CENG - 414 HW1

1) a) Nominal - level. It's not quantitative value.

Chp 2 b) ordinal - level. It's non-mathematical and scalable value.

c) Ratio - level. Both differences and ratios are meaningful.

d) If it's letter grading, it's ordinal-level since value is not numeric. If it's numerical grading, it's interval - level.

e) Nominal-level. Values only distinguish.

f) Ratio - level. Value is numeric and ratios and differences are meaningful.

g) Ratio-level. Both differences and ratios are meaningful.

h) If temperature is in celsius or Fahrenheit, it's interval-levels since absolute value of zero is arbitrary. If it's in Kelvin, it's ratio - level.

Chp 2

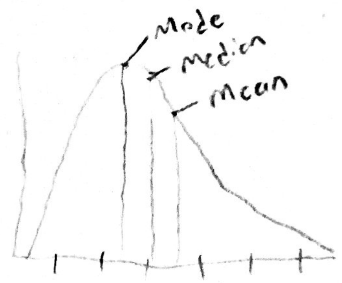
2) Since it's easy to use and it's accurate representation of the larger population. Also there is no need to divide the population into sub-populations.

3) Indep. variable : Condition that you change in an experiment.
Dep. " : " " " measure " " "

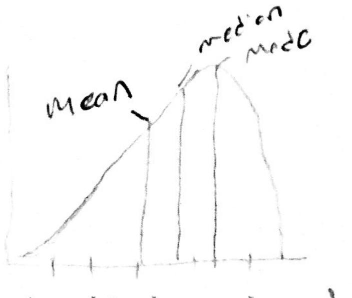
a) Indep. variable ⟹ time of workout
Dep. " ⟹ risk of catching a cold.

b) Indep. variable ⟹ meditation
Dep. " ⟹ type of decision

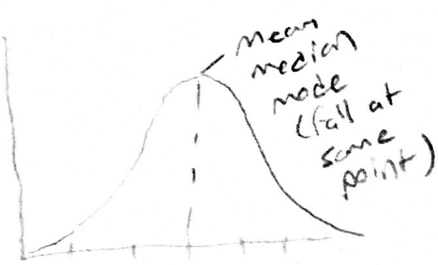
4)



Positively skewed
(right ")



Negatively skewed
(Left ")



Symmetric

5) Mean ⇒ Average of set of data.
Median ⇒ Middle of a set of data.
mode ⇒ Most common value of a set of data.

a) Mode b) Median c) Mean d) Mode e) median

6) No mode, zero mode or mode = 0.

7) Mean $= \frac{\sum x}{n} = \frac{59+52+28+26+19+19+18+17+17+17}{10} = 27.2$

Median = 19

Mode = 17

8) Mean $= \frac{9+10+14+7+8+3}{6} = 8.5$

Sample Variance $= \frac{\sum (x-\bar{x})^2}{5} = \frac{\left[(9-8.5)^2+(10-8.5)^2+(14-8.5)^2+(7-8.5)^2+(8-8.5)^2+(3-8.5)^2\right]}{5}$

Sample Variance $= 13.1 = s^2$

Std. deviation $= s = \sqrt{13.1} \approx 3.6$

9) Range = highest value − lowest value = 316 − 10 = 306

Mean $= \bar{x} = \frac{\sum x}{n} = \frac{33+10+62+132+123+316+123+133+18+150+26+138}{12} \approx 105.3$

Population Variance $= \sigma^2 = \frac{\sum (x-\bar{x})^2}{n}$

$$\approx \frac{\left[\begin{array}{l}(33-105.3)^2+(10-105.3)^2+(62-105.3)^2+(132-105.3)^2+\\(123-105.3)^2+(316-105.3)^2+(123-105.3)^2+(133-105.3)^2+\\(18-105.3)^2+(150-105.3)^2+(26-105.3)^2+(138-105.3)^2\end{array}\right]}{12}$$

≈ 6638.5

Population Std. dev. $= \sigma \approx \sqrt{6638.5} \approx 81.4$

10) Events are mutually exclusive if the occurence of one event excludes the occurence of other(s). Mutually exclusive events cannot happen at the same time.

