MAST90105 Lab and Workshop 10 Solutions

The Lab and Workshop this week covers problems arising from Module 7. The problems will be assigned to groups this week.

1 Lab

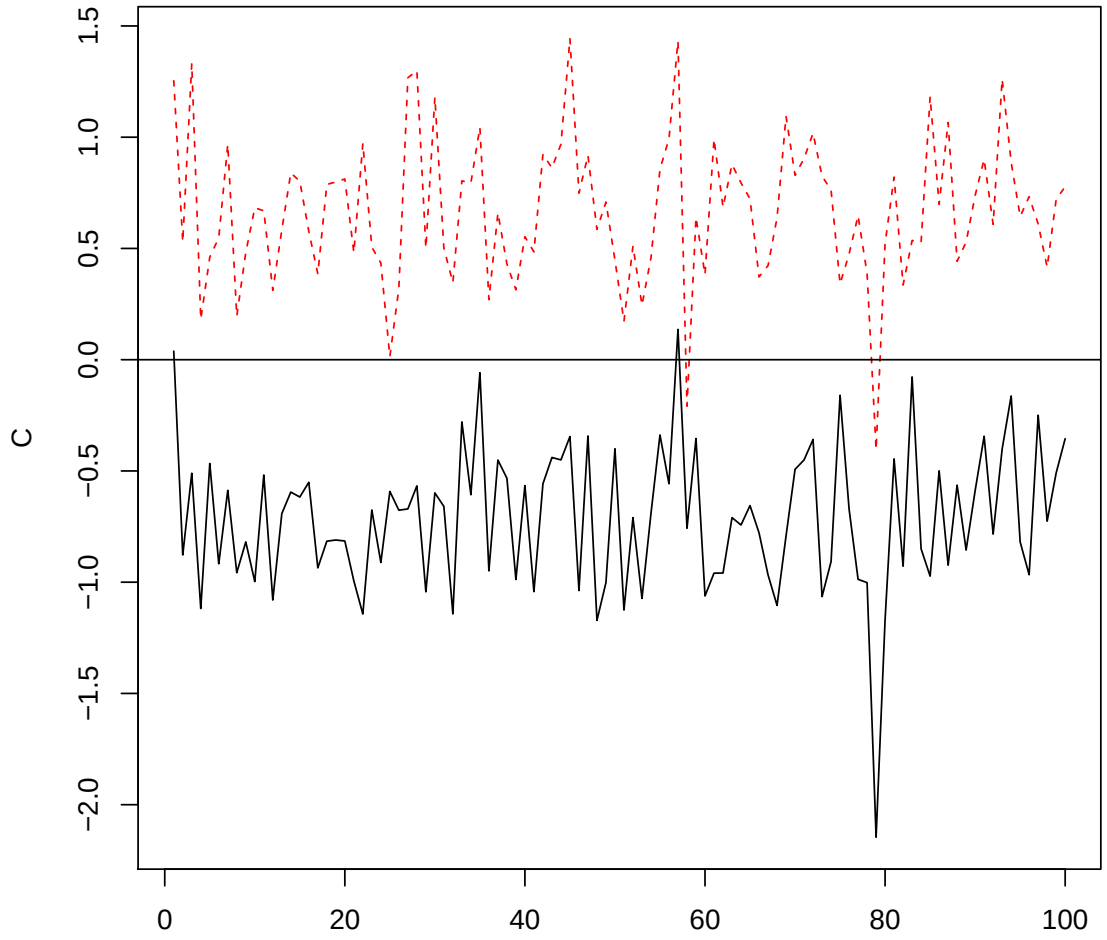
Last week's lab had a lot of Mathematica and R problems. So this week only 2 have been set.

1. How good are confidence intervals? If we repeat the experiment a large number of times we expect 95% of the confidence intervals for contain the parameter values. We can check this using simulations. Enter the following commands:

```
x = t.test(rnorm(10))
x


##
##  One Sample t-test
##
## data:  rnorm(10)
## t = -0.031222, df = 9, p-value = 0.9758
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.7460613  0.7257474
## sample estimates:
##   mean of x
## -0.01015698

names(x)

## [1] "statistic"   "parameter"   "p.value"
## [4] "conf.int"    "estimate"    "null.value"
## [7] "alternative" "method"      "data.name"

x$conf.int

## [1] -0.7460613  0.7257474
## attr(,"conf.level")
## [1] 0.95
```

