University of Waterloo STAT 221 – W. H. Cherry

Assignment 1 Outline Solution

A1 – 1. This question involves the binomial and negative binomial distributions as probability models.

(a) Let the random variable Y represent the number of instantaneous recoveries in a group of n patients; we use the model $Y \sim \mathrm{Bi}(\mathrm{n}, 0.1)$.

Then: $\Pr(y \text{ patients in n recover}) = \binom{\mathrm{n}}{y}(0.1)^{\mathrm{n}}(0.9)^{\mathrm{n}-y} \quad y = 1, 2, \ldots., \mathrm{n},$

so that: $\Pr(\text{at least 1 patient in n recovers}) = 1 - \Pr(0 \text{ patients in n recover}).$

$\therefore \quad 1 - \binom{\mathrm{n}}{0}(0.1)^0(0.9)^{\mathrm{n}} = 0.5$ so that: $\mathrm{n} = \dfrac{\ln(0.5)}{\ln(0.9)} \simeq 6.579 \simeq 7$; *i.e.*, about 7 patients are required.

(b) $\Pr(7\text{th patient is } 1st \text{ recovery}) = \Pr(\text{no recoveries in 6 patients}) \times \Pr(\text{recovery})$

$= \binom{6}{0}(0.1)^0(0.9)^6 \times 0.1 = (0.9)^0(0.1)^1 \simeq 0.05314;$

(c) $\Pr(10\text{th patient is } 2nd \text{ recovery}) = \Pr(1 \text{ recovery in 9 patients}) \times \Pr(\text{recovery})$

$= \binom{9}{1}(0.1)^1(0.9)^8 \times 0.1 = (0.1)^2(0.9)^0 \simeq 0.03874.$

A1 – 2. This question involves the negative binomial distribution as a probability model.

(a) Pr(exactly 8 applicants have to be interviewed for 5 positions)

= Pr(4 successful applicants in 7 interviews) × Pr(successful applicant)

$= \binom{7}{4}\left(\tfrac{2}{3}\right)^4\left(\tfrac{1}{3}\right)^3 \times \left(\tfrac{2}{3}\right) = \dfrac{1{,}120}{6{,}561} \simeq 0.1707;$

(b) Pr(8 or fewer applicants have to be interviewed for 5 positions)

= Pr(8 or 7 or 6 or 5 applicants have to be interviewed for 5 positions)

$= \sum_{y=5}^{8} \Pr(y \text{ applicants have to be interviewed for 5 positions})$

$= \left(\tfrac{2}{3}\right)^5\left[1 + \tfrac{5}{3} + \tfrac{15}{9} + \tfrac{35}{27}\right] = \dfrac{4{,}864}{6{,}561} \simeq 0.7414;$ [reasoning as in (a) for each probability].

A1 – 3. This question involves the Poisson distribution as a probability model.
Let the random variable Y represent the number of calls reaching the switchboard in a time interval of length t – we use the model $Y \sim \mathrm{Po}(\lambda\mathrm{t})$;

i.e., $\Pr(y \text{ calls reach switchboard in time t}) = \dfrac{(\lambda\mathrm{t})^y e^{-\lambda\mathrm{t}}}{y!}; \quad y = 0, 1, 2, 3, \ldots.,$

where: λ = average rate at which calls reach the switchboard
= 600 per hour = 10 per minute = 5/3 per ten seconds.

(a) Pr(switchboard is overtaxed during a given minute)

= Pr(21 or more calls reach the switchboard in the minute)

= 1 – Pr(20 or fewer calls reach the switchboard in the minute)

$= \sum_{y=0}^{20} \Pr(y \text{ calls reach the switchboard in the minute})$

$= \sum_{y=0}^{20} \dfrac{(10)^y e^{-10}}{y!} \simeq 0.001\,588;$

NOTE: In *this* question, four significant figures for the final probability (surprisingly) can be obtained by summing *fewer* terms by avoiding use of the complement.

