Statistics

Week 1: Introduction (Chapter 1)

ESD, SUTD

Term 5, 2017

Outline

Introduction

Lecturer: James Wan, james_wan@sutd.edu.sg

TA: Qifang Bao, qfbao@mit.edu, 1.715-S10

Office hour (with either lecturer or TA): by appointment

Read the *Course Description* and *Project Options* on eDimension.

Course structure

Lectures: 2×2 hours weekly, focus on theory, contain some short activities. **Start on time**, with 10 minute break in the middle.

Recitations: 1 hour weekly, focus on applications, contain longer activities.

In weeks 2, 4, 6, 8, 10, 12, there will be an **additional recitation** for the project (Thursdays, 2–3pm, TT21).

Homework sets: contain problems that require software.

We will mainly use *Excel*, and occasionally *R*.

Exams: hand computations and proofs (no software allowed).

Makeup class: Wednesday 1 February, 11am–1pm, TT21.

Course notes

These lecture slides will provide a sufficient outline of the course material. They are *not* meant to be complete or comprehensive, nor will they be uploaded far ahead of time.

To *improve* your study skills, you will need to take your own notes to complement the slides, based on what is covered in class.

For a fuller understanding of the course, you should refer to the text book, *Statistics and Data Analysis: From Elementary to Intermediate*, by Tamhane and Dunlop.

Feel free to use online resources, but stay away from ones that *spoon-feed* you. We are aiming for conceptual understanding, not memorization.

Outline

Data

In everyday life we both generate and receive data, for instance when we access websites, use social media, fill out forms, . . .

We have the need to analyze and interpret data: in daily life, we need to understand what the media gives us; managers and governments need data to *identify patterns* and *make decisions*.

Some people, such as engineers, analysts, scientists, design *models* using data. Models aid in decision-making by trying to *predict* future events, and finding *relationships* that may not be readily apparent.

In statistics, we typically collect data $\rightarrow$ organize and summarize data $\rightarrow$ analyze and interpret data.

Example

Reporting of numbers in the media

What does a headline such as

'*Study shows that you can only concentrate for 20 minutes at a time*'

really mean?

One probable interpretation: the average is 20 minutes.

To get the full story, one should track down and read the original study.

Too much data...

- The number of photos taken is growing exponentially. There are around 300 million photo uploads on Facebook per day.
- Google receives around 4 million search queries per minute.
- Twitter users tweet about 450 000 times every minute.
- Email users send over 200 million messages every minute.
- About 300 hours of video are uploaded to YouTube every minute (compare with 100 hours in 2014).

Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom. – Clifford Stoll

Data sources

- Department of Statistics – http://www.singstat.gov.sg/
- Singapore Government – http://data.gov.sg/
- World Bank – http://data.worldbank.org/
- Gapminder – http://www.gapminder.org/data/
- World Factbook – https://www.cia.gov/library/publications/the-world-factbook/
- Interactive visualizations – http://www.informationisbeautiful.net/
- Correlation vs causation: http://www.tylervigen.com/

Outline

Relationship between probability and statistics

Probability: You know how the population behaves; from that information, you use probability to infer how a sample behaves.

Statistics: You know how a sample behaves; from that information, you use statistics to infer how the population behaves.

Learning statistics

Consistently, many university graduates find statistics to be among the most useful subjects offered.

To be successful in statistics, you need a good understanding of *probability*, and also *common sense*.

Therefore you need to retain everything covered in the Probability course, as well as some calculus and linear algebra.

In particular,

$\mathsf{E}(c_1\,X_1 + c_2\,X_2) = c_1\,\mathsf{E}(X_1) + c_2\,\mathsf{E}(X_2)$, for any constants c_1, c_2 and *any random variables* X_1, X_2.

$\mathsf{Var}(c_1\,X_1 + c_2\,X_2) = c_1^2\,\mathsf{Var}(X_1) + c_2^2\,\mathsf{Var}(X_2)$, for any constants c_1, c_2 and *uncorrelated random variables* X_1, X_2.

Also, binomial distribution, normal distribution, CLT, ...

Examples

Example 1: Among 1 million items, 100 of them are defective. In a random sample of 10000, how many do we expect to be defective?

In a sample of 10000 items, we observe 5 defective ones. Estimate the number of defective items in the population of 1 million.

Example 2: Suppose adult female height in Singapore is normally distributed with mean 160cm and standard deviation 5.0cm. Find the probability that 10 randomly selected females are all between 150cm and 170cm tall.

Given the heights of 10 randomly selected females, estimate the average height of the female population.

Exercise – intro to hypothesis testing

A man claims to be psychic. He demonstrates this by correctly predicting the outcomes of 4 out of 5 coin tosses.

Your task: find the probability that one can do *at least as well as* him by random guessing, and decide if his claim is likely to be true.

Solution

'At least as well' means correctly predicting 4 out of 5 tosses, or 5 out of 5 tosses.

We use the binomial distribution with $p = 1/2$ and $n = 5$:

$$\binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 + \binom{5}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0 = 0.1875,$$

which is quite high (e. g. more likely than guessing the outcome of a dice roll), so we have no reason to believe his claim.

Outline

Survey questions

A statistical study may be *observational* or *experimental*.

- Observations studies often require a sample *survey*.

  Unless the sample equals the population (in which case it is called a *census*), there will be errors in any conclusion we draw about the population (sampling errors).

  Errors can also come from the survey design (such as *selection bias*).

- In experiments, errors can come from faulty equipment, measurement, design, etc.

Poorly designed surveys

What is wrong with . . .

- Conducting a survey on how many telephones the average household has, by dialing people selected randomly from a phone book?

  Truman was famously predicted to lose the 1948 US presidential election. The mistake came from reliance on the results of a phone survey.

- Conducting a survey on the average family size, by asking people how many children their parents had?

- Conducting a survey about what jobs people have, by choosing a random suburb, and knocking on doors whose street number is divisible by 5, between 11am and 4pm?

Poorly designed surveys

- What is wrong with the following survey question?

Which Chinese dialect do you speak at home?

- Mandarin
- Hokkien
- Teochew
- Cantonese

- What is wrong with asking the respondents to rate a new type of coffee on a scale from 0 to 100?

Example

Use of loaded wording

Which of the following would you rather choose?

(1) An operation with a survival rate of 90%?

(2) An operation with a mortality rate of 10%?

Surveys with sensitive questions

The *wording* of a survey may prompt respondents to answer in a particular way. For example, an early version of the Scottish independence referendum read,

'Do you agree that Scotland should be an independent country?'

In some surveys, the questions asked are sensitive no matter how you phrased them:

- Have you taken illicit drugs?
- Are you a racist?

People often do not respond or give false answers. Techniques such as *unmatched count* and *randomized response* can be used to reduce this.

Unmatched count

In the 1991 survey, white Americans were questioned to measure racial hatred against black Americans. Respondents were randomly divided into two groups, and asked the following questions:

Now I'm going to read you three things that sometimes make people angry or upset. After I read all three, just tell me HOW MANY of them upset you. (I don't want to know which ones, just how many.)

(1) "the federal government increasing the tax on gasoline;"
(2) "professional athletes getting million-dollar-plus salaries;"
(3) "large corporations polluting the environment."

Now I'm going to read you four things that sometimes make people angry or upset. After I read all four, just tell me HOW MANY of them upset you. (I don't want to know which ones, just how many.)

(1) "the federal government increasing the tax on gasoline;"
(2) "professional athletes getting million-dollar-plus salaries;"
(3) "large corporations polluting the environment;"
(4) "a black family moving next door to you."

Suppose there are n people in each group, and the total number of things that upset people are respectively u_1 and u_2. Then the proportion of people with racial hatred can be estimated by $\frac{u_2 - u_1}{n}$.

Randomized response

To estimate the proportion of the people who have consumed marijuana, a survey contains two questions:

Q1: "I have consumed marijuana." Answer YES or NO.

Q2: "I have never consumed marijuana." Answer YES or NO.

The interviewee is asked to secretly throw a dice, and answer Q1 if they throw a 6, otherwise answer Q2.

The interviewer does *not* know which question is answered.

Exercise:

If 70% of the responses are YES, estimate the proportion of the population that has consumed marijuana.

Solution

Let the proportion of marijuana consumers be p.

To get a YES, either a marijuana consumer throws a 6, or a marijuana non-consumer throws a non-6.

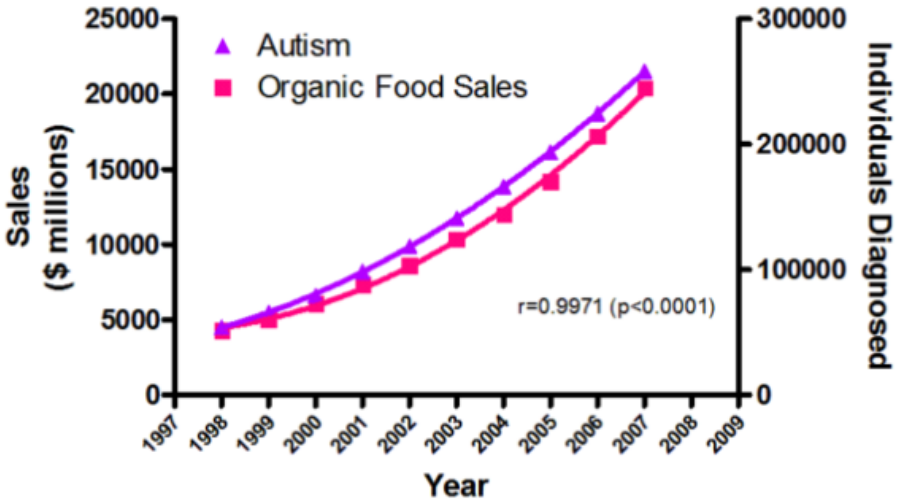
The probability of the above scenarios is

$$p \cdot \frac{1}{6} + (1-p) \cdot \frac{5}{6} = \frac{5-4p}{6}.$$

From the data given, $(5-4p)/6 = 0.7$, so $p = 0.2$.

Outline

Correlation does not mean causation!



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

A strong relationship (*correlation*) between two sets of data does not imply *anything* about causation, even if the study is error-free.

Both sets of data may be influenced by another variable (a *confounding* variable), or the correlation could occur by chance.

Research (of non-statistical nature) needs to be carried out to prove causation. E. g. from the strong correlation between smoking and cancer *alone*, we cannot conclude that 'smoking causes cancer' or 'cancer causes smoking'.

Excel exercise

- Open the Excel file on stork population vs human birth rate.
- Construct a *scatter plot* and comment on the correlation. Offer an explanation.
- Do you notice something suspicious about the data?
- Go to File $\rightarrow$ Options $\rightarrow$ Add-Ins $\rightarrow$ Go, and select the Analysis ToolPak.

Statistics

Week 1: Collecting Data (Chapter 3)

ESD, SUTD

Term 5, 2017

Outline

Excel

If you use *Excel* on a Mac:

Download StatPlus:mac LE for free from AnalystSoft, and then use it while running *Excel*.

It should work similarly to the Analysis ToolPak.

Unmatched count

For the 1991 Race and Politics Survey, see the paper

J. H. Kuklinski and M. D. Cobb, *Racial Attitudes and the "New South"*, The Journal of Politics, **59** (1997), p323–349.

Some conclusions:

- For the non-South regions, unmatched count failed to give meaningful results (many people upset about everything). Other techniques gave racial discrimination at about 10%.

- For the South, unmatched count gave meaningful results. Racial discrimination at around 40%.

Randomized response

See pdf file on *eDimension*.

For this question, you can partially *check* your answer by considering some simple (boundary) cases:

- What would be the proportion of YES if the consumption rate was 0%?
- What would be the proportion of YES if the consumption rate was 100%?

Being able to check your own work is extremely important.

Outline

Sampling

In many applications, the population is so large that sampling is the only practical option.

In some applications, destructive testing (such as crash tests) is required, so sampling is the only feasible option.

Before you sample, you need to:

- Define the population of concern.
- Specify what you want to measure.
- Specify a sampling method.
- Determine the sample size.

Questions

1. Why do we have to stick to the sample size? (Why can't we stop whenever we want?)
2. Why do we have to specify what to measure? (Multiple testing.)

Answers

1. So we can't 'cheat' by taking advantage of random fluctuations – for instance, if we wish to show that a coin is not fair by tossing it repeatedly, then stopping at a convenient point when there are significantly more H's than T's would be cheating.

2. Multiple testing – for instance, suppose we simultaneously test for 100 rare diseases, each has only a 1% chance of randomly occurring. Assuming that the diseases are independent, then detecting at least one of the diseases has probability as high as

$$1 - \left(1 - \frac{1}{100}\right)^{100} \approx 0.63.$$

Note that this probability is independent of any experiment we might wish to perform.

