

COMMERCE MENTORSHIP PROGRAM

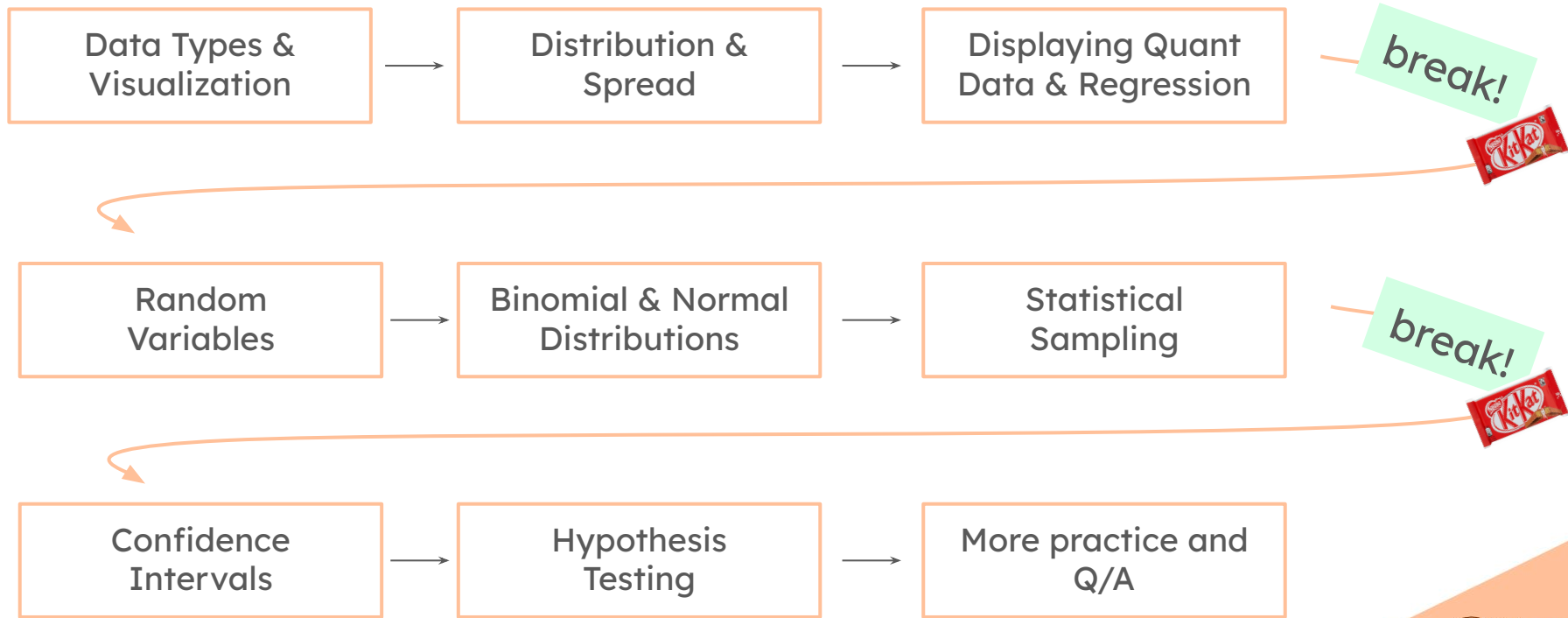
MIDTERM REVIEW SESSION

COMM 191



PREPARED BY
Rebekah Redlich

TABLE OF CONTENTS



@ubccmp



@ubccmp



cmp.cus.ca

Data & Data Visualization

Record (by a
respondent or
subject)

	Variable		Value
	Favourite Colour	Marital Status	COMM191 Score
Anthony	Red	Not Married	88
Benedict	Orange	Married	73
Collin	Yellow	Married	78
Daphne	Green	Married	85
Eloise	Blue	Not Married	65
Francesca	Indigo	Not Married	66
Gregory	Violet	Not Married	72
Hyacinth	Pink	Not Married	79

