Probability and Statistics
For Final Year Engineering Students
Home Work Project #1

Ben Smith - 6161189
Simon Lehman - 6164668

June 12, 2011

1. *Normal approximation to the binomial distribution*
   Let B_1, B_1, ... be a sequence of binomial random variables with B_n having a number of trials parameter equal to n and success probability parameter p (the same value for all random variables in the sequence). Let μ_n be the sequence of means, $E[B_n] = \mu_n$ and let σ_n^2 be the sequence of variances, $V(B_n) = \sigma_n^2$. Denote $\tilde{B_n} = \frac{B_n - \mu_n}{\sigma_n}$.

   (a) For $p = \frac{1}{3}$ and $p = \frac{1}{2}$, plot the sequence of PDF's of $\tilde{B_n}$for $n = 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 100$.

   The probability that B_n will have a count k is

$$P(B_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

   The PDF of $\tilde{B_n}$, is the normalized PDF of B_n, and can be plotted by calculating the probabilities that B_n will have a count k and plotting these values on a transformed x axis, using $z = \frac{k-\mu_n}{\sigma_n}$, giving $P\left(\tilde{B_n} = z\right)$. To calculate z for all values of k the means and standard deviations all need to be calculated.
   The mean of a binomial sequence is $\mu_n = np$.
   The variance for a binomial sequence is $Var(B_n) = np(1-p)$.
   Below is the Matlab code for plotting the PDF of $\tilde{B_n}$ for $p = \frac{1}{2}$ and $n = 50$

```
p = 1 / 2;                  % The probability parameter
n = 50;                     % The number of trials
k = [0:n];                  % The k values for P(B_n = k)

mean = n * p;               % The sequence of means for each trial
                            % size
var = n * p * (1 - p);      % The sequence of variances

z = (k - mean) / sqrt(var);       % The sequence of z values for the
                                  % normalized PDF

% The normalized PDF
f = factorial(n) ./ (factorial(k) .* factorial(n - k)) .* ...
    p .^ k .* (1 - p) .^ (n - k);

stem(z, f)                  % Plot
```

   This is repeated for all n and p values, resulting in the table on the next page.

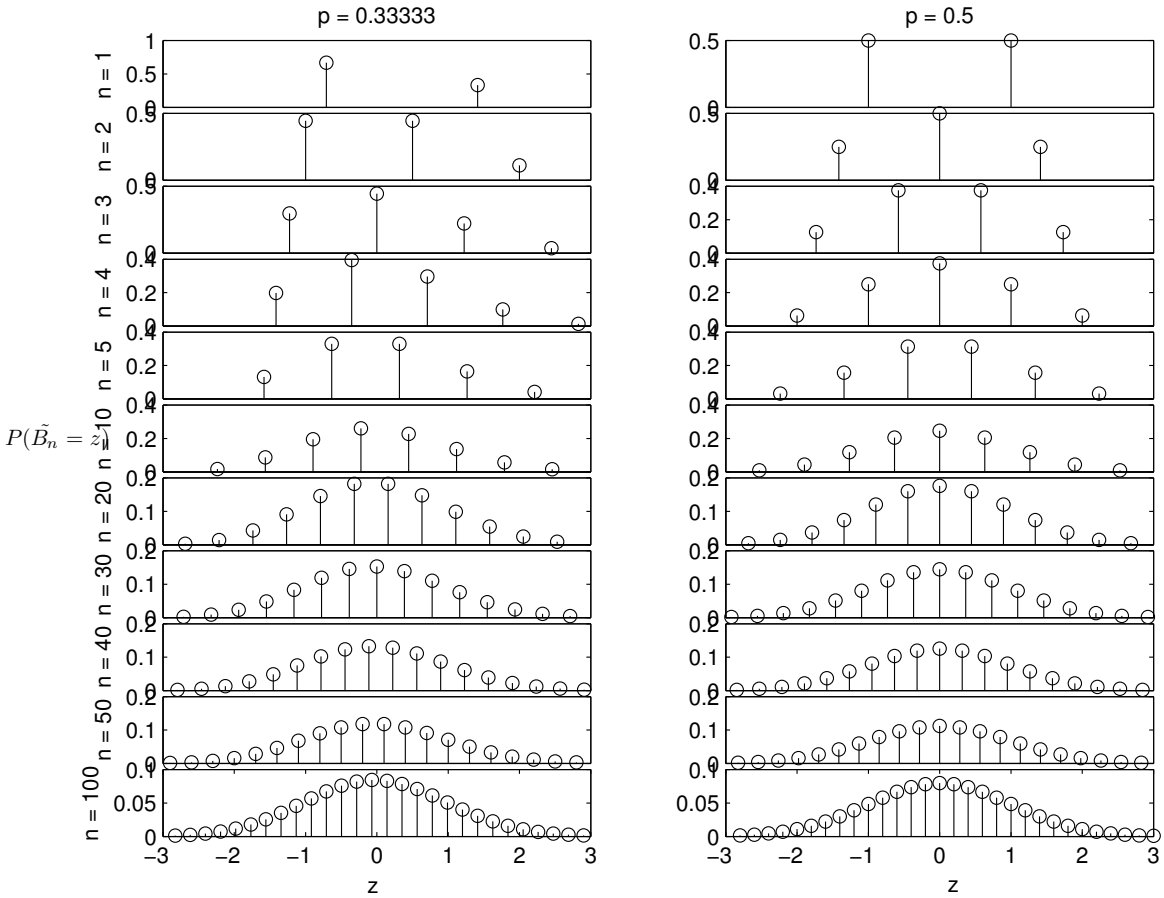

p = 0.33333
p = 0.5
n = 1
n = 2
n = 3
n = 4
n = 5
n = 10
n = 20
n = 30
n = 40
n = 50
n = 100
$P(\tilde{B}_n = \tilde{z})$
z
z

Figure 1

(b) Let $q_n = P(B_n \geq \mu_n + 2\sigma_n)$. For $p = \frac{1}{3}$ and $p = \frac{1}{2}$, calculate q_n for $n = 1, ..., 100$, put the results in a table.

