Commerce Mentorship Program

COMM 291

Midterm Review Package

Prepared by Angelica Cabrera

1. INTRODUCTION TO DATA AND VARIABLES

Categorical vs. Quantitative Data

	Categorical	Quantitative
Possible values for variable	Limited number – distinct categories	Large number
Measurement units?	No	Yes

EXAMPLE. Which variables are quantitative and which are categorical?

Employee #	Age (years)	Annual Income (in 1,000s of dollars)	Performance Rating (1-5 scale)	Job Type
5543	48	50 – 100	4.5	Management
2431	34	20 – 49	3.9	Clerical
7281	31	0 – 19	3.4	Maintenance

_____ _____ _____ _____ _____

2. SURVEYS AND SAMPLING

Population: _____ individuals with a common characteristic that you want to generalize about

Sample: _____ of population

Parameter: fact or characteristic about _____

Statistic: fact or characteristic about _____

EXAMPLE. Mattel claims that less than 5% of all its Hot Wheels toys are defective. When testing 100 Hot Wheels toys from a production run of 7000 toys, 7% were found to be defective. What is the:

a) Population? _____ b) Statistic? _____

c) Parameter? _____ d) Sample? _____

Poor (Biased) Sampling

- **Convenience sampling**: Choosing respondents that are _____ to obtain
- **Voluntary response**: Respondents volunteer, so those with _____ opinions are more likely to respond

Sampling Designs

1. **Simple Random Sampling (SRS)**: Every individual has an equal chance of being selected
2. **Stratified Random Sampling**: Divide population into _____ subgroups and randomly select from each stratum
3. **Cluster Random Sampling**: Divide population into _____ subgroups that are representative of population and select a few clusters
4. **Systematic Sampling**: with a random starting point, select at regular intervals

***EXAMPLE*.** You are considering ways to randomly sample UBC varsity athletes to learn about types of sports drinks they would prefer. What kind of sampling designs are the following?

a) Select ten individuals from each sport team at UBC (Ex. hockey, basketball, rowing, etc.) _____
b) Randomly select 50 athletes using their student numbers _____
c) Randomly select a faculty and survey all of athletes in that faculty _____
d) Select every third name in alphabetized list of all varsity sports athletes at UBC, starting with a random name _____

3. DISPLAYING AND DESCRIBING CATEGORICAL DATA

- Best represented in a _____ graph

Contingency Tables

- Counts can be converted into:

_____ Percentages: $\frac{\text{Cell Value}}{\text{Row Total}}$

_____ Percentages: $\frac{\text{Cell Value}}{\text{Column Total}}$

EXAMPLE. A survey collected teenagers' preferences for soft drinks.

Soft Drink	***Male***	***Female***	***Total***
Pepsi	55	87	**142**
Sprite	99	150	**249**
Coke	196	113	**309**
Total	**350**	**350**	**700**

a) What percentage of teenagers preferred Pepsi? _____

b) What percentage of teenagers who preferred Coke were males? _____

c) Of teenagers who are females, what percentage preferred Sprite? _____

Simpson's Paradox: an association that holds for all of several groups can _____ direction when the data are combined to form a single group.

4. DISPLAYING AND DESCRIBING QUANTITATIVE DATA

- Best represented in a _____, and an alternative is a _____ plot

Measures of Centre

1. **Mean:** average of data
2. **Median:** middle of data
3. **Mode:** most frequent data value

Measures of Spread

- **Range** = Maximum value – minimum value
- **Standard deviation (SD)**: "typical" distance from the data value to the mean
- **Variance** = $(SD)^2$
- **Percentile:** value below which % of data values fall
- **IQR** = Q3 – Q1

Histogram Shapes

Symmetric **Skewed**

Left: *Right:*

Mean ___median Mean ___median Mean ___median

Best Measures of Centre and Spread

- For symmetric distributions, use _____ and _____
- For skewed distributions, use _____ and _____

Box-and-Whisker Plots

- How to draw a box-and-whisker plot

1. Plot points given 2. Draw IQR 3. Find inner fences • **Lower inner fence** = Q1 – 1.5 (IQR) • **Outer inner fence** = Q3 + 1.5 (IQR)	4. Draw whiskers (last values within the inner fences) 5. Draw outliers (values outside inner fences)