(b) $\Pr(\text{no calls in 10 seconds}) = \dfrac{(5/3)^0 e^{-5/3}}{0!} = e^{-5/3} \simeq 0.1889;$

A1 – 4. This question involves the Poisson distribution as a probability model.
Let the random variable Y represent the number of blackflies in a volume of v m^3 – we use the model $Y \sim \mathrm{Po}(15\mathrm{v})$;

i.e., $\Pr(y \text{ blackflies in volume v m}^3) = \dfrac{(15\mathrm{v})^y e^{-15\mathrm{v}}}{y!}; \quad y = 0, 1, 2, 3, \ldots.,$

(a) In 1.2 m^3, we expect 18 blackflies;

$\therefore$ $\Pr(\text{exactly 3 blackflies in a tent of volume 1.2 m}^3) = \dfrac{(18)^3 e^{-18}}{3!} \simeq 0.000\,014\,804.$

(b) In 0.04 m^3, we expect 0.6 blackflies;

$\therefore$ $\Pr(\text{at least 1 blackfly in a box of volume 0.04 m}^3) = 1 - \Pr(\text{no blackflies in the box}) = 1 - e^{-0.6} \simeq 0.4512.$

(*continued overleaf*)

Assignment 1 Outline Solution (continued 1)

A1 – 4. (cont.) (c)(i) Regarding the three tents as a total volume of 3.6 m^3, we expect 54 blackflies in them;

$$\therefore \quad \Pr(\text{exactly 40 blackflies in the three tents}) = \frac{(54)^{40}e^{-54}}{40!} \simeq 0.008\,555;$$

(c)(ii) We use the probability from (a) [π, say] in a binomial distribution with n = 3;

$$\therefore \quad \Pr(\text{exactly 2 tents contain exactly 3 blackflies each}) = \binom{3}{2}\pi^2(1-\pi)^1 \simeq 6.574\times10^{-10}.$$

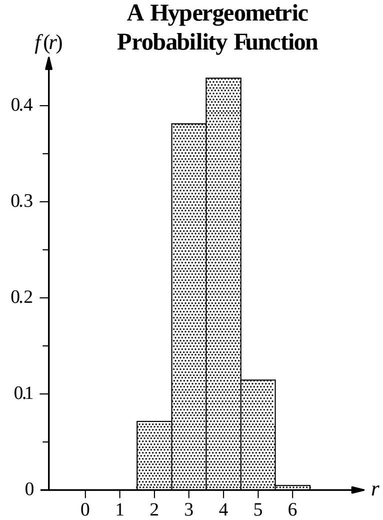
A1 – 5. (a) The **hypergeometric** distribution can be used to model the process of selecting a sample of n balls at random with*out* replacement from an urn containing a total of $\mathbf{N}$ balls of which $\mathbf{R}$ are red (and $\mathbf{N}-\mathbf{R}$ are black). If the random variable R represents the number of red balls in the sample, its probability function is as given at the right ('urn' here corresponds to 'population' in statistics).

$$f(r) = \frac{\binom{\mathbf{R}}{r}\binom{\mathbf{N}-\mathbf{R}}{\mathrm{n}-r}}{\binom{\mathbf{N}}{\mathrm{n}}}, \quad r = 0, 1, 2, \ldots, \mathrm{n}.$$

When $\mathbf{N} = 10$, $\mathbf{R} = 6$ and n = 6, we can tabulate and graph $f(r)$ as follows:

r	$f(r)$
0	0
1	0
2	$\frac{15}{210} \simeq 0.0714$
3	$\frac{80}{210} \simeq 0.3810$
4	$\frac{90}{210} \simeq 0.4286$
5	$\frac{24}{210} \simeq 0.1143$
6	$\frac{1}{210} \simeq 0.0048$

A Hypergeometric Probability Function



(b) The **binomial** distribution can be used to model a process called a Bernoulli trial, which has the following (restricted) chracteristics:

- a Bernoulli trial has *two* outcomes (generically called 'success' and 'failure');
- under repetition, the outcomes of Bernoulli trials ('success' and 'failure') are *probabilistically indepentent* events;
- under repetition, the probability of 'success' (and, hence, the probability of 'failure') is *constant* [π, say, and $1-\pi$].

an example of Bernoulli trials is selecting from an urn of red and black balls *with* replacement [to be contrasted with the hypergeometric distribution for selecting with*out* replacement in (a) above]; a more familiar example of a Benoulli trial is tossing a coin.

The binomal probability function (for the random variable Y) is:

$$\Pr(y\ \text{'successes' in n trials}) = \binom{\mathrm{n}}{y}\pi^y(1-\pi)^{\mathrm{n}-y}; \qquad y = 0, 1, 2, \ldots, \mathrm{n}.$$

When n = 10 and π successively takes the values 0.2, 0.5 and 0.9, the probability functions [$f(w)$, $f(y)$, $f(z)$] of the respective random variables W, Y and Z are tabulated and graphed as shown on page 0.9d.