Sampling methods

Convenience sampling: use a sample that is readily available. (Not accurate. E. g. many psychological studies are done on psychology students.)

Snowball sampling: the first respondent refers to a friend. This friend refers to the next friend, etc.

Simple random sampling: a sample of size n is drawn from a population of size N without replacement, such that each possible sample of size n has the same chance of being chosen.

Demo: we can do this in Excel by creating a list of random numbers and sorting them.

More sampling methods

Systematic sampling: select every kth unit (useful for items coming off an assembly line).

Cluster sampling: divide the population into heterogeneous clusters (e. g. geographic areas), then draw simple random samples from each one. This method saves cost.

Stratified sampling: divide the population into homogeneous groups/strata (e. g. ethnicity, age group), then draw simple random samples. This method is more accurate.

For example, the General Household Survey 2015 used stratified sampling based on dwelling types and planning areas. Within each group, systematic sampling was used.

Exercise

How would you sample people to determine the proportion of various natural hair colours in a European country?

Hint: define the target population.

Possible design:

- Observe people from streets or dwellings (pick a range of different locations), or photos from a database. Do not use (say) Internet surveys, since self-reporting may be unreliable .
- Target the 20-40 year old age group, since people much older than this may be balding or going gray, and people much younger than this may experience darkening of hair colour.
- Target males, since they are less likely to dye their hair.

Outline

Treatment and placebo

Often, the aim of an experiment is to determine the effectiveness of a particular *treatment*.

Examples: medical treatments, diets, new ways of teaching, different work conditions, different production techniques, . . .

Problem: the *placebo* effect. A placebo is a simulated and otherwise ineffectual treatment.

Frequently, a patient given a placebo treatment (without knowing it) will have a perceived or actual improvement.

In medicine, common placebos include inert tablets (such as sugar pills) or inert injections.

Examples of placebo

- Placebos that are perceived to be more expensive tend to work better.
- When (falsely) told that a placebo has a certain smell/taste, some patients start to believe that they can smell/taste it.
- When (falsely) told that a placebo has a side-effect (e. g. numbs pain, or causes a rash), some patients actually experience the side-effect.
- Red placebo pills work better as stimulants while blue pills work better as depressants (e. g. sleeping pills).
- Even renaming a medication will temporarily make it more effective due to the novelty.

Examples of placebo

Occasionally, the placebo effect may be advantageous, for instance if a clinic runs out of pain killers.

For a real life example of the placebo effect, watch www.youtube.com/watch?v=udJ31KKXBKk from 1:50 onwards.

There are many related effects, such as the Hawthorne effect: any change in work conditions will temporarily increase productivity, due to the novelty and the perception that the workers are getting attention.

Control group

Therefore, it is important in a study to have a *control group*, which receives either no treatment, the standard treatment, or an ineffectual (placebo) treatment, whichever is most appropriate. This group provides a baseline for comparison.

The control group and treatment group should be allocated randomly.

A study is called *double blind* if both the researcher and the subject are kept unaware of which group they belong to. This removes psychological effects and is more accurate.

However, using placebos in an experiment raises ethical questions as it can be seen as a form of deception. Therefore, any research that involves human subjects must be reviewed and approved by the appropriately authority (at SUTD: the IRB).

Real life examples

- Salk polio vaccine trial: $> 200\,000$ people each in placebo and treatment groups. '6 sigma' (the meaning will be explained later).
- For many vaccines, roughly the same percentage of people in placebo and treatment groups experience side effects.
- 'Although there have been reports of an MSG-sensitive subset of the population, this has not been demonstrated in placebo-controlled trials.'

  From M. Freeman, *Reconsidering the effects of monosodium glutamate: a literature review*, J Am Acad Nurse Pract, **18** (2006), p482-486.

Exercise

How would you design a study to test whether video games improve one's reflexes?

Think about: whether the study should be observational or experimental, whether to use control groups, etc.

Answer

An observational study may establish a positive correlation between video games and reflexes, but one possible explanation for this correlation may be that people with better reflexes are better at video games, and hence play them more often. Since we are interested in whether videos games *improve* one's reflexes, such a study is not ideal.

An experimental study is preferred. Subjects should have no prior experience with video games. They should be randomly divided into a control group and a treatment group. Ideally, the two groups should be as similar as possible.

The treatment group gets to play video games, while the control group doesn't. After a predetermined period of time, the reflexes of both groups should be measured and compared.

(We can also *block* by the age group of the test subjects.)

Outline

Randomized block design

Example: suppose we want to compare three medical treatments, A, B and C, for their effectiveness.

It is known that the treatments may affect people of different ages differently.

We are only interested in A, B and C (the *treatment factor*), not in the ages (the *noise factor*).

How to design an experiment? As in stratified sampling, put people of the same age group in a 'block'. Randomize treatment within each block.

11–20	21–30	31–40	41–50	51–60
A	B	C	A	B
C	A	B	C	C
B	C	A	B	A

Why randomize? To reduce the effects of any hidden variables not accounted for by blocking.

Latin square design

What if there are two noise factors?

Example: suppose we want to compare four types of air fresheners (labeled 1, 2, 3, 4). The tests are done in 4 different rooms and in 4 different months. Both the room and the month have an effect on air quality.

	R_1	R_2	R_3	R_4
M_1				
M_2				
M_3				
M_4				

We can fill in the grid as if it were a (simplified) sudoku puzzle. This is an example of a *Latin square* design.

Latin square answer

One possible design is:

	R_1	R_2	R_3	R_4
M_1	1	2	3	4
M_2	2	3	4	1
M_3	3	4	1	2
M_4	4	1	2	3

Each number appears exactly once in every row and column.

Statistics

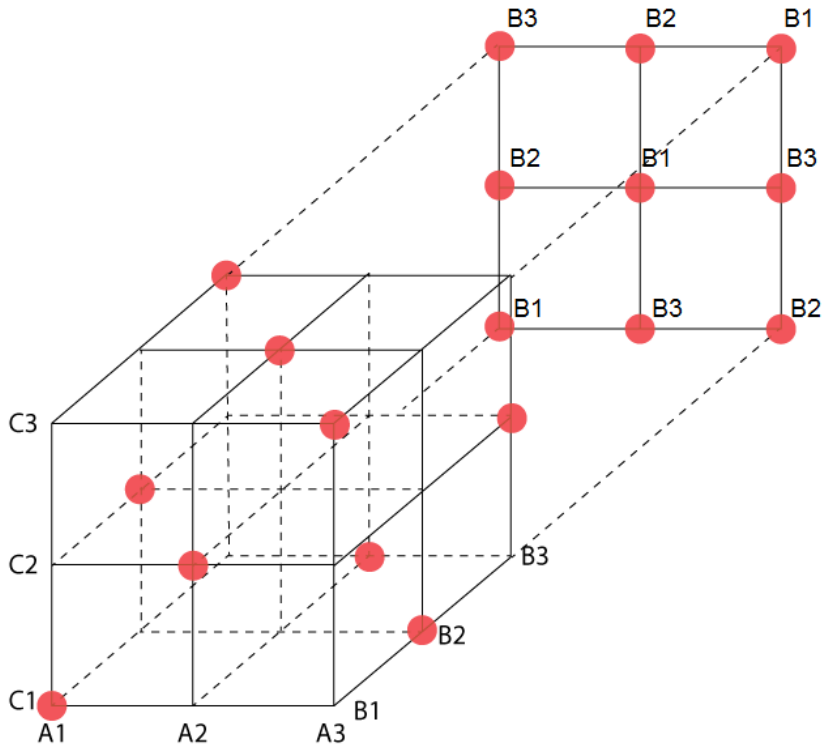
Week 2: Summarizing and Exploring Data (Chapter 4)

ESD, SUTD

Term 5, 2017

Recap

Last week, factorial design:



The projection is a Latin square!

Outline

Types of data

Data can be classified into two types: **categorical** (qualitative) and **numerical** (quantitative).

Categorical data can be either *nominal* (distinct labels, such as red, blue, yellow) or *ordinal* (ranked, such as disagree, neutral, agree).

Numerical data can be either *discrete* (results of counting, such as number of people) or *continuous* (results of measurement, such as distance).

Frequency table

A frequency table can be used to show the number of occurrences for each category. The relative frequency (the proportion in each category) can also be given.

Example: a survey of 100 people is conducted on the top 2 reasons why they are late to class or work.

Reason	Frequency	Relative frequency (%)
Bad weather	24	12
Overslept	58	29
Alarm failure	36	18
Family issue	6	3
Traffic	68	34
Other	8	4
Total	200	100

Bar chart

A bar chart uses rectangular bars to denote the frequencies.

Example: use *Excel*'s column chart or bar chart option to construct a bar chart for the previous table.

Note: the bars in a bar chart should not touch, in order to separate the different categories.

Pie chart

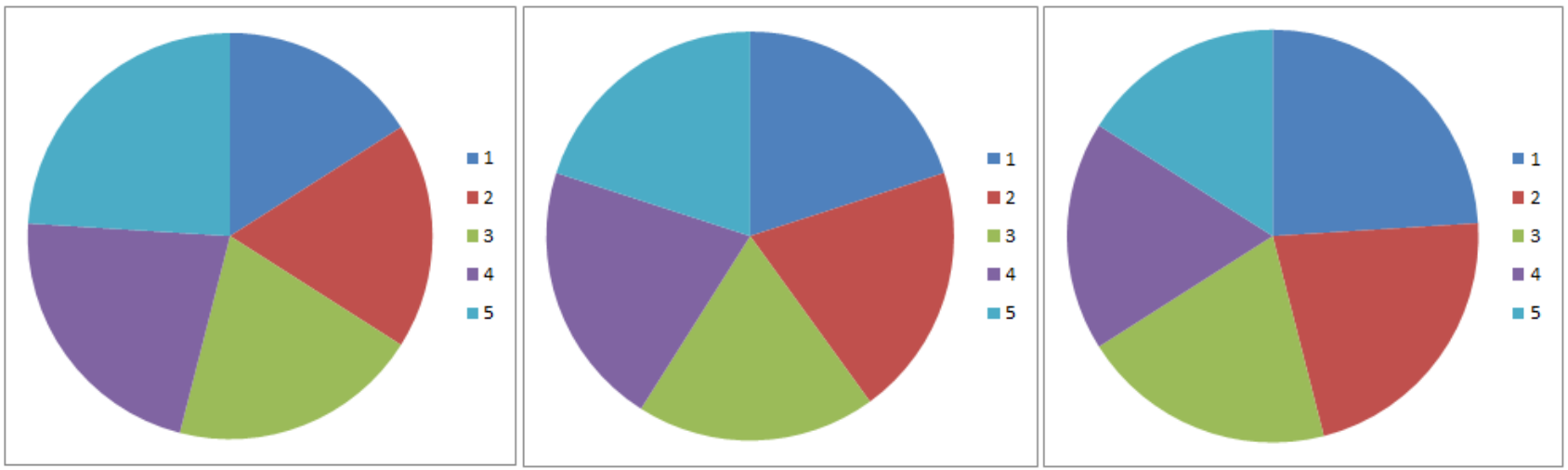
A pie chart uses the relative area, or angle, of the sectors represent the relative frequencies. An informative pie chart should label the frequencies.

Many experts *do not recommend* the use of pie charts. There are several reasons for this: pie charts are very often misused; also, humans are not good at comparing angles.

Do not use pie charts for too many or too few (such as 2) categories. Do not use 3D pie charts.

