Distributions and their Characteristics

1. A **distribution** of a variable is merely a list of all values the variable can take, and the corresponding frequencies. Distributions may be **represented** or **displayed** in myriad ways:

- a Frequency / Relative Frequency / Percentile **Table**;
- a Frequency / Relative Frequency **Histogram**;
- an Ogive: Cumulative Frequency Ogive or Cumulative Relative Frequency Ogive
- a Dotplot;
- a Stemplot;
- a Boxplot

2. The p-th percentile for a given X-value, a, is such that p% of ALL values [incomes, weights, heights, length of pregnancies, etc.] are $\leq a$: **$P(X \leq a) = p\%$**.

INTERPRETATION: If the 17th percentile of shopping expenditures was \$19.56, then 17% of expenses were \$19.56 or less **OR** 17% of shoppers spent \$19.56 or less.

You should be able to calculate percentiles [given X-values] and X-values [given percentiles]:

- from a list of numbers e.g. distribution of shopping
- from a frequency **or** probability [relative frequency] table e.g. distribution of Scottish militiamen chest-sizes, distribution of household size
- from a frequency **or** probability [relative frequency] histogram e.g. distribution of president's ages at inauguration, distribution of lengths of Shakespeare's words,
- from Ogives distribution of president's ages at inauguration
- from stemplots and dotplots e.g. caffeine content

Given *any* data set or distribution, we can find
i) the percentile corresponding to a given X by using the definition $P(X \leq a)$ and...common-sense.
ii) the X corresponding to a given percentile by finding the *position* of the X-value 1st.

Given 2 data-sets, we can determine and compare the Percentiles corresponding to a given value, a, for both sets: $P1 = P(X \leq a)$ and $P2 = P(Y \leq a)$. We can then determine what value of the 2nd data-set corresponds to a comparable value in the 1st.

3. To calculate the Mean, Median, Q1, Q3, IQR, Sx for a data set, use **1-Varstat** command:

- Given *only* **X-values:** use **1-Varstat L1** with L1: x-values
- Given **X-values** and **Frequencies or Relative Frequencies:** use **1-Varstat L1, L2** with L1: x-values and L2: Frequencies or Relative Frequencies
- Given **Class Intervals** and **Frequencies or Relative Frequencies:** use **1-Varstat L1, L2** with L1: mid-value of the Class Intervals and L2: Frequencies or Relative Frequencies

4. Terminology: These terms have the same connotations:

- Average ~ Mean ~ Expected Value
- Relative Frequency ~ Percentage ~ Probability ~ Proportion
- Percentile ~ Relative Cumulative Frequency ~ Cumulative Relative Frequency

5. The common plots / graphs are: dotplots, stemplots, boxplots, and histograms. In general, ***if applicable,*** label the axes, identify i.e. label the distributions, title the distributions and provide a Key for a stemplot.

Dotplots and stemplots display *actual* values of the variable. Histograms may do so [when the variable takes integer values] or may not [when the data is described by class-intervals]. Boxplots do not reveal the values of the variable, nor do they reveal the sample size, *n* [which dotplots, stemplots and histograms do: simply add up the frequencies!].

6. To describe univariate (1-variable) distributions: *in context,* discuss
a) The **Centre** of a Distribution which is any representative value of the data-set in terms of average value [Mean], the middle value [Median] and most frequently occurring value [Mode in the case of a Stemplot / Histogram / Dotplot]. Examine the context of the problem – e.g. pollution levels, weight-loss, score improvements, mileage, points missed, etc. – to determine if a [higher or lower] Centre of a distribution is preferable.

b) The **Spread** of a Distribution – denotes the variability or dispersion of a data-set, and answers the question: how spread out are the data? One might describe the Spread using IQR and the s.d. and also the Range, though it only offers a rough [and incomplete] picture. In general, a lower spread is preferable since it reveals consistency and predictability and reliability.

c) The **Shape** of the Distribution is addressed by discussing whether there is a specific *pattern* in the data. Are the numbers heavy on one end? Are there several clusters in the data-set? Are there gaps in the data? Is the distribution left-skewed, right-skewed or symmetric? Is the distribution unimodal (one peak) or bimodal (2 peaks) [***in the case of a stemplot / histogram***]? Are there any other obvious / unusual features...that are blatantly obvious? Are the outliers? Is the Q1 of one distribution comparable to, say, Q3 of the other?!

d) **Outliers**, if any, which are values that stand out from the overall pattern. For severely skewed distributions, observations *beyond* Q3 + 1.5 IQR and Q1 – 1.5 IQR are considered outliers. The Outlier limits **can** be large negative numbers [lower outlier limit] or very large positive numbers [upper outlier limit]. But these are only the limits...that **doesnt** mean there *are* outliers. After calculating the suggested cut-offs for the outliers, scan the data to determine if there *are* indeed outliers!

Note: all data-sets dont have Outliers...but we check for them nevertheless. There *may* be outliers at the lower end only **OR** outliers at the upper end only **OR** outliers at both ends **OR** no outliers at all.

