Part III — Statistics

Based on lectures by Brian

Notes taken by Dexter Chua

Lent 2017-2018

These notes are not endorsed by the lecturers, and I have modified them (often significantly) after lectures. They are nowhere near accurate representations of what was actually lectured, and in particular, all errors are almost surely mine.

Contents

1 Representation and summary of data - location

1.1 Basic Concepts of Variable

Definition (Quantitative variables and Qualitative variables)**.** Quantitative variable associated with numerical observation. Qualitative variables associated with non-numerical observations.

Definition (Continuous variable and discrete variable)**.** Continuous variable can take ant value in given range. Discrete can take only specific values in a given range.

1.2 Grouped data

Definition (Grouped data)**.** The groups are more commonly known as classes.

- class boundaries.
- mid-point of a class.
- class width.

Example. Example 5-6

Definition (Frequency and cumulative frequency)**.** Number of anything; example is how many sheeps. It is sometimes helpful to add a column to the table showing the running total of the frequencies. This is called the cumulative frequency

Definition (Ungrouped data)**.** Show all data

1.3 Mean , mode and median

Definition (Mode)**.** The mode is the value that occurs most often

Definition (Median)**.** n/2 term or 1 term above

Definition (Mean)**.**

$$\bar{x} = \frac{\sum_i^n x_i}{n}$$

1.4 Linear interpolation

Example. Example 14-15

1.5 Coding

Example. pick 1 example

2 Representation and summary of data - measures of dispersion

2.1 Range and interquartile range

The list of formula:

- Range = Upper value − Lowest value

Example. example 3

2.2 Percentiles split the data into 100 parts

Example. example 4

2.3 Range and Interquartile range

Example (Linear Interpolation)**.**

2.4 Variance and standard deviation

Definition (Variance)**.** Let f stand for the frequency, then $n = \sum f$ and

$$\text{Variance} = \frac{\sum f(x-\bar{x})^2}{\sum f} \text{ or } \frac{\sum fx^2}{\sum f} - (\frac{\sum fx^2}{\sum f})$$

2.5 Variance and standard deviation for grouped data

Definition.

Example. example 7-8

2.6 Coding

Example. example 9-11

3 Representation of data

3.1 Stem and Leaf diagrams

3.2 Outlier

Definition. An outlier is an extreme value that lies outside the overall pattern of the data.

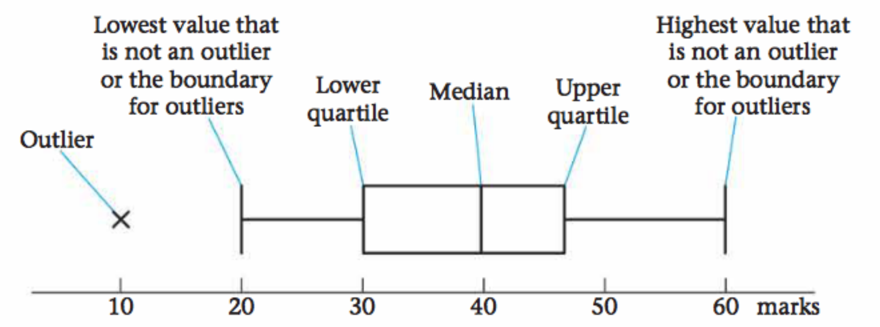
An outlier is any value, which is

$$\text{greater than the upper quartile} + 1.5 \times \text{interquartile range}$$

OR

$$\text{less than the lower quartile} + 1.5 \times \text{interquartile range}$$

3.3 Box plot



3.4 Histogram

Definition (Frequency density)**.**

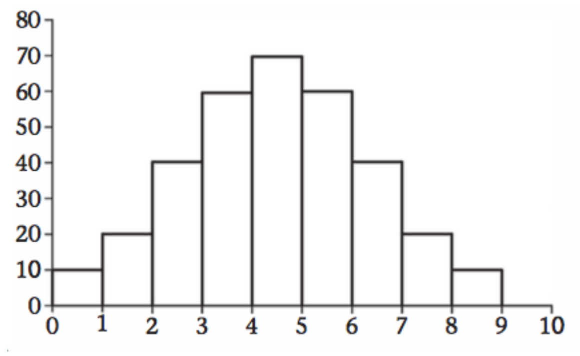
$$\text{frequency density} = \frac{\text{frequency}}{\text{class width}}$$