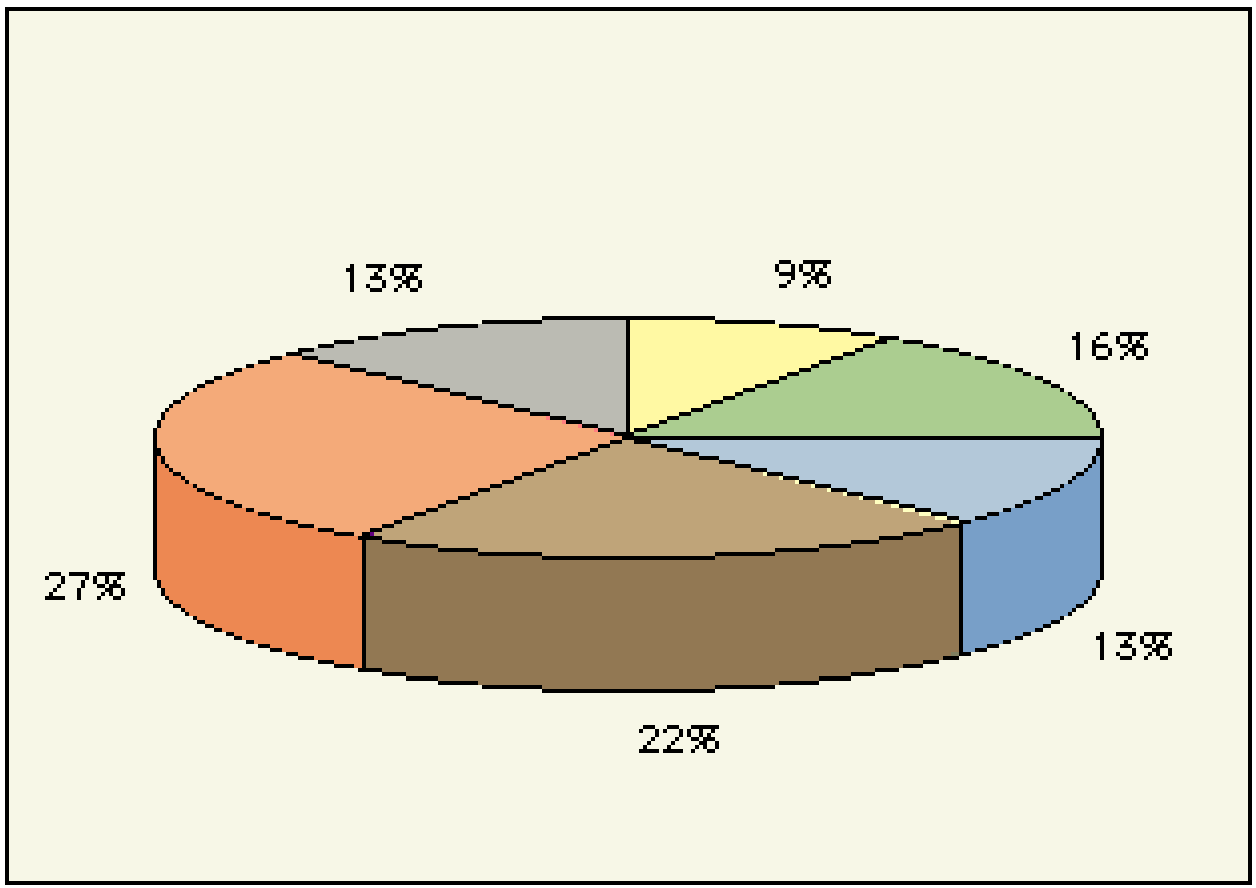
Misuse 1

The charts below represent the results from a local election with five candidates at three different locations. What are some of the problems with using pie charts here?



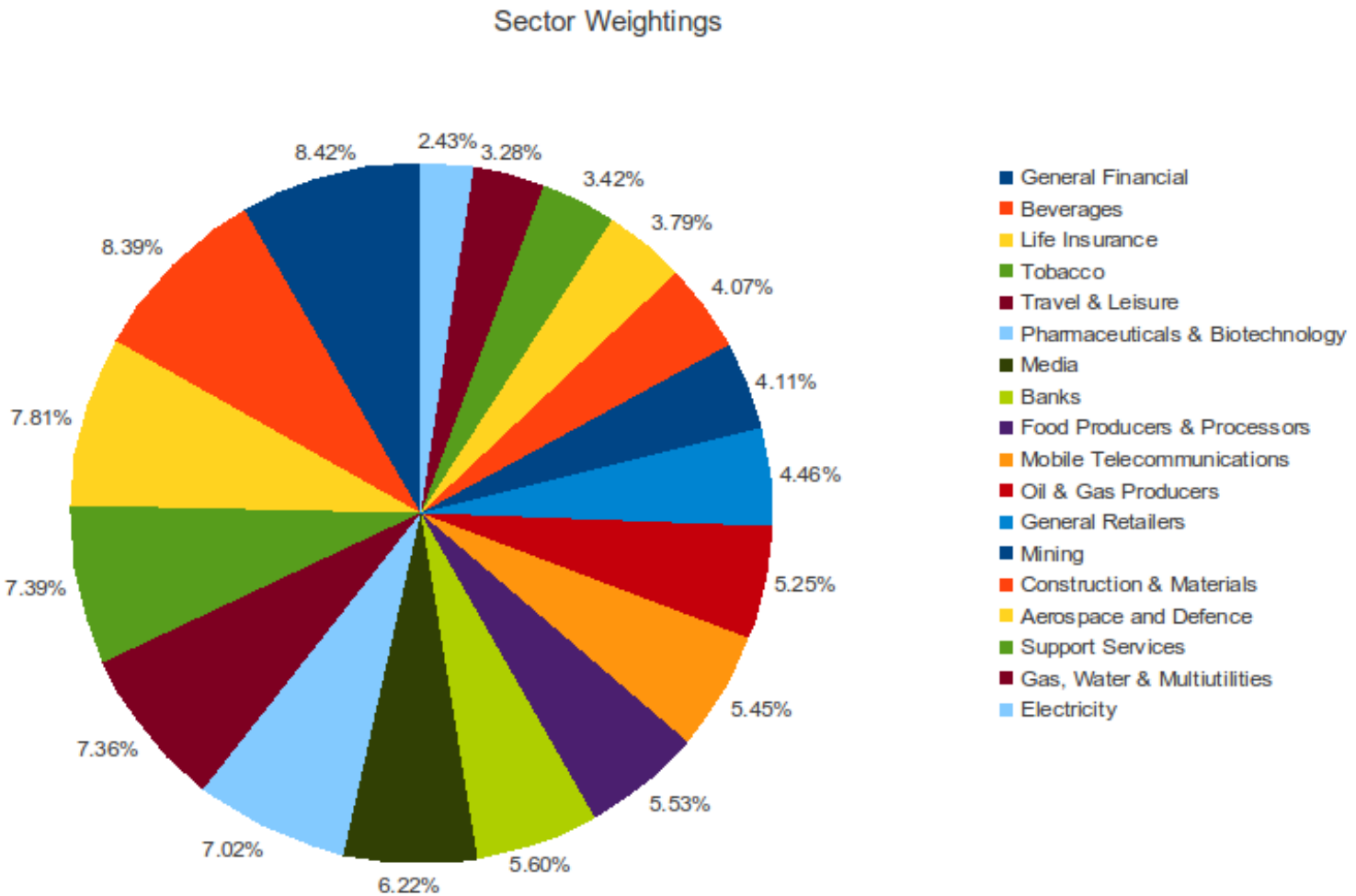
Misuse 2

Is this pie chart is a good representation of the data?



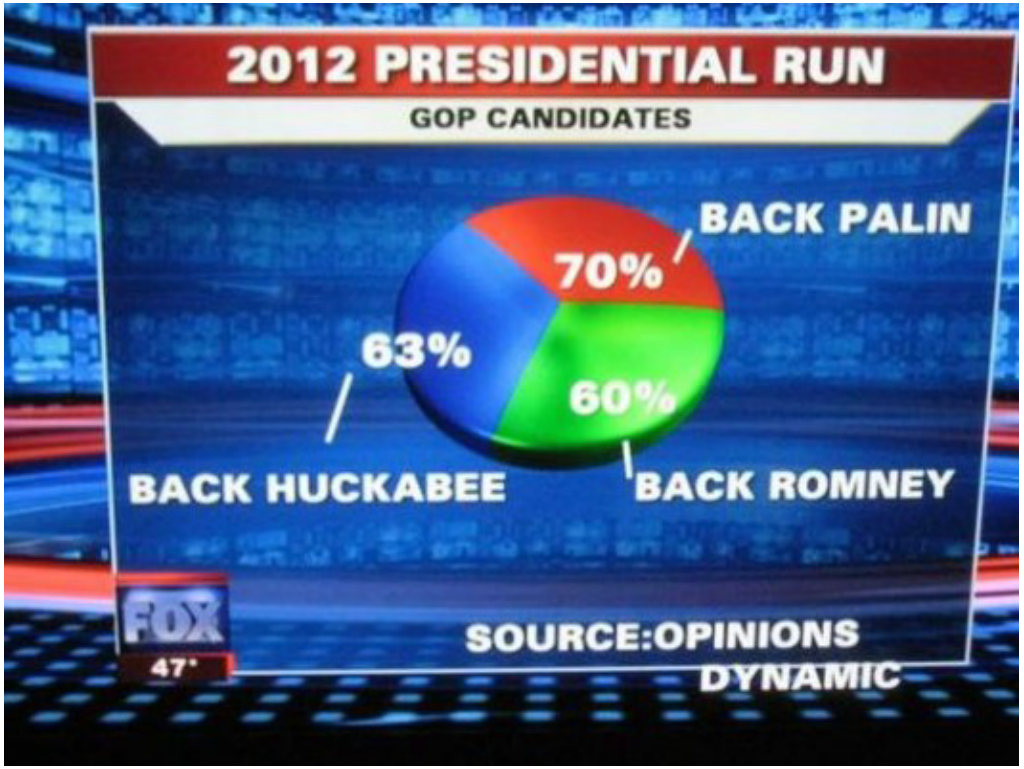
Misuse 3

Is this pie chart a good representation of the data?



Misuse 4

This chart was used on FOX News.



Outline

Summary statistics – measures of centre

Mean (average)

Given data values $x_1, x_2, \ldots, x_n$, the (sample) mean is defined as

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

The sample mean is used to estimate the population mean μ (more on this later).

Median

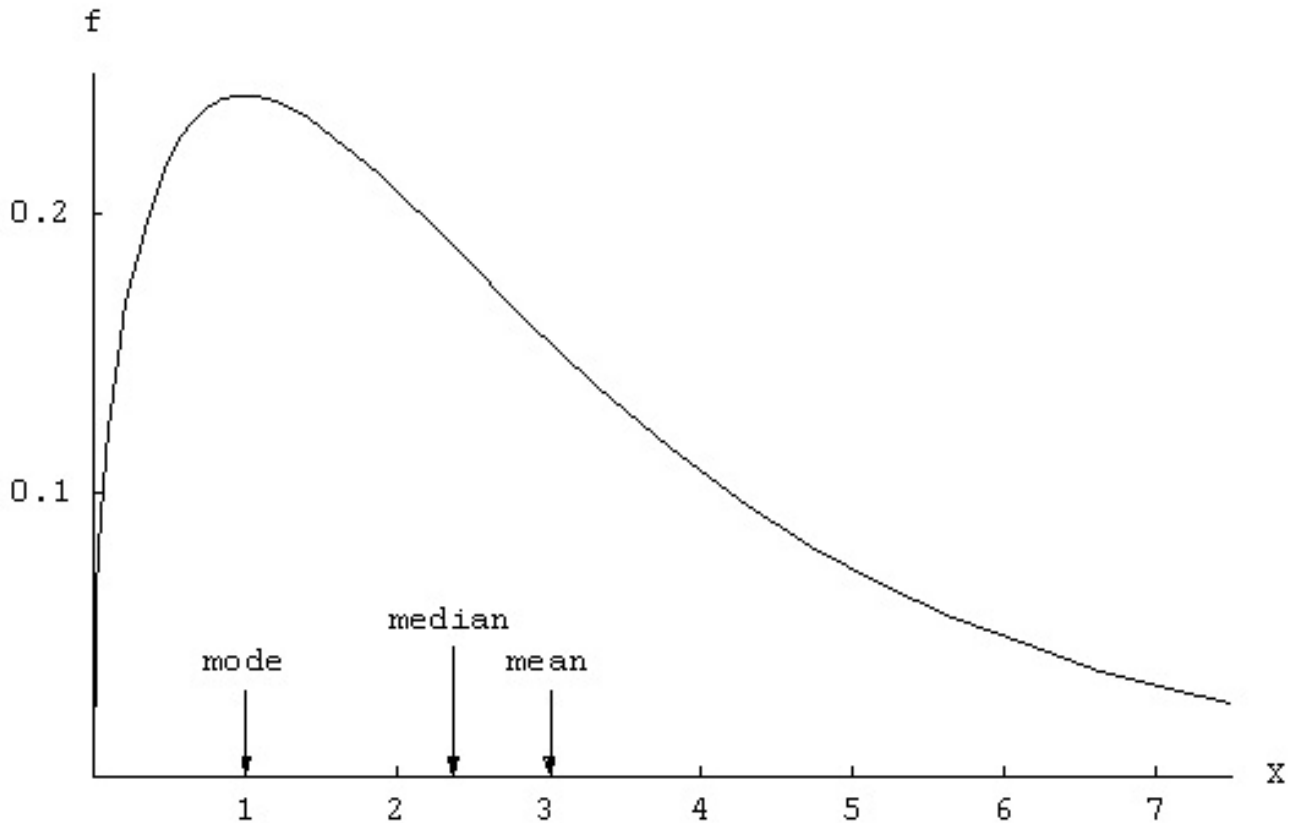
Given data values $x_1, x_2, \ldots, x_n$, order them as follows: $x_{\text{min}} = x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} = x_{\text{max}}$. The median is defined as

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right] & \text{if } n \text{ is even.} \end{cases}$$

Mean, median and mode

The *mode* is a value that appears most often in a data set. It is not always a good measure of center, nor is it always unique.

