THE LARGE DATA SET

KEY WORDS & DEFINITIONS

1. Daily Mean Temperature
The average of hourly temperature readings in a 24hour period, in Celsius.

2. Daily Total Rainfall
The depth of precipitation as a liquid. All precipitation is included, not just rainfall, but it is melted if necessary for the measurement. Heights less than 0.05mm are recorded as a "trace" or "tr".

3. Daily Total Sunshine
Recorded to the nearest 10th of an hour (6 minutes).

4. Daily Mean Wind Direction
Given as a bearing and/or in cardinal (compass) directions.

5. Daily Mean Windspeed
Averaged over 24 hours of a day (midnight to midnight), in Knots, nautical miles per hour where 1 Knot = 1.15mph. Can also be categorised by the Beaufort Scale.

6. Daily Maximum Gust
The highest instantaneous windspeed recorded, in Knots.

7. Daily Maximum Gust Direction
The direction of the maximum gust of wind recorded.

8. Daily Maximum Relative Humidity
A percentage of air saturation with water vapour. Relative humidities above 95% result in mist or fog.

9. Daily Mean Cloud Cover
Measured in eighths of the sky that is covered (Oktas).

10. Daily Mean Visibility
The greatest horizontal distance at which an object can be seen in daylight, measured in decametres (Dm).

11. Daily Mean Pressure
Measured in hectopascals (hPa)

WHAT DO I NEED TO KNOW?

1. What the Large Data Set is about
The Edexcel LDS has samples on weather data in different locations for certain time periods. The data is provided by the Met Office.
The LDS contains the weather data for 5 UK weather stations and 3 weather stations overseas.

2. How to clean the data
N/A should be removed before calculations
tr (trace) should be turned to 0

3. Locations
Learn maps and understand geographical significance of North, South, Coastal etc,

4. Dates
Remember the Large Data Set only has information from May—October 1987 and May—October 2015. Anything between November and April is outside the range of our data.

5. Understand OKTAS
A measure of the fraction of the celestial dome covered by cloud, measured in
eighths. 0 oktas represents a clear sky,
while a value of 8 indicates complete overcast.

6. How to convert units
1 knot = 1.151 mph

7. Limitations
These stations do not tell us about the whole UK

THE BEAUFORT SCALE

Beaufort Scale	Description	Av. Wind Speed 10m above ground
0	Calm	< 1 Knot
1-3	Light	1 – 10 Knots
4	Moderate	11 – 16 Knots
5	Fresh	17 – 21 Knots

UK DATA

Location (N to S)	Temp Range (°C)	Wind Speed Range (kn)
Leuchars	4 – 9	3 – 23
Leeming	4 – 23	3 – 17
Heathrow	8 - 29	3 – 19
Hurn	6 - 24	2 – 19
Camborne	10 - 20	3 – 18

@Lucyjc1612

DATA COLLECTION

KEY WORDS & DEFINITIONS

1. Population
Whole set of items that could be sampled.
2. Census
Observations taken from the entire population.
3. Sample
Observations taken from a subset of the population.
4. Sampling Unit
One individual observation set from the population.
5. Sampling Frame
A numbered (or named) list of individual sampling units.
6. Strata
A subset of the population.

TYPES OF SAMPLING

1. Simple Random Sampling
Every sample of a specified size has an equal chance of being selected from a sampling frame.
2. Systematic Sampling
Items are chosen at regular intervals from a sampling frame.
3. Stratified Sampling
Random samples are taken proportionally from mutually exclusive groups or strata.
4. Quota Sampling
Non-random sample is taken to fulfil predetermined quotas for different categories.
5. Opportunity Sampling
Non-random sample is selected from available sampling units.

TYPES OF DATA

1. Quantitative Data
Variables or data associated with a numerical value.
2. Qualitative Data
Variables or data associated with a non-numerical value.
3. Continuous
Variables that can take any value.. **Measured**.
4. Discrete
Variables that can only take specific values.. **Counted**.

CENSUS VS SAMPLE

	Census	Sample
Advantages	Includes every member of the population to give a fully representative set of data	Less time consuming to collect and process data. Fewer people needed therefore cheaper to conduct.
Disadvantages	Time consuming & expensive. Cannot be used when testing process destroys the item being tested	May not be fully representative of population. Outliers or whole subgroups possibly excluded

WHAT DO I NEED TO KNOW?

