MA 214: Applied Statistics

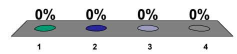
Instructor: Ashis Gangopadhyay

Introduction

First Clicker Question!!!!

I am taking MA 214 because….

1. I am going for an easy "A".
2. It is required for my major (talk about cruel and unusual punishment ☹).
3. The future belongs to quantitative knowledge and I want to be prepared.
4. No clue. Get me out of here!



1.1 Some Key Statistical Concepts

- **Statistics** – the art of data analysis, involving data collection, organization and interpretation of data
- **Population** – a collection of well-defined data that characterizes some phenomenon

Example:

The collection of all current BU students

1.1 Some Key Statistical Concepts

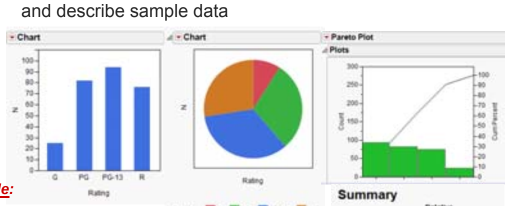
- **Sample** – a subset of the population

Example:

Only those BU students who are currently taking MA 214

1.1 Some Key Statistical Concepts

- **Descriptive Statistics** – Techniques used to organize and describe sample data



Summary

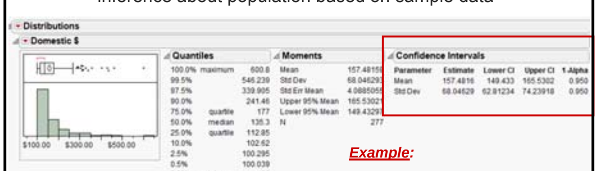
Rating	Frequency	Relative Frequency	Percentage
G	25	0.09025	9.02527
PG	82	0.29603	29.6029
PG-13	94	0.33935	33.935
R	76	0.27437	27.4368

Example:

Visual & numerical summaries of ratings of movies from JMP Sample Data set

1.1 Some Key Statistical Concepts

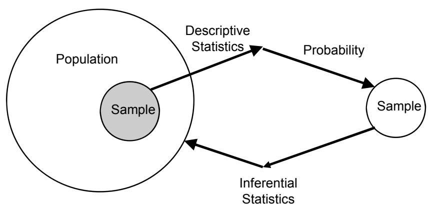
- **Inferential Statistics** – Techniques used to draw inference about population based on sample data



Example:

Confidence intervals for the average domestic profit of movies in JMP Sample Data set

1.1 Some Key Statistical Concepts



1.1 Some Key Statistical Concepts

Example:

Suppose we take a simple random sample of BU students and record their GPA:

GPA
4.0
3.1
3.3
3.6
2.0

Variable – one specific characteristic of the units in a given population

Value – a value assigned to the variable

$\mu = \text{true mean GPA}$ **Parameter** – Descriptive measure of a population

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{16}{5} = 3.2$ **Statistic** – Descriptive measure of a sample

1.1 Some Key Statistical Concepts

Example:

Suppose we take a simple random sample of BU students and record their GPA:

GPA
4.0
3.1
3.3
3.6
2.0

$\mu = \text{true mean GPA}$

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{16}{5} = 3.2$

Statistical Inference – process of making an estimate, prediction or decision about a population based on sample data

Measure of Reliability – a statement about the degree of uncertainty

$\boxed{?\ \ \overline{X} \approx \mu\ \ ?}$

1.1.1 Example: Cola Wars

“Cola Wars” is a term that describes intense competition between Coca Cola and Pepsi. The "war" is fuelled by massive advertising campaigns by the two companies, and claims of consumer preferences for one brand of cola or the other.

Suppose a blind taste test, in which the two brand names are disguised, was conducted to determine the consumer preference for each of these two brands of cola. Each of the 1000 participants were asked to state their gender, age, and their preference for either brand A or brand B.

1.1.1 Example: Cola Wars

Population	Collection of all cola consumers
Sample	1000 customers selected from the population
Variable	Age, Gender, and the preference (A or B) of cola
Inference of Interest	To generalize the consumer preference of 1000 sampled customers to the whole population of cola consumers

1.1.1 Example: Cola Wars

Suppose in the sample, 600 consumers preferred the taste of Pepsi; i.e., 60% of the consumers in the sample preferred Pepsi.

Can we conclude that 60% of the consumers in the population also preferred Pepsi?

1.1.1 Example: Cola Wars

Suppose in the sample, 600 consumers preferred the taste of Pepsi; i.e., 60% of the consumers in the sample preferred Pepsi.

Can we conclude that 60% of the consumers in the population also preferred Pepsi? ***The answer is NO!***

It does not follow, nor it is likely, that exactly 60% of the consumers in the population prefer Pepsi.

