpha)\%$ confidence interval for σ_1/σ_2 is obtained by taking the square root of each endpoint of the interval for σ_1^2/σ_2^2.

Example 10

In Example 7, a 95% C.I. was constructed for the difference in the average chemical content at two stations, assuming that the observations came from normal populations with unequal variances. Justify this assumption by constructing 98% C.I. for σ_1^2/σ_2^2, where σ_1^2 and σ_2^2 are the variance of populations of chemical contents at station 1 and station 2, respectively.

Solution

Station 1: $\bar{x}_1 = 3.84, s_1 = 3.07, n_1 = 15$

Station 2: $\bar{x}_2 = 1.49, s_2 = 0.80, n_2 = 12$

From statistical table, $f_{0.01}(14{,}11) \approx 4.30$ and $f_{0.01}(11{,}14) \approx 3.87$. Therefore, the 98% C.I. for σ_1^2/σ_2^2 is

$$\left(\frac{3.07^2}{0.80^2}\right)\left(\frac{1}{4.30}\right) < \frac{\sigma_1^2}{\sigma_2^2} < \left(\frac{3.07^2}{0.80^2}\right)(3.87)$$

$$\Rightarrow 3.425 < \frac{\sigma_1^2}{\sigma_2^2} < 56.991$$

Since this interval does not contain 1, it is reasonable to conclude based on the data that the two population variances are unequal, subject to statistical error. ■

5.3 Summary

- How to construct confidence intervals.
- Knowledge of the significance of confidence intervals
- Ability to use confidence intervals to draw conclusions

Chapter 6: One- and Two-Sample Tests of Hypotheses

6.1 Statistical Hypotheses: General Concepts

6.1.1 The Null and Alternative Hypotheses

6.1.2 One- and Two-Tailed Test

6.1.3 Steps in Performing a Hypothesis Test Using p-Value

6.1.4 Steps in Performing a Hypothesis Test Using Significance Level α

6.4 One Sample: Tests on a Mean

6.5 Two Samples: Tests on Two Means

6.6 One- Sample and Two-Sample: Tests on Variances

6.7 Goodness-of-Fit Test

6.8 Test for Independence (Categorical Data)

6.9 Type I and Type II Error

6.9.1 Power of a Test

6.10 Summary

Chapter 6: One- and Two-Sample Tests of Hypotheses

Often we are not into the estimation of a population parameter, as discussed in Chapter Two.

Instead, we would use a data-based decision procedure to make conclusions.

Examples:

1. To use experimental evidence to decide if drinking coffee increases the risk of cancer in humans
2. To decide on the basis of sample data whether there is performance difference between two types of machines
3. To conclude using data collected as to whether a person’s blood type and eye colour are independent

In each case, we state an assertion or conjecture about one or more populations.

This chapter is about establishing and testing hypotheses about the population of interest.

6.1 Statistical Hypotheses: General Concepts

- Reading task: Moore-McCabe-Craig Chapter ***6.2-3***

Definition

A statistical hypothesis is an assertion or conjecture concerning one or more populations.

We can never know the truth or falsity of a statistical hypothesis unless we examine the entire population. Instead, we take a random sample from this population and use it to provide evidence that either supports or does not support the hypothesis.

The hypothesis is a statement about the unknown population parameters.

Evidence from sample that is inconsistent with the stated hypothesis leads to a *rejection* of the hypothesis.

6.1.1 The Null and Alternative Hypotheses

In hypothesis testing, we have the **null hypothesis** H_0 which is some hypothesis we wish to *find evidence against.* Rejection of H_0 leads to the acceptance of an **alternative hypothesis** H_1.

The null hypothesis H_0 states that a population parameter is equal to a value θ_0. The alternative Hypothesis H_1 states that the population parameter is different from the value θ_0.

In all hypothesis tests, we will arrive at one of the two following conclusions:

1) **reject $\boldsymbol{H_0}$** in favor of H_1 because of sufficient evidence in the data, or
2) **fail to reject $\boldsymbol{H_0}$** because of insufficient evidence

Note The conclusions do not say "*accept* H_0**".** Failing to find evidence against H_0 means only that the data are consistent with H_0, not that we have clear evidence that H_0 is true. so we can only reject or fail to reject H_0.

Example 1

The concept of hypothesis testing is practised in a court. In a jury trial, the hypotheses are:

$$H_0: \text{ defendent is innocent}$$
$$H_1: \text{ defendent is guilty}$$

The defendant is called into court because of the suspicion of guilt.
The hypothesis H_0 (status quo) stands in opposition to H_1 and is upheld, unless there is evidence “beyond a reasonable doubt”.

Note that “failure to reject H_0” in this case **does not imply innocence**; instead it means the evidence is insufficient to overturn the status quo that the defendant is innocent. The rational is that the legal system would rather to allow a guilty person to get away unconvicted than to have an innocent person convicted. ■

6.1.2 One- and Two-Tailed Test

Any statistical hypothesis where the alternative is **one sided**, e.g.

$$\begin{cases} H_0: \ \theta = \theta_0 \\ H_1: \ \theta > \theta_0 \end{cases}, \qquad \text{or}$$

$$\begin{cases} H_0: \ \theta = \theta_0 \\ H_1: \ \theta < \theta_0 \end{cases}$$

is called a **one-tailed test**.

Any statistical hypothesis where the alternative is **two sided**, e.g.

$$\begin{cases} H_0: \ \theta = \theta_0 \\ H_1: \ \theta \neq \theta_0 \end{cases}$$

is called a **two-tailed test**.

Note

For **one-tailed tests**, they are really

$$\begin{cases} H_0\text{:}\;\; \theta \le \theta_0 \\ H_1\text{:}\;\; \theta > \theta_0 \end{cases}, \qquad \text{or}$$

$$\begin{cases} H_0\text{:}\;\; \theta \ge \theta_0 \\ H_1\text{:}\;\; \theta < \theta_0 \end{cases}$$

but we write H_0 as equality, i.e. $\theta = \theta_0$. This is because under H_0 we choose θ to be as close as possible to the value of the same parameter under H_1, i.e. we look at the p-value for the most conservative case where the sampling distributions under H_0 and H_1 are closest, i.e .you make the biggest error in hypothesis testing.

Example 2

A manufacturer of a certain brand of rice claims that the average saturated fat content does not exceed 1.5 mg per serving. State the null and alternative hypotheses to be used in testing this claim. Is the test one-tailed or two-tailed?

Solution

The manufacturer’s claim should be rejected only if μ is greater than 1.5 mg and should not be rejected if μ is less than or equal to 1.5 mg. We test

$$\begin{cases} H_0: \ \ \mu = 1.5 \\ H_1: \ \ \mu > 1.5 \end{cases}$$

Non-rejection of H_0 does not rule out values less than 1.5 mg. This is a one-tailed test. ■

Note The importance of rejecting H_0 is not that we reject the single value $\mu = 1.5$, but that we reject all values $\mu \leq 1.5$.

6.1.3 Steps in Performing a Hypothesis Test Using p-Value

1. Define H_0 and H_1.

2. Assume H_0 to be true.

3. Compute a **test statistic**. A test statistic is a statistic that measures the compatibility between H_0 and data. It contains information about the disagreement between H_0 and data.

4. Compute the ***p*-value** of the test statistic. The p-value is the probability, assuming H_0 to be true, that the test statistic would take a value as extreme or more extreme than that actually observed. Put simply, the p-value quantifies the evidence against H_0.