You should know how to find the (population) mean, median and mode for a continuous or discrete distribution.



Optimization

Exercise: (1) Show that $\bar{x}$ is the optimal solution to

$$\min_x \sum_{i=1}^{n} (x_i - x)^2.$$

(2) Show that $\tilde{x}$ is an optimal solution to

$$\min_x \sum_{i=1}^{n} |x_i - x|.$$

Hint: draw a picture, and use the triangle inequality, $|a| + |b| \geq |a + b|$.

Answer to (2)

Order the data values as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The proof here assumes that n is even; the odd case is very similar.

We have

$$\begin{aligned}\sum_{i=1}^{n} |x_i - x| &= \Big(|x_{(1)} - x| + |x_{(n)} - x|\Big) + \Big(|x_{(2)} - x| + |x_{(n-1)} - x|\Big) \\ &\quad + \cdots + \Big(|x_{(n/2)} - x| + |x_{(n/2+1)} - x|\Big) \\ &\geq |x_{(1)} - x_{(n)}| + |x_{(2)} - x_{(n-1)}| + \cdots + |x_{(n/2)} - x_{(n/2+1)}|,\end{aligned}$$

where we have applied the triangle inequality $n/2$ times.

Note that the last line depends only on the data values, and not on x, so it is fixed; let us call this fixed value X.

Answer to (2)

So we have

$$\sum_{i=1}^{n} |x_i - x| \geq X,$$

and equality is achieved if and only if x lies between $x_{(1)}$ and $x_{(n)}$, and between $x_{(2)}$ and $x_{(n-1)}$, ..., and between $x_{(n/2)}$ and $x_{(n/2+1)}$.

In other words, equality is achieved if and only if x is between $x_{(n/2)}$ and $x_{(n/2+1)}$ (since the data values are ordered).

Therefore $\sum_{i=1}^{n} |x_i - x|$ achieves its minimum value when x is between $x_{(n/2)}$ and $x_{(n/2+1)}$, which occurs, for instance, when x equals the median $\tilde{x}$.

Use of the mean

Example 1: SUTD jelly bean contest.

The entries were:

404, 225, 1228, 1119, 1117, 1234, 1125, 5566, 1234, 920, 1695, 987, 1400, 1467, 1425, 1650, 1545, 1600, 1250, 1272, 1350, 1783, 1199, 2359, 777, 777, 1500, 908, 1317, 1445, 1876, 888, 1370, 1560, 1000, 688, 1360, 1275, 1700, 2215, 1234, 911, 1028, 1524, 888, 945, 159, 1212, 1518, 999, 1456, 1200, 1313, 1086, 2359, 1763, 1800, 1452, 1500, 857, 1239

If we take out the 'obviously' too high guess of 5566, and the 'obviously' too low guesses of 159 and 225, then the sample mean of the remaining guesses is 1315.57, which is closer than the winning entry of 1317.

Use and misuse of the mean

Example 2: a stopped clock shows the correct time twice a day, while a clock that is one minute too slow never shows the correct time.

Naïvely, it may seem that the stopped clock is more accurate. However, if we compare the *mean* error of each clock, then the slow clock is far more accurate.

Example 3:

"The average annual income of leading research mathematicians (those, say, with at least three articles in the Annals of Mathematics) is about 10,000,000 USD." – James Simons

Outlier and robustness

Outlier

An outlier is a data point that is 'distant' from the main body of data points. Outliers can occur by chance in any distribution, but they are often indicative either of measurement *error* or that the population has a *heavy-tailed* distribution.

There is no hard and fast rule for detecting outliers.

The median is a *robust* statistic, meaning that it is resistant to outliers, while the mean is not robust.

If an outlier is a measurement error, then we should discard it or use a robust statistic. If an outlier results from a heavy-tailed distribution, then we should be cautious in using tools that assume a normal distribution.

Summary statistics – measures of spread

Sample variance

Given data values $x_1, x_2, \ldots, x_n$, the sample variance is defined as

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

The sample standard deviation is s.

The population variance is denoted by σ^2.

We will explain the '$n-1$' later.

Range, interquartile range

The *range* is define as $x_{\max} - x_{\min}$.

The *interquartile range* IQR is defined as $Q_3 - Q_1$. *One* definition of Q_1 (resp. Q_3) is the median of the lower (resp. upper) half of the ordered data. If there are an odd number of data points, do not include the median in either half.

The five number summary is $\{x_{\min}, Q_1, \tilde{x}, Q_3, x_{\max}\}$. They can be used to construct a **box plot**, which is useful for comparing different data sets.

For a box plot, data values that are more than 1.5 IQR below Q_1 or above Q_3 are usually considered outliers.

Box plots are poorly supported in *Excel*, but can be done with *R*.

Quantiles

Exercise: which of standard deviation, range, and IQR are robust?

The pth quantile $\tilde{x}_p$ is a value such that fraction p of the data are less than or equal to it.

There are different definitions! The textbook uses

$$\tilde{x}_p = x_{(m)} + \big[p(n+1) - m\big]\,\big(x_{(m+1)} - x_{(m)}\big),$$

where m denotes the integer part of $p(n+1)$. (For some values of p, $\tilde{x}_p$ is not defined.)

You may check that $\tilde{x} = \tilde{x}_{0.5}$. On the other hand, another definition for the quartiles is $Q_1 = \tilde{x}_{0.25}$ and $Q_3 = \tilde{x}_{0.75}$, but this is *different* from the one given on the last slide. The *Excel* command is `quartile.exc`.

Histogram

To construct a histogram, first divide up the range of values into intervals (*bins*), then counts how many values fall into each bin.

For each bin, a rectangle is drawn with height proportional to the count and width equal to the bin size. The rectangles should touch each other.

A histogram can be used to estimate the probability distribution of a continuous variable.

It is often a good idea to use at least 2 different plots to explore a data set.

Exercise: see *Excel* file on speed of light data.

- Find the mean, standard deviation, the five number summary, and any outliers.
- Make a box plot (by hand).
- Make a histogram.

Statistics

Week 2: Summarizing and Exploring Data (Chapter 4)

ESD, SUTD

Term 5, 2017

Outline

Histogram, continued

In an *Excel* histogram, you should:

- Properly label the bins (the values generated are only the upper boundaries of the bins).
- Change the Gap Width to 0%.

There is no universal formula for choosing the number of bins.

- *Excel* uses $[\sqrt{n}]$ bins.
- Another recommendations is to use $[\log_2 n] + 1$ bins.
- Yet another rule is to set the bin width to $2\,\mathrm{IQR}/n^{1/3}$.

In practice, aim for between 5 and 20 bins, and make the boundaries 'nice' numbers.

Other measures of spread

The *sample coefficient of variation* is defined as

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

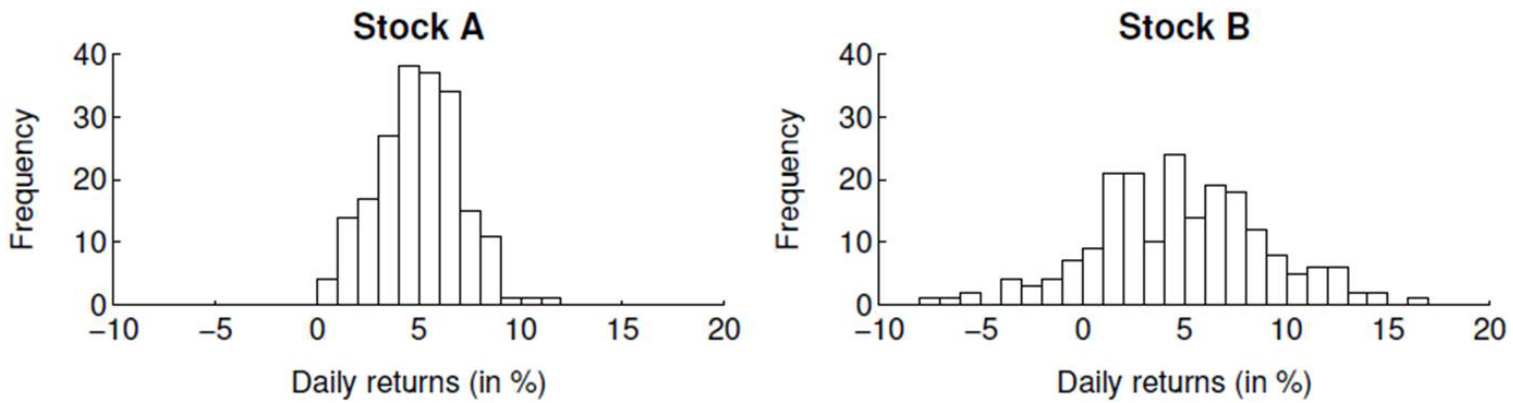
It is used in, for example, queueing theory (for an exponential distribution, CV should be 1).

The *z-score* or standard score calculates how many standard deviations a data value is above the sample mean:

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

You have seen this used in the normal distribution. It is useful for comparing different data sets.

Why study spread – example



- Are these two stocks similar for investors?
- Which one would you invest in?

Sample covariance

Bivariate data can be represented on a scatter plot.

Recall that the covariance of two random variables X and Y is given by $\mathsf{E}[(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])]$.

Sample covariance and correlation

Given data values (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$, the sample covariance is defined as

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

The sample correlation coefficient is given by

$$r = \frac{s_{xy}}{s_x s_y}.$$

We will use these when studying linear regression.

Demo: look up Anscombe's quartet.

Tables

Bivariate data can also be represented in table form.

Before making any conclusions from tables, be careful of the way samples are drawn.

Example: a respiratory problem is studied by first finding 500 smokers and 500 non-smokers and then determining whether or not each individual has the problem. The results are shown below.

	Yes	No	Row total
Smokers	250	250	500
Non-smokers	50	450	500
Column total	300	700	1000

Exercise: are the following statements true?

About $5/6$ of all people with the respiratory problem are smokers.

About $1/2$ of all smokers have the respiratory problem.

Simpson's paradox

Real life example comparing two treatments for kidney stones:

	Treatment A	Treatment B
Small stones	93% $(81/87)$	87% $(234/270)$
Large stones	73% $(192/263)$	69% $(55/80)$
Both	78% $(273/350)$	83% $(289/350)$

This can occur when the group sizes are uneven, so watch out.

Q-Q plot

A **Q-Q plot** compares two probability distributions by plotting their quantiles against each other.

A point (x, y) on the plot corresponds to a quantile of the 2nd distribution plotted against the same quantile of the 1st one.

The *normal probability plot* is a special case of the Q-Q plot, when the 2nd distribution is the standard normal.

If a normal probability plot is close to a *straight line*, then the 1st distribution is approximately normal (since all normal distributions are related by linear transformations).

Normal probability plot

Consider some ordered data values $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

Then $x_{(i)}$ is the $\frac{i}{n+1}$ quantile.

We plot $x_{(i)}$ against the $\frac{i}{n+1}$ quantile of the standard normal distribution, which is given by $\Phi^{-1}\left(\frac{i}{n+1}\right)$.

Intuition for using $n+1$ and not n: (1) imagine drawing $x_{(i)}$ from a distribution... (2) $\Phi^{-1}\left(\frac{n}{n}\right) = \infty$.

See *Excel* demo on speed of light data. For Φ^{-1}, use `norm.s.inv`.