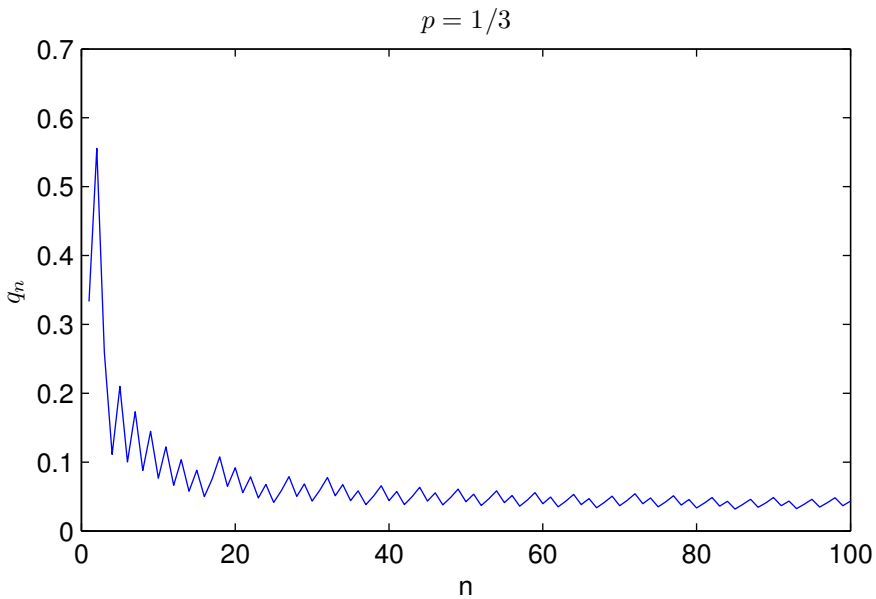
q_n can be evaluated by summing up all the probabilities in the probability mass function of B_n where k is greater or equal to:

$$\mu_n + 2\sigma_n = np + 2\sqrt{np(1-p)}$$

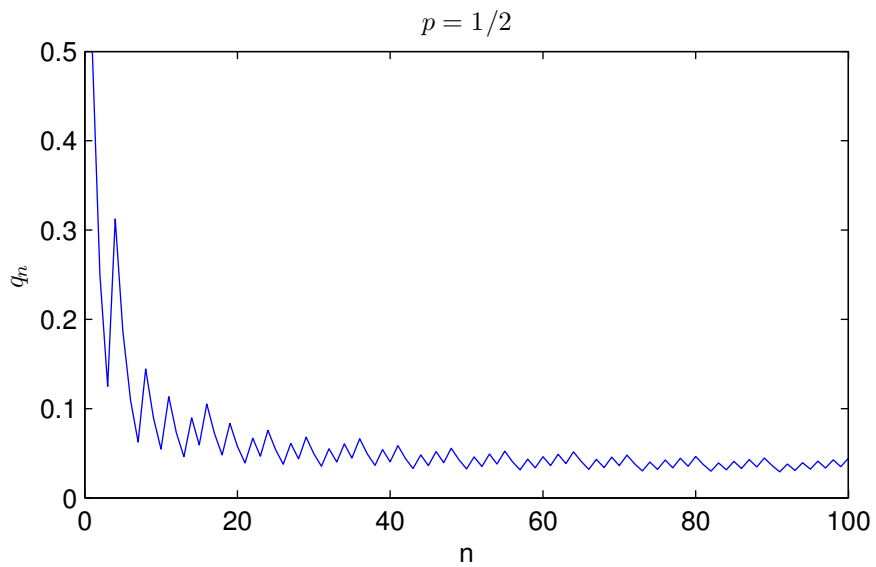
Below is the code to calculate this probability for $p = \frac{1}{2}$ and $n = 100$

```
p = 1 / 2
n = 100
k = [0:n]

f = factorial(n) ./ (factorial(k) .* factorial(n - k)) ...
    .* p .^ k .* (1 - p) .^ (n - k);

mean = n * p;
var = n * p * (1 - p);
sd = sqrt(var);

qn = sum(f(ceil(mean + 2 * sd):end))
```

This can be done for values of n and p as shown in Table 1 on page 5.



(a) q_n values for $p = 1/3$



(b) q_n values for $p = 1/2$

Figure 2: Binomial q_n values

n	q_n	n	q_n
1	0.3333	51	0.0535
2	0.5556	52	0.0371
3	0.2593	53	0.0469
4	0.1111	54	0.0584
5	0.2099	55	0.0411
6	0.1001	56	0.0513
7	0.1733	57	0.0361
8	0.0879	58	0.0451
9	0.1448	59	0.0557
10	0.0766	60	0.0397
11	0.1221	61	0.0491
12	0.0664	62	0.0349
13	0.1035	63	0.0433
14	0.0576	64	0.0531
15	0.0882	65	0.0382
16	0.0500	66	0.0469
17	0.0755	67	0.0337
18	0.1076	68	0.0415
19	0.0648	69	0.0505
20	0.0919	70	0.0366
21	0.0557	71	0.0447
22	0.0787	72	0.0541
23	0.0480	73	0.0396
24	0.0677	74	0.0480
25	0.0415	75	0.0350
26	0.0583	76	0.0426
27	0.0790	77	0.0512
28	0.0503	78	0.0378
29	0.0682	79	0.0455
30	0.0435	80	0.0335
31	0.0589	81	0.0405
32	0.0777	82	0.0485
33	0.0510	83	0.0360
34	0.0673	84	0.0432
35	0.0442	85	0.0319
36	0.0584	86	0.0384
37	0.0384	87	0.0459
38	0.0507	88	0.0342
39	0.0656	89	0.0409
40	0.0441	90	0.0486
41	0.0571	91	0.0365
42	0.0384	92	0.0434
43	0.0498	93	0.0325
44	0.0634	94	0.0388
45	0.0434	95	0.0459
46	0.0554	96	0.0346
47	0.0379	97	0.0411
48	0.0485	98	0.0484
49	0.0610	99	0.0367
50	0.0424	100	0.0434

(a) q_n values for $p = \frac{1}{3}$

n	q_n	n	q_n
1	0.5000	51	0.0460
2	0.2500	52	0.0352
3	0.1250	53	0.0492
4	0.3125	54	0.0380
5	0.1875	55	0.0524
6	0.1094	56	0.0407
7	0.0625	57	0.0314
8	0.1445	58	0.0435
9	0.0898	59	0.0337
10	0.0547	60	0.0462
11	0.1133	61	0.0361
12	0.0730	62	0.0490
13	0.0461	63	0.0385
14	0.0898	64	0.0517
15	0.0592	65	0.0408
16	0.1051	66	0.0320
17	0.0717	67	0.0432
18	0.0481	68	0.0341
19	0.0835	69	0.0456
20	0.0577	70	0.0361
21	0.0392	71	0.0480
22	0.0669	72	0.0382
23	0.0466	73	0.0302
24	0.0758	74	0.0403
25	0.0539	75	0.0320
26	0.0378	76	0.0423
27	0.0610	77	0.0338
28	0.0436	78	0.0444
29	0.0680	79	0.0356
30	0.0494	80	0.0465
31	0.0354	81	0.0374
32	0.0551	82	0.0299
33	0.0401	83	0.0392
34	0.0607	84	0.0315
35	0.0448	85	0.0410
36	0.0662	86	0.0331
37	0.0494	87	0.0428
38	0.0365	88	0.0347
39	0.0541	89	0.0447
40	0.0403	90	0.0363
41	0.0586	91	0.0293
42	0.0442	92	0.0379
43	0.0330	93	0.0307
44	0.0481	94	0.0395
45	0.0362	95	0.0321
46	0.0519	96	0.0411
47	0.0395	97	0.0335
48	0.0557	98	0.0427
49	0.0427	99	0.0350
50	0.0325	100	0.0443

(b) q_n values for $p = \frac{1}{2}$

Table 1: Binomial q_n values

(c) Use a normal (CLT) approximation to approximate the results of (b). Put the results in a table, showing the errors.

