Mathematics 206 Probability & Statistics II

Professor Peter M. Higgins

October 31, 2018

This unit continues our study of probability and statistics topics that began in MA105, with both fields featuring in equal measure. We begin with some probability *ballot and random walk* problems that introduce the *Reflection principle*. Set 2 introduces the basic notion of *maximum likelihood estimation*, which is a fundamental technique used for estimating parameters from observations of a random variable. Set 3 introduces *significance testing* and *confidence intervals* based on facets of the *normal distribution*. Set 4 introduces the study of *joint distributions* and the associated ideas of *marginal and conditional distributions* and the *covariance* between two random variables.

Set 5 introduces the *chi-squared goodness of fit statistic* and Students *t-distribution* for data with unknown parameters. Set 6 introduces another staple of statistical methods, that being the method of *Least squares regression*. Set 7 is about the *Pearson regression coefficient* that is used to estimate the linear correlation between data sets.

In the final three sets we return to probability theory, beginning in Set 8 with *Markov chains*, which describe *memoryless stochastic processes*, where the probability of passage to any particular state depends only on the current state (and not on the history of the process before that). In Set 9 we introduce the z-transform, which here is used as an alternative to the *Moment generaring function* of a random variable that we met in the problems of MA105. The final set develops the basic theory of the sum of two independent random variables.

Problem Set 1 Birthday and ballot problems

1. *Birthday problem*

(a) How many people are required before the chances that two or more of them were born on the same day of the week?

(b) Answer the question again for having a birthday coincidence. (Leap years intrusions make little difference so just assume all years have 365 days.)

2. *Ballot problems* Two election candidates A and B poll p and q votes respectively with $p > q$. The votes are counted and we draw a *path* for the count: each stage is represented by a point $(n+m, n-m)$ where A has n votes and B has m votes at this point. We draw a line between each such point and the next (which will be either $(n+m+1, n-m+1)$ or $(n+m+1, n-m-1)$) according as the next vote is for A or for B. Characterize those paths representing vote counts where:

(a) A leads right through the count;

(b) A never falls behind in the count.

(c) What is the total number of possible paths representing counts of this vote?

3. Consider the reverse count (in which the votes are counted in the reverse order to some given count). Characterize the nature of the counts which are the reverse of types (a) and (b) of Question 2.

4. (*Reflection principle*). Let a, b and c be positive integers. Show that the number of paths from $(0, a)$ to (b, c) that touch or cross the x-axis equals the total number of paths from $(0, -a)$ to (b, c).

5. What is the total number of paths from $(0, -a)$ to (b, c) $(a, b, c \in \mathbb{Z}^+)$?

6. What is the probability that B leads at some point in the count?

7. What is the probability that A leads the count all the way through?

8. Explain why the answer to Question 7 is 1−P(the count is tied at some point) and deduce the answer using this.

Random walks Beginning at the origin we make moves of one unit either up or down with probability $\frac{1}{2}$ for each of these possibilities. The evolution of the walk may be recorded as a path in the fashion used above for ballot problems.

9(a) After n steps the coordinates of the endpoint of our path is (n, y) say. Show that $n \equiv y \pmod 2$.

(b) For a random walk, find the probability u_{2n} of returning to the origin (perhaps not for the first time) after $2n$ steps.

10. Show that u_{2n} is also equal to the probability of no return to the origin at any time during the first $2n$ steps.

Problem Set 2 Maximum likelihood estimators

Maximum likelihood estimators Let $x_1, x_2, \cdots, x_n$ represent a random sample from n i.i.d. (*independent and identically distributed*) random variables $X_1, X_2, \cdots, X_n$ with common distribution pdf $f(x;\theta)$ for some unknown parameter θ. The *maximum likelihood estimate* $\hat{\theta}$ *of* θ is $\hat{\theta} = u(X_1, X_2, \cdots, X_n)$ where $\theta = u(x_1, x_2, \cdots, x_n)$ maximizes the *likelihood function*;

$$L(\theta; x_1, x_2, \cdots, x_n) = \Pi_{i=1}^{n} f(x_i;\theta).$$

Since the likelihood function involves products, it is often better to work with $\log L$, which has the same maximum. Find $\hat{\theta} = u(x_1, x_2, \cdots, x_n)$, a function u of the given observations, for each of the following distributions.