Categorical
(Counted in “bins”)

Binary

Quantitative
(Measurable in units)

Nominal
Ordinal
Identifier
String
Cross-Sectional
Time Series



@ubccmp



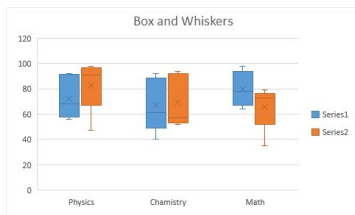
@ubccmp



cmp.cus.ca

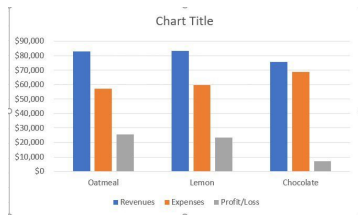
Data & Data Visualization

Box Plots



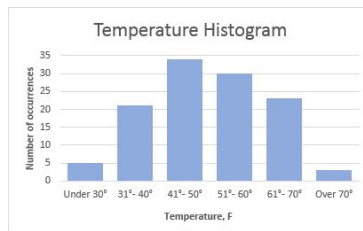
Used to analyze and compare distributions of quantitative* data.

Bar Graphs



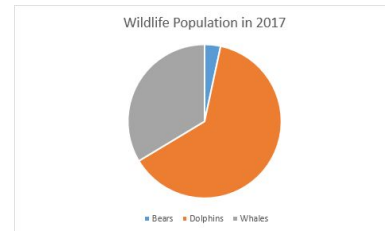
Used to show and compare frequency counts (typically in lower volumes).

Histograms



Puts data into "bins" allowing for distribution analysis, usually larger amt of data.

Pie Charts



Show % components of a whole, don't use these! Too tough to interpret.

Good Graphs: Informative headings, center alignment, limit sig figs, legible.



@ubccmp



@ubccmp



cmp.cus.ca

Data & Data Visualization

Frequency Tables: Shows the number of cases per category (can also show relativity by showing percentages).

Contingency Tables: In a contingency table, when the distribution of 1 variable is the same for all categories of another variable, they are independent.

Marginal Distribution: A type of probability distribution that reflects the relative percentage breakdowns of a contingency table.



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

Data Quality

HOW was the data collected?

WHERE did the data come from?

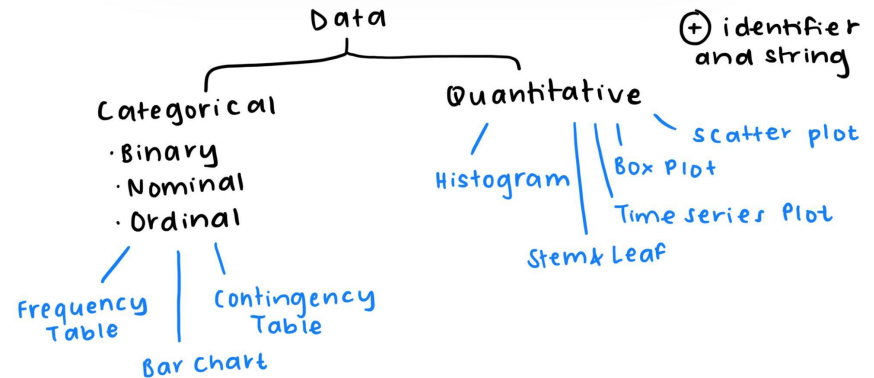
WHY are you examining the data?

WHEN was the data collected?

WHAT does each variable refer to?

WHO is being studied?

The Area Principle: The size of the area shown should always correspond with relative magnitude.



Practice Q

	Sauder	Engineering	Total
Great Dane	202	63	265
Loafe	43	109	152
Blue Chip	150	120	270
Total	395	292	687

What is the best chart to display this information?

What percent of students who prefer Great Dane are Sauder students?


What percent of Engineers prefer Loafe?


What percent of students prefer Blue Chip?