Example. 7

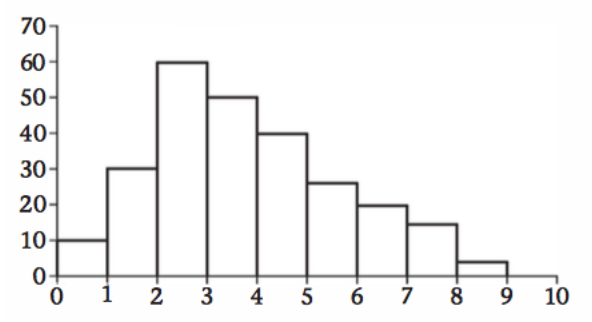
3.5 Skewness (Shape)

A distribution can be symmetrical , have positive skew or have negative skew

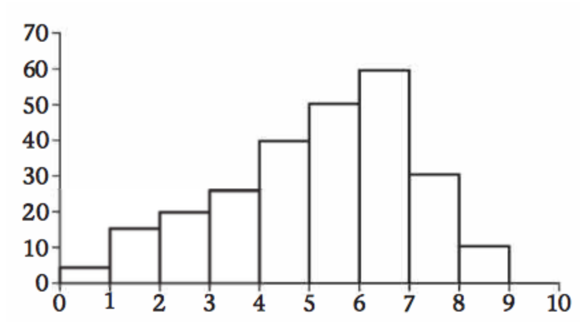
symmetrical $Q_2 - Q_1 = Q_3 - Q_2$ or mode=median=mean



positive :$Q_2 - Q_1 < Q_3 - Q_2$ or mode<median<mean



negative :$Q_2 - Q_1 > Q_3 - Q_2$ or mode>median>mean



Or you can calculate:

$$\frac{3(\text{mean} - \text{median})}{\text{SD}}$$

3.6 What!?

Example. example 10-12

4 Probability

4.1 Classical Probability

4.2 Venn diagram and their rules

Definition (Complementary Probability)**.**

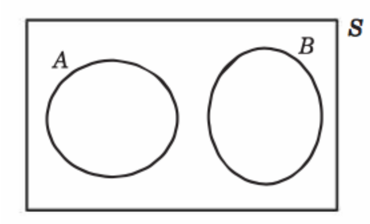
4.3 Conditional Probabilites

4.3.1 Vann diagram

4.3.2 Tree diagram

4.4 Special Events of Probabilites

Definition (Mutually exclusive)**.** When events have no outcomes in common, they are mutually exclusive.



There is no intersection of A and B, so $P(A \cap B) = 0$

We can use $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
result is

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

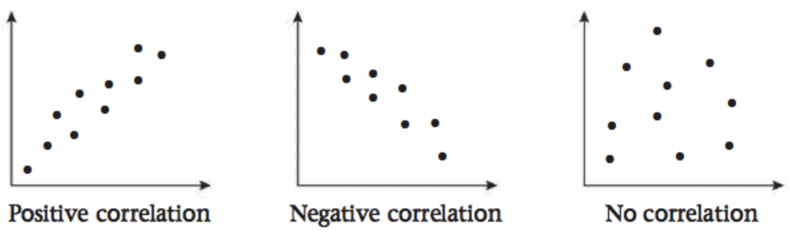
Definition (Independent events)**.** When one event has no effect on another, they are independent so $P(A|B) = P(A)$

by $\frac{P(A \cap B)}{P(B)} = P(A)$ we have:

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

5 Correlation

5.1 Correlation



5.2 Bivariate data

Recall this formula :

$$\text{Variance} = \frac{\sum(x-\bar{x})^2}{n}$$

In correlation we write:

$$S_{xx} = \sum(x-\bar{x})^2$$

$$S_{yy} = \sum(y-\bar{y})^2$$

so

$$\text{Variance} = \frac{S_{xx}}{n}$$

Definition (Co-Variance)**.**

$$S_{xy} = \frac{\sum(x-\bar{x})(x-\bar{y})}{n}$$

5.3 Product moment Correlation coefficient r

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

The value of r varies between -1 and 1

If $r = 1$, positive linear correlation

If $r = -1$, nagative linear correlation

If $r = 0$, no linear correlation

limitation:

