STATISTICS

✓ Final 6 - Randomly selected from homework
4 from portfolio
2 Omni

- Text book - Basic business statistics.

BASIC PROBABILITY

→ probability - the extent to wlc something is likely to occur
- the most likely Cause of something.

→ sample space - the set of all possible outcomes.

→ event - subset of sample space, the set of all outcomes that produce a specific result.

Simple (a single outcome) | joint (compound)

→ A complement ('A') - subset of outcomes that are not part of event.

⇒ Mutually exclusive events
- a set of events that can't occur at the same time eg 'needs an hair

⇒ collectively exhaustive events
- if one of the events must occur

eg. in a die roll
event 1 - getting Odd num.
event 2. getting even num.
- then may be mutually exclusive or not.

Types of probability

1. Priori probability/classical
is based on having a prior knowledge of the outcomes that can occur.
eg. tossing a coin.

2. Empirical probability.
- based on the observed data
eg. football odds
- mainly identified by survey.

3. Subjective Probability
- a probability that differs from person to person.
eg. probability of you marrying.

4. Axiomatic probability

⇒ Simple probability
- Probability of occurence of single event
$P = \frac{x}{N}$ ← no of outcomes of event / ← all possible outcomes.

⇒ Joint probability
- The probability of occurence of one two or more events.

⇒ Marginal probability
- is an event consisting a set of joint probabilities.
eg.

	A	K
B		

$P(A) = P(A \text{ and } h) + P(A \text{ and } K)$

⇒ General addition rule
$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$

Conditional probability ⇒ A given B

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ ← joint prob. / ← marginal prob

- The probability of occurence of event A given that event B has already occured.

(*) $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- If conditional probability of event B is known.

(*) $P(A|B) = P(A)$ or $P(A \cap B) = P(A).P(B)$.
if they are independent of one another.

(*) $P(A|B) = P(A)$ ⊙
if they are mutually exclusive

⇒ $P(A \cap B) = P(A|B).P(B)$

→ Marginal probability using multiplication rule.

$$P(A) = (A \cap B_1) + (A \cap B_2) + (A_n \cap B_n)$$

→ $P(A) = (P(A|B_1) \cdot P(B_1) + P(A|B_2) \cdot P(B_2) \cdots)$

Where $B_1, B_2 \cdots B_n$ are mutually exclusive collectively exhaustive

→ Bayes' theorem

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) \cdots P(A|B_k)P(B_k)}$$

where B_i is i^{th} event out of K mutually exclusive and collectively exhaustive events.

Counting rules

rule 1

- no of possible outcomes if anyone of k different mutually exclusive and collectively exhaustive ~~eleme~~ events. in n trials is

$$k^n$$

eg. variation of license plates having 3 nos and 3 letters is

- $10^3 \cdot 26^3 = \cdots$

rule 2.

- if there are k_1 events in 1^{st} trial k_2 events in 2^{nd} trial and k^n in n^{th} trial

then. no of possible outcomes.

$$(k_1)(k_2) \cdots (k_n)$$

rule 3

- The no of ways that all n items can be ordered.

$$n!$$

rule 4 - permutation.

- no of ways of arranging x items selected from n items.

$${}_nP_x = \frac{n!}{(n-x)!}$$

rule 5 - combination.

- no of ways of selecting x items from n items irrespective of order.

$${}_nC_x = \frac{n!}{x!(n-x)!}$$

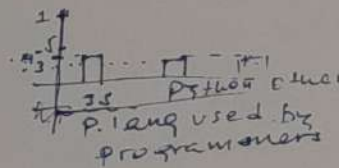
CH-5

Discrete probability distribution

- is mutually exclusive list of all possible numerical outcomes along with the probability of each outcome.

