OCR Statistics 4 Module Revision Sheet

The S4 exam is 1 hour 30 minutes long. You are allowed a graphics calculator.

Before you go into the exam make sure you are fully aware of the contents of the formula booklet you receive. Also be sure not to panic; it is not uncommon to get stuck on a question (I've been there!). Just continue with what you can do and return at the end to the question(s) you have found hard. If you have time check all your work, especially the first question you attempted... always an area prone to error.

$\mathscr{J}.\mathscr{M}.\mathscr{S}.$

Preliminaries

- Your pure maths needs to be far stronger for S4 than in any other Statistics module.
- You must be strong on general binomial expansion from C4.

  $$(1+x)^n = 1 + nx + \frac{n(n-1)}{2}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \frac{n(n-1)(n-2)(n-3)}{2}x^4 + \cdots$$

  This is valid only for $|x| < 1$. This is important for probability/moment generating functions.
- In particular you must be good at 'plucking out' specific coefficients (which may represent probabilities). For example find the x^8 coefficient in $\frac{x(3+x^2)}{\sqrt{4+2x}}$.

  $$\begin{aligned}\frac{x(3+x^2)}{\sqrt{4+2x}} &= (3x+x^3)(4+2x)^{-\frac{1}{2}} \\ &= (3x+x^3)\left(4(1+\frac{x}{2})\right)^{-\frac{1}{2}} \\ &= \frac{1}{2}(3x+x^3)\left(1+\frac{x}{2}\right)^{-\frac{1}{2}} \\ &= \frac{1}{2}(3x+x^3)\left(1-\frac{x}{4}+\cdots-\frac{63}{8192}x^5+\cdots-\frac{429}{262,144}x^7+\ldots\right)\end{aligned}$$

  So the x^8 coefficient will be $-\frac{1}{2}(3\times\frac{429}{262,144}+1\times\frac{63}{8192}) = -\frac{3303}{524,288}$. It helps hugely to be thinking ahead about what coefficients you are going to need.
- Recall from S3 that $\mathbb{E}(g(X)) = \sum g(x_i)p_i$ for discrete random variables and $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx$ for continuous random variables.
- Recall also that $\mathrm{Var}(X) \equiv \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

Probability

- There are three very useful ways of representing information in probability questions. Venn diagrams, tree diagrams and two-way tables. You must think hard about which approach is going to be most helpful in the question you are to answer. Read the whole question before you start!
- Set theory is very important in probability. Know the following

- '$A \cap B$' is the intersection of the sets A and B. The overlap between the two sets. "AND"
- '$A \cup B$' is the union of the sets A and B. Anything that lies in either A or B (or both). "OR"
- A' means 'not A'. Everything outside A.
- $\{\,\}$ (or $\emptyset$) denotes the empty set. For example $A \cap A' = \{\,\}$

- Events A and B are *mutually exclusive* if both A and B cannot both happen. Represented by a Venn diagram of non-overlapping circles. Here

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

- However in the general case where A and B are not mutually exclusive we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

  This is because we are overcounting the overlap. It is called the *addition law.*

  For three events the addition law becomes A, B and C we have (in general)

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

  Again this drops out easily from a Venn diagram.

- Events $A_1, A_2, \ldots$ are said to be exhaustive if $\mathbb{P}(A_1 \cup A_2 \cup \ldots) = 1$. In other words the events $A_1, A_2, \ldots$ contain all the possibilities.

- If A and B are *independent* events then

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B).$$

- We read $\mathbb{P}(A|B)$ as the probability of A given that B has occured. It is defined

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

  However this formula is not always easy to apply, so Mr Stone's patented 'collapsing universes' approach from a Venn or tree diagram is often more intuitive.

- Using $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ and $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$ we discover

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) = \mathbb{P}(B)\mathbb{P}(A|B).$$

  This is called the *multiplication law* of probability and is incredibly useful in converting $\mathbb{P}(A|B)$ into $\mathbb{P}(B|A)$ and vice versa. The multiplication law drops out readily from a tree diagram.