5.4 Coding

does not effect r

6 Regression

6.1 Linear

let $y = a + bx$ be a regression line
where

$$b = \frac{S_{xy}}{S_{xx}} \text{ and } a = \bar{y} - b\bar{x}$$

6.2 Coding

6.3 Interpolation and Extrapolation

7 Discrete random variables

7.1 Probability distribution

Definition (Mean / Expected value)**.**

$$E(X) = \sum xp(x)$$

when we find $E(X^n)$:

$$E(X^n) = \sum x^n p(x)$$

Definition (Variable)**.**

$$Var(X) = E(X^2) - (E(X))^2$$

The constant a and b affect on $E(X)$ and $Var(X)$

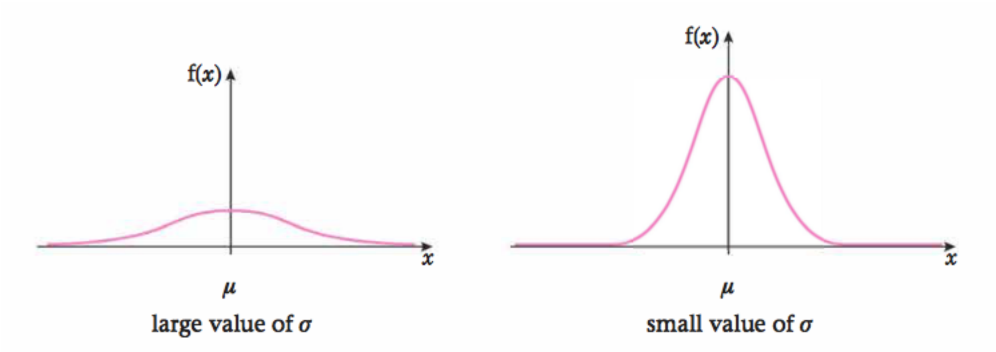
$$E(aX + b) = aE(x) + b$$

$$Var(aX + b) = a^2 Var(X)$$

Definition (Uniform distribution)**.** The distribution is uniform when all the probabilities is the same of all values.

8 The normal distribution

$Z \sim N(\mu, \sigma^2)$ represent the normal distribution.



The random variable X can be written as $X \sim N(\mu, \sigma^2)$

you can transformed X to Z by this formula

$$z = \frac{X - \mu}{\sigma}$$

Example. Example 8-9

9 Binomial distribution

9.1 Basic Concept

$X \sim B(n,p)$ represent the Binomial distribution, then

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

9.2 Mean and Variance

If $X \sim B(n,p)$ then

$$E(X) = \mu = np$$
$$Var(X) = \sigma^2 = np(1-p)$$

Example. example 9-14

10 Poisson distribution

10.1 Basic Concepts

Recall the exponential function

$$e^x = 1 + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + + \frac{x^r}{r!} + \cdots$$

If you let $x = \lambda$ and remember that $\lambda^0 = 1$ this gives

$$e^\lambda = \lambda^0 + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots + + \frac{\lambda^r}{r!} + \cdots$$

Dividing by e^λ gives

$$\frac{e^\lambda}{e^\lambda} = \lambda^0 + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^3 e^{-\lambda}}{3!} + \cdots + \frac{\lambda^r e^{-\lambda}}{r!} + \cdots$$

And the probability function is

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

We say that X has a Poisson distribution with parameter λ abd write

$$X \sim Po(\lambda)$$

10.2 Mean and Variance

$$Var(X) = E(X) = \mu = \sigma^2 = \lambda$$

Lemma. If mean and standard deviation square is same, we usually use Poisson distribution.

Example. example 5-6

10.3 Approximate a Binomial with Poisson

If $X \sim B(n, p)$ and

- n is large
- p is small

then X can be approximated by

$$Po(np)$$

Example. example 7-8 9-10

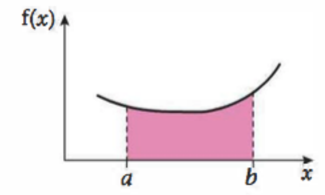
11 Continuous random variables

11.1 Continous random variable

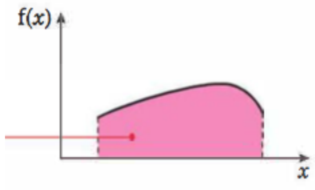
The Continuous random variables with p.d.f $f(x)$ satisfied the following properties:

(i) $f(x) \geq 0$ since we cannot have negative probabilities

(ii) $P(a < X < b) =$ shaded area $= \int_a^b f(x)\,\mathrm{d}x$



(iii) $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$ since the area under the curve $= 1$.