The CLT approximation states that a normal bell curve can be used to approximate a binomial distribution and that this approximation gets better as the binomial has more trials. The probability that a normally distributed set of random variables is greater than two standard deviations from the mean can be read off a normal distribution table at 0.0228. This is shown in Table 2 on page 7 and Table 3 on page 8 with an error calculated by subtracting the CLT approximation q from the actual binomial probability q_n.
This table shows that as there are more trials, the approximation gets closer to the actual binomial probability.

n	q_n	q (CLT)	error
1	0.3333	0.0228	0.3105
2	0.5556	0.0228	0.5328
3	0.2593	0.0228	0.2365
4	0.1111	0.0228	0.0883
5	0.2099	0.0228	0.1871
6	0.1001	0.0228	0.0773
7	0.1733	0.0228	0.1505
8	0.0879	0.0228	0.0651
9	0.1448	0.0228	0.1220
10	0.0766	0.0228	0.0538
11	0.1221	0.0228	0.0993
12	0.0664	0.0228	0.0436
13	0.1035	0.0228	0.0807
14	0.0576	0.0228	0.0348
15	0.0882	0.0228	0.0654
16	0.0500	0.0228	0.0272
17	0.0755	0.0228	0.0527
18	0.1076	0.0228	0.0848
19	0.0648	0.0228	0.0420
20	0.0919	0.0228	0.0691
21	0.0557	0.0228	0.0329
22	0.0787	0.0228	0.0559
23	0.0480	0.0228	0.0252
24	0.0677	0.0228	0.0449
25	0.0415	0.0228	0.0187
26	0.0583	0.0228	0.0355
27	0.0790	0.0228	0.0562
28	0.0503	0.0228	0.0275
29	0.0682	0.0228	0.0454
30	0.0435	0.0228	0.0207
31	0.0589	0.0228	0.0361
32	0.0777	0.0228	0.0549
33	0.0510	0.0228	0.0282
34	0.0673	0.0228	0.0445
35	0.0442	0.0228	0.0214
36	0.0584	0.0228	0.0356
37	0.0384	0.0228	0.0156
38	0.0507	0.0228	0.0279
39	0.0656	0.0228	0.0428
40	0.0441	0.0228	0.0213
41	0.0571	0.0228	0.0343
42	0.0384	0.0228	0.0156
43	0.0498	0.0228	0.0270
44	0.0634	0.0228	0.0406
45	0.0434	0.0228	0.0206
46	0.0554	0.0228	0.0326
47	0.0379	0.0228	0.0151
48	0.0485	0.0228	0.0257
49	0.0610	0.0228	0.0382
50	0.0424	0.0228	0.0196

n	q_n	q (CLT)	error
51	0.0535	0.0228	0.0307
52	0.0371	0.0228	0.0143
53	0.0469	0.0228	0.0241
54	0.0584	0.0228	0.0356
55	0.0411	0.0228	0.0183
56	0.0513	0.0228	0.0285
57	0.0361	0.0228	0.0133
58	0.0451	0.0228	0.0223
59	0.0557	0.0228	0.0329
60	0.0397	0.0228	0.0169
61	0.0491	0.0228	0.0263
62	0.0349	0.0228	0.0121
63	0.0433	0.0228	0.0205
64	0.0531	0.0228	0.0303
65	0.0382	0.0228	0.0154
66	0.0469	0.0228	0.0241
67	0.0337	0.0228	0.0109
68	0.0415	0.0228	0.0187
69	0.0505	0.0228	0.0277
70	0.0366	0.0228	0.0138
71	0.0447	0.0228	0.0219
72	0.0541	0.0228	0.0313
73	0.0396	0.0228	0.0168
74	0.0480	0.0228	0.0252
75	0.0350	0.0228	0.0122
76	0.0426	0.0228	0.0198
77	0.0512	0.0228	0.0284
78	0.0378	0.0228	0.0150
79	0.0455	0.0228	0.0227
80	0.0335	0.0228	0.0107
81	0.0405	0.0228	0.0177
82	0.0485	0.0228	0.0257
83	0.0360	0.0228	0.0132
84	0.0432	0.0228	0.0204
85	0.0319	0.0228	0.0091
86	0.0384	0.0228	0.0156
87	0.0459	0.0228	0.0231
88	0.0342	0.0228	0.0114
89	0.0409	0.0228	0.0181
90	0.0486	0.0228	0.0258
91	0.0365	0.0228	0.0137
92	0.0434	0.0228	0.0206
93	0.0325	0.0228	0.0097
94	0.0388	0.0228	0.0160
95	0.0459	0.0228	0.0231
96	0.0346	0.0228	0.0118
97	0.0411	0.0228	0.0183
98	0.0484	0.0228	0.0256
99	0.0367	0.0228	0.0139
100	0.0434	0.0228	0.0206