1. $N(\theta, 1)$ a normal distribution with unknown mean θ so that

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} \ (-\infty < x < \infty).$$

2. A Poisson distribution,

$$f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!},\ x = 0, 1, 2, \cdots,\ 0 \le \theta < \infty,\ \text{zero elsewhere, where } f(0;0) = 1.$$

3.
$$f(x;\theta) = \theta x^{\theta-1},\ 0 < x < \infty,\ 0 \text{ elsewhere.}$$

4.
$$f(x;\theta) = (1/\theta) e^{-x/\theta},\ 0 < x < \infty,\ 0 \text{ elsewhere.}$$

5.
$$f(x;\theta) = e^{-(x-\theta)},\ \theta \le x < \infty,\ -\infty < \theta < \infty.$$

6.
$$f(x;\theta) = \frac{1}{2} e^{-|x-\theta|},\ -\infty < x < \infty,\ -\infty < \theta < \infty.$$

7.
$$f(x;\theta) = \frac{1}{\theta},\ 0 < x \le \theta,\ 0 < \theta < \infty.$$

8. Show for the example of Question 7 that $E[\max(X_i)] = \frac{(n+1)\theta}{n}$ and so $\hat{\theta}$ is a biased estimate of θ in this case.

9. The uniform distribution,

$$f(x;\theta) = 1,\ \theta - \frac{1}{2} \le x \le \theta + \frac{1}{2},\ 0 \text{ elsewhere } -\infty < \theta < \infty.$$

10.

$$f(x;\theta) = \theta^x (1-\theta)^{1-x},\ 0 \le x \le 1,\ 0 \le \theta \le 1,\ 0 \text{ elsewhere, } f(0;0) = f(1;1) = 1.$$

Problem Set 3: Significance testing

1. A random variable X is such that $X \sim N(\mu, 100)$. A random observation x of X yields $x = 172$. Test the null hypothesis $H_0 : \mu = 150$ against $H_1 : \mu \neq 150$ at the 5% level.

2. A distributor claims that 90% of the seeds sold can be expected to germinate. In a sample of 100 seeds, 83 germinate. Use the normal approximation to the binomial distribution to test the claim at the 5% level.

3. A machine produces piping measured in centimetres according to a normal distribution $N(420, 12^2)$. After maintainance a sample of 100 pipes have a mean length of 423cm. Is there evidence at the 5% level of a change in the mean length of pipes produced?

4. It is claimed that machine parts are produced with a mean mass of 6g and a standard deviation of $0 \cdot 8$g. What is the range of possible values of the mean mass obtained from a random sample of 50 for this claim to be accepted at the 5% level?

5. A coin will be accepted as fair if, in 120 tosses, the number of heads does not lie outside the range of 50 to 70.

(i) Find the probability of *Type I error*, which is rejecting the hypothesis when it is correct.

(ii) Given that the coin is biased with head probability of $0 \cdot 6$, find the probability of *Type II error*, which is accepting the null hypothesis when it is false.

6. Suppose that X is a Poisson random variable with mean λ and that $H_0 : \lambda = 6 \cdot 5$ and $H_1 : \lambda \neq 6 \cdot 5$.

(i) If $x = 2$, test this at the 5% level.

(ii) With the same observation, test H_0 against $H_1 : \lambda < 6 \cdot 5$.

7. A loading machine breaks down on average 3 times per month. The company that leases the machine will refund the clients money if it breaks down n times in a year. Use a normal approximation to a Poisson distribution to find the least value of n that ensures that the probability of having to make a refund is less than 2%.

8. *The sign test* A teacher seeks to justify her claim that her students performed better in the final examination than they did in the 'mock' examination. She looks at their marks and writes a + or − according as the student performed better in the final examination or did not perform better. The result was:

$$-++--+++++++++.$$

Use the binomial distribution to test at the 5% level the null hypothesis H_0 : there is no difference between the marks, against H_1 : the students did better on their final examination.

If $X_i \sim N(\mu_i, \sigma_i^2)$ $(i = 1, 2)$ then $\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$. When the variances are *known* then the statistic for testing significant difference between the means is

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

When testing for μ with σ unknown and n large we use as test statistic:

$$Z = \frac{\overline{X} - \mu}{\hat{\sigma}/\sqrt{n}},\ \hat{\sigma}^2 = \frac{nS^2}{n-1},\ \text{where } S^2 = \tfrac{1}{n}\textstyle\sum(X - \overline{X})^2.$$

9. Random samples of respective sizes 100 and 80 are taken from two normal populations with variances $\sigma_1^2 = 40$ and $\sigma_2^2 = 30$. The respective sample means are $\overline{x}_1 = 38 \cdot 3$ and $\overline{x}_2 = 40 \cdot 1$. Test at the 5% level for a significant difference in the (unknown) populations means $\mu_1$and μ_2.

10. A particular normal random variable has had in the past a mean of 50. A random sample of size 100 gave a mean of $52 \cdot 6$ and a standard deviation of $14 \cdot 5$. Is there evidence at the 1% level that the mean has increased?

Problem Set 4: Joint distributions and Covariance

When discussing general results, we take our random variables to be continuous so that formal calculations will be in terms of integrals (as opposed to summation notation).

1. Show that

$$f(x,y) = 2 - x - y,\ 0 < x < 1,\ 0 < y < 1$$

and 0 elsewhere is a probability density function and find *marginal distribution*s $f_1(x)$ and $f_1(y)$ respectively.