5. State a conclusion about the strength of the evidence against H_0:

$p > 0.1$	No evidence against the null hypothesis.
$0.05 < p \leq 0.1$	Weak evidence against the null hypothesis
$0.01 < p \leq 0.05$	Moderate evidence against the null hypothesis
$0.001 < p \leq 0.01$	Strong evidence against the null hypothesis
$p \leq 0.001$	Very strong evidence against the null hypothesis

Note

1. The smaller the p-value, the stronger the evidence is against H_0.
2. In the table, each interval is arbitrary and another person might use something different.
3. For an illustration, we shall show how the p-value is used to test the hypothesis about a population mean.

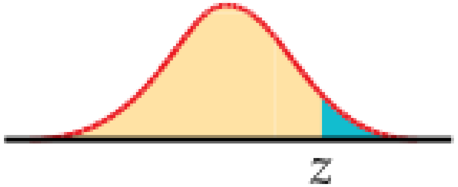
From a population with unknown mean μ and known variance σ^2, suppose a random sample $x_1, x_2, \cdots, x_n$ is obtained. The test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

Case 1

$$H_0: \ \mu = \mu_0$$
$$H_1: \mu > \mu_0$$

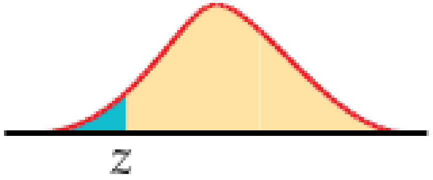
The corresponding p-value is $P(Z > z)$.



.

Case 2

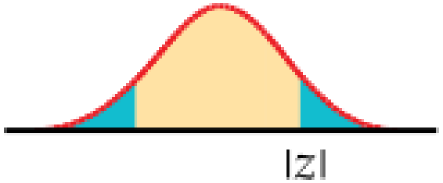
$$H_0: \ \mu = \mu_0$$
$$H_1: \mu < \mu_0$$



The corresponding p-value is $P(Z < z)$.

Case 3

$$H_0: \ \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$



The corresponding p-value is $2P(Z > |z|)$.

The location of the shaded area depends on the nature of inequality sign in the alternative hypothesis H_1.

You use the same strategy in one-sample and two-sample tests, i.e. obtain the required p-value based on the test statistic and alternative hypothesis H_1.

Statistical Significance of the p-Value

Whenever the *p*-value is less than a particular threshold, the result is said to be “statistically significant” at that level. For example, if $p \leq 0.05$, the result is statistically significant at the 5% level; if $p \leq 0.01$, the result is statistically significant at the 1% level, and so on.

If a result is statistically significant at the $100\alpha\%$ level, we can also say that the H_0 is “rejected at level $100\alpha\%$.”

6.1.4 Steps in Performing a Hypothesis Test Using Significance Level α

We shall introduce another way to carry out hypothesis testing using a fixed significance level α.

1. Define H_0 and H_1.

2. Assume H_0 to be true.

3. Compute a **test statistic**. A test statistic is a statistic that measures the compatibility between H_0 and data. It contains information about the disagreement between H_0 and data.

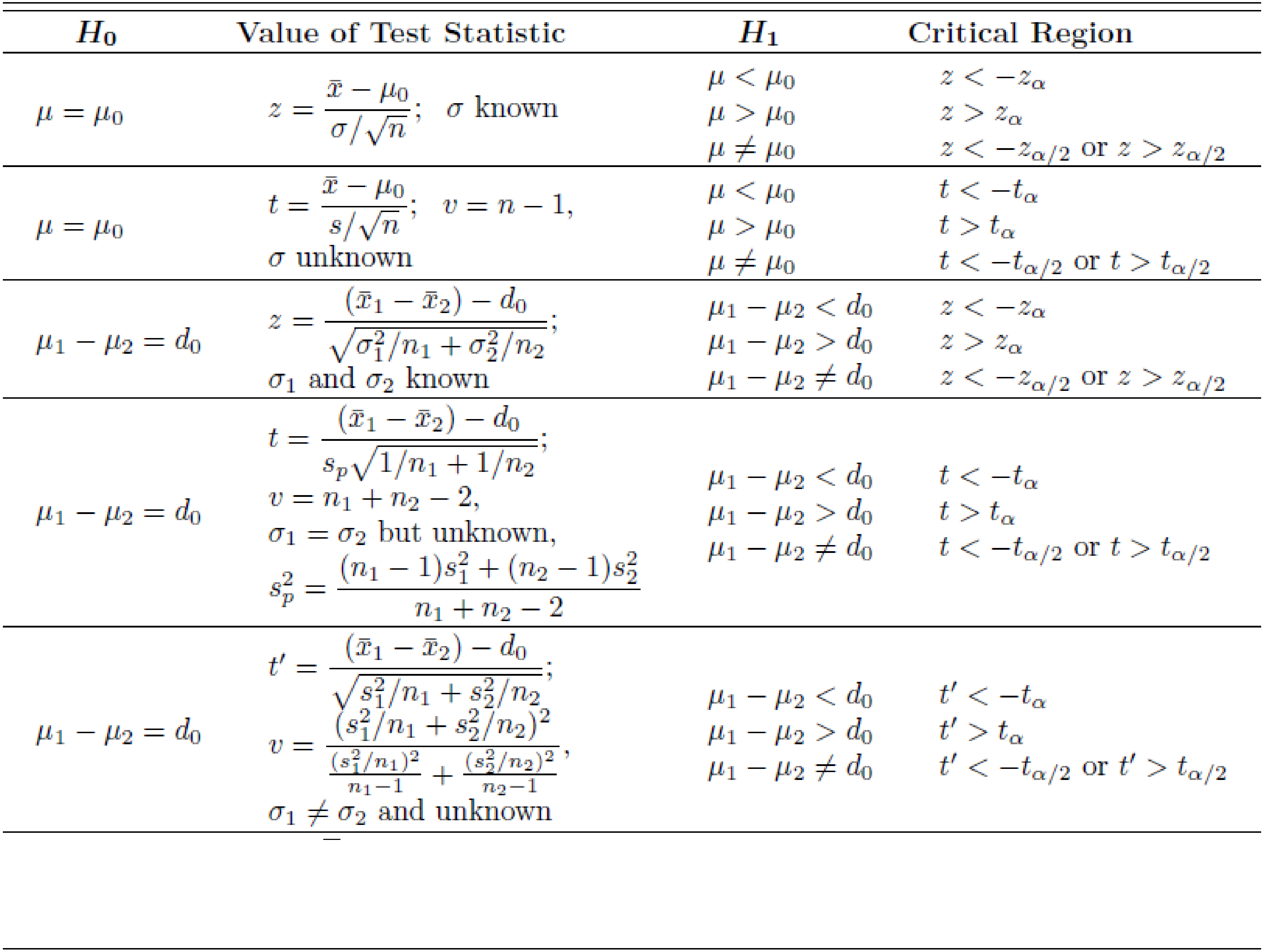
4. Choose (or you may possibly be given) a fixed significance level α which establishes the critical region based on H_1

5. Reject H_0 if the computed test statistic is in the critical region. Otherwise, do not reject H_0.

Note

This approach gives either a "reject H_0" or a "do not reject H_0" conclusion. In contrast, the approach using the p-value requires the subjective judgment to arrive at a conclusion.

