

ETC 2420/5242 Lab 4 2017

SOLUTION

Week 4

Purpose

This lab is to examine different statistical distributions, fit distributions to samples by estimating the parameters by maximum likelihood and checking the fit with QQ-plots.

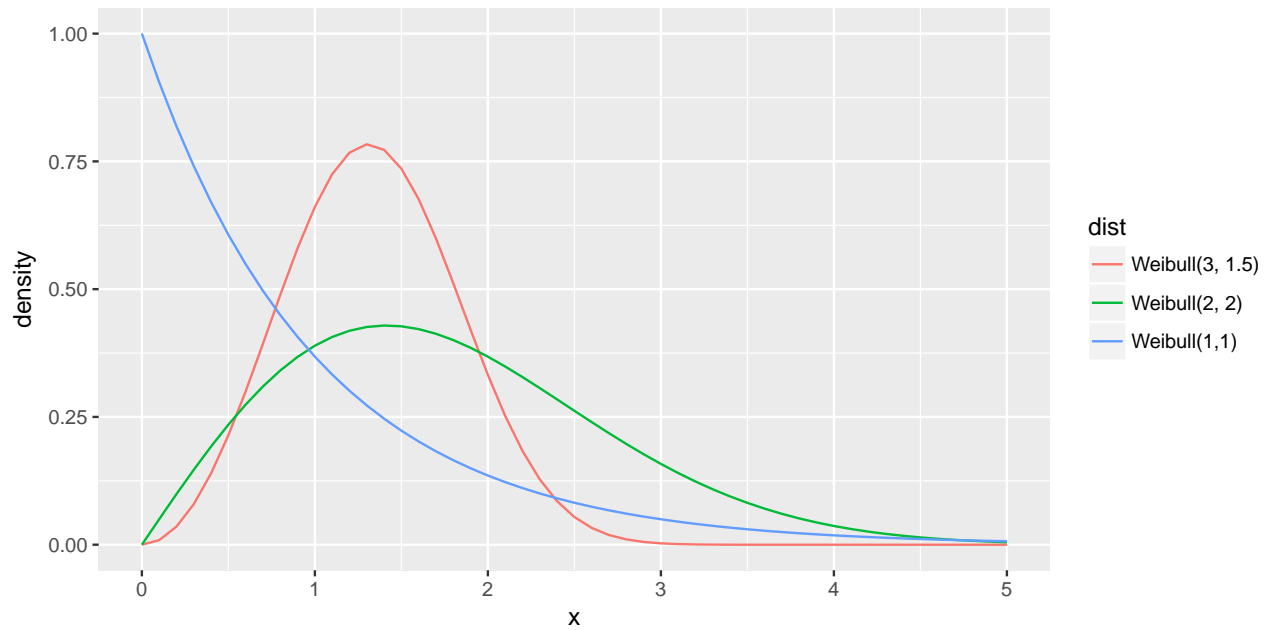
Reading

Read the material on maximum likelihood estimation at <https://onlinecourses.science.psu.edu/stat414/node/191>.

Read the code in the lecture notes from Week 3. Particularly look at the functions for making QQ-plots, computing and plotting the likelihood functions.