Table 2: CLT approximation for $p = \frac{1}{3}$

n	q_n	q (CLT)	error
1	0.5000	0.0228	0.4772
2	0.2500	0.0228	0.2272
3	0.1250	0.0228	0.1022
4	0.3125	0.0228	0.2897
5	0.1875	0.0228	0.1647
6	0.1094	0.0228	0.0866
7	0.0625	0.0228	0.0397
8	0.1445	0.0228	0.1217
9	0.0898	0.0228	0.0670
10	0.0547	0.0228	0.0319
11	0.1133	0.0228	0.0905
12	0.0730	0.0228	0.0502
13	0.0461	0.0228	0.0233
14	0.0898	0.0228	0.0670
15	0.0592	0.0228	0.0364
16	0.1051	0.0228	0.0823
17	0.0717	0.0228	0.0489
18	0.0481	0.0228	0.0253
19	0.0835	0.0228	0.0607
20	0.0577	0.0228	0.0349
21	0.0392	0.0228	0.0164
22	0.0669	0.0228	0.0441
23	0.0466	0.0228	0.0238
24	0.0758	0.0228	0.0530
25	0.0539	0.0228	0.0311
26	0.0378	0.0228	0.0150
27	0.0610	0.0228	0.0382
28	0.0436	0.0228	0.0208
29	0.0680	0.0228	0.0452
30	0.0494	0.0228	0.0266
31	0.0354	0.0228	0.0126
32	0.0551	0.0228	0.0323
33	0.0401	0.0228	0.0173
34	0.0607	0.0228	0.0379
35	0.0448	0.0228	0.0220
36	0.0662	0.0228	0.0434
37	0.0494	0.0228	0.0266
38	0.0365	0.0228	0.0137
39	0.0541	0.0228	0.0313
40	0.0403	0.0228	0.0175
41	0.0586	0.0228	0.0358
42	0.0442	0.0228	0.0214
43	0.0330	0.0228	0.0102
44	0.0481	0.0228	0.0253
45	0.0362	0.0228	0.0134
46	0.0519	0.0228	0.0291
47	0.0395	0.0228	0.0167
48	0.0557	0.0228	0.0329
49	0.0427	0.0228	0.0199
50	0.0325	0.0228	0.0097

n	q_n	q (CLT)	error
51	0.0460	0.0228	0.0232
52	0.0352	0.0228	0.0124
53	0.0492	0.0228	0.0264
54	0.0380	0.0228	0.0152
55	0.0524	0.0228	0.0296
56	0.0407	0.0228	0.0179
57	0.0314	0.0228	0.0086
58	0.0435	0.0228	0.0207
59	0.0337	0.0228	0.0109
60	0.0462	0.0228	0.0234
61	0.0361	0.0228	0.0133
62	0.0490	0.0228	0.0262
63	0.0385	0.0228	0.0157
64	0.0517	0.0228	0.0289
65	0.0408	0.0228	0.0180
66	0.0320	0.0228	0.0092
67	0.0432	0.0228	0.0204
68	0.0341	0.0228	0.0113
69	0.0456	0.0228	0.0228
70	0.0361	0.0228	0.0133
71	0.0480	0.0228	0.0252
72	0.0382	0.0228	0.0154
73	0.0302	0.0228	0.0074
74	0.0403	0.0228	0.0175
75	0.0320	0.0228	0.0092
76	0.0423	0.0228	0.0195
77	0.0338	0.0228	0.0110
78	0.0444	0.0228	0.0216
79	0.0356	0.0228	0.0128
80	0.0465	0.0228	0.0237
81	0.0374	0.0228	0.0146
82	0.0299	0.0228	0.0071
83	0.0392	0.0228	0.0164
84	0.0315	0.0228	0.0087
85	0.0410	0.0228	0.0182
86	0.0331	0.0228	0.0103
87	0.0428	0.0228	0.0200
88	0.0347	0.0228	0.0119
89	0.0447	0.0228	0.0219
90	0.0363	0.0228	0.0135
91	0.0293	0.0228	0.0065
92	0.0379	0.0228	0.0151
93	0.0307	0.0228	0.0079
94	0.0395	0.0228	0.0167
95	0.0321	0.0228	0.0093
96	0.0411	0.0228	0.0183
97	0.0335	0.0228	0.0107
98	0.0427	0.0228	0.0199
99	0.0350	0.0228	0.0122
100	0.0443	0.0228	0.0215

Table 3: CLT approximation for $p = \frac{1}{2}$

2. *Analytical investigation of the exponential distribution.*
X is an exponential random variable with parameter $\lambda > 0$ if the density is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and 0 elsewhere.

(a) Find the CDF of X and show your calculation.

The CDF of X can be defined in terms of its PDF

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)dt \\ &= \int_{-\infty}^{0} 0dt + \int_{0}^{x} \lambda e^{-\lambda t} dt \\ &= \lambda \left[\frac{-e^{-\lambda t}}{\lambda} \right]_0^x \\ &= -e^{-\lambda x} + e^0 \\ &= 1 - e^{-\lambda x} \end{aligned}$$

(b) Calculate the mean of X.

The mean value of a continous random variable is given by

$$\begin{aligned} \mu &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \lambda \int_{0}^{\infty} x e^{-\lambda x} dx \end{aligned}$$

solve using integration by parts

$$\int u dv = uv - \int v du$$

$$\begin{aligned} u &= x \\ du &= dx \end{aligned}$$

$$\begin{aligned} dv &= e^{-\lambda x} dx \\ v &= \frac{-e^{-\lambda x}}{\lambda} \end{aligned}$$

Using the integration by parts formula

$$\begin{aligned} \mu &= \lambda \left(\left[-x \frac{e^{-\lambda x}}{\lambda} \right]_0^{\infty} - \int_0^{\infty} \frac{-e^{-\lambda x}}{\lambda} dx \right) \\ &= \left[-x e^{-\lambda x} \right]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx \\ &= [0 - 0] + \left[\frac{-e^{-\lambda x}}{\lambda} \right]_0^{\infty} \\ &= \lambda^{-1} \end{aligned}$$

(c) Calculate the variance of X.

The variance of a continous random variable is defined by

$$\begin{aligned}
Var(X) &= \int (x-\mu)^2 f(x)dx \\
&= \int_0^\infty \left(x-\lambda^{-1}\right)^2 \lambda e^{-\lambda x}dx \\
&= \int_0^\infty \left(x^2 - 2x\lambda^{-1} + \lambda^{-2}\right) \lambda e^{-\lambda x}dx \\
&= \int_0^\infty \lambda x^2 e^{-\lambda x} - 2xe^{-\lambda x} + \lambda^{-1}e^{-\lambda x}dx
\end{aligned}$$

Solve first integral using intergration by parts
let:

$$\begin{aligned}
u &= x^2 \\
du &= 2xdx
\end{aligned}$$

$$\begin{aligned}
dv &= e^{-\lambda x}dx \\
v &= \frac{-e^{-\lambda x}}{\lambda}
\end{aligned}$$

Sub back into variance equation, the center integrals will cancel

$$\begin{aligned}
Var(X) &= \left(\lambda\left[\frac{-x^2e^{-\lambda x}}{\lambda}\right]_0^\infty + \lambda\int_0^\infty 2x\frac{e^{-\lambda x}}{\lambda}dx\right) - \int_0^\infty 2xe^{-\lambda x}dx + \int_0^\infty \lambda^{-1}e^{-\lambda x}dx \\
&= \lambda\left[\frac{-x^2e^{-\lambda x}}{\lambda}\right]_0^\infty + \int_0^\infty \lambda^{-1}e^{-\lambda x}dx \\
&= [0-0] + \left[\frac{-e^{-\lambda x}}{\lambda^2}\right]_0^\infty \\
&= \lambda^{-2}
\end{aligned}$$