   You should use the help function or your tutor to understand what each command does. Note that rnorm simulates values from $N(0, 1)$ so we know the true mean is zero. Then automate the process

```
 f=function(t){x=t.test(rnorm(t));as.vector(x$conf.int)};
f(10);

## [1] -0.03521054  0.92537722

 f(20);

## [1] -0.5737504  0.6221041

t <- as.matrix(rep(10,100));
C <- t(apply(t,1,f)); #this is a trick so we don't have to program
 matplot(C,type="l");#a matrix plot
abline(0,0)#includes a line at 0
```



Each column of the matrix C is the lower and upper bounds of a 95% confidence interval. From your plot determine how many of these intervals contain the true mean zero. Is it close to 95%? You can check as follows:

```
num = (C[, 1] < 0) & (C[, 2] > 0)
sum(num)/nrow(C)


## [1] 0.96
```

2. In class, we discussed the Newspoll outcomes from March 20 and April 3 2017. The March 20 poll reported that 675 of 1824 voters would vote first for the Government if an election were held then, and on April 3 it was 615 out of 1708 voters.

a. Starting with a uniform distribution over (0,1), find the posterior distribution for the population proportion after the March 20 .

- *Using Module 7 pp. 17-19, the posterior distribution after the March 20 poll is Beta(675+1,1824-675+1) = Beta(676,1150).*

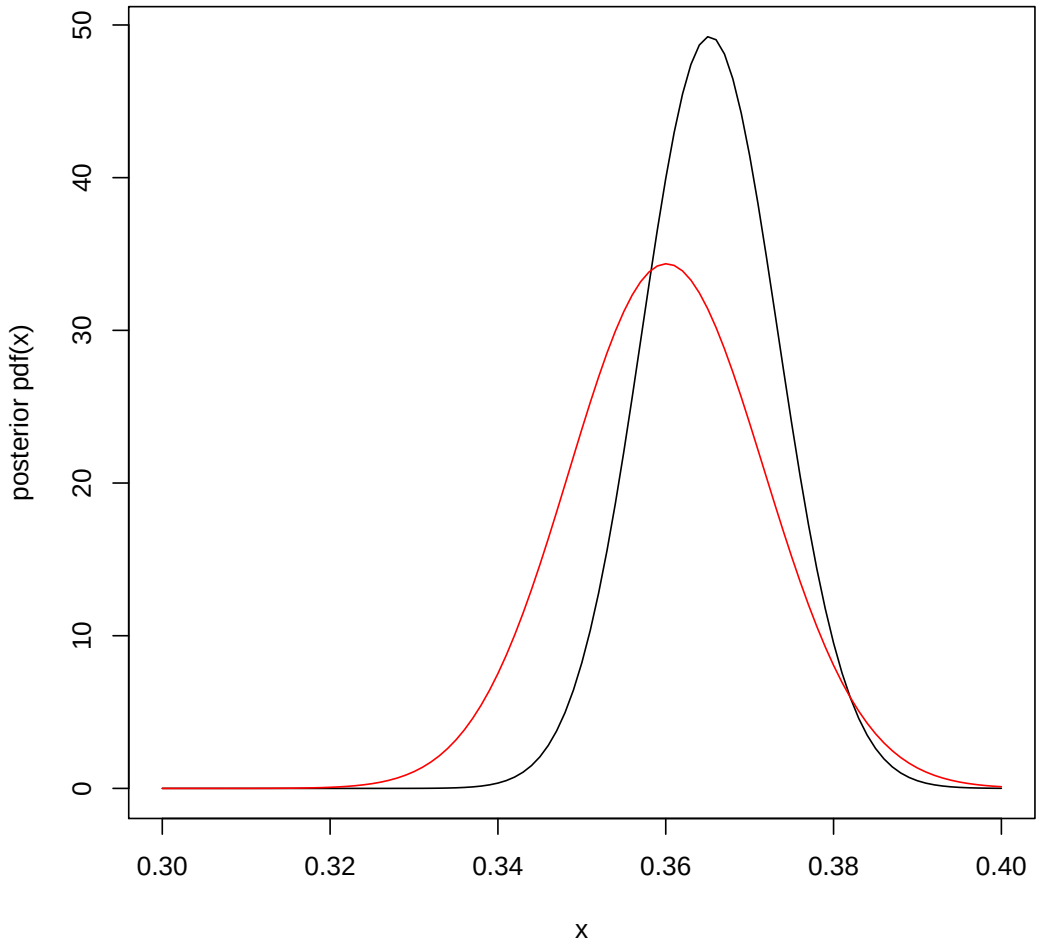
b. Use this posterior as a prior distribution for the April 3 Newspoll and find the resulting posterior distribution.

- *Using the same analysis, the posterior distribution after the April 3 poll, using Beta(676,1150) as the prior, is Beta(676+615,1150+1708-615) = Beta(1291,2243).*

c. Plot this density with the posterior density obtained in lectures from a uniform prior.