However, statistical inference techniques tell us that the "true" percentage of the population who prefer Pepsi is almost certainly within a specified limit of the sample estimate.

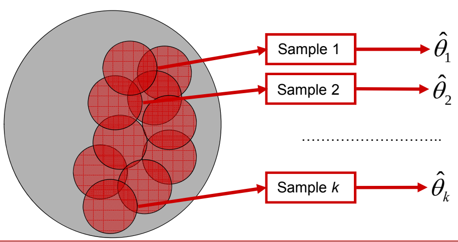
1.1.1 Example: Cola Wars

For example, our analysis may suggest that the sample estimate of the preference for Pepsi is within 3% of the "true" value; implying that the actual preference for Pepsi in the population is between 57% (= 60%-3%) and 63% (= 60% +3%).

This interval represents a measure of reliability of the inference.

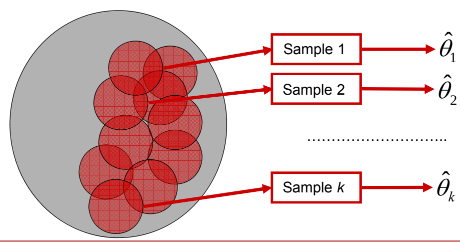
1.2 Sampling Distribution

Suppose we repeatedly draw samples of fixed size from the population and calculate sample estimate $\hat{\theta}$ of population parameter θ



1.2 Sampling Distribution

The distribution of the collection of estimates $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k\}$ is called the sampling distribution of $\hat{\theta}$



1.2.1 Central Limit Theorem (CLT): Sampling Distribution of $\overline{X}$

One of the most important results in statistics is the sampling distribution of the sample mean $\overline{X}$ as an estimate of the population mean μ

$$\hat{\theta}_1 = \overline{X}_1$$
$$\hat{\theta}_2 = \overline{X}_2$$
$$\ldots$$
$$\hat{\theta}_k = \overline{X}_k$$

Central Limit Theorem

Suppose n observations are randomly selected from ***any*** large population with mean μ and standard deviation σ. Then, for n sufficiently large,

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

1.2.1 Central Limit Theorem (CLT): Sampling Distribution of $\overline{X}$

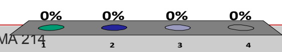
Some important observations regarding CLT

- If the population distribution is normal, then the distribution of $\overline{X}$ is normal for any sample size. However, if the distribution of the population is not normal, then for large sample size (say, for $n \geq 30$), the distribution of $\overline{X}$ is approximately normal.
- $\mu_{\overline{X}} = \mu$
- $\sigma_{\overline{X}} = \sigma / \sqrt{n}$
- $\left(\overline{X} - \mu_{\overline{X}}\right) / \sigma_{\overline{X}} = Z$ where Z is the standard normal r.v.

Clicker Question

Suppose in a large population of college-educated adults, the mean IQ is 118 with a standard deviation of 20. A random sample of 100 adults are selected from the population for a market research campaign. The distribution of sample mean IQ is

1. Approx normal with mean 118 and standard deviation 20.
2. ✓ Approx normal with mean 118 and standard deviation 2.
3. Approx normal with mean 11.8 and standard deviation 2.
4. Can't say as we don't know population distribution.



1.2.1 Central Limit Theorem (CLT): Sampling Distribution of $\overline{X}$

Example 1

- Suppose we are trying to estimate the average weight of a certain animal based on a random sample of 100 observations. Suppose it is known that the standard deviation of the weight is 5. What is the probability that the estimate is within 2 units of the true mean?

1.2.1 Central Limit Theorem (CLT): Sampling Distribution of $\overline{X}$

Example 1

- <u>**Solution**</u>:

$$\text{Note } \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{100}} = 0.5$$

$$\text{So } P\left(-2 \le \overline{X} - \mu \le 2\right) = P\left(\frac{-2}{\sigma_{\overline{X}}} \le \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \le \frac{2}{\sigma_{\overline{X}}}\right)$$

$$= P\left(-4 \le Z \le 4\right) \approx 1$$

- So we are guaranteed to have the sample estimate be within 2 units of the true mean.

1.2.1 Central Limit Theorem (CLT): Sampling Distribution of $\overline{X}$

Example 2

- Suppose a bank manager wants to find out the "true" rate at which customers arrive within a 10 minute period at the branch between 12 – 1 PM. The manager collected data between 12 – 1 PM for 30 consecutive days (i.e. *n* = 180), and found $\overline{X} = 8.1$ arrivals per 10 minutes with a standard deviation $s = 2.99$. Does this data support the manager's hypothesis that the actual mean arrival rate $\mu = 9$?

