1. Since it is difficult ~~to~~ and rather unrealistic to find points that anger different subjects, I choose to perform experiments on subjects registered for anger management.

The experimental steps are given as below :

Calming Task → Angering Task ⟶ Venting Task ⟶ Anger Measurement

Control Group ⇒ Calming Task → Angering Task ⟶ Anger Measurement

* **Procedure**

1. Calming Task : Each subject will be allowed to do something that soothes them to ensure a balanced state of mind before starting the experiment. This can be done till the anger levels of the subjects is lesser than a threshold.

2. Angering Task : Each subject will undergo a series of tasks that are annoying to most people as an attempt to increase the anger levels of the subjects.

   These tasks can include waiting for a long time without anything to do.

interrupting a lot while performing a puzzling task, etc.

3. Venting Task : The subject will be asked to punch a punching bag or pillow, or some other tasks can be performed that ~~ca~~ are claimed to reduce anger level.

4. Anger Measurement : The therapist will be asked to measure the anger levels of the subjects belonging to the control group and the other group as well.

*The hypothesis can be stated True if the average anger levels of the ~~non~~ control group is higher than the ~~avg~~ average anger levels of the other group.

* Variables

Dependent : Final anger level of the subject.

Independent : Venting Task performed or not, characteristics of venting task.

Control : Initial (before angering task) anger level

___/___/___

* Inclusion Criteria

Only anger management patients with age within the range '20-30' will be considered to ensure similar maturity level and intellectual level.

2. (a) Tree Diagram

Population

0.2 % | 99.8 %

Infected | Not Infected

98% | 2% | 15% | 85%

	Test Positive	Test Negative	Test Positive	Test Negative
Probabilies	P = 0.00196	P = 0.00004	P = 0.1497	P = 0.8483
Events →	A	B	C	D.

(b) $P(E \text{ and } F) = P(A) = \cancel{0.000}\ \boxed{0.00196}$

(c) $P(F) = P(A \text{ or } C)$

$= P(A) + P(C)$

$= 0.00196 + 0.1497$

$= \boxed{0.15166}$

(d) $P(E-F) = P(E \text{ and not } F)$

$= P(B)$

$= \boxed{0.00004}$

3. (a), (b) and (c) - plots in the zip file.
↳ (code / q.3 / plots)

(d) The correlation is statistically significant (> 0.4) for all combinations, however is lesser between assets and profiles.

4. a) $\bar{X}$ [n → sample size]

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[1/n \sum_{i=0}^{n} X_i \right]$$

$$= 1/n \sum_{i=0}^{n} \mathbb{E}[X_i]$$

Since $\{X_i\}$ are i.i.d with mean μ, we have

$$\mathbb{E}[\bar{X}] = 1/n \sum_{i=0}^{n} \mu = \mu$$

$\therefore$ $\bar{X}$ is unbiased

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$ (From lectures)

as $n \to \infty$, we here

$$\lim_{n \to \infty} Var(\bar{X}) = \lim_{n \to \infty} \frac{\sigma^2}{n} = 0$$

$\therefore$ $\bar{X}$ is consistent.

b) Randomly chosen value from the sample

Suppose we represent as a ~~the~~ random variable 'I' ~~the~~ the ~~the~~ index of the sample to use.

Then the estimator is given as X_I.

$$\mathbb{E}[X_I] = \mathbb{E}\left[\sum_{i=0}^{n} \mathbb{I}[i=I]\, X_i\right]$$

$$= \sum_{i=0}^{n} p(i=I)\, \mathbb{E}[X_i]$$

$$= \mu$$

$\therefore$ The estimator is unbiased

$$Var[X_I] = Var[X_i]$$

$$= \sigma^2$$

$\therefore$ Estimator is not consistent

c) Average of largest and smallest value in sample.

Since the expectation cannot be computed, we cannot say anything about the bias of the estimator. Since the variance is not inversely proportional to the sample size, the estimator is not consistent d). $\bar{X} + 2/n$

$$\mathbb{E}[\bar{X} + 2/n] = \mathbb{E}[\bar{X}] + 2/n$$
$$= \mu + 2/n$$

$\therefore$ ~~not consistent~~ not unbiased

$$Var(\bar{X} + 2/n) = Var(\bar{X})$$
$$= \frac{\sigma^2}{n}$$

$\therefore$ consistent (From a)

Since $Var(\bar{X}) < Var(X_1)$,
The efficiency of (a) is greater than (b). We cannot compare with (c) and (d) as they are not unbiased.

_/_/_

5. (a) Plots in the zip file
↳ (code / q5 / plots)

(b) Pearson correlation (r)= 0.737345.
(Computed using python code)

(c) Since the correlation is high (based on r) we can say that if no other factors were changed (i.e all levels are constant), smoking would be ~~the~~ a likely cause for deaths due to lung cancer.

However this is not necessarily true ~~for~~ in the real world because there would be confounding factors and we cannot conclude causation in that case.