1. Advantages & Disadvantages
Why is one type of sampling more appropriate than another. Consider time, cost, bias, ease, accuracy of population representation.
2. How to work with Grouped Data
Understand inequalities. Find maximum, minimum & midpoint of each group.
3. How to use the Large Data Set
Be able to clean data, take samples and comment on findings.

MEASURES OF LOCATION & SPREAD

KEY WORDS & DEFINITIONS

1. **Measure of Location**
A single value which describes a position in a data set.
2. **Measure of Central Tendency**
A measure of location which describes the central position in a data set.
3. **Measure of Spread or Dispersion**
A value which describes how spread out the data is.
4. **Mean**
The sum of all the data divided by how many pieces of data there are. Includes all pieces of data. Affected by outliers.
5. **Median Q_2**
The middle value when the data values are put in order. Does not include all pieces of data. Not affected by outliers.
6. **Mode**
The value that occurs most often in the data. Good for non-numerical data.
7. **Modal class**
The class that has the highest frequency in grouped data.
8. **Lower Quartile Q_1**
A measure of location that is one quarter of the way through the data set.
9. **Upper Quartile Q_3**
A measure of location that is three-quarters of the way through the data set.
10. **Percentile**
A measure of location that is the specified percentage of the way through the data set.
11. **Range**
The difference between the largest and smallest values in a data set. Affected by outliers.
12. **Inter-quartile Range**
The difference between the upper and lower quartiles in a data set. $Q_3 - Q_1$
Not affected by outliers.

IMPORTANT FORMULAE

Mean: $\bar{x} = \frac{\Sigma x}{n}$

Mean from Frequency Table: $\bar{x} = \frac{\Sigma f x}{\Sigma f}$

Variance σ^2:

$$\frac{\sum (x - \bar{x})^2}{n} = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

Standard Deviation σ = $\sqrt{\text{Variance}}$

CODING

If data is coded using $y = \frac{x - a}{b}$

Mean of coded data = $\bar{y} = \frac{\bar{x} - a}{b}$

s.d. of coded data = $\sigma_y = \frac{\sigma_x}{b}$

To find mean & s.d. of original data use:

$$\bar{x} = b\bar{y} + a$$

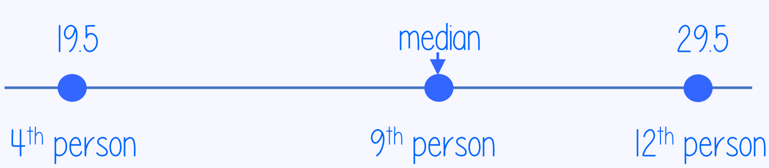
$$\sigma_x = b\sigma_y$$

INTERPOLATION

Assume data values are evenly distributed within each class then estimate median or percentile values using proportional reasoning.

Age	10 – 19	20 – 29	30 - 39
Frequency	4	8	5
Cumulative Freq	4	12	17

17 people ∴ median is 9th person
9th person is in 20 – 29 group
Take boundaries to be 19.5 & 29.5



$$\frac{m - 19.5}{29.5 - 19.5} = \frac{9 - 4}{12 - 4}$$

$$m = 25.75$$

REPRESENTATIONS OF DATA

KEY WORDS & DEFINITIONS

1. Outlier

A data value that lies beyond expected extremities. These are usually calculated as a multiple of the interquartile range above the upper quartile or below the lower quartile.
i.e. either greater than $Q_3 + k(Q_3 - Q_1)$
or less than $Q_1 - k(Q_3 - Q_1)$

2. Cleaning

The process of removing anomalies from the data set.