(d) Let $M_k = E[X^k]$. What is M_0? Find a recursive formula for M_k in terms of M_{k-1}.

$$\begin{aligned}
M_k &= E\left[X^k\right] \\
&= E\left[g(X)\right]
\end{aligned}$$

where

$$g(X) = X^k$$

and

$$f(x) = \lambda e^{-\lambda x}$$

using $E\left[g\left(X\right)\right] = \int_{-\infty}^{\infty} g\left(x\right) f\left(x\right) dx$

$$M_k = \int_0^\infty x^k \lambda e^{-\lambda x}dx$$

calculate M_0

$$\begin{aligned}
M_0 &= \int_0^\infty x^0 \lambda e^{-\lambda x}dx \\
&= \left[-e^{-\lambda x}\right]_0^\infty \\
&= 1
\end{aligned}$$

To find a recursive formula we will first try to solve unsing integration by parts

$$M_k = \int_0^\infty x^k \lambda e^{-\lambda x} dx$$

$$\begin{aligned} u &= x^k \\ du &= kx^{k-1} dx \end{aligned}$$

$$\begin{aligned} dv &= \lambda e^{-\lambda x} dx \\ v &= -e^{-\lambda x} \end{aligned}$$

$$\begin{aligned} M_k &= \left[-x^k e^{-\lambda x}\right]_0^\infty - \int_0^\infty -e^{-\lambda x} k x^{k-1} dx \\ &= k \int_0^\infty e^{-\lambda x} x^{k-1} dx \end{aligned}$$

Now see what M_{k-1} equals

$$\begin{aligned} M_{k-1} &= \int_0^\infty x^{k-1} \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^\infty e^{-\lambda x} x^{k-1} dx \end{aligned}$$

sub back to get the solution

$$M_k = k \frac{M_{k-1}}{\lambda}$$

3. *Properties of variance estimators.*
Let $X_1, ..., X_n$ be a sequence of independent uniform $[0, 1]$ random variables.

(a) Calculate $\sigma^2 = Var(X)$

The variance of a continuous random variable is given by:

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

where

$$f(x) = 1$$

and

$$\begin{aligned} \mu &= \int_0^1 x f(x) dx \\ &= \int_0^1 x dx \\ &= \left[\frac{x^2}{2} \right]_0^1 \\ &= \frac{1}{2} \end{aligned}$$

therefore

$$\begin{aligned} Var(X) &= \int_0^1 (x - \frac{1}{2})^2 dx \\ &= \int_0^1 x^2 - x + \frac{1}{4} dx \\ &= \left[\frac{x^3}{3} - \frac{x^2}{2} + \frac{1}{4} x \right]_0^1 \\ &= \frac{1}{3} - \frac{1}{2} + \frac{1}{4} \\ &= \frac{1}{12} \end{aligned}$$

(b) Estimate σ^2 using the sample variance. Plot your estimate as an increasing function of the sample size, n.

The sample variance is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$

This can be plotted with simulation

```
N = 200;
X = rand(1, N);

var = zeros(N, 1);

for n = 1:N
    mean = sum(X(1:n)) / n;
    var(n) = sum((X(1:n) - mean) .^ 2) / n;
end
```

```
scatter(1:N, var)
xlabel('$$n$$', 'interpreter', 'latex')
ylabel('$$E[\sigma^2]=s^2$$', 'interpreter', 'latex')
```

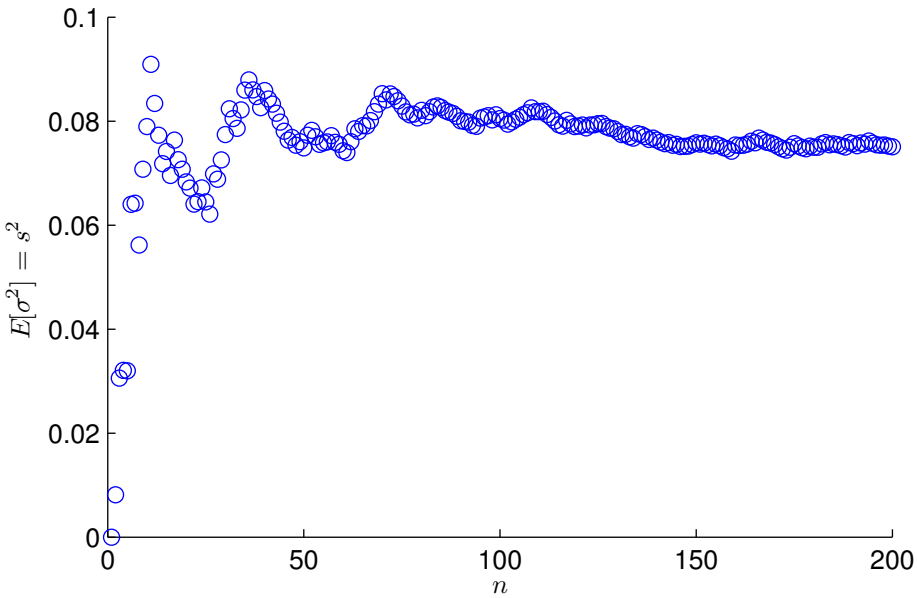


Figure 3: Plot of estimation of population variance

And it can be seen that as the number of samples is increased the estimator of the population variance approaches the actual value of $\frac{1}{12}$

(c) Fix $n = 5$. Use simulation to plot the distribution of the sample variance. What is the mean of the distribution?

```
n = 5;
n_simulations = 100000;

var = zeros(1, n_simulations);

for i = 1:n_simulations
        X = rand(1, n);
    mean  = sum(X) / n;
    var(i) = sum((X - mean) .^ 2) / (n - 1);
end

n_bins = 20;

[a x] = hist(var, n_bins);
bar(x, a ./ sum(a), 'hist');

xlabel('Var(X)')
ylabel('proportion');
```