Steps to determine Outliers:
I Calculate the Lower Outlier Limit: $Q1 - 1.5(Q3 - Q1)$ →
II Scan data-set: values *below* the Lower Outlier Limit are considered Outliers. **Identify any outliers.**
III Calculate the Upper Outlier Limit: $Q3 + 1.5(Q3 - Q1)$ →
IV Scan data-set: values *above* the Upper Outlier Limit are considered Outliers. **Identify any outliers.**

In boxplots, all outliers are depicted by dots; and the 1st non-outlier at the lower and upper ends...marks the end of the whisker. **If there are no outliers,** proceed naturally without any "complications": the 1st non-outliers are simply the Min and / or Max values.

7. **IQR**: is the spread / dispersion / variability / range of the middle 50% of the data-set, $Q_3 - Q_1 = 75^{th} - 25^{th}$ Percentiles

INTERPRETATION: If the IQR of amounts spent by shoppers is \$25.13, then the spread or range of the middle 50% of amounts was \$25.13.

8. **Boxplots**: are constructed using the **5 Number Summary** of Min, Q1, Q2, Q3 and Max values. Boxplots reveal *degree* of symmetricity ONLY; *not* precise shape of distribution (⇨ use Histogram / Stemplot / Dotplot). Boxplots do not suggest sample size [*n*] either. Symmetric Boxplots / distributions need *not* be bell-shaped.

For width of **scale ≈ (**≈Max – ≈Min**)**/5
For instance, in a data-set, if the Minimum value is 20 and the Maximum is 520, then use a scale of (550 – 0)/5 ≈100

9. **Histograms**: reveal rough center, shape and spread; *not* Numerical Summaries ⇨ use 1-Varstat L1, L2.

10. Mean and s.d. are immediately affected by the presence of extreme values or outliers since they are based on the *actual* values [as opposed to Median and IQR which are only based on *position*]. In a skewed distribution, the Mean *tends to* get "pulled" in the direction of the skew so that extreme values raise or lower the Mean towards themselves.

For **skewed distributions** do *not* use Mean or s.d. as measures of Centre and Spread, since Mean is influenced and hence "distorted" by extreme values ⇨ use Median and IQR since they are relatively unaffected.

For roughly **symmetric distributions**, use Mean and s.d.

Measures that depend on Position e.g. Percentiles, Median and IQR, are resistant to outliers since they are based on relative *position* and not on the actual values.

11. For **left-skewed** distributions,
- the smaller values of X occur with low frequencies while larger values, more frequently:
- Mode > Median > Mean.
- Q2 – Min >> Max – Q2.

For **right-skewed** distributions,
- the larger values of X occur rarely, while the smaller values occur more often
- Mode < Median < Mean.
- Q2 – Min << Max – Q2.

12. The **Standard Deviation**, $Sx = \sqrt{[\sum(X - Mean)^2/(n - 1)]}$, is a measure of the variability or

dispersion of each X-value from the sample mean **OR** it indicates how far, on average, each X-value is away / from the Mean.

INTERPRETATION: If the s.d. of amounts spent by shoppers is $5.46 when the mean amount is $84.58, then
a) The s.d. of $5.46 is a measure of the average variability of each amount from the mean expenditure of $84.58 **OR**
b) The s.d. of $5.46 indicates that, on average, each shopping amount is about $5.46 from the mean of $84.58.
c) The s.d. of $5.46 indicates that, on average, the shopping amounts differ from the mean of $84.58 by about $5.46.
d) The s.d. of $5.46 indicates that, on average, the difference between the individual shopping amounts and the mean of $84.58 is about $5.46.

13. **Linear Transformations**: Measures of Centre and Position [Mean, Median, Percentiles] and Measures of Spread [Range, IQR and s.d.] are affected by transformations in a similar fashion.

I *Adding* a constant, ± *k*, to each data value, X: changes the Measures of Center [Mean, Median, Mode] and Position [Percentiles] by ± k units; Measures of Spread [Range, s.d., IQR] are *not* affected.

II *Multiplying* each data value, X, by a constant *k*: changes Measures of Center [Mean, Median, Mode], Position [Percentiles] *and* measures of Spread [Range, s.d., IQR] by *k* units.

Example 1

1. What are the 3 measures of 3 centre to describe a distribution?
2. What are the 3 measures of 3 spread to describe a distribution?
3. To describe a skewed distribution, which measures of Centre and Spread *should* you use? *Why*? What advantage have they got over the other measures?
4. For which plots would you employ Mode to describe the Centre of a distribution?
5. What is the Outlier Rule of Thumb?
6. In case of a Histogram / Stemplot / Dotplot, what 4 aspects of the distribution would you discuss *re* **Shape**?
7. For which type of distribution would the Mean tends to be higher than the Median? [**Note:** this is <u>not</u> always true; hence the phrase "tends to be"...there *are* counterexamples!]
8. Suppose the 61st percentile of Incomes was $73,000. Interpret it in simple English [*without* using statistical terminology].
9. Suppose the Mean of incomes was $432,765 with a s.d. of $98,236. Interpret the s.d. in simple English.
10. Suppose, for a country, most families practice contraception / birth control.
a) What would the shape of the distribution of <u>Family Size</u> be? Right-skewed or left-skewed? Explain.
b) Would the Mean Family Size be higher or lower than the Median?
11. Suppose the distribution of <u>Number of Shots</u> made by a basketball player is left-skewed.
a) Is he a good player or bad? Explain.
b) Would the Mean number of baskets be higher or lower than the Median?
12. Suppose a teacher scales a hard test by raising everybody's by 8%. If these are the *erstwhile* Summary Statistics, determine the corresponding *new* ones: Min = 14, Q1 = 31, Med = 48, Q3 = 57,