1.2.1 Central Limit Theorem (CLT): Sampling Distribution of $\overline{X}$

Example 2

- <u>**Solution**</u>: The question is that if the hypothesis is true, and the mean $\mu = 9$, then is it possible to observe a sample mean $\overline{X}$ to be 8.1 or less?

$$P\left(\overline{X} \le 8.1\right) = P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \le \frac{8.1 - 9}{0.223}\right) = P\left(Z \le -4.04\right) \approx 0$$

2 Confidence Interval

- A **confidence interval** estimate provides a range of plausible estimates of the parameter of interest

- We use sample information to construct a confidence interval, which, with (1-α)*100% confidence, contains the true parameter. Thus if an experiment is repeated, and a 95% confidence interval is constructed each time, then we should expect 95% of those intervals to cover the true value of the parameter

2.1 General form of a confidence interval

- Often, confidence intervals are given by the form

  Parameter Estimate $\pm$ Maximum Error

  Note: "Maximum Error" is also called "Margin of Error"

- Maximum Error depends on the sampling distribution of the parameter of interest. Suppose we want to estimate a parameter θ, and the sampling distribution of its estimator $\hat{\theta}$ is normal, then a general form of a confidence interval is given by $\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$

2.1 General form of a confidence interval

- In particular: a large sample ($n \geq 30$) confidence interval for a population mean μ is given by:

$$\overline{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

- Note that a narrower interval is better than a wider interval, as a narrow interval contains more information
  - Larger confidence level leads to wider interval
  - Larger sample size leads to narrower interval

2.1 General form of a confidence interval

- **Example:** Suppose a software company wishes to estimate the average number of employees absent per day. A review of records from the last 100 days revealed that the average number of employees absent per day was $\overline{X} = 5.1$ with a standard deviation of $s = 2.0$. Compute a 95% confidence interval for μ, the true average number of employees absent per day.

2.1 General form of a confidence interval

- **Solution:** Since 1 – α = 0.95, $z_{\alpha/2} = z_{0.025} = 1.96$

  Thus, a 95% confidence interval is given by:

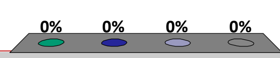
$$5.1 \pm 1.96\left(\frac{2}{\sqrt{100}}\right) = (5.1 - 0.39, 5.1 + 0.39) = (4.71, 5.49)$$

  Is it possible that the true average number of employees absent per day is more than 7?

Clicker Question…..

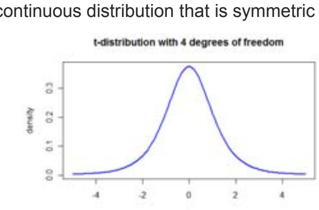
Which of the following actions would reduce the width of a confidence interval. Choose the best answer.

1. Increase the sample size.
2. Decrease the confidence level
3. Increase the sample size AND decrease the confidence level.
4. ✓ All of the above.



2.2 Student's t distribution

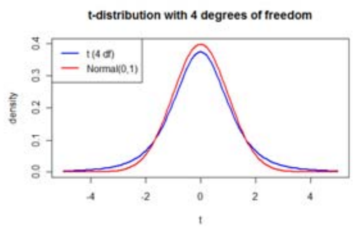
- A continuous distribution that is symmetric about zero



- Characterized by a parameter *v*, called *degrees of freedom* (*df*). Degrees of freedom *v* assumes positive integer values.

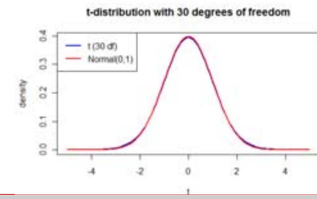
2.2 Student's t distribution

- A t-distribution with low *df* has longer tail compared to standard normal distribution.



2.2 Student's t distribution

- As the *df* increases, the tails of the t-distribution get shorter. For $v \geq 30$, there is almost no difference between the t-distribution and the standard normal distribution.



2.2 Student's t distribution

- *Fact*: If a random sample of size *n* is selected from a *normal population* with mean μ, the distribution of

  the statistic $\dfrac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ has a t-distribution with (n-1)

  degrees of freedom.

2.2 Student's t distribution

- How do we verify that a population is normally distributed?
  - Histogram
  - Normal Probability Plot
  - Various tests for normality (more on this later)

2.3 Small sample confidence interval for μ

- Assuming that the sample observations are from a normal population, a (1-α)*100% confidence interval for μ is given by

$$\overline{X} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

2.3 Small sample confidence interval for μ

- **Example:** Suppose a statistics professor wants to estimate the average number of classes the undergraduate students miss each semester. She randomly selected ten undergraduate students and asked them how many times they missed last semester. The students' responses are as follows: 4,7,2,0,1,0,10,2,0,3.
- Use the data to find a 95% confidence interval for μ, the average number of classes the students were absent last semester.

