UQF2101I Cheatsheet

for test 1, by ning

Learning Objectives

- Approach a problem quantitatively
- Critically think about problem w/ QR tools
- To solve (or make a good effort) the problem
- Criteria air pollutants and their adverse effects
- Know how air quality is quantified and reported
- Recognise key indicators of beach water quality
- Know how beach water quality is quantified
- Qualitatively discuss observations/questions
- Able to formulate hypothesis
- Able to deduce predictions for hypotheses
- Describe the different methods used to collect data for testing our predictions
- Explain the advantages/disadvantages of the various methods of data collection
- Identify the various factors that can affect the validity of outcomes in our study
- Relate probability to quantitative reasoning
- Able to perform probability calculations
- Conditional probability & independence
- Preliminary analysis of data
- Descriptive statistics
- Graphical methods

Quantitative Reasoning is *the way in which we can use numbers to provide evidence for our arguments.* It has 10 elements,

1. Scientific method
2. Hypothesis
3. Sampling
4. Data
5. Operationalize
6. Variable
7. Modelling
8. Analysis
9. Deduction
10. Critical Thinking

This module examines 'pressing environmental issues that *continue to have a significant impact on human health.* QR provides. . .

- more clarity, better structured, objective, and useful analysis of issues,
- so that decision-making and arguments can be guided and supported by quantitative proof and rigorous analysis
- proof with numbers that a problem exists, and the magnitude of that problem

The **scientific method** involves four generally sequential steps:

1. **Observation/Question**—start off by noticing a phenomenon that poses a problem/question
2. **Hypothesis**—form a *suggested explanation* of the observation above; made on the basis of limited evidence or survey of existing literature; must be falsifiable
3. **Prediction**—then, deduce a specific prediction based on the hypothesis
4. **Empirical Test**—using everything else

In order to test the hypothesis, we need to translate theoretical concepts to actual measures through **operationalisation**. By measuring these to obtain values, we can quantify those concepts.

A **variable** is something whose value can vary

- A random variable is a variable that can take on one or more values, each associated with a certain probability
- Variables are either dependent or independent
- Dependent variables corresponds to the outcomes we want to explain, they are determined by other variables
- Independent variables vary independently of other variables, they determine (fully or partially) the dependent variable

Air Quality problems all begin with the emission of pollutants. Pollutants are substances in the air that change the original composition of the air around us, and has an adverse effect on human health, or the environment.

- Excessive levels of pollutants can cause adverse health effects. Singapore tracks 6 criteria health pollutants: $CO, NO_2, O_3, SO_2, PM_{2.5}, PM_{10}$.
- Sources of these air pollutants can be categorised as mobile or stationary, and primary or secondary. Primary sources directly emit pollutants into the air, secondary sources are the formation of pollutants through physical or chemical processes in the air.
- Air quality control measures aim to limit airborne concentrations to a level where no adverse health effects are expected.
- Air control measures can be source reduction, or air-scrubbing (indirect)

Carbon Monoxide, CO

- Non-reactive species in urban air
- Does not react rapidly with surfaces
- Low solubility in air
- Atmospheric concentrations vary directly with emissions
- Main fate is oxidation to CO_2
- Binds irreversibly with haemoglobin
- Impairs visual perception, work capacity, manual dexterity, learning ability, and performance of complex tasks
- Life-threatening at high levels (about 600 ppm)
- Elderly, young, and those with pre-existing cardiovascular diseases most sensitive
- From incomplete combustion of hydrocarbon-based fuels
- From (predominantly) vehicle exhaust
- From cigarette smoke (indoor pollutant)
- Control by catalytic converters to treat exhaust
- Control by inspection programmes (smog check)
- Control by using reformulated, oxygenated fuels

Nitrogen Dioxide, NO_2

- Most nitrous oxide missions are in the form NO, with remainder NO_2. However, since NO quickly oxidises to NO_2 in the air, nitrous oxide concentrations are reported and measured as NO_2 concentrations
- Respiratory irritant that lowers resistance to respiratory infections when inhaled
- Long-term exposure to elevated levels may cause increased incidence of acute respiratory diseases in children
- Contributes to formation of ozone
- Causes visibility problems (reddish-brown colour in smog)
- Forms acid deposition (rain): $NO_2 + OH \longrightarrow HNO_3$
- From oxidation of NO in the air, and NO is a combustion by-product from industries, power plants, and motor vehicles
- Control by combustion manipulation—using the minimum mix of atmospheric air (rich in N_2) for complete combustion
- Control by catalytic converters reducing NO_X back to N_2