Max = 73, Mean = 41, s.d. = 18. ***Also***, determine the new Range and IQR.
13. Suppose that a truck rental company rents a moving truck such that the costs, C, are related to the the Number of Miles driven, m, by the relationship: C = $0.20m + $29. If these are the *erstwhile* Summary Statistics for the Number of miles driven, determine the corresponding Cost, C: Min = 19mi, Q1 = 25mi, Med = 29mi, Q3 = 67mi, Max = 82mi, Mean = 58mi, s.d. = 15mi. ***Also***, determine the Range and IQR of Costs. **Show minimal work.**

Solution.

1. Median, Mean, Mode.
2. Range, IQR, s.d.
3. Median, IQR since they are **resistant** to extreme values / outliers unlike Mean and s.d.
4. Histograms, Stemplots, Dotplots
5. Observations < Q1 - 1.5IQR and those > Q3 + 1.5IQR.
6. Symmetricity, Modality [if possible!], Outliers, Gaps / Clusters
7. Right-skewed distribution.
8. The 61st percentile of Incomes being $73,000 indicates 61% of incomes were $73,000 or less **or** 61% of individuals had incomes of $73,000 or less.
9. The s.d. of $98,236 is a measure of the average spread / variability of the individual [or each] income[s] from the mean of $432,765. **OR** The s.d. of $98,236 suggests that each income differs from the mean of $432,765 by about $98,236.
10a) Since most families practice contraception, theyd tend to have small families...with a few with large ones. This would correspond to a Right-skewed distribution.
b) For a Right-skewed distribution, the Mean tends to be higher than the Median.
11. Since the distribution of Number of Shots made is left-skewed, most of the time he make a large Number of Shots and rarely does he make a few. Ergo, he *is* a good player.
b) Mean < Median since the Mean would get pulled in the direction of the [left] skew i.e. bottom.
12. A raise of 8% ~ multiplying by 1.08 => ALL statistics [measures of Centre / position ***and*** Spread would be affected]. Do this yourselves.
13. Multiplication [0.20] shall affect ALL statistics [Centre and Spread] while addition [29] shall influence *only* measure of Centre / Position: Minimum Cost = 19·0.2 + 29 = $32.80 and Median Cost = 29·0.2 + 29 = $34.80 while s.d. Cost = 0.2·15 = $3 and IQR(Cost) = 0.2·42. Do the rest yourselves.

Example 2.

a) Suppose that youre the manager of a sports team. Suppose youre recruiting a new member. Recalling the shape of the Income distribution, which measure of Centre (mean / median) would you use if you want to attract him / her?
b) Imagine a distribution of scores of an extremely competent teacher. The shape of the distribution would be (left- / right-) skewed?
c) In a left-skewed distribution, most values are on the (lower / upper) end?
d) At a community college, the distribution of SAT scores of those admitted is likely (left- / right-) skewed?
e) For a strongly skewed distribution, which measure of Centre **should** one employ? which measure of Spread **should** one employ?
f) If in a distribution, most values are in the lower-end, that would be a (left- / right-) skewed distribution?
g) Suppose the Mean of a data-set is 5lbs with a s.d. of 1.2lbs. Interpret the s.d. *in context*.

h) The 3 measures of Spread are:
i) Imagine a distribution of property-values in an impoverished city [Detroit!]. The shape of the distribution would be (left- / right-) skewed?
j) At MIT, re the distribution of SAT scores of those admitted, the Mean score is likely (higher / lower) than the Median?
k) Sketch a smooth curve depicting a rough right-skewed distribution. Determine the relative positions of Mean, Median and Mode.
l) Suppose youre a terrible basketball player. Which measure of Centre would you prefer to report: (mean / median) number of baskets?
m) The 3 measures of Centre are:
n) For a strongly skewed distribution, which measure of Centre should one **not** employ? which measure of Spread should one **not** employ?
o) Sketch a smooth curve depicting a rough left-skewed distribution. Determine the relative positions of Mean, Median and Mode.

Solution.

a) Income distributions are Right-skewed; to impress the recruit, report the Mean earnings...which shall be pulled up, in the direction of the skew!
b) For an extremely competent teacher, most scores shall be on the high side => left- skewed.
c) In a left-skewed distribution, most values are on the upper end.
d) At a community college, the distribution of SAT scores of those admitted is likely right skewed -- since most scores shall be low while a few very high!
e) For a strongly skewed distribution, employ the IQR since it is unaffected by extreme values!
f) If in a distribution, most values are in the lower-end, that would be a left skewed distribution.
g) The s.d. of 1.2lbs is a measure of the average difference of each weight from the mean weight of 5lbs **OR** The s.d. of 1.2lbs indicates that each weight differs from the mean weight of 5lbs by about 1.2lbs.
h)The 3 measures of Spread are IQR, Range and s.d.
i) Imagine a distribution of property-values in an impoverished city [Detroit!]. The shape of the distribution would be right skewed since most values shall be lower while a few, really high.
j) At MIT, re the distribution of SAT scores of those admitted, the Mean score is likely lower than the Median since most scores shall be on the higher end (~2400!) while a few shall be lower making it a left-skewed ditribution => mean would get pulled in the direction of the skew!
k) For a right-skewed distribution, Mean > Median > Mode.
l) Suppose youre a terrible basketball player. Report the Mean number of baskets since most of the time youd be scoring poorly, and rarely doing really well => mean would get pulled up in the direction of the skew!
m) The 3 measures of Centre are: Mean, Median and Mode.
n) For a strongly skewed distribution, do **not** the Mean re Centre. For Spread, do **not** employ the Range or s.d.
o) For a right-skewed distribution, Mean < Median < Mode.