2. Repeat Question 1 for

$$f(x,y) = \frac{xy}{2},\ 0 < x < 2,\ 0 < y < x.$$

3. The *conditional distribution of* Y given X is defined as that with density

$$f(y|x) = \frac{f(x,y)}{f_1(x)},$$

where $f(x,y)$ is the joint density function of X and Y and $f_1(x)$ is the density function of X. Show that $f(y|x)$ is a (probability) density function.

4. Let $f(x,y) = 2$ for all $0 < x < y < 1$ and be 0 elsewhere. Find the conditional distributions $f(x|y)$ and $f(y|x)$.

5. For the random variables X and Y of Question 2, find $\mu_{Y|x} = E(Y|X)$.

Two random variables X and Y are *(stochastically) independent* if their joint density function $f(x,y)$ has the form $g(x)h(y)$.

6. Show that if X and Y are independent then

$$P(a \le X \le b, c \le Y \le d) = P(a \le X \le b)P(c \le Y \le d).$$

7. Show that for any two random variables, X and Y, $E(X+Y) = E(X) + E(Y)$.

The *covariance* and *correlation* ρ of two random variables X and Y are defined as

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y),\ \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{(\mathrm{Var}X)^{\frac{1}{2}}(\mathrm{Var}Y)^{\frac{1}{2}}}$$

8. Show that $\mathrm{Cov}(X,Y) = E((X - EX)(Y - EY))$.

9. Show that $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X,Y)$.

10. By considering the discriminant of $\mathrm{Var}(X + tY)$ (regarding it as a quadratic function in the free real variable t) and using Question 9 show that $-1 \le \rho(X,Y) \le 1$.

Problem Set 5: χ^2 - and t- tests

If $n_1, \cdots, n_k$ and $e_1, \cdots, e_k$ represent observed and expected frequencies respectively for the k possible outcomes of an experiment performed k times, then the limiting distribution of the statistic:

$$\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$

is χ^2 on $k-1$ degrees of freedom. If the cell probabilities have unknown parameters then the corresponding statistic is χ^2 on $k-2$ degrees of freedom if the unknown parameters are replaced by their maximum likelihood estimators.

1. An experiment yield molecules of four types A, B, C and D in respective quantities $120, 48, 36$ and 18. Theory predicts that the expected ratios should be $9:3:3:1$. Are the experimental results consistent with this theory?

2. Fitness levels of police recruits are categorized in increasing order of fitness into five categories. The categories have been set up from long-standing data so that each category is expected to be equally represented. A group of 200 recruits are tested and the numbers that fall into each category (from least fit to most fit) are respectively $54, 44, 40, 35, 27$. Is it correct to conclude that this cohort is inferior in terms of fitness?

3. A point is chosen at random from the unit interval according to the pdf $f(x) = 2x$. Let A_1, A_2, A_3, A_4 be the four quarter-intervals of $[0, 1]$ in the natural order. A random sample of 80 choices hits the four intervals $5, 15, 25$ and 35 times respectively. Are these observation consistent at the 0.025 level with the given model?

4. According to *Mendelian inheritance* offspring of a certain plant crossing should be red, black, or white in the ratios $9:3:4$. If an experiment gave 70, 36, and 38 offspring of each respective colour, would that contradict the theory?

5. According to the *Hardy-Weinberg formula*, the number of flies resulting from certain breeding conditions should be in proportions $q^2 : 2pq : p^2$ where $p + q = 1$. In an actual experiment the observed frequencies were 40, 50 and 20. Is this consistent with the theory given that q is estimated by the maximum likelihood estimate of

$$\hat{q} = \frac{n_1 + \frac{n_2}{2}}{n_1 + n_2 + n_3}?$$

Confidence interval of p for σ^2 sampled from an normal distribution is:

$$\frac{nS^2}{\chi_2^2} < \sigma^2 < \frac{nS^2}{\chi_1^2}$$

where the probablility that that χ^2 on $n-1$ degrees of freedom exceeds χ_2^2 is $(1-p)/2$ and the probability that χ^2 is less than χ_1^2 is also $(1-p)/2$.

6. Suppose that sample values based on size 5 were $s_1^2 = 237$, $s_2^2 = 320$, $s_3^2 = 853$, $s_4^2 = 296$, $s_5^2 = 141$.

(a) How many degrees of freedom are possessed by the sum $\sum \frac{n_i S_i^2}{\sigma^2}$?

(b) From the appropriate χ^2 table, find that $\chi_1^2 = 9 \cdot 237$ and $\chi_2^2 = 35 \cdot 02$ and hence find a 96% confidence interval for σ^2.