- Bayes' Theorem[1] states

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j)\mathbb{P}(B|A_j)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A_j)\mathbb{P}(B|A_j)}{\sum_{\text{all } i} \mathbb{P}(A_i)\mathbb{P}(B|A_i)}.$$

  This looks scary, but drops out from a tree diagram. The formal statement is not required for S4, but is very important.

[1] Reverend Thomas Bayes from my home town of Tunbridge Wells. Wrote a document defending Newton's calculus hence a rather good bloke.

Non-Parametric Tests

- All of the hypothesis tests studied in Stats 2 & 3 required knowledge (or at the very least an assumption) of some kind of underlying distribution for you to carry out the test. However sometimes you have no knowledge about the underlying population. Statisticians therefore developed a series of *non-parametric* tests for situations where you have no knowledge of the underlying population.

- The *sign test* is a test about the *median* (i.e. the point at which you have an equal number of data points either side). If H_0 : median = 10, say, then under H_0, whether a random piece of data lies above or below 10 has probability $\frac{1}{2}$. For n pieces of data we therefore have a binomial $B(n, \frac{1}{2})$. Rather than work out critical values, the best approach is probably to calculate (under H_0) the probability of what you have observed and anything more extreme. For example test at the 5% level whether the median of the data

  $$1,\ 1,\ 2,\ 3,\ 6,\ 7,\ 8,\ 9,\ 9,\ 9,\ 10,\ 10,\ 11,\ 13$$

  is 5. Note that there are four pieces of data less than 5.

  H_0 : The median of the data is 5.

  H_1 : The median of the data is not 5.

  $\alpha = 5\%$. Two tailed test.

  Under H_0, $X \sim B(14, \frac{1}{2})$.

  $\mathbb{P}(X \leqslant 4) = 0.0898 > 0.025$, so at the 5% level there is insufficient evidence to reject H_0 and we conclude that the median of the data is probably 5. [You could have also gone through the rigmarole of demonstrating that the critical value is 2 (or 12) but my way is quicker and life's short.]

- Although there is no example in your textbook I see no reason why they couldn't ask a question where you had a large enough sample to require the normal approximation to $B(n, \frac{1}{2})$... don't forget your continuity correction.

- The sign test is a very crude test because it takes absolutely no account of how far away the data lies on either side of the median. If you want to take account of the magnitude of the deviations you need to use...

- ... the *Wilcoxon signed-rank test*. Here it is assumed that the data is *symmetric*; therefore it is a test about both the median or the mean because for symmetric data the median and mean are the same.

  You calculate the deviations from the median/mean, rank the size of the deviations and then sum the positive ranks to get P and sum the negative ranks to get Q. The test statistic is T, where T is the smaller of P or Q. For example test at the 5% level whether the mean of

  $$1.3,\ 2.1,\ 7.3,\ 4.9,\ 3.2,\ 1.6,\ 5.6,\ 5.7$$

  is 3.

  The data sort of looks symmetric, so OK to proceed with Wilcoxon.

  H_0 : The mean of the data is 3.

  H_1 : The mean of the data is greater than 3.

  $\alpha = 5\%$. One tailed test.

Data	1.3	2.1	7.3	4.9	3.2	1.6	5.6	5.7
Deviation	-1.7	-0.9	$+4.3$	$+1.9$	$+0.2$	-1.4	$+2.6$	$+2.7$
Rank	4	2	8	5	1	3	6	7
Signed Rank	-4	-2	$+8$	$+5$	$+1$	-3	$+6$	$+7$

So $P = 27$, $Q = 9$, so $T_{\text{obs}} = 9$. The lower T is, the worse it is for H_0 and the tables give the *largest* value at which you would reject H_0. $T_{\text{crit}} = 5$. $9 > 5$, so at the 5% level we have insufficient evidence to reject H_0 and conclude that the mean is probably 3.