(d) A few confounding factors could be high pollution, ~~higher~~ ~~stress~~, high stress among smokers, lower income among most smokers and therefore lower health level.

(e) • Consumption of alcohol and lung cancer.
→ Although there is no biological dependency of one with ~~the~~ other, there is a chance of high correlation.

__/__/__

- Number of phones – No. of smart TVs
- Number of hospitals – ~~[illegible]~~ Deforestation

6\.

Frequencies ↓	Sample	Mean	Median	Avg of min and max
2x	1, 1, 2	4/3	1	3/2
1x	1, 1, 3	5/3	1	4/2
2x	1, 2, 2	5/3	2	3/2
4x	1, 2, 3	6/3	2	4/2
1x	2, 2, 3	7/3	2	5/2.

- Mean – Sampling Distribution.

Statistic	Probability
4/3	2/10
5/3	3/10
6/3	4/10
7/3	1/10

Average / Expected value → $\frac{4\times2}{30} + \frac{5\times3}{30} + \frac{6\times4}{30} + \frac{7\times1}{30}$

$= 54/30$

$= 1.8$

- Median - Sampling Distribution

Statistic	Probability
1	3/10
2	7/10
~~3~~	

Average / Expected Value → $\frac{1 \times 3}{10} + \frac{2 \times 7}{10} = 1.7$

- Avg. of Largest and Smallest - Sampling Distribution

Statistic	Probability
3/2	4/10
4/2	5/10
5/2	1/10.

Average / Expected Value → $\frac{3 \times 4}{20} + \frac{4 \times 5}{20} + \frac{5}{20} = \frac{37}{20}$

$= 18.5.$

* Since only the sample mean estimator is unbiased it would be best to use that ~~stt~~ as our population mean estimator.

7. Plots in the zip file
↳ (code / q7 / plots)

8. Probability of false alarm = 25% = 0.25

Suppose the random variables are given as $\{X_1 \dots X_{100}\}$. we have $X = X_1 + X_2 \dots X_{100}$.
where each X_i $(i = 1 \dots 100)$ is given as bernoulli random variable with 1 representing a false alarm.
$\therefore$ $X = \sum X_i$ represents the number of false alarms.
Also, we have $X = 100\bar{X}$ where $\bar{X}$ is the sample mean.

Since the length of sample (100) is large, we can approximate the probabilities using the sample statistic's sample distribution (Approximated as a gaussian)
Also, we have $\mu = 0.25$, $\sigma = \sqrt{\frac{p(1-p)}{100}} = 0.0433$

(a) $P(20 \leq X \leq 30)$
$= P(X \leq 30) - P(X \leq 19)$
$= P(\bar{X} \leq 0.3) - P(\bar{X} \leq 0.19)$

Approximating this using a standard gaussian variable Z, we have.

[$\Phi \rightarrow$ cdf of standard gaussian]

$$P(\bar{X} \leq 0.3) \approx P\left(Z \leq \frac{0.3 - 0.25}{0.0433}\right)$$

$$= \Phi(1.1547)$$

Similarly, $P(X \leq 0.19) = \Phi(-1.3857)$

$$\Rightarrow P(20 \leq X \leq 30) = \Phi(1.1547) - \Phi(-1.3857)$$

$$= \boxed{0.7930}$$

(b) $P(20 < X < 30)$

$$= P(X \leq 29) - P(X \leq 20)$$

$$= \Phi(0.9238) - \Phi(-1.1547)$$

$$= \boxed{0.6981}$$

(c) $P(35 \leq X)$

$$= 1 - P(X \leq 34)$$

$$= 1 - \Phi(2.0785)$$

$$= \boxed{0.0188}$$

__/__/__

(d) $P(\bar{X} \notin [0.25 - 2\times0.0433, 0.25 + 2\times0.0433])$

$= P(\bar{X} \le 0.16) + P(\bar{X} \ge 0.34)$

$= 1 + P(\bar{X} \le 0.16) - P(X \le 0.33)$

$\approx \boxed{0.0512}$

9. We have $\mu = 160$, $\sigma = 15$.

(a) $P(X > 197.5) = 1 - P(X \le 197.5)$

$= 1 - \Phi\left(\frac{197.5 - 160}{15}\right)$

$= 1 - \Phi(2.5)$

$= \boxed{0.0062}$

(b) $P(145 \le X \le 175 \text{ and } 135 \le X \le 185)$

$= P(145 \le X \le 175)$

$= \Phi(1) - \Phi(-1)$

$= \boxed{0.6827}$

__/__/__

(c) P(at leest one person is short)

$= 1 - P($ noone is short$)$

Suppose if n is sample size.

$\Rightarrow P = 1 - P\left(\prod_{i=1}^{n} X_i \geq 145\right)$

$= 1 - \prod_{i=1}^{n} P(X_i \geq 145)$ $\{X_i\}$ are i.i.d.

$= \boxed{1 - p^n}$

where $p = 1 - P(X_i < 145)$

$= 1 - \Phi(-1)$

$= \boxed{0.8413}$