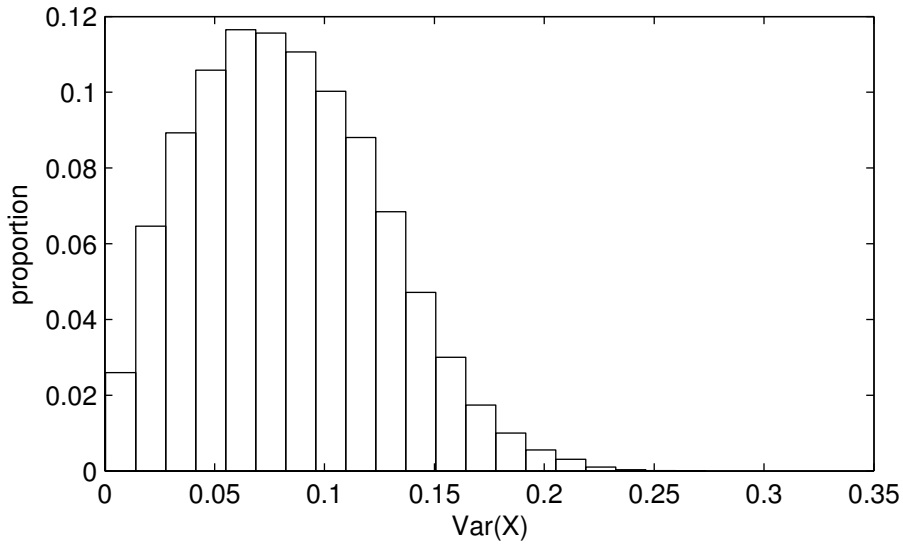


Figure 4: Distribution of $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ variance estimator for $n = 5$ and 100000 simulations

The mean of this distribution can be calculated with the following matlab script

```
sum(var) / n_simulations
```

And gives the value 0.0832 which is close to the population variance of $\frac{1}{12} \sim 0.0833$

(d) Let $S^2 = \frac{\sum_{i=1}^{n}(X_i - \frac{1}{2})^2}{n}$. Is this an unbiased estimator of the variance? Prove your result.

The $\frac{1}{2}$ is the population mean μ

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

To check if it is biased we will see if the expected value of the estimator is the same as the population variance

$$\begin{aligned}
E\left[S^2\right] &= E\left[\frac{\sum_{i=1}^{n}\left(X_i - \mu\right)^2}{n}\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i^2 - 2\mu X_i + \mu^2\right)\right] \\
nE\left[S^2\right] &= E\left[\sum_{i=1}^{n}X_i^2 - 2\mu\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\mu^2\right] \\
&= E\left[\sum_{i=1}^{n}X_i^2\right] - E\left[2\mu\sum_{i=1}^{n}X_i\right] + E\left[\sum_{i=1}^{n}\mu^2\right] \\
&= E\left[\sum_{i=1}^{n}X_i^2\right] - E\left[2\mu\left(nX\right)\right] + E\left[n\mu^2\right] \\
&= nE\left[X^2\right] - 2n\mu E\left[X\right] + n\mu^2 \\
E\left[S^2\right] &= E\left[X^2\right] - 2\mu E\left[X\right] + \mu^2
\end{aligned}$$

The expected value of X is the population mean μ
From the definition of variance

$$\begin{aligned} Var(X) &= E\left[(X-\mu)^2\right] \\ &= E\left[X^2\right] - E\left[X\right]^2 \end{aligned}$$

$$E\left[X^2\right] = \sigma^2 + \mu^2$$

sub back in

$$\begin{aligned} E\left[S^2\right] &= \left(\sigma^2 + \mu^2\right) - 2\mu\left(\mu\right) + \mu^2 \\ &= \sigma^2 \end{aligned}$$

Therefore the estimator is unbiased as it gives the actual population variance. The reason is because it is using the population mean rather than the sample mean, which is why the $s_n^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$ estimator is said to be biased.

Show your result by means of simulation (use $n = 5$).

This can be simulated using the same code as for question 3c, just changing the estimator.

```
n = 5;
n_simulations = 10000000;

mu = 0.5;

var = zeros(1, n_simulations);

for i = 1:n_simulations
    X = rand(1, n);
    var(i) = sum((X - mu) .^ 2) / n;
end

n_bins = 20;
[a x] = hist(var, n_bins);
bar(x, a ./ sum(a), 'hist');

xlabel('$$E[S^2]$$', 'interpreter', 'latex')
ylabel('proportion');
```

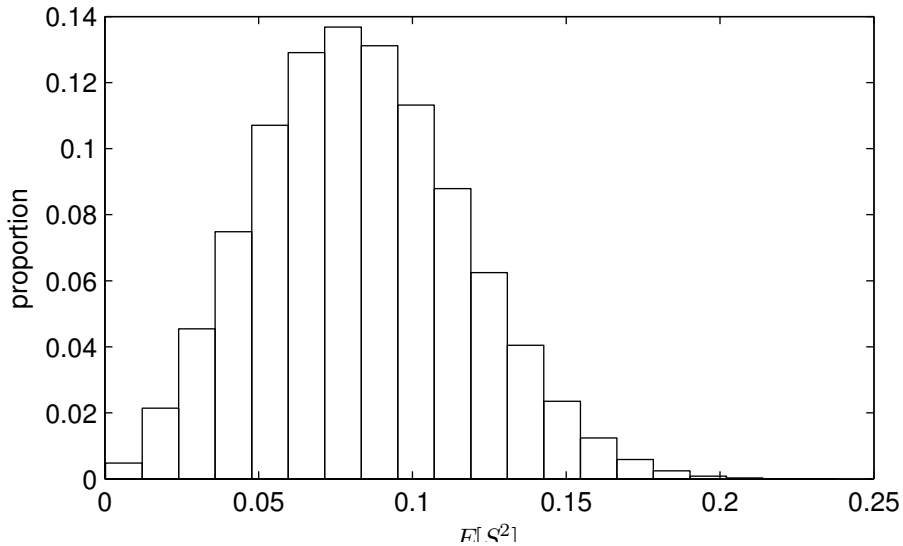


Figure 5: Distribution of $S^2 = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}$ variance estimator for $n = 5$ and 10000000 simulations

And the mean of this variance estimator is 0.0833 which is the same as the population variance, showing that it is indeed unbiased.

4. *The Gamma distribution.*
X is a Gamma random variable with parameters $\lambda > 0$ and $\alpha > 0$, if the density is $f(x) = Cx^{\alpha-1}e^{-\lambda x}$ for $x \geq 0$ and some $C > 0$.

(a) Find an expression for the normalizing constant C.