EXAMPLE 2. **Below is a five-number summary for hourly wages for managers at AEKI, a furniture store.**

Min	Q1	Median	Q3	Max
20.94	37.64	44.77	49.24	67.11

a) This distribution is skewed to the:

b) What is the IQR? _____

c) What is the lower inner fence? _____

d) What is the upper inner fence? _____

e) Where do the outliers lie? _____

f) There was an error and the lowest hourly wage for sales managers was $18.15 instead of $20.94. How would this affect:

The mean? _____ The range? _____

The median? _____ The IQR? _____

5. SCATTERPLOTS, CORRELATION, AND LINEAR REGRESSION

Correlation (*r*): how strong the linear clustering is around a line

- Only for _____ data with a _____ pattern
- $-1 \leq r \leq +1$
- NO units
- Correlation of X and Y = Correlation of Y and X
- Correlation ≠ causation
- Beware of:
    1. **Lurking variables**: third variable causing X and Y
    2. **Extrapolation**: extending results beyond your range of data

Linear Regression helps us find the best-fitting straight line through the scatterplot

- How to find a regression line
    1. Find $\bar{x}$, $\bar{y}$, S_x, S_y, r
    2. Calculate $b_1 = r\frac{S_y}{S_x}$
    3. Calculate $b_0 = \bar{y} - b_1\bar{x}$.
    4. Find regression line $\hat{y} = b_0 + b_1 x$
- **Regression effect:** variables fluctuate naturally and regress towards the mean

R-Squared (r^2): percentage of variation in the Y-values that can be explained by the model

- $0 \leq r^2 \leq 1$

Residual Plots

- Residual = $e_i = y_i - \hat{y}_i$, calculated for each data point
- Adequate fit if there is a _____ and _____ band around 0 (the x-axis)
- If is curved, not a _____ trend
- If is a _____ trend, you made an error

EXAMPLE. Data was collected on several employees: their stress levels (X, on a scale of 0 to 10), and productivity levels (Y, in parts made per hour).

$\bar{x}$ = 5.4	S_x = 3.3
$\bar{y}$ = 57.5	S_y = 11.1
b_1 = -3.19	S_e = 4.3

a) What is the estimated regression equation?

b) What is r^2? _____

c) For each additional unit on the stress scale, the productivity level, on average, decreases by _____ parts per hour. _____

d) i. What is the productivity of an individual with a stress level of 8?

_____

ii. Suppose this employee has an actual productivity level of 64 parts per hour. The SD of residuals is 4.3. Is this data point an outlier?

Outliers vs. Influential Points

- **Outlier:** "Far away" from the data
- **Influential observation:** if removed, changes _____

6. COMBINATIONS OF RANDOM VARIABLES

Random variable (RV): numerical outcome of a random phenomenon

- Mean of a RV: Long-run average outcome = μ
- SD of a RV: Long-run standard deviation of outcome = σ

Properties of Combinations of Random Variables

Sum of two INDEPENDENT RVs: X+Y
Mean (X+Y) = Mean (X+Y)
Var (X+Y) = Var(X)+Var(Y)
SD(X+Y) = $\sqrt{}$ Var(X)+Var(Y)

Difference of two INDEPENDENT RVs: X–Y
Mean (X-Y) = Mean (X-Y)
Var (X-Y) = Var(X)+Var(Y)
SD(X-Y) = $\sqrt{}$ Var(X)+Var(Y)

EXAMPLE. A juice box has a mean weight of 9.60 oz and a SD of 0.80 oz. Suppose a carton can fit 24 juice boxes. An empty carton has a mean weight of 30 oz and a SD of 2.20 oz. What is the mean and SD of a filled carton?