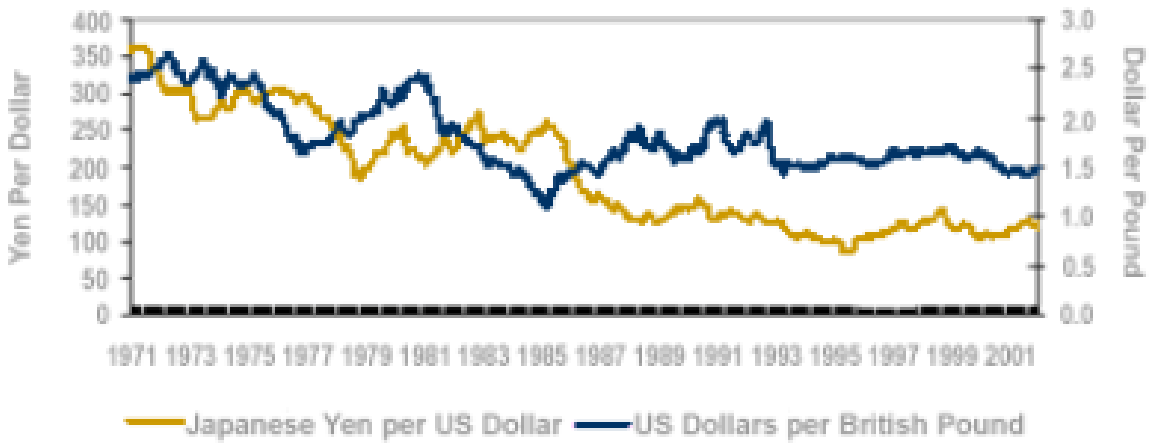
Outline

Time-series data

A **time series** is a sequence of data points $x_1, x_2, x_3, \ldots$, measured at successive points in time (typically spaced at uniform intervals). For examples, daily closing value of a stock, or annual rainfall.

Usually, a time-series has the following components: stable, trend (long-term pattern), seasonal (short-term, periodic fluctuation), random.

There are many examples from economics and finance:



Forecasting

We now describe some methods to

- Smooth out short-term fluctuations and highlight long-term trends in a time series, and/or
- Attempt to predict (forecast) the value of a time series at the next point in time.

A naïve way to forecast is to use the last data point:

$$F_{t+1} = x_t.$$

A more sophisticated approach is the **moving average**:

$$F_{t+1} = \frac{x_{t-w+1} + \cdots + x_{t-1} + x_t}{w}.$$

This also allows us to smooth out the time series, but can introduce a lag.

Exponentially weighted moving average

Weighted moving average: α_i are the weights; the idea is to give more importance to more recent data.

$$F_{t+1} = \frac{\alpha_{w-1}x_{t-w+1} + \cdots + \alpha_1 x_{t-1} + \alpha_0 x_t}{\alpha_{w-1} + \cdots + \alpha_1 + \alpha_0}.$$

We can choose α_i to be decreasing **exponentially**.

Let $\alpha \in (0, 1)$, then define $F_{t+1} = \mathrm{EWMA}_t$, where

$$\mathrm{EWMA}_t = \alpha\, x_t + (1-\alpha)\, \mathrm{EWMA}_{t-1},$$

with $\mathrm{EWMA}_0 = x_1$.

If we apply this formula repeatedly, then after simplification,

$$\mathrm{EWMA}_t = \alpha \left[x_t + (1-\alpha)x_{t-1} + (1-\alpha)^2 x_{t-2} + \cdots + (1-\alpha)^{t-1} x_1\right] + (1-\alpha)^t x_1.$$

Forecasting error

How do we pick α?

The error of the forecast is $e_t = x_t - F_t$.

α can be chosen to minimize some total error.

One commonly used measure of total error is the *mean absolute percent error*, defined as

$$\text{MAPE} = \frac{1}{T-1} \sum_{t=2}^{T} \left| \frac{e_t}{x_t} \right| \times 100\%.$$

In *Excel*, we can use Solver to find the value of α that minimizes MAPE.

Statistics

Week 3: Sampling Distributions (Chapter 5)

ESD, SUTD

Term 5, 2017

Outline

Random variable

Motivating example: before an election, we wish to estimate the population proportion p of people who plan to vote for a particular candidate. We can survey 100 random voters and estimate p.

Each time we run such a survey, the sequence of responses, and the resulting estimate, will generally be different.

Therefore, the data points we are getting are essentially acting like random variables, and can be treated (modeled) as such. In this case, each response can be modeled as a Bernoulli random variable with parameter p; moreover, the responses are independent.

The statistic we are interested in (the estimate for p) is also a random variable.

A *sampling distribution* is the probability distribution of a given statistic based on a random sample.

Sample mean

Consider a population with mean μ and variance σ^2. Let $X_1, X_2, \ldots, X_n$ be a random sample drawn from the population; note that X_i are independent and identically distributed (iid).

Notation: upper case letters (such as X_1, X_2) denote random variables; lower case letters (such as x_1, x_2) denote data values, which are the observed values of the random variables. However, when the context is clear, we will sometimes abuse the notation and use lower case letters to represent both.

How does the sample mean, $\bar{X}$, behave?

$\mathsf{E}(\bar{X}) =$

$\mathsf{Var}(\bar{X}) =$

Moral: larger n means better approximation to μ.

Sample mean – example

Suppose we have 2 electronic scales, whose readings are iid with mean = actual weight of object, and variance = σ^2.

Then it is actually more precise to weigh the same object on both scales, and take the average reading, than it is to just weigh it once.

We can demonstrate this effect using a 'toy' scale and some simple numbers...

Central limit theorem

Some special cases:

- If $X_i \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \sigma^2/n)$.
- If X_i are Bernoulli with parameter p, then $P(\bar{X} = x/n) = P\left(\sum_{i=1}^n X_i = x\right) = \binom{n}{x} p^x (1-p)^{n-x}$ (binomial).

What about X_i drawn from an arbitrary distribution?

Central limit theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be iid random variables drawn from an arbitrary distribution with finite mean μ and variance σ^2. Then as $n \to \infty$,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0, 1).$$

Excel demo: average of n loaded dice.

CLT; binomial

As a rule of thumb, the central limit theorem is used when the sample size is at least 30.

You should also recall the more elementary, but useful result:

Normal approximation to binomial

Suppose X is a Bin(n, p) random variable. For large n,

$$\frac{X - np}{\sqrt{np(1-p)}} \approx N(0, 1).$$

In practice, do not forget the continuity correction.

Exercise: how is the above a special case of the CLT?

Outline

Point estimator

Let x_i be random samples drawn from a population with an unknown parameter θ. A point *estimator* $\hat{\theta}$ is a statistic computed from x_i, and is used to estimate θ.

Example: if θ is the population mean μ, then we can use $\hat{\theta} = \bar{x}$.

The **bias** of an estimator is defined as $\mathrm{E}(\hat{\theta}) - \theta$. If the bias is 0, then the estimator is called *unbiased*.

Note: do not confuse 'bias' here with selection bias in sampling, which occurs when randomization is not achieved.

Exercise: is $\bar{x}$ an unbiased estimator of μ?

Unbiased estimator

How do we estimate σ^2 when only a sample of size n is taken?

We could try the unbiased estimator

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

the problem is that we do not know μ, and only know $\bar{x}$!

Exercise: check that the above estimator is unbiased, using $\mathrm{Var}(X) = \mathrm{E}((X-\mu)^2)$.

What about

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

Intuitively, a sample may miss some extreme values, so this estimator seems too small.

Sample variance

We can modify the factor in front of the $\sum$ to make the estimator unbiased.

Tricky calculation

$$\mathsf{E}\left[\sum_{i=1}^{n}(x_i-\mu)^2\right] = \mathsf{E}\left[\sum_{i=1}^{n}\Big((x_i-\bar{x})^2+(\bar{x}-\mu)^2+2(x_i-\bar{x})(\bar{x}-\mu)\Big)\right]$$

$$\Rightarrow\ n\,\sigma^2 = \mathsf{E}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] + n\,\mathsf{Var}(\bar{x}) + 0$$

$$\Rightarrow\ (n-1)\,\sigma^2 = \mathsf{E}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right].$$

Therefore

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

is an unbiased estimator of σ^2, namely $\mathsf{E}(s^2)=\sigma^2$.

German tank problem

During World War II, the Allies had the serial numbers of some captured/destroyed German tanks and wanted to estimate the total number of German tanks.

They correctly guessed that the serial numbers ran from 1 to N, where N was the total number of tanks produced. How can one estimate N?

Month of production	Intelligence estimate	Statistical estimate	Actual (from post-war captured documents)
June 1940	1000	169	122
June 1941	1550	244	271
August 1942	1550	327	342

Exercise

Goal: estimate the maximum value N of a discrete uniform random variable taking values from $\{1, 2, \ldots, N\}$, given a sample of size n drawn without replacement.

Let the sample be $\{x_1, x_2, \ldots, x_n\}$. Some possible estimators for N are:

- $\bar{x} + 3s$
- $2\bar{x} - 1$
- $x_{(n)}$
- $x_{(n)} + x_{(1)} - 1$
- $\dfrac{n+1}{n} x_{(n)} - 1$

The *Excel* sheet provides 500 simulated samples, each of size $n = 5$, from a population of $N = 250$. Evaluate the estimators and compare the bias.

Solution

(1) This estimator is inspired by the normal distribution. Biased.

(2) $\mathsf{E}(\bar{x}) = \frac{1}{5}\big(\mathsf{E}(x_1) + \cdots + \mathsf{E}(x_5)\big) = \frac{N+1}{2}$. Hence $\mathsf{E}(2\bar{x} - 1) = N$ (unbiased). Note however that this may be less than $x_{(n)}$.

(3) $x_{(n)} \leq N$ always, so it is a biased estimator.

(4) This estimator is based on the last one, as well as symmetry. Unbiased.

(5) It turns out that this is unbiased, and has the *least variance* among all unbiased estimators.

Proof sketch of unbiasedness: the probability that the maximum, $x_{(n)}$, equals m is given by

$$P(x_{(n)} = m) = \frac{\binom{m-1}{n-1}}{\binom{N}{n}}.$$

Multiply this by m and sum:

$$\begin{aligned}
\mathsf{E}(x_{(n)}) &= \sum_{m=n}^{N} m\, P(x_{(n)} = m) \\
&= \sum_{m=n}^{N} m \frac{\binom{m-1}{n-1}}{\binom{N}{n}} = \frac{n}{\binom{N}{n}} \sum_{m=n}^{N} \binom{m}{n} \\
&= \frac{n}{\binom{N}{n}} \binom{N+1}{n+1} = \frac{n(N+1)}{n+1}.
\end{aligned}$$

Rearranging gives the required result.

Statistics

Week 3: Sampling Distributions (Chapter 5), Estimation (Chapter 6)

ESD, SUTD

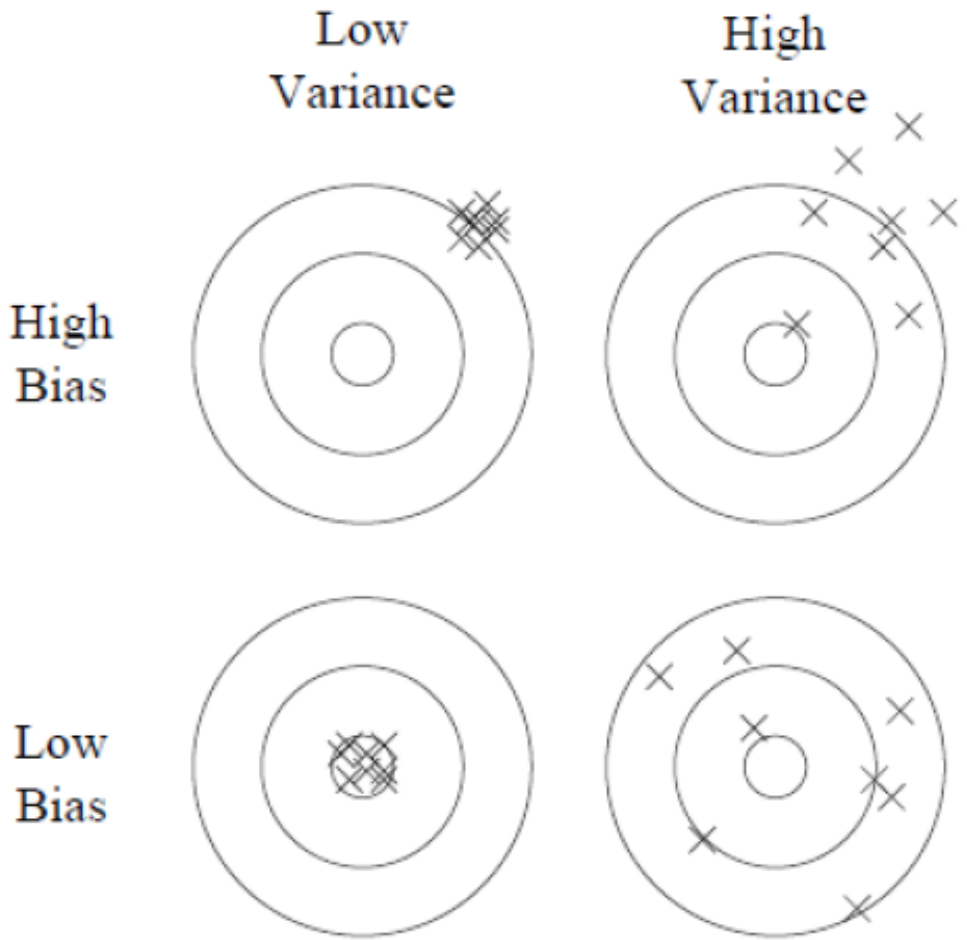
Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Outline

Bias and variance of an estimator



The relationship between bias and variance (of an estimator) is analogous to the relationship between *accuracy* and *precision*.