- For large samples (i.e. when the tables don't give the values you want; running out of values) a normal approximation can be used where

$$Z = \frac{T + 0.5 - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}.$$

Note that because T is the smaller of P and Q that Z will always be negative (both Z_{crit} and Z_{obs}). For example if you had 100 pieces of data and you were testing at the 1% level whether the mean was some value (against H_1 of the mean not being some value) and $P = 2000$ and $Q = 3050$ then $T = 2000$. So

$$\begin{aligned} Z_{\text{obs}} &= \frac{T_{\text{obs}} + 0.5 - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \\ &= \frac{2000 + 0.5 - \frac{1}{4} \times 100 \times 101}{\sqrt{\frac{1}{24} \times 100 \times 101 \times 201}} \\ &= -1.803 \end{aligned}$$

Because it is a two-tailed 1% test we reverse look-up 0.995 to obtain $Z_{\text{crit}} = -2.576$. Finally $-1.803 > -2.576$, so at the 1% level there is insufficient evidence to reject H_0 and conclude that the mean is probably whatever we thought it was under H_0.

- The *Wilcoxon rank-sum test* is the non-parametric equivalent of the two-sample t-test from S3. It tests whether two different sets of data are drawn from identical populations. The central idea for the theory is that if X and Y are drawn from identical distributions, then $P(X < Y) = \frac{1}{2}$. The tables are then constructed from tedious consideration of all the possible arrangements of the ranks (called the 'sampling distribution').

  Given two sets of data, let m be the number of pieces of data from the smaller data set and n be the number of pieces of data from the larger data set (if they are both the same size it's up to you which is m and which n). Then rank *all* the data and sum the ranks of the 'm' population; call this total R_m. Also calculate $m(n+m+1) - R_m$ and let the test statistic W be the smaller of R_m and $m(n+m+1) - R_m$. The smaller W is, the more likely we are to reject H_0 and the tables give the largest W at which we reject H_0.

  For example test at the 5% level whether the following are drawn from identical populations.

A	23	14	42	12	30	40
B	16	21	9	35		

  H_0 : Data drawn from identical distributions.

  H_1 : Data not drawn from identical distributions.

  $\alpha = 5\%$. Two tailed test.

Data	9	12	14	16	21	23	30	35	40	42
Rank	1	2	3	4	5	6	7	8	9	10

So $m = 4$, $n = 6$, $R_m = 18$, $m(n + m + 1) - R_m = 26$, $W_{\text{obs}} = 18$. Looking at the tables we see $W_{\text{crit}} = 12$, and $18 > 12$, so at the 5% level there is insufficient evidence to reject H_0 and we conclude that the data is probably drawn from identical distributions.

- For large samples (i.e. when the tables don't give the values you want; running out of values) a normal approximation can be used where

$$Z = \frac{W + 0.5 - \frac{1}{2}m(m+n+1)}{\sqrt{\frac{1}{12}mn(m+n+1)}}.$$

Probability Generating Functions

- In Stats 1 & 2 you met discrete random variables (DRVs) such that each outcome had a probability attached. Sometimes there were rules which related the probability to the outcome (binomial, geometric, Poisson). However, in general we had:

x	x_1	x_2	x_3	x_4	$\ldots$
$\mathbb{P}(X = x)$	p_1	p_2	p_3	p_4	$\ldots$

Recall that $\sum p_i = 1$ because the sum of all the probabilities must total 1 and that $\mathbb{E}(X) = \sum p_i x_i$. Also $\mathbb{E}(f(X)) = \sum p_i f(x_i)$ from Stats 3 and $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \sum p_i x_i^2 - (\sum p_i x_i)^2$ from Stats 2.

- At some point some bright spark decided to consider the properties of

$$G_X(t) = \mathbb{E}(t^X) = \sum p_i t^{x_i} = p_1 t^{x_1} + p_2 t^{x_2} + p_3 t^{x_3} + p_4 t^{x_4} + \cdots$$

where t is a 'dummy variable' unrelated to x. You can see that this will create either a finite or infinite series. This is called the probability generating function of X. It is a single function that contains within it all of the (potentially infinite) probabilities of X.

For example given

x	-2	-1	0	1	2
$\mathbb{P}(X = x)$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{8}$	$\frac{1}{8}$

the generating function is $G_X(t) = p_1 t^{x_1} + p_2 t^{x_2} + p_3 t^{x_3} + p_4 t^{x_4} + \cdots = \frac{1}{6}t^{-2} + \frac{1}{4}t^{-1} + \frac{1}{3} + \frac{1}{8}t + \frac{1}{8}t^2$. We can therefore see that if (say) we saw a term $\frac{5}{24}t^6$, then we can see that $\mathbb{P}(X = 6) = \frac{5}{24}$. Note that if you see a constant term then that tells you $\mathbb{P}(X = 0)$ because $t^0 = 1$.