Tests Concerning Means



H_0	Value of Test Statistic	H_1	Critical Region
$\mu = \mu_0$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ known	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n - 1$, σ unknown	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$, $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ or $t' > t_{\alpha/2}$

6.4 One Sample: Tests on a Mean

Reading task: Moore-McCabe-Craig Chapter ***7.1***

We know that the hypothesis test on a single sample differs based on

1) Variance σ^2 is known
2) Variance σ^2 is unknown

Remember that the inequality sign in H_1 will indicate the location of the rejection region.

Example 3

A random sample of 100 recorded deaths during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years?

1) Use the p-value in your conclusion.
2) Use the 0.05 significance level in your conclusion.

Solution

Given: $\bar{x} = 71.8, \sigma = 8.9, n = 100, \alpha = 0.05$

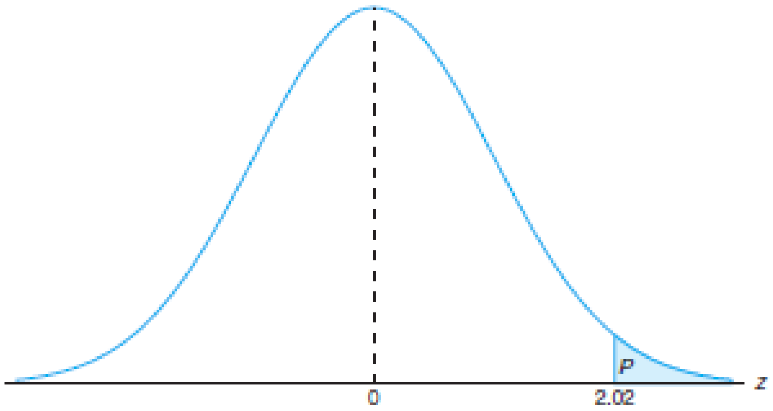
We have

$$H_0: \ \mu = 70$$
$$H_1: \mu > 70$$

Test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{71.8 - 70}{8.9/\sqrt{100}} = 2.02$$

1) Corresponding p-value is $P(Z > 2.02) = 0.0217$.



There is moderate evidence against H_0 in favor of H_1. Hence there is indication from the data that the mean life span today is significantly greater than 70 years.

2) Critical region is $z > 1.645$ where $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.

Since test statistic $z = 2.02$ is in critical region, therefore we reject H_0 at 0.05 level of significance. We conclude based on our data that the mean life span today is greater than 70 years at 0.05 level of significance. ■

Example 4

A manufacturer of sports equipment has developed a new synthetic fishing line that the company claims has a mean breaking strength of 8 kg with a standard deviation of 0.5 kg. Test the hypothesis that $\mu = 8$ kg against the alternative that $\mu \neq 8$ kg if a sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kg.

1) Use the p-value in your conclusion.
2) Use a 0.01 significance level in your conclusion.

Solution

Given: $\bar{x} = 7.8, \sigma = 0.5, n = 50, \alpha = 0.01$

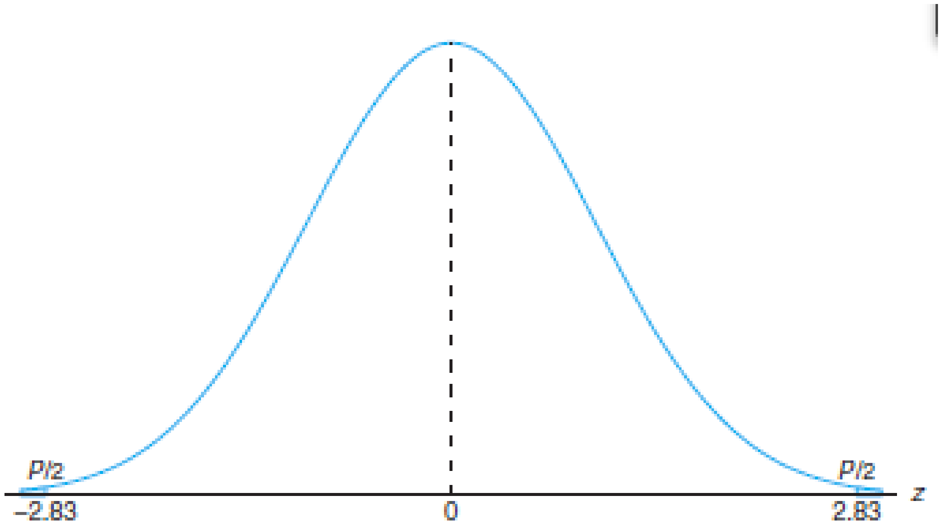
We have

$$H_0\colon\ \mu = 8$$
$$H_1\colon \mu \neq 8$$

Test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{7.8 - 8}{0.5/\sqrt{50}} = -2.83$$

1) Corresponding p-value is $2P(Z < -2.83) = 0.0046$.



There is strong evidence against H_0 in favor of H_1. Hence we conclude based on the observations that the mean breaking strength is significantly not 8 kg.

2) Critical region is $z < -2.575$ or $z > 2.575$ where $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.

Since test statistic $z = -2.83$ is in critical region, therefore we reject H_0 at 0.01 level of significance. We conclude based on our observations that the mean breaking strength is not 8 kg at 0.01 level of significance. ■

Example 5

An electric company has published figures on the number of kilowatt hours used annually by various home appliances. It is claimed that a vacuum cleaner uses an average of 46 kilowatt hours per year. If a random sample of 12 homes included in a planned study indicates that vacuum cleaners use an average of 42 kilowatt hours per year with a standard deviation of 11.9 kilowatt hours, does this suggest that vacuum cleaners use, on average, less than 46 kilowatt hours annually? Assume the population of kilowatt hours to be normal.

1) Use the p-value in your conclusion.
2) Use a 0.05 significance level in your conclusion.

Solution

Given: $\bar{x} = 42, s = 11.9, n = 12, \alpha = 0.05$

We have

$$H_0\text{:}\;\; \mu = 46$$
$$H_1\text{:}\; \mu < 46$$

Test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16$$

1) Corresponding p-value is $P(T < -1.16) \approx 0.135$.

There is no evidence against H_0 in favor of H_1. Hence we conclude based on the observations that the average kilowatt hours used annually by home vacuum cleaners is not significantly less than 46.

2) Critical region is $t < -1.796$ where $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ with 11 degrees of freedom.

Since test statistic $t = -1.16$ is not in critical region, therefore we do not reject H_0 at 0.05 level of significance. We conclude based on our observations that the average kilowatt hours used annually by home vacuum cleaners is not less than 46 at 0.05 level of significance. ■

6.5 Two Samples: Tests on Two Means

Reading task: Moore-McCabe-Craig Chapter ***7.2***

Here we shall consider two independent random samples of sizes n_1 and n_2 respectively drawn from two populations with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2. We assume that n_1 and n_2 are sufficiently large that the Central Limit Theorem applies. Therefore

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1^2).$$

This statistic serves as a starting point for the development of the test procedures involving two means.

The null hypothesis (for all cases) is

$$H_0\colon\ \mu_1 - \mu_2 = d_0$$

and the alternative hypothesis is

$$\begin{cases} H_1\colon \mu_1 - \mu_2 > d_0, \text{or} \\ H_1\colon \mu_1 - \mu_2 \neq d_0, \text{or} \\ H_1\colon \mu_1 - \mu_2 < d_0 \end{cases}$$

Example 6

An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear, Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average wear of 85 units with a standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample deviation of 5.

Can we conclude that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units? Assume the population to be normally distributed with equal variances.