Student's T-statistic If $X \sim N(\mu, \sigma^2)$. Let $\overline{X}$ and S be the sample mean and sample variance based on a random sample of size n. Then

$$T = \frac{(\overline{X} - \mu)\sqrt{n-1}}{S}$$

has a *t-distribution on* $n-1$ *degrees of freedom.* In order to compare the difference of two means $\overline{X}$ and $\overline{Y}$ then T has the t-distribution on $\nu = n_X + n_Y - 2$ degrees of freedom where

$$T = \frac{\overline{(X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{n_X S_X^2 + n_Y S_y^2}} \sqrt{\frac{n_X n_Y (n_X - n_Y - 2)}{n_X + n_Y}}.$$

7. Ten patients test a new drug meant to allow them to sleep for longer. The results are tabled below.

Patient	1	2	3	4	5	6	7	8	9	10
Hours gained	0.7	-1.1	-0.2	1.2	0.1	3.4	3.7	0.8	1.8	2.0

Test the null hypothesis of no gain in sleep against the alternative of a positive gain.

8. Given a 95% confidence interval for μ for the data of Question 7.

9. The yield of corn in bushels/plot on 20 plots is as follows:

Treated	6.2	5.7	6.5	6.0	6.3	5.8	5.7	6.0	6.0	5.8
Untreated	5.6	5.9	5.6	5.7	5.8	5.7	6.0	5.5	5.7	5.5

Find the observed means $\overline{x}$ and $\overline{y}$ and the sample standard deviations $n_X s_X^2$ and $n_Y s_Y^2$.

10. Hence compute the T statistic for the data of Question 9 and so test the null hypothesis $H_0 : \mu_X = \mu_Y$ against the alternative $H_1 : u_X > \mu_Y$.

Problem Set 6: Least Squares Regression

The *Least Squares Regression Line* for y on x (with a data set of order n) is $y = a + bx$ where $\sum y = na + b\sum x$ and $\sum xy = a\sum x + b\sum x^2$.

1. Find the equations for the regression line of x on y and y on x and show that the two regression lines meet at $(\overline{x}, \overline{y})$.

2. By solving the normal equations, calculate the equation of the least squares regression line for y on x for the data set:

x	1	2	4	6	7	8	10
y	10	14	12	13	15	12	13

3. Repeat Question 2 for the regression line of x on y.

4. For a given data set we have $\overline{x} = 10$ and $\overline{y} = 4$ with the gradient of the Regression line $0 \cdot 6$. Find the equation of the line and estimate y when $x = 12$.

5. The *covariance* of a data set $(x_1, y_1), \cdots, (x_n, y_n)$ is

$$s_{xy} = \frac{1}{n}\sum(x - \overline{x})(y - \overline{y}).$$

(Similarly $s_x^2 = \frac{1}{n}\sum(x - \overline{x})^2$.) Show that $s_{xy} = \frac{\sum xy}{n} - \overline{xy}$.

6. Prove that least squares regression line of y on x can be expressed as:

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x}).$$

7. A theatre manager records the following information for the performance of a play over 12 nights.

x = Advanced sales	53	10	79	59	33	13	72	69	22	45	36	18
y = door sales	71	32	98	79	49	33	91	84	42	63	54	37

Obtain the equation of the regression line using the formula you derived in Question 6, working to 3 decimal places. ($\sum x^2 = 27,963$, $\sum xy = 37,249$).

8(a) For the data set of Question 7 estimate the audience figure if the theatre takes 82 advanced ticket sales.

(b) Given that the venue can seat only 120, what is the maximum number of advanced tickets sales that can be sold before the regression line predicts that customer might have to be turned away on the night?

9. *Derivation of Least Squares formula* Let $\hat{y}_i = a + b(x_i - \overline{x})$ and consider $Q = \sum(y_i - \hat{y}_i)^2$. Find $\frac{\partial Q}{\partial a}$ and hence show that $a = \overline{y}$.

10. With the known value of a, find $\frac{\partial Q}{\partial b}$ show that the optimum value of b is $b = \frac{s_{xy}}{s_x^2}$, all in accord with Question 6.

Problem Set 7: Pearson Regression coefficient

Also known as the *product-moment relational coefficient,* $r = \frac{s_{xy}}{s_2 s_y}$ is a measure of the degree of scatter of a bivariate data set; r measures the degree of *linear* correlation (how the data matches a straight line).

1. Express r explicitly in the symbols $a_i = x_i - \overline{x}, b_i = y - \overline{y}$.

2. Consider the inequality $\sum(za_i + b_i)^2 \geq 0$, where z is a real variable. Show, by considering the associated quadratic equation, that if this inequality is strict then $-1 < r < 1$.

3. Show that the sum is identically zero implies that $r = \pm 1$.

4. Ten candidates tooks exams in Mathematics and in Physics. Find the value of the relational correlation coefficient r and comment on the degree of correlation.