Warmup exercises

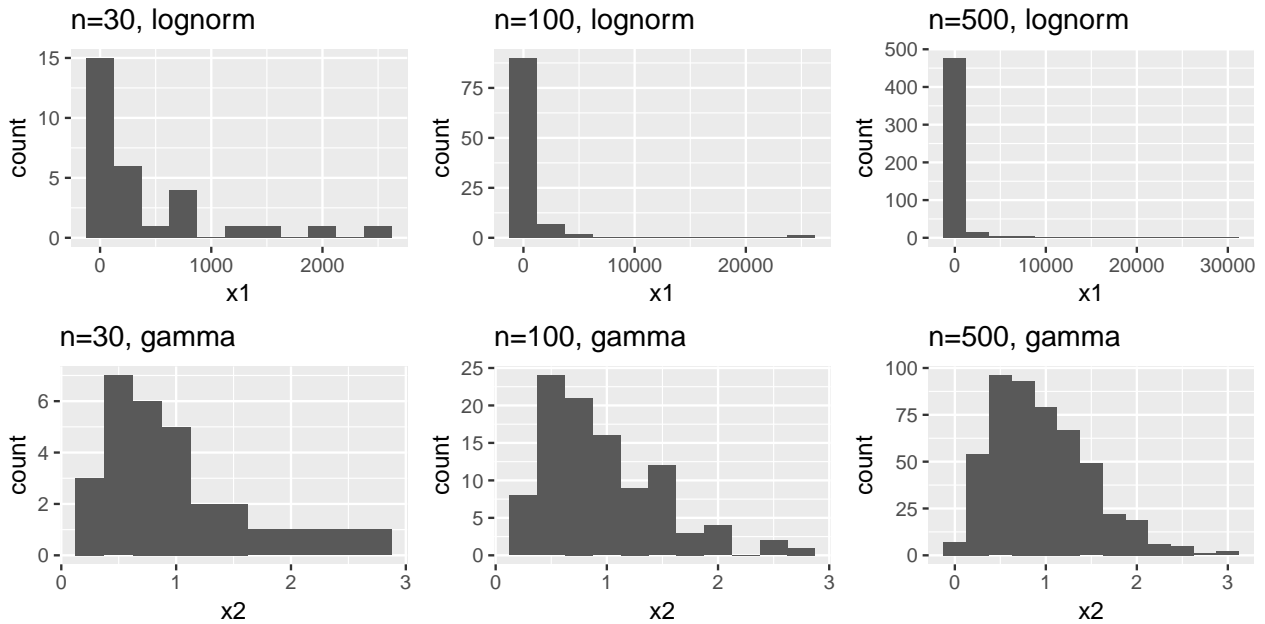
- Compute these probabilities for $X \sim N(3.3, 1.1)$
 - $P(X < 1.3)$ 0.0345182
 - $P(X > 1.9)$ 0.8984426
 - $P(1.8 < X < 2.2)$ 0.0723142
- Compute the quantile value (X) for $X \sim N(-10, 4)$ which matches these probabilities
 - $P(X < x) = 0.53$ -9.6989206
 - $P(X < x) = 0.12$ -14.6999472
 - $P(X < x) = 0.84$ -6.0221685
 - $P(X < x) = 1.2$ Not possible
- Compute the value of the density function for a $N(12, 5)$ corresponding to $X =$
 - 13.0 0.0782085
 - 4.0 0.0221842
 - 20.0 0.0221842
- Plot the density curves a
 - $Weibull(3, 1.5)$
 - $Weibull(2, 2)$
 - $Weibull(1, 1)$ on the same plot.



Question 1 (9pts)

- a. (2pt) Simulate samples of size $n = 30, 100, 500$ from these distributions
- i. Lognormal(4, 2)
 - ii. Gamma(3, 3)

```
library(gridExtra)
set.seed(4)
df_30 <- data.frame(x1=rlnorm(30, 4, 2), x2=rgamma(30, 3, 3))
p1 <- ggplot(df_30, aes(x=x1)) + geom_histogram(binwidth=250) + ggtitle("n=30, lognorm")
p2 <- ggplot(df_30, aes(x=x2)) + geom_histogram(binwidth=0.25) + ggtitle("n=30, gamma")
df_100 <- data.frame(x1=rlnorm(100, 4, 2), x2=rgamma(100, 3, 3))
p3 <- ggplot(df_100, aes(x=x1)) + geom_histogram(binwidth=2500) + ggtitle("n=100, lognorm")
p4 <- ggplot(df_100, aes(x=x2)) + geom_histogram(binwidth=0.25) + ggtitle("n=100, gamma")
df_500 <- data.frame(x1=rlnorm(500, 4, 2), x2=rgamma(500, 3, 3))
p5 <- ggplot(df_500, aes(x=x1)) + geom_histogram(binwidth=2500) + ggtitle("n=500, lognorm")
p6 <- ggplot(df_500, aes(x=x2)) + geom_histogram(binwidth=0.25) + ggtitle("n=500, gamma")
grid.arrange(p1, p3, p5, p2, p4, p6, ncol=3)
```



b. (2pt) Do an internet search to find an example of where a lognormal distribution, and a gamma distribution might be used.

Lognormal: (<http://www.statisticshowto.com/lognormal-distribution/>)

- Milk production by cows.
- Lives of industrial units with failure modes that are characterized by fatigue-stress.
- Amounts of rainfall.
- Size distributions of rainfall droplets.
- The volume of gas in a petroleum reserve.

gamma: (https://en.wikipedia.org/wiki/Gamma_distribution)

the waiting time until death

Be sure to provide the reference link for your information.