1) Use the p-value in your conclusion.
2) Use a 0.05 significance level in your conclusion.

Solution

Given: $\bar{x}_1 = 85, \bar{x}_2 = 81, s_1 = 4, s_2 = 5, n_1 = 12, n_2 = 10, \alpha = 0.05$

We have

$$H_0:\ \mu_1 - \mu_2 = d_0 = 2$$
$$H_1:\ \mu_1 - \mu_2 = d_0 > 2$$

Test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ with } n_1 + n_2 - 2 \text{ degrees of freedom.}$$

We find

$$s_p = \sqrt{\frac{(11)(16) + (9)(25)}{12 + 10 - 2}} = 4.478$$

$$t = \frac{(85 - 81) - 2}{4.478\sqrt{\frac{1}{12} + \frac{1}{10}}} = 1.04 \text{ with } 20 \text{ degrees of freedom}$$

1) Corresponding p-value is $P(T > 1.04) \approx 0.16$.

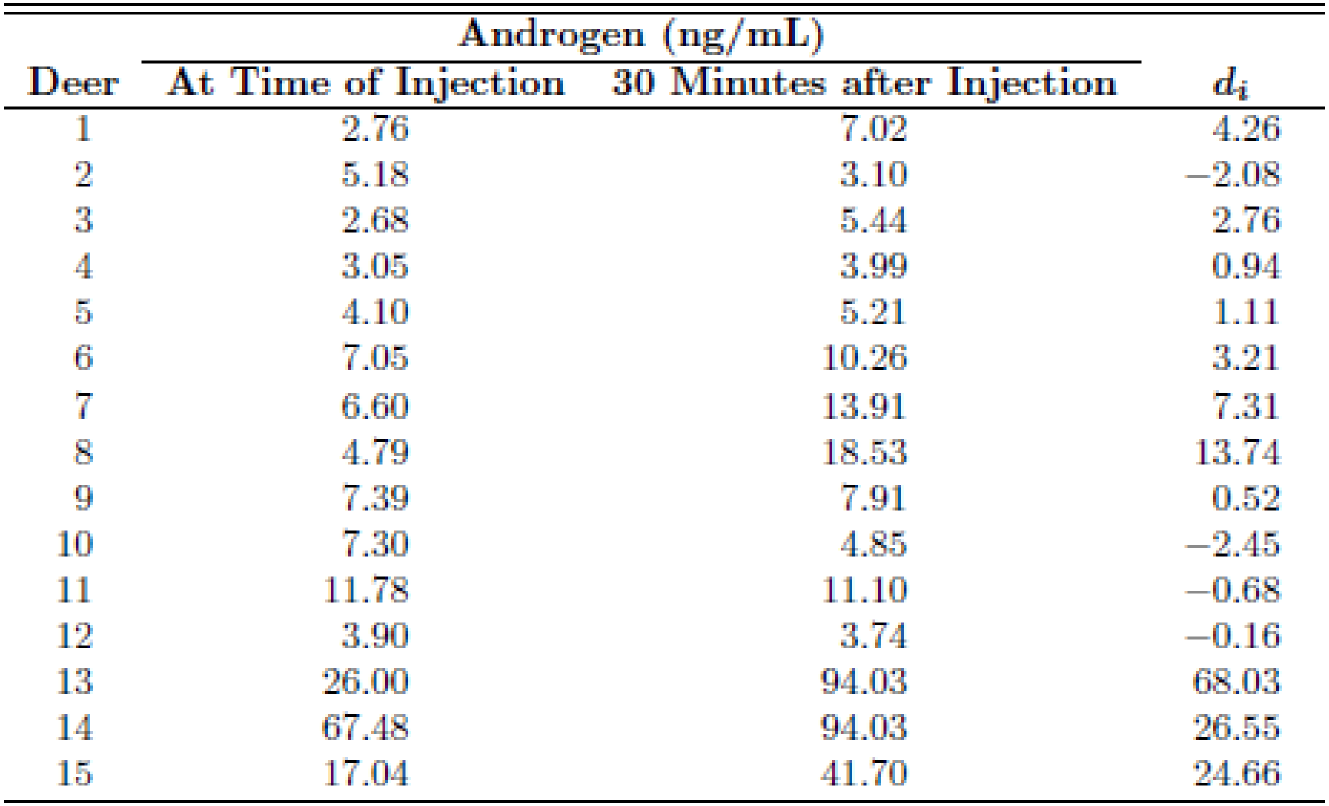
There is no evidence against H_0 in favor of H_1. Hence we conclude based on the observations that it is not significant for the abrasive wear of material 1 to exceed that of material 2 by more than 2 units. .

2) Critical region is $t > 1.725$ where $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ with 20 degrees of freedom.

Since test statistic $t = 1.04$ is not in critical region, therefore we do not reject H_0 at 0.05 level of significance. We conclude based on our observations that the abrasive wear of material 1 does not exceed that of material 2 by more than 2 units at 0.05 level of significance. ■

Example 7

The influence of the drug succinylcholine on the circulation levels of androgens in the blood was investigated. Blood samples were taken from wild deer immediately after they received an injection of succinylcholine. A second blood sample was obtained from each deer 30 minutes after the first sample. The levels of androgens (in nanograms per milliliter) are monitored for 15 deer with the results given below.



	Androgen (ng/mL)		
Deer	At Time of Injection	30 Minutes after Injection	d_i
1	2.76	7.02	4.26
2	5.18	3.10	−2.08
3	2.68	5.44	2.76
4	3.05	3.99	0.94
5	4.10	5.21	1.11
6	7.05	10.26	3.21
7	6.60	13.91	7.31
8	4.79	18.53	13.74
9	7.39	7.91	0.52
10	7.30	4.85	−2.45
11	11.78	11.10	−0.68
12	3.90	3.74	−0.16
13	26.00	94.03	68.03
14	67.48	94.03	26.55
15	17.04	41.70	24.66

Assuming that the population of androgen levels at the time of injection and 30 minutes later are normally distributed, test whether the androgen concentrations are altered after 30 minutes.

1) Use the p-value in your conclusion.
2) Use a 0.05 significance level in your conclusion.

Solution

Let μ_1 and μ_2 respectively be the average androgen concentration at the time of injection and 30 minutes later.

We compute sample mean $\bar{d} = 9.848$ and standard deviation $s_d = 18.474$

We have

$$H_0\colon\ d_0 = \mu_1 - \mu_2 = 0$$
$$H_1\colon\ d_0 = \mu_1 - \mu_2 \neq 0$$

Test statistic is

$$t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}} = \frac{9.848 - 0}{18.474/\sqrt{15}} = 2.06 \text{ with } n - 1 = 14 \text{ degrees of freedom.}$$

1) Corresponding p-value is $P(|T| > 2.06) \approx 0.06$.

There is weak evidence against H_0 in favor of H_1. Hence we conclude based on the observations that there is a significant difference in mean circulating levels of androgen.

2) Critical region is $t < -2.145$ or $t > 2.145$ where $t = \frac{\bar{d}-d_0}{s_d/\sqrt{n}}$ with 14 degrees of freedom.

Since test statistic $t = 2.06$ is not in critical region, therefore we do not reject H_0 at 0.05 level of significance. We conclude based on our observations that concentrations are not significantly altered after 30 minutes 0.05 level of significance. ■

6.6 One- Sample and Two-Sample: Tests on Variances

- Reading task: Moore-McCabe-Craig Chapter ***7.3***

In this section, we are concerned with testing hypotheses concerning population variances. For one sample test, we test if the population variance σ^2 takes the value σ_0^2.