Mathematics (x)	18	20	30	40	46	54	60	80	88	92
Physics (y)	42	54	60	54	62	68	80	66	80	100

5. Find the value of r for the data set of Question 7 of Set 6.

6. Let $y = ax + b$ and $x = cy + d$ be the respective regression lines of y on x and x on y.
(a) Show that $bd = r^2$.
(b) Show that b and d have the same sign and that, if non-zero, r shares that sign.

7. Apply the results of Question 6 to the data of Question 2 of Set 6 to find the value of r and comment on its strength.

8. Show that if $r = 0$ then the two regression lines are orthogonal.

9. *Testing for no correlation* The test statistic $t = r\sqrt{\frac{n-2}{1-r^2}}$ is approximately t-distributed on $n-2$ degrees of freedom. Test the null hypothesis $H_0 : \rho = 0$ for the data set of Question 7 of Set 6.

10. Find a 95% confidence interval for the correlation in the data of Question 4.

Problem Set 8: Markov chains

A *Markov process* M is one where a system consists of a collection of states $\{E_i\}$ and, given the system is in state i, there is a fixed probability p_{ij} that it passes to state j during the next time period. When the set of states is finite and the time units used are discrete (as opposed to continuous) we speak of a *Markov chain*. In this case we may represent M as an $n \times n$ *transition matrix* P with entries p_{ij}.

1. Show that each of the rows of P sum to 1.

2. *Chapman-Kolmogorov* equations Let $x^{(0)} = (x_1, \cdots, x_n)$ be an initial list of probabilities that the system is in state i $(1 \leq i \leq n)$ and let $x^{(n)}$ be the corresponding list of probabilities after the passing of n time units. Show that $x^{(n)} = x^{(0)} P^n$.

3. Suppose that $x = \lim_{n\to\infty} x^{(n)}$ exists; we call x the *steady state vector* or the *stationary distribution* of the process. Explain why x is an eigenvector of P with eigenvalue 1.

4. Simple weather model. All days are labelled either *dry* (no precipitation) or *wet* otherwise. If today is dry (resp. wet) then the probability that tomorrow is dry (resp. wet) is $0 \cdot 8$ (resp. $0 \cdot 6$).

(a) Write down the transition matrix of this process.

(b) Given that Wednesday is dry, what is the probability that Sunday is likewise.

(c) Repeat part (b), replacing dry by wet.

(d) What is the long term proportion of dry days?

5(a) Two power companies Pow and Zap compete for customers. Each year Pow retains 75% of its customers, with the remainder switching to Zap. Similarly Zap loses 10% of its customers annually to Pow. Show that, in the long run, 5 out of 7 people will be with Zap.

(b) In general, for non-zero α, β, find the stationary distribution of the Markov chain with transition matrix:

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}.$$

6. Find the steady state vector of the Markov chain with transition matrix:

$$P = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{5} & \frac{3}{5} & \frac{1}{5} \end{bmatrix}.$$

7. It can be shown that if a Markov chain has a stationary distribution with no zero entries, $\pi = (\pi_i)$, then the mean number of steps, m_{ii}, for the process to return to a given state E_i from E_i is π_i^{-1}.

(a) Let m_{ij} denote the mean number of time steps to move from E_i to E_j in such a Markov chain. Show that

$$m_{ij} = 1 + \sum_{k \neq j} p_{ik} m_{kj}.$$

(b) Show that for a two-state Markov chain we have for $i \neq j$:

$$m_{ij} = \frac{1}{1 - p_{ii}}.$$

(c) Hence calculate the matrix $M = [m_{ij}]$ for the general Markov chain of Question 5(b).

(d) Calculate M explicitly for Question 5(a) and state how often a Zap customer will, on average, switch to Pow.

8. *First return times* Let $f_{jj}^{(n)}$ denote the probability that, starting from state E_j, the system first returns to E_j occurs at the nth step. Writing $p_{ij}^{(n)}$ for the probability of passing from E_i to E_j after n steps, prove by induction on n that

$$f_{jj}^{(n)} = p_{jj}^{(n)} - \sum_{m=1}^{n-1} f_{jj}^{(m)} p_{jj}^{(n-m)}.$$

9. Apply the result of Question 8 to find the values of the $f_{jj}^{(2)}$ for the Markov chain of Question 4.

10. Consider the Markov chain M with transition matrix:

$$P = \begin{bmatrix} 0 \cdot 5 & 0 \cdot 5 & 0 \\ 0 & 0 \cdot 3 & 0 \cdot 7 \\ 0 & 0 \cdot 6 & 0 \cdot 4 \end{bmatrix}.$$

(a) Find the value of $F_{11} = \sum_{n=1}^{\infty} f_{11}^{(n)}$.

(b) Find the stationary distribution π of M.

Problem Set 9: z-transforms

The *z-transform* of a discrete pdf with probabilities p_n is $Z(p_n) = P(z) = \sum_{n=0}^{\infty} p_n z^n$. In general, $z \in \mathbb{C}$. Verify the following properties of $P(z)$.