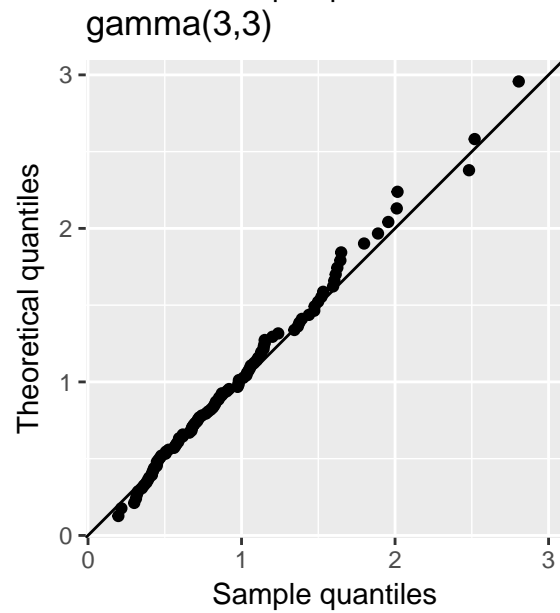
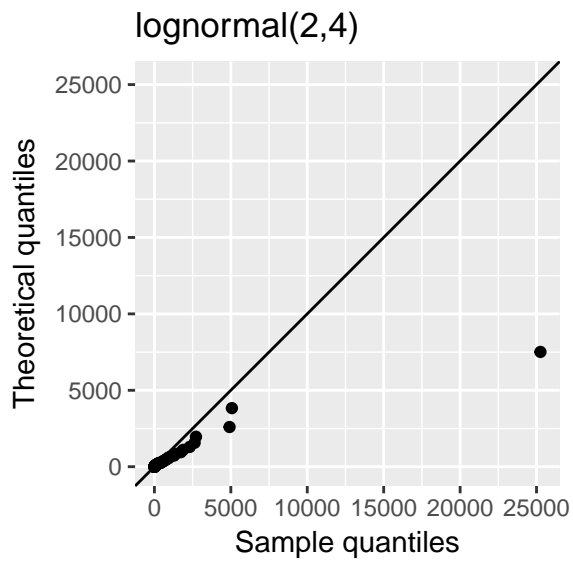
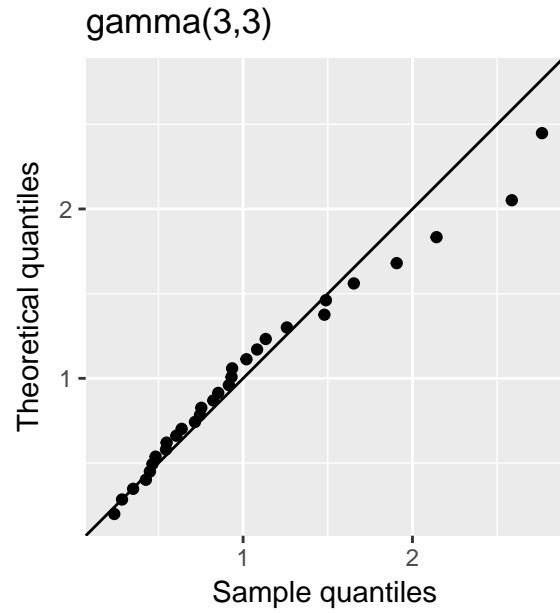
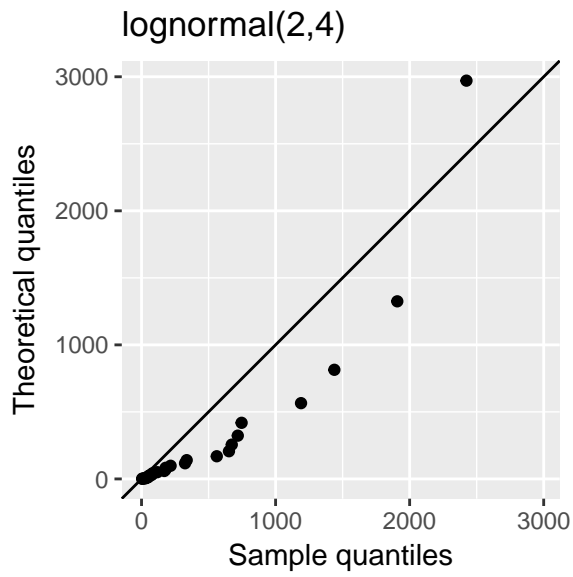
c. (5pt) Make a QQ-plot of each these samples. Explain how closely the samples, of different sizes, appears to match the theoretical distribution.

