MAS131/231: Introduction to Statistics

Lecturer: Dr Phil Ansell

Office: M511, Phone: 6344

Email: p.s.ansell@ncl.ac.uk

Lectures: Lectures will take place on Wednesdays at 11 in Herschel Building (Lecture Theatre 1) and Thursdays at 12 in Claremont Tower (CLT120). Although brief notes for the course will be distributed in lectures, additional information and most examples (similar to those on the exercise sheets and the exam), will only be covered in the lectures. Regular lecture attendance is vital to performing well in this module.

Tutorials: Tutorials will take place in ODD weeks. Computer practical sessions will take place in WEEKS 2, 6 and 8.

Assessment: Tests in Tutorials will count for 10%. Fortnightly homework will count for 10%. Project work will count for 10%. The 90 minute exam at the end of the semester will count for 70%.

Announcements: Announcements relating to the course will be made via email. You should check your email at least twice a week.

Notes:

1. Handouts/Tutorial sheets/Solutions will only be distributed in lectures. Further copies of handouts/tutorial sheets/solutions will only be available from

   `http://www.mas.ncl.ac.uk/~npsa2/Teaching/MAS131/home.html`

   or via the 'Additional Teaching Material' link to be found on the School of Mathematics and Statistics home page.

2. Homework must be handed in by the given deadline. Late homework will not be accepted unless a *good* reason (*eg.* illness) is provided.

3. Homework not handed in will count as zero for assessment purposes unless there is a good reason and a note from your tutor is provided.

4. The University requires that you attend all lectures, tutorials and practical sessions.

Part I
Continuous Probability Models

1 Introduction

Semester 1 should have given you a fairly good understanding of discrete probability models. When each value of the random variable as well as its probability of occurring can be listed, the random variable is discrete. In this part of the course, we will discuss the other type, namely continuous random variables. Continuous random quantities are random quantities with a sample space which is neither finite nor countably infinite. The sample space is usually taken to be the real line, or a part thereof. Continuous probability models are appropriate when the result of an experiment is a continuous *measurement*, rather than a *count* of a discrete set. Example of continuous random variables include such variables as X = height, X = weight and X = time. If X is a continuous random quantity with sample space S_X, then for any particular $a \in S_X$, we generally have that

$$P(X = a) = 0.$$

This is because the sample space is so "large" and every possible outcome is so "small" that the probability of any "particular" value is vanishingly small. Therefore the probability mass function we defined for discrete random quantities is inappropriate for understanding continuous random quantities. In order to understand continuous random quantities, we need a little calculus.

2 The probability density function

If X is a *continuous* random quantity, then there exists a function $f_X(x)$, called the *probability density function* (PDF), which satisfies the following:

1. $f_X(x) \geq 0, \ \forall x$;
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$;
3. $P(a \leq X \leq b) = \int_a^b f_X(x)dx$ for any a and b.

Consequently we have

$$\begin{aligned} P(x \leq X \leq x + \delta x) &= \int_x^{x+\delta x} f_X(y)dy \\ &\simeq f_X(x)\delta x \qquad \text{for small } \delta x \\ \Rightarrow f_X(x) &\simeq \frac{P(x \leq X \leq x + \delta x)}{\delta x}. \end{aligned}$$

and so we may interpret the PDF as

$$f_X(x) = \lim_{\delta x \to 0} \frac{P(x \leq X \leq x + \delta x)}{\delta x}.$$

2.1 Example

The length of time required by students to complete a 1-hour exam is a random variable with a density function given by

$$f_Y(y) = \begin{cases} cy^2 + y, & 0 \leq y \leq 1, \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Find c and sketch $f_Y(y)$.

(b) Find the probability that a student takes between 30 and 45 minutes to finish the exam.

Notes

1. Remember that PDFs are *not* probabilities. For example, the density can take values greater than 1 in some regions as long as it still integrates to 1.

2. It is sometimes helpful to think of a PDF as the limit of a relative frequency histogram for many realisations of the random quantity, where the number of realisations is very large and the bin widths are very small.

3. Because $P(X = a) = 0$, we have $P(X \leq k) = P(X < k)$ for continuous random quantities.

3 The distribution function

In Semester 1 the cumulative distribution function of a random variable X was defined to be

$$F_X(x) = P(X \leq x), \qquad \forall x.$$

This definition works just as well for continuous random quantities, and is one of the many reasons why the distribution function is so useful. For a discrete random quantity we had

$$F_X(x) = P(X \leq x) = \sum_{\{y \in S_X \,|\, y \leq x\}} P(X = y)$$

but for a continuous random quantity we have the continuous analogue

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(-\infty \leq X \leq x) \\ &= \int_{-\infty}^{x} f_X(z) dz. \end{aligned}$$

Just as in the discrete case, the distribution function is defined for all $x \in \mathbb{R}$ even if the sample space S_X is not the whole of the real line.

3.1 Properties

1. Since it represents a probability, $F_X(x) \in [0, 1]$.
2. $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.
3. When X is continuous, $F_X(x)$ is continuous. Also, by the Fundamental Theorem of Calculus, we have
$$\frac{d}{dx}F_X(x) = f_X(x),$$
and so the *slope* of the CDF $F_X(x)$ is the PDF $f_X(x)$.

3.2 Example

For Example 2.1, where the probability density function was given by

$$f_Y(y) = \begin{cases} \frac{3y^2}{2} + y, & 0 \leq y \leq 1, \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Find and sketch $F_Y(y)$.

(b) Find the probability that a student finishes in less than half an hour.

(c) Given that a student needs at least 15 minutes to complete the exam, find the probability that she will require at least 30 minutes to finish.

4 Medians and quartiles

The *median* of a random quantity is the "middle" of the distribution. That is, it is the value m such that

$$P(X \leq m) = P(X \geq m) = \frac{1}{2}.$$

Equivalently, it is the value, m such that

$$F_X(m) = 0.5.$$

Similarly, the *lower quartile* of a random quantity is the value l such that

$$F_X(l) = 0.25.$$

and the *upper quartile* is the value such that

$$F_X(u) = 0.75.$$

4.1 Example

The proportion of time, Y, that an industrial robot is in operation during a 40-hour week is a random variable with probability density function

$$f_Y(y) = \begin{cases} 2y, & 0 \leq y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Find the median, upper and lower quartiles of the distribution.

5 Expectation of continuous random quantities

The *expectation* or *mean* of a continuous random quantity X is given by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

which is just the continuous analogue of the corresponding formula for discrete random quantities. Similarly, the *variance* is given by

$$Var(X) \;=\; \int_{-\infty}^{\infty} \{x - E[X]\}^2 f_X(x) dx = \int_{-\infty}^{\infty} x^2 f_X(x) dx - \{E[X]\}^2 .$$

Note that the expectation of $g(X)$ is given by

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

and so the variance is just

$$Var(X) = E\left[(X - E[X])^2\right] = E[X^2] - \{E[X]\}^2 .$$

5.1 Example

Weekly CPU time used by an accounting firm has a PDF (measured in hours) given by

$$f_X(x) = \begin{cases} \frac{3}{64}x^2(4-x), & 0 \le x \le 4, \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Check that this is a valid PDF (integrates to 1).

(b) Find the expected value and variance of weekly CPU time.

6 PDF and CDF of a linear transformation

Let X be a continuous random quantity with PDF $f_X(x)$ and CDF $F_X(x)$. Let $Y = aX + b$ where $a > 0$. The CDF of Y is

$$F_Y(y) \;=\; P(Y \le y) = F_X\left(\frac{y-b}{a}\right)$$

and by differentiating both sides with respect to y we get

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

6.1 Example

For Example 5.1, where the weekly CPU time used by an accounting firm has a PDF (measured in hours) given by

$$f_X(x) = \begin{cases} \frac{3}{64}x^2(4-x), & 0 \le x \le 4, \\ 0, & \text{elsewhere.} \end{cases}$$

The CPU time costs the firm £200 per hour and a weekly setup cost of £50. What is the probability that the weekly cost of CPU time exceeds £650?

7 The uniform distribution

Now that we understand the basic properties of continuous random quantities, we can look at some of the important standard continuous probability models. The simplest of these is the uniform distribution. The uniform distribution is very useful for computer simulation, as random quantities from many different distributions can be obtained from $U(0,1)$ random quantities.

7.1 Definition

The random quantity X has a *uniform distribution* over the range $[a,b]$, written

$$X \sim U(a,b)$$

if the PDF is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

7.2 Result

We can show that the CDF of a uniform random quantity defined on the range $[a,b]$ is given by

$$F_X(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x \leq b, \\ 1, & x > b. \end{cases}$$

7.3 Result

The lower quartile, median and upper quartile of the uniform distribution are

$$\frac{3}{4}a + \frac{1}{4}b, \qquad \frac{a+b}{2}, \qquad \frac{1}{4}a + \frac{3}{4}b,$$

respectively.

7.4 Result

The expectation and variance of a uniform random quantity are

$$E[X] = \frac{a+b}{2} \text{ and } Var(X) = \frac{(b-a)^2}{12}.$$

7.5 Example

A parachutist lands at a random point on a line between markers A and B.

(a) Find the probability that she is closer to A than B.

(b) Find the probability that her distance from A is more than three times her distance to B.

(c) Suppose that three parachutists operate independently as described above. What is the probability that exactly one of the three lands past the midpoint between A and B?

8 The exponential distribution

8.1 Definition

The random variable X has an *exponential distribution* with parameter $\lambda > 0$, written

$$X \sim Exp(\lambda)$$

if it has PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

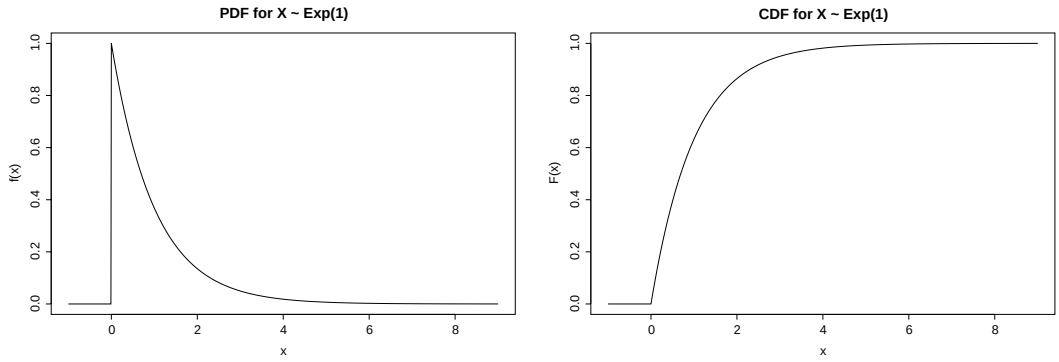
8.2 Result

The distribution function $F_X(x)$ is therefore given by

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

Comment

The PDF and CDF for an $Exp(1)$ are shown on below.



8.3 Result

The expectation and variance of the exponential distribution is

$$E[X] = \frac{1}{\lambda} \text{ and } Var(X) = \frac{1}{\lambda^2}.$$

Comment

This means that the expectation and standard deviation are both $\frac{1}{\lambda}$.

Notes

1. As λ increases, the probability of small values of X increases and the mean decreases.

2. The median m is given by

$$m = \frac{\log 2}{\lambda} = \log 2 E[X] < E[X].$$

3. The exponential distribution is often used to model lifetimes and times between random events.

8.4 Example

The magnitudes of earthquakes recorded in a region of North America can be modelled as having an exponential distribution with mean 2.4, as measured on the Richter scale. Find the probability that an earthquake striking this region will

(a) exceed 3.0 on the Richter scale,

(b) fall between 2.0 and 3.0 on the Richter scale,

(c) Out of the next 10 earthquakes to strike this region, what is the probability that at least one will exceed 5.0 on the Richter scale?

8.5 Relationship with the Poisson process

The exponential distribution with parameter λ is the time between events of a Poisson process with rate λ. Let X be the number of events in the interval $(0, t)$. In Semester 1 we saw that $X \sim P(\lambda t)$. Let T be the time to the first event. Then

$$F_T(t) = P(T \leq t) = 1 - e^{-\lambda t}.$$

This is the distribution function of an $Exp(\lambda)$ random quantity, and so $T \sim Exp(\lambda)$.

8.6 Example

Consider the Poisson process for calls arriving at an ISP at rate 5 per minute. Let T be the time between two consecutive calls. Then we have

$$T \sim Exp(5)$$

and so $E[T] = SD(T) = 1/5$ minutes.

8.7 Result (the memoryless property)

If $X \sim Exp(\lambda)$, then

$$P(X > s + t | X > t) = P(X > s).$$

9 The normal distribution

9.1 Definition

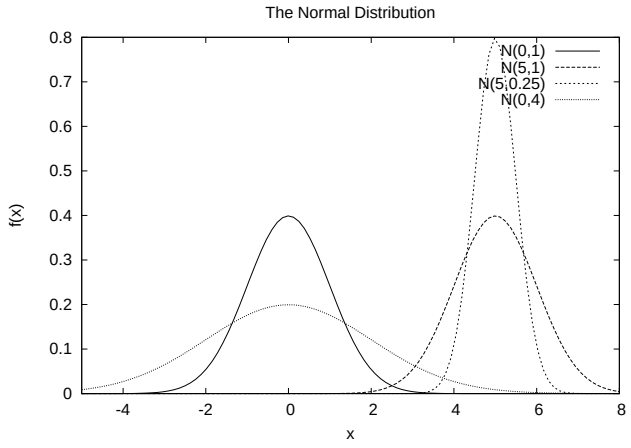
A random quantity X has a normal distribution with parameters μ and σ^2, written

$$X \sim N(\mu, \sigma^2)$$

if it has PDF

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ x \in \mathbb{R},$$

for $\mu \in \mathbb{R}$ and $\sigma > 0$. Note that $f_X(x)$ is symmetric about $x = \mu$ and so (provided the density integrates to 1) the median of the distribution will be μ. The PDFs for a range of $N(\mu, \sigma^2)$ random quantities are given in the plot below.