- An important property is that $G_X(1) = 1$ because $G_X(1)$ is just the sum of all the probabilities of X, i.e. $\sum p_i$.

- Another useful thing to do is consider the derivative $G'_X(t)$ with respect to t;

$$G'_X(t) = \sum p_i x_i t^{x_i - 1} = p_1 x_1 t^{x_1 - 1} + p_2 x_2 t^{x_2 - 1} + p_3 x_3 t^{x_3 - 1} + \cdots$$

Again, if we consider $G'(1)$ we obtain

$$G'(1) = \sum p_i x_i = p_1 x_1 + p_2 x_2 + p_2 x_2 + p_3 x_3 + \cdots = \mathbb{E}(X).$$

- Variances can also be calculated by

$$\mathrm{Var}(X) = G_X''(1) + G_X'(1) - (G_X'(1))^2.$$

- Some standard pgfs are given in the formula book:

Distribution	$B(n,p)$	$Po(\lambda)$	$Geo(p)$
pgf	$(1-p+pt)^n$	$e^{\lambda(t-1)}$	$\frac{pt}{1-(1-p)t}$

  Any good candidate should be able to derive these...

- For two *independent* random variables X and Y (with pgfs $G_X(t)$ and $G_Y(t)$ respectively) the pgf of $X+Y$ is $G_{X+Y}(t) = G_X(t) \times G_Y(t)$. This extends to three or more *independent* random variables.

Moment Generating Functions

- You will recall from FP2 that the Maclaurin expansion for e^x is

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots .$$

  This is valid for all values of x (and you should know why from your Pure teachings). An alternative notation used is $e^x \equiv \exp(x)$.

- The nth moment of a distribution is $\mathbb{E}(X^n)$. So the first moment is just $\mathbb{E}(X)$. The second moment is $\mathbb{E}(X^2)$, which is useful in calculating variances. The zeroth moment is $\mathbb{E}(X^0) = \mathbb{E}(1) = 1$.

- The moment generating function (mgf) is defined for

x	x_1	x_2	x_3	x_4	...
$\mathbb{P}(X=x)$	p_1	p_2	p_3	p_4	...

  by

$$\begin{aligned}
M_X(t) = \mathbb{E}(e^{tX}) = \sum p_i e^{x_i t} &= p_1 e^{x_1 t} + p_2 e^{x_2 t} + p_3 e^{x_3 t} + p_4 e^{x_4 t} + \cdots \\
&= \quad p_1 + p_1 x_1 t + p_1 \frac{x_1^2 t^2}{2!} + p_1 \frac{x_1^2 t^3}{3!} + \cdots \\
&\quad + p_2 + p_2 x_2 t + p_2 \frac{x_2^2 t^2}{2!} + p_2 \frac{x_2^2 t^3}{3!} + \cdots \\
&\quad + p_3 + p_3 x_3 t + p_3 \frac{x_3^2 t^2}{2!} + p_3 \frac{x_3^2 t^3}{3!} + \cdots \\
&\quad + p_4 + p_4 x_2 t + p_4 \frac{x_4^2 t^2}{2!} + p_4 \frac{x_4^2 t^3}{3!} + \cdots \\
&= \quad (p_1 + p_2 + p_3 + p_4 + \cdots) \\
&\quad + (p_1 x_1 + p_2 x_2 + p_3 x_3 + p_4 x_4 + \cdots)\, t \\
&\quad + \left(p_1 x_1^2 + p_2 x_2^2 + p_3 x_3^2 + p_4 x_4^2 + \cdots\right) \frac{t^2}{2!} \\
&\quad + \left(p_1 x_1^3 + p_2 x_2^3 + p_3 x_3^3 + p_4 x_4^3 + \cdots\right) \frac{t^3}{3!} \\
&\quad + \cdots \\
&= \mathbb{E}(1) + \mathbb{E}(X)t + \mathbb{E}\left(X^2\right) \frac{t^2}{2!} + \mathbb{E}\left(X^3\right) \frac{t^3}{3!} + E\left(X^4\right) \frac{t^4}{4!} + \cdots
\end{aligned}$$

