Math (Statistics) Notes

Chang Si Yuan

August 17, 2020

(ver 1.0)

This compilation of notes are to be used as a reference for the GCE "A"-level Mathematics paper, both as a refresher in theories as well as for general descriptions of presentation form. These notes are meant for free, public use, but at the reader's own risk.
Good luck with your exams.

1 Combinatorics and Probability

1.1 Permutations and Combinations

$$^nP_r = \frac{n!}{(n-r)!}$$

^nP_r gives the number of ways to order r objects out of a possible n objects.

If a group of objects to be ordered contains $k_1, k_2 \dots$ similar objects, divide the number of permutations by $k_1!k_2!\dots$.

If n objects are placed in a ring, find the number of different ways to organize the n objects in a line and then divide by n as each permutation is overcounted for each n possible rotations.

$$^nC_r = \frac{n!}{(n-r)!r!}$$

^nC_r gives the number of ways to choose r objects out of a possible n objects.

1.2 P & C Heuristics

1.2.1 'Group' Method

When arranging n objects, of which k must placed next to each other, consider the k objects as one single unit. Arrange the $n-k+1$ objects, then multiply by the number of ways the k objects can be arranged among themselves.

1.2.2 'Slot' Method

When arranging n objects, of which k cannot be placed next to each other, arrange the remaining $n-k$ objects and then 'slot' each of the k objects in the gaps between the $n-k$ objects.

1.3 Probability and Search Space

For a number of events which have equal chance to occur, the function P represents the fraction of cases where this condition is true.

$$\mathrm{P}(\text{condition}) = \frac{\text{Cases where condition is true}}{\text{Count of all possible cases}}$$

As a result, $\mathrm{P}(\text{Guaranteed true condition}) = 1$ and $\mathrm{P}(\text{Guaranteed false condition}) = 0$.

1.4 Inverse, Unions and Intersections

The probability of multiple conditions can be compounded using their Inverse, Unions and Intersections.

The inverse A' of an event is the set of events which do not appear in the event A.

The union A ∪ B of two events is the set of events which appear at least once in A or B.

The intersection A ∩ B of two events is the set of events which appear in both A and B.

Two events are mutually exclusive if the intersection between the two events is empty.

$$\mathrm{P}(\text{A}) = 1 - \mathrm{P}(\text{A'})$$
$$\mathrm{P}(\text{A} \cup \text{B}) = \mathrm{P}(\text{A}) + \mathrm{P}(\text{B}) - \mathrm{P}(\text{A} \cap \text{B})$$
$$\mathrm{P}(\text{A}) = \mathrm{P}(\text{A}|\text{B}) + \mathrm{P}(\text{A}|\text{B'})$$

1.5 Conditional Probability

Conditional probability considers the probability of an event given that another event has already occurred.

$$\begin{aligned} &\mathrm{P}(\text{B}|\text{A}) \\ &= \text{Probability that B occurs given that A has occurred} \\ &= \frac{\mathrm{P}(\text{B} \cap \text{A})}{\mathrm{P}\, A} \end{aligned}$$

1.6 Independence

Two events are independent if one event occurring does not affect the probability of another from occurring.

$$\begin{aligned} &\text{A and B are Independent} \\ \iff &\mathrm{P}(\text{B}|\text{A}) = \mathrm{P}(\text{B}) \\ \iff &\mathrm{P}(\text{A}|\text{B}) = \mathrm{P}(\text{A}) \\ \iff &\mathrm{P}(\text{A} \cap \text{B}) = \mathrm{P}(\text{A})\,\mathrm{P}(\text{B}) \end{aligned}$$

2 Discrete Random Variable

2.1 Probability Distribution

A discrete random variable X (or denoted by other capital letters) is a variable which can take a finite and distinct set of values. Its probability distribution is a description of the probability that it takes some value.

x	1	2	...	5
$\mathrm{P}(X = x)$	0.2	0.3		0.1

2.2 Expectation and Variance

The expectation value of a discrete random variable is the sum of the products of its possible values and the probability for it to be some value.

$$\mathrm{E}(X) = \sum_x \mathrm{P}(X = x) \times x$$

The variance of a discrete random variable is a quantification of how spread out are the possible values of X from its mean.

$$\begin{aligned}\mathrm{Var}(X) &= \sum_x \mathrm{P}(X=x) \times (x - \mathrm{E}(X))^2 \\ &= \mathrm{E}(X^2) - \mathrm{E}(X)^2\end{aligned}$$

The standard deviation σ of a discrete random variable is another representation of spread from its mean, but more accurate to the scale of the original random variable.