eg

No of meals per day	probability
0	1%
1	15%
2	29%
3	40%
4	15%



* Expected value of discrete variables
  - mean based on probability

$$\mu = E(X) = \sum_{i=1}^{n} x_i P(X = x_i)$$

* Variance

$$\sigma^2 = \sum_{i=1}^{N} [x_i - E(x)]^2 P(X = x_i)$$

* Standard deviation

$$\sigma = \sqrt{\sigma^2}$$

Binomial distribution

- π is the probability of event of interest.

$$P(X = x | n, \pi) = \frac{n!}{x!(n-x)!} \cdot \pi^x (1-\pi)^{n-x}$$

($\frac{n!}{x!(n-x)!}$: no of combinations)

- n - no of observation
- x - no of event of interest in the sample
- π - probability of an event of interest.

* mean of binomial distribution.

$$\mu = E(x) = n\pi$$

* standard deviation of b dist

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var(x)} = \sqrt{n\pi(1-\pi)}$$

Poisson distribution

probability of an event in area or interval of time given mean (expected no of event per unit)

$$P(X = x | \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

$e = 2.7$

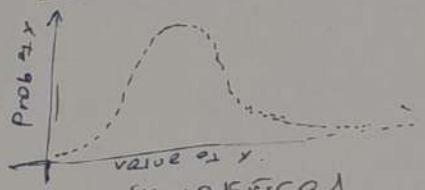
x = no of events.

CH-6

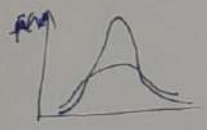
The normal distribution and other continuous distribution

- it doesn't have finite no variables or events
- Area under the curve is probability.
- The whole A. under the curve = 1
- The single value can't have a prob because do not have area

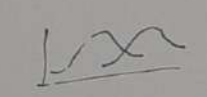
Normal distribution



- Symetrical
- Its mean and median are equa
- Its range is infinite $(-\infty, \infty)$



same mean dif σ



same σ dif μ

* Normal probability density function.

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

x - any value of the variab

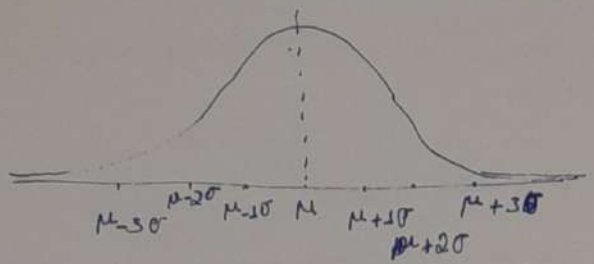
$$z = \frac{x - \mu}{\sigma}$$

Converts normally distributed variable into standardized norma variables

- μ of $Z = 0$
- σ of $Z = 1$

but normally distributed variables have their own μ and σ (x)

If the ave load time(μ) for a website is 7 sec and standard deviation of 2



x scale	1	3	5	7	9	11	13
z scale	-3	-2	-1	0	+1	+2	+3

- so load time of 1 sec is -3 standardized unit (3 standard deviations) below the mean value.

- to find the probability in a normal distribution you can find the z index and find prob from the commulative distribution table.
and if x is asked given probability you can do the same.

⇒ Uniform distribution.

- values are evenly distributed in the range b/h the smallest value A and the larges value B.

uniform prob density fun. $f(x) = \frac{1}{(b-a)}$ if $a \le x \le b$

mean $\frac{a+b}{2}$

$\sigma^2 = \frac{(b-a)^2}{12}$

$\sigma = \sqrt{\frac{(b-a)^2}{12}}$

CH-7

Sampling the distribution of the mean

sampling distribution - is the distribution of the result if you actually selected all possible samples.

sampling distribution of mean
- is distribution of results if you actually selected means of all possible samples
- the mean of all possible sample means is equal to the popn mean.
- it is always a normal distribution (becomes less normal if n of sample is less)