Ozone, O_3

- Refers to low altitude ozone
- Strong oxidant that cause chemical damage to lung tissue and makes lungs more sensitive to other pollutants/irritants
- Can cause chest pain, coughing and nausea
- Long-term exposure may lead to permanent structural damage of lungs
- Damages flora (crops, trees) as well
- From secondary sources only: photochemical reactions in the atmosphere, with NO_X and VOC (volatile organic compounds) as reactants
- VOC from various sources, such as vehicle exhaust, fuel/solvent evaporation, industrial processes, plant emissions
- Control by reducing emissions of reactants NO_X and VOCs
- Control by controlling VOC through product substitution/ reformulation, sorption by activated carbon, better vapour containment, or catalytic oxidation

Sulfur Dioxide, SO_2

- Highly water soluble, likely to be absorbed by moist surfaces of upper respiratory tract
- But, can travel deeper into lungs when entrained in particles
- Affects breathing, causes respiratory illnesses, alters defence mechanisms of lungs, aggravates pre-existing symptoms
- Dominant precursor for acid deposition
- From combustion of sulfur-containing fuels (for electricity/industry)
- From combustion of high sulfur fuel by vehicles and off-road equipment (minor source)
- Control by reducing sulfur content in fuels
- Treating flue gas to remove SO_2

Particulate Matter, PM_{10} & $PM_{2.5}$

- Number following 'PM' is the 'size' of the particle in μm
- Particle behaviour dependent on particle size
- Course particles $> 2.5\mu m$ are controlled by mass and inertia, have lifetime of days to weeks before settling or impacting; therefore, effect mostly on areas near its source
- Ultrafine particles $< 0.1\mu m$ readily diffuse and coalesce with other particles; have lifetime of a few days to weeks
- Intermediate particles $0.1 - 2.5\mu m$ too small to settle, too light to impact, too large to diffuse; have lifetime on the order of weeks; can be transported across large distances before depositing
- Adverse effects on breathing, aggravates pre-existing respiratory and cardiovascular diseases, particularly asthma
- Impairs immune system, mechanical damage to lung tissue
- Major contributor against atmospheric visibility
- From primary sources, for coarse: road dust, construction, industrial, and agricultural emissions
- From primary sources, for fine: industrial emissions, diesel soot from vehicles, wood burning
- From secondary sources, for fine: oxidation of hydrocarbons into organics, formation of aerosol salts from SO_2 and NO_X
- Control by filtration mechanisms such as electrostatic precipitation, mechanical filtration

Pollutant Standards Index, PSI

Air Q.	PSI	24-hr PM_{10}	24-hr SO_2
Good	0 – 50	0 – 50	0 – 80
Mod.	51 – 100	51 – 150	81 – 365
Unh.	101 – 200	151 – 350	366 – 800
V. Unh.	201 – 300	351 – 420	801 – 1600
Hazard.	301 – 400	421 – 500	1601 – 2100
Hazard.	401 – 500	501 – 600	2101 – 2620

Air Q.	PSI	8-hr CO	8-hr O_3
Good	0 – 50	0.0 – 5.0	0 – 118
Mod.	51 – 100	5.1 – 10.0	119 – 157
Unh.	101 – 200	10.1 – 17.0	158 – 235
V. Unh.	201 – 300	17.1 – 34.0	236 – 785
Hazard.	301 – 400	34.1 – 46.0	786 – 980
Hazard.	401 – 500	46.1 – 57.5	981 – 1180

Air Q.	PSI	1-hr NO_2	24-hr $PM_{2.5}$
Good	0 – 50	–	0 – 12
Mod.	51 – 100	–	13 – 55
Unh.	101 – 200	1130	56 – 150
V. Unh.	201 – 300	1131 – 2260	151 – 250
Hazard.	301 – 400	2261 – 3000	251 – 350
Hazard.	401 – 500	3001 – 3750	251 – 500