Example 3

1. *Suppose* the 40th %ile of amounts spent on groceries at Store A was $50 and at Store B, the 40th %ile was $60. Write 2-3 sentences to explain -- in simple English -- which store you'd rather be the manager of. [Tip! How would you *interpret* both %iles? Write a conclusion...]
2. *Suppose again* at Store A the 40th %ile of amounts spent on groceries was $50 and at Store B,

the 60th %ile was $50. Write 2-3 sentences to explain -- in simple English -- which store you'd rather be the manager of. [Tip! How would you *interpret* both %iles? Write a conclusion...]

Solution.

1. 40% of the shoppers spent $50 or less in Store A while 40% of the shoppers spent $60 or less in Store B. Since the same % of shoppers spent a higher amount in B, being the manager of store B is preferable. [Alternately, 60% of shoppers spent > $50 in store A while the same % spent > $60 in B...]
2. 40% of the shoppers spent $50 or less in Store A while 60% of the shoppers spent $50 or less in Store B. Since a greater % of shoppers spent the same amount in B than A, being the manager of store A is preferable. [Alternately, 60% of shoppers spent > $50 in store A while only 40% spent > $50 in B...]

General Principles for Describing & Comparing Distributions

-> ALWAYS **compare** CENTRE, SHAPE, SPREAD and OUTLIERS in context: ie. What do those observations represent? Height, weight, income, flexibility ratings?
-> Use numerical summaries (ie. NUMBERS!) for each of the above. Do **NOT** say: "From the graph/plot, it is *obvious...***What numbers did you examine to make your decision about center/spread/etc.?**
-> ALWAYS write a sentence giving the Big Picture for BOTH center and spread (variability / consistency)!
-> Write a *final* sentence giving your conclusions about center and spread.

Center:

1. For Boxplots, Describe/Compare Mean and Median. Interpret in context!
2. For Stemplots and Histograms, ADDITIONALLY provide the modal value or class interval. (What was the value or class interval for which the frequency was *highest*?)
3. Do NOT confuse Median with 'Average': *Average* means Mean! When you use *Average* you'd better be talking about mean, and NOT median!
4. **In general, for <u>strongly</u> skewed distributions, do NOT use Mean** (since it's afftected in the direction of the skew!)**: use Median!**

Spread: Describe/Compare

1. Range = Max-Min
2. IQR = spread of the middle 50% of the observations
3. Use terms like Variability and Consistency. For eg. The student race completion times were less variable and more consistent than the faculty completion times.
4. For roughly symmetrical distributions, ADDITIONALLY Mention / Compare, *Sx*, the standard deviation~ spread/deviations of observations about the *mean.*

Shape:

1. Use terms like Slightly/Roughly/Clearly. Use them CAREFULLY, however!
2. Unless you draw a Stemplot/Histogram and you know the shape of the disbn, **do NOT confuse Normal for Symmetric**! Normal distributions are Symmetric while Symmetric distributions are not *always* Normal. **<---------- OMG, common source of confusion!**
3. For left skewed distributions (most observations are on the right!): the left whisker is long on the BoxPlot. Also, Q2, Q3 and Max are close.

4. For right skewed distributions (most observations are on the left!): the right whisker is longer. Also, Q1, Q2 and Min are close.
5. For Symmetric distributions, Q1 and Q3 are equidistant from Q2. Min and Max are *also* equidistant from Q2.
6. Mention if the distribution is Uni/B-Modal.
7. Discuss any Gaps and Clusters in the distribution. **<---------- Remember this!**

Outliers:

1. **Values Below:** Q1 – 1.5·IQR and **Values Above:** Q3 + 1.5IQR
2. For Roughly Symmetric distributions, use **Values BEYOND Mean ± 3·Standard Deviations**

Percentiles

Given an X-value and a **Percentile**, the Percentile gives you what % of ALL values are ≤ the X-value.
Technical Definition that requires us to round up to the nearest integer: If the p-th percentile is X, then ***at least*** p% of ALL values are ≤ X and ***at least*** (1 – p)% are ≥ X.
Note: some texts use a slightly alternate definition: If the p-th percentile is X, then ***at least*** p% of ALL values are < X and ***at least*** (1 – p)% are > X. Either is fine!

Finding an X-value...given a Percentile

If the p-th percentile is X, then p% of ALL values are ≤ X.