So you can see that the constant term of $M_X(t)$ should always be $\mathbb{E}(1) = 1$ because it represents the sum of the probabilities. The coefficient of t will be $\mathbb{E}(X)$ and the coefficient of $\frac{t^2}{2!}$ (*not* just the coefficient of t^2) will be $\mathbb{E}(X^2)$. In general the coefficient of $\frac{t^n}{n!}$ will be $\mathbb{E}(X^n)$, that is, the nth moment.

- As with pgfs, differentiating mgfs (with respect to t) is a 'good thing'. However, instead of letting $t = 1$ we let $t = 0$ (because $a^0 = 1$). So differentiating $M_X(t)$ we find:

$$\begin{aligned} M_X(t) &= p_1e^{x_1t} + p_2e^{x_2t} + p_3e^{x_3t} + p_4e^{x_4t} + \cdots \\ M'_X(t) &= p_1x_1e^{x_1t} + p_2x_2e^{x_2t} + p_3x_3e^{x_3t} + p_4x_4e^{x_4t} + \cdots \\ M'_X(0) &= p_1x_1 + p_2x_2 + p_3x_3 + p_4x_4 + \cdots \\ &= \sum x_ip_i = \mathbb{E}(X). \end{aligned}$$

So $M'_X(0) = \mathbb{E}(X)$.

Differentiating again we find:

$$\begin{aligned} M'_X(t) &= p_1x_1e^{x_1t} + p_2x_2e^{x_2t} + p_3x_3e^{x_3t} + p_4x_4e^{x_4t} + \cdots \\ M''_X(t) &= p_1x_1^2e^{x_1t} + p_2x_2^2e^{x_2t} + p_3x_3^2e^{x_3t} + p_4x_4^2e^{x_4t} + \cdots \\ M''_X(0) &= p_1x_1^2 + p_2x_2^2 + p_3x_3^2 + p_4x_4^2 + \cdots \\ &= \sum x_i^2p_i = \mathbb{E}(X^2). \end{aligned}$$

So using $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ we find $\mathrm{Var}(X) = M''_X(0) - (M'_X(0))^2$.

- Notice that with mgfs there are two ways to obtain the expectation and variance of your random variable. All things being equal I would choose the differentiation method, but you must ensure that your mgf is defined for $t = 0$. Also read the question carefully to see what they are wanting.

- Moment generating functions can also be defined for continuous random variables:

$$M_X(t) = \int_{-\infty}^{\infty} f(x)e^{tx}\,dx.$$

As before $M_X(0) = 1$, $M'_X(0) = \mathbb{E}(X)$, $M''_X(0) = \mathbb{E}(X^2)$. Convergence issues can arise EXAMPLE!!!!!!

- Some standard mgfs are given in the formula book:

Distribution	Uniform on $[a, b]$	Exponential	$N(\mu, \sigma^2)$
mgf	$\frac{e^{bt}-e^{at}}{(b-a)t}$	$\frac{\lambda}{\lambda-t}$	$e^{\mu t+\frac{1}{2}\sigma^2t^2}$

Any good candidate should be able to derive these too...

- As with pgfs, for two *independent* random variables X and Y (with mgfs $G_X(t)$ and $G_Y(t)$ respectively) the mgf of $X + Y$ is $M_{X+Y}(t) = M_X(t) \times M_Y(t)$. This extends to three or more *independent* random variables.

Estimators

- It is vital to recall here that $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$ (by definition).

- Given a population there may be many parameters that we may wish to know. For example we might like to know the mean μ, the variance σ^2, the median M, the maximum or minimum, the IQR, etc. In general we shall call this parameter θ.

  Usually we will never know θ because we won't have the whole population. But we will be able to take a random sample from the population. From this sample we can calculate a quantity U which we shall use to estimate θ. We call U an estimator of θ.