9.2 Result

If $X \sim N(\mu, \sigma^2)$, i.e. X has a normal distribution with parameters μ and σ^2, then

$$E[X] = \mu \text{ and } Var(X) = \sigma^2.$$

9.3 Definition

A *standard normal* random quantity, $Z \sim N(0, 1)$, is a normal random quantity with mean zero and variance equal to one. The PDF is denoted $\phi(z)$ and is therefore,

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2},\ z \in \mathbb{R},$$

N.B. This is symmetric about zero. The CDF is denoted $\Phi(z)$ and is given by

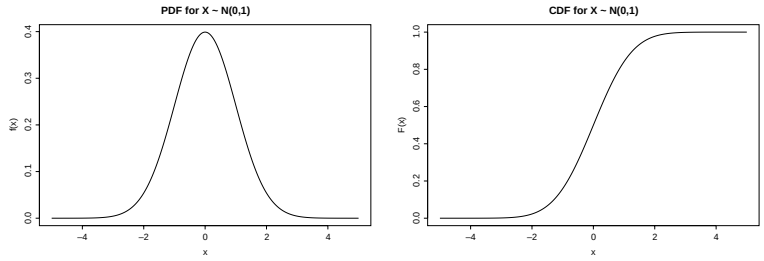
$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(x)dx.$$

Comment

Note that there is no analytic expression for $\Phi(z)$, so tabulated values are used. The following can all be useful for calculations.

$$\Phi(-\infty) = 0,\ \Phi(\infty) = 1,\ \Phi(0) = \frac{1}{2},\ \Phi(-z) = 1 - \Phi(z).$$

The PDF and CDF for a $N(0,1)$ are given below.



9.4 Example

Use tables to compute $\Phi(1.5)$ and $\Phi(1.56)$.

9.5 Example

Use tables to compute $\Phi(-1.2),\ \Phi(-1.23)$.

9.6 Result

The standard normal distribution is important because it is easy to transform any normal random quantity by means of a simple linear scaling. We use the result for the PDF of a linear transformation. If $X \sim N(\mu, \sigma^2)$, then the CDF of X is given by

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

9.7 Example

If $X \sim N(3, 2^2)$ compute

(a) $P(X < 5)$;

(b) $P(2 < X < 4)$.

9.8 Example

MENSA have established that IQ levels in Britain can be modelled by a normal distribution with parameters $\mu = 100$ and $\sigma^2 = 400 = 20^2$, i.e., $X \sim N(100, 20^2)$, where X is the IQ level of people in Britain.

(a) Compute $P(X > 150)$;

(b) What IQ level do you need to be in the top 2.5% of the population?

9.9 Example

(a) If $X \sim N(\mu, 10^2)$ and $P(X > 20) = 0.1$, what is μ?

(b) If $X \sim N(20, \sigma^2)$ and $P(X > 40) = 0.01$, what is σ?

(c) If $X \sim N(\mu, \sigma^2)$, $P(X < 0) = 0.1$ and $P(X > 10) = 0.05$, what are μ and σ^2?

z	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	-0.00
-2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
-2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
-2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
-2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
-2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
-2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
-2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446
-1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
-1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
-1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
-1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357
-1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
-0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
-0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
-0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420
-0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
-0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
-0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
-0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821
-0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
-0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000

Table 1: **The Standard Normal Distribution. Values of $P(Z \leq z)$, $z \leq 0$, where $Z \sim N(0,1)$**

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Table 2: **The Standard Normal Distribution. Values of $P(Z \leq z)$, $z \geq 0$, where $Z \sim N(0,1)$**

$p = \Phi(z)$	$z = \Phi^{-1}(p)$
0.5000	0.000
0.8000	0.842
0.9000	1.282
0.9500	1.645
0.9750	1.960
0.9900	2.326
0.9950	2.576
0.9990	3.090
0.9995	3.291

Table 3: **Quantiles of Standard Normal Distribution**

Part II
Expectation

1 Functions of a single random variable

Before we can develop methods for estimating parameters and drawing inferences, we need some results on expectation. Recall from Semester 1 that the expectation (mean) of a discrete random variable X with probability function $p(x)$ and sample space S is

$$E[X] = \sum_{x \in S} x\, p(x).$$

More generally, the expectation of any function of X, say $g(X)$, is

$$E[g(X)] = \sum_{x \in S} g(x)\, p(x),$$

that is, the expectation of $g(X)$ is a sum of all values of $g(x)$ weighted by how likely the value x is to occur.

A similar result holds for continuous random variables. If X is a continuous random variable with probability density function $f_X(x)$ then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx.$$

An example of this generalisation is the average squared deviation about the mean, that is,

$$Var(X) = E\left[(X - \mu)^2\right],$$

where $E[X] = \mu$ and $g(X) = (X - \mu)^2$.

1.1 Example

Suppose the discrete random variable X has probability function

x	-2	0	1	4
$p(x)$	0.1	0.4	0.3	0.2

Find $E[X]$, $E[X^2]$ and $E[e^X]$.

1.2 Example

Suppose the continuous random variable X has an exponential distribution with parameter $\theta > 0$ and probability density function

$$f_X(x) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Find $E[e^{-X}]$.

The following result holds for both discrete and continuous random quantities:

1.3 Result

(a) $E[aX + b] = aE[X] + b$,

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

where a and b are constants.

1.4 Example

Let X be the maximum daily temperature (in Celsius) in Newcastle during February.

(a) If $E[X] = 10^oC$ then what is the expected temperature in Fahrenheit?

(b) Is temperature more or less variable on the Fahrenheit scale than on the Celsius scale?

2 Linear combinations of independent random variables

Recall (from Semester 1) that two events E and F are independent if

$$P(E \text{ and } F) = P(E) \times P(F).$$

Consider two *independent* discrete random variables X and Y with sample spaces S_X and S_Y respectively. If we define the events as being two particular outcomes of these random variables, namely

$$E = \{X = x\} \text{ and } F = \{Y = y\}$$

then these events are independent, and so

$$P(X = x \text{ and } Y = y) = P(X = x) \times P(Y = y),\ x \in S_X,\ y \in S_Y.$$

The l.h.s. is called the *joint probability function* of X and Y. It describes how likely pairs of values are to occur.

The continuous analogue of this is: if X and Y are *independent* continuous random variables with probability density function $f_X(x)$ and $F_Y(y)$ then the *joint probability density function* of X and Y is

$$f_{X,Y}(x, y) = f_X(x) \times f_Y(y) \qquad -\infty < x, y < \infty.$$

We now quote some results which will be useful in studying the properties of random samples.

2.1 Result

(i) If X and Y are random variables then

$$E[X+Y] = E[X] + E[Y].$$

(ii) If X and Y are *independent* random variables then

$$E[XY] = E[X]E[Y].$$

(iii) If X and Y are *independent* random variables then

$$Var(X+Y) = Var(X) + Var(Y).$$

2.2 Result

(i) If $X_1, X_2, \ldots, X_n$ are random variables then

$$\begin{aligned} &E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] \\ &\qquad = a_1E[X_1] + a_2E[X_2] + \cdots + a_nE[X_n]. \end{aligned}$$

(ii) If $X_1, X_2, \ldots, X_n$ are *independent* random variables then

$$\begin{aligned} &Var(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &\qquad = a_1^2Var(X_1) + a_2^2Var(X_2) + \cdots + a_n^2Var(X_n). \end{aligned}$$

2.3 Example

Suppose X_1, X_2 and X_3 are independent random variables with means -3, 2 and 5, and variances 1, 3 and 2 respectively. Find the mean and variance of

(i) $Y = X_1 + 3X_2 + X_3$;

(ii) $Y = 2X_1 - X_2 - 4X_3$.

2.4 Result

Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables with

$$E[X_i] = \mu,\ i = 1, \ldots, n$$

and

$$Var(X_i) = \sigma^2,\ i = 1, \ldots, n.$$

(a) Show that $E[\bar{X}] = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$.

(b) Show that $E[S^2] = \sigma^2$.

Part III
Statistical Inference

1 Introduction

Statistical inference is the study of how best to draw conclusions from a limited amount of data. For example,

(a) The performance of a new drug to combat cancer.

(b) Daily demand for beds in a hospital ward.

The statistical problem in both the above examples is how to generalise from the conclusions concerning a relatively small amount of data to a much larger (effectively infinite) population. Naturally, the larger the initial experiment, the more reliable the conclusions of the experiment will be when applied to the population. We talk of making *inferences* from the data, and quantify the accuracy or reliability of the inferences.

The general area of statistical inference is very broad and ranges from using simple techniques of exploratory data analysis (EDA), including graphical and numerical summaries, to analysing very sophisticated and complex statistical models. In this course we develop the central ideas of statistical inference by studying some simple statistical experiments.

The first stage of any analysis of data is to consider how the data were collected and what is a plausible statistical model for the population.

1.1 Example

In the cancer experiment we may be interested in the survival time (time until death) of patients and an exponential distribution with probability density function

$$f(x|\theta) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0 & x < 0, \end{cases}$$

may be a satisfactory statistical model for describing the survival time X in the population of cancer patients. The method of choosing (*estimating*) θ must take into account how the data were collected, and most importantly, whether they are representative of the population. For instance, inferences drawn from data collected on a male only ward may not apply to females. The best way of ensuring that the data are representative of the population is to take a *random sample*. This is a collection of data in which all members of the population are *equally likely* to be chosen. In this part of the course, we consider how to make inferences about population quantities using random samples of data. Important extensions of these techniques to those in which the data are not independent or not identically distributed are considered in second, third and fourth year modules. To be successful, you must be familiar with

1. the differences between populations and samples. For example, do you understand the difference between the sample mean $\bar{x}$ and the population mean μ?

2. the differences between random variables (written in capitals) and observations on random variables (written in lower case). For example, do you understand that random variables have distributions and observations do not? Do you understand that the notation $X = x$ represents the random variable X taking the value/observation x?

1.2 Random Samples

Many statistical investigations involve taking random samples to obtain information about, or survey opinion in, a population. For example, opinion polls are used not only by political parties to assess voters reaction to possible new policies and by newspapers in election periods to gauge the popularity of political parties but also by manufacturers to assess the impact of an advertising campaign or to find out why customers use a rival product. The actual sampling method used can be quite sophisticated. For example, it can ensure that the sample contains known proportions of certain target groups of the population, such as social classes A, B, C and D. However, the central mechanism of all statistically valid polling schemes is to take a random sample from the population (or group within the population).

Suppose we are interested in the cigarette smoking habits of the 1000 smokers on a remote island. In order to gain some idea of the level of nicotine in these smokers, it is decided to take a random sample of 5 smokers and measure their blood plasma nicotine level. Table 1.1 contains the nicotine levels of all 1000 smokers (measured in nanograms per millilitre, ng/ml), written in blocks of 100 smokers. Note that the population mean level is $\mu = 320\, ng/ml$. We shall pretend that all this information is not available to us, and see how we can take random samples and possibly draw inferences about μ.

First we must decide exactly how we can take a random sample of size 5 – each member of the population must have the same probability of being chosen. We begin by numbering the population 1–1000: the top left-hand block will be smokers 1–100, counting 1–10 along the first row, then 11–20 along the second row, and so on. The top right-hand block will contain smokers 101–200, counting again along rows. Repeating this for the other blocks gives a unique label to each smoker in the population.