2.3 Small sample confidence interval for μ

- **Solution:** Note $\sum X = 29, \sum X^2 = 183$, Thus

$$\overline{X} = \frac{29}{10} = 2.9$$

$$s^2 = \frac{1}{n-1}\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right] = \frac{1}{9}\left[183 - \frac{29^2}{10}\right] = 10.98$$

$$s = \sqrt{10.98} = 3.31$$

2.3 Small sample confidence interval for μ

- Also $(1-\alpha) = 0.95, t_{\frac{\alpha}{2}} = t_{0.025} = 2.262$ based on *df* = 9

- 95% confidence interval for mean is given by:

$$2.9 \pm 2.262 * \left(\frac{3.31}{\sqrt{10}}\right) = 2.9 \pm 2.37 = (0.53, 5.27)$$

2.4 Confidence interval for population proportion *p*

- (1-α)*100% confidence interval for *p* is given by

$$\widehat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

where $\widehat{p} = \frac{x}{n}$ is the sample proportion

2.4 Confidence interval for population proportion *p*

- **Note:** Sample size should be large enough such that

$n\widehat{p} \geq 15$ and $n(1-\widehat{p}) \geq 15$

- *What do we do if the above condition is not met?*

Use modified sample proportion $\widetilde{p} = \frac{x+2}{n+4}$

and construct confidence interval by replacing $\widehat{p}$ by $\widetilde{p}$

2.4 Confidence interval for population proportion *p*

- **Example:** In a study of 100 accidents that required treatment in an emergency room, 60 involved children under the age of 10. Find the 95% confidence interval of the true proportion of accidents that involve children under the age of 10.

2.4 Confidence interval for population proportion *p*

- **Solution:** Is the sample size adequate?
- # of successes (x = 60) ≥ 15
- #of failures (n – x =40) ≥ 15
- Thus the sample size is adequate.

2.4 Confidence interval for population proportion *p*

- **Solution:**

  95% confidence interval for *p* is given by:

$$0.6 \pm 1.96\sqrt{\frac{0.6 \times 0.4}{100}} = 0.6 \pm 0.096$$
$$= (0.504, 0.696)$$

2.5 Sample size determination

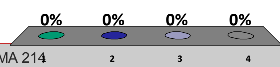
- Sample size *n* required to estimate a population mean *μ* within *E* units of its true value (i.e., estimate should be within (*μ-E, μ+E*)):

$$n = \left[\frac{z_{\frac{\alpha}{2}}\sigma}{E}\right]^2$$

Clicker Question…….

In determining the required sample size to estimate a population mean, which of the following can be ignored?

1. Variability of the population.
2. ✓ Size of the population.
3. Desired accuracy of your estimation.
4. Level of confidence that your estimate meets the accuracy requirements.



2.5 Sample size determination

- How to find σ?
  1. Use a value of σ suggested in previous surveys on the same variable.
  2. If an approximate range, *R*, on the variable is available, then a crude estimate of σ is obtained by

$$\sigma = \frac{R}{4}$$

2.5 Sample size determination

- **Example:** We want to be 99% sure that a random sample of IQ scores yields a mean that is within 2.0 of the true mean. How large should the sample be? Assume that σ is 15.

2.5 Sample size determination

- **Solution:** Since 1-α = 0.99, $z_{\frac{\alpha}{2}} = z_{0.005} = 2.57$ (from the normal table). Thus

$$n = \left[\frac{2.57 \times 15}{2}\right]^2 = 371.52 \approx 372$$

2.5 Sample size determination

- Sample size *n* required to estimate a population mean *p* within *E* units of its true value (i.e., estimate should be within (*p-E, p+E*)):

$$n = \frac{z_{\frac{\alpha}{2}}^2 p(1-p)}{E^2}$$

- **Note:** If prior information about *p* is available, we can use that value of *p* to estimate the sample size. If no information is available, use *p* = 0.5.

2.5 Sample size determination

- **Example:** We want to estimate, with a maximum error of 0.03, the true proportion of all TV household turned to a particular show, and we want 95% confidence in our result. How many TV households should be sampled if no prior information about *p* is available?

2.5 Sample size determination

- **Solution:** $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$ Thus,

$$n = \frac{(1.96)^2 0.5(1-0.5)}{0.03^2} = 1067.11 \approx 1068$$

2.5 Sample size determination

- **Example:** Suppose in the last example it is known (from an earlier survey) that the true proportion *p* is approximately 0.4. Use this information to recalculate the sample size.

- **Solution:**

$$n = \frac{(1.96)^2 0.4(1-0.4)}{0.03^2} = 1024.43 \approx 1025$$