H_0	Value of Test Statistic	H_1	Critical Region
$\sigma^2 = \sigma_0^2$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}; \quad v = n-1$	$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi_{1-\alpha}^2$
		$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi_{\alpha}^2$
		$\sigma^2 \neq \sigma_0^2$	$\chi^2 < \chi_{1-\alpha/2}^2$ or $\chi^2 > \chi_{\alpha/2}^2$

For two sample test, we test if two population variances σ_1^2 and σ_2^2 are equal.

H_0	Value of Test Statistic	H_1	Critical Region
$\sigma_1^2 = \sigma_2^2$	$f = \frac{s_1^2}{s_2^2};$ $v_1 = n_1 - 1$ $v_2 = n_2 - 1$	$\sigma_1^2 < \sigma_2^2$	$f < f_{1-\alpha}(v_1, v_2)$
		$\sigma_1^2 > \sigma_2^2$	$f > f_{\alpha}(v_1, v_2)$
		$\sigma_1^2 \neq \sigma_2^2$	$f < f_{1-\alpha/2}(v_1, v_2)$ or $f > f_{\alpha/2}(v_1, v_2)$

Example 8

A manufacturer of car batteries claims that the life of the company's batteries is normally distributed with a standard deviation equal to 0.9 year. If random sample of 10 of these batteries has a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ year?

1) Use the p-value in your conclusion.
2) Use a 0.05 significance level in your conclusion.

Solution

Given: $s = 1.2, n = 10, \alpha = 0.05$

We have

$$H_0:\ \ \sigma^2 = 0.81$$
$$H_1:\ \sigma^2 > 0.81$$

Test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(9)(1.44)}{0.81} = 16.0 \text{ with 9 degrees of freedom.}$$

.

1) Corresponding p-value is $P(\chi^2 > 16.0) \approx 0.07$.

There is weak evidence against H_0 in favor of H_1. Hence we conclude based on the observations that it is significant that $\sigma > 0.9$ year.

2) Critical region is $\chi^2 > 16.919$ where $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$ with 9 degrees of freedom.

Since test statistic $\chi^2 = 16.0$ is not in critical region, therefore we do not reject H_0 at 0.05 level of significance. We conclude based on our observations that it is not significant that $\sigma > 0.9$ at 0.05 level of significance. ■

Example 9

In testing for the difference in the abrasive of two materials, twelve pieces of material 1 and ten pieces of material 2 were sampled. Each piece is exposed to a machine measuring wear. Given the sample variance of material 1 and 2 are $s_1 = 4$ and $s_2 = 5$ respectively, is there evidence to suggest that the two unknown population variances are unequal? Use a 0.10 level of significance.

Solution

Given: $s_1 = 4, s_2 = 5, n_1 = 12, n_2 = 10, \alpha = 0.10$

We have

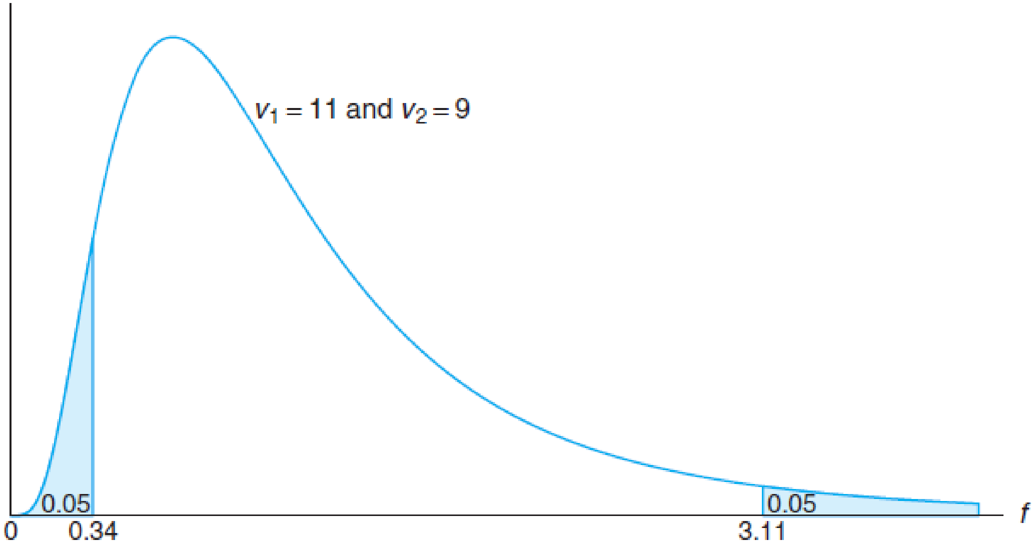
$$H_0:\ \sigma_1^2 = \sigma_2^2$$
$$H_1:\ \sigma_1^2 \neq \sigma_2^2$$

Test statistic is

$$f = \frac{s_1^2}{s_2^2} = \frac{16}{25} = 0.64 \text{ with } 11, 9 \text{ degrees of freedom.}$$

From the diagram below, we see that $f_{0.05}(11,9) = 3.11$. Furthermore,

$$f_{0.95}(11,9) = \frac{1}{f_{0.05}(9,11)} = 0.34$$



Therefore H_0 is rejected when $f < 0.34$ or $f > 3.11$ where $f = \frac{s_1^2}{s_2^2}$ with 11, 9 degrees of freedom.

Since test statistic $f = 0.64$ is not in critical region, therefore we do not reject H_0 at 0.10 level of significance. We conclude based on our observations that it is not significant that $\sigma_1^2 \neq \sigma_2^2$ at 0.10 level of significance. ■

6.7 Goodness-of-Fit Test

- Reading task: Moore-McCabe-Craig Chapter ***9.3***

We shall consider a test to determine whether a frequency distribution fits an expected distribution. The hypothesis test is:

H_0: Frequency distribution fits the specified distribution.
H_1: Frequency distribution does not fit the specified distribution.

A goodness-of-fit test between observed and expected frequencies is based on the statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where χ^2 follows a chi-squared distribution (approximate) with $k-1$ degrees of freedom. Each outcome of an experiment is called a category. There are k categories where o_i and e_i represent the observed and expected frequencies in i^{th} category, $i = 1,2 \cdots, k$. The test is **right-tailed** only.

Note

To apply the Goodness-of-Fit Test, the following must be satisfied.

1. The observed frequencies must be obtained by using a random sample.
2. Each expected frequency must be ≥ 5. **This restriction may require combining of adjacent categories, resulting in a reduction in the number of degrees of freedom.**

Example 10

It is claimed that the tax preparation methods of the population is distributed as follows.

Each outcome is classified into **categories**.

Distribution of tax preparation methods	
Accountant	25%
By hand	20%
Computer software	35%
Friend/family	5%
Tax preparation service	15%

The probability for each possible outcome is fixed.

It randomly selects 300 adults and asks them how they prepare their taxes. The observed frequency for each category is given below.

Survey results (n = 300)	
Accountant	71
By hand	40
Computer software	101
Friend/family	35
Tax preparation service	53

observed frequency

Test the validity of the claim at 0.01 level of significance.

Solution

To test the claim, the following hypotheses is used.

H_0: The distribution of tax preparation methods is 25% by accountant, 20% by hand, 35% by computer software, 5% by friend or family, and 15 % by tax preparation service. (claim)

H_1: The distribution of tax preparation methods differs from the claimed or expected distribution.