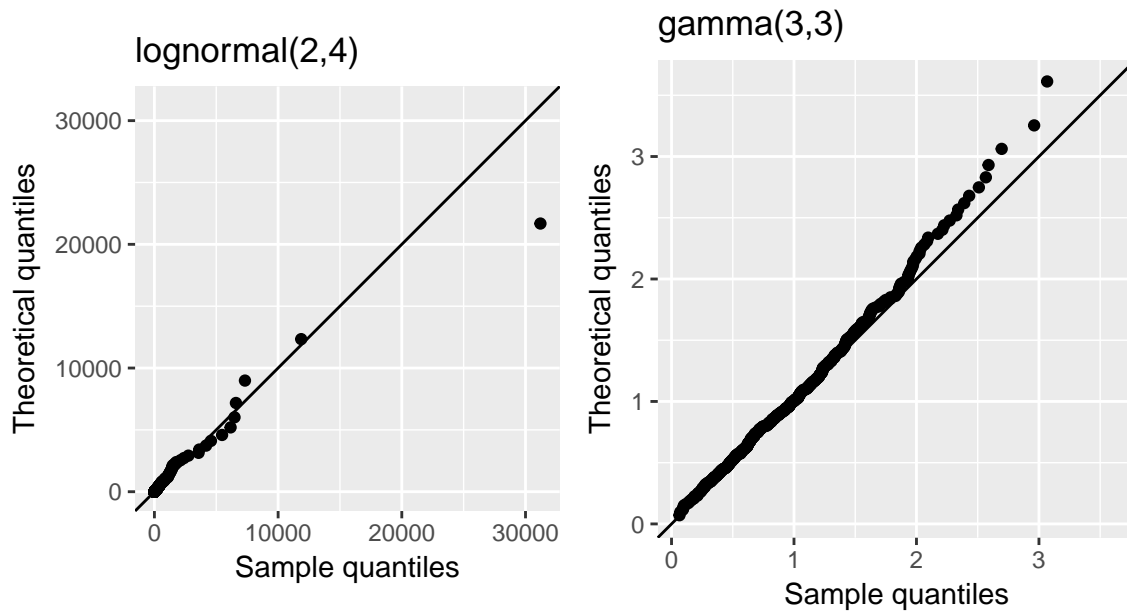
```
n <- 30
df_30$x1q = qlnorm(c(1 - 0.5^(1/n), (2:(n-1) - 0.3175) /
                    (n + 0.365), 0.5^(1/n)), 4, 2)
df_30$x2q = qgamma(c(1 - 0.5^(1/n), (2:(n-1) - 0.3175) /
                    (n + 0.365), 0.5^(1/n)), 3, 3)
lims1 <- c(min(min(df_30$x1), min(df_30$x1q)), max(max(df_30$x1), max(df_30$x1q)))
lims2 <- c(min(min(df_30$x2), min(df_30$x2q)), max(max(df_30$x2), max(df_30$x2q)))
p1 <- ggplot(df_30, aes(x=sort(x1), y=x1q)) +
  geom_abline(intercept=0, slope=1) +
  geom_point() + coord_equal() +
  xlim(lims1) + ylim(lims1) +
  xlab("Sample quantiles") + ylab("Theoretical quantiles") +
  ggtitle("lognormal(2,4)")
p2 <- ggplot(df_30, aes(x=sort(x2), y=x2q)) +
  geom_abline(intercept=0, slope=1) +
  geom_point() + coord_equal() +
  xlim(lims2) + ylim(lims2) +
  xlab("Sample quantiles") + ylab("Theoretical quantiles") +
  ggtitle("gamma(3,3)")
```

```

grid.arrange(p1, p2, ncol=2)
n <- 100
df_100$x1q = qlnorm(c(1 - 0.5^(1/n), (2:(n-1) - 0.3175) /
                    (n + 0.365), 0.5^(1/n)), 4, 2)
df_100$x2q = qgamma(c(1 - 0.5^(1/n), (2:(n-1) - 0.3175) /
                    (n + 0.365), 0.5^(1/n)), 3, 3)
lims1 <- c(min(min(df_100$x1), min(df_100$x1q)), max(max(df_100$x1), max(df_100$x1q)))
lims2 <- c(min(min(df_100$x2), min(df_100$x2q)), max(max(df_100$x2), max(df_100$x2q)))
p1 <- ggplot(df_100, aes(x=sort(x1), y=x1q)) +
  geom_abline(intercept=0, slope=1) +
  geom_point() + coord_equal() +
  xlim(lims1) + ylim(lims1) +
  xlab("Sample quantiles") + ylab("Theoretical quantiles") +
  ggtitle("lognormal(2,4)")
p2 <- ggplot(df_100, aes(x=sort(x2), y=x2q)) +
  geom_abline(intercept=0, slope=1) +
  geom_point() + coord_equal() +
  xlim(lims2) + ylim(lims2) +
  xlab("Sample quantiles") + ylab("Theoretical quantiles") +
  ggtitle("gamma(3,3)")
grid.arrange(p1, p2, ncol=2)
n <- 500
df_500$x1q = qlnorm(c(1 - 0.5^(1/n), (2:(n-1) - 0.3175) /
                    (n + 0.365), 0.5^(1/n)), 4, 2)
df_500$x2q = qgamma(c(1 - 0.5^(1/n), (2:(n-1) - 0.3175) /
                    (n + 0.365), 0.5^(1/n)), 3, 3)
lims1 <- c(min(min(df_500$x1), min(df_500$x1q)), max(max(df_500$x1), max(df_500$x1q)))
lims2 <- c(min(min(df_500$x2), min(df_500$x2q)), max(max(df_500$x2), max(df_500$x2q)))
p1 <- ggplot(df_500, aes(x=sort(x1), y=x1q)) +
  geom_abline(intercept=0, slope=1) +
  geom_point() + coord_equal() +
  xlim(lims1) + ylim(lims1) +
  xlab("Sample quantiles") + ylab("Theoretical quantiles") +
  ggtitle("lognormal(2,4)")
p2 <- ggplot(df_500, aes(x=sort(x2), y=x2q)) +
  geom_abline(intercept=0, slope=1) +
  geom_point() + coord_equal() +
  xlim(lims2) + ylim(lims2) +
  xlab("Sample quantiles") + ylab("Theoretical quantiles") +
  ggtitle("gamma(3,3)")
grid.arrange(p1, p2, ncol=2)