The area under the density function must be 1

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x)dx \\ &= C\int_{0}^{\infty} x^{\alpha-1}e^{-\lambda x}dx \end{aligned}$$

Find solution to the integral
let

$$G_\alpha = \int_{0}^{\infty} x^{\alpha-1}e^{-\lambda x}dx$$

solve using integration by parts

$$\begin{aligned} u &= x^{\alpha-1} \\ du &= (\alpha-1)\,x^{\alpha-2}dx \end{aligned}$$

$$\begin{aligned} dv &= e^{-\lambda x}dx \\ v &= \frac{-e^{-\lambda x}}{\lambda} \end{aligned}$$

$$\begin{aligned} G_\alpha &= \left[\frac{-x^{\alpha-1}e^{-\lambda x}}{\lambda}\right]_0^{\infty} + \int_0^{\infty} \frac{e^{-\lambda x}}{\lambda}(\alpha-1)\,x^{\alpha-2}dx \\ &= \left(\frac{\alpha-1}{\lambda}\right)\int_0^{\infty} x^{\alpha-2}e^{-\lambda x}dx \end{aligned}$$

try to get something that looks like the integral

$$\begin{aligned} G_{\alpha-1} &= \int_0^{\infty} x^{(\alpha-1)-1}e^{-\lambda x}dx \\ &= \int_0^{\infty} x^{\alpha-2}e^{-\lambda x}dx \end{aligned}$$

sub back

$$\begin{aligned} G_\alpha &= \left(\frac{\alpha-1}{\lambda}\right)\int_0^{\infty} x^{\alpha-2}e^{-\lambda x}dx \\ &= \left(\frac{\alpha-1}{\lambda}\right)G_{\alpha-1} \end{aligned}$$

evaluate the first few

$$\begin{aligned} G_1 &= \int_0^{\infty} x^{1-1}e^{-\lambda x}dx \\ &= \left[\frac{-e^{-\lambda x}}{\lambda}\right]_0^{\infty} \\ &= \frac{1}{\lambda} \end{aligned}$$

$$
\begin{aligned}
G_2 &= \left(\frac{\alpha-1}{\lambda}\right)\left(\frac{1}{\lambda}\right) \\
&= \frac{1}{\lambda^2}
\end{aligned}
$$

$$
\begin{aligned}
G_3 &= \left(\frac{\alpha-1}{\lambda}\right)\left(\frac{1}{\lambda^2}\right) \\
&= \frac{2}{\lambda^3}
\end{aligned}
$$

$$
\begin{aligned}
G_4 &= \left(\frac{\alpha-1}{\lambda}\right)\left(\frac{2}{\lambda^3}\right) \\
&= \frac{6}{\lambda^4}
\end{aligned}
$$

pattern emerges

$$
G_\alpha = \frac{(\alpha-1)!}{\lambda^\alpha}
$$

sub back

$$
\begin{aligned}
1 &= C\int_0^\infty x^{\alpha-1}e^{-\lambda x}dx \\
&= C\left(\frac{(\alpha-1)!}{\lambda^\alpha}\right) \\
C &= \frac{\lambda^\alpha}{(\alpha-1)!}
\end{aligned}
$$

(b) Plot gamma densities for different parameter values.

Plots for differnt λ and α values can be made with the following Matlab script

```
x = 0:0.1:3;

alpha = 1;
style = {'-' '--' 'o' '*'}

for lambda = 1:4
    C = lambda ^ alpha / factorial(alpha - 1);
    f = C .* x .^ (alpha - 1) .* exp(-lambda .* x);

    plot(x, f, style{lambda})
    hold on
end
xlabel('$$x$$', 'interpreter', 'latex')
ylabel('$$f(x)=Cx^{\alpha-1}e^{-\lambda x}$$', 'interpreter', ...
       'latex')
h = legend('$$\lambda=1$$', '$$\lambda=2$$', '$$\lambda=3$$', ...
           '$$\lambda=4$$');
set(h, 'interpreter', 'latex')
```

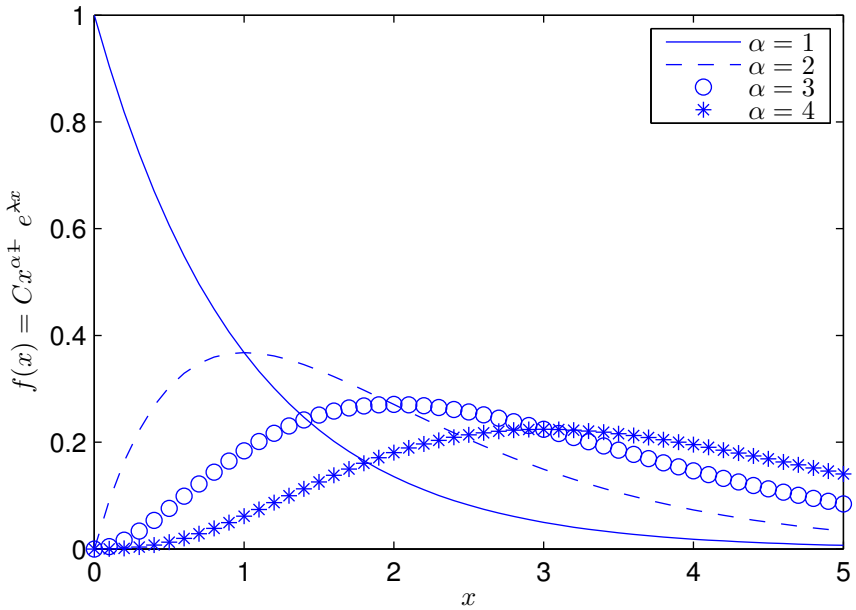


Figure 6: Plot for different α values (with $\lambda = 1$)

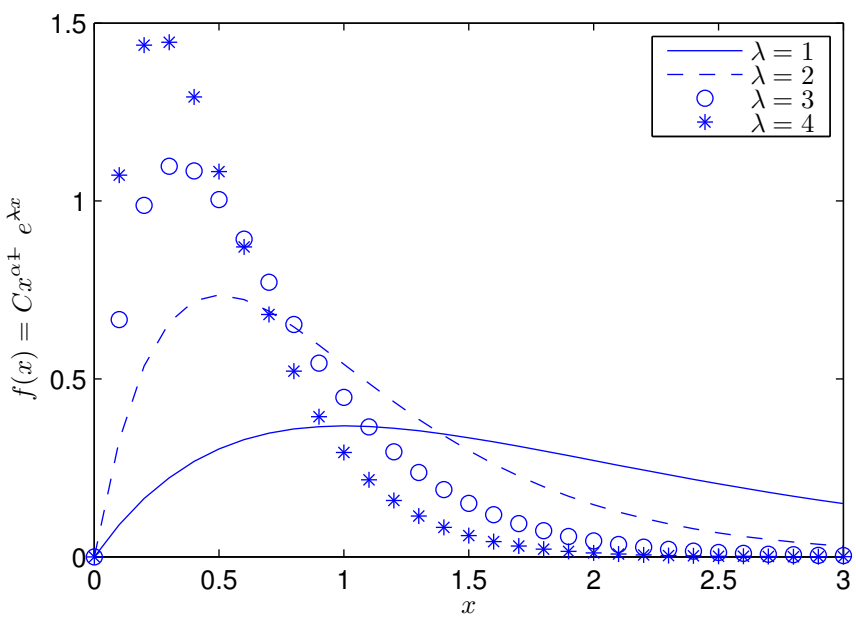


Figure 7: Plot for different λ values (with $\alpha = 2$

(c) Is the exponential distribution a special case?