Are faculty and cafe preference independent?



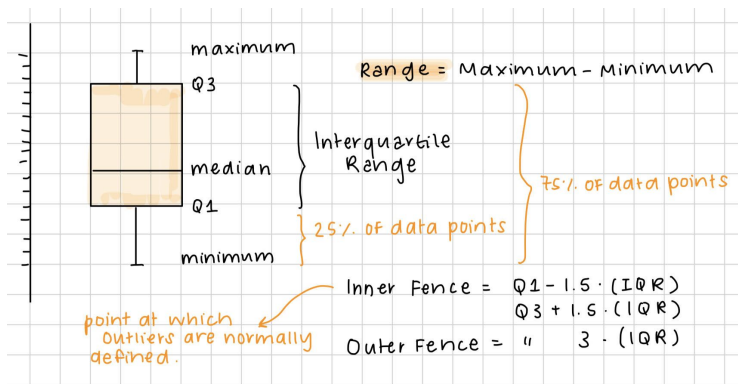
 @ubccmp

 @ubccmp

 <http://cmp.cus.ca>

Distribution

Box Plots:



Questions Related To:

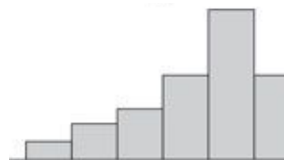
- Identifying key values
- Variance/spread - theory (boxplots are good for comparing distribution)
- Symmetry

Conceptual Check: Can you identify skew from the 5 number summary?

Histograms:

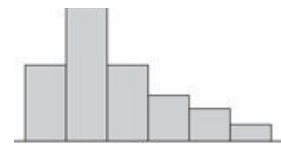
Skew: Measure of distribution asymmetry.

Left Skew (left tail):



Mode > Median > Mean

Right Skew (right tail):



Mean > Median > Mode

Questions Related To:

- Key Values (median, range, IQR)
- Variance/spread - theory
- Count/frequency



@ubccmp



@ubccmp



<http://cmp.cus.ca>

Bias

Simpson's Paradox: A phenomenon that arises when average percentages are taken across different groups which contradict the aggregate.

!! Example: UC Berkeley admits 44% of male applicants but 35% of female applicants. Yet, there was a statistically significant gender bias towards female applicants in 4/6 faculties. *Why?* Women tended to apply to more competitive programs!

Non-Response Bias: Non-participant, common among people who have less extreme opinions.

Self-Selection: Pro-participant, common among people who have more extreme opinions.

Undercoverage: Leaving out sub groups.

Bad Sample Frame: Wrong population chosen.



@ubccmp



@ubccmp



cmp.cus.ca

Central Tendency & Spread

Mean: Like the centre of gravity, the *average* value (*note that mean is limited as it does not show spread).

- Arithmetic (the one we use), harmonic and geometric.

Median: The middle value of a data set (much more resistant to outliers/robust)

Range: Maximum - Minimum Value

Variance / St.Dev: Average distance of each data point from the mean **(ALWAYS USE STEDV.S NOT STDEV.P IN EXCEL)**

- Skews and outliers will alter standard deviation and mean. To deal with skew you can...
- Take a logarithm
- Standardize


$$z \text{ score} = \frac{(\text{point} - \text{mean})}{\text{standard deviation}}$$

**Symmetric
Distribution?** Use
mean and SD

**Skewed
Distribution?** Use
median and IQR



@ubccmp



@ubccmp



cmp.cus.ca

Practice Q

In a group of 100 Taylor Swift fans, one fan's daily listening time of 70 minutes corresponds to a z-score of 0 and a second fan's daily listening time of 82 minutes corresponds to a z-score of 1.5. What is the standard deviation of the listening times?



@ubccmp



@ubccmp



cmp.cus.ca

Statistical Relationships & Scatter Plots

Association: A statistical relationship, expressed in terms of categorical variables.

Response Variable: variable—a variable that measures the outcome; also called “outcome variable” or “dependent variable” or “output variable.”