- U is said to be an *unbiased estimator* of θ if

$$\mathbb{E}(U) = \theta.$$

  i.e. if we take an average of *all possible* U (remember that U is a random variable) we will get the desired θ. If $\mathbb{E}(U) \neq \theta$ then the estimator is said to be biased (not giving the desired result on average).

  For example to show that $K = \frac{X_1+2X_2+5X_3}{8}$ is an unbiased estimator of μ we merely consider $\mathbb{E}(K)$ and keep whittling down as far as we can go (using S3 expectation and variance algebra)

$$\begin{aligned}
\mathbb{E}(K) &= \mathbb{E}\left(\frac{X_1 + 2X_2 + 5X_3}{8}\right) \\
&= \mathbb{E}\left(\frac{X_1}{8} + \frac{X_2}{4} + \frac{5X_3}{8}\right) \\
&= \frac{1}{8}\mathbb{E}(X_1) + \frac{1}{4}\mathbb{E}(X_2) + \frac{5}{8}\mathbb{E}(X_3) \\
&= \frac{1}{8}\mu + \frac{1}{4}\mu + \frac{5}{8}\mu = \mu.
\end{aligned}$$

- For continuous random variables just remember that $\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)\,dx$. For example find the value of k which makes $L = k(X_1 + X_2)$ an unbiased estimator of θ for

$$f(x) = \begin{cases} \frac{2}{\theta}\left(1 - \frac{x}{\theta}\right) & 0 \leqslant x \leqslant \theta \\ 0 & \text{otherwise} \end{cases}$$

  First calculate $\mathbb{E}(X)$ *à la* S2:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_0^{\theta} x\,\frac{2}{\theta}\left(1 - \frac{x}{\theta}\right)\,dx = \left[\frac{x^2}{\theta} - \frac{2x^3}{3\theta^2}\right]_0^{\theta} = \frac{\theta}{3}.$$

  So for L to be unbiased we need $\mathbb{E}(L) = \theta$, so

$$\begin{aligned}
\mathbb{E}(L) &= \theta \\
\mathbb{E}(k(X_1 + X_2) &= \theta \\
k(\mathbb{E}(X_1) + \mathbb{E}(X_2)) &= \theta \\
k(2 \times \mathbb{E}(X)) &= \theta \\
k\left(\frac{2\theta}{3}\right) &= \theta \\
k &= \frac{3}{2}.
\end{aligned}$$

- Given two *unbiased* estimators the *most efficient* estimator (of the two) is the one where $\mathrm{Var}(U)$ is smaller. A smaller variance is a 'good thing'.

- Sometimes you may need calculus to work out the most efficient estimator from an infinite family. For example X_1, X_2 and X_3 are three independent measurements of X.

$$S = \frac{aX_1 + 2X_2 + 4X_3}{a+6} \quad (\text{with } a \neq -6)$$

is suggested as an estimator for μ. Prove that S is unbiased whatever the value of a and find the value of a which makes S most efficient. So

$$\begin{aligned}
\mathbb{E}(S) &= \mathbb{E}\left(\frac{aX_1 + 2X_2 + 4X_3}{a+6}\right) \\
&= \frac{1}{a+6}\mathbb{E}(aX_1 + 2X_2 + 4X_3) \\
&= \frac{1}{a+6}[a\mathbb{E}(X_1) + 2\mathbb{E}(X_2) + 4\mathbb{E}(X_3)] \\
&= \frac{1}{a+6}[a\mu + 2\mu + 4\mu] \\
&= \frac{\mu}{a+6}(a+6) = \mu.
\end{aligned}$$

So S is unbiased for all values of a. Now consider

$$\begin{aligned}
\text{Var}(S) &= \text{Var}\left(\frac{aX_1 + 2X_2 + 4X_3}{a+6}\right) \\
&= \frac{1}{(a+6)^2}\text{Var}(aX_1 + 2X_2 + 4X_3) \\
&= \frac{1}{(a+6)^2}[a^2\text{Var}(X_1) + 4\text{Var}(X_2) + 16\text{Var}(X_3)] \\
&= \frac{a^2+20}{(a+6)^2}\sigma^2.
\end{aligned}$$