WHAT DO I NEED TO KNOW

Comparing 2 sets of data:

Calculate & compare the measures of location
Calculate & compare the measures of spread
Compare outliers if applicable
Mean & s.d go together
Median & IQR go together.
Ensure all comparisons are done IN CONTEXT

Histograms

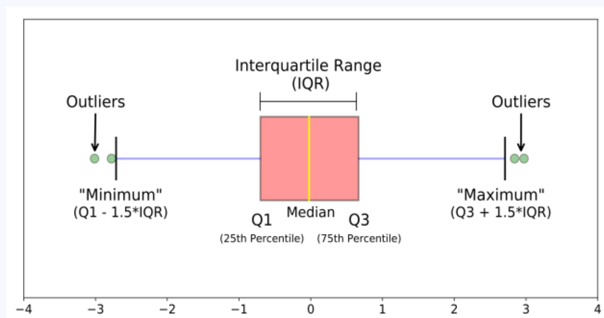
Area of bar $\propto$ Frequency so
Area of bar = k x Frequency
Area does NOT always = Frequency

BOX PLOTS

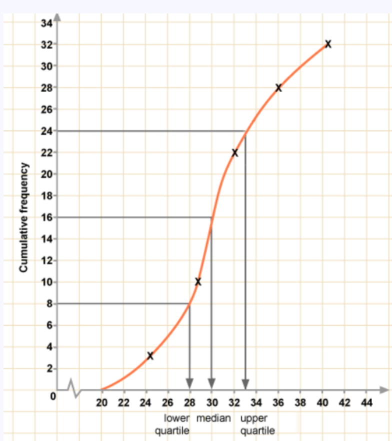
Box plots are rarely symmetrical

25% of the data lies within each section

Always use the same scale when comparing box plots



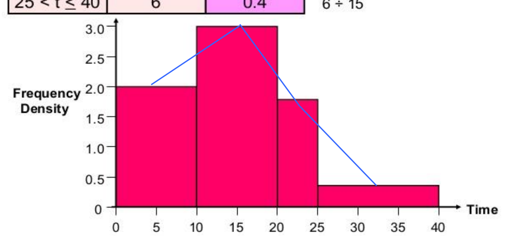
CUMULATIVE FREQUENCY



Plot points at the upper limits of group boundaries
Ensure it makes sense to extrapolate the curve at the beginning
Be careful of questions that ask "How many are **more** than..."

HISTOGRAMS

Time	Frequency	Frequency Density	Frequency Density = Frequency / Class width
$0 < t \leq 10$	20	2	20 ÷ 10
$10 < t \leq 15$	15	3	15 ÷ 5
$15 < t \leq 20$	10	2	10 ÷ 5
$20 < t \leq 25$	9	1.8	9 ÷ 5
$25 < t \leq 40$	6	0.4	6 ÷ 15



Histograms are used to represent grouped continuous data
Area of bar = k x frequency
If k = 1, then frequency density = $\frac{\text{frequency}}{\text{class width}}$

You may need to find the areas of parts of bars if questions don't use the class boundaries.
Joining the middle of the tops of each bar in a histogram forms a frequency polygon

@Lucyjc1612

CORRELATION & REGRESSION

KEY WORDS & DEFINITIONS

1. **Correlation** A description of the linear relationship between two variables.
2. **Bivariate data** Pairs of values for two variables
3. **Causal relationship** Where a change in a variable causes a change in another. Not always true.
4. **Least squares regression line**
A type of line of best fit which is a straight line in the form $y = a + bx$
5. **'b' of a regression line**
The gradient of the line; indicating positive correlation if it is positive and negative correlation if it is negative..
6. **Independent or Explanatory variable**
The variable which occurs regardless of the other variable (e.g. time passing). Plotted on the x axis.
7. **Dependent or Response variable**
The variable whose value depends on the independent variable's data points.
8. **Interpolation** Estimating a value within the range of the data. Reliable.
9. **Extrapolation** Estimating a value outside of the range of the data. NOT reliable.
10. **Product Moment Correlation Coefficient**
A measure of the strength and type of correlation.