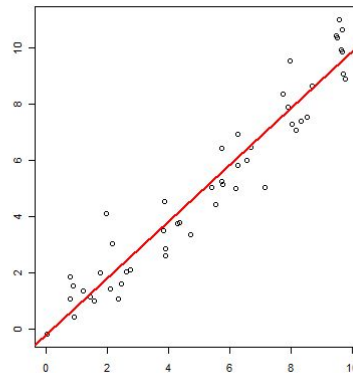
Explanatory Variable: variable—a variable that tries to explain or predict the observed outcome; also called “predictor variable” or “independent variable” or “input variable.”

Case: An individual subject on which measurements are taken and recorded.

Correlation: Linear association between two QUANTITATIVE variables.

- Each (x, y) is a case. Assessed for direction, form, strength and outliers.

👁️ **Common Exam Q:** Given categorical variables and asked if correlation (rather than association) is shown.



@ubccmp

@ubccmp

<http://cmp.cus.ca>

Regression

The regression equation, b_1 and b_0 values are important, keep them handy during the exam!

$$\hat{y} = b_0 + b_1x \text{ (with } e_i = y_i - \tilde{y}_i \text{)}$$

- Least squares regression line takes the sum of the error terms squared in order to estimate data.

Why do we use this technique? Alternatives such as minimizing absolute values etc. don't make sense.

$$b_1 = r * \frac{s_y}{s_x} , \quad b_0 = \tilde{y} - b_1\bar{x}$$

- For each value of x , the regression line passes through the average value of y .
- Symmetry \neq regression.



@ubccmp



@ubccmp



cmp.cus.ca

The Correlation Coefficient - r

- A value $-1 \leq r \leq 1$ with “standard” units (aka no unit), +/- indicates positivity/negativity of the regression line.
- CORRELATION \neq CAUSATION
- The roles of the X and Y variables are interchangeable (the correlation between X and Y is the same as Y and X).
- It is sensitive to outliers.

Checks

- ☐ Quantitative
- ☐ “Perils” of aggregated data (like Simpson’s paradox)
- ☐ Extrapolation
- ☐ Lurking variables



@ubccmp



@ubccmp



cmp.cus.ca

Regression

Residual: $e_i = y_i - \hat{y}_i$ (or observed - predicted), the vertical distance or deviation from the observation to the regression line. For every value of X, the regression line goes through the average of y.

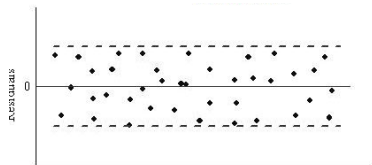
Goodness of Fit: R^2 : Fraction of variation in the y-values that is explained by the regression of Y on X ($0 \leq r^2 \leq 1$)

Influential Observation: Often thought of as horizontal or x variance, markedly changes regression!

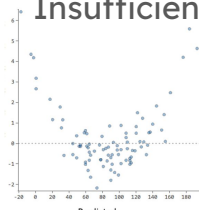
Outlier: Often thought of as vertical or y variance, doesn't follow the data pattern.

Residual Plots: *(less likely to see on exam but keep this in your notes!)*

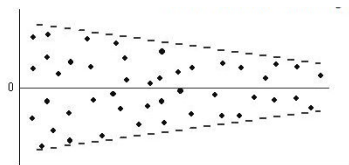
Good Fit



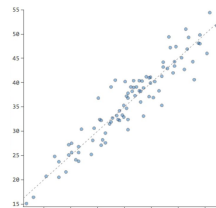
Linearity =
Insufficient



Skewness



Wrong Line!



@ubccmp



@ubccmp

[http://](http://cmp.cus.ca)

cmp.cus.ca

Additional Notes

Regression To The Mean: Regressions tend to be dense in the middle, most phenomena tend to centralize.

Regression Effect: Things fall back to the middle.

Regression Fallacy: Effect must be due to something important, not just line spread.