- *The lectures posterior is in red and the one using the previous poterior is in black.*

```
curve(dbeta(x, shape1 = 1291, shape2 = 2243), from = 0.3,
    to = 0.4, ylab = "posterior pdf(x)")

curve(dbeta(x, shape1 = 616, shape2 = 1094), from = 0.3,
    to = 0.4, add = TRUE, col = "red")
```



d. Find a 95% posterior probability interval from your posterior distribution and compare this to the one from lectures.

- *The following compares with the interval (0.337,0.383). The greater certainty shifts the interval and narrows it.*

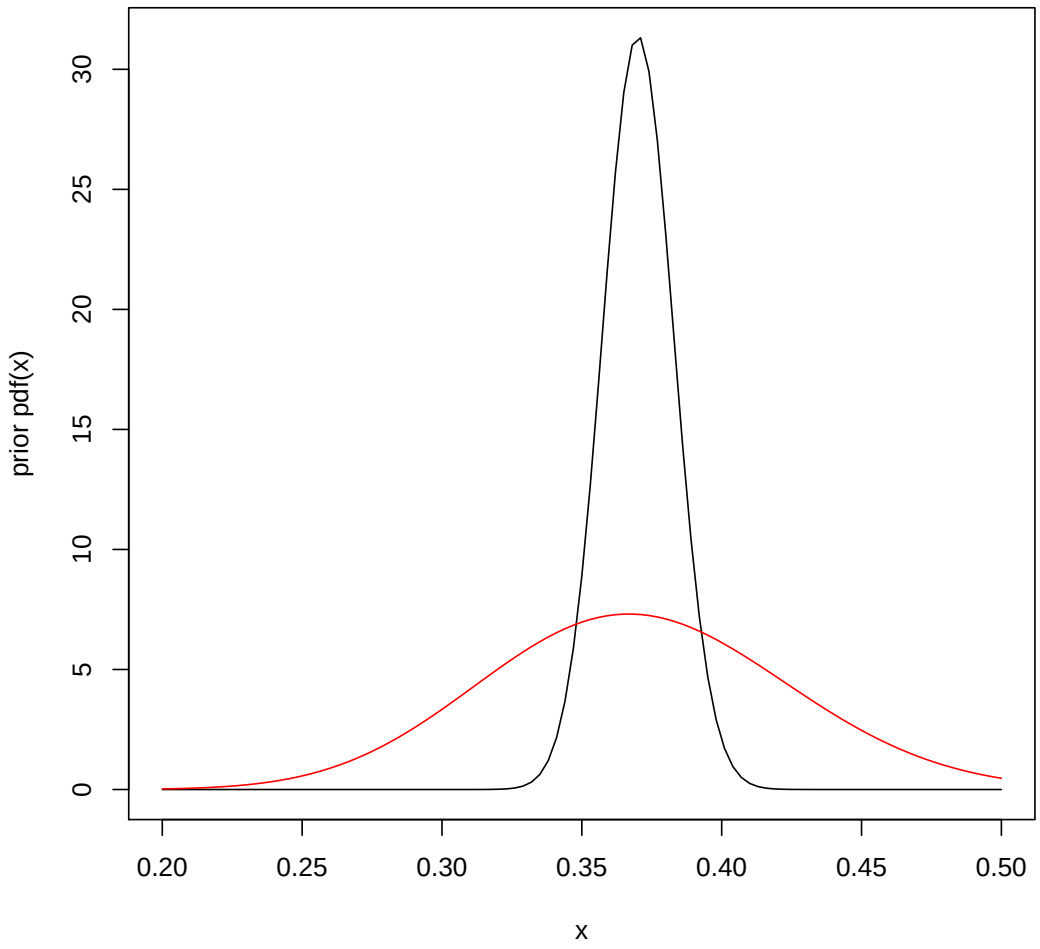
```
c(qbeta(0.025, shape1 = 1291, shape2 = 2243), qbeta(0.975,
    shape1 = 1291, shape2 = 2243))

## [1] 0.3495085 0.3812528
```

e. Construct a Beta distribution as a prior for the data that arrived on April 3 based on your Bayes estimates from the previous poll so that there is 99% probablity that the true proportion is less than (a) 50% (b) 40%. Compute the posterior in each case.

- *The R function* **`uniroot`** *finds roots of equations so can find these priors as follows.*
- *The mean from the posterior after the first survey is 676/(676+1150)*
- *Since the mean of a beta*(α, β) *distribution is* $\frac{\alpha}{\alpha+\beta}$*, a beta distribution with parameters* $\alpha = x, \beta = x * 1150/676$ *always has mean 676/(676+1150