The expected frequency of a category is calculated using

$$e_i = np_i$$

where n is the number of trials (the sample size) and p_i is the probability of the i^{th} category.

Tax preparation method	% of people	Observed frequency	Expected frequency
Accountant	25%	71	300(0.25) = 75
By hand	20%	40	300(0.20) = 60
Computer Software	35%	101	300(0.35) = 105
Friend/family	5%	35	300(0.05) = 15
Tax preparation service	15%	53	300(0.15) = 45

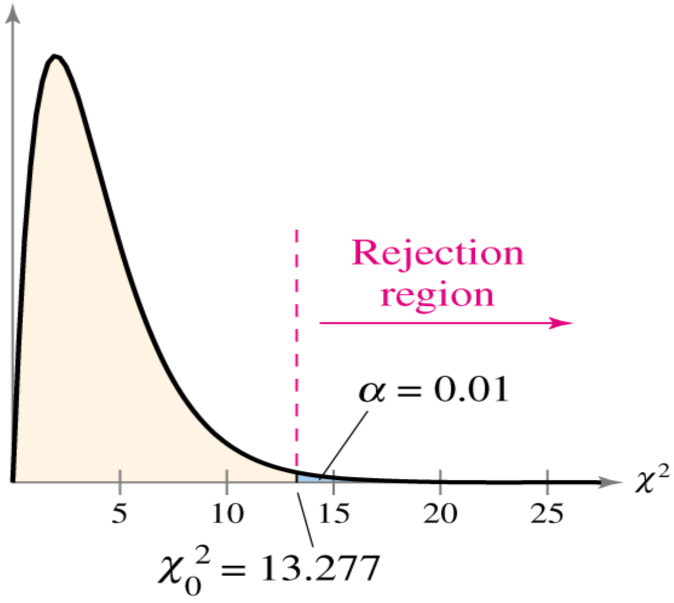
Given: $k = 5, \alpha = 0.01$

Test statistic is

$$\chi^2 = \sum_{i=1}^{5} \frac{(o_i - e_i)^2}{e_i} = \frac{(71-75)^2}{75} + \frac{(40-60)^2}{60} + \frac{(101-105)^2}{105} + \frac{(35-15)^2}{15} + \frac{(53-45)^2}{45}$$

$= 35.121$ with 4 degrees of freedom.

Critical region is $\chi^2 > 13.277$ with 4 degrees of freedom



Since test statistic $\chi^2 = 35.121$ is in critical region, we reject H_0 at 0.01 level of significance. We conclude based on our observations that the actual distribution of tax preparation methods differs from what is claimed at 0.01 level of significance. ■

Example 11

It is claimed that the frequency distribution of battery lives may be approximated by a normal distribution with mean $\mu = 3.5$ and standard deviation $\sigma = 0.7$. Test this claim at 0.05 level of significance.

Frequency Distribution of Battery Lives	
Interval for Battery Lives	Frequency
$1.5 - 1.9$	2
$2.0 - 2.4$	1
$2.5 - 2.9$	4
$3.0 - 3.4$	15
$3.5 - 3.9$	10
$4.0 - 4.4$	5
$4.5 - 4.9$	3

Solution

Given: $k = 7, \alpha = 0.05, n = 40$

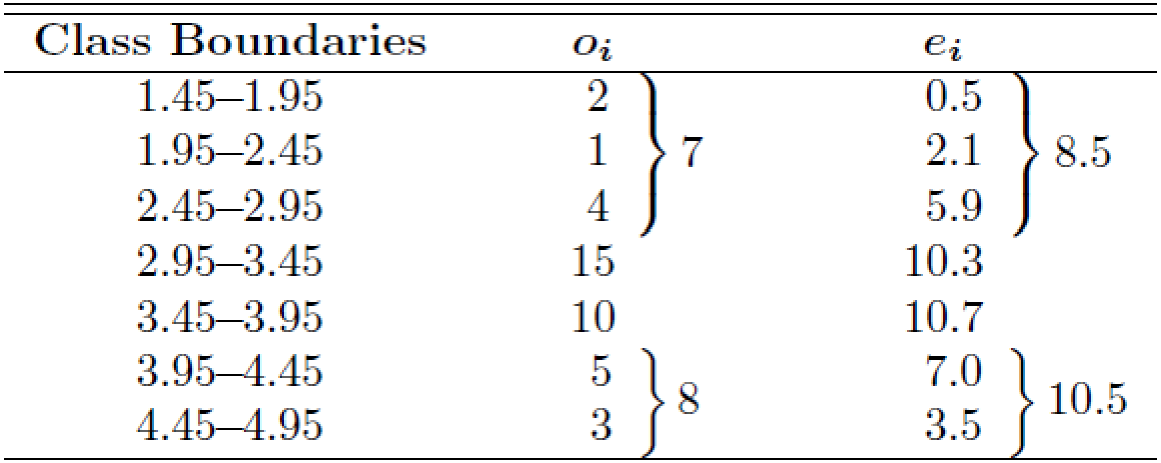
Let the random variable $X \sim N(3.5, 0.7^2)$ be the live of a battery. We define the class boundaries as follows to enable us obtain the corresponding probabilities and hence the expected frequencies.

Class Boundaries	$\boldsymbol{o_i}$	Probability	$\boldsymbol{e_i}$
$1.45 - 1.95$	2	$P(X \leq 1.95) = 0.0125$	0.5
$1.95 - 2.45$	1	$P(1.95 < X \leq 2.45) = 0.0525$	2.1
$2.45 - 2.95$	4	$P(2.45 < X \leq 2.95) = 0.1475$	5.9
$2.95 - 3.45$	15	$P(2.95 < X \leq 3.45) = 0.258$	10.3
$3.45 - 3.95$	10	$P(3.45 < X \leq 3.95) = 0.268$	10.7
$3.45 - 4.45$	5	$P(3.45 < X \leq 4.45) = 0.175$	7.0
$4.45 - 4.95$	3	$P(4.45 < X) = 0.0875$	3.5

Note

1) The first and last probabilities are calculated differently.
2) We do not have ≥ 5 frequencies in all categories

We combine adjacent categories so that the frequencies are all ≥ 5. Consequently the total number of intervals is reduced from 7 to 4, resulting in $v = 3$ degrees of freedom:



Class Boundaries	o_i		e_i	
1.45–1.95	2	7	0.5	8.5
1.95–2.45	1		2.1	
2.45–2.95	4		5.9	
2.95–3.45	15		10.3	
3.45–3.95	10		10.7	
3.95–4.45	5	8	7.0	10.5
4.45–4.95	3		3.5	

The χ^2 test statistic is then

$$\chi^2 = \frac{(7-8.5)^2}{8.5} + \frac{(15-10.3)^2}{10.3} + \frac{(10-10.7)^2}{10.7} + \frac{(8-10.5)^2}{10.5} = 3.05$$

with 3 degrees of freedom.