Law of Large Numbers: There is stability in long-term results.

Law of Averages / Gambler's Fallacy: “I’ve lost a few rounds of poker which means I’m gonna win this one!” DOESN’T WORK.

Standard Deviation of the Residuals:

$$s_e = \sqrt{\frac{\sum e_i^2}{n - 2}}$$



Practice Q

A model to predict price of a used car from mileage for a random sample of 50 cars gave the following line:

$$y = 24\,000 - 0.15x$$

For each additional 1000 miles, on average this will result in an

_____ of _____ dollars in price.

For a car with mileage of 55,000, regression would predict the price of the car to be _____.

The actual price was \$16,250. What is the residual for the data value?

What % of the variation in price is explained by mileage?



@ubccmp



@ubccmp



cmp.cus.ca

Practice Q

TRUE OR FALSE:

When X is at its mean, a regression line will yield the mean of Y .

A negative residual means that the model's predicted y -value was lower than the actual value.

If the correlation between X and Y is r then the correlation between Y and X is $-r$.

A correlation of 0 means X and Y are not related at all.



@ubccmp



@ubccmp



cmp.cus.ca

Random Variables

- Variable with a numerical outcome for some random phenomenon or set of probabilistic outcomes. **Possible outcomes & probabilities!**
- Can be discrete (categorical) or continuous (quantitative)
- If x is RV: $E(x) = \text{Mean} = \mu$ $\text{Var}(x) = \sigma^2$ (long form $\sum (x-\mu)^2 * p(x)$)
- If x is a discrete R.V.: $E(x) = \sum(x) * P(x)$

Combining Random Variables

Sum - Independent

$$E(X + Y) = E(X) + E(Y)$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\text{SD}(X + Y) = \sqrt{\text{Var}(X) + \text{Var}(Y)}$$

Difference - Independent

$$E(X - Y) = E(X) - E(Y)$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

Linear Combination - Ind.

$$E(aX + bY) = aE(X) + bE(Y)$$

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

$$\text{SD}(aX + bY) = \sqrt{a^2\text{Var}(X) + b^2\text{Var}(Y)}$$

Linear Transformation:

$$E(a + bX) = a + bE(X)$$

$$\text{Var}(a + bX) = b^2\text{Var}(X)$$

$$\text{SD}(a + bX) = |b|\text{SD}(X)$$

Linear Combination - Dep.

$$E(aX + bY) = aE(X) + bE(Y)$$

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{SD}(X)\text{SD}(Y)r,$$

***Never add standard deviations**



@ubccmp



@ubccmp



cmp.cus.ca

Practice Q

Four students work to assemble pizzas at an Italian restaurant. All staff members work independently, the time to finish each team member's task follows a normal distribution with a mean of 8 minutes and a standard deviation of 2 minutes.

What is the expected amount of time to complete one pizza?

What is the standard deviation for the time taken to complete a pizza?

The Binomial Distribution

Bernoulli Trial: *(for a Binomial Distribution)*

- ❑ Two possible outcomes
- ❑ Probability of success is p (failure is $1-p$ or q)
- ❑ Probability is the same from one trial to the next
- ❑ Trials are independent.

Uniform distribution: All outcomes are of equal likelihood.

- The Binomial distribution models a set of “ n ” bernoulli trials, where success is the statistic you’re looking for.
- x = # of successes in “ n ” independent trials, p = probability

$$\Pr (x=k) = \binom{n}{k} p^k q^{n-k} \quad \mathbf{E(X) = \mu = np, \text{Var}(X) = \sigma^2 = npq}$$



Binomial Distribution Thinking Question

Toss a fair coin n times and record the number of heads, X . If the standard deviation of X is 2, what is the expected number of heads?



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

The Normal Distribution

Models the majority of natural phenomena! Uses continuous probability.

The Standard Normal Distribution

$N(0,1)$ or mean = 0, SD = 1.

Excel

NORM.DIST = Given x, find p.

NORM.INV = Given p, find x.