All measures are in $\mu g/m^3$, except for CO which is in mg/m^3.

$$I_i = \frac{I_{i,j+1} - I_{i,j}}{X_{i,j+1} - X_{i,j}} \cdot (X_i - X_{i,j}) + I_{i,j}$$

- Relates health effects with varying levels of pollution
- Provides health advisories to the public on the necessary precautions when pollution levels go into unsafe levels
- Average pollutant concentration is converted into a sub-index, highest sub-index for a region is taken as its PSI
- Monitored through *Telemetric Air Quality Monitoring Network*, 22 stations islandwide place according to population density

Beach Water Quality

- Most common pollutant is faecal pollution
- From sources such as sewage leaks, decomposition, wildlife, humans, marine vessels dumping waste, industrial wastes, surface run-off
- Can cause diarrhea, vomiting, respiratory infections, infections of the eyes/ears/nose/skin; or gastroenteritis
- Levels of faecal pollution can be measured by indicators of coliform bacterial or faecal *streptococci*
- Total coliforms—can occur in human faeces but also naturally e.g. in soil or submerged wood; therefore not recommended
- Faecal coliforms—a subset of total coliforms more specific to human faeces, but contains genus *Klebsiella* which may also occur naturally e.g. textiles, pulp, and paper mill wastes
- *E. Coli*—a species of faecal coliform specific to human and mammal faeces; most specific measure of coliform bacteria; currently recommended by USEPA
- Faecal *streptococci*—generally occur in digestive systems of mammals
- *Enterococci*—a genus more human-specific than faecal *streptococci*, previously categorised under the genus *streptococcus*; currently recommended by USEPA as a indicator for salt water, since these more more able to survive in salt water environments

Singapore Recreational Water Guidelines

- 95% of the time, *Entorococcus* counts for the last three years must be less than or equal 200 counts per 100ml; these are rated 'Good' or better
- Also takes into account susceptibility of the location to faecal influence
- Only 'Good' or better beaches are considered suitable for primary contact activities
- Beaches are sampled weekly

Hong Kong Beach Grading System

Grd	BWQ	n/100ml	Ill/1000ppl
1	Good	≤ 24	–
2	Fair	$24 - 180$	≤ 10
3	Poor	$181 - 610$	$11 - 15$
4	V. Pr	< 610 or last > 1600	> 15

- Based on geometric mean of 5 most recent samples, or if last reading very high
- Public advised not to swim in Grade 4 beaches until quality improves
- During bathing season (March to October), beaches monitored at least three times a month, interval between sampling 3 to 14 days
- During non-bathing season (November to February), beaches which remain open still monitored at least three times a month, but other beaches monitored once per month

Empirical Tests are conducted on predictions made (through the scientific method), using data...

- Data—numerical values of physical quantities; collected through retrospective studies, observational studies, or designed experiments
- Retrospective studies—looks at limited historical records
- Hard to determine a 'sole impact', because lack of control over variables
- Independent variable may not have varied historically
- Form of data may not be ideal (want daily, have yearly)
- May not get most relevant data (want diesel vehicles, have all vehicles)
- Overall, difficult to obtain solid and reliable results; hard, if not impossible to conclusively prove causal relationships
- Observational studies—observe population of interest with minimal disturbance
- Usually carried out over a short period of time
- Allows measurements of some variables normally unavailable to retrospective studies
- Designed experiments—able to make deliberate changes to independent variables, in order to observe change in dependent variable
- Good control over what to measure, and how
- Time range to collect influenced by practicality, periodic trends (e.g. seasonal), relevance to current conditions (e.g. data from 1900s vs 2000s)
- Type of data chosen influenced by relevance (e.g. rainfall near *E. Coli* sampling sites); can include control measurements
- Generally, three factors leading to invalidation of data
- Measurement—such as counting errors, contamination
- Sampling—such as proximity of rainfall station, sample representative *E. Coli* samples of the whole beach, using outdated data
- Operationalise—such as choosing 12h, 24h, 48h, or 96h rainfall
- Followed by modelling, analysis, and critical thinking (e.g. examine assumptions made)