```





With larger sample sizes the points should lie more closely on the guide line. For the most part this is

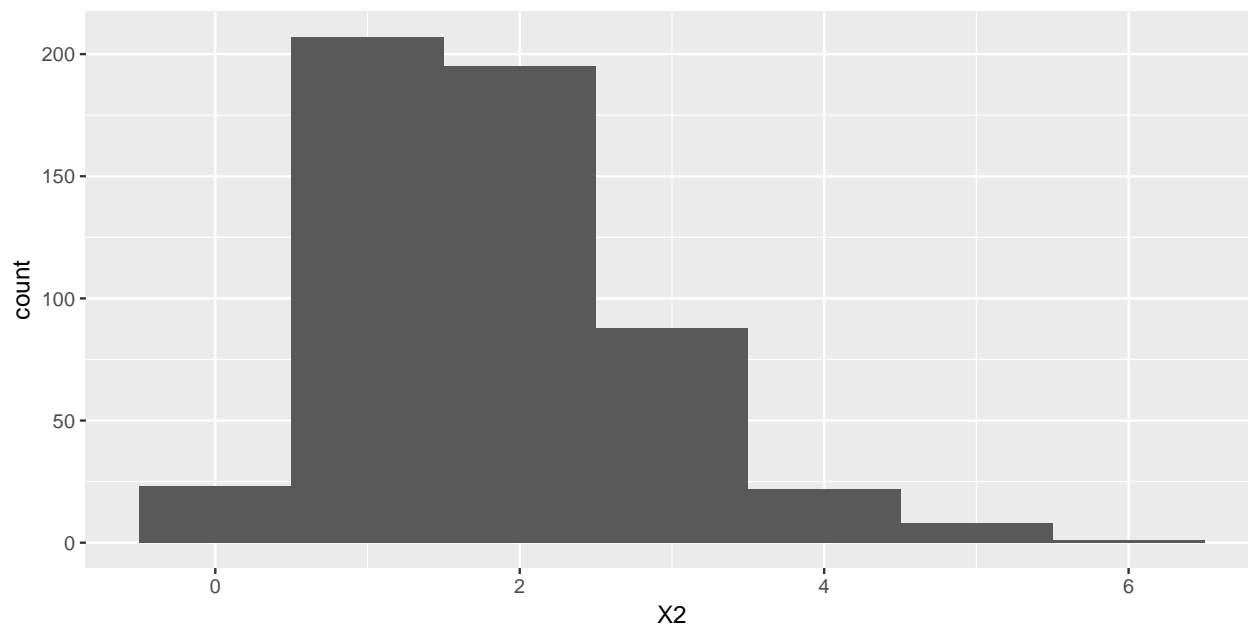
Question 2 (12pts)