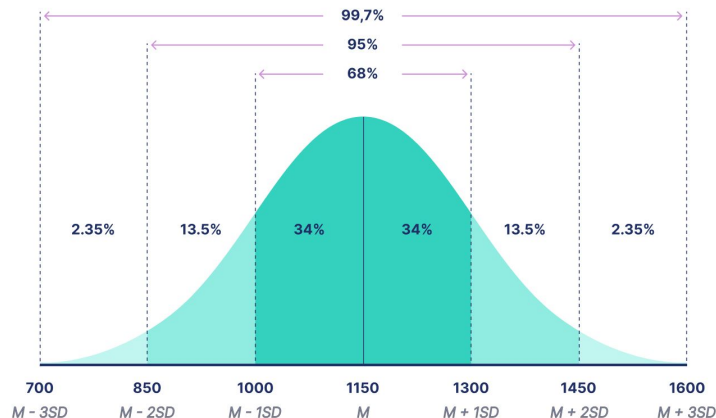
NORM.S.DIST/INV: Uses standardized distribution.

Rule of thumb - Standard Dev and Range:

$n > 20: s \approx \text{range} / 6$ $n \leq 20: s \approx \text{range} / 4$ 🐼

Remember z-scores?

The Empirical Rule



Biggest Tip: Always draw a picture 🧡



@ubccmp

@ubccmp

<http://cmp.cus.ca>

Practice Q

A cafe owner finds that the number of coffee cups sold per week is normally distributed with a mean of 800 cups and a standard deviation of 100 cups.

What is the probability that in any given week at least 880 cups will be sold?

To ensure the cafe will not run out of coffee more than 10% of the weeks, what is the minimum number of cups that should be available?

Another location has mean sales of 800 cups but an unknown standard deviation. If we know that in a randomly chosen week the probability of the number of coffee cups sold exceeding 880 is 0.1870 what is the standard deviation?



@ubccmp



@ubccmp



cmp.cus.ca

Practice Q

A survey of 24 students was done to determine how many hours they spend studying during the week. The highest studying time per week was 30 hours and the lowest was 3. Estimate the standard deviation of the data set.

What if the sample size was 18 students?

The Normal Approximation to the Binomial

- If $np > 10$ and $nq > 1$, x is binomial as well as normal via standardization.
- The chance that \bar{x} (mean of the sample) = μ (mean of population) is 0.

If X is Binomial:
$$\Pr(a < X < b) = \Pr\left(\frac{a-np}{\sqrt{npq}} < Z < \frac{b-np}{\sqrt{npq}}\right)$$



@ubccmp



@ubccmp



cmp.cus.ca

Statistical Sampling

- 1.) Examine part of the whole
- 2.) Randomization - prevents bias
 - a.) Differences = sampling variability.
- 3.) Size of the sample matters but not the size of the population
(greater sample = more accurate)

Sampling Error: Variability in data from sample to sample.

Law of Diminishing Returns: Increasing sample sizes generates declining marginal benefit (time? cost?).



@ubccmp



@ubccmp



cmp.cus.ca

Statistical Sampling Elements

Population: Universe of interest.

Sampling Frame: Portion of the population you have access to and can choose your sample from.

Parameter: Characteristic of the population.

Sample: Subset of the population.

- Your sample should be no less than 10% of your population.

Statistic: Characteristic of the sample used to *estimate* a parameter.

Census: Entire population surveyed.



@ubccmp



@ubccmp



cmp.cus.ca

Sampling Designs

Systematic: Units chosen on a regular interval.

Convenience: Easiest to reach.

Stratified: Homogeneous subgroups, individuals from each group.

Cluster: Heterogeneous groups, whole random group chosen.

Random: Random number generation.

Multi-Stage: Multiple rounds of surveying completed.

Practice Q

What sample design is being used for each of the following?