Critical region is $\chi^2 > 7.815$ with 3 degrees of freedom. Since test statistic $\chi^2 = 3.05$ is not in critical region, therefore we do not reject H_0 at 0.05 level of significance. We conclude based on our observations that the normal distribution $N(3.5, 0.7^2)$ provides good fit for the distribution of battery lives at 0.05 level of significance. ■

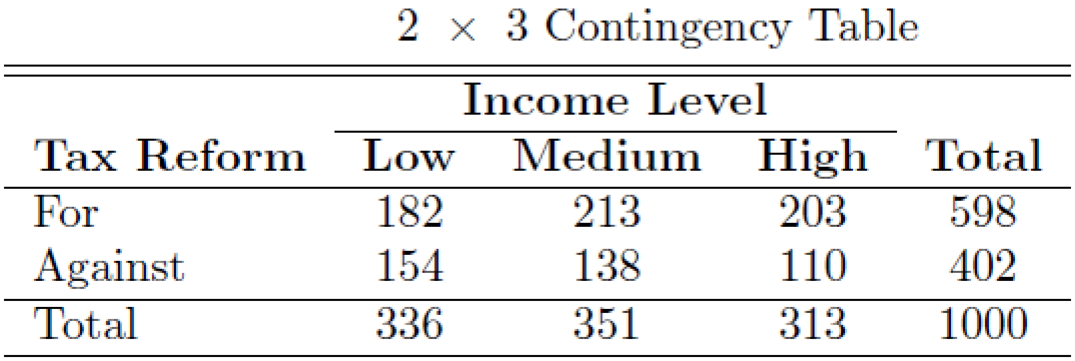
6.8 Test for Independence (Categorical Data)

- Reading task: Moore-McCabe-Craig Chapter ***9.1-2***

The Chi-Squared test procedure can also be used to test the hypothesis of independence of two variables of categorical data. Such data are presented in a $r \times c$ contingency table.

- Shows the observed frequencies for two variables.
- The observed frequencies are arranged in *r* rows and *c* columns.
- The intersection of a row and a column is called a **cell**.

An example of a 2×3 contingency table is given below.



2×3 Contingency Table

	Income Level			
Tax Reform	Low	Medium	High	Total
For	182	213	203	598
Against	154	138	110	402
Total	336	351	313	1000

Note

To apply the Chi-Squared Independence Test, the following must be satisfied.

1. The observed frequencies must be obtained by using a random sample.
2. Each expected frequency must be ≥ 5. **This restriction may require combining of adjacent cells, resulting in a reduction in the number of degrees of freedom.**

The hypothesis test is:

H_0: Independence between two variables
H_1: No independence between two variables.

The Chi-Squared Independence test between observed and expected frequencies is based on the statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where χ^2 follows a chi-squared distribution (approximate) with $(r-1)(c-1)$ degrees of freedom. In the formula, o_i and e_i represent the observed and expected frequencies in i^{th} cell, $i = 1,2 \cdots, rc$.

The general rule for obtaining expected frequency of any cell is

$$\text{expected frequency} = \frac{(\text{row total}) \times (\text{column total})}{\text{grand total}}$$

The hypothesis test is **right-tailed** only.

Example 12

A survey is carried out on 2200 adults to understand their preferred way of eating ice-cream. Using the contingency table below, can you conclude that the preferred ways to eat ice cream are related to gender? Use 0.01 level of significance. The expected frequencies are shown in parentheses.

	Favorite way to eat ice cream					
Gender	Cup	Cone	Sundae	Sandwich	Other	Total
Male	600 (550.91)	288 (342.55)	204 (209.45)	24 (24)	84 (73.09)	1200
Female	410 (459.09)	340 (285.45)	180 (174.55)	20 (20)	50 (60.91)	1000
Total	1010	628	384	44	134	2200

Given: $r = 2, c = 5, \alpha = 0.01$

The hypotheses are

H_0: Independence between two variables

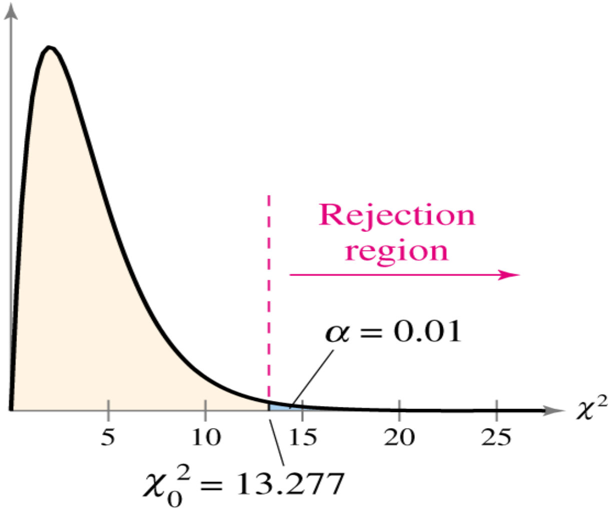
H_1: No independence between two variables

Test statistic is

$$\chi^2 = \sum_{i=1} \frac{(o_i - e_i)^2}{e_i} = \frac{(600 - 550.91)^2}{550.91} + \cdots + \frac{(20 - 20)^2}{20} + \frac{(50 - 60.91)^2}{60.91}$$

$= 32.63$ with $(2 - 1)(5 - 1) = 4$ degrees of freedom.

Critical region is $\chi^2 > 13.277$ with 4 degrees of freedom

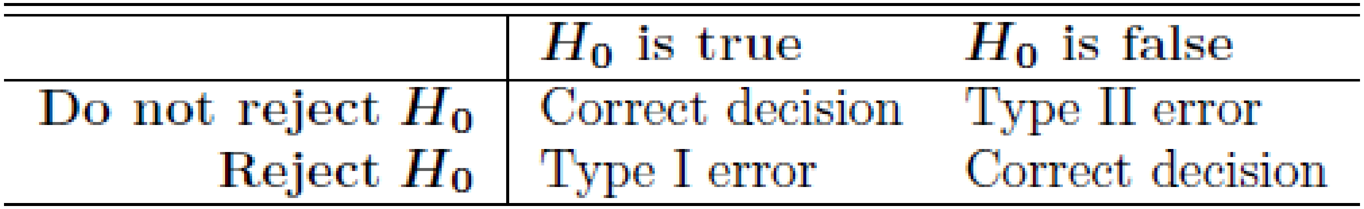


Since test statistic $\chi^2 = 32.630$ is in critical region, we reject H_0 at 0.01 level of significance. We conclude based on our observations that the preferred ways to eat ice cream are related to gender at 0.01 level of significance. ■

6.9 Type I and Type II Error

- Reading task: Moore-McCabe-Craig Chapters ***6.4***

When testing a hypothesis H_0 with level of significance α, there are two ways in which decision can be wrong.



	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

We should try to make the probabilities of Type I and Type II errors reasonably small when designing our experiment. There is no use to conduct an experiment with large probability of leading to an incorrect decision.

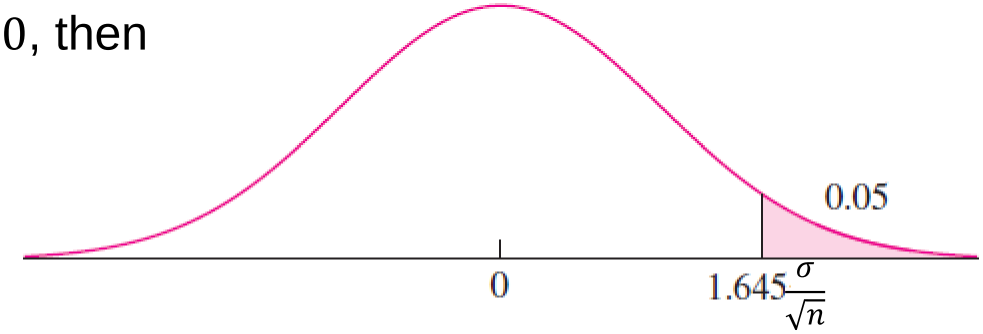
It turns out that we can control the probability of Type I error easily thru α.