WHAT DO I NEED TO KNOW

Interpreting 'b' of a regression line:
Refer to the change in the variable y for each unit change of the variable x IN CONTEXT

PMCC, r is the PMCC for a population sample

PMCC, ρ is the PMCC for the entire population

Range of PMCC, r: $-1 \leq r \leq 1$

Hypotheses for one tailed test on PMCC:
$H_0: \rho = 0$
$H_1: \rho > 0$ or $H_1: \rho < 0$

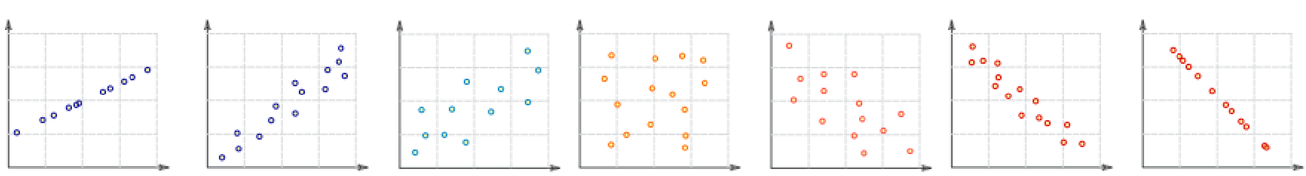
Hypotheses for two tailed test on PMCC:
$H_0: \rho = 0$
$H_1: \rho \neq 0$

Check sample size is big enough to draw a valid conclusion and comment on it if not.

A regression line is only a valid model when the data shows linear correlation.

Only make predictions for the dependent variable using the regression line of y on x within the range of the original data

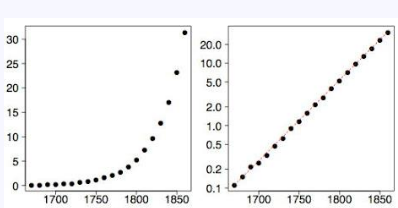
Perfect positive correlation	Strong positive correlation	Weak positive correlation	No correlation	Weak negative correlation	Strong negative correlation	Perfect negative correlation



$r = 1$	$r = 0.8$	$r = 0.3$	$r = 0$	$r = -0.3$	$r = -0.8$	$r = -1$

EXPONENTIAL MODELS

You can use logarithms and coding to transform graphs and examine trends in non-linear data



If $y = ax^n$ then $\log y = \log a + n \log x$

If $y = kb^x$ then $\log y = \log k + x \log b$

@Lucyjc1612

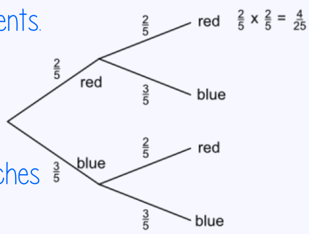
PROBABILITY

KEY WORDS & DEFINITIONS

1. **Experiment** A repeatable process that results in a number of outcomes.
2. **Event** A collection of one or more outcomes.
3. **Sample Space** The set of all possible outcomes. ξ is the universal set.
4. **Mutually Exclusive** Events that have no outcomes in common.
5. **Independent** When events have no effect on another.
6. **Intersection** When two or more events all happen.
7. **Union** When one or both events happen.
8. **Complement** When an event does not happen.

TREE DIAGRAMS

You can use tree diagrams to show the outcome of 2 or more successive events.



Multiply ALONG the branches

Add all the favourable final probabilities.

WHAT DO I NEED TO KNOW

Probabilities of all possible outcomes add to 1

Probability values must be between 0 and 1

Intersection $A \cap B$ $\Rightarrow$ A AND B happen

Union $A \cup B$ $\Rightarrow$ A OR B OR BOTH happen

Complement of A is A' $\Rightarrow$ NOT A

$P(A') = 1 - P(A)$

Mutually Exclusive events:

$P(A \cup B) = P(A) + P(B)$

Independent Events:

$P(A \cap B) = P(A) \times P(B)$

Probability of B, given A has occurred:

$P(B | A)$

For independent events:

$P(A | B) = P(A | B') = P(A)$

In formulae book:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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

VENN DIAGRAMS

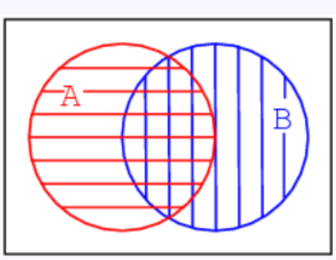
Venn diagrams can be used to show either probabilities or the number of outcomes.

n(A) is the number of outcomes while P(A) is the probability of an outcome

e.g. n(Aces) = 4   P(Ace) = 4/52

Use cross hatch shading to help you work out probabilities.

Focus on one condition at a time, ignoring the other condition completely when you shade.