The next step is to select the 5 smokers for our random sample. What we need are 5 random numbers from the discrete uniform distribution on $\{1, 2, \ldots, 1000\}$. We can generate these using values from a uniform $U(0, 1)$ distribution – these are the random numbers 0.000–0.999 given by a standard calculator. Taking the first three digits after the decimal point and then adding one will give values from the required discrete uniform distribution. For example, if the calculator gives $u = 0.636$, then we select smoker 637, giving our first observation as $x_1 = 374$. Repeating this on my calculator produces the random sample:

$$\begin{array}{lll} u = 0.636 & smoker = 637 & x_1 = 374 \\ u = 0.326 & smoker = 327 & x_2 = 452 \\ u = 0.848 & smoker = 849 & x_3 = 271 \\ u = 0.665 & smoker = 666 & x_4 = 419 \\ u = 0.679 & smoker = 680 & x_5 = 643 \end{array}$$

282	258	399	271	343	285	247	513	171	123	168	327	430	240	410	341	90	512	245	336
290	263	446	185	330	111	243	376	139	351	311	389	546	321	393	487	287	514	149	315
264	320	217	257	349	640	97	298	393	454	363	354	360	326	199	502	154	273	213	413
293	407	362	270	344	263	290	263	50	253	345	581	229	264	304	394	246	235	417	452
499	276	412	323	310	177	248	178	409	275	278	307	495	515	232	432	577	269	370	248
339	404	371	262	336	218	274	483	211	245	316	381	432	233	223	447	412	250	262	337
202	133	356	408	224	379	197	278	235	509	171	232	429	315	326	602	63	290	230	121
242	389	219	206	393	437	306	152	294	271	230	398	346	344	379	347	468	300	325	237
305	174	291	261	214	532	335	63	100	357	190	347	208	420	322	463	203	216	356	504
389	236	445	378	255	301	308	150	289	453	464	273	211	450	222	250	214	259	296	356
320	420	357	160	372	99	316	218	248	322	145	399	433	393	403	361	241	234	388	255
261	279	369	342	168	322	304	254	99	503	303	212	105	166	257	422	460	331	288	410
346	370	235	355	65	340	420	338	568	644	164	288	319	159	324	208	452	297	305	259
268	340	305	361	319	519	293	380	286	431	402	329	363	330	612	248	302	592	589	349
446	588	304	454	164	240	293	478	540	339	245	257	222	471	469	273	244	126	174	183
277	216	555	401	380	338	212	476	77	363	140	451	329	66	217	461	435	380	314	324
522	111	119	316	116	471	142	336	277	101	518	264	226	256	539	324	320	292	476	324
333	332	404	362	202	204	341	80	333	267	439	136	343	389	244	370	268	362	317	400
372	595	314	182	470	192	555	374	368	192	225	321	435	403	316	312	307	368	236	452
192	63	407	125	253	89	70	186	491	342	122	367	106	334	161	177	180	355	356	317
454	122	286	39	361	262	316	272	285	201	191	162	229	334	278	231	154	290	277	392
644	297	398	118	246	148	478	167	337	344	395	334	255	401	504	304	408	204	673	126
192	507	41	457	405	306	282	446	195	512	252	510	557	191	321	404	542	438	291	449
377	240	441	308	346	265	375	332	580	130	353	426	95	588	332	109	467	333	388	309
263	529	172	529	315	257	481	260	297	382	438	64	226	185	369	275	320	126	321	375
190	340	337	224	363	212	371	229	175	388	332	315	389	452	266	393	323	253	280	420
219	400	378	241	616	551	359	489	314	450	645	224	320	405	182	251	370	341	318	232
240	471	293	240	184	296	617	565	206	147	169	401	140	462	389	310	262	334	263	269
323	351	187	544	387	425	353	175	378	484	205	295	413	189	559	251	480	283	262	304
213	574	579	325	246	206	419	306	471	264	270	300	278	131	561	328	440	514	280	391
281	403	256	348	183	161	444	482	338	268	313	252	179	414	444	266	400	435	433	506
203	269	450	322	459	183	212	242	144	406	401	174	605	270	487	494	235	316	368	319
260	254	157	377	145	284	401	220	452	59	335	467	251	192	371	298	317	382	363	397
282	303	328	378	363	636	374	143	495	239	423	496	411	462	282	411	203	395	590	388
278	272	417	666	233	316	287	268	186	247	339	397	276	291	324	81	271	399	129	325
247	152	315	224	130	323	352	276	398	338	231	258	310	421	215	85	237	356	439	348
507	277	240	188	321	419	370	374	211	224	340	264	441	226	563	279	297	114	546	277
281	196	498	375	348	234	469	103	324	643	315	293	444	109	408	100	477	293	138	206
485	279	494	513	97	293	669	312	425	70	181	210	241	187	448	55	253	564	404	382
31	496	234	200	411	386	218	382	483	405	435	414	379	360	194	291	393	247	314	285
204	188	444	416	106	485	276	250	248	200	352	463	251	197	197	456	293	333	373	240
295	297	271	141	319	256	197	110	338	237	249	291	393	437	432	274	202	182	176	212
482	96	272	296	323	289	285	160	203	336	217	321	202	266	253	436	390	259	596	383
236	291	226	250	270	439	360	310	326	415	447	336	354	273	243	390	213	318	346	599
637	255	61	393	324	492	484	259	271	150	550	185	224	352	387	441	232	261	313	410
246	529	97	448	369	199	140	498	287	293	258	431	267	396	217	340	278	297	387	281
162	237	305	239	246	412	632	385	342	340	673	414	298	383	152	438	408	452	492	603
439	223	404	466	380	214	155	410	291	234	248	325	391	338	416	262	361	358	484	129
152	363	90	383	365	500	362	190	343	138	233	179	200	476	128	308	221	649	278	152
525	275	355	585	394	183	488	323	312	595	257	434	160	375	478	353	239	331	426	477

Table 1.1: Blood plasma nicotine levels for 1000 smokers (ng/ml)

Obviously, care must be taken not to select any smoker more than once. Therefore, when selecting a random sample of smokers, if a smoker is selected that is already in the sample, this additional selection should be rejected and another smoker selected (using this algorithm). This technique is called *sampling without replacement*.

Thus, we have a random sample on which to try to draw inferences about μ, the mean nicotine level in the population as a whole. The most obvious best guess for μ is the sample mean $\bar{x} = 431.8\, ng/ml$. Clearly, this is some way from the correct population value $\mu = 320\, ng/ml$. This result begs the question of how accurate our sample mean $\bar{x}$ can be in estimating the population mean μ.

Take another random sample of size $n = 5$:

$u = 0.557$	$smoker = 558$	$x_1 = 253$
$u = 0.427$	$smoker = 428$	$x_2 = 446$
$u = 0.902$	$smoker = 903$	$x_3 = 251$
$u = 0.427$	$smoker = 428$	try again
$u = 0.363$	$smoker = 364$	$x_4 = 256$
$u = 0.013$	$smoker = 14$	$x_5 = 185$

giving a sample mean $\bar{x} = 278.30\, ng/ml$. This sample mean is much closer to the population mean but still not very close. Also, it is quite different from the mean of the previous random sample.

Repeat this procedure yourself (filling the tables below) to select three more random samples of size $n = 5$ and calculate the sample means. How close are these sample means to the correct population value $\mu = 320\, ng/ml$?

Your random sample 1

$u =$	$smoker =$	$x_1 =$
$u =$	$smoker =$	$x_2 =$
$u =$	$smoker =$	$x_3 =$
$u =$	$smoker =$	$x_4 =$
$u =$	$smoker =$	$x_5 =$
		$\bar{x} =$

Your random sample 2

$u =$	$smoker =$	$x_1 =$
$u =$	$smoker =$	$x_2 =$
$u =$	$smoker =$	$x_3 =$
$u =$	$smoker =$	$x_4 =$
$u =$	$smoker =$	$x_5 =$
		$\bar{x} =$

Your random sample 3

$u =$	$smoker =$	$x_1 =$
$u =$	$smoker =$	$x_2 =$
$u =$	$smoker =$	$x_3 =$
$u =$	$smoker =$	$x_4 =$
$u =$	$smoker =$	$x_5 =$
		$\bar{x} =$

1.3 Examples of Statistical Inference

1. Returning to the cancer example, suppose we have data in the form of two independent random samples, one for males and one for females. Suppose the variation in the data looks like an exponential distribution is appropriate, but with (possibly) different parameters θ_M for males and θ_F for females. Questions of interest may include

(a) Using the data, what are the best guesses at the values of θ_M and θ_F?

(b) How accurate are these guesses? We know that our guesses won't have 100% accuracy since it is very likely that we would get different data if we repeated the experiment.

(c) How plausible is it that the drug affects males and females in a similar way, that is, is $\theta_M = \theta_F$?

(d) Are the data consistent with exponential distributions?

(e) Suppose the initial experiment is a small-scale pilot study. How many males and females should be recruited into the main study in order that our final conclusions will be reliable?

2. Suppose we have data on the number X of attempts required for people to pass a driving test. If the result of driving tests are independent of one another, each with success probability θ, then

$$\begin{aligned} Pr(X = i) &= Pr(\text{fail } i-1 \text{ tests and pass the } i^{\text{th}} \text{ test}) \\ &= (1-\theta)^{i-1}\theta, \quad i = 1, 2, 3, \ldots . \end{aligned}$$

This is called the *Geometric*(θ) distribution. This distribution is the simplest one which can be used to model data of this type. It is the *independence* assumption between test results which makes this model relatively simple (and perhaps unrealistic). A more realistic model may take into account that learning to drive may be a cumulative process with diminishing returns, that is, people learn from taking tests but if they don't pass after say 3 tests then their chance of passing reduces with each test taken. Such a model would have

$$Pr(X = 1) < Pr(X = 2) < Pr(X = 3) > Pr(X = 4) > Pr(X = 5) > \ldots$$

whereas, the (simple) geometric model has decreasing probabilities

$$Pr(X = 1) > Pr(X = 2) > Pr(X = 3) > Pr(X = 4) > Pr(X = 5) > \ldots \ .$$

Suppose the data are recorded by age group and sex, and that a (simple) geometric model is thought to be correct. We may be able to answer:

(a) What are the estimates of the values of θ in the different groups? How accurate are these estimates?

(b) Are there any obvious patterns in the estimates? For example, are there any differences between males and females? Is there a consistent pattern in the estimates across ages?

(c) Are the data consistent with geometric distributions? If not, is θ a decreasing function in age? Can we find a (fairly simple) distribution which is more consistent with the data and which satisfies our "learning" model equations?

2 Estimation of Population Quantities

2.1 Introduction

Suppose we are interested in determining some summary measure of a characteristic X in a very large population, such as the mean μ or variance σ^2 of X. For example, X may be annual wages in the U.K. or the amount of alcohol drunk weekly by students. We cannot obtain all values of X in the population because of its size, so we sample the values in a small proportion of the population. We choose people randomly to make sure that the sample is truly representative of the population. Suppose we take a random sample of size n. We write this as $x_1, x_2, \ldots, x_n$. As these observations are made on people chosen at random, we can think of them as observations on independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots, X_n$. Sometimes, we refer to $X_1, X_2, \ldots, X_n$ as the random sample. For example, if we are measuring wages in the UK, then X_1 represents the wage of the first person to be chosen in the random sample, X_2 the wage of the second person chosen and so on. Before we obtain our sample, the value of X_1 (say) is unknown but does have a distribution, namely the distribution of wages in the UK. Once we we have sampled, we observe the value x_1 (say £15000) on X_1. This is also the case for the other random variables and so $X_1, X_2, \ldots, X_n$ represent the possible values of the wages before we take the sample and the observations $x_1, x_2, \ldots, x_n$ the actual values observed in the sample.

We will now see how random samples can be used to estimate the population mean μ and the population variance σ^2. The key statistical properties of the random sample we shall be using are that $X_1, X_2, \ldots, X_n$ are *independent* random variables, each with the same distribution (the population distribution), and, in particular, that they have the same mean and the same variance:

$$E[X_1] = E[X_2] = \cdots = E[X_n] = \mu$$

and

$$Var(X_1) = Var(X_2) = \cdots = Var(X_n) = \sigma^2.$$

2.2 Estimation of the Population Mean

An obvious estimate of the population mean μ is the sample mean $\bar{x}$. But how good an estimate is it? Does it make best use of all the information in the data? Each time a sample is taken from the population, the values in the sample will change because different members of the population will be selected and so we will get different values for $\bar{x}$ in different samples. But which one should we use? They can't all be correct! Is it possible to get a value of $\bar{x}$ which is "miles away" from μ? To answer such questions we study the distribution of all possible values of $\bar{x}$ we can get when sampling the population, that is, we study the distribution of the random variable

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Sometimes this distribution is called the *sampling distribution* of $\bar{X}$ to reflect the origin of the random variation. We say that $\bar{X}$ is the *estimator* of μ and $\bar{x}$ is the *estimate* of μ (from the current sample). Two useful properties of an estimator are that

(i) on average the estimator gives the (true) parameter value, and

(i) the estimator has small variance.

2.3 Definition

An estimator T is *unbiased* for a parameter θ if $E[T] = \theta$.

2.4 Example

We have shown that the variance estimator S^2 is unbiased for population variance σ^2. We might therefore think that S is a good (unbiased) estimator for σ. But is this so?

Comment: The above results show that S^2 is an unbiased estimator for σ^2 and that S_*^2 is biased. This is one of the reasons for preferring S^2 as an estimator. Also, it can be shown that, when taking random samples from a population whose characteristic X follows a *normal distribution*,

$$Var(S^2) = \frac{2\sigma^4}{n-1}$$

and so $Var(S^2)$ decreases as n increases. Therefore, large samples produce more accurate estimates of σ^2 than small samples.

2.5 Example

Consider a queueing system in which we are interested in the arrival and departure rates. It is common to model the time X between arrivals by an exponential distribution with parameter θ, and probability density function

$$f(x) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Here θ represents the arrival rate to the queue. How do we estimate θ from a random sample? We know that

$$\mu = E[X] = \frac{1}{\theta}.$$

If we use the mean estimator $\bar{X}$ to estimate the population mean μ then perhaps we should estimate θ by $1/\bar{X}$. However, is this estimator a good one?