Probability

- Relative frequency of an outcome settles down to one value over the long run
- The one value is then the probability of that outcome

Basic Statistics

- Presents data economically
- Set of numbers to summarise data and its characteristics
- Three types: Measures of central tendency, dispersion and asymmetry
- Central tendency—a value that many systems tend to cluster around; can be taken as being representative of data collected
- Sample arithmetic mean—strongly influenced by outliers
- Median—also the 50^{th} percentile; less sensitive to outliers (although 95^{th} percentile is *more* sensitive to outliers
- Geometric mean—less sensitive to outliers, always $\leq$ the arithmetic mean; used in averaging values that represent rate of change
- Mode—most frequently occurring values, can be more than one
- A. Mean = Median $\implies$ symmetrical distribution; other direction not always true
- Range—extremely susceptible to outliers $\implies$ mainly used for small n
- Interquartile range—difference of the 1^{st} and 3^{rd} percentile; more resistant to outliers
- Variance (s^2/σ^2) and standard deviation (s/σ) —sensitive to extreme values
- Data can be considered to be dispersed if *coefficient of variation* > 1

Graphical Methods

- Dotplot—good for small pool of data, for spotting unusual data points, bad for spotting dispersion pattern
- Boxplot—good for highlighting distribution features, for comparing data measuring related or similar characteristics; box shows 1^{st}, 2^{nd} and 3^{rd} quartiles, whiskers to first data point within 1.5 IQR of box borders, rest are outliers
- Histogram—good for observing shape of distribution, central tendency and scatter, for guessing type of probability distribution, bad for small bin size; recommended number of bins is $\sqrt{n}$, traditionally bottom inclusive and top exclusive (but reversed for excel)
- Data points can be considered to be outliers if outside $[\bar{x} - 3s, \bar{x} + 3s]$
- Skew—a measure of symmetry; positive indicates right skew, tail to the right, negative indicates left skew, tail to the left; zero is symmetrical
- Time series—a plot of data against time, since time may be an important factor in variability; can compare trends in mean and dispersion through time

Other Statistical Terms

- Variability—can be higher-end, lower-end
- Sampling—choosing subset of population to characterise properties of population
- Sample space—all possible outcomes
- Event—subset of the sample space
- Outliers—data that do not seem to conform to distribution of the rest of the data; due to errors in data collection *(above)*

Formulas

Percentile Calculations

$$\text{Percentile Rank, } \mathrm{r} = 1 + p\,(n-1)$$

$$\mathrm{p}^{\text{th}} \text{ percentile} = (1 - (r \bmod 1))\, X_{\mathrm{int}(r)} + (r \bmod 1) \cdot X_{\mathrm{int}(r)+1}$$

Basic Probability

$$\begin{aligned} E_1 \ \text{ OR } \ E_2 &= E_1 \cup E_2 \\ E_1 \text{ AND } E_2 &= E_1 \cap E_2 \\ \text{NOT } \ E &= E' \end{aligned}$$

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}$$

Mutual Exclusivity

$$\begin{aligned} \text{if } \quad & E_1 \cap E_2 = 0 \\ \text{then } & \mathrm{P}(E_1 \cup E_2) = \mathrm{P}(E_1) + \mathrm{P}(E_2) \\ \text{and } & \mathrm{P}(E_1 \cap E_2) = 0 \end{aligned}$$

Independence

$$\begin{aligned} & \mathrm{E}_1 \text{ and } \mathrm{E}_2 \text{ are independent} \\ \iff & \mathrm{P}(E_1|E_2) = \mathrm{P}(E_1) \\ \iff & \mathrm{P}(E_2|E_1) = \mathrm{P}(E_2) \\ \iff & \mathrm{P}(E_1 \cap E_2) = \mathrm{P}(E_1)\mathrm{P}(E_2) \end{aligned}$$

Basic Statistics

$$\begin{aligned} \bar{x}_a &= \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{x}_g &= \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} \\ &= \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \end{aligned}$$

$$\begin{aligned} s^2 &= \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{(n-1)}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] \end{aligned}$$

$$\mathrm{CV} = \frac{s}{\bar{x}} \times 100\% \text{ or } \frac{\sigma}{\mu} \times 100\%$$