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

2.3 Binomial Distribution

A binomial distribution $X \sim B(n,p)$ is the probability distribution describing the number of successful trials over n total trials, each with a probability of success p.

$$\begin{aligned}\mathrm{P}(X=x) &= ^n C_x \times p^x \times (1-p)^{n-x} \\ \mathrm{E}(X) &= np \\ \mathrm{Var}(X) &= np(1-p)\end{aligned}$$

3 Continuous Random Variable

3.1 Continuous Random Variable

Unlike discrete random variables, continuous random variables are described by a probability distribution function (p.d.f.). The actual probability of the random variable having a value between a range is its direct integral of the p.d.f. within the range.

A p.d.f., when integrated across the range $(-\infty, \infty)$ should have a total probability of 1. Such a function is said to be normalized.

3.2 Expectation and Variance

The expectation value of a continuous random variable is the integral of the product of its instantaneous value and the probability distribution function at that value.

$$\mathrm{E}(X) = \int \mathrm{P}(X=x)x\ dx$$

The variance of a continuous random variable is obtained in a similar fashion to discrete random variables.

$$\begin{aligned}\mathrm{Var}(X) &= \int \mathrm{P}(X=x) \times (x - \mathrm{E}(X))^2\ dx \\ &= \mathrm{E}(X^2) - \mathrm{E}(X)^2\end{aligned}$$

3.3 Normal Distribution

A normal distribution is a graph of the form e^{-x^2}, with added terms to ensure that its mean and standard deviation is equal to a specified value. Normal distributions are defined as $X \sim N(\mu, \sigma^2)$ by assigning them a mean and a standard deviation.

3.4 Combination of Random Variables

$$\begin{gathered}X \sim N(\mu_X, \sigma_X^2) \qquad Y \sim N(\mu_Y, \sigma_Y^2) \\ X + c \sim N(\mu_X + c, \sigma_X^2) \\ X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \\ X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2) \\ k_1X + k_2Y \sim N(k_1\mu_X + k_2\mu_Y, k_1^2\sigma_X^2 + k_2^2\sigma_Y^2)\end{gathered}$$

- Adding a constant to a normal distribution only adjusts its mean
- Adding two normal distributions adds their means together and adds their variances
- Subtracting a normal distribution from another requires subtracting their mean but adding their variances
- Multiplying a random variable by a constant factor k requires adding its mean with a factor k and adding its variance with an additional factor of k^2

3.5 Standard Normal Distribution

The standard normal distribution Z is a specially defined normal distribution as $Z \sim N(0,1)$. Other normal random variables can be adjusted in order to equal the standard normal distribution, by offsetting $\bar{X}$ to be centered at 0 and then dividing by its standard deviation σ.

$$Z = \frac{N - \mu}{\sigma}$$

From this, the calculation of probabilities can be done through the use of data tables or the inverse normal distribution function.

3.6 Modeling as Normal Distribution

To model a random variable as a normal distribution:

- Random variable should have bell curve-like distribution
- Random variable should 'make sense' with values from $(-\infty, \infty)$, or at least minimize the probability of a nonsensical value
- Multiple observations of the random variable should be independent of each other
- Observations of multiple different random variables should be independent of each other

4 Sampling

4.1 Populations and Samples

A population is the entire collection of data to be studied. When populations are too large, too fluid/changing to measure at once or the amount of data to be collected is limited, a sample may be taken to represent the distribution of the population instead.

Given a population described by a random variable X, multiple independent observations of that random variable can be used to obtain a sample. From this sample, an approximate of the population's parameters of mean and variance can be obtained.

4.2 Sample Parameters

When a sample of size n is taken from a population X with unknown population mean μ and unknown population variance σ^2, sample mean and sample variance can be obtained.

From the sample parameters, an unbiased estimate of the population mean $\bar{x}$ and an unbiased estimate of the variance s^2 can be obtained.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum x}{n}$$
$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$$

4.3 Sample Mean

Different samples of a population may get a different unbiased estimate of mean $\bar{x}$. As such, the distribution of such sample means can be said to be a random variable $\bar{X}$, defined as

$$\bar{X} = \frac{1}{n}\sum X = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

with $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

4.4 Central Limit Theorem

The central limit theorem states that the sample mean from of large sample size $n \geq 50$, regardless of the distribution of the population, will *approximately* follow a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

5 Hypothesis Testing

5.1 Hypotheses

A null hypothesis $H_0 : \mu = k$ is a proposed hypothesis of what property a random variable should hold.