To minimise $\text{Var}(S)$ we need $\frac{d}{da}\text{Var}(S) = 0$. So

$$\begin{aligned}
0 = \frac{d}{da}\left(\frac{a^2+20}{(a+6)^2}\sigma^2\right) &= \frac{2a(a+6)^2 - 2(a+6)(a^2+20)}{(a+6)^4}\sigma^2 \\
\text{So } 0 &= 2a(a+6)^2 - 2(a+6)(a^2+20) \\
0 &= 2(a+6)[a(a+6) - (a^2+20)] \\
0 &= (a+6)(6a-20).
\end{aligned}$$

So $a = -6$ or $a = \frac{10}{3}$, but $a \neq -6$ so $a = \frac{10}{3}$ is the value of a that makes S most efficient[2].

- Here is a tough type of problem that caught me out the first two (or three (or four $(\ldots)$)) times I saw it. Slot away the method just in case. For example consider

$$f(x) = \begin{cases} \frac{2x}{\theta^2} & 0 \leqslant x \leqslant \theta \\ 0 & \text{otherwise} \end{cases}$$

An estimate of θ is required and a suggestion is made to calculate $\frac{5L}{4}$ where L is the maximum of two independent observations of X (X_1 and X_2). Show that this estimator is unbiased.

The thing to remember is that for L to be the maximum of X_1 and X_2, then X_1 *and* X_2 must *both* be less than or equal to L; i.e. we are going to calculate a cdf. So

$$\mathbb{P}(L \leqslant l) = \mathbb{P}(X_1 \leqslant l) \times \mathbb{P}(X_2 \leqslant l).$$

[2] I suppose we should consider the second derivative to show that this value of a minimises rather than maximises the variance, but life's too short...

(This can be extended to three or more independent samplings of X.)

By sketching $f(x)$ we can see that the probability that one observation is less than or equal to l is given by a triangle in this case of area $\frac{l^2}{\theta^2}$ (or by the integral $\int_0^l f(x)\,dx$ for a more general $f(x)$). So $\mathbb{P}(L \leqslant l) = \mathbb{P}(X_1 \leqslant l) \times \mathbb{P}(X_2 \leqslant l) = \frac{l^2}{\theta^2} \times \frac{l^2}{\theta^2} = \frac{l^4}{\theta^4}$. Differentiating wrt to l we find the pdf of l to be

$$f(l) = \begin{cases} \frac{4l^3}{\theta^4} & 0 \leqslant l \leqslant \theta \\ 0 & \text{otherwise} \end{cases}$$

Therefore we calculate $\mathbb{E}\left(\frac{5L}{4}\right)$ as follows:

$$\begin{aligned} \mathbb{E}\left(\frac{5L}{4}\right) &= \frac{5}{4}\mathbb{E}(L) \\ &= \frac{5}{4}\int_0^\theta l \times \frac{4l^3}{\theta^4}\,dl \\ &= \frac{5}{4}\left[\frac{4l^5}{5\theta^4}\right]_0^\theta = \theta. \end{aligned}$$

Therefore $\frac{5L}{4}$ is an unbiased estimator of θ. I will leave it as an exercise for the reader to demonstrate that $\text{Var}\left(\frac{5L}{4}\right) = \frac{\theta^2}{24}$.

Discrete Bivariate Distributions

- The discrete random variables you have met thus far have been in one variable only. For example

x	2	3	5	7
$\mathbb{P}(X = x)$	$\frac{1}{2}$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{18}$

However we can have discrete *bivariate* distributions. For example

		X		
		2	3	5
Y	4	0	$\frac{1}{2}$	$\frac{1}{10}$
	5	$\frac{1}{5}$	$\frac{3}{20}$	$\frac{1}{20}$

From this we can see, say, $\mathbb{P}(X = 3, Y = 5) = \frac{3}{20}$.