Exercise (adapted from 2015 exam)

Let the iid random variables X_1, X_2, X_3 be drawn from a distribution with mean μ and variance σ^2.

(1) If the estimator for μ,

$$\hat{\mu} = c_1 X_1 + c_2 X_2 + c_3 X_3$$

is unbiased, then what relation must the constants c_1, c_2, c_3 satisfy?

(2) Find, with proof, the values of c_1, c_2, c_3 such that $\mathrm{Var}(\hat{\mu})$ is minimized.

Answers

(1) $\mathsf{E}(\hat{\mu}) = c_1\,\mathsf{E}(X_1) + c_2\,\mathsf{E}(X_2) + c_3\,\mathsf{E}(X_3) = (c_1 + c_2 + c_3)\mu$, so $c_1 + c_2 + c_3 = 1$.

(2) $\mathsf{Var}(\hat{\mu}) = (c_1^2 + c_2^2 + c_3^2)\sigma^2$.

To minimize $c_1^2 + c_2^2 + c_3^2$ subject to the constraint $c_1 + c_2 + c_3 = 1$, we may use Lagrange multipliers (among other methods).

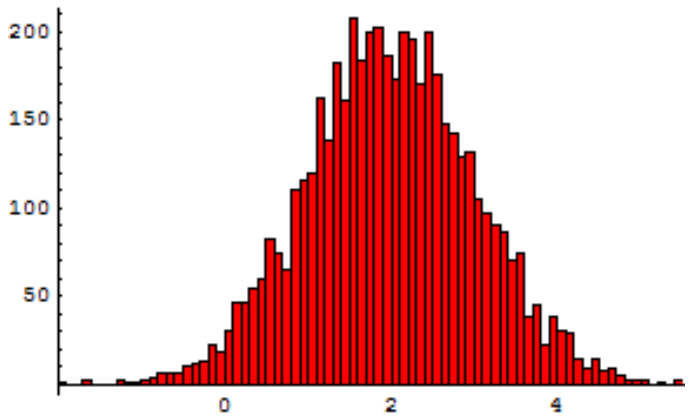
The minimum is achieved when $c_1 = c_2 = c_3 = \dfrac{1}{3}$, that is, when $\hat{\mu} = \bar{x}$.

Outline

Quality control – toy example

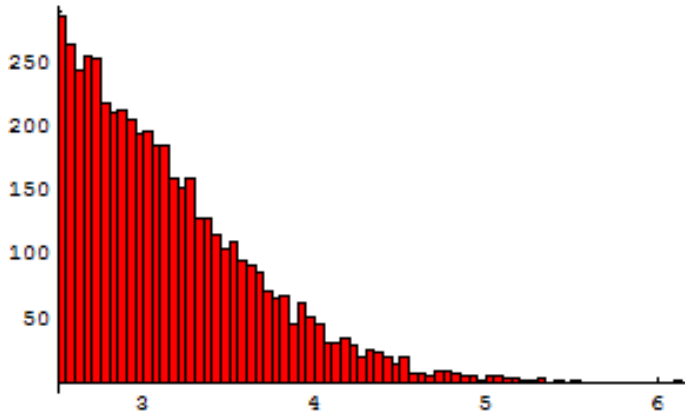
You ask a supplier to give you 5000 of their products, selected at random, to test if the mean weight is 2.5. You know that the weight is normally distributed with variance σ^2.

You plot the weights on a *histogram*.

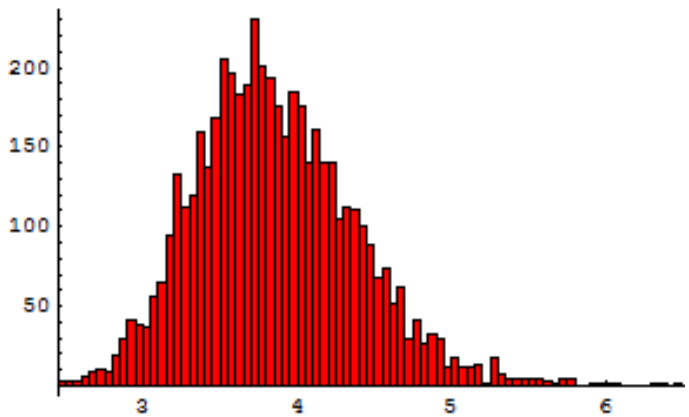


So they fail. The next month, they again give you 5000 products, but not completely randomly: only products with weight > 2.5 are selected.

Their scheming is discovered using another histogram.



The month after, they grow more cunning, and each product given actually has the *maximum* weight out of a batch of 10.



The histogram looks normal. . . How to uncover their cheating?

Answer: use a Q-Q plot, or study the *variance*.

Chi-squared distribution

The CLT gives the approximate distribution for the sample mean when the sample size is large. Unfortunately, there is no such theorem for the sample variance drawn from an arbitrary distribution.

However, if the distribution is *normal*, then the behaviour of the sample variance s^2 is well-understood, in terms of the **chi-squared distribution**.

Chi-squared random variable

A chi-squared random variable with n *degrees of freedom*, denoted by χ^2_n, is defined as the sum of squares of n iid standard normal random variables.

Chi-squared – pdf

The probability density function of a chi-squared random variable with n degrees of freedom is given by

$$\frac{1}{2^{n/2}\,\Gamma(n/2)}\,x^{n/2-1}e^{-x/2}.$$

- Note 1: $\Gamma(n)$ denotes the Gamma function, which is a continuous interpolation of the factorial function. $\Gamma(n+1) = n\,\Gamma(n)$, with $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$.
- Note 2: although we will not directly work with the above formula, the existence of a closed form for the pdf means it is easy to implement in computer programs, and hence useful in actual calculations.

Chi-squared – proof

Step 1: let Z be a standard normal r. v. Let F be the cdf of Z^2 and Φ be the cdf of Z. Then

$$F(x) = P(Z^2 \leq x) = P(-\sqrt{x} \leq Z \leq \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}).$$

Differentiating the first term and the last term with respect to x, we find that the pdf of Z^2 is

$$f(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2}.$$

Step 2: from the above pdf, we find that the *moment generating function* of Z^2 is

$$M_{Z^2}(t) = (1-2t)^{-1/2}.$$

On the other hand, from the pdf for χ^2_n, the corresponding mgf is $(1-2t)^{-n/2}$.

Step 3: use the fact that when X and Y are independent, $M_{X+Y}(t) = M_X(t)M_Y(t)$.

Chi-squared and variance

Let $X_1, \ldots, X_n$ be iid normal random variables with mean μ and variance σ^2, then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

The proof is similar to the calculation of $\mathrm{E}(s^2)$

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \sum_{i=1}^{n} \left(\frac{\bar{X} - \mu}{\sigma}\right)^2 + 0$$

$$= \frac{(n-1)s^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2.$$

LHS $\sim \chi^2_n$, last term $\sim \chi^2_1$.

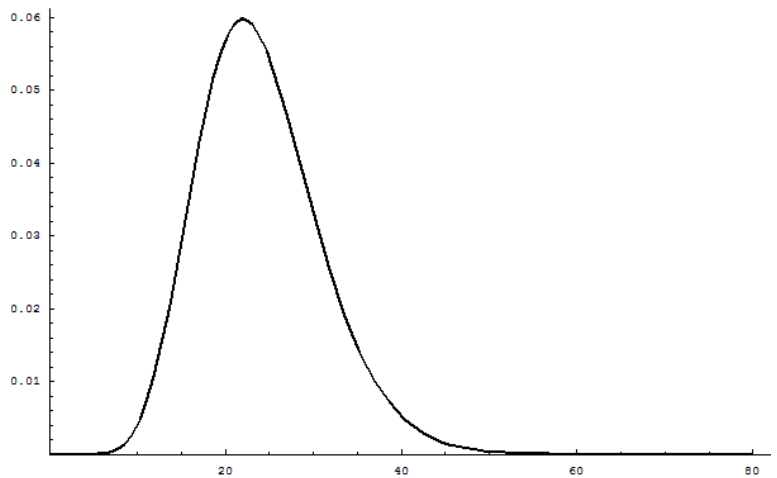
Also, it follows from the variance of χ^2_{n-1} that $\mathrm{Var}(s^2) = \dfrac{2\sigma^4}{n-1}$.

Chi-squared – exercise

The waiting times in a bank are normally distributed with a standard deviation of 8.2 minutes. What is the probability that for a random sample of 25 customers, the sample standard deviation is greater than 10 minutes?

Use the *Excel* command `chisq.dist`.

Answer: 0.0588



t-distribution

Let $X_1, \ldots, X_n$ be iid normal random variables with mean μ and variance σ^2. Let the random variable T_{n-1} be

$$T_{n-1} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}.$$

Definition: T_{n-1} follows a **Student t-distribution** with $(n-1)$ degrees of freedom.

The t-distribution is symmetric and bell-shaped, but has heavier tails than the standard normal distribution; it converges to the standard normal as $n \to \infty$.

t-distribution: formula and history

Using transformation of random variables (from Probability), we can show that the pdf for T_{n-1} is

$$f_{n-1}(t) = \frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi}\,\Gamma(\frac{n-1}{2})}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}.$$

The t-distribution was popularized by William Gosset, who was a researcher at Guinness Brewery and published under the pen name 'Student', since employees of the brewery were forbidden to publish papers lest they revealed trade secrets.

Statistics

Week 4: Confidence Intervals (Chapter 6)

ESD, SUTD

Term 5, 2017

Established in collaboration with MIT

Outline

Motivation

So far we have studied *point* estimators for a statistic (such as μ), which give a single value estimate (such as $\bar{x}$) for that statistic.

However, it would be more useful to give an *interval* estimate, so that we could make statements such as: 'most' of the time, μ will lie between the values L and U. (We will quantify this soon.)

Consider the following problem: a new training technique is believed to improve running times. After a month of training with this technique, six runners from a team recorded times of 50.1, 50.3, 50.3, 51.2, 51.5, 51.6 (in s). Is this *significantly* lower than the team average before the technique was introduced, $51.7s$?

We will look at two approaches to this problem: **confidence intervals**, and **hypothesis testing**.

Confidence interval

Confidence interval (CI)

A confidence interval is an interval estimate of a parameter θ, and can be used to indicate the reliability of an estimate.

It is an interval $[L, U]$ such that

$$P(L \leq \theta \leq U) = 1 - \alpha,$$

where $\alpha \in (0, 1)$, and L, U are functions of the sample X_i.

$(1 - \alpha)$ is called the *confidence level*.

Commonly used values for α include 0.1, 0.05 and 0.01. By convention, 0.05 is very common, though the value of α you pick should depend on the nature of the problem.

Two-sided confidence interval

Consider a random sample $X_1, X_2, \ldots, X_n$ drawn from a distribution with mean μ and variance σ^2. Suppose that n is large (e. g. $n > 30$), μ is unknown and to be estimated, but σ^2 is *known*.

From CLT, $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is approximately standard normal, and hence (approximately)

$$P(-1.96 \leq Z \leq 1.96) = 0.95.$$

Rearranging, we obtain

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

This gives a 95%, *two-sided* confidence interval *for* the mean μ,

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right].$$

Interpretation

Note 1: if we know that the population distribution is normal, then Z is exactly normal for any n.

Note 2: once the confidence interval (CI) is calculated, the true mean μ either lies inside it, or it doesn't. So, technically speaking, it is *incorrect* to say that μ lies inside $\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ with 95% probability.

One acceptable interpretation for the CI: if we repeatedly draw samples of size n from the same population, and calculate the CI using the same method each time, then the proportion of CIs that contains μ will be 95%.

One-sided confidence intervals

We can also construct *one-sided* confidence intervals.

Lower 95% CI: $P\left(\mu \geq \bar{X} - 1.645\frac{\sigma}{\sqrt{n}}\right) = 0.95,$

so the interval is $\left[\bar{X} - 1.645\frac{\sigma}{\sqrt{n}}, \infty\right).$

Upper 95% CI: $P\left(\mu \leq \bar{X} + 1.645\frac{\sigma}{\sqrt{n}}\right) = 0.95,$

so the interval is $\left(-\infty, \bar{X} + 1.645\frac{\sigma}{\sqrt{n}}\right].$

Exercises: (1) Where does the constant 1.645 come from? (Draw a picture!)

