Statistics 1

Edexcel IAL S1 (WST01/01); data, probability, regression, discrete random variables and the Normal

Formulae to memorise

Family	**Must know**
means	$\overline{x} = \frac{\sum x}{n}$ or $\frac{\sum fx}{\sum f}$
probability	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
conditional	$P(A \cap B) = P(A)P(B \mid A)$
coding	$E(aX + b) = aE(X) + b$, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
Normal	$Z = \frac{X-\mu}{\sigma}$

Why use a model

> Quicker and cheaper, repeatable, and it exposes relationships.

- A simplified description of reality.
- "To make predictions" alone is not enough.

Histograms: area is frequency

Second bar: $\frac{10.8}{18} \times 10 = 6$ cm

- Height is frequency density.
- Compare bars by ratio.

Worked: combining two classes

1. **Total** $\sum x = 54(45) = 2430$
2. **Class B** $\sum x_B = 2430 - 24(47) = 1302$: $b = 43.4$
3. **Combined SD** $\sqrt{\frac{66876+73826}{54} - 45^2} = 24.1$

Decision

B up by 2: B's **variance** unchanged, combined SD falls.

Worked: quartiles by interpolation

1. **Median** 112.5th, in 25–30: $25 + 5 \times \frac{112.5-54}{90} = 28.25$
2. **Upper quartile** 168.75th, in 30–40: $30 + 10 \times \frac{168.75-144}{60} = 34.125$
3. **IQR** $34.125 - 25.14$

Decision

$IQR = 8.985$

Reading skew from the box

Positive skew: a Normal model would be refused.

- Compare the two halves of the box.
- Or mean against **median**.

Fill the Venn diagram

$P(C \cap D) = 0.09,\ P(C) = 0.23$

- $P(C \cup D) = 0.59,\ P(D) = 0.45$
- $P(C \mid D) = 0.2$

A tree without replacement

Exactly one yellow: $\frac{18}{72} + \frac{18}{72} = \frac{1}{2}$

- 3 white, 6 yellow; two drawn.
- Second layer drops by one.

Worked: reversing the tree

1. **Condition** at least one white: $\frac{7}{12}$
2. **Both white** $\frac{3}{9} \cdot \frac{2}{8} = \frac{1}{12}$
3. **Divide** $\frac{1/12}{7/12}$

Decision

$P(\text{both white} \mid \text{at least one white}) = \frac{1}{7}$

Worked: coded regression

1. **Sww** $10036.45 - \frac{339.25^2}{12} = 445.57$
2. **r** $\frac{-91.55}{\sqrt{26.43 \times 445.57}} = -0.844$
3. **w on t** $b = \frac{-91.55}{26.43} = -3.46$, $w = -3.46t + 30.9$

Decision

Divide by 11.5: $d = -0.301t + 2.69$

What a regression line can do

Can

predict y from x

within the data range

Cannot

predict x from y

extrapolate safely

Worked: the discrete routine

1. **Inequality** $3X - 3 < X + 2$: $X < 2.5$, so $P = 0.5$
2. **Moments** $E(X) = 2.35$, $E(X^2) = 6.65$
3. **Variance** $6.65 - 2.35^2 = 1.1275$

Decision

$\mathrm{Var}(5 - 2X) = 4(1.1275) = 4.51$

The Normal tail

$$P(X > 508) = P(Z > 2) = 0.0228$$

- Standardise, then read the table.
- Symmetry does half the work.

Worked: two conditions, one k

1. **Find** σ $z_{0.85} = 1.0364$: $\sigma = \frac{14.51}{1.0364} = 14$
2. **Equal tails** $P(X > k) = P(Y < 2k)$: $\frac{k-502}{3} = \frac{1024-2k}{14}$
3. **Solve** $14k - 7028 = 3072 - 6k$: $k = 505$

Decision

$p = P(Z > 1) = 0.159$

Check yourself

Question	**Answer**
Histogram bar height?	frequency density
$Q_3 - Q_2 > Q_2 - Q_1$?	positive skew
$P(B \mid A)$?	$\frac{P(A \cap B)}{P(A)}$
$\mathrm{Var}(aX + b)$?	$a^2\mathrm{Var}(X)$
Regression of y on x predicts?	y from x, in range