(c) The **negative binomial** distribution is also used to model repetition of Bernoulli trials, but it differs from the binomial distribution in its concern with the probability that the r*th* 'success' occurs on the n*th* trial or, equivalently, that v 'failures' occur before the r*th* 'success'. The negative binomial probability function (for random variable V) is:

$$\Pr(v\ \text{'failures' before the r}th\ \text{'successes'}) = \binom{\mathrm{r}+v-1}{\mathrm{r}-1}\pi^{\mathrm{r}}(1-\pi)^v \quad ; \quad v = 0, 1, 2, 3, \ldots.$$

When r = 2 and $\pi = 0.4$, the probability function $f(v)$ of V is tabulated and graphed on page 9c. [Terms of the infinite set of values of $f(v)$ are evaluated up to $V = 15$, to show the behaviour of the right tail of this distribution.]

(d) The **Poisson** distribution is used to model processes involving 'outcomes occurring unpredictably (on an *individual* basis) in space or time'; the more common shorter description of modelling *random events in space or time* has the *dis*advantage of using 'random' and 'event' in their colloquial (*not* their probabilistic) sense.

The derivation of the Poisson probability function involves three assumptions:

(*continued*)

Assignment 1 Outline Solution (continued 2)

A1 – 5. (cont.) (d

- *independence*: events (which are *numbers of outcomes in an interval*) are *probabilistically independent* for *non-*overlapping (spatial or temporal) intervals; the shorter version of this assumption – the *numbers* of outcomes comprising events in non-overlapping intervals are independent – risks confusing probabilistic independence of events with independence of numbers.
- *individuality*: outcomes occur singly, not in groups; *i.e.*, the probability of more than one outcome in a sufficiently small interval is negligible;
- *homogeneity*: outcomes occur at a uniform rate over the entire interval under consideration so that, in a (sub-)interval of length t, the expected number of outcomes is λt, where the *intensity parameter* λ is the *average rate of outcomes over the entire interval.*

  This *mathematical* assumption has troublesome real-world implications – a *uniform* underlying process exhibiting *unpredictable* actual individual outcomes.

The Poisson probability function (for the random variable S) is:

$$\Pr(s \text{ outcomes in an interval of length t}) = \frac{(\lambda \mathrm{t})^s e^{-\lambda \mathrm{t}}}{s!} = \frac{\mu^s e^{-\mu}}{s!} \qquad s = 0, 1, 2, 3, \ldots.$$

In contrast to Bernoulli trials where *a person* is usually involved in repetition (*e.g.*, in coin tossing), a Poisson process often repeats with*out* human intervention and the 'repetitions' arise from the (spatial or temporal) intervals *imposed* on the outcomes.

When μ takes the two values 1 and 10, the probability functions [$f(s)$, $f(t)$] of the two random variables S and T are tabulated and graphed as shown on the page 9e. [Terms of the infinite sets of values of $f(s)$ and $f(t)$ and are evaluated out to $S = 10$ and $T = 20$, respectively, to show the behaviour of the right tails of these distributions.]

***A1 – 6.** From the data given, we can find values, for good and for bad drivers, of the proportion of 2-car accidents *involving the respective types of drivers* in which they are at fault; these proportions can be used as values of the corresponding probabilities. That is:

for good drivers: $\Pr(\text{at fault}) \simeq \dfrac{5,000 + 15,000}{5,000 + 5,000 + 15,000 + 25,000} = \frac{2}{5} = 0.4;$

for bad drivers: $\Pr(\text{at fault}) \simeq \dfrac{25,000 + 55,000}{15,000 + 25,000 + 55,000 + 55,000} = \frac{8}{15} = 0.5\dot{3}.$

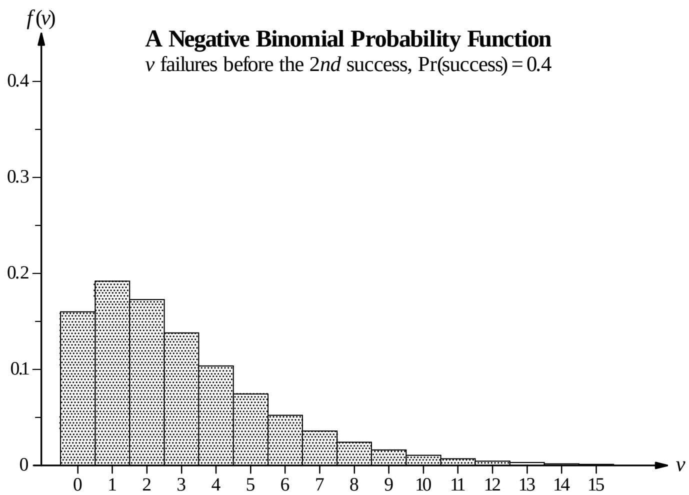
Then: Pr(license revoked immediately after 5th accident) =
Pr(driver at fault in 2 of their first 4 accidents *and* in their 5th accident) = $\binom{4}{2}\pi^3(1-\pi)^2$.