Using this code, generate a sample of size $n = 544$ from a $Gamma(3.2, 1.7)$ distribution.

```
set.seed(123)
X2 <- rgamma(n=544, 3.2, 1.7)
```

- a. (2pt) Plot the sample, using a histogram, describe the shape of the distribution. unimodal, slightly right-skewed

```
ggplot(data.frame(X2), aes(x=X2)) + geom_histogram(binwidth=1)
```



- b. (1pt) What parameters of the gamma distribution were used to simulate the sample? ($\alpha = 3.2, \beta = 1.7$)
- c. (1pt) If we are to use maximum likelihood distribution what values would we expect to get as the parameter estimates? (3.2, 1.7)
- d. (2pt) Write a function to compute the likelihood function.

```
nmle <- function(x, a, b) {  
  f <- prod(dgamma(x, a, b))  
  return(f)  
}
```

- e. (3pt) Plot the likelihood function for a range of values of α, β that shows the maximum likelihood estimates for each parameter.

```
a <- seq(2.5, 4.2, 0.05)  
b <- seq(1.5, 2.5, 0.05)  
g <- expand.grid(x=a, y=b)  
g$f <- 0  
for (i in 1:nrow(g)) {  
  cat(i, "\n")  
  g$f[i] <- nmle(X2, g$x[i], g$y[i])  
}  
# 1  
# 2  
# 3  
# 4  
# 5  
# 6  
# 7  
# 8  
# 9  
# 10  
# 11  
# 12  
# 13  
# 14  
# 15  
# 16  
# 17  
# 18  
# 19  
# 20  
# 21  
# 22  
# 23  
# 24  
# 25  
# 26  
# 27  
# 28  
# 29  
# 30  
# 31  
# 32  
# 33  
# 34
```

35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87

88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140

141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193

194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246

247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299

300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352

353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405

406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458

459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511

512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564

565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617

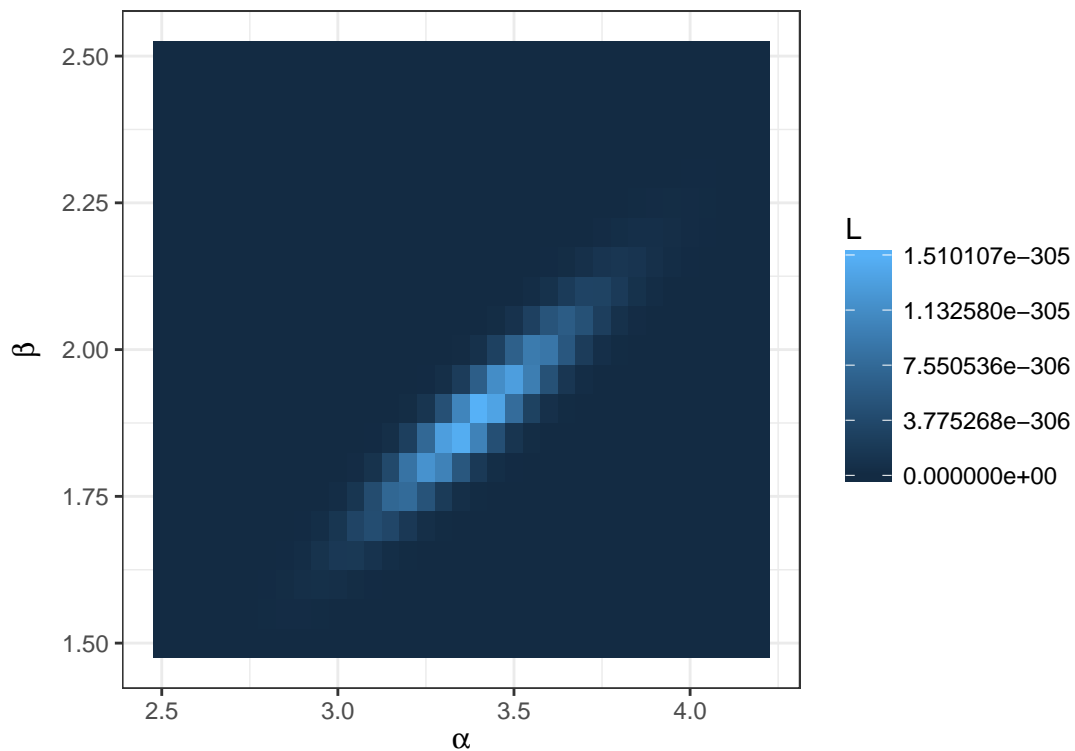
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670