1. Show that $P(z)$ converges if $|z| \leq 1$.

2. $P(0) = p_0$ and $P(1) = 1$.

3. $p_k = \frac{1}{k!}P^{(k)}(0)$, where $P^{(k)}(z)$ means $\frac{\partial^k Pz)}{\partial z^k}$.

4. $\mathbb{E}(n) = P^{(1)}(1)$, $\mathbb{V}ar(n) = P^{(2)}(1) + P^{(1)}(1) - (P^{(1)}(1))^2$.

We now extend the definition of z-transform to any sequence of non-negative numbers. Verify each of the following properties.

5(a) $Z(p_{n-1}) = zP(z)$;
(b) $Z(p_{n+1}) = \frac{1}{z}(P(z) - p_0)$;
(c) $Z(ap_n + bq_n) = aP(z) + bQ(z)$, where a and b are constants.

6. Show that the z-transform of the *convolution* $y_n = q_n * p_n = \sum_{k=0}^{n} q_k p_{n-k}$ is given by $Z(y_n) = Z(p_n)Z(q_n)$.

7(a) Find the z-transform of the Poisson distribution $p_n = \frac{\lambda^n e^{-\lambda}}{n!}$, $n = 0, 1, 2, \cdots$.
(b) Use Question 4 to find the mean and variance of the corresponding Poisson random variable X.

8. Use the fact that the z-transform for a pdf is unique, so that $p_n = Z^{-1}(P(z))$, to find the distribution of a convolution of two independent Poisson distributions $p_n = e^{-\lambda_1}\frac{\lambda_1^n}{n!}$ and $q_n = e^{-\lambda_2}\frac{\lambda_2^n}{n!}$.

9. Find the z-transform of a binomial distribution $X \sim B(n, p)$ and thus determine the mean and variance of X.

10. Use the z-transform to solve for x_n in terms of y_n where

$$x_n = ax_{n-1} + by_n, \ n = 2, 3, \cdots$$

$$x_1 = y_1, \ x_0 = 0.$$

Problem Set 10: Sum of two random variables

This problem set concerns the sum $Y = X_1 + X_2$ of two independent random variables, X_1 and X_2. We first examine the case where the X_i are discrete and defined on the non-negative integers. By considering all the ways that $Y = n$ is possible, we arrive at the *convolution formula* $P(Y = n) = \sum_{k=0}^{n} P(X_1 = k)P(X_2 = n - k)$. Find the pdf of Y in each of the following examples.

1. *Vandermonde's Identity* By expanding both sides of $(1 + x)^{n_1+n_2} = (1 + x)^{n_1}(1 + x)^{n_2}$ prove that

$$= \sum_{k=0}^{m} \binom{n_1}{k}\binom{n_2}{m-k} = \binom{n_1+n_2}{m}.$$

2(a) $X_1 \sim B(n_1, p)$, $X_2 \sim B(n_2, p)$.
(b) Extend this result to the sum of m such random variables.

3(a) $X_1 \sim \mathrm{Po}(\lambda_1)$, $X_2 \sim \mathrm{Po}(\lambda_2)$.
(b) Extend this result to the sum of m such random variables.

4. $X_i \sim \mathrm{G}(p)$ $(i = 1, 2)$, meaning that $PX_i = n) = pq^{n-1}$, where $q = 1 - p$.

5. Repeat Question 4 with $X_1 \sim$G(p_1) and $X_2 \sim$G(p_2) with $p_1 \neq p_2$.

For two independent continous random variables X and Y with density functions $f(x)$ and $g(y)$ repsectively, the density function $f_Z(x)$ of $Z = X + Y$ is the *convolution* of $f(x)$ and $f(y)$ given by:

$$f_Z(x) = f(x) * g(x) = \int_{-\infty}^{\infty} f(tx)g(x - t)\, dt$$

6. By using a simple substitution, show that $f(x) * g(x) = g(x) * f(x)$.

Find the density function of $Z = X + Y$ in each of the following cases, where X and Y are i.i.d. random variables (*independent and identically distributed*) random variables with the following density functions.

7. The *uniform density* $f(x) = 1$ on the unit interval $0 \leq x \leq 1$.

8. The *exponential density* $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$.

9. The *standard normal density* $N(0, 1)$.

10(a) Find the pdf of $Z = X + Y$ where X and Y have respective pdf of $f(x) = 2x$ and $g(x) = 1 - 2x$ $(0 \leq x \leq 1)$.
(b) Verify directly that your answer to (a) is a pdf.

Hints for Problems

Problem Set 1

1. What is the probability in each case that the people all have different birthday types?

4. The clue is in the name: consider what happens when a path does meet the x-axis for the first time and reflect that initial portion of the path in that axis.