Using the 'at fault' values above for π, we obtain: probability license revoked for good drivers: 0.138 24;
probability license revoked for bad drivers: 0.198 226.

- The accident data in the statement of this question indicate a *substantial* difference between the accident records of good and bad drivers, yet these two final probabilities are relatively close in value. Explain briefly why this is so.

A1 – 5. (cont.) (c) A negative binomial distribution probability function (r = 2, π = 0.4):

v	$f(v)$
0	0.16
1	0.192
2	0.172 8
3	0.138 24
4	0.103 68
5	0.074 649 6
6	0.052 254 72
7	0.035 831 808
8	0.024 186 470
9	0.016 124 314
10	0.010 642 047
11	0.006 965 703
12	0.004 527 707
13	0.002 925 595
14	0.001 880 740
15	0.001 203 674
.	
.	(0.997 912 378)

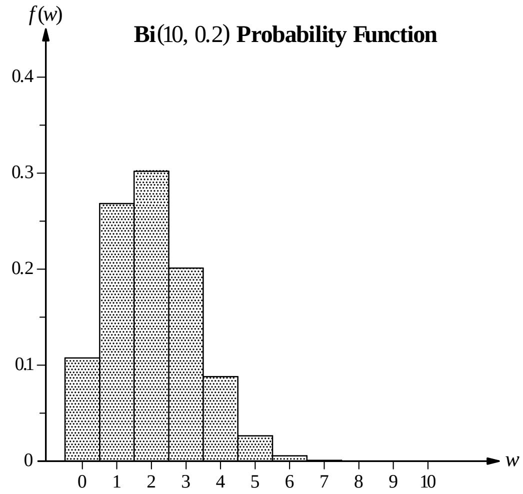


(*continued overleaf*)

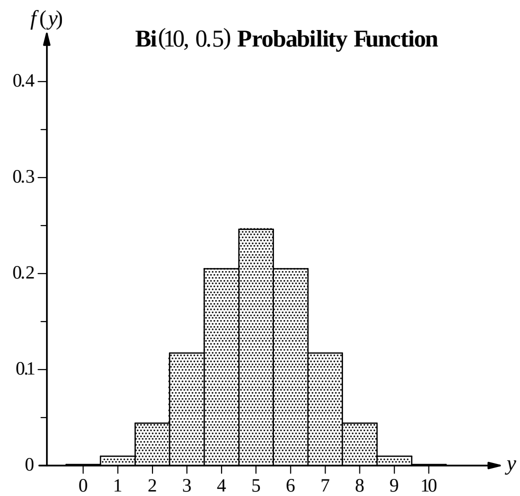
Assignment 1 Outline Solution (continued 3)

A1 – 5. (b) **(cont.)**

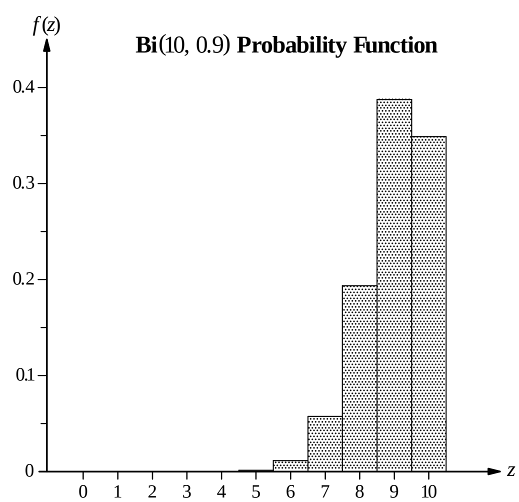
w	$f(w)$
0	0.107 374 182
1	0.268 435 456
2	0.301 989 888
3	0.201 326 592
4	0.088 080 384
5	0.026 424 115
6	0.005 505 024
7	0.000 786 432
8	0.000 073 728
9	0.000 004 096
10	0.000 000 102
	(1.)



y	$f(y)$
0	0.000 976 563
1	0.009 765 625
2	0.043 945 313
3	0.117 187 500
4	0.205 078 125
5	0.246 093 750
6	0.205 078 125
7	0.117 187 500
8	0.043 945 313
9	0.009 765 625
10	0.000 976 563
	(1.)



z	$f(z)$
0	0.000 000 000
1	0.000 000 009
2	0.000 000 365
3	0.000 008 748
4	0.000 137 781
5	0.001 488 035
6	0.011 160 261
7	0.057 395 628
8	0.193 710 245
9	0.387 420 489
10	0.348 678 440
	(1.)