That is, we need an X-value such that p% [given] of ALL values are ≤ the X-value.
If we knew its *position*, we should be able to locate it! For this, calculate p% of N [and round *up*].
No, go ahead and locate the corresponding X-value.
Understand this well! Do <u>not</u> memorize the process...it should be intuitive i.e. make obvious sense!

Example: find the 62nd percentile of ages for 31 children.
Required: an age, X, such that 62% of ALL ages [or children] are X years old or younger. In other words, 62% of the 31 [or children] are X years old or younger.

Now, IF there were 100 children, we would just need the age of the 62nd child [position = 62]. But since there are 31 children, we need the age of the 62% of 31 = 19.22th child ~ the 20th child! [**Always round up since rounding down shall not yield the required percentile – using the definition.]** We can then scan the data to locate the age of *that* child.

BIG IDEA: If we could find the POSITION of the X-value, we can easily find the X-value itself! Well, P% of N yields the desired POSITION → simply count up to reach the desired X-value...

Finding the Percentile...given the X-value:

If the p-th percentile is X, then p% of ALL values are ≤ X.

A percentile is a type of percentage! So, we just need to find out what % of values are at X are below it. We *know* the position of the X-value [by scanning the sorted data!]. Let this be: *p*. To find the p-th percentile, simply calculate: *p*/N => *That* would give you a percentage. **<----------**
Understand this well! Do <u>not</u> memorize the process...it should be intuitive i.e. make obvious sense!

Example: what percentile does a child that is 15years old correspond to, if there are 31 children.
Required: the percentile, P, such that P% of ALL ages [or children] are 15 years or younger.

For the percentile, lets first determine *how many* kids are 15 years or younger. **Suppose** 8 were. [That is, the 15 year old lies in the 8^{th} position amongst 31 kids.]

So, 8 kids are 15 years old or younger...so 15 years corresponds to the $8/31 = 25.8\% \sim 25.8^{th}$ percentile!

BIG IDEA: If we could find the POSITION of the X-value, we can easily find the percentile via: $p/N \cdot 100$.

A couple of common sources of confusion:

- For percentiles, we are interested in the actual X-value, *not* the position. [We employ the position to *find* the X-value!] So, we talk in terms of the 43rd percentile of ***ages***, the 36th percentile of ***incomes***, the 90th percentile of ***weights***.
- When using a TABLE with X-values and Frequencies, do not mix-up the frequencies with the X-values. By "counting up" the frequencies, we're finding the *position* of the X-value. But we *still* need the corresponding X-value!
- Always provide **units**, when applicable.

Interpreting Percentiles: This is the mad-libs version of it.
If the p-th percentile of N [x-variable] is X, then p% of N [individuals] have an [x-variable] of X or less.

Examples
If the 20th percentile of the 60 life-expectancies is 45 years, then 20% of women [/ countries?] have a life-expectancy of 45 years or less.
If 100lbs corresponds to the 31st percentile, then 31% of weights / children have a weight of 100lbs or less.
If 12% of the population being seniors corresponds to the 62nd percentile, then 62% of states [counties?] have 12% of fewer of their population as seniors.

Example 4

The numbers below denote the horsepower of n = 38 vehicles. I've sorted it for you below [*What a guy!*]: 65, 65, 68, 68, 69, 70, 71, 71, 75, 78, 80, 80, 85, 88, 90, 90, 95, 97, 97, 103, 105, 109, 110, 110, 115, 115, 115, 120, 125, 125, 129, 130, 133, 135, 138, 142, 150, 155

a) Find the 1^{st} and 3^{rd} Quartiles. Interpret the 1^{st} Quartile. [Tip! The 3 Quartiles are the 25th, 50th and 75th percentiles.]
c) Find the 68th and 91st percentile and interpret the *former*.
d) Two vehicles with horsepower of 103hp and 120hp lie in what percentiles? Interpret the *former*.

Solution.

a) For Q1, $0.25 \cdot 38 = 9.5 \rightarrow$ 10th value => 78hp
Interpretation: A car with a horsepower of 78 corresponds to the 25th percentile which means that 25% of all [given] vehicles have a horsepower of 78hp *or less*.

Q3 = 129hp
c) 0.68·38 = 25.84 ~ 26th value → 115hp
Interpretation: A car with a horsepower of 115 corresponds to the 68th percentile which means that 68% of all [given] vehicles have a horsepower of 115hp *or less*.
Do this yourselves: find the 91st percentile.
d) P(X ≤ 103) = 20/38 = 52.63rd percentile
Interpretation: A car with a horsepower of 103 corresponds to the 52.63rd percentile which means that 52.63% of all [given] vehicles have a horsepower of 103hp *or less*.
Do this yourselves: find the percentile of a car with 120hp.

Example 5 The literacy rates for 19 North African nations is as (in %):
34.3, 48.9, 54.6, 59.6, 77.8, 83.9, 86.6, 88.5, 88.9, 91.6, 92.2, 92.5, 94.9, 96.1, 96.9, 97.4, 97.8, 98.6, 98.8
That for the 23 Central African nations is:
34.6, 34.8, 35.5, 38.6, 40.9, 43, 45.4, 53.9, 58.3, 58.4, 61, 61.9, 61.9, 63.4, 63.8, 66.7, 69.9, 73.9, 74.4, 79.6, 85, 97.4, 98.9
1. Compute the 5 Number Summary for both data-sets: Min, Max, 25th, 50th and 75th percentiles.
2. Calculate the Lower and Upper Outlier Limits for North Africa and the corresponding Outliers. Also, identify the 1st non-outliers.
3. Calculate and interpret the Inter-Quartile Range for literacy rates in North Africa.