(2) Find the expressions for the 99% one-sided CIs. (Use the *Excel* command `norm.s.inv`).

Outline

Unknown variance

In most applications, however, σ^2 is *unknown*, and is estimated using s^2.

Remember that for *normal* X_i, $\dfrac{\bar{X} - \mu}{s/\sqrt{n}}$ follows a t-distribution.

Therefore, to calculate a two-sided CI for μ, we use

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

where $t_{n-1,\,1-\alpha/2}$ is a critical point (inverse of the cdf) of the t-distribution with $(n-1)$ degrees of freedom.

In *Excel*: use `t.inv`$(1-\alpha/2,\ n-1)$.

Unknown variance, continued

$s/\sqrt{n}$ is called the standard error.

Using the t-distribution results in wider CIs (the distribution has heavier tails; using s instead of σ introduces more uncertainty).

One-sided confidence intervals can be similarly obtained.

Whether to use two-sided or one-sided CI depends on the context of the problem.

Example: suppose 14 sheets of rubber have sample mean strength $\bar{x} = 33.7$ and sd $s = 0.80$. If you wish to make a statement like 'with 95% confidence, the population mean strength of the rubber is at least L', then we require a one-sided confidence interval, with

$$L = \bar{x} - t_{13,\,0.95}\,\frac{s}{\sqrt{14}} \approx 33.3$$

t-distribution vs z-distribution

Short summary:

If σ^2 is **known**, and the population is **normal**, then we use the z-distribution (standard normal) to find the CI.

If σ^2 is **known**, the population is not normal but n is large, CLT allows us to also use the z-distribution.

If σ^2 is **unknown**, and the population is **normal**, then we use the t-distribution to find the CI.

If σ^2 is **unknown**, the population is not normal but n is large, we may again apply the CLT and observe that $s \approx \sigma$. However, a *conservative* approach is to still use the t-distribution for this case.

Outline

Confidence interval for variance

Suppose $X_1, X_2, \ldots, X_n$ are random samples drawn from a normal population with variance σ^2.

Recall that $\dfrac{(n-1)s^2}{\sigma^2}$ is a χ^2_{n-1} random variable. Therefore, to find a $(1-\alpha)$ CI for σ^2:

$$P\left(\chi^2_{n-1,\,\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{n-1,\,1-\alpha/2}\right) = 1-\alpha,$$

$$P\left(\frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}}\right) = 1-\alpha.$$

Here $\chi^2_{n-1,\,\alpha/2}$ is a critical point of the distribution. In *Excel*: use `chisq.inv(`$\alpha/2$, $n-1$`)`.

One-sided CIs can be similarly obtained.

Exercise

A bottling company uses a filling machine, and the amount filled is normally distributed. Based on 16 samples, the sample standard deviation of the amount filled is 0.700ml. Find a 95% two-sided confidence interval for σ.

Statistics

Week 4: Hypothesis Testing (Chapter 6)

ESD, SUTD

Term 5, 2017

Outline

Hypothesis

A **hypothesis** is a claim. In *hypothesis testing*, we attempt to answer the following:

Given some data from a sample, does it provide statistically significant evidence to prove (beyond reasonable doubt) a hypothesis about the population, or could it have arisen due to random chance?

As a generic example, a hypothesis could be that a particular treatment has a real effect (e. g. better than an existing treatment, or placebo, or doing nothing).

Null and alternative hypotheses

More specifically, using the sample data, we test the validity of a claim about the population, against a counter claim. We set up these two competing claims as follows:

- The **null** hypothesis, H_0, is the claim of no difference or no effect; usually, H_0 is the status quo.
- The **alternative** hypothesis, H_1, is the claim that there is a difference or effect (usually it is the claim you are interested to prove).

Rejecting the null hypothesis is a primary task in scientific research.

Exercise: write down H_0 and H_1 for the training technique example from last class.

'Proof' by contradiction

The standard approach is to first *assume* H_0 *is true*. Then, perform a calculation to determine whether the data *contradicts* this assumption beyond reasonable doubt.

- If Yes, then reject H_0. We may also accept H_1.
- If No, then do not reject H_0. We cannot rule out H_0 as an explanation for the data, but we have not proven it either. So we *do not accept either* hypothesis.

So if we fail to prove H_1, then it *may* be because H_0 is true, or it *may* be the case that H_1 is true, but there is *insufficient* information to rule out random chance as an alternative explanation for the data.

In this case, we take the conservative stance and 'do not reject' H_0 – the data, after all, may still be consistent with null hypothesis.

Analogies

Analogy 1: in most legal systems, a person is assumed innocent until proven guilty. The burden of proof is on the one who makes the (extraordinary) claim that the person is guilty.

H_0: innocent; H_1: guilty.

If there is not enough evidence to establish guilt, it does not prove that the person is innocent.

Analogy 2: in general, H_0 is usually a negative statement, such as 'telepathy does not exist', and it is very hard to prove negative statements. However, a person who makes the (extraordinary) claim that he is telepathic (H_1) needs to prove it.

'Extraordinary claims require extraordinary evidence.'

Example

Example: a sample of 50 tins of tomatoes are tested, to see if their average weight deviates from the acceptable value of $\mu_0 = 350$g. State the hypotheses.

Answer: $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$.

Suppose the weights satisfy $\sigma = 10$ and $\bar{x} = 355.2$. Take 'statistically significant' to mean 95% confidence.

Assuming that H_0 is true, we have

$$P\left(\mu_0 - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95,$$

$$P\left(350 - 1.96\frac{10}{\sqrt{50}} \leq \bar{X} \leq 350 + 1.96\frac{10}{\sqrt{50}}\right) = 0.95.$$

Connection with CI

Assuming H_0, then with 95% probability, the sample mean lies between 347.2 and 352.8. As $\bar{x} = 355.2$, we reject H_0 (at the 5% significance level) and accept H_1.

Note that the inequalities on the last slide are *equivalent* to those involved in the confidence interval calculation for μ. This relationship also holds for one-sided tests and one-sided CIs.

Hypothesis test for μ

We reject H_0 at *significance level* α if and only if μ_0 falls outside the appropriate $(1 - \alpha)$-level CI for μ.

Meaning of α: type I error

The significance level α is the (maximum) probability of accepting H_1 when H_0 is in fact true.

This type of error is known as a **type I error**, or a false positive.

Examples: (1) An innocent person is convicted to be guilty.

(2) A test shows a patient to have a rare disease when in fact she does not have it.

(3) A spam filter wrongly classifies a legitimate email as spam.

During an experimental set up, and before any hypothesis test is performed, we need to clearly specify H_0, H_1, as well as α.

Type II error and power

A **type II error** occurs when a test fails to reject H_0 when H_1 is actually true. It is also known as a false negative. Its probability is denoted by β.

Examples: (1) Baggage screening in airport security fails to pick up explosives.

(2) A person is guilty but the courtroom fails to identify it.

Exercises: (a) Is one type of error always more serious than the other?

(b) What does $(1 - \beta)$ represent?

$(1 - \beta)$ is called the *power* of a test. Usually a power of 80% is acceptable; 90% is desirable.

p-value

We have seen how to perform a hypothesis test using a CI.

Another approach to hypothesis testing is to ask the question: What is the probability of observing a sample statistic *at least as extreme* as the one observed, assuming H_0 is true?

Intuition for using 'at least as extreme': think of it as an area outside a confidence interval.

This probability is known as the p-value. If the p-value $\leq \alpha$, then reject H_0.

We have already computed a p-value back in Week 1.

Exercise: Compute the p-value for the tomatoes example.

p-value, properties

- The smaller the p-value, the more significant is the test result. Therefore, it is a good practice to quote the p-value after you perform a hypothesis test.
- The p-value is also the smallest α at which H_0 can be rejected.
- The p-value computation may be one- or two-sided, depending on the hypotheses.
- Sometimes the p-value is quoted as a number of standard deviations away from the mean in a normal distribution.

  For example, the 2012 discovery of the Higgs boson has a significance of 5 sigma (p-value $\approx 1/3.5$ million); $n \approx 300$ trillion proton-proton collisions were analyzed.

Statistics

Week 5: Inference for Single Samples (Chapters 6 & 7)

ESD, SUTD

Term 5, 2017

Outline

Inequalities in H_0

For a one-sided alternative hypothesis, such as $H_1 : \mu > \mu_0$, it does not matter if we use

$$H_0 : \mu = \mu_0 \quad \text{or} \quad H_0 : \mu \leq \mu_0.$$

If we use the latter, then the maximum p-value is still obtained at the boundary, when $\mu = \mu_0$.

One-sided tests are used when the deviation is expected to be in a particular direction. They should *not* be used as a device to make a statistically non-significant result significant.

Hypothesis test: a summary

Here are some equivalent ways to perform a hypothesis test for the mean μ, assuming that σ is known and the sample size is large:

- Calculate the appropriate $(1-\alpha)$-level confidence interval around $\bar{x}$, and check if μ_0 falls outside it.
- Calculate the appropriate p-value and compare it with α.
- Calculate the z-statistic, $(\bar{x}-\mu_0)/(\sigma/\sqrt{n})$, and compare it with the appropriate critical value ($z_{1-\alpha}$ or $z_{1-\alpha/2}$).

Hypothesis test: exercise

The procedure is similar if σ is unknown but the population is normal: we use the t-distribution instead.

Exercise: suppose that you selected a random sample of 36 SUTD students, and found that on average, they spend 20.0 hours on homework per week, with a sample standard deviation of 3.0 hours. Assume normality.

For the hypotheses $H_0 : \mu = 19$ vs $H_1 : \mu > 19$,

(1) Find the p-value.

(2) Can H_0 be rejected if $\alpha = 5\%$? $\alpha = 1\%$?

Outline

Power: exercise 1

Exercise: previous research showed that the amount of time children spend watching TV per week had $\mu = 22.6$h and $\sigma = 6.1$h. A market research firm believes that the stated mean is now too low. A random sample of 60 children are taken to measure the number of hours they watch TV. A hypothesis test at the $\alpha = 0.01$ level is carried out.

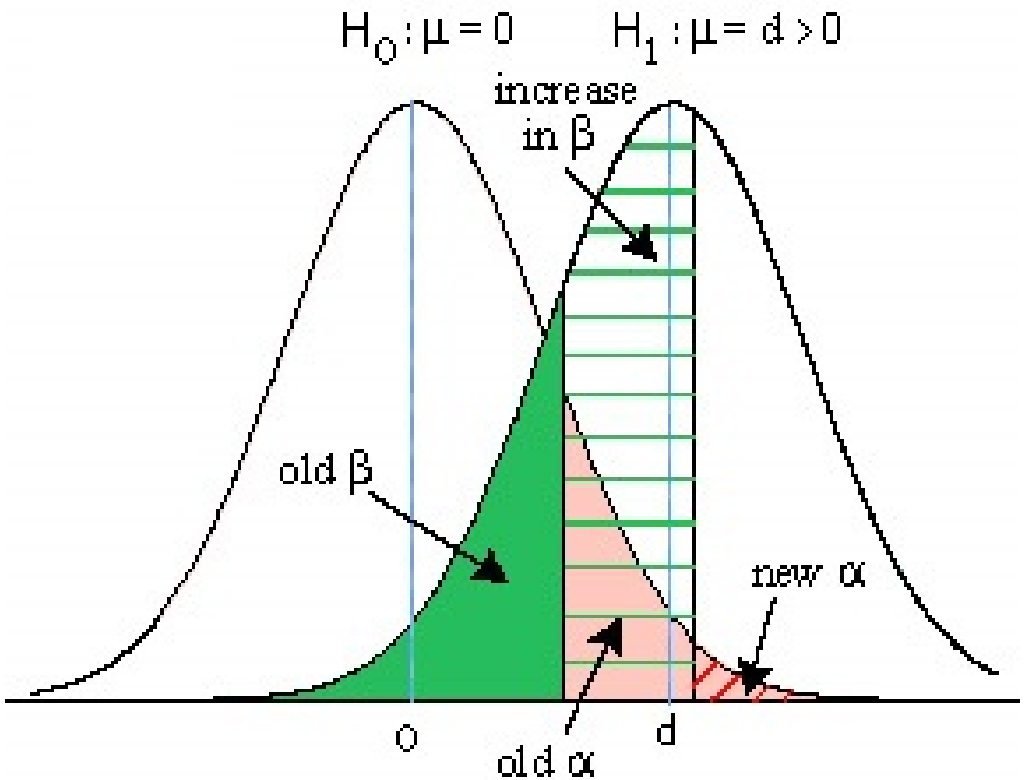
(1) State H_0 and H_1.