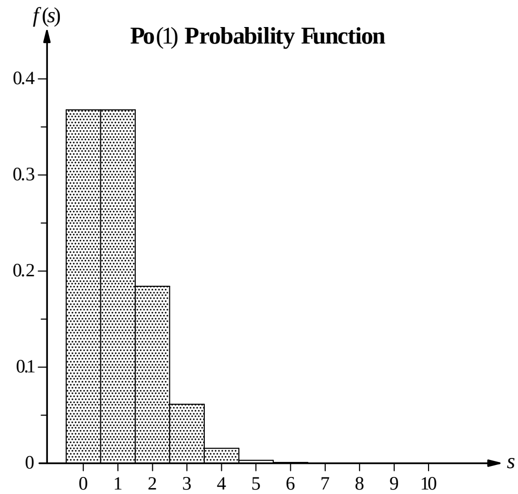


We see that, in each case, the most probable value of the random variable, corresponding to the *highest* bar of each histogram, is given by the product, $n \times \pi$, of the binomial distribution parameters; this is (of course) the *mean* of the distribution.

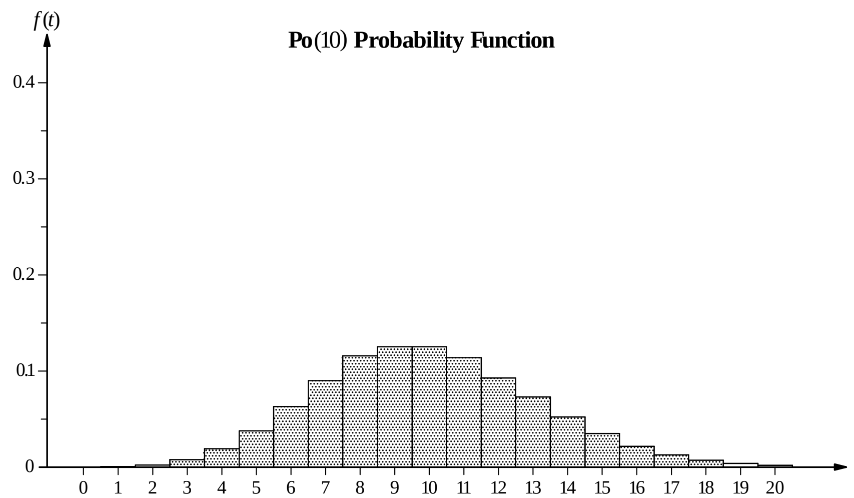
Assignment 1 Outline Solution (continued 4)

A1 – 5. (c)
(cont.)

s	$f(s)$
0	0.367 879 441
1	0.367 879 441
2	0.183 939 721
3	0.061 313 240
4	0.015 328 310
5	0.003 065 662
6	0.000 510 944
7	0.000 072 992
8	0.000 009 124
9	0.000 001 014
10	0.000 000 101
.	
.	(0.999 999 990)



t	$f(t)$
0	0.000 045 400
1	0.000 453 999
2	0.002 269 996
3	0.007 566 655
4	0.018 916 637
5	0.037 833 275
6	0.063 055 458
7	0.090 079 226
8	0.112 599 032
9	0.125 110 036
10	0.125 110 036
11	0.113 736 396
12	0.094 780 330
13	0.072 907 946
14	0.052 077 104
15	0.034 718 070
16	0.021 698 794
17	0.012 763 996
18	0.007 091 109
19	0.003 732 163
20	0.001 866 081
.	
.	(0.998 411 739)



We see that, in both cases, the most probable value of the random variable, corresponding to the *highest* bar of each histogram, is the integer part of the value of the parameter (here, the mean) μ; if this value is an integer, then the bar immediately to the *left* of this tallest bar is also of the same height (*i.e.*, there are *two* most probable values of the random variable if the mean of the Poisson distribution is an integer).

NOTE: The Po(1) distribtiion is highly asymmetric, the Po(10) distribution is relatively *symmetrical* and '*bell-shaped*,' reminding us of the *normal* approximation for finding values for (sums of) Poisson probabilities when μ is large (say 20 or more).