3 Comparison of Estimators

3.1 Introduction

Various *unbiased* estimators were proposed

(i) $X_1 + X_2 - 2X_3 + X_4$, variance $7\sigma^2$;

(ii) $(2X_1 + X_2 - 2X_3 + X_4)/2$, variance $5\sigma^2/2$;

(iii) $(2X_1 + X_2 + X_3 + \cdots + X_n)/(n+1)$, variance $(n+3)\sigma^2/(n+1)^2$;

(iv) $\bar{X}$, variance σ^2/n.

and it was shown that amongst unbiased estimators of the form $a_1X_1 + a_2X_2 + \cdots + a_nX_n$, the estimator with smallest variance is $\bar{X}$. Are there any unbiased estimators that are better than $\bar{X}$? If the distribution of X is symmetric about its mean (as is the normal distribution), then there are unbiased estimators of μ which are not linear combinations of the X's. We will study two such estimators

(a) the sample median, M,

(a) the sample mid-range, $MR = [\min(X_i) + \max(X_i)]/2$,

and compare their performance to that of $\bar{X}$. But before we can do this we must consider the attributes of a "good" estimator.

3.2 What Makes a Good Estimator?

The merits of an estimator T are judged by looking at its performance over all possible samples. In other words, by looking at the *sampling distribution of* T. There are many properties which characterise a "good" estimator, some theoretical and some practical. There is rarely a "best" estimator. The following properties are desirable:

(i) *Unbiasedness*
On average the estimator gives the correct answer: $E[T] = \mu$.

(ii) *Efficiency*
The estimator has small variance: $Var(T)$ is small.

(iii) *Consistency*
Larger samples give more precise estimates: $E[T] \to \mu$ and $Var(T) \to 0$ as $n \to \infty$.

(iv) *Robustness/resistance*
The estimator will perform well even if the assumed model is not quite correct or there are outlying values in the data.

(v) *Ease of calculation*
An estimator is preferred if it is easy to calculate and to understand.

For many distributions it is possible to derive the sampling distribution of an estimator theoretically. However, such techniques go beyond the scope of this course.

Part IV
Likelihood Methods

1 Introduction

In Part III, we saw that sample means and variances were good estimators for population means and variances. In most practical situations, simply being able to estimate the population mean and variance is not enough. The results of an experiment may need to be described by a probability distribution which depends on some unknown parameters. The statistical problem becomes one of how to estimate these unknown parameters. Sometimes it is obvious what estimator to use, other times it is far from clear. Consider how you would estimate the parameter in the following example.

1.1 Example

The scene is a hospital consultant's office. A patient is waiting to find out whether the consultant has detected early the onset of some disease. Fortunately, treatment after an early detection of the disease results in a cure. It is possible to detect whether the patient has the disease by waiting to see if certain symptoms appear; however, once they have, treatment is more problematic. The consultant has recently discovered that the disease is caused by the mutation of a certain type of cell; the mutation causes the cell to be larger than its non-mutated form. Healthy patients have very few mutated cells. Therefore, the consultant wants to know what proportion of mutated cells the patient has in order to detect whether they have the disease. Unfortunately, it is not easy to detect which cells are mutated and which are not as some non-mutated cells are large and some mutated cells are small. However, the mutation can be detected using very expensive equipment – too expensive to be used on a day-to-day basis. This equipment reveals that the size (in μm) of non-mutated cells follows a normal $N(50, 10^2)$ distribution and those of mutated cells, a normal $N(80, 10^2)$ distribution; see Figure 1. If the proportion of mutated cells is p then, using the Law of Total Probability, the overall distribution of cell sizes X has density

$$f(x) = pf_{mutated}(x) + (1-p)f_{normal}(x);$$

see Figure 2.

Typical histograms of (random samples of) the cell sizes of healthy and ill patients are given in Figures 3 and 4.

The problem for the clinician/statistician is that given data from a patient (such as that displayed in Figure 4), can we determine the correct value for p? Comparing this histogram with Figure 2, it looks as if $p > 0.3$ and $p < 0.5$. But can we get a more accurate answer? Obtaining the correct value of p may be crucial in deciding which treatment to give the patient.

In this part of the course, we consider a general method for estimating parameters, such as p in the above example, using likelihood methods. We begin by developing the concept of likelihood for very simple problems and then consider more complicated problems involving random samples of data.

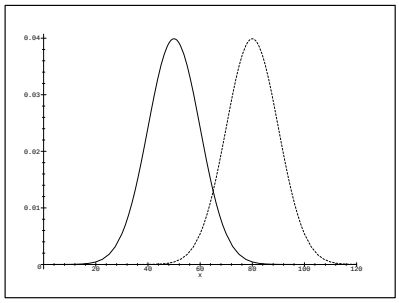


Figure 1: Distribution of cell sizes for normal cells (solid line) and mutated cells (dashed line)

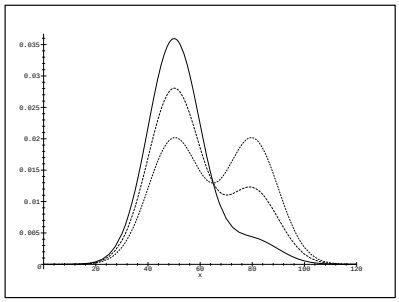


Figure 2: Distribution of cell sizes with $p = 0.1$ (solid line), $p = 0.3$ (line with long dashes) and $p = 0.5$ (line with short dashes)

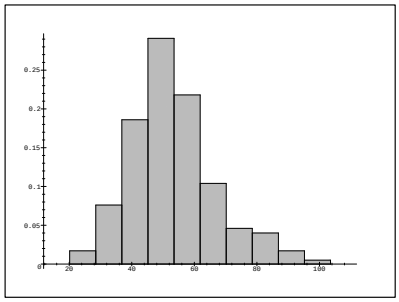


Figure 3: Distribution of cell sizes for a healthy patient

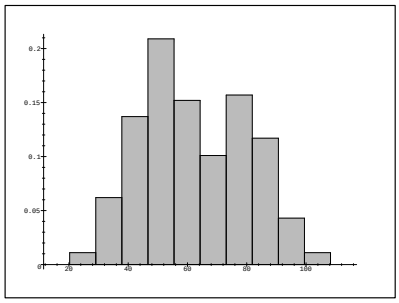


Figure 4: Distribution of cell sizes for an ill patient

1.2 Example (Single Observation)

Suppose that your car suffers from two intermittent problems, one caused by a fault in the engine (θ_1) and the other due to a fault in the gearbox (θ_2). When examined by a garage mechanic your car exhibits one of the following symptoms

$$\begin{aligned} x_1 &: \text{overheating only,} \\ x_2 &: \text{irregular traction only,} \\ x_3 &: \text{both.} \end{aligned}$$

Suppose it is known in the garage trade that these symptoms occur with the following probabilities

	O/H	I/T	Both
$Pr(X = x\|\theta)$	x_1	x_2	x_3
θ_1: fault in engine	0.1	0.4	0.5
θ_2: fault in gearbox	0.5	0.3	0.2

Construct a diagnostic rule which will help the garage mechanic to determine faults.

1.3 Example

Suppose we are interested in the proportion θ of people in Newcastle who have been to the Metro Centre in the past year. In a sample of 10 randomly chosen people, 6 responded that they had been to the Metro Centre. What value of θ is most consistent with these data?

1.4 Definition

The *likelihood function* $L(\theta|x)$ for θ is the probability (density) of observing the data, regarded as a function of θ.

1.5 Result

Suppose the data consist of a single observation x on a discrete random variable X with probability function $p(x|\theta)$ then

$$L(\theta|x) = p(x|\theta).$$

If X is a continuous random variable with probability density function $f(x|\theta)$ then

$$L(\theta|x) = f(x|\theta).$$

1.6 Definition

The *maximum likelihood estimate* (m.l.e.) for θ is any value maximising the likelihood function $L(\theta|x)$. The m.l.e. is written as $\hat{\theta}$.

1.7 Example

Consider a queueing system at a supermarket checkout. Suppose that times X between arrivals can be described by an exponential distribution with parameter θ and that we observe one such time $x = 2$. What value of θ is most consistent with this observation? What would the most likely value of θ if we had to wait x minutes before seeing the first arrival?

2 Likelihood (Random Samples)

In the examples we have looked at so far we have used only one observation to estimate the parameter. However, in most practical situations we have a random sample of observations with which to estimate the parameter. How do we combine the information in the sample to produce an estimate? The answer lies in the definition of the likelihood function. Recall that the likelihood function equals the probability (density) function of observing the data $x_1, x_2, \ldots, x_n$.

2.1 Result

The *likelihood function* $L(\theta|\mathbf{x})$ for θ given observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ on a random sample $x_1, x_2, \ldots, x_n$ is

$$\begin{aligned} L(\theta|\mathbf{x}) &= Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|\theta) \\ &= p(x_1|\theta) \times p(x_2|\theta) \times \cdots \times p(x_n|\theta) \end{aligned}$$

when the Xs are *discrete* random variables, each with probability function $p(x|\theta)$, and

$$L(\theta|\mathbf{x}) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)$$

when the Xs are *continuous* random variables, each with probability density function $f(x|\theta)$.

Our aim is to use the likelihood function to determine the most likely estimate for θ, given the data. The value of θ which maximises the likelihood function $L(\theta|\mathbf{x})$ will also be the same as the value of θ which maximises the *log-likelihood function*

$$\ell(\theta|\mathbf{x}) = \log_e L(\theta|\mathbf{x}).$$

In many cases, the calculations involved in maximising the log-likelihood function are easier than those for the likelihood function, and so we generally determine m.l.e.s using the log-likelihood function.

2.2 Example

Suppose that the numbers of arrivals at the queue for the fish counter in a supermarket in consecutive 10 minute periods are 3, 1, 3, 2, 0 and 3. If people arrive at the queue randomly (in time) then it can be shown that these observations are a random sample from a Poisson distribution. We write the data as $\mathbf{x} = (3, 1, 3, 2, 0, 3)$. Suppose we are interested in mean arrival rate θ. What is the maximum likelihood estimate of θ?

2.3 Definition

Consider an experiment which consists of n independent trials, each of which has one of r possible outcomes $E_1, E_2, \ldots, E_r$, and the probability of outcome E_i is p_i for each trial, where $\sum_{i=1}^{r} p_i = 1$. Let X_i be the number of times outcome E_i occurs in the n trials $(i = 1, 2, \ldots, r)$. The result of a typical experiment will be the number of times each outcome occurs, that is, $\mathbf{x} = (X_1, X_2, \ldots, X_r)$. The random quantity $\mathbf{x}$ has a *multinomial distribution* $M(n; p_1, p_2, \ldots, p_r)$ with index n and parameters $p_1, p_2, \ldots, p_r$, and has probability function

$$Pr(X_1 = x_1, X_2 = x_2, \ldots, X_r = x_r) = \frac{n!}{x_1!\, x_2! \cdots x_r!}\, p_1^{x_1} p_2^{x_2} \cdots p_r^{x_r}$$

for $x_1, x_2, \ldots, x_r = 0, 1, \ldots, n$ and $\sum_{i=1}^{r} x_i = n$.

Note that this distribution is a generalisation of the binomial distribution: if we have $r = 2$ possible outcomes (*success* and *failure*) then $x_2 = n - x_1$ and $p_2 = 1 - p_1$ and we obtain binomial probabilities.

2.4 Example

In a genetic experiment concerning the leaf characteristics of the Indian creeper plant *Pharbitis nil*, four different combination of leaf-types were possible. In a sample of 290 leaves the following frequencies were observed

Type	Frequency
A	187
B	35
C	37
D	31

The standard theory suggested that these types are produced independently with probabilities

$$\frac{9}{16} \;:\; \frac{3}{16} \;:\; \frac{3}{16} \;:\; \frac{1}{16}.$$

However, if this were true then we would expect the frequencies to look like

$$163.125 \;:\; 54.375 \;:\; 54.375 \;:\; 18.125,$$

and so the theory was rejected.

An alternative theory which allows for genetic linkage suggests that these types are produced independently with probabilities

$$\frac{9}{16} + \theta \;:\; \frac{3}{16} - \theta \;:\; \frac{3}{16} - \theta \;:\; \frac{1}{16} + \theta,$$

where $0 < \theta < 3/16$. If this is true, what is the most likely value for θ?

2.5 Example

We now consider the general case of the problem posed in Example 2.2. Suppose that the numbers of arrivals at the queue in n consecutive 10 minute periods are $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and that they form a random sample from a Poisson distribution with mean parameter θ. What is the maximum likelihood estimate for θ?

2.6 Example

We now consider the general case of the problem posed in Example 1.7. Suppose now that we have the times $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ between successive arrivals to the queue and that these times are a random sample from an exponential distribution with parameter θ. What is the maximum likelihood estimate for θ?

2.7 Example

Suppose we have a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from a normal $N(\theta, 1)$ distribution, with probability density function

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\theta)^2\right\}.$$

What is the maximum likelihood estimate for θ?

2.8 Example

Suppose we have a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from a Pareto(θ) distribution, with probability density function

$$f(x|\theta) = \begin{cases} \frac{\theta}{x^{\theta+1}}, & \text{if } x \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

A version of this distribution is often used to model wage distributions. What is the maximum likelihood estimate for θ?