- 1) From all employee IDs, take every 20th employee starting from a randomly chosen one.
- 2) Pick the first 50 employees you meet.
- 3) From 1,000 employees, pick 75 randomly.
- 4) From female staff choose 25 at random and from male staff choose 25 at random.



@ubccmp



@ubccmp



cmp.cus.ca

The Sampling Distribution

- A probability distribution of a statistic that comes from choosing random samples of a given population.

The Sample Proportion: \hat{p} (categorical) **The Sample Mean:** \bar{x} (quant)

Where X = # of successes: $\hat{p} = \frac{X}{n}$ SD: $\sqrt{\frac{pq}{n}}$ $Z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$

- If values are random, the sampling distribution for \hat{p} is normal if $n > 30$.

Central Limit Theorem: If the population is normal then the sampling distribution of \bar{x} is exactly normal. If there is any deviation, it is approximately normal if $n > 30$.



@ubccmp



@ubccmp



cmp.cus.ca

The Sampling Distribution

The same logic can apply to determining the sample means.

For a sample mean, the histogram will be $E(\bar{x}) = \mu$, $SD = \sigma/\sqrt{n}$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \qquad Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

The Sampling Distribution - Assumptions

- 1) **Independence:** All trials must be independent.
- 2) **Sample Size:** Typically (but not always), $n > 30$.
- 3) **Randomization:** Must be random.
- 4) **10% Condition:** the population should be atleast $10n$.
- 5) **Success/Failure Condition:** $np > 10$ and $nq > 10$



@ubccmp



@ubccmp



cmp.cus.ca

Practice Q

The amount of time teens spend on social media has a mean of 9 hours and standard deviation of 4 hours weekly. From a random sample of 64 teens, what is the probability that the mean amount of time spent on social media exceeds 10 hours per week?



@ubccmp



@ubccmp



cmp.cus.ca

Confidence Intervals

In the real world we don't actually know p , so SD is found using an estimate, \hat{p} .

Standard Error: Estimate of the standard deviation of the sampling distribution.

$$\text{Standard Error} = SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} \quad SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

Margin of Error = estimate \pm ME = $2SE(\hat{p})$ for 95% confidence

“We are 95% confident that the true proportion lies within the interval...”



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

Confidence Intervals

About 95% of p-hats will be within 2 SD of p, so the formula below has a 95% chance of capturing the true p.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Some Helpful Formulas:

Solving for Sample Size:

$$n = \frac{(z^*)^2 \hat{p}\hat{q}}{(ME)^2}$$

$$ME \approx \frac{1}{\sqrt{n}} \quad (\text{for 95\% confidence})$$


For E = $\pm 10\%$, need $n \approx 100$
For E = $\pm 5\%$, need $n \approx 400$
For E = $\pm 3\%$, need $n \approx 1,100$
For E = $\pm 2\%$, need $n \approx 2,500$

Common Critical Values:

CI	z^*
90%	1.645
95%	1.95
99%	2.576



 @ubccmp

 @ubccmp

 <http://cmp.cus.ca>

Practice Q

A fair coin is tossed n times. For the probability to be at least 68% that the proportion of heads is between 0.45 and 0.55, how many times should the coin be tossed?



@ubccmp



@ubccmp



cmp.cus.ca

Confidence Intervals Ctd.

MORE CONFIDENT = LARGER MARGIN OF ERROR = WIDER INTERVAL

- If you have a narrower interval without giving up confidence, then you have less variability.
- A given sample size will always yield the same margin of error regardless of varying population sizes.

Critical Value: Number of SE's to move away from the mean of the sampling distribution to correspond to a specific confidence value.

- If $|z\text{-stat}| > z^*$, reject H_0 .

Significance Level ($100\alpha\%$)	Alpha (α)	One-sided	Two-sided
5%	0.05	$z^* = 1.645$	$z^* = 1.96$
1%	0.01	$z^* = 2.33$	$z^* = 2.576$



Practice Q

A confidence interval estimate for a single proportion based on a random sample of 500 is 0.372 to 0.428. From this information, what level of confidence level was used in this estimate?