Solution.

1. The IQR of literacy rates for N. Africa was (96.9 - 77.8) = 19.1% which is the spread of the middle 50% of literacy rates.
2. For N. Africa,
the LOWER Outlier Limit is Literacy Rates < Q1 - 1.5IQR = 77.8 - 1.5(96.9 - 77.8) = 49.15%
the UPPER Outlier Limit is Literacy Rates > Q3 + 1.5IQR = 96.9 + 1.5(96.9 - 77.8) = 125.55%
The countries w Literacy Rates 34.3% and 48.9% are Lower Outliers for N. Africa.
The 1st non-outlier for N. Africa is: 54.6%. <---------- constitutes the end of the whisker at the upper-end in the boxplot.
3. The 5 Number Summary for the N. African nations **n = 19**:
34.3, 48.9, 54.6, 59.6, **77.8**, 83.9, 86.6, 88.5, 88.9, **91.6**, 92.2, 92.5, 94.9, 96.1, **96.9**, 97.4, 97.8, 98.6, **98.8**
The 5 Number Summary for the C. African nations **n = 23**:
34.6, 34.8, 35.5, 38.6, 40.9, **43**, 45.4, 53.9, 58.3, 58.4, 61, **61.9**, 61.9, 63.4, 63.8, 66.7, 69.9, **73.9**, 74.4, 79.6, 85, 97.4, **98.9**

Example 6

The probability distribution for the number of repairs, *N*, a brand of refrigerator requires over a 5-year period is:

N	0	1	2	3	4	5	6
P(*N*)	0.17	0.32	0.23	0.15	0.08	0.04	0.01

a) Estimate the percentile corresponding to 3 repairs. Interpret it, in context.
b) Determine the 3 quartiles. Then, calculate and interpret the IQR, in context.
c) Which measure***s*** of Centre ought to be used to accurately describe the number of repairs needed over a 5-year period? Explain.

d) Do you expect the Mean number of repairs to be higher or lower than the Median? Why?
e) Calculate the expected number of repairs needed over a 5-year period. [What is being asked?]
f) Calculate the s.d. of the number of repairs needed over a 5-year period. Interpret it, in context.

Solution.

a) $P(X \leq 3) = P(X = 0) + ...P(X = 3) = 0.17 + \ldots 0.15 = 0.87$. The 87th percentile being 3 repairs indicates that 87% of repairs over a 5-year period were 3 or less **or better still** The 87th percentile being 3 repairs indicates that 87% of refrigerators needed / had 3 or less over a 5-year period. **Observe the detailed context!**
b) Q1 = 1 since $P(X \leq 1) \geq 0.25$ and $P(X \geq 1) \geq 0.75$. <---------- This is the technical definition of percentile, p, corresponding to $X = a$: **$P(X \leq a) \geq p$ and $P(X \geq a) \geq 1 - p$**...and this definition permits us to "cross" the percentage!
Q2 = 2 since $P(X \leq 2) \geq 0.50$ and $P(X \geq 2) \geq 0.50$.
Q3 = 3 since $P(X \leq 3) \geq 0.75$ and $P(X \geq 3) \geq 0.25$.
IQR = 3 – 1 = 2 is the spread or range of the middle 50% of the number of refrigerator repairs needed in a 5-year period. **Observe the detailed context!**
c) Median and Mode since since this a right-skewed distribution.
d) The Mean shall likely be higher than the Median since it is a right-skewed distribution, so the Mean shall be "pulled up" by the extreme values of X on the upper-end. **Observe the detailed context!**
e) $E(X) \approx$ Average $= \Sigma X \cdot P(X) = 0 \cdot 0.17 + ...6 \cdot 0.15 = 1.81$ using 1-Var Stats L1, L2 with X-values in L1, and probabilities in L2 **You must have mastered the Topic II Notes emailed on Fri and Sat** *as well as* **the CW Notes!**
f) $\sigma(X) = 1.3978$ using 1-Var Stats L1, L2 with X-values in L1, and probabilities in L2.

BE FAMILIAR WITH THE INTERPRETATIONS BELOW!

- The s.d. of 1.3978 is a measure of the average variability of the different / each number of repair(s) from the mean number of repairs of 1.81 **Observe the detailed context! OR**
- The s.d. of 1.3978 indicates that, on average, the different / each number of repair(s) is about 1.3978 from the mean number of repairs of 1.81**Observe the detailed context! OR**
- The s.d. of 1.3978 indicates that, on average, the different / each number of repair(s) differs from the mean number of repairs of 1.81 by about 1.3978 **Observe the detailed context! OR**
- The s.d. of 1.3978 indicates that, on average, the difference between the different / each number of repair(s) and the mean numbers of repair of 1.81 is about 1.3978. **Observe the detailed context!**

Example 7

The table below is a Relative Frequency distribution of the Number of Accidents by bus drivers. For example, 16.5% of bus drivers had no accident whereas 0.1% of bus drivers had 11 accidents.