2.9 Example

Recall that we wanted to determine the proportion of mutated cells using a random sample of cell sizes from the distribution with density

$$f(x|p) = pf_{mutated}(x) + (1-p)f_{normal}(x),$$

where $f_{normal}(x) \equiv N(50, 10^2)$ and $f_{mutated}(x) \equiv N(80, 10^2)$. The likelihood function is

$$\begin{aligned} L(p|\mathbf{x}) &= f(x_1|p) \times f(x_2|p) \times \cdots \times f(x_n|p) \\ &= \{pf_{mutated}(x_1) + (1-p)f_{normal}(x_1)\} \\ &\quad \times \{pf_{mutated}(x_2) + (1-p)f_{normal}(x_2)\} \\ &\quad \times \cdots \times \{pf_{mutated}(x_n) + (1-p)f_{normal}(x_n)\}. \end{aligned}$$

This function is rather complicated and is not particularly simplified when we take logs: the log–likelihood function is

$$\ell(p|\mathbf{x}) = \log L(p|\mathbf{x}) = \sum_{i=1}^{n} \log\{pf_{mutated}(x_i) + (1-p)f_{normal}(x_i)\}.$$

It is rather tricky to determine the maximum point as this has to be done using numerical methods. However, it is easily plotted for a given set of data. Figures 5 and 6 show the likelihood function and log-likelihood function for the data of the ill patient displayed in Figure 4. The maximum point looks to be between $p = 0.3$ and $p = 0.35$. It is, in fact, at $\hat{p} = 0.321$ (3 *d.p.*).

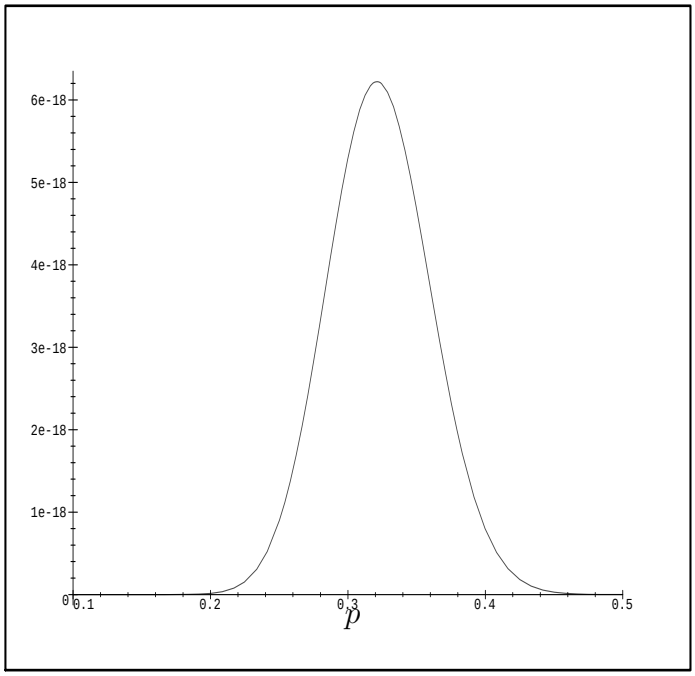


Figure 5: Likelihood function for p for the ill patient

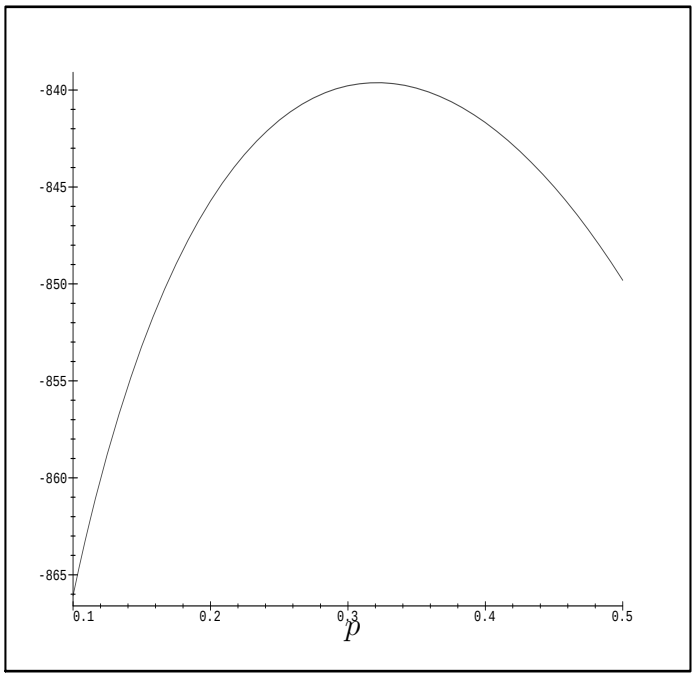


Figure 6: Log-likelihood function for p for the ill patient

2.10 Properties of Maximum Likelihood Estimators

Earlier in the course we found unbiased estimators for the population mean μ and variance σ^2. These estimators also had the property that increasing the sample made them more accurate, for example, $Var(\bar{X}) = \sigma^2/n \to 0$ as $n \to \infty$.

Maximum likelihood estimators also possess "good" properties, including

(i) they are often *unbiased* ($E[\hat{\theta}] = \theta$); if not, then they are *asymptotically unbiased*, that is

$$E[\hat{\theta}] \to \theta \quad \text{as } n \to \infty;$$

(ii) their variance decreases with increasing sample size, and in particular

$$Var(\hat{\theta}) \to 0 \quad \text{as } n \to \infty;$$

(iii) they are invariant under 1-1 transformations, that is,

$$\text{if} \quad \hat{\theta} \text{ is the m.l.e. for } \theta \quad \text{then} \quad g(\hat{\theta}) \text{ is the m.l.e. for } g(\theta)$$

Property (iii) appears to be rather technical, but in fact provides a very useful result. In Example 2.6 we calculated the m.l.e. for the arrival rate θ in a queue, assuming exponential times X between arrivals. Here $\hat{\theta} = 1/\bar{x}$. Suppose now we are interested in calculating the average time μ between arrivals. Because $X \sim Exp(\theta)$, we have

$$\mu = E(X) = \frac{1}{\theta}.$$

How should we estimate μ? Property (iii) tells us that m.l.e. for μ is $\hat{\mu} = 1/\hat{\theta} = \bar{x}$. If instead we were interested in β, the probability that times between arrivals exceed 1, then since

$$\beta = Pr(X > 1) = e^{-\theta},$$

the m.l.e. for β is

$$\hat{\beta} = e^{-\hat{\theta}}.$$

2.11 Example (A Two Parameter Problem)

The problems we have considered so far have concerned how we can determine the most likely value of a single parameter θ. In most realistic situations, the variation in the data is sufficiently complex that we need to use distributions with many more parameters. Here we give an example of how the likelihood method works when we have a random sample from a two-parameter distribution.

Suppose we have a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from a normal $N(\mu, \sigma^2)$ distribution, with probability density function

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}.$$

What are the maximum likelihood estimates for μ and σ?

2.12 Example

The straw yield data have summary statistics $n = 50$, $\bar{x} = 6.862\,kg$ and $s = 0.8456\,kg$ ($s_* = 0.8371\,kg$). Assuming that straw yields follow a normal $N(\mu, \sigma^2)$ distribution, the log-likelihood function is

$$\ell(\mu, \sigma|\mathbf{x}) = -\frac{n}{2}\log(2\pi) - 50\log\sigma - \frac{17.52 + 25(6.862 - \mu)^2}{\sigma^2}.$$

Plots of the log-likelihood surface and contours are given in Figure 7, and those for the likelihood function in Figure 8. Note that the likelihood function here has been scaled so that `Maple` produces a better looking plot – the scaling doesn't change the shape or the location of the maximum point.

How do the likelihood and log-likelihood functions depend on sample size? Suppose we obtained the same data summaries ($\bar{x}$ and s) from a sample of size n. The log-likelihood function would be

$$\begin{aligned}
\ell_n(\mu, \sigma|\mathbf{x}) &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{n\{0.8371^2 + (6.862 - \mu)^2\}}{\sigma^2} \\
&= n\left(-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{\{0.8371^2 + (6.862 - \mu)^2\}}{\sigma^2}\right) \\
&= n\ell_1(\mu, \sigma|\mathbf{x}),
\end{aligned}$$

that is, the log-likelihood from n observations is n times that from a single observation. Thinking now about the likelihood function itself, we have

$$\begin{aligned}
L_n(\mu, \sigma|\mathbf{x}) &= \exp\{\ell_n(\mu, \sigma|\mathbf{x})\} \\
&= \exp\{n\ell_1(\mu, \sigma|\mathbf{x})\} \\
&= [\exp\{\ell_1(\mu, \sigma|\mathbf{x})\}]^n \\
&= \{L_1(\mu, \sigma|\mathbf{x})\}^n,
\end{aligned}$$

that is, the likelihood from n observations is that from a single observation raised to the nth power. Figures 9 and 10 show the likelihood function for $n = 10$ and $n = 500$ respectively using the same data summaries as in Figure 8. Notice that, as the sample size increases (and hence the information we have about μ and σ), the likelihood function becomes more concentrated around its mode. In fact, as you will see in future modules, it is possible to determine the accuracy of the m.l.e.s using the curvature of the surface at its mode: the higher the curvature, the more accurate the estimates. The plots were drawn using the following `Maple` commands. Note that \ is the `Maple` continuation symbol.

```
loglik:=(mu,sigma,n)->-n*log(2*evalf(Pi))/2-n*log(sigma)\
   -n*(0.8371^2+(6.862-mu)^2)/(2*sigma^2);
with(plots);
plot3d(loglik(mu,sigma,50),mu=6.3..7.3,sigma=.5..1.5,\
   orientation=[-140,60],axes=NORMAL,style=PATCHCONTOUR);
contourplot(loglik(mu,sigma,50),mu=6.3..7.3,sigma=.5..1.5,\
   grid=[50,50]);
plot3d(exp(55+loglik(mu,sigma,50)),mu=6.3..7.3,sigma=.5..1.5,\
   orientation=[-140,60],axes=NORMAL,style=PATCHCONTOUR);
contourplot(exp(55+loglik(mu,sigma,50)),mu=6.3..7.3,sigma=.5..1.5,\
```

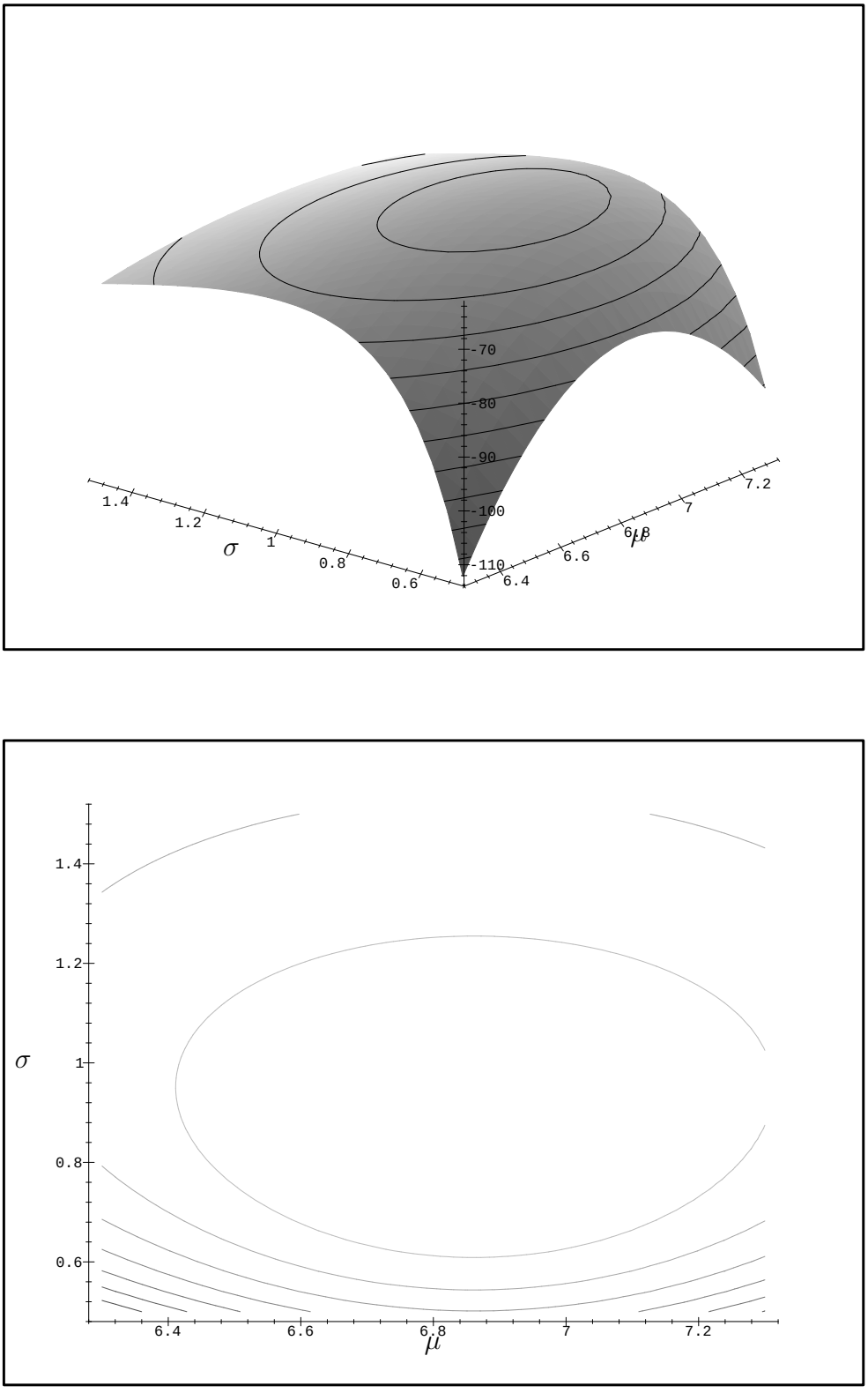


Figure 7: Log-likelihood function $\ell(\mu, \sigma|\mathbf{x})$ after observing the straw yield data