671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723

```

# 724
# 725
# 726
# 727
# 728
# 729
# 730
# 731
# 732
# 733
# 734
# 735
ggplot(g, aes(x=x, y=y, fill=f)) + geom_tile() + xlab(expression(alpha)) + ylab(expression(beta)) + the
scale_fill_continuous("L") +
theme(aspect.ratio=1)

```



```

g[which.max(g$f),]
#      x      y      f
# 299 3.4  1.9 1.510107e-305

```

- f. (3pt) Look up the function `fitdistr` from the MASS library. Explain what this does. Use it to find the MLE estimates for α, β . How do these compare with the values you read off your plot? The values from the plot are very similar.

```

library(MASS)
fitdistr(X2, "gamma")
#      shape      rate
# 3.3879099  1.8832870
# (0.1961500) (0.1175305)

```

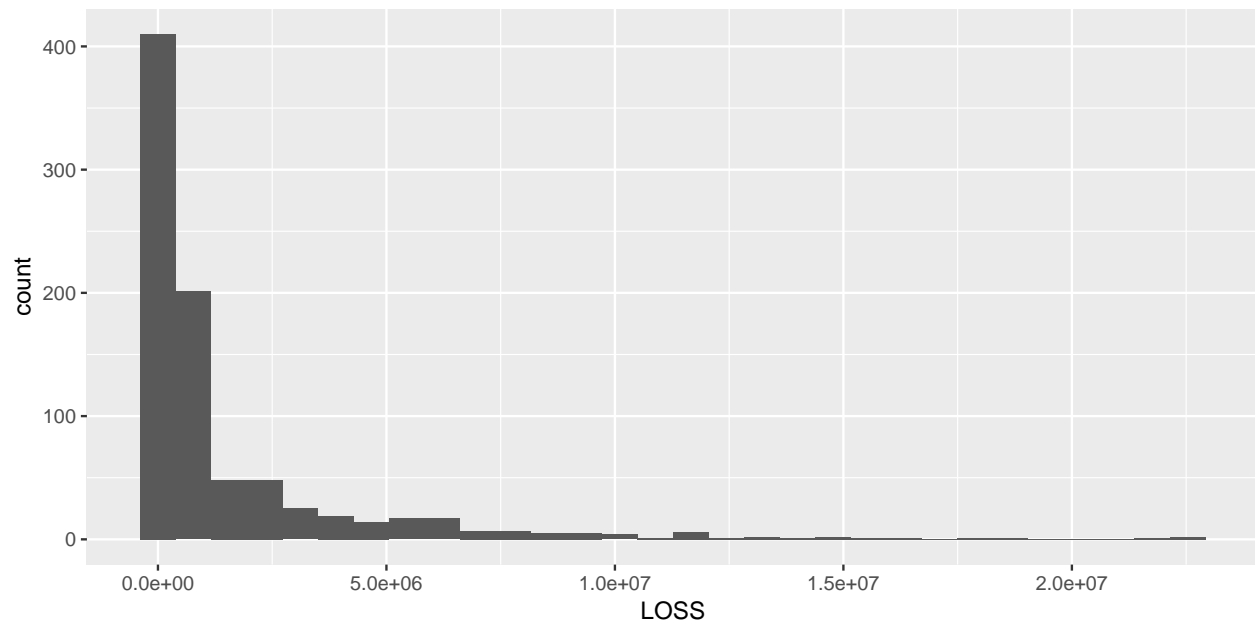
Question 3 (9pts)

Take a look at the data set `usworkcomp` from the `CASdatasets` library. Read the documentation about this data on <http://cas.uqam.ca/pub/R/web/CASdatasets-manual.pdf>.

A copy of the data is provided with the lab, in case the `CASdatasets` are not all available.