Number of Accidents	Relative Frequency (%)
0	16.5
1	22.2

2	22.3
3	16.2
4	11
5	6.2
6	3
7	1
8	0.8
9	0.1
10	0.4
11	0.1
	N = 708 bus drivers

1a) Determine if the Distribution of Number of Accidents is left-skewed or right. **Tip!** Imagine a Histogram...**Explain [1 sentence, using the definition].**
b) Do you expect the Mean to be higher or lower than the Median? Explain.
2. *How many* bus drivers had 4 accidents or less?
3. What proportion of bus drivers had between 3 and 6 accidents [inclusive]?
4. Calculate the percentile corresponding to 5 accidents. **Interpret this.**
5. Calculate the Quartiles for the Number of Accidents.
6. Sketch, label and Title a Boxplot for the distribution of Number of Accidents. For this, which accidents are Outliers? Calculate the Outlier Limits, 1^{st}.
7. Calculate the percentiles for 1, 3, 5, 7, 9 and 11 accidents.
8. Enter the values into your calculator:

L1: Number of Accidents	**L2: Relative Frequency (%)**

For L2: enter the Relative Frequencies as ***decimals***.
E.g. 16.5% ~ 0.165, 1% ~ 0.01, and 0.1% ~ 0.001. Use STAT → CALC → 1-Var Stats L1, L2. Compute the Mean, the s.d., Sx.

Solution.

1a) Since less than 10% of bus drivers had $\geq$ 5 accidents, most had fewer accidents while only a few had a large number, it's a right-skewed distribution! **Alternately,** most accidents were on the lower end / most drivers had few accidents [0-4] whereas very few accidents were on the higher end / few drivers had large numbers of accidents, ergo, a right-skewed distribution.
b) Since it's a right-skewed distribution, the Mean is likely higher than the Median.
2. $P(X \leq 4) = 88.2\%$ so that 88.2% * 708 ~ 624drivers.
3. $P(3 \leq X \leq 6) = 36.4\%$
4. Translated, the Q is asking, what % of drivers had $\leq$ 5 accidents → 94.4% ~ 94.4^{th} percentile, indicating that 94.4% of drivers had 5 accidents or less.
5. Since the Relative Frequencies (%) are GIVEN, we just need to add them to get the percentiles!
Q1 ~ 25^{th} percentile = 1,
Q2 ~ 50^{th} percentile = 2,

Q3 ~ 75th percentile = 3
Note: Q1, Q2 and Q3 being so close and *low* while the Max = 11 obviously indicates a Right-skewed distribution!
6. Do this yourselves, adhering strictly to AP Expectations!
Note: All accidents *exceeding* 6: 7, 8, 9, 10 and 11 – are outliers since values > Q3 = 1.5(Q3 – Q1) = 6 and shall be denoted with DOTS while the 1st non-outlier i.e. 6, shall be the end of the right whisker. There are no outliers on the lower end → the Min = 0 shall get the end of the left whisker!
7. Using the X-axis as 1, 3, 5, 7, 9 and 11. Plot:

Number of Accidents	Percentiles (%)
≤0	16.5
≤1	38.7
≤3	77.2
≤5	94.4
≤7	98.4
≤9	99.3
≤11	99.8*

* Doesnt add up to 100% because of Rounding Errors in the *original* table, relax.
8. XBar = 2.2825, Sx = 1.8366

Example 8
Consider the distribution of weights of WalMart shoppers below.

Weights (lbs)	Percentiles (%)
≤0	0
≤50	1
≤100	6
≤150	20
≤200	38
≤250	65
≤300	86
≤350	95
≤400	99
≤450	99
≤500	100

a) Identify the class-intervals containing the 5 Number summary.
b) What proportion of shoppers weigh 300lbs or less?

c) What proportion of shoppers weigh more than 150lbs?
d) What proportion of shoppers weigh 450lbs or less?
e) What weight lies in the 95th percentile of shopper's weights?
f) What proportion of shoppers weigh more than 300lbs?
g) What proportion of shoppers weigh between [excluding] 200lbs and 350lbs [inclusive]?
h) The 99th percentile of weights was 400lbs *and* 450lbs. What does this indicate?
i) Which 2 class-intervals corresponds to the middle 60% of weights?
j) Which class-interval corresponds to the top 30% of weights?

Solution.

a) Q1: 150-200lbs; Q2: 200-250lbs; Q3: 250-300lbs
b) $P(X \leq 300) = 0.86$ [since the 86th percentile is 300lbs!]
c) $P(X > 150) = 0.80$ [since the 20th percentile is 150lbs!]
d) $P(X \leq 450) = 0.99$ [since the 99th percentile is 450lbs!]
e) 350lbs
f) $P(X > 300) = 0.14$ [since the 86th percentile is 300lbs!]
g) $P(200 < X \leq 350) = 0.57$ since 200 is at the 38th percentile and 350 is at the 95th.
h) This indicates that 0% ~ nobody had weights between 400 and 450lbs. **Note:** You had a Q like this on the weekend re *Amount of Change in Students' pockets*.
i) We need the 20th and 80th percentiles: *a* = 100-150lbs and *b* = 250-300lbs
j) We need the 70th percentile: 250-300lbs.