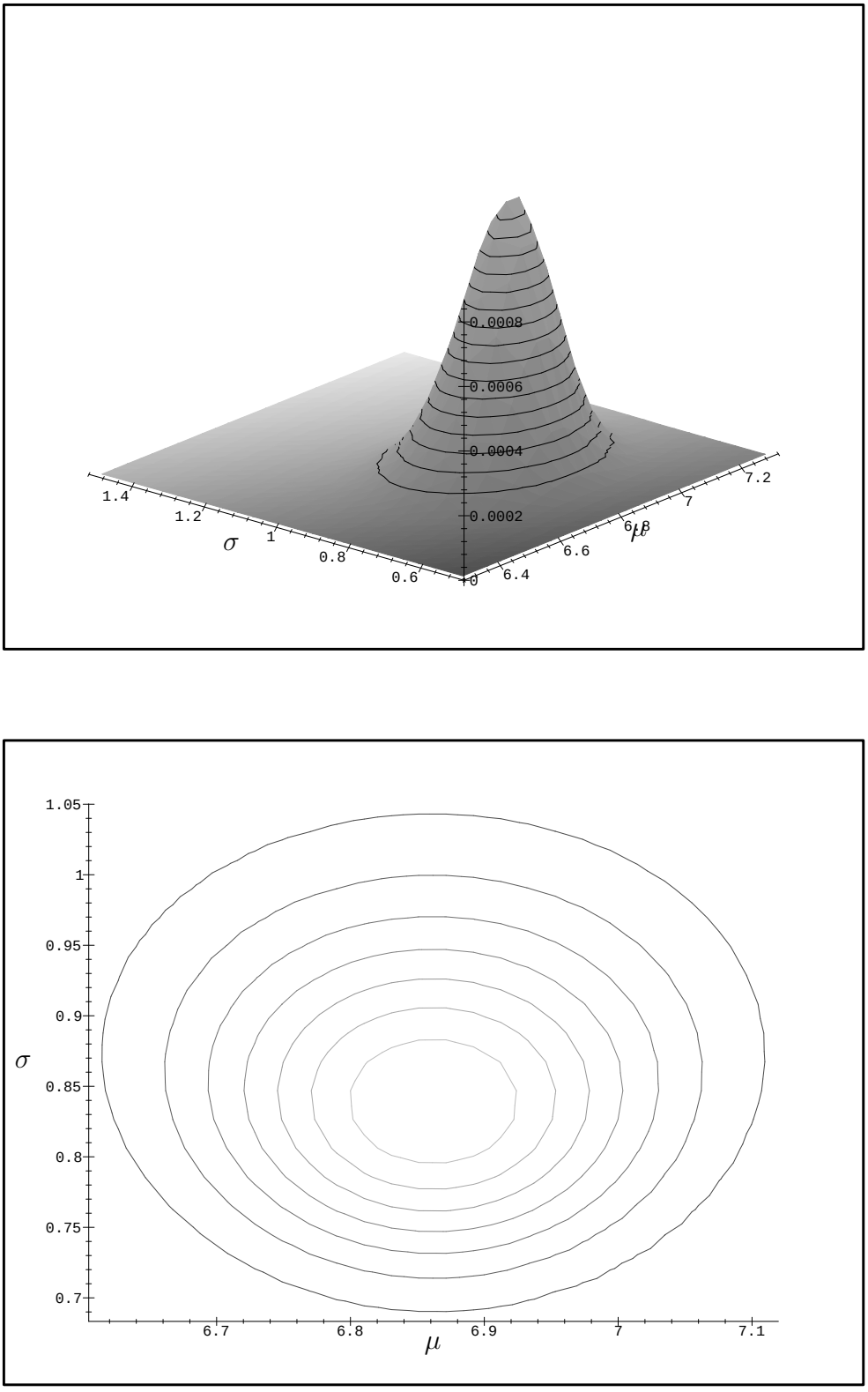


Figure 8: (Scaled) Likelihood function $L(\mu, \sigma|\mathbf{x})$ after observing the straw yield data with $n = 50$

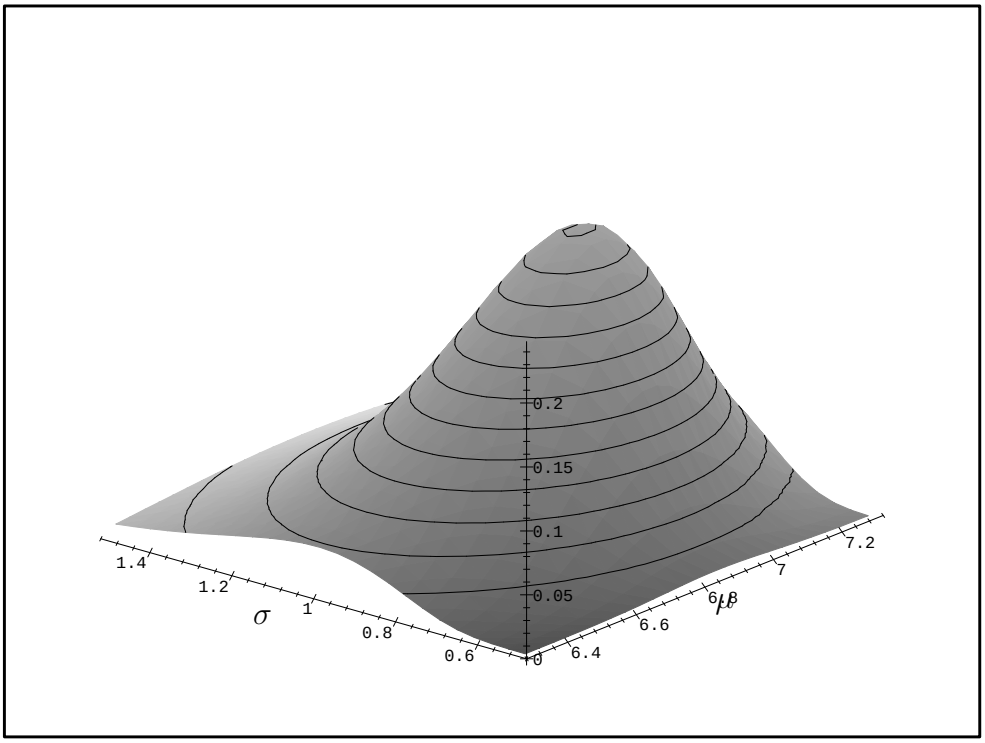


Figure 9: (Scaled) Likelihood function $L(\mu, \sigma|\mathbf{x})$ if $n = 10$

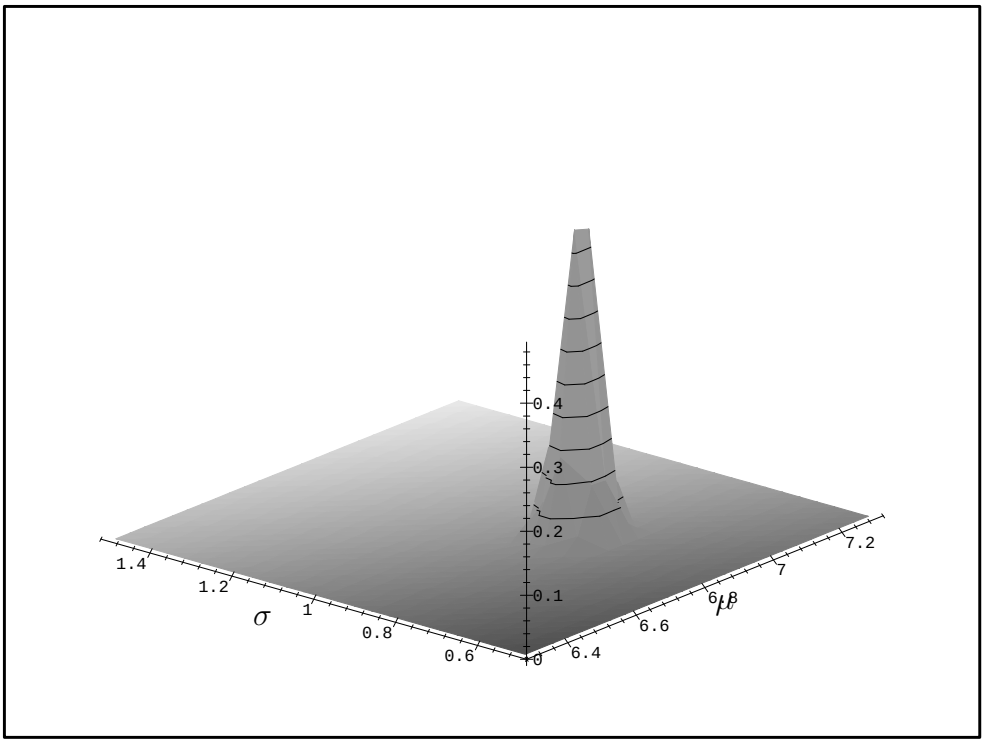


Figure 10: (Scaled) Likelihood function $L(\mu, \sigma|\mathbf{x})$ if $n = 500$

```
   grid=[50,50]);
plot3d(exp(11+loglik(mu,sigma,10)),mu=6.3..7.3,sigma=.5..1.5,\
   orientation=[-140,60],axes=NORMAL,style=PATCHCONTOUR);
plot3d(exp(620+loglik(mu,sigma,500)),mu=6.3..7.3,sigma=.5..1.5,\
   orientation=[-140,60],axes=NORMAL,style=PATCHCONTOUR);
```

(1) Suppose that a random variable Y has CDF given by

$$F_Y(y) = \begin{cases} 0, & y < 0, \\ y^2, & 0 \leq y \leq 1, \\ 1, & y \geq 1. \end{cases}$$

(i) Determine the PDF of Y.

(ii) Determine $E[Y]$.

(2) Suppose that four random variables X_1, X_2, X_3 and X_4 form a random sample from a population whose mean is 7 and variance is 2.

(i) Consider $Y_1 = X_1 - 1$;

Does $E[Y_1]$ = (a) 5, (b) 6, (c) 7, (d) 8?

Does $Var(Y_1)$ = (a) 0, (b) 1, (c) 2, (d) 3?

(ii) Consider $Y_2 = 2X_1 - X_2 - X_3 - X_4 + 7$;

Does $E[Y_2]$ = (a) 0, (b) 5, (c) 12, (d) 56?

Does $Var(Y_2)$ = (a) -7, (b) 0, (c) 1, (d) 14, (e) 21?

(3) Suppose that four random variables X_1, X_2, X_3 and X_4 form a random sample from a population whose mean is μ and variance is σ^2.

(i) Consider $Y_1 = X_1 + 2X_2 + 3X_3 + 4X_4$;

Does $E[Y_1]$ = (a) 10, (b) 10μ, (c) 0, (d) 30μ?

Does $Var(Y_1)$ = (a) 0, (b) $10\sigma^2$, (c) $30\sigma^2$, (d) 30?

(ii) Consider $Y_2 = 2X_2 - 2X_4 + 3$

Does $E[Y_2]$ = (a) 3μ, (b) $4\mu + 3$, (c) 3, (d) $4\mu + 3$?

Does $Var(Y_2)$ = (a) 0, (b) 3, (c) $4\sigma^2$, (d) $8\sigma^2$, (e) $8\sigma^2 + 3$?

(4) A random sample is taken from a population that can be described by a geometric probability model with $p_X(x) = (1-p)^{x-1}p,\ x = 1, 2, 3, \ldots$. There are four observations, namely, 3,2,1,3.

(i) Show that $L(p|\underline{x}) = (1-p)^5 p^4$.

(ii) Obtain $\ell(p|\underline{x})$ and maximise it.

Part V
Other Important Continuous Random Variables

1 Introduction

In this short section of the course, we will discuss some other important families of continuous random variables. These distributions will come into their own when we start to discuss Bayesian Statistics in the final part of the course. The first one that we will consider is the *gamma* family of random variables. The gamma distribution is a generalisation of the exponential distribution and has a wide range of applications in statistics, acturial science and engineering.

1.1 Result (The Gamma Function)

To be able to work with the gamma distribution we need to look at the gamma function which is denote by Γ and is defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx, \qquad \alpha > 0.$$

The gamma function has a number of important properties, namely,

(i) $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$,

(ii) $\Gamma(1) = 1$,

(iii) $\Gamma(\alpha) = (\alpha-1)!$ for $\alpha =$ positive integer.

2 Gamma Distribution

A continuous random variable X is called a *gamma random variable* if it has probability density function (PDF) given by

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, \qquad x > 0,$$

and $f_X(x) = 0$ otherwise, where α and λ are positive real numbers. We write $X \sim \Gamma(\alpha, \lambda)$ to denote that X has a gamma distribution with parameters α and λ.

2.1 Comments

(i) The overall shape of the gamma PDF depends on its α parameter; its λ parameter affects scale only.