- a. (2pt) Make a histogram of the `LOSS`. Describe the shape. Heavily right-skewed, unimodal

```
ggplot(usworkcomp, aes(x=LOSS)) + geom_histogram()
```



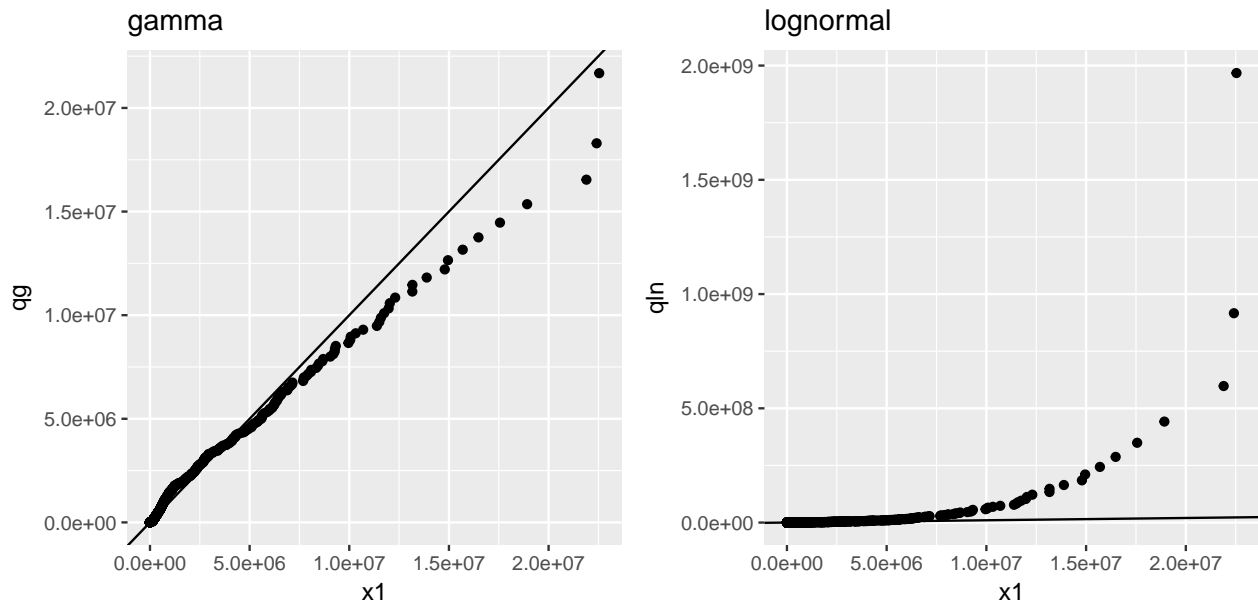
- b. (3pt) Fit both a gamma and lognormal distribution to the sample, i.e. find the MLEs.

```
fitdistr((usworkcomp$LOSS+100)/100000, "gamma")
#      shape      rate
# 0.367774667 0.023499352
# (0.014346994) (0.001614735)
fitdistr((usworkcomp$LOSS+100)/100000, "lognormal")
#      meanlog      sdlog
# 0.936916777 2.84185608
# (0.09764733) (0.06904709)
```

- c. (2pt) Produce a QQ-plot for each of the distributions.

```
n <- nrow(usworkcomp)
df <- data.frame(x1=sort(usworkcomp$LOSS))
df$qg = (qgamma(c(1 - 0.5^(1/n), (2:(n-1) - 0.3175) /
  (n + 0.365), 0.5^(1/n)), 0.367774667, 0.023499352))*100000-100
df$qln = (qlnorm(c(1 - 0.5^(1/n), (2:(n-1) - 0.3175) /
  (n + 0.365), 0.5^(1/n)), 0.936916777, 2.84185608))*100000-100
p1 <- ggplot(df, aes(x=x1, y=qg)) +
  geom_abline(intercept=0, slope=1) +
  geom_point() + theme(aspect.ratio=1) +
  ggtitle("gamma")
p2 <- ggplot(df, aes(x=x1, y=qln)) +
  geom_abline(intercept=0, slope=1) +
```

```
geom_point() + theme(aspect.ratio=1) +
ggtitle("lognormal")
grid.arrange(p1, p2, ncol=2)
```



d. (2pt) Which is the better fit to the sample? Gamma. It is fairly close, except for high values.

TURN IN

- Your .Rmd file
- Your html file that results from knitting the Rmd.
- Make sure your group members are listed as authors, one person per group will turn in the report