If P(A) = // and P(B) = \\

$P(A \cap B)$ = #

$P(A \cup B)$ = // + \\ + #

@Lucyjc1612

STATISTICAL DISTRIBUTIONS

KEY WORDS & DEFINITIONS

1 **Random variable** A variable whose outcome depends on a random event.

2 **Sample space** The range of values a variable can take.

3 **Discrete variable** A variable that can only take specific values.

4 **Probability Distribution** A full description of the probability of all possible outcomes in a sample space.

5 **Uniform distribution** When the probabilities in a distribution are all equal.

6 **Binomial Distribution** A distribution where the random variable, X, represents the number of successful trials in an experiment.

7 **Cumulative probability distribution** The sum of probabilities up to and including the given value..

WHAT DO I NEED TO KNOW

Probabilities of all possible outcomes add to 1

$\sum P(X = x) = 1$ for all x

Probability distributions can be described in different ways. E.g. if X = the score when a fair die is rolled

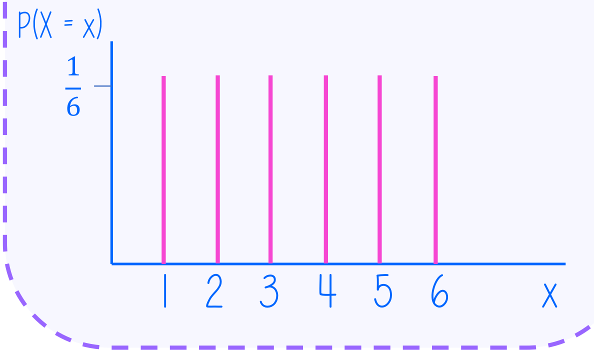
Table:

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

Probability Mass Function:

$$P(X = x) = \begin{cases} \frac{1}{6}, & x = 1,2,3,4,5,6 \\ 0 & \text{otherwise} \end{cases}$$

Diagram:



BINOMIAL DISTRIBUTION

Conditions for a binomial distribution B(n , p)

- Only two possible outcomes (success/failure)
- Fixed number of trials, n
- Fixed probability of success, p
- Trials are independent of each other .

Probability mass function of a Binomial distribution

$$p(X = r) = \binom{n}{r} p^r (1 - p)^{n-r}$$

Binomial Cumulative Probability Function

The sum of all the individual probabilities up to and including the given value of x in the calculation for $P(X \leq x)$

These values can be found in the tables or on a calculator.

Phrase	Means	Calculation
Greater than 5	$X > 5$	$1 - P(X \leq 5)$
No more than 3	$X \leq 3$	$P(X \leq 3)$
At least 7	$X \geq 7$	$1 - P(X \leq 6)$
Fewer than 10	$X < 10$	$P(X \leq 9)$
At most 8	$X \leq 8$	$P(X \leq 8)$

CALCULATORS FOR BINOMIAL

Casio fx-991EX:

Menu 7 – Binomial CD or Binomial PD

Casio CG50:

Menu 2 - F5 Dist – F5 Binomial – Bpd or Bcd

@Lucyjc1612

HYPOTHESIS TESTING

KEY WORDS & DEFINITIONS

1 Hypothesis Test
A process that considers the probability of an observed (or calculated) value occurring.

2 Null Hypothesis, H_0
The hypothesis about the parameter that is assumed to be correct.

3 Alternative Hypothesis, H_1
The hypothesis about the parameter if the assumption is not correct.

4 Test Statistic
The result of an experiment, or the value calculated from a sample.

5 One-tailed Test
A hypothesis test that involves the alternative hypothesis describing the parameter as being less than or greater than the null hypothesis value.

6 Two-tailed test
A hypothesis test that involves the alternative hypothesis describing the parameter as taking any value that is not the null hypothesis value.

7 Critical Region
The region of the probability distribution where the test statistic value would result in the null hypothesis being rejected.

8 Critical value
The first value of the test statistic that could fall in the critical region.

9 Significance Level
The total probability of incorrectly rejecting the null hypothesis.

WHAT DO I NEED TO KNOW

To carry out a Hypothesis Test, assume H_0 is true, then consider how likely the observed value of the test statistic was to occur. Remember we need it to be **even more unlikely** than the significance level in order to be 'significant' and to reject H_0.