(ii) If α is a positive integer then the gamma distribution can be thought of as the distribution of the time to the α^{th} event in a Poisson process with rate λ.

(iii) In the special case when $\alpha = 1$, the Gamma distribution is equivalent to the *exponential distribution* with parameter λ.

(iv) The $\Gamma(\frac{\nu}{2}, \frac{1}{2})$ distribution is also known as the *chi-square distribution* with ν degress of freedom.

(v) PDFs of gamma random variables for various choices of the parameter α and $\lambda = 2$ are given in Figure 11.

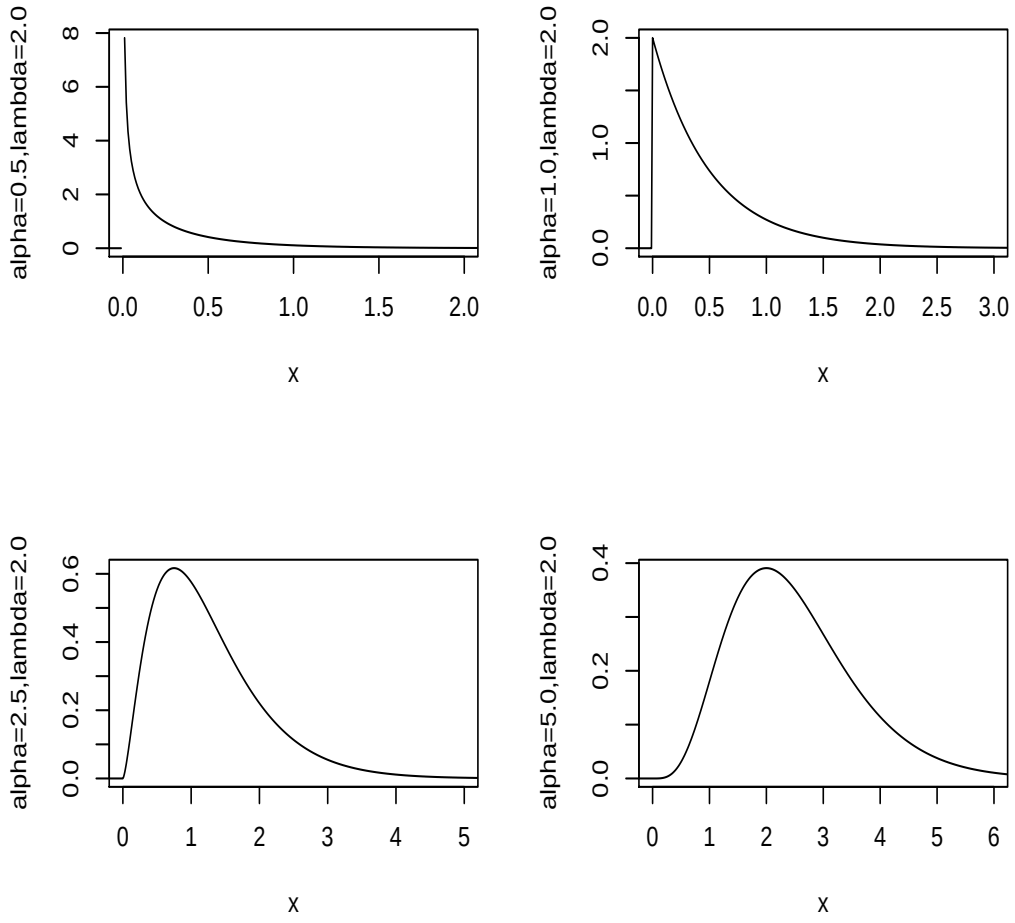


Figure 11: PDFs of gamma random variables for various choices of the parameter α and $\lambda = 2$

2.2 Example

A piece of electrical equipment has two components - one active, the other as a backup. If the first component fails, the second is automatically brought into action. Suppose that the piece of equipment is expected to be used continuously for at most 50 hours. According to the manufacturers specifications, the components are expected to fail once every 100 hours. What are the chances that the equipment would not remain functioning for the full 50 hours?

3 Beta Distribution

We now consider the family of *beta random variables*. The range of a beta random variable is the interval of real numbers between 0 and 1 which makes beta distributions particularly useful for modelling proportions, percentages or probabilities. For example, we might use a beta distribution to model

(i) the proportion of customers who are satisfied with their service each month,

(ii) the percentage of defective items in a shipment,

(iii) the percentage of data-entry errors for a particular task,

(iv) an unknown success probability in Bernoulli trials.

A random variable X is called a *beta random variable* if it has PDF given by

$$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1,$$

and $f_X(x) = 0$ otherwise, where α and β are positive real numbers and

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

is the Beta function. Note that

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

3.1 Comments

(i) Beta distributions are particularly useful because a wide variety of random phenonema can be modelled by varying the paramaters appropriately.

(ii) When $\alpha = \beta = 1$ the beta distribution is equivalent to the $U(0,1)$ distribution.

(iii) When $\alpha = \beta$ the beta distribution is symmetric about $x = 1/2$, otherwise the beta PDF is skewed.

(iv) PDFs of beta random variables for various choices of the parameter α and β are given in Figure 12.

(v) Beta and Gamma distributions (along with others) are widely used as prior distributions in Bayesian Statistics (see later).

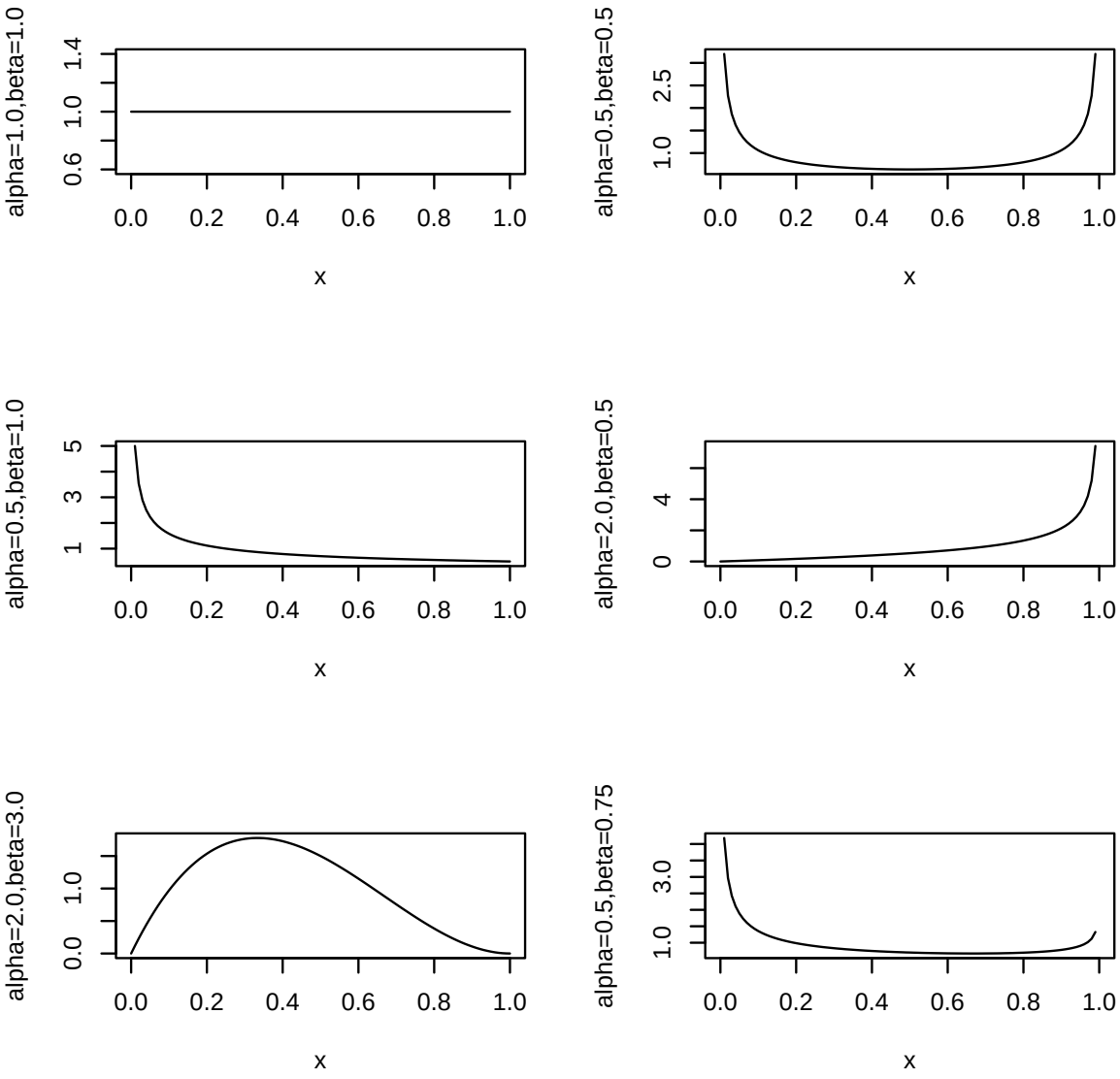


Figure 12: PDFs of beta random variables for various choices of the parameter α and β

4 Expectation and variance of Gamma and Beta random variables

The following table shows the mean and variance of gamma and beta random variables

Family	Parameters	Expected Value	Variance
Gamma	α and λ	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Beta	α and β	$\frac{\alpha}{(\alpha+\beta)}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Part VI
Introduction to Bayesian Statistics

1 Introduction

The result in which Bayesian Statistics rests - Bayes Theorem - is uncontroversial. It is simply a result in elementary probability theory. It was originally given by the Reverend Thomas Bayes (1702-61) and can be expressed in several different ways.

However, for now, let us take a step back and think about our understanding of the concept of probability.

2 Probability

Probability as a concept has been around in one form or another for a very long time. As you might expect, probability theory has developed through games of chance and gambling. The Eygptians were playing games of chance with cubical dice as early as 2000 B.C. The mathematical theory of probability was started around the 17th century when people like Galilei, Bernoulli and De Moivre tried to understand why some bets lead to the winning of more money than others. There are three main ways of understanding and thinking about probability.

2.1 Classical probability

If the outcome of an experiment must be one of n different outcomes and these outcomes are equally likely then the probability of each outcome is $1/n$.

2.2 Frequentist probability

The probability of an outcome is the long-run proportion of times that the event occurs in a large number of replications of the experiment under similar conditions. For example, if a coin is tossed 1,000,000 times and a head appears n times then

$$Pr(Head) = \frac{n}{1,000,000}$$

2.3 Subjective probability

This measures an individuals uncertainty in an event and may very form individual to individual. Your subjective probability represents your own judgement of the likelihood that the outcome will occur. You will (hopefully) base your judgement on the information that you have at the time.

2.4 Example

Which of the interpretations of probability could be used to determine the probability of the following events?

(i) The probability that England win the toss at the Boxing Day Ashes test match,

(ii) The probability that Andrew Murray wins at Wimbledon this year,

(iii) The probability that a student chosen at random was born in April.

2.5 Bayes Theorem

If $A_1, A_2, \ldots A_n$ form a partition of the sample space S and B is any event with $P(B) > 0$ then

$$\begin{aligned} P(A_i|B) &= \frac{P(B|A_i)P(A_i)}{P(B)} \\ &= \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}, \quad i = 1, 2, \ldots, n. \end{aligned}$$

2.6 Example

A virus can be one of two strains A and B, and an attempt is being made to classify the strain from the symptoms displayed by the person suffering from the virus instead of by costly medical tests. There are three symptoms and a large study of people who have caught known strains of the virus suggest the following probability model for symptom variability.

	Symptom		
Strain	Fever	Headache	Fever and Headache
A	0.4	0.2	0.4
B	0.6	0.2	0.2

On the assumption that strain B is twice as likely to occur than strain A, evaluate the probability that the virus type is A when the symptom observed is

(i) Fever,

(ii) Headache,

(iii) Fever and Headache.

3 Bayesian statistics

In Bayesian statistics we calibrate our prior information about unknown quantities by constructing a probability distribution which describes how likely we believe different values are to occur. This prior information is then combined with that from experimental data using Bayes Theorem. The key ingredients are then:

- a statistical model for the experimental data,
- quantifiable prior information about any unknown parameters.

4 Bayes Theorem for distributions

Suppose that we have a single parameter θ. Recall that the Bayesian regards θ as a random variable and that before the experiment he picks a *prior distribution* for θ with probability (density) function $\pi(\theta)$ to describe his beliefs about θ. Precise prior knowledge implies a sharply peaked prior. Vague prior knowledge gives a flatter prior.

To get information about θ we observe $x_1, x_2, \ldots, x_n$ from a distribution with probability (density) function $f(x|\theta)$. The likelihood function for θ is therefore given by:

$$L(\theta|\underline{x}) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta).$$

Our revised beliefs about θ are then given by the *posterior distribution*, which is the conditional distribution of θ given $\underline{X} = \underline{x}$. We combine both pieces of information by using the following version of Bayes Theorem.