Example 9

The table below gives the distribution of the Number of Rooms for Owner-occupied units in San Jose, California:

# of Rooms	1	2	3	4	5	6	7	8	9	10
Owned	0.003	0.002	0.023	0.104	0.210	0.224	0.197	0.149	0.053	0.035

a) Calculate the 5-Number Summary for the distribution.
b) What percentiles do 4 and 7 rooms correspond to? Interpret the percentile for 4rooms, in context.
c) [Imagining a histogram with X-axis ~ # of rooms, Y-axis: Probabilities], what it is the approximate shape of the distribution: left-skewed, right-skewed or reasonably symmetric?
d) Use the results of a) to *quickly* make a boxplot.
e) Confirm your results for c) ***visually:*** by using d).

Solution.

a) Min = 1, Q1 = 5, Q2 = 6, Q3 = 7, Max = 10
b) Do this yourselves, showing work.
c) Reasonably symmetric since the tails at both ends [1-4 and 8-10] are almost flat and gently rising, with most X-values peaking at the middle. Perhaps, slightly left-skewed since the left probabilities [1-3] are *much* lower than the right [8-10].
d) Do this yourselves. Make a scale 1st, then label the 5-Number Summary.
e) Do this yourselves.

Example 10

a) Suppose that youre the manager of a sports team. Suppose youre recruiting a new member. Recalling the shape of the Income distribution, which measure – mean or median – would you use if you want to attract him / her to join? Explain briefly.
b) Imagine a distribution of scores of on a hard exam.

- The shape of the distribution would be (left- / right-) skewed? Explain briefly.
- The Mean score shall be [lower / higher] than the Median? Explain briefly.

c) In a left-skewed distribution, most values are on the (lower / upper) end?
d) At a community college [**Note**: *everybody* is accepted!],

- the distribution of SAT scores of those admitted is likely (left- / right-) skewed? Explain briefly.
- The Median score shall be [lower / higher] than the Mean? Explain briefly.

e) For a strongly skewed distribution, which measure of Centre – Mean or Median – should one <u>not</u> employ? **Tip!** Which of the 2 shall get affected by extreme values, so that itd *distort* the true picture?
f) If in a distribution, most values are in the lower-end, that would be a (left- / right-) skewed distribution?
g) Suppose this was the Distribution of Ages of Individuals in a Park:
0 years, 0^{th} percentile
60 years, 25^{th} percentile
75 years, 50^{th} percentile
80 years, 75^{th} percentile
100 years, 100^{th} percentile
Just by examining the data, is the distribution, left- or right-skewed? How can you tell?
h) Imagine a distribution of property-values in an impoverished city [Detroit!].

- The shape of the distribution would be (left- / right-) skewed? Explain briefly.
- The Mean values shall be [lower / higher] than the Median? Explain briefly.

i) At MIT, re the distribution of SAT scores of those admitted, the Mean score is likely (higher / lower) than the Median? Explain briefly.
j) Suppose this was the Distribution of Ages of Individuals in a Park:
0 years, 0percentile
3 years, 25^{th} percentile
5 years, 50^{th} percentile
6 years, 75^{th} percentile
25 years, 100^{th} percentile
Just by examining the data, is the distribution, left- or right-skewed? How can you tell?
k) Sketch a smooth curve depicting a rough left-skewed distribution. Determine the relative positions of Mean, Median and Mode.
l) For a strongly skewed distribution, which measure – Mean or Median – **should** one employ? **Tip!** Which one shall *not* easily be affected by extreme values / outliers?
m) Sketch a smooth curve depicting a rough right-skewed distribution. Determine the relative positions of Mean, Median and Mode.

Solution.

a) Income distribution is Right-skewed since most values are on the lower end, in general → to impress, report the Mean since it'd be higher.
b) Right-skewed → most-values would be low and a few high, so Mean shall be higher as it gets

pulled in the direction of the extreme values at the higher end.
c) Upper-end.
d) Most scores shall be on the lower end → Right-skewed → Median shall be lower since Mean shall be higher as it gets pulled in the direction of the extreme values at the higher end.
e) Mean since it is affected by extreme values. [Median isnt since it is only based on the value in the middle ***Position***...and **not** on the actual values!]
f) Right-skewed.
g) Since most individuals are very old – 75% of chaps are $\geq$ 60 years – and only 25% of values are 0-60years, it's a left-skewed distribution. Alternately, calculating the lengths of whiskers: left whisker, Q1 – Min = 60 >> right whisker, Max – Q3 = 20years.
h) Property values behave like Incomes: most values shall be low, few high → Right-skewed → Mean would be higher.
i) Most values shall be high, a few lower → Left-skewed → Mean would be lower than the Median.
j) Since most individuals are very young and few are old – 75% of chaps are $\leq$ 6 years and only 25% of values are 6-25 years – it's a right-skewed distribution.
Alternately, calculating the lengths of whiskers: left whisker, Q1 – Min = 3 << right whisker, Max – Q3 = 19years.
k) Mean < Median < Mode, usually.
l) Median, since Mean would get affected by outliers.
m) Mean > Median > Mode, usually.