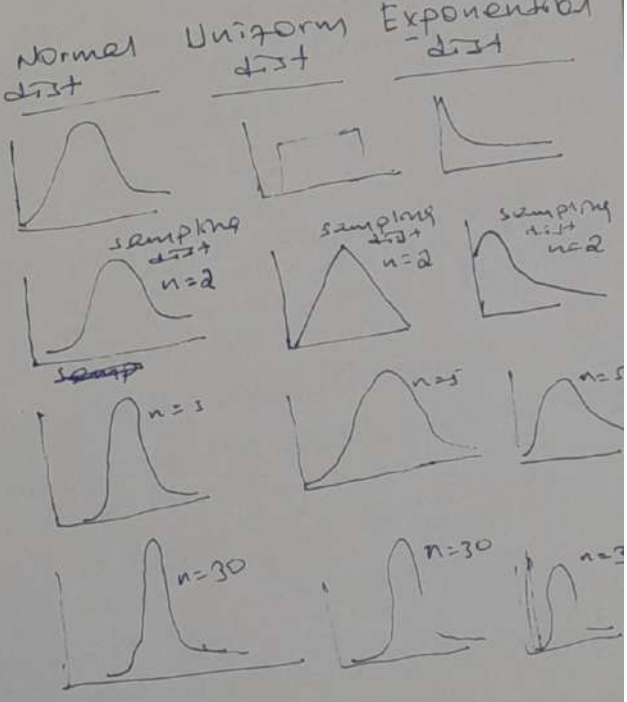
standard error of the mean (standard deviation of sample)

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ n- no of elmt of the sample.

Z-score of the sampling dist of the mean

$z = \frac{(\bar{x} - \mu_{\bar{x}})}{\sigma_{\bar{x}}} = \frac{(\bar{x} - \mu_x)\sqrt{n}}{\sigma}$

$\bar{X} = \mu + Z\frac{\sigma}{\sqrt{n}}$



...variance ... are called parameter of popⁿ.

sampling proportion of the Proportion

sampling proportion (p) = $\frac{\text{No of elmt of variable of interest}}{\text{sample size}}$

$$p = \frac{X}{n}$$

- used to estimat the popⁿ propo rtion π
- use when dealing with a categorical variables.

standard error of the proportion

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

$$Z = \frac{p - \pi}{\sigma_p}$$

CH-8

Confidence Interval Estimation.

- is a range of no's that a given parameter of a popⁿ is true at a given probability. (most of the times 95% probability.
- The variation of sampling statistics from sample to sample is called sampling error.

conf interval for mean (σ known)

$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ — $Z_{\alpha/2}$ - critical value; $\alpha = 1$ - given prob; $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ - sampling error.

confidence interval of mean (σ unknown)

$\bar{x} \pm t_{v\alpha/2}\left(\frac{\sigma_x}{\sqrt{n}}\right)$ — $v = n-1$ then find $t_{v\alpha/2}$ using v & $\alpha/2$ from t-score table.

confidence interval for proportion.

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

determining sample size.

sampling error $e = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

so $n = \frac{Z_{\alpha/2}^2 \sigma^2}{e^2}$

for proportion replace σ with $\sqrt{\pi(1-\pi)}$

$$n = \frac{Z_{\alpha/2}^2(\pi(1-\pi))}{e^2}$$

CH-9.

Fundamentals of hypothesis Testing - One sample tests.

null hypothesis (H_0) - states the status quo claim.

- refers to a popⁿ parameter such as μ σ like $\mu = 4.5$ min
- always includes equal sign.

collectively exhaustive & mutually ...

alternative hypothesis (H_1)

- states a claim that is contrary to the null hypothesis.

risks

⇒ TYPE 1 Error - if you reject the null.... when H_0 is true and should not be rejected.
- is a false alarm.
- its probability is α

⇒ TYPE 2 Error - if you don't reject the H_0 when H_0 is false and should be rejected.
- is missed opportunity
- its prob is β risk

- α - level of significance
  confidence coefficient ($1-\alpha$)
- β - risk.
  power of statistical test ($1-\beta$)