If the test is two-tailed there are two critical regions, one at each end of the distribution. We therefore need to halve the significance level at the end we are testing.

If the test statistic is $X \sim B(n, p)$ then the **expected** outcome is **np**

If the observed value lies in critical region we say there is sufficient evidence to reject H_0 and conclude that H_1 is correct.

If observed value is not in critical region we say there is insufficient evidence to reject H_0.

ALWAYS add a final line in your conclusion in the **context** of the question

Beware of questions that say 'The probability in the tail should be as close as possible to the significance level'. In these cases we may choose a value that is actually *slightly* more likely than the significance level.

@Lucyjc1612

THE NORMAL DISTRIBUTION

KEY WORDS & DEFINITIONS

The Normal Distribution

A continuous probability distribution that can be used to model variables that are more likely to be grouped around a central value than at extremities.

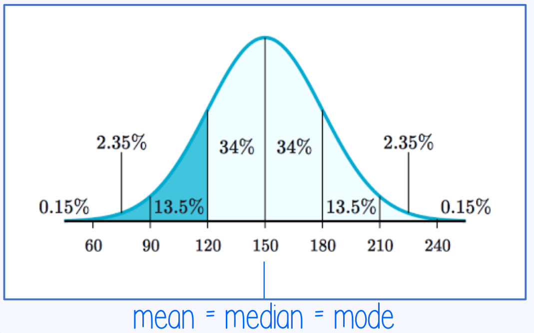
THE NORMAL DISTRIBUTION CURVE

Symmetrically bell-shaped, with asymptotes at each end.

68% percent of data is within one s.d. of μ

95% percent of data is within two s.d. of μ

99.7% percent of data is within three s.d. of μ



mean = median = mode

THE NORMAL DISTRIBUTION TABLE

To find z-values that correspond to given probabilities, i.e. $P(Z > z) = p$ use this table:

p	z	p	z
0.5000	0.0000	0.0500	1.6449
0.4000	0.2533	0.0250	1.9600
0.3000	0.5244	0.0100	2.3263
0.2000	0.8416	0.0050	2.5758
0.1500	1.0364	0.0010	3.0902
0.1000	1.2816	0.0005	3.2905

CALCULATORS FOR NORMAL DISTRIBUTION

Casio fx-991EX:

Menu 7 – Normal PD, Normal CD or Inverse Normal

Casio CG50:

Menu 2 - F5 Dist – F1 Normal – Npd, Ncd or InvN

Choose extremely large or small values for upper or lower limits as appropriate

WHAT DO I NEED TO KNOW

1. The area under a continuous probability distribution curve = 1

2. If X is a normally distributed random variable, with population mean, μ, and population variance, σ^2 we say $X \sim N(\mu, \sigma^2)$

3. To find an unknown value that is a limit for a given probability value, use the inverse normal distribution function on the calculator.

4. The notation of the standard normal variable Z is $Z \sim N(0, 1^2)$

5. The formula to standardise X is $z = \frac{x-\mu}{\sigma}$

6. The notation for the probability $P(Z < a)$ is $\phi(a)$

7. To find an unknown mean or standard deviation use coding and the standard normal variable, Z.

8. Conditions for a Binomial distribution to be approximated by a Normal distribution:
n must be large
p must be close to 0.5

9. The mean calculated from an approximated Binomial distribution is $\mu = np$

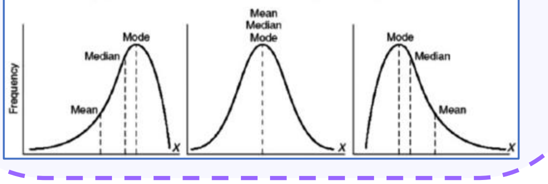
10. The variance calculated from an approximated Binomial distribution is $\sigma^2 = np(1-p)$

11. Apply a continuity correction when calculating probabilities from an approximated Binomial distribution using limits so that the integers are completely included or excluded, as required.

12. The mean of a sample from normally distributed population, is distributed as:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{then} \quad z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

13. Skewed data is NOT 'Normal'



@Lucyjc1612