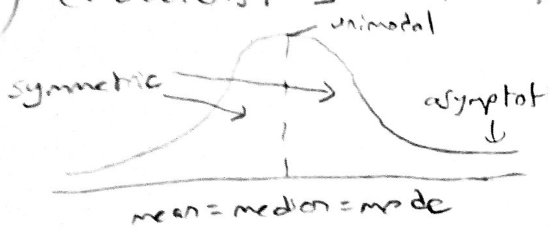
For mut. exc. events A and B: $P(A \wedge B) = 0$

Events are indep. if the occurence of one event doesn't influence the occurence of other(s).

For indep. events A and B: $P(A \cap B) = P(A)P(B)$

Mutually exclusive events cannot be independent, and indep. events cannot be mutually exclusive.

11) Characteristics of normal distribution:



1) It's unimodal, it has only one mode.
2) Symmetric about the mean.
3) It's asymptotic, it never touches the x-axis.
4) Mean, median and mode are all equal.

Position of curve depends on mean μ.
Shape of curve depends on std. dev. σ. (larger std. deviation means more dispersed distribution.)

12) Standart normal distribution is a special case of the normal distribution. It occurs when a normal random variable has $\mu = 0$ and $\sigma = 1$.

Total area under the normal dis. curve = 1.

13) T-test: Refers to a type of parametric test that is applied to identify how the two means of two sets of data differ from one another when variance is not given.
It's used in small sample size.

$$T\text{-test} = \frac{x - \mu}{s/\sqrt{n}}$$

Z-test : Implies a hypothesis test which specifies if the means of two datasets are different from each other when variance is given.

$$z\text{-test} = \frac{x-\mu}{\sigma/\sqrt{n}}$$ It's used in large sample size.

Chi-square test : Procedure for testing if two categorical variables are related in some population.

$$\chi^2 = \Sigma\left[\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right]$$

$$e_{ij} = \frac{O_i \cdot O_j}{N}$$

$e_{ij} \rightarrow$ expected freq.

$O_i \rightarrow$ marginal column freq.

$O_j \rightarrow$ marginal row freq.

$N \rightarrow$ total sample size

14) F-test is used for comparing two variances or standard deviations.

$F = S_1^2/S_2^2$

15) If two variables (related with correlation) is positive, both variables tend to move in same direction: if one variable increase, the other tends to increase. Same for decreasing case.

If two variables is negative, both variables tend to move in the opposite direction: if one variable increases, other tends to decrease.

16) Total variation $= \Sigma[(y-\bar{y})^2]$, $y \rightarrow$ y-value $\bar{y} \rightarrow$ mean of y

Explained variation $= \Sigma[(\hat{y}-\bar{y})^2]$, $\hat{y} \rightarrow$ predicted y-value

Unexplained variation $= \Sigma[(y-\hat{y})^2]$

Total variation = Explained variation + Unexplained variation

17) Correlation coefficient, denoted by r, measures how strong a relationship is between two variables.
If $r=0$, it means there is no relationship between two variables.

18) Only A always changes, since median and mode are only states the value but mean is equation of all values.

19) a) F - In equation of std. dev.:

$$\sqrt{\frac{\sum(x-\mu)^2}{n}} \quad \begin{matrix} \rightarrow \text{numerator} > 0 \\ \rightarrow \text{denominator} > 0 \end{matrix}$$

So, std. dev. must be positive.

b) F - Very high or very low value (outlier) would increase standart dev. as it would be very different from the mean. Hence outliers will effect std. dev. (not robust (strong)).

c) T

d) T - It's right-skewed.

20) $MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$, $n=5$, $x_i \rightarrow$ measurement, $x \rightarrow$ true value

For this question:

$$MAE = \frac{1}{5}\cdot\left[|45-65|+|76-70|+|78-75|+|87-80|+|79-85|\right]$$

$$= \frac{20+6+3+7+6}{5} = \frac{42}{5} = \boxed{8.4}$$