Z - test for the mean (σ known)

step 1 - find critical values $\pm Z_{\alpha/2}$

step 2. - evaluat Z_{STAT}.

$$Z_{STAT} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

step. if Z_{STAT} is within the range of critical value then accept the null hypothesis. and if it is not reject the H_0.

t - test of the mean (σ known)

same with the above procedure but replace the formula by

$$t_{STAT} = \frac{\bar{x} - \mu}{\frac{\sigma_x}{\sqrt{n}}}$$

σ_x sometimes called S. use $v = n-1$. and used t-dist instead of Z-dist

One tail test

when you face < > ≥ ≤ probs

- if the rejection area is in the upper tail check $t_{STAT} > t_{v\alpha/2}$ if false don't reject (use Z if σ is known)
- if the rejection area is in the lower tail check $t_{STAT} < t_{v\alpha/2}$ if false don't reject.

Z- Test ~~for~~ hypothesis for Proportion

The use one of the above steps then use..

$$Z_{STAT} = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$ (*)

CH-10

TWO SAMPLE TESTS

(1) pooled variance t test for diff bw means

- find critical values using $\alpha/2$ and $v = n_1 + n_2 - 2$.
- identify H_0, H_1 and rejection region.
- evaluate

$$t_{STAT} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

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

(pooled variance)

- Reject if test stat on the rejection area.

Population consists of all the individuals which you want to reach conclusion.

Variable → a characteristic of an item or individual

Data are d/t values associated with variable

Defining and collecting data

Classifying variables by type

Numerical variable → whose data represent a counted or measured quantity.

Categorical variable → whose data represent categories (Eg:- Gender)

sample - is a portion of population selected for analysis.

Numerical variable

Discrete
↳ Have data that arise from a counting process (Represent a number of sth.)

continuous
↳ Have data that arise from a measuring process (Eg: The time spent waiting on a checkout line. It represents timing measurements)

Measurment scales

Numerical variables
- Interval scale ⇒ expresses a difference b/n measurments that donot include a true zero point.
- Ratio → an ordered scale that includes a scale true zero point.

⇒ For both interval and ratio scales, what the d/ce of 1unit represents remains the same among pairs of values.

Categorical variable
- nominal scale → category values express no order of ranking (weakest form of measurment because you cannot specify any ranking across the various categories)
- ordinal scale → an ordering or ranking of category values is implied. (good, better, best)

A parameter is a measure that describes a characteristic of a pop□.

A statistic: is a measure that describes a characteristic of a sample.

Parameter: summarizes the value of a pop for a specific variable. Collecting data

Statistic - summarizes a specific variable for sample data.

A sample contains only a portion of interest

Types of sampling methods

Frame - is a complete or partial listing of items

⇒ Non-probability sample: you select items (individuals) without knowing their probabilities of selection.

- Convenience sample: you select items that are easy, inexpensive or convenient to sample.
- Judgment sample: you collect the opinions of preselected experts in the subject matter.

⇒ Probability sample: you select items based on known probabilities.

⇒ Simple random sample → every item from a frame has the same chance of selection as every other item and every sample of a fixed size has the same chance of selection as every other sample of that size.

[You use "n" to represent the sample size.
"N" to represent the Frame size.]

⇒ Sampling with replacement → After you select an item you return it to the frame, where it has the prob of being selected again.

⇒ Sampling without replacement → Once you selected an item you cannot select it again.

[On the first selection $1/N$
On the 2nd " $1/N-1$]

n=40 N=800 ⇒ you partition the frame of 800 into 40 each of which contain 20 employees.

$k = \frac{800}{40} = 20$

032, 052, 060, 080, 108, 128.

⇒ Systematic sample: you partition the N-items in the frame into n-groups of k-items, where $k = \frac{N}{n}$ (you round k into the nearest integer)

- Assign number to every pop sample
- Select a random number
- Select samples at regular interval

⇒ Stratified sample

↳ You first subdivide the N-items in the frame into separate sub-populations or strata

2⇒ More efficient than either single random & systematic sampling

Stratified: members of strata have similar characteristics

⇒ Cluster sample

↳ You divide N-items in the frame into clusters that contain several items.