(2) Can we use the CLT?

(3) Suppose the *true* mean for this population is 25 hours. What is β, and what is the power in this case? (Draw a picture!)

α and β

α and β *cannot* be reduced simultaneously, unless we increase the sample size.



Power calculation – formula

Assume that σ is known, and that n is large so we may use the z-distribution.

Consider the problem of testing $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$. Then the power $(1 - \beta)$, as a function of μ, is given by

$$1 - \beta = \Phi\left(\frac{(\mu - \mu_0)\sqrt{n}}{\sigma} - z_{1-\alpha}\right).$$

Proof: generalize from Exercise 1. You should also figure out the corresponding formulas for $H_1 : \mu < \mu_0$ and $H_1 : \mu \neq \mu_0$

Note: in situations where we need to use the t-distribution, the power calculation is less straightforward.

Sample size determination – part 1

We can now relate the required sample size to α and β.

With the assumptions on the previous slides, the minimum sample size required for an α-level hypothesis test with power of $(1-\beta)$ is

$$n = \left(\frac{(z_{1-\alpha} + z_{1-\beta})\sigma}{\mu - \mu_0} \right)^2,$$

rounded to the next integer.

Sample size determination – part 2

Consider a $(1-\alpha)$ two-sided confidence interval for μ using the z-distribution. What is the relationship between the width of the interval and the sample size?

If the width of the CI is $2E$, then we require the minimal sample size to be

$$n = \left(\frac{z_{1-\alpha/2}\,\sigma}{E} \right)^2,$$

rounded to the next integer.

Exercise: Find the required sample size for a 95% CI, whose width is $\sigma/4$.

Sample size and error

- 'Conservative' tests tend to have lower α, and hence higher β.
- Discovery of exoplanets using time series data have very high α (by some estimates, as high as 0.5).
- Many popular science programs on TV use very small sample sizes (such as $n <= 3$), and thus their results are unreliable (large α or large β).

Side question: all the exoplanets discovered are much larger than earth. Does this make earth an outlier in terms of planetary size?

Power: exercise 2

Changes in test scores for students retaking the SAT without coaching has $\mu = 15$ and $\sigma = 40$. The changes in the scores are roughly normally distributed. A coaching program claims that on average it can improve the mean score by at least 35 points. A 0.01-level test of $H_0 : \mu = 15$ vs $H_1 : \mu > 15$ is to be conducted. Find the number of students that must be tested in order to have at least 90% power for detecting an increase of 35 points or more.

Statistics

Week 5: Inference for Two Samples (Chapter 8)

ESD, SUTD

Term 5, 2017

Outline

Independent samples design

Suppose we wish to check if two treatments are significantly different, e. g., compare the effectiveness between two teaching methods, or investigate the salary gap between men and women.

In the 1st case, we can divide a class randomly into two groups, and use different methods to teach them. In the 2nd case, we can randomly sample some men and women from the same company.

Such situations can be modeled as follows. Take random samples from two populations:

Sample 1: $x_1, x_2, \ldots, x_n$

Sample 2: $y_1, y_2, \ldots, y_m$

The two samples are statistically *independent*, and n, m do *not* necessarily equal.

Graphical methods

Before doing any statistical analysis, one should investigate the two samples graphically, for example using

- Side-by-side box plots,
- A Q-Q plot. This is particularly easy if $n = m$, since it involves plotting $x_{(i)}$ vs $y_{(i)}$.

Why is a scatter plot not a good idea?

Compare the means

Suppose the two populations have means μ_1 and μ_2, and standard deviations σ_1 and σ_2, all of which are unknown. We are interested in the difference $\mu_1 - \mu_2$.

Consider the sample means $\bar{X}$ and $\bar{Y}$. We have:

$$\mathsf{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2,$$

$$\mathsf{Var}(\bar{X} - \bar{Y}) = \mathsf{Var}(\bar{X}) + \mathsf{Var}(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}.$$

(Make sure that you understand why these formulas are true.)

Large samples, CI

When both n and m are *large*, the random variable

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

is approximately standard normal, as a consequence of the CLT.

Since n and m are large, we can approximate σ_i by s_i.

Therefore, the $(1-\alpha)$ two-sided confidence interval for $\mu_1 - \mu_2$ is given by:

$$\bar{x} - \bar{y} - z_{1-\alpha/2}\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + z_{1-\alpha/2}\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}.$$

Hypothesis testing

We are often interested in testing
$H_0 : \mu_1 - \mu_2 = \delta_0$ vs $H_1 : \mu_1 - \mu_2 \neq \delta_0$,
where δ_0 is some specified value (often, $\delta_0 = 0$ is used).

We may use the CI to perform the hypothesis test. Alternatively, we can compute the statistic

$$z = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}},$$

and find the p-value, given by $P(|Z| \geq |z|)$.

Reject H_0 if the p-value is less than the predetermined value of α.

One-sided confidence intervals and hypothesis testing can be done similarly (try writing down the formulas yourself).

Further reading

What if n and m are not large? Then the formulas become more involved, and depend on whether the populations variances equal.

You can read about them in Section 8.3 of the textbook, under the heading 'Inferences for small samples'. This section is not a part of the course, but you may find it useful for your project.

Outline

Introduction

Independent samples design has a disadvantage: randomization may not ensure that the two groups are equal on all attributes (except for the treatment).

To overcome this, we can use **matched pairs design**. Form n matched pairs, (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n). Each pair is chosen to be *as similar as possible*, thus improving precision of the study.

Examples: to investigate the salary gap, for each pair, choose 1 man and 1 woman with the same job, qualification and experience.

To compare customers' preference between two types of coffee, ask each customer to drink both types.

To compare the rate of respiratory problems between smokers and non-smokers, find identical twins, only one in each pair smokes.

Drawback: it may not be easy or possible to form matched pairs.

Mean and variance

Assume that $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_i \sim N(\mu_2, \sigma_2^2)$, and that n is not necessarily large.

Now that the samples are paired up, it makes sense the consider the normal random variables $D_i = X_i - Y_i$.

$$\mathsf{E}(D_i) = \mu_1 - \mu_2,$$

$$\mathsf{Var}(D_i) = \mathsf{Var}(X_i) + \mathsf{Var}(Y_i) - 2\,\mathsf{Cov}(X_i, Y_i),$$

which is usually *smaller* than the sum of the variances, if the matching is done successfully. (Why?)

Consider the observed values x_i and y_i. Let

$$d_i = x_i - y_i, \quad \text{and} \quad \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i.$$

Confidence interval

Also, compute the sample standard deviation of the d_i,

$$s_d = \left[\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2\right]^{1/2}.$$

As $\dfrac{\bar{d} - (\mu_1 - \mu_2)}{s_d/\sqrt{n}}$ follows a t-distribution, the $(1-\alpha)$ two-sided confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{d} - t_{n-1,\,1-\alpha/2}\,\frac{s_d}{\sqrt{n}} \le \mu_1 - \mu_2 \le \bar{d} + t_{n-1,\,1-\alpha/2}\,\frac{s_d}{\sqrt{n}}.$$

As the standard error here is expected to be less than the one in independent samples design, we expect matched pairs to give smaller CI.

Hypothesis testing

The CI can be use for hypothesis testing. Alternatively, compute

$$t = \frac{\bar{d} - \delta_0}{s_d/\sqrt{n}},$$

and then find the p-value.

Exercise:

1. Open the *Excel* file. We want to check if the body temperatures of males and females are different, using $\alpha = 0.05$. Which type of experimental design is this?
2. Set up H_0 and H_1 and perform the hypothesis test.

Statistics

Week 6: Inference for Two Samples (Chapter 8)

ESD, SUTD

Term 5, 2017

Outline

Independent samples design, small sample size

Suppose $x_1, x_2, \ldots, x_n$ come from an $N(\mu_1, \sigma^2)$ distribution, and $y_1, y_2, \ldots, y_n$ come from an $N(\mu_2, \sigma^2)$ distribution; n is not necessarily large.

Note: we are assuming that the variances are the same.

Then the $(1-\alpha)$ two-sided confidence interval for $\mu_1 - \mu_2$ is:

$$\bar{x} - \bar{y} - t_{2n-2,\,1-\alpha/2}\sqrt{\frac{s_1^2 + s_2^2}{n}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + t_{2n-2,\,1-\alpha/2}\sqrt{\frac{s_1^2 + s_2^2}{n}}.$$

General rule for calculating the *degree of freedom*: total number of data points minus the number of contraints.

Comparing variances

Therefore it is useful to test whether two populations have the same variance.

Set up: assume that $x_1, x_2, \ldots, x_{n_1}$ come from an $N(\mu_1, \sigma_1^2)$ distribution, and $y_1, y_2, \ldots, y_{n_2}$ come from an $N(\mu_2, \sigma_2^2)$ distribution.

To test whether the variances are the same, we use the ratio σ_1^2/σ_2^2, which is estimated using s_1^2/s_2^2.

Terminology: when not all the random variables in a collection have the same variance, they are called *heteroscedastic*.

Snedecor's F distribution

Define the random variable

$$F_{u,v} = \frac{\chi^2_u/u}{\chi^2_v/v}.$$

This random variable has an F distribution with degrees of freedom u and v.

Its probability density function is

$$F_{u,v}(x) = \frac{\Gamma((u+v)/2)}{\Gamma(u/2)\Gamma(v/2)} \left(\frac{u}{v}\right)^{u/2} x^{u/2-1} \left(1 + \frac{u}{v}x\right)^{-(u+v)/2}.$$

Comparing variances – CI

Recall that $\dfrac{(n_i-1)s_i^2}{\sigma_i^2}$ is a $\chi^2_{n_i-1}$ random variable.

Therefore $\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is an $F_{n_1-1,\,n_2-1}$ random variable.

It follows that

$$P\left(f_{n_1-1,\,n_2-1,\,\alpha/2} \le \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \le f_{n_1-1,\,n_2-1,\,1-\alpha/2}\right) = 1-\alpha,$$

where $f_{n_1-1,\,n_2-1,\,x}$ is a critical point, given by `F.INV`$(x,\ n_1-1,\ n_2-1)$ in *Excel*.

Hence, a two-sided $(1-\alpha)$ CI for σ_1^2/σ_2^2 is

$$\left[\frac{1}{f_{n_1-1,\,n_2-1,\,1-\alpha/2}}\frac{s_1^2}{s_2^2},\ \frac{1}{f_{n_1-1,\,n_2-1,\,\alpha/2}}\frac{s_1^2}{s_2^2}\right].$$

Comparing variances – example

Suppose $s_1 = 1.1$, $s_2 = 0.9$, and $n_1 = n_2 = 50$.

How do we test $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$, using $\alpha = 0.05$?

Comparing variances – example

Solution: check if 1 lies inside the confidence interval.

The CI is $[0.848, 2.63]$.

Bonus question: fixing s_i, how big must $n_1 = n_2$ be for us to reject H_0?

Answer: 98.

Outline

Non-comprehensive summary of what we covered

- Experiments and surveys: sensitive questions, sampling methods, placebo effect, control group, randomized block design, Latin square, Simpson's paradox.

  Excel: Data Analysis → `Random Number Generation`; `Sampling` (systematic; with replacement).

- Summary statistics: mean, median, standard deviation, IQR, outlier, covariance.

  Excel: Data Analysis → `Descriptive Statistics`.

- Graphical methods: bar chart, histogram, box plot, Q-Q plot, time series, EWMA.

  Excel histogram: Data Analysis → `Histogram`. EWMA: `Exponential Smoothing` (Damping factor $= 1 - \alpha$).

- Unbiased estimators: s^2, German tank problem.

- Distributions: normal, CLT, χ^2, Student t, F.

  Excel: `norm.s.inv`, `norm.s.dist`, etc.

- Confidence intervals, single sample: for mean with known and unknown σ; for variance; one and two sided.

- Confidence intervals, two samples: for mean (independent samples, matched pair); for variance.

  Excel: Data Analysis $\rightarrow$ `z-Test: Two Sample for Means`, `t-Test: Paired Two Sample for Means`

- Hypothesis testing: H_0 vs H_1; reject/do not reject H_0; one and two sided; for μ and σ; connection with CI; p-value; α and β; power and sample size calculation.

After the break

- Hypothesis testing involving proportions.
- Chi-squared test: how to show quantitatively if your data fits a certain distribution?
- Regression (trend lines): linear, multiple, logistic.
- ANOVA: when there are more than two treatments, how can we know if one is significantly better than the others?
- Non-parametric tests: bootstrap, permutation test, sign test.
- Test for randomness: how can you tell if some data (e. g. sequence of coin tosses) is made up?
- Maximum likelihood.