An alternative hypothesis $H_1 : \mu \neq k$ is a proposed hypothesis which contradicts a null hypothesis, and describes how a null hypothesis will be tested.

The Level of Significance α of a test is the probability of incorrectly rejecting the null hypothesis, i.e. the the chance a false positive is identified. Uses of α can range from 0.000001 in physical sciences to 0.05 in general sciences to 0.3 in social sciences.

5.2 Z-Test

A Z-test is a statistical test for the validity of a null hypothesis which follows a normal distribution.

A null hypothesis is first established with its mean μ and variance σ^2ORs^2, either from known data or using an unbiased estimate of a sample.

A sample mean with n samples and mean $\bar{x}$ is taken. This sample mean is then compared to the distribution of the null hypothesis with distribution $X \sim N(\mu, \frac{\sigma^2}{n})$.

The sample mean is then verified if it lies at the extremities of the null hypothesis' probability distribution, by finding its relative probability (p-value) or finding the critical value at which results then are beyond the level of significance (z-value).

5.2.1 Syntax for Z-Test

1. Define random variables if necessary.

   'Let X be the random variable denoting ___.'

2. Hypotheses H_0 and H_1 with their definitions.

   '$H_0 : \mu =$ ___ vs $H_1 : \mu \neq$ ___'; OR
   '$H_0 : \mu =$ ___ vs $H_1 : \mu >$ ___'; OR
   '$H_0 : \mu =$ ___ vs $H_1 : \mu <$ ___'

3. Type of test, level of significance.

   'Conduct a 2-tailed test at a $\alpha =$ ___% level of significance'; OR
   'Conduct a 1-tailed test at a $\alpha =$ ___% level of significance'

4. Describe random variable in terms of known and unknown quantities, stating the values of known quantities; state the sample used and its derived quantities.

   If variance is given:

   'Under H_0, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ where $\mu =$ ___ and $\sigma^2 =$ ___.

From sample: $\bar{x} = \_\_\_$, $n = \_\_\_$'

If variance is estimated from sample:

'Under H_0, $\bar{X} \sim N\left(\mu, \frac{s^2}{n}\right)$ where $\mu = \_\_\_$.
From sample: $\bar{x} = \_\_\_$, $n = \_\_\_$, $s^2 = \_\_\_$,'

5. Calculate p-value or z-value and compare to level of significance.

   p-value approach: calculate the area of the probability distribution between the extremities and $\bar{x}$, then compare to level of significance.

   1-tailed: 'p-value $= \mathrm{P}(\bar{X} < \bar{x}) = \_\_\_$'
   2-tailed: 'p-value $= \mathrm{P}(\bar{X} < \bar{x}) \times 2 = \_\_\_$'
   In 2-tailed tests, p-value is two times the probability calculated as two 'tails' of probability are compared with α, do NOT halve α instead.
   End with comparing to α, writing less or greater than.

   z-value approach: find the 'critical value' corresponding to α and the standard normal distribution Z, then transform to the given distribution $\bar{X}$ and compare with $\bar{x}$.
   1-tailed: 'z-value $= \mathrm{P}(Z < \frac{c-\mu}{\sigma}) = \alpha, c = \_\_\_$'
   2-tailed: 'z-value $= \mathrm{P}(Z < \frac{c-\mu}{\sigma}) = \frac{\alpha}{2}, c = \_\_\_$'
   In 2-tailed tests, z-value is calculated with α divided by 2 because the probability is distributed across two tails.
   End with comparing to $\bar{x}$, writing less or greater than.

6. Conclude test with context.

   'Reject H_0 because there is sufficient evidence at a ___% level of significance to conclude that ____ '; OR

   'Reject H_0 because there is insufficient evidence at a ___% level of significance to conclude that ____ '

6 Linear Regression

6.1 2-Variable Data

Previous statistics regarding probability distributions and hypothesis testing typically involve data points with one dimension of quantity, but often there is a use to determine a relation between two different quantities of collected data. The approach of correlation and regression attempts to identify these relationships in a consistent and objective manner while handling statistical interference.

Data collected with two quantities at one point are called bivariate data. Sometimes one quantity may be controlled to be measured at predetermined values. In this scenario, the controlled variable is the 'independent' variable and the other is the 'dependent' variable.

6.2 Scatter Plots

A scatter plot displays points on a graph where data has been measured, with each set of values varying the x and y coordinates respectively. When drawing scatter plots:

- Draw and label axis with their variables x, y, arrow pointing towards increasing.
- Mark the minimum and maximum values of each axis and label their values.
- Attempt to maintain the same shape of the trend
- Preserve relative heights of plots, preserve even or uneven spacing and grouping.
- Otherwise, no need to be too precise.