- The marginal distribution is what one obtains if one of the variables is 'ignored'. In the above example the marginal distribution of X can be written

x	2	3	5
$\mathbb{P}(X = x)$	$\frac{1}{5}$	$\frac{13}{20}$	$\frac{3}{20}$

This can be added to the bivariate distribution thus:

		X		
		2	3	5
Y	4	0	$\frac{1}{2}$	$\frac{1}{10}$
	5	$\frac{1}{5}$	$\frac{3}{20}$	$\frac{1}{20}$
		$\frac{1}{5}$	$\frac{13}{20}$	$\frac{3}{20}$

$\mathbb{E}(X)$ and $\mathrm{Var}(X)$ can be calculated in the usual way obtaining $\mathbb{E}(X) = \frac{31}{10}$ and $\mathrm{Var}(X) = \frac{79}{100}$ (do it!). Similarly you can work out the marginal distribution of Y if you are so inclined.

- The *conditional* distribution of a bivariate distribution can be calculated *given that* one of the variables (X or Y) has taken a specific value. For the above example the "distribution of X conditional on $Y = 4$" is calculated by rewriting the 4 row with all the values divided by $\mathbb{P}(Y = 4) = \frac{3}{5}$.

x	2	3	5
$\mathbb{P}(X = x \mid Y = 4)$	0	$\frac{5}{6}$	$\frac{1}{6}$

This is all from our friend $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

- A way to check whether X and Y are *independent* of each other in a bivariate distribution is to check whether every entry in the distribution is the product of the two relevant marginal probabilities. For example

		X			
		1	2	3	
	1	$\frac{1}{3}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{2}{3}$
Y	2	$\frac{1}{6}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{3}$
		$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	

Here we see $\mathbb{P}(X = 2, Y = 1) = \frac{2}{9}$ is the same as $\mathbb{P}(X = 2) \times \mathbb{P}(Y = 1) = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}$. The same is true for *every* entry in the table, so X and Y are independent. It only takes one entry not to satisfy this to ensure X and Y are *not* independent.

- The *covariance* of a discrete bivariate distribution is defined

$$\mathrm{Cov}(X, Y) \equiv \mathbb{E}((X - \mu_X)(Y - \mu_Y)).$$

However this tends to be cumbersome to calculate so we use the equivalent formula

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mu_X \mu_Y.$$

The covariance can be thought of as the correlation coefficient (r from Stats 1) for two probability distributions (sort of). The covariance can be both positive or negative (like the correlation coefficient).

- To calculate the covariance, first create the marginal distributions:

		X		
		1	3	4
	2	$\frac{1}{3}$	$\frac{1}{4}$	0
Y	5	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{8}$

$\Rightarrow$

		X			
		1	3	4	
	2	$\frac{1}{3}$	$\frac{1}{4}$	0	$\frac{7}{12}$
Y	5	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{5}{12}$
		$\frac{1}{2}$	$\frac{3}{8}$	$\frac{1}{8}$	

Then use the marginal distributions to calculate μ_X and μ_Y.

$\mu_X = \mathbb{E}(X) = \sum xp = 1 \times \frac{1}{2} + 3 \times \frac{3}{8} + 4 \times \frac{1}{8} = \frac{17}{8}$.

$\mu_Y = \mathbb{E}(Y) = \sum yp = 2 \times \frac{7}{12} + 5 \times \frac{5}{12} = \frac{13}{4}$.

Now we use this to calculate the covariance thus:

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}(XY) - \mu_X \mu_Y \\ &= (1 \times 2 \times \tfrac{1}{3}) + (1 \times 5 \times \tfrac{1}{6}) + (3 \times 2 \times \tfrac{1}{4}) + (3 \times 5 \times \tfrac{1}{8}) + (4 \times 2 \times 0) + (4 \times 5 \times \tfrac{1}{8}) - \tfrac{17}{8} \times \tfrac{13}{4} \\ &= \frac{15}{32}.\end{aligned}$$

- If X and Y are independent then $\text{Cov}(X, Y) = 0$. However, if $\text{Cov}(X, Y) = 0$ this does not *necessarily* mean that X and Y are independent. But if $\text{Cov}(X, Y) \neq 0$ then X and Y cannot be independent.

- With an understanding of covariance we can write the relationship for $\text{Var}(aX \pm bY)$ when X and Y are *not* independent:

$$\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) \pm 2ab\text{Cov}(X, Y).$$

  Notice the extra term at the end of the formula we are used to from S3 for *independent* X and Y.