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

Hypothesis Testing

Null Hypothesis (H_0): Statement of no change from traditional value, claim being assessed, $H_0: p = p_0$

Alternative Hypothesis ($H_{1/A}$): Proposes what should be concluded if the null hypothesis is found unlikely.

- If p_0 is a reasonable estimate for p , it should behave like Z .

Test Statistic: $z =$

$$\frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

CI for p :

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Level Of significance / Alpha Level: Threshold for when p is small.

IF P-VALUE $< \alpha$ REJECT H_0 (If p_0 lies outside the $1 - \alpha$ confidence interval)

IF P-VALUE $> \alpha$ DO NOT REJECT H_0

Two Sided: P-value = $2 \times \Pr(z > |z\text{-stat}|)$ One Sided: $\Pr(z > |z\text{-stat}|)$



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

One Proportion Z Testing

- The p-value assumes that H_0 is true. The smaller the p-value, the greater the evidence against the null hypothesis.

Complete Steps:

- 1) Hypotheses
- 2) Test statistic
- 3) P-value
- 4) Conclusion:
 - a) There is sufficient evidence to reject the null hypothesis
 - b) There is insufficient evidence to reject the null hypothesis.



@ubccmp



@ubccmp



cmp.cus.ca

An Overview of Hypothesis Tests & CI's

① CI: $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$

note the use of \hat{p}

• If p falls in interval, do not reject.

z^* = critical value

② $z\text{-stat} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$

test statistic

note the use of p_0

• if $|z\text{-stat}| > z^*$, reject H_0
(often $z^* = 1.96$)

③ p-value:

$1 - \alpha$ = CI level
if two sided. $\alpha \cdot (\Pr(z > |z\text{-stat}|))$

= $[2 \times] [1 - \text{NORM.S.DIST}(z\text{-stat}, \text{TRUE})]$

Excel

↳ case dependent.

• if p-value $< \alpha$, reject H_0

Walk Through - Two Sided

A study found that 30% of people were willing to buy “ugly vegetables” at the grocery store that are normally thrown away if they were on sale.

In a random sample of 100 save-on shoppers, 28 respondents agreed to purchase such products. Test if sample differs from the proportion.



@ubccmp



@ubccmp



cmp.cus.ca

Walk Through - One Sided

Hinge wants to advertise that more than 50% of matches which result in a first date lead to marriage. A random sample of 240 recent matches indicated that 138 of these couples had ended up getting married. Use 5% significance level to test.



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

Practice Q

One of top universities in Canada had only a 60% graduation rate in 2022. From a random sample of 750 seniors in 2023, 480 graduated. Does this provide sufficient evidence that the school is doing better in 2023 than 2022 at the 0.05 alpha level?

What are the null and alternative hypotheses?

What is the standard deviation of \hat{p} ?

What is the value of the test statistic?

What is the critical value?

What is the 95% confidence interval to estimate the true proportion of students that graduate in 2023?



@ubccmp





@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca



 @ubccmp

 @ubccmp

 [http://](http://cmp.cus.ca) cmp.cus.ca

General Advice

- Take all the things you don't know / can't remember and put them in one, easy to read spot
 - More people struggle with the timing than the content.
- Practice, practice, practice
 - You'll see very familiar questions 🙄🙄
- Check your answers
 - Avoid silly mistakes, practice like you test!
- Read questions carefully
 - Identify what the question is asking for.
- Understanding things conceptually = speed.
- Always do sanity checks.
- I did not use a master excel so I don't vouch for it, however some people find it helpful.

Me in COMM 191 last year!



I promise you can do well on this midterm !!!



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

Questions?

Please Give Me Feedback!!

Did you find this review session helpful?

Yes 😊

Great, let me know!


No 😞


I'm sorry, let me know
how I can do better!

FILL IN THE FORM (please :)

If you still have questions or need help please let me know!



 @ubccmp

 @ubccmp

 <http://cmp.cus.ca>