Example. The random variable X has probability density function

$$f(x) = \begin{cases} kx(4-x) & 2 \leq x \leq 4 \\ 0 & \text{otherwise.} \end{cases}$$

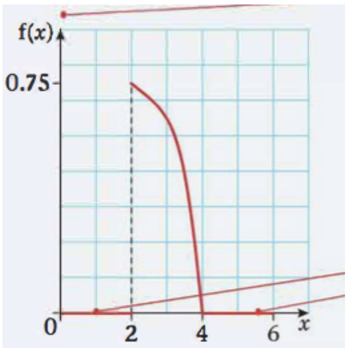
Find the value of k and sketch the p.d.f

Proof.

$$\int_2^4 k(4x - x^2)\,\mathrm{d}x = 1$$

$$k[2x^2 - \frac{x^3}{3}]_2^4 = 1$$

$$k = (\frac{3}{16})$$



∎

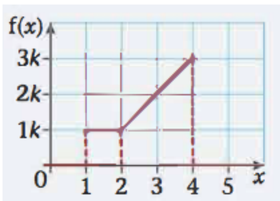
Example. The random variable X has probability density function

$$f(x) = \begin{cases} k & 1 < x < 2 \\ k(x-1) & 2 \le x \le 4 \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of k and sketch the p.d.f

Proof.

$$\int_1^2 k \, \mathrm{d}x + \int_2^4 k(x-1) \, \mathrm{d}x = 1$$
$$k = \frac{1}{5}$$

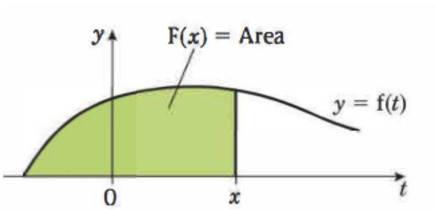


∎

11.2 Cumulative distribution function

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t) \, \mathrm{d}t$$

where $F(x) = P(X \le x) = 1$



If X us a Continous random variable with c.d.f. $F(x)$ and p.d.f $f(x)$

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}t}F(x) \text{ and } F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

Example. example 5-6

11.3 Mean and Variance

If X is a Continuous random variable with p.d.f $f(x)$

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x$$

$$\sigma^2 = E(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - \mu^2$$

Remark. The range is the range of that function instead of negative infinity to infinity.

11.4 Mode, median and quartiles

The median m or Q_2 satisfies $F(m) = F(Q_2) = 0.5$
The lower quartile Q_1 satisfies $F(Q_1) = 0.25$
The lower quartile Q_1 satisfies $F(Q_3) = 0.75$

The mode is the x value at the highest point of the p.d.f $f(x)$

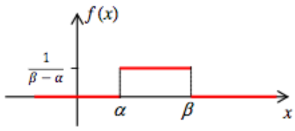
12 Continuous uniform distribution

12.1 Continuous uniform distribution

Definition

A continuous uniform distribution has **constant** probability density over a fixed interval.

Thus $f(x) = \dfrac{1}{\beta - \alpha}$ is the continuous uniform p.d.f. over the interval $[\alpha, \beta]$ and has a rectangular shape.



Median

By symmetry the median is $\dfrac{\alpha + \beta}{2}$

Mean and Variance

The expected mean is $\mathrm{E}[X] = \mu = \dfrac{\alpha + \beta}{2}$, which is the same as the median.

and the expected variance is $\mathrm{Var}[X] = \sigma^2 = \dfrac{(\beta - \alpha)^2}{12}$.