A scatter plot is good for how tightly-spaced a group of data are to its best fit line and also the type of relation (linear, quadratic, exponential etc.) it displays.

6.3 Product Moment Correlation Coefficient

The Product Moment Correlation Coefficient r is a quantity describing whether a set of bivariate data follows a increasing $(r \approx 1)$ or decreasing $(r \approx -1)$ trend. The formula for calculation can be found in MF26, or alternatively calculated through your GC using the "2-Var Stats" function.

$r = 1$ indicates a perfect positive linear correlation with all points on the line, $r = -1$ indicates a perfect negative linear correlation with all points on the line. $r = 0$ indicates no linear correlation, meaning variables could be not related OR have a nonlinear relationship.

The value of r^2 removes the negative sign on r and its value used to determine if a relation present.

6.4 Linear Regression

After identifying a linear correlation between a set of bivariate data, the data can then be given a best fit line. The method of Linear Least-Squares Regression is an objective and consistent method to obtain an equation of the form $y = a + bx$ and calculating the necessary constants a and b.

The least squares regression of y on x assumes that the obtained data has NO error in x and that all deviation of data points from the best fit are due to errors in y. A linear graph is then fit to the data to minimize the sum of square of all residuals. At a data point (x_1, y_1) and point on the best fit (x_1, y_1'), the residual is $e_1 = y_1' - y_1$ and the sum of squares of residuals is $\sum e_r^2$. As a result of this approach, x is the independent variable and y is the dependent variable.

The process of obtaining this line of best fit is not through trial and error, but rather through selecting a central point and calculating the most suitable gradient to minimize the sum of squares of residuals. For a set of data with points (x_i, y_i), the point $(\bar{x}, \bar{y})$ is first selected and the gradient $b$is calculated with:

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

The linear regression of x against y treats x as the dependent variable and y as the independent variable. This alternative approach is useful if data given goes against conventions of labeling x and y, but also in scenarios where neither set of data are reliably controlled. This regression assumes NO error in obtained values of y, and that all deviation of data points from the best fit are due to to errors in x.

6.4.1 Linearization

Even if two variables are not related by a linear relation of the form $y = mx + c$, some types of relationships can be manipulated into that form and to find a linear relationship between them.

$$y = ax^2 + b \xRightarrow{X=x^2} y = aX + b$$

$$\sqrt{y} = \frac{a}{x} + b \xRightarrow[Y=\sqrt{y}]{X=\frac{1}{x}} Y = aX + b$$

$$\begin{aligned} y = ab^x &\iff \ln(y) = \ln(a) + x\ln(b) \\ &\xRightarrow[Y=\ln(y)]{} Y = \ln(a) + x\ln(b) \end{aligned}$$

$$\begin{aligned} y = ax^b &\iff \ln(y) = \ln(a) + b\ln(x) \\ &\xRightarrow[Y=\ln(y)]{X=\ln(x)} Y = \ln(a) + bX \end{aligned}$$

Given a set of data and multiple ways to linearize it, the linearization which obtains a product moment correlation coefficient closes to ±1 or the linearization which has a similar shape and trend in gradient is the more suitable one.

6.5 Applications of Linear Regression

6.5.1 Determining a Correlation

Whether a correlation is present in data can be determined by inspection of its scatter plot and its product moment correlation coefficient.

Scatter plots are used to identify the shape of a relationship and verify if a relation is linear or otherwise. Scatter plots are also effective in identifying outliers in data which may arise due to errors in data collection, after which specific data points can be ignored.

The product moment correlation coefficient is a quantitative measure of how well the data points fit in a linear relation. Calculating values of r for different methods of linearizing a set of data can help to find the most suitable linearization.

6.5.2 Predicting Values

After obtaining the equation of the linear regression between two values, one can predict and estimate the value of one variable when given the other.

If a set of data is given with a distinct independent variable x (whether stated in the problem, or if it takes 'nice'/whole values in the data given), use the equation obtained from the linear regression of y on x to predict future points.

If a set of data is given with no distinct independent variable, the linear regression equation to use depends on what information is given in the prediction. If finding y for a given x, use the regression of y against x, otherwise use the regression of x against y. This corresponds to the approach of calculating the linear regression: when predicting y for a given x, there is zero error in given value of x and hence the best regression to use would be when the data is assumed to have zero error in x, and vice versa.