Cluster: members of cluster have d/t characteristics

Data cleaning ~~is a process to~~

⇒ Even if you follow proper procedure to collect data, it may contain incorrect (inconsistent data) that could affect statistical results. Data cleaning corrects such defects & ensure your data contains suitable quality for your needs.

⇒ Seeks to correct the fit irregularities (being incorrect by simple scanning technique)

- Invalid variable values, including
  - ⇒ Non-numerical data for numerical variable
  - ⇒ Invalid categorical values of a categorical variable
  - ⇒ Numeric values outside a defined range
- coding errors (Poor recording of data)
  - ⇒ Inconsistent categorical values
  - ⇒ Inconsistent case for categorical values
  - ⇒ Extraneous characters.
- Data integration error.
  - ⇒ Redundant columns
  - ⇒ Duplicated rows
  - ⇒ Differing column lengths
  - ⇒ Different units of measure or scale for numerical variables

Missing values → values that were not collected. for a variable.

U-2

Organizing and visualizing variables

Organizing categorical variables

Summary table

↳ Helps you see the d/d among the category by displaying the frequency, amount or percentage of items in a set of categories by displaying the frequency amount or percentage of

Summary table

↳ Helps you see

Contingency table: cross-tabulates, or tallies jointly, the data of two or more categorical variables allowing you to study patterns that may exist b/n the variables.

Frequency distribution → tallies the values of a numerical variable into a set of numerically ordered classes.

Class interval width

$$\text{Interval width} = \frac{\text{highest value} - \text{lowest value}}{\text{number of classes.}}$$

Computing the proportion or Relative frequency

$$\text{Proportion} = \text{relative frequency} = \frac{\text{number of values in each category}}{\text{total number of values}}$$

The lower the CV, the less variability of random variable relative to it's mean. and vice-versa

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

Sample standard deviation $s = \sqrt{s^2}$

The coefficient of variation (CV)

↳ measures the scatter in the data relative to the mean. (expressed always in percentage)

$$CV = \left(\frac{s}{\bar{x}}\right) \times 100\%$$

Z-scores

$$z = \frac{x - \bar{x}}{s}$$

⇒ A z-score of zero indicates that the value is the same as the mean.

⇒ If it's positive or negative, indicates whether the value is above or below the mean and by how many standard deviations.

⇒ Helps identify outliers (The values that seem excessively d/t)

⇒ z-score greater than +3.0 / less than −3.0 } indicates outlier values.

Shape: skewness

Skewness - measures the extent to which the data values are not symmetrical around the mean

- Mean < median, negative (left skewed dist.)
- mean = median, symmetrical dist. (zero skewness)
- Mean > median, postive (right skewed dist.)

Shape: kurtosis

Kurtosis: measures the peakedness of the curve of the dist., how sharply the curve rises approaching the center of distr.

09

Bell-shaped normal distribution ⇒ zero kurtosis, of sense

⇒ Leptokurtic: A dist. that has sharper rising center peak (higher concentration of the values near the mean of the distribution)
⇒ Postive kurtosis. (fatter tails (ends) ; many more values in tails)

⇒ Platykurtic: A slower rising (flatter) center peak (lower concentration compared to a normal distribution)
⇒ negative kurtosis.