These formulae are proved in the appendix

12.2 Mean and Variance

Example. example 4-7

12.3 Choosing the right model

Example. example 8-10

13 Normal approximation

13.1 Approximating binomial by normal

If $X \sim B(n,p)$ and

- n is large
- p is close to 0.5

Then X can be approximated by

$$Y \sim N(np, np(1-p))$$

Example. $X \sim B(120, 0.25)$ approximated to $Y \sim N(30, (\sqrt{22.5})^2)$

Example. example 4

13.2 Approximating Poisson by normal

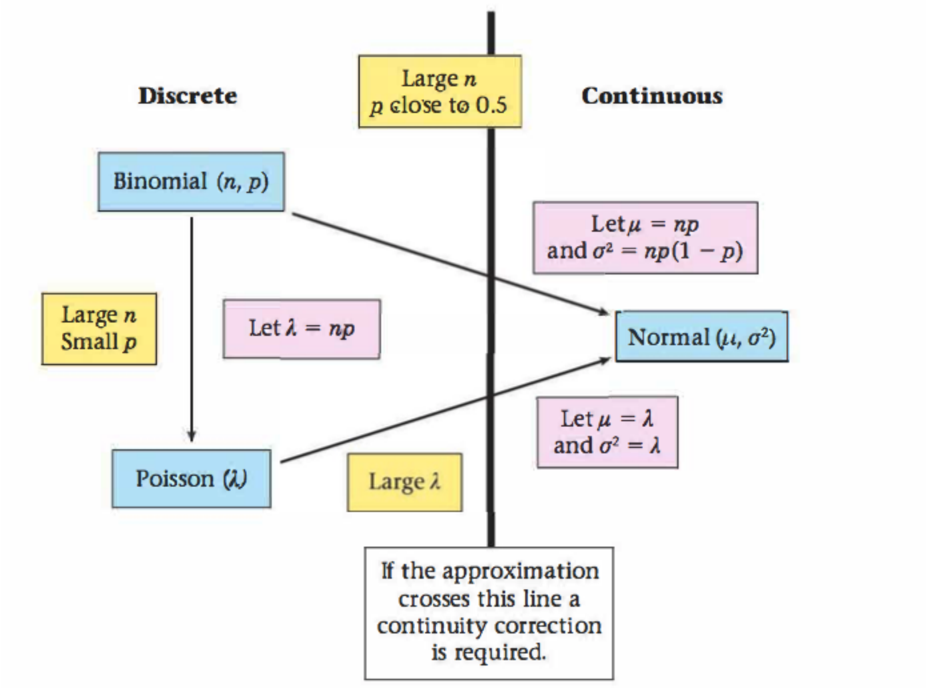
If λ is large

$$X \sim Po(\lambda)\text{to}Y \sim N(\lambda, (\sqrt{\lambda})^2)$$

Example. $X \sim Po(25)$ transformed to $Y \sim N(25, 5^2)$

Example. example 6

13.3 Choosing the appropriate approximation



Example. example 7

14 Population and samples

14.1 The Concept of population and samples

List of the possible samples and find their probabilities and distribution.

Example. example 5 6

15 Hypothesis testing

15.1 Concept of hypothesis testing

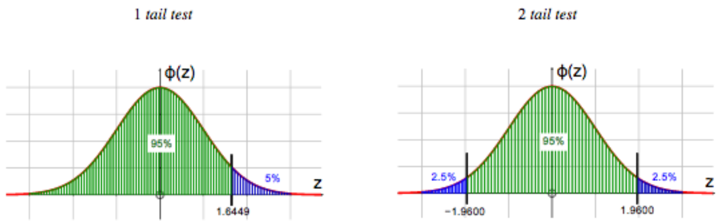
Definition (Null hypothesis H_0)**.** The hypothesis which is assumed to be correct unless shown otherwise.

Definition (Alternarive hypothesis H_1)**.** This is the conclusion that should be made if H_0 is rejected

Definition (Critical region)**.** The range of values which would lead you to reject the null hypothesis, H_0

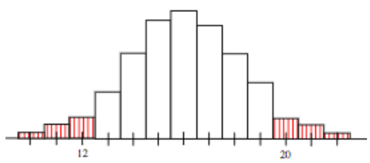
Definition (Significance level)**.** The actual significance level is the probability of rejecting H_0 when it is in fact true.

From your observed result (test statistic) you decide whether to reject or not to reject the null hypothesis H_0



The test statistic is significantat 5%, or that we reject H_0. Thus H_0could actually be true but we still reject it. Thus, the significance level, 5%, is the probability that we reject H_0 when it is in fact true, or the probability of incorrectly rejecting H_0.
When we reject the null hypothesis, H_0, we use the alternative hypothesis to write the conclusion.

The Poisson and Binomial distributions are discrete, and we look at probability histograms.



In the diagram, the critical region (shown by the shaded areas) is $X \leq 12$ or $X \geq 20$.

We include the whole bar around $X = 12$ and around $X = 20$

So $P(X \leq 12)$ is the area to the left of 12.5,
and $P(X \geq 20)$ is the area to the right of 19.5,

If $P(X \leq 12) = 0.0234$ and $P(X \geq 20) = 0.0217$, then
the actual signifaicance level is $0.0234 + 0.0217 = 0.0451 = 4.51\%$
Thus the probability of incorrectly rejecting H_0 is 0.0451.

15.2 One- and two-tailed tests

The One-tail test is

$$\begin{aligned} H_0 &: a = b \\ H_1 &: a > b \text{ or } a < b \end{aligned}$$

The Two-tail test is

$$\begin{aligned} H_0 &: a = b \\ H_1 &: a \neq b \end{aligned}$$

Example. 3

Example. 4-13

16 Combination of random variables

17 Sampling

18 Estimation , confidence intervals and tests

19 Goodness of fit and contingency tables

20 Regression and correlation

21 Quality of tests and estimators

22 One-sample procedures

23 Two-sample procedures