6 & 7. Apply the Reflection principle.

10. Let W be a walk that returns to the origin (perhaps not for the first time) after $2n$ steps. Let the leftmost minimum point of W be $M = (k, m)$. Reflect the section from the origin to M along the vertical line $y = k$ and slide the reflected portion to the point $(2n, 0)$ of W. Taking M as the origin of a new coordinate system, the new path W' leads from the origin M to the point $(2n, 2m)$ and has all vertices strictly above the x-axis.

Problem Set 2

1 & 2. Put the derivative of the log likelihood equal to zero.

5. In this case, the derivative of the log likelihood is never equal to 0; we still need the maximum in the prescribed range of θ.

6. In this you need to find the minimum of $\sum_{i=1}^{n} |x_i - \theta|$, which is the *median* of the x_i, but you need to explain why.

7. A similar difficulty to that of Question 5.

8. You will need to find the mean of the pdf for $\max(X_i)$, which is the derivative of the cdf (*cumulative density function*).

9. Here the answer is not at all unique. All statistics in a certain range (which you should find) will do.

Problem Set 3

1. $H_0 : \mu = 150$, $H_1 : \mu \neq 150$. The $2 \cdot 5$ % significance level is $z = \frac{x-\mu}{\sigma} = 1 \cdot 96$.

2. The $2 \cdot 5\%$ significnance level is $z = -1 \cdot 645$. Remember the continuity correction.

3. Use $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Problem Set 4

1. Integrate out y in $f(x, y)$ to get $f_1(x)$.
2. Be careful with the inner limits when finding $f_1(x)$ and $f_1(y)$.
4. To find $f(y|x)$ you will first need to find $f_1(x)$.
5. To find $\mu_{Y||X}$ you will first need to find $f(y|x)$.
6. Just follow the definition of the joint probability as a double integral and separate variables.
10. Remember that $\mathrm{Var}(tY) = t^2\mathrm{Var}(Y)$ and similarly $\mathrm{Cov}(X, tY) = t\mathrm{Cov}(X, Y)$.

Problem Set 5

7. The form of the confidence interval is $\overline{X} \pm t_{0{\cdot}05}|\frac{S}{\sqrt{n-1}}$.

Problem Set 6

1. Just interchange the roles of x and y. 2. & 3. Solve the simultaneous equations.
6. Solve the normal equations in general.
9. & 10. Put each of the partial derivatives equal to 0 and then solve for a and for b respectively.

Problem Set 7

5. r will be positive if b and d have the same sign and otherwise not.

Problem Set 8

5(a) Represent the process as a two-state Markov chain.
7. We may pass from E_i to E_j in the first transition, or first pass to E_k with $k \neq j$.
10. Can E_1 appear in the long term process?

Problem Set 9

6. Write down $Z(y_n)$ as a double sum, change the order of summation, taking care to use the correct limits.

8. Use Questions 6 and 7.

10. Establish that $Z(x_n) = bZ(a^n) * Z(y_n) + (1-b)y_1 Z(a^{n-1})$.

Problem Set 10

2(a) You will need the Vandermonde identity to simplify the sum.

3(a) You will need to work in the Binomial theorem to simplify the expression that arises.

4(b) Sum the finite geometric series that arises: the result does cancel down quite neatly.

7. $X + Y$ takes on values in the interval $0 \leq x \leq 2$ and you need to treat the cases $x \leq 1$ and $1 \leq x$ separately as the limits of integration changes as you cross that boundary.

9. You will need to complete the square in the power of the exponential in order to reduce the integral to that of a normal distribution. Remember multiplier terms entirely in x can be taken outside the integral.

10. Repeat the approach of Question 7. For the interval $1 \leq x \leq 2$, you may exploit the symmetry of the pdf of $X + Y$ in the line $x = 1$ to simplify the calculations.

Answers to the Problems

Problem Set 1

1(a) 4 (b) 23. 2(a) path always lies above the x-axis (b) path never crosses below the x-axis. (c) $\binom{p+q}{p}$. 3. A's lead does not reach $p-q$ until the final vote is counted; A's lead never exceeds $p-q$. 5.$\binom{b}{\frac{a+b+c}{2}}$. 6. $\frac{p+1-q}{p}$. 7. $\frac{p-q}{p+q}$. 9. $\frac{\binom{2n}{n}}{2^n}$.

Problem Set 2

1. & 2. $\hat{\theta} = \bar{X}$. 3. $-\frac{n}{\ln(X_1 X_2 \cdots X_n)}$. 4. $\bar{X}$. 5. $\min(X_i)$. 6. the median of the X_i. 7. $\max(X_i)$. 9. $\frac{1}{2}(\min(X_i) + \max(X_i))$ is one solution. 10. $\bar{X}$.