⇒ Quartiles (Q_1 = 25% Q_3 = 75% Q_2 = 50% = median)

(*) $Q_1 = \frac{n+1}{4}$ ranked value

(*) $Q_3 = \frac{3(n+1)}{4}$ ranked value

(If the ranked value ends in .25 rounded to the next whole number; .75 rounded to the previous whole number)

⇒ Percentiles (split into 100 equal parts)

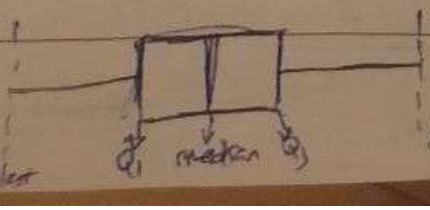
⇒ Interquartile = $Q_3 - Q_1$ (also called ranges (mispread))

⇒ Five number summary

X_{small} Q_1 median Q_2 $X_{largest}$

	Left skewed	Right skewed
$X_{largest}$ to median / median to $X_{largest}$	$X_{smallest}$ to median > median to $X_{largest}$	$X_{smallest}$ to median < median to $X_{largest}$
$X_{smallest}$ to Q_1, Q_3 to $X_{largest}$	$X_{smallest}$ to Q_1 > Q_3 to $X_{largest}$	$X_{smallest}$ to Q_1 < Q_3 to $X_{largest}$
Q_1, median to Q_3	Q_1 to median > median to Q_3	Q_1 to median < median to Q_3

Box plot



Numerical Descriptive measures of pop^n

Pop^n mean $\Rightarrow$ $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

pop^n variance & S.D ($\Rightarrow$ measure variation in the pop^n)

$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)}{N}$ $\quad \sigma = \sqrt{\sigma^2}$

- The Empirical rule

- Approximately 68% of the values are within ± 1 S.D of the mean
- 95% $\rightarrow \pm 2$ s.d from the mean.
- 99.7% $\rightarrow \pm 3$ s.d from the mean.

~~1 out of 20.~~

$\Rightarrow$ Implies about 1 out of 20 values will be beyond 2 standard deviations from the mean in either dir^n.

$\mu \pm \sigma \rightarrow 68\%$

$\rightarrow$ values not found in the interval $\rightarrow \mu \pm 2\sigma \rightarrow$ potential outliers. $\Rightarrow 95\%$

$\Rightarrow$ 3 in 1000 will be beyond 3 s.d from the mean.

$\mu \pm 3\sigma \rightarrow$ values not found in the interval are almost always outliers. $\Rightarrow 99.7\%$

- Chebyshev's theorm $\rightarrow$ For heavily skewed sets of data & data sets that do not appear normally distributed.

$\left(1 - \frac{1}{k^2}\right) \times 100\%$

$\Rightarrow$ states that for any data set, regardless of shape, the percentage of values that are found within distances of k S.D from the mean must be at least

chebyshev's theorem
$\Rightarrow$ Generally applies to any set of data

Empirical
$\Rightarrow$ If data approximately bell shaped
$\Rightarrow$ It'll more accurately describe the greater concentration of data close to the mean

09

$\mu \pm \sigma$	Chebby at least 0%	Empi Approx. 68%
$\mu \pm 2\sigma$	at least 75%	Approx. 95%
$\mu \pm 3\sigma$	At least 88.89%	Approx. 99.7%

→ To understand how data are distributed around the mean when you have sample data

The covariance & the coefficient of correlation:

Sample Covariance (measures the strength of the linear relationship b/n two numerical variables) → X and Y.

$$\text{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

A strong correlation can be produced by chance; by the effect of lurking variable

Coefficient of correlation

↳ measures the relative strength of linear relationship. (strong or weak)

↳ Correlation ~~doesn't~~ cannot prove that there is a causation effect (The change in one variable causes the change of the other variable)

Perfect negative (-1) (when X increases Y decreases)

Perfect positive (+1) (X increases the Y increase)

No correlation (X increases Y no change)

⇒ You can say causation implies correlation, but correlation does not imply causation.

$$r = \frac{\text{Cov}(X,Y)}{S_X S_Y}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n-1}} \qquad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})}{n-1}}$$

Unit-4

Basic probability

⇒ Mutually exclusive :- can not occur at the same time.

⇒ Collectively exhaustive → one of the events must occur.

The two event cover all (the) possible outcomes

0.40