The exponential distribution is defined as

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

If we let $\alpha = 1$, the gamma distribution becomes

$$\begin{aligned} f(x) &= \begin{cases} \frac{\lambda^\alpha}{(\alpha-1)!} x^{\alpha-1} e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \\ &= \begin{cases} \frac{\lambda^1}{0!} x^0 e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \\ &= \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \end{aligned}$$

Therefore the exponential distribution is a special case of the gamma distribution, having $\alpha = 1$

(d) Find the mean and variance of the gamma distribution - show your calculations.

The expected value of a continous random variable is given by

$$\begin{aligned} E[x] &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_0^{\infty} x \left(C x^{\alpha-1} e^{-\lambda x} \right) dx \\ &= \frac{\lambda^\alpha}{(\alpha-1)!} \int_0^{\infty} x^\alpha e^{-\lambda x} dx \end{aligned}$$

from the calculation of C we know that

$$\begin{aligned} \frac{\lambda^\alpha}{(\alpha-1)!} \int_0^{\infty} x^{\alpha-1} e^{-\lambda x} dx &= 1 \\ \int_0^{\infty} x^{\alpha-1} e^{-\lambda x} dx &= \frac{(\alpha-1)!}{\lambda^\alpha} \end{aligned}$$

if α is increased by 1

$$\begin{aligned} \int_0^{\infty} x^{(\alpha+1)-1} e^{-\lambda x} dx &= \frac{((\alpha+1)-1)!}{\lambda^{(\alpha+1)}} \\ \int_0^{\infty} x^\alpha e^{-\lambda x} dx &= \frac{\alpha!}{\lambda^{\alpha+1}} \end{aligned}$$

sub this back into the formula for the $E[X]$

$$\begin{aligned} E[x] &= \frac{\lambda^\alpha}{(\alpha-1)!} \int_0^{\infty} x^\alpha e^{-\lambda x} dx \\ &= \frac{\lambda^\alpha}{(\alpha-1)!} \left(\frac{\alpha!}{\lambda^{\alpha+1}} \right) \\ &= \frac{\lambda^\alpha}{(\alpha-1)!} \left(\frac{\alpha(\alpha-1)!}{\lambda^{\alpha+1}} \right) \\ &= \frac{\alpha}{\lambda} \end{aligned}$$

The variance of a continous random variable is given by

$$\begin{aligned} Var(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \\ &= \int_0^{\infty} \left(x - \frac{\alpha}{\lambda} \right)^2 \left(C x^{\alpha-1} e^{-\lambda x} \right) dx \\ &= C \int_0^{\infty} \left(x^2 - \frac{2\alpha x}{\lambda} + \frac{\alpha^2}{\lambda^2} \right) \left(x^{\alpha-1} e^{-\lambda x} \right) dx \\ \frac{Var(X)}{C} &= \int_0^{\infty} x^{\alpha+1} e^{-\lambda x} dx - \frac{2\alpha}{\lambda} \int_0^{\infty} x e^{-\lambda x} dx + \frac{\alpha^2}{\lambda^2} \int_0^{\infty} x^{\alpha-1} e^{-\lambda x} dx \end{aligned}$$

using previous results

$$\begin{aligned}
\frac{Var(X)}{C} &= \left(\frac{(\alpha+1)!}{\lambda^{\alpha+2}}\right) - \frac{2\alpha}{\lambda}\left(\frac{\alpha!}{\lambda^{\alpha+1}}\right) + \frac{\alpha^2}{\lambda^2}\left(\frac{(\alpha-1)!}{\lambda^{\alpha}}\right) \\
&= \frac{(\alpha+1)!}{\lambda^2\lambda^{\alpha}} - \frac{2\alpha\cdot\alpha!}{\lambda^2\lambda^{\alpha}} + \frac{\alpha^2(\alpha-1)!}{\lambda^2\lambda^{\alpha}} \\
Var(X) &= \frac{\lambda^{\alpha}}{(\alpha-1)!}\left(\frac{(\alpha+1)! - 2\alpha^2(\alpha-1)! + \alpha^2(\alpha-1)!}{\lambda^2\lambda^{\alpha}}\right) \\
&= \frac{1}{(\alpha-1)!}\left(\frac{\alpha(\alpha+1)(\alpha-1)! - \alpha^2(\alpha-1)!}{\lambda^2}\right) \\
&= \frac{\alpha(\alpha+1) - \alpha^2}{\lambda^2} \\
&= \frac{\alpha}{\lambda^2}
\end{aligned}$$

(e) Assume now that $X_1, ..., X_n$ is a random sample from a gamma distribution with unknown parameters λ and α. One way to estimate the parameters is the method of moments. In the method, you equate the sample mean and sample variance to the mean and variance expressions and solve for λ and α. Write the equations for the method of moments and then write the resulting estimators (functions of the random sample), $\hat{\lambda}$ and $\hat{\alpha}$.

sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

equate to mean

$$\begin{aligned}
\overline{x} &= \frac{\hat{\alpha}}{\hat{\lambda}} \\
\hat{\alpha} &= \overline{x}\hat{\lambda} \qquad (1)
\end{aligned}$$

sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

equate to variance

$$\begin{aligned}
s^2 &= \frac{\hat{\alpha}}{\hat{\lambda}^2} \\
\hat{\alpha} &= s^2\hat{\lambda}^2 \qquad (2)
\end{aligned}$$

equate 1 and 2 to remove α

$$\begin{aligned}
\overline{x}\hat{\lambda} &= s^2\hat{\lambda}^2 \\
\hat{\lambda} &= \frac{\overline{x}}{s^2}
\end{aligned}$$

sub $\hat{\lambda}$ into 2

$$\begin{aligned}
\hat{\alpha} &= s^2\left(\frac{\overline{x}}{s^2}\right)^2 \\
&= \frac{\overline{x}^2}{s^2}
\end{aligned}$$

(f) Use simulation to check if the estimators in (d) are biased/asymptotically unbiased.

The following script estimates the gamma parameters from simulation

```
alpha = 10;
lambda = 6;
n = 1000000;

X = random('gamma', alpha, 1/lambda, 1, n);

sample_mean = sum(X) / n;
sample_var = sum((X - sample_mean) .^ 2) / (n - 1);

alpha_hat = sample_mean ^ 2 / sample_var
lambda_hat = sample_mean / sample_var
```

The results are

set	α	$\hat{\alpha}$	λ	$\hat{\lambda}$
1	1	0.9997	1	1.0014
2	10	10.0057	6	6.0023
3	2	2.0014	8	7.9982

Table 4: Estimator check