7. THE NORMAL DISTRIBUTION

- Represents all possible values of a random variable, so total area under the curve: _____%

The 68-95-99.7% Rule

For hypothetical distribution
For actual data
For Z (standardized)

Doing Normal Curve Calculations

- Important: Table Z gives probabilities to the _____ of the z-score
- How to Solve Normal Curve Calculations

To find a probability

1. Find Z = $\frac{X-\mu}{\sigma}$
2. Find area corresponding to Z-score using Table Z
3. Draw a picture
4. Find area

To find x

1. Identify desired area
2. Find (approximate) z-score corresponding to this area, using Table Z
3. Draw a picture
4. "Unstandardize" to find $X = \mu + Z\sigma$

EXAMPLE. The wingspan of penguins at the Vancouver Aquarium are normally distributed. The probability that a penguin has a wingspan of more than 20 cm is 0.5, while the probability of a wingspan of more than 30 cm is 0.1587. What are the mean and SD of the wingspan of penguins at the Aquarium?

EXAMPLE. A psychology exam found that females had a mean score of 120 and a SD of 28. For males, the mean score was 105 with a SD of 35. Scores are normally distributed.

a) What percentage of female students had scores greater than 162?

b) What score is exceeded by only 10% of female students?

c) What percentage of male students had a score between 70 and 140?

d) i. Suppose you select a single female student and a single male student at random. What are mean and SD of the DIFFERENCE between their scores?

ii. What is the probability that a chosen female has a higher score than a chosen male?

8. SAMPLING DISTRIBUTIONS FOR MEANS AND PROPORTIONS

- **Sampling distribution:** theoretical distribution of all the values taken by a statistic if a large number of samples of the same size were drawn from the same population.

Sampling Distribution for a Proportion

- Only for _____ data
- _____ possible outcomes

Sample proportion	**Mean**	**SD**	**Z-score**
$\hat{p} = \frac{X}{n}$	p	$SD(\hat{p}) = \sqrt{\frac{pq}{n}}$	$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$

ASSUMPTIONS – Sampling distribution (if values are a random sample) is _____ NORMAL if n is "large enough":

a) 10% Condition: n should be no more than 10% of the population
b) Success/Failure Condition: np > 10 and nq >10

- "Mean of $\hat{p}$ is p" = long-run average value of $\hat{p}$ is p

Sampling Distribution for a Mean

- Only for _____ data

Sample mean	**Population mean**	**SD**	**Z-score**
$\bar{x}$	μ	$SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$	$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

ASSUMPTIONS – Sampling distribution (if values are a random sample) is:

- _____ Normal if original population is normally distributed,
- _____ Normal even if the original population is not normally distributed, as long as n is large enough. This result is called

the _____

Large enough if:

a) n = 30 (typically)
b) 10% Condition

- "Mean of $\hat{p}$ is p" = long-run average value of $\hat{p}$ is p
- The mean of several observations has a _____ SD than a single observation

EXAMPLE. For Sky Cable, an Internet provider, past data indicate there is a probability of 0.70 that service troubles can be repaired within the same day of being reported over the phone.

a) If the company receives 100 trouble calls on a particular day, what is the chance that more than 80% of will receive same-day repairs?

b) It is known that the repair time has a mean of 480 minutes and a SD of 250 minutes. A random sample of 400 trouble calls was taken. Compute the probability that the mean of the 400 repair times is less than 500 minutes.

9. CONFIDENCE INTERVALS

- Estimates like $\hat{p}$ and $\bar{x}$ have _____ and confidence intervals quantify this by expressing a margin of error
- How to Compute Confidence Intervals

For proportions	For means
1. Find $\hat{p}$ 2. Choose z* **CI** **z*** 90% 1.645 95% 1.96 99% 2.576 3. Find CI = $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$	**1.** Find $\bar{x}$ **2.** Choose t^*_{n-1} using Table T **3.** Find CI = $\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$

• How to Compute *n* for a required ME $n = \frac{(z^*)^2 \hat{p}\hat{q}}{ME^2}$	

EXAMPLE. TV Town, a television retailer, reported that 12% of its LCD TVs required warranty service. TV Town wants to know if this percentage is the same for its plasma TVs.

a) i. TV Town wants to estimate the true percentage of warranties for plasma TVs within ±4% with 99% confidence. How many LCD TVs should they sample?

ii. Suppose TV Town wants this margin of error to be ±2%instead. What would they have to do?

_____ Sample size _____ Level of confidence

b) i. Ignore your answer above and assume TV Town selects a sample of 450 plasma TVs. What is the 95% CI?

ii.State the conclusion of your CI in one sentence.