Bayes Theorem

The posterior (density) function for θ is

$$\pi(\theta|\underline{x}) = \frac{\pi(\theta)L(\theta|\underline{x})}{f(\underline{x})}$$

where

$$f(\underline{x}) = \begin{cases} \int_\Theta \pi(\theta)L(\theta|\underline{x})d\theta & \text{if } \theta \text{ is continuous} \\ \sum_\Theta \pi(\theta)L(\theta|\underline{x}) & \text{if } \theta \text{ is discrete} \end{cases}$$

Notice that, as $f(\underline{x})$ is not a function of θ, it is often simplest to ignore it initially and use Bayes Theorem in the form

$$\pi(\theta|\underline{x}) \propto \pi(\theta)L(\theta|\underline{x})$$

discarding any factors which do not depend on θ, i.e.

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

Then use the fact that the posterior (density) function must integrate to 1 to find the normalising constant. In the case of standard distributions, the normalising constant can be inserted by inspection, if necessary.

4.1 Example

Max, a video game pirate (and Bayesian) is trying to identify the percentage of potential customers, θ who might be interested in buying "World Cup Zombie Manager" during the summer holidays. Suppose that Max believes that all values of θ are equally likely. Suppose that he asks 5 potential customers and only 1 of them would be willing to buy the game from him. Using this information, what is Max's posterior distribution?

In the previous summer, Max sold his previous game "Ashes Fever (Zombie Edition)" to 10% of the customers who came to his stall. Obtain a sensible prior distribution for the parameter θ. With this prior distribution, what is the posterior distribution?

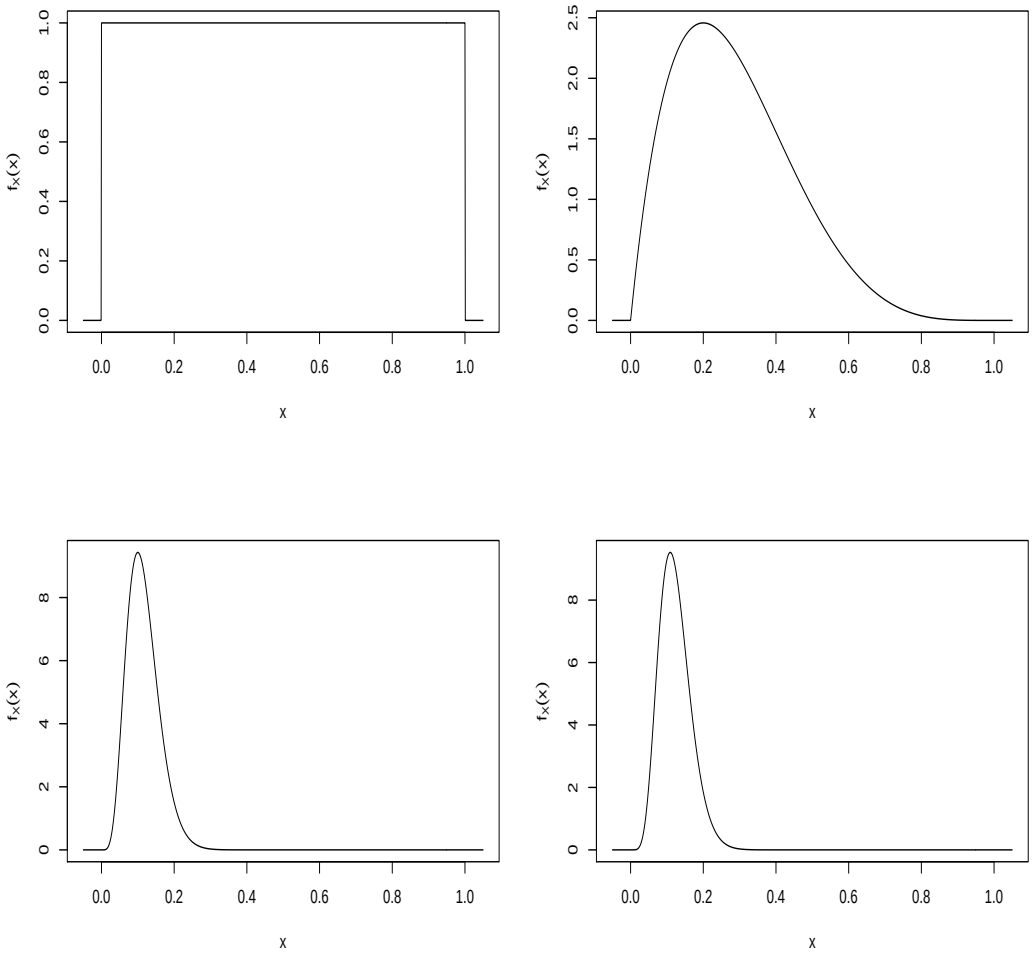


Figure 13: Priors and Posteriors for Example 4.1

4.2 Example

Let $x_1, x_2, \ldots, x_n$ be a random sample from a Poisson distribution with unknown mean θ. Suppose that we take a gamma prior $\Gamma(\alpha, \lambda)$ where α and λ are known.

Obtain the posterior distribution of θ.

4.3 Example

Let $x_1, x_2, \ldots, x_n$ be a random sample from a $N(\theta, \frac{1}{w})$, where we assume that $\frac{1}{w}$ is known. As our prior we choose a specific Normal distribution $N(\theta_0, \frac{1}{kw})$, where k can be any positive number (i.e. the prior variance is not subject to any restriction).

Obtain the posterior distribution for θ.

4.4 Example

Two physicists A and B want more accurate estimates of some physical constant θ, previously known approximately. Both observe a random variable

$$Y \sim N(\theta, 40^2)$$

i.e. they see the result of the same experiment.

Physicist A has more experience in the field of study then B. A chooses the prior

$$\theta \sim N(900, 20^2)$$

and B chooses the prior

$$\theta \sim N(800, 80^2).$$

(i) Suppose that they observe a single observation, $y = 850$. Using this information, compute the physicist's posterior distributions.

(ii) Suppose that 100 independent observations of Y are taken and that $\bar{y} = 870$. What are the posterior distributions now?

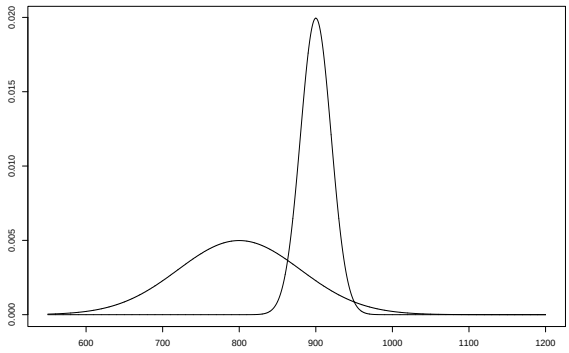


Figure 14: Priors for Example 4.4

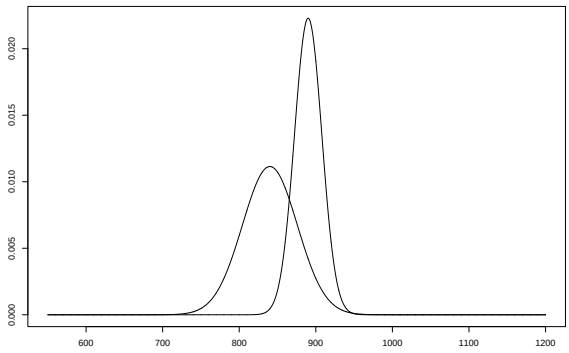


Figure 15: Posteriors for Example 4.4(i)

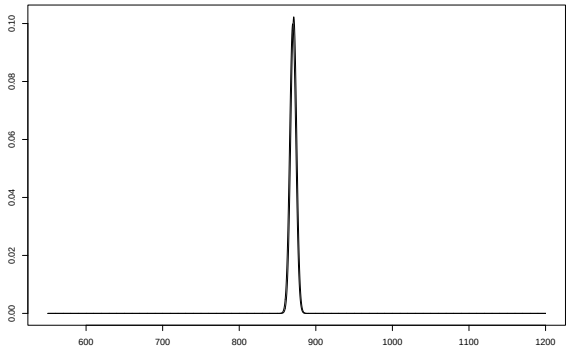


Figure 16: Posteriors for Example 4.4(ii)

Point and Interval Estimates

Since the posterior incorporates all the available information, the informative conclusion to an experiment is to state the posterior, or to provide a graphical representation. If we attempt to summarise the posterior distribution, we will inevitably waste information. However, if a summary is necessary, point or interval estimates can be found.

4.5 Point Estimates

When a point estimate is required for the unknown parameter we will often use the mean, median or mode of the posterior distribution.

4.6 Interval Estimates

An interval estimate is often a more useful way of summarising the posterior distribution as it reflects the variation of the distribution. In the Bayesian framework, a confidence interval is a conceptually simple idea.

4.7 A Bayesian Confidence Interval

A $100(1-\alpha)\%$ Bayesian Confidence Interval for θ is any region (a, b) such that

$$\int_a^b \pi(\theta|\underline{x})d\theta = 1-\alpha.$$

Bayesian confidence intervals are sometimes called *credible regions* or *plausible regions*. Clearly these regions will not be unique, since there will be many intervals with the correct probability coverage for a given posterior distribution.

4.8 A Highest Density Interval (H.D.I)

A $100(1-\alpha)\%$ Highest Density Interval is a Bayesian Confidence Interval which also has the property that for $\theta_1 \in (a, b)$ and $\theta_2 \notin (a, b)$, $\pi(\theta_1|\underline{x}) \geq \pi(\theta_2|\underline{x})$.

4.9 Example

Suppose that the posterior distribution for θ is (i) a $Beta(1, 24)$; (ii) a $Beta(2, 23)$ distribution. In each case, obtain a 95% H.D.I for θ.

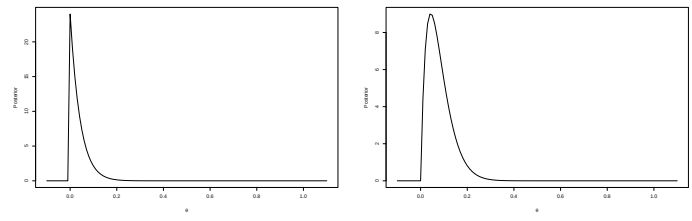


Figure 17: Posteriors for Example 4.9

Examples

(i) Telephone calls are received at a telephone switchboard at a constant rate. Let X denote the number of calls received per day.

(ii) Suppose that we want to model the percentage of customers who would want to receive email correspondence from a bank. Let X denote our beliefs about the value of that percentage.

(iii) A manufacturing process makes electrical components. If 5% of the components are defective, and a batch of 1000 components is taken, let X denote the number of defective items.

(iv) A post office opens at 9am. Customers arrive at a constant rate. Let X be the time until the first customer arrives at the shop.

(v) A person suffering from a recurrent illness will be put on medication after their third bout of the illness. The rate at which bouts occur is assumed to be constant. Let X denote the be the time until the patient is put on medication.

(vi) A commuter train arrives punctually at a station every half hour. Each morning John leaves his house and casually strolls to the train station. Let X denote the time, in minutes, that John has to wait for the train from the time he reaches the station.

(vii) A six-sided die is rolled repeatedly. Let X denote the number of rolls until the first six is obtained.

Summary of Discrete Distributions

Distribution	PMF	$E[X]$	$Var(X)$
Binomial	$X \sim Bin(n,\theta) \Rightarrow P(X=x) = \binom{n}{x}\theta^x(1-\theta)^{n-x},\ x=0,1,2,\ldots,n$	$n\theta$	$n\theta(1-\theta$
Poisson	$X \sim Poisson(\lambda) \Rightarrow P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!},\ x=0,1,\ldots,\ \lambda>0$	λ	λ
Geometric	$X \sim Geometric(p) \Rightarrow P(X=x) = (1-p)^{x-1}p, \qquad x=1,2,\ldots,$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$

Summary of Continuous Distributions

Distribution	PDF	$E[X]$	V
Uniform	$X \sim U(a,b) \Rightarrow f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$	$\frac{a+b}{2}$	
Exponential	$X \sim Exp(\lambda) \Rightarrow f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0, \\ 0, & \text{otherwise.} \end{cases}$	$\frac{1}{\lambda}$	
Normal	$X \sim N(\mu, \sigma^2) \Rightarrow f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ x \in \mathbb{R}, \qquad \mu \in \mathbb{R}$ and $\sigma > 0.$	μ	
Gamma	$X \sim \Gamma(\alpha, \lambda) \Rightarrow f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \qquad x, \alpha, \lambda > 0$	$\frac{\alpha}{\lambda}$	
Beta	$X \sim Beta(\alpha, \beta) \Rightarrow f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1, \alpha, \beta > 0$	$\frac{\alpha}{(\alpha+\beta)}$	$\frac{}{(\alpha+\beta}$

(Bx) The number of scratches per item for $n = 150$ newly manufactured items was recorded. The observations are regarded as a sample from a Poisson distribution, with mean θ. Also suppose that a $\Gamma(\alpha, \lambda)$ random variable, with probability density function (PDF) given by

$$\pi(\theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda\theta} \qquad \theta, \alpha, \lambda > 0,$$

with $\alpha = 40$ and $\lambda = 10$ is chosen as the prior distribution for θ.

(a) What is the prior mean and variance?

(b) Obtain the likelihood function, $L(\theta|\underline{y})$.

(c) What is the posterior distribution $\pi(\theta|\underline{y})$?

(d) What is the posterior mean and variance? Show that the posterior mean can be written as a weighted average of the prior mean and the sample mean.

(e) Show that any value of $\bar{y} > 4$ would lead to the posterior mean being greater than the prior mean?