If α is the significance level that has been chosen for the test, then the probability of a Type I error is never greater than α.

We shall give a simple illustration. Let $X_1, X_2, \cdots, X_n$ be a random sample from a population with mean μ and variance σ^2. Suppose sample size $n \geq 30$. Then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Assume that we test H_0: $\mu = 0$ against H_1: $\mu > 0$ at $\alpha = 0.05$ level of significance. The sampling distribution of $\bar{X}$ under H_0 is $\bar{X} \sim N\left(0, \frac{\sigma^2}{n}\right)$.

Using $\alpha = 0.05$, the critical region is defined as $\bar{x} \geq 1.645\frac{\sigma}{\sqrt{n}}$.



Assuming H_0 to be true, with $\mu = 0$, then

$P(\text{Type I error})$

$= P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true})$

$= P\left(\bar{X} \geq 1.645\frac{\sigma}{\sqrt{n}}\right)$

$= 0.05$

Next, assuming H_0 to be true, with $\mu < 0$, then the distribution of $\bar{X}$ is obtained by shifting the curve to the left. Therefore

$P(\text{Type I error}) = \left(\bar{X} \geq 1.645\frac{\sigma}{\sqrt{n}}\right) < 0.05$ ■

Therefore ,

we can control Type I error since $P(\text{Type I error})$ is never greater than the significance level α that we choose.

However,

Type I error and Type II error are related. A decrease in the probability of one generally results in the increase in the probability of the other.

The usual strategy is to begin by choosing a value for α so that the probability of a Type I error will be reasonably small. Then one computes the probability of a Type II error and hopes that it is not too large.

When obtaining $P(\text{Type II error})$, we need to assume a “true” mean μ under H_1.

If $P(\text{Type II error})$ is large, we can decrease it by:

1. increasing sample size n
2. increasing significance level α

6.9.1 Power of Test

A hypothesis test results in a Type II error if H_0 is not rejected when it is false.

The **power of a test** (or simply, *power*) is the probability of rejecting H_0 when it is false.

Therefore

$$\boxed{\text{Power} = 1 - P(\text{Type II error})}$$

The power is a succinct measure of how sensitive the test is for detecting differences between a hypothesized mean μ_0 and a “true” mean μ under H_1.

Computing the power involves two steps:

1. Compute the rejection region
2. Compute the probability that the test statistic falls in the rejection region if H_1 is true. This is the power.

In general, test with power greater than 0.80 are considered acceptable.

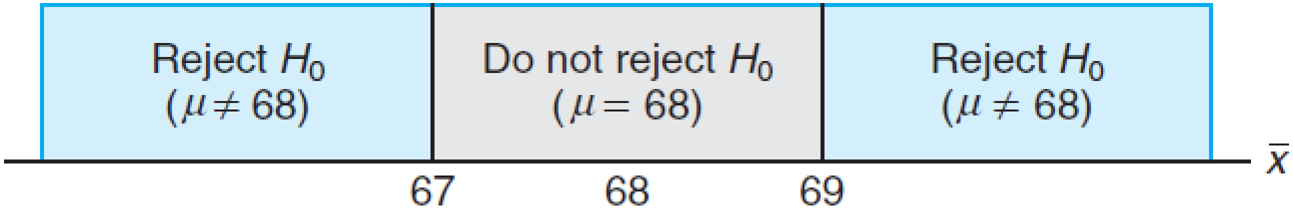
Example 13

Consider the null hypothesis that the average weight of male students in a certain university is 68 kg against the alternative hypothesis that it is unequal to 68 kg. That is, we wish to test

$$\begin{cases} H_0: \mu = 68 \\ H_1: \mu \neq 68 \end{cases}$$

for a population with standard deviation $\sigma = 3.6$.

Suppose a random sample of size $n = 36$ is obtained and the critical region is chosen to be $\bar{x} < 67$ or $\bar{x} > 69$.



1) Obtain the probability of Type I error.

2) Obtain the probability of Type II error if, in fact, the true mean $\mu = 70$ kg.

Solution

1) $P(\text{Type I error})$

$= P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true})$

$= P(\bar{X} < 67$ or $\bar{X} > 69 \text{ when } \mu = 68)$

$= P(\bar{X} < 67 \text{ when } \mu = 68) + P(\bar{X} > 69 \text{ when } \mu = 68)$

$= P(Z < -1.67) + P(Z > 1.67)$

$= 2P(Z < -1.67)$ by symmetry

$= 0.0950$

2) $P(\text{Type II error})$

$= P(\text{not rejecting } H_0 \text{ when } H_1 \text{ is true})$

$= P(67 \le \bar{X} \le 69 \text{ when } \mu = 70)$

$= P(-5 \le Z \le -1.67)$

$= P(Z \le -1.67) - P(Z < -5) \approx 0.0475 - 0 = 0.0475$ ■

Example 14

Assuming $n = 50$ and $\sigma = 5$, find the power of the test of

$$\begin{cases} H_0: \mu = 80 \\ H_1: \mu > 80 \end{cases}$$

for the mean yield of the new process if, in fact, the true $\mu = 82$. Use a 0.05 level of significance.

Solution

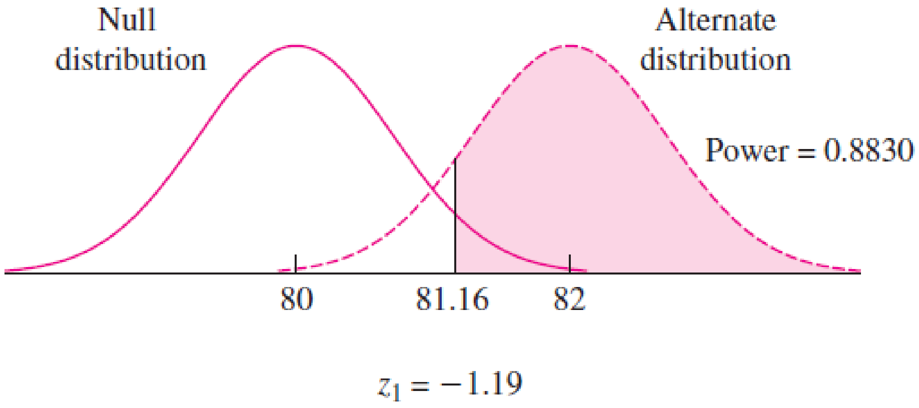
Rejection region is right tail with probability 0.05 . Since $z_{0.05} = 1.645$, so the rejection region (under H_0) is $\bar{x} > 80 + 1.645\frac{\sigma}{\sqrt{n}} = 81.16$, therefore

$$\text{Power} = P(\text{Reject } H_0 \text{ when } H_1 \text{ is true})$$

$$= P(\bar{X} > 81.16 \text{ when } \mu = 82)$$

$$= P\left(Z > \frac{81.16-82}{5/\sqrt{50}} \text{ when } \mu = 82\right)$$

$$= P(Z > -1.19) = 0.8830 \quad \blacksquare$$



6.10 Summary

- How to set up hypotheses
- Defining correct critical region based on H_1
- Using the appropriate hypothesis test
- Use of the p-level and level of significance α in hypothesis testing
- Knowledge of Type I, Type II errors and the power of test.