10. HYPOTHESIS TESTS FOR PROPORTIONS AND MEANS

- Hypothesis tests show which of two hypothesis about a parameter is better supported by the data.

H_0 → null hypothesis ("status quo")
H_a → alternative (research) hypothesis

We assume that H_0 is true, then we calculate the P-value, or probability of our observed data occurring in that hypothetical world.

Hypothesis Test for Proportions

- How to do a Hypothesis Test

a) Two-sided test	b) One-sided test
1. State hypotheses H_0: $p = p_0$ H_a. $p \neq p_0$	**1.** State hypotheses H_0: $p = p_0$ H_a: $p > p_0$ or H_a: $p < p_0$
2. Find $\hat{p} = \frac{X}{n}$	
3. Calculate test statistic (z-stat) $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$	

4. Calculate P-value = 2 x Pr (z > |z-stat|)

4. Calculate P-value = Pr (z > |z-stat|)

5. Compare to alpha level

If P-value < α, Reject H_0

If P-value > α, Don't reject H_0

c) Using Confidence Intervals

1. State hypotheses

2. Find $\hat{p}$

3. Choose z* according to your desired level of confidence

Significance Level	Confidence Level	One-sided	Two-sided
5%	95%	z* = 1.645	z* = 1.96
1%	99%	z* = 2.33	z* = 2.576
0.1%	99.9%	z* = 3.09	z* = 3.29

4. Find CI = $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$

5. Compare to confidence level

- If p_0 lies outside the CI, Reject H_0
- If p_0 lies inside the CI, Don't reject H_0

EXAMPLE. A national newspaper reported that 28% of Canadians had difficulty in making mortgage payments. A newspaper in Vancouver, BC randomly sampled 400 residents and found that 136 of them reported this difficulty. Does this mean that this problem is more severe in Vancouver?

a) What are the null and alternative hypotheses?

b) What is the test statistic?

c) What is the P-value?

d) At α = 0.5, state your conclusion in one sentence.

e) Based on a 95% CI, what would your conclusion be?

Hypothesis Test for Means

- Introduce Student's t-distribution, with n-1 degrees of freedom

  - As n approaches infinity, t-distribution becomes more _____

- How to do a Hypothesis Test

a) Two-sided test	b) One-sided test
1. State hypotheses H_0: $\mu = \mu_0$ H_a. $\mu \neq \mu_0$	**1.** State hypotheses H_0: $\mu = \mu_0$ H_a. $\mu > \mu_0$ or H_a. $\mu < \mu_0$
2. Find $\bar{x}$	
3. Calculate test statistic (t-stat) $t = \dfrac{\bar{x} - \mu_o}{s/\sqrt{n}}$	
4. Calculate P-value = 2 x Pr (t > |t-stat|)	**4.** Calculate P-value = Pr (t > |t-stat|)
5. Compare to alpha level – If P-value < α, Reject H_0 – If P-value > α, Don't reject H_0	

c) Using Confidence Intervals
1. State hypotheses
2. Find $\bar{x}$
3. Choose t^*_{n-1} according to your desired level of confidence and df, using Table T
4. Find CI = $\bar{x} \pm t^*_{n-1}\dfrac{s}{\sqrt{n}}$
3. Compare to confidence level – If μ_0 lies outside the CI, Reject H_0 – If μ_0 lies inside the CI, Don't reject H_0

EXAMPLE. 15 years ago, Canada Post found that the average time that their employees worked for the company was 7.5 years. Recently, a sample of 100 Canada Post employees found that the mean length of time they have worked for the company was 7.0 years with SD of 2.0 years. Management wants to know if this value had changed from 15 years ago.

a) What are the null and alternative hypotheses?

b) What is the test statistic?

c) What is the range where the P-value is located?

d) At the 1% significance level, state your conclusion in one sentence.

e) Based on a 95% CI, what would your conclusion be?

Errors in Hypothesis Testing

		Decision	
		Accept H_0	**Reject H_o**
True State	**H_o True**	Correct	Type I Error
	H_o False	Type II Error	Correct

EXAMPLE.

Null Hypothesis (H_0)	**Type I Error**	**Type II Error**
Defendant is innocent until proven guilty		