Problem Set 3

1. $z = 2 \cdot 2$, reject H_0. 2. $z = -2 \cdot 33$ reject H_0. 3. $z = 2 \cdot 2$, reject H_0. 4. $(5 \cdot 78\text{cm}, 6 \cdot 22\text{cm})$.5 (i) $0 \cdot 0554$, (ii) $0 \cdot 390$. 6 (i) accept H_0; (ii) reject H_0. 7 48. 8. reject H_0. 9. $z = -2 \cdot 04$, reject H_0. 10. $z = 1 \cdot 784$, accept H_0.

Problem Set 4

1. $\frac{3}{2} - x$. 2. $\frac{x^3}{4}$. 4. $\frac{1}{1-x}$, $2y$. 5. $\frac{2x}{3}$.

Problem Set 5

1. $\chi^2 = 1 \cdot 9$, H_0 accepted. 2. $\chi^2 = 10 \cdot 1$, H_0 rejected. 3. $\chi^2 = 1 \cdot 83$, H_0 accepted. 4. $\chi^2 = 4 \cdot 60$, H_0 accepted. 5. $\chi^2 = 0 \cdot 3341$, H_0 accepted. 6(a) 20 (b) $264 < \sigma^2 < 1000$. 7. $t = 2 \cdot 57$, $\nu = 9$; reject H_0 at the 5% level. 8. $(0 \cdot 78, 1 \cdot 69)$. 9. $\overline{x} = 6 \cdot 0$ $\overline{y} = 5 \cdot 7$ $n_X s_X^2 = 0 \cdot 64$, $n_Y s_Y^2 = 0 \cdot 24$. 10. $t = 3 \cdot 03$, $\nu = 18$.The critical $\cdot 005$ value of t is $t = 2 \cdot 878$, H_0 is rejected.

Problem Set 6

1. $\sum y = nc + d\sum y$ and $\sum xy = c\sum y + d\sum y^2$. 2. $y = 11 \cdot 7 + 0 \cdot 186x$. 3. $x = -4 \cdot 34 + 0 \cdot 769y$. 4. $y = -2 + 0 \cdot 6x$; $y = 5 \cdot 2$.

Problem Set 7

4. $0 \cdot 86$. 5. $0 \cdot 2879$. 7. $0 \cdot 38$. 9. Accept H_0. 10. $(-1 \cdot 85, 2 \cdot 66)$.

Problem Set 8

4(a) $\begin{bmatrix} 0 \cdot 8 & 0 \cdot 2 \\ 0 \cdot 4 & 0 \cdot 6 \end{bmatrix}$. (b) $\begin{bmatrix} 0 \cdot 6752 & 0 \cdot 3248 \\ 0 \cdot 6496 & 0 \cdot 3504 \end{bmatrix}$, $0 \cdot 6752$. (c) 0.6496. (d) $\frac{2}{3}$. 5(a) $\frac{5}{7}$. (b) $(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$. 6. $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. 7. (c) $\begin{bmatrix} 1+\frac{\alpha}{\beta} & \frac{1}{\alpha} \\ \frac{1}{\beta} & 1+\frac{\beta}{\alpha} \end{bmatrix}$. (d) $\begin{bmatrix} 3 \cdot 5 & 4 \\ 10 & 1 \cdot 4 \end{bmatrix}$. 9. $0 \cdot 08$, $0 \cdot 08$. 10. $0 \cdot 5$, $(0, \frac{6}{13}, \frac{7}{13})$.

Problem Set 9

7. $e^{\lambda(z-1)}$. 8. $\mathrm{Po}(\lambda_1 + \lambda_2)$. 9. $(q + pz)^n$. 10. $x_n = b(y_1 a^{n-1} + y_2 a^{n-2} + \cdots + y_n) + (1-b)y_1 a^{n-1}$.

Problem Set 10

2(c) $B(n_1 + \cdots + n_m, p)$. 3(c) $\mathrm{Po}(\lambda_1 + \cdots + \lambda_m)$. 4(a) $(n-1)p^2(1-p)^{n-2}$, $n = 2, 3, \cdots$. (b) $\frac{p_1 p_2}{p_1 - p_2} \cdot \Big((1-p_2)^{n-1} - (1-p_1)^{n-1}\Big)$, $n = 2, 3, \cdots$. 5. $\frac{p_1 p_2}{p_1 - p_2} \cdot \Big((1-p_2)^{n-1} - (1-p_1)^{n-1}\Big)$, $n = 2, 3, \cdots$. 7. $f_Z(x) = x$ if $0 \le x \le 1$ and $f_Z(x) = 2 - x$ if $1 \le x \le 2$. 8. $\lambda^2 x e^{-\lambda x}$, $x \ge 0$. 9. $N(0, 2)$. 10(a) $= 2x^2(1 - \frac{x}{3})$ $(0 \le x \le 1)$ $2(2-x)^2(\frac{1+x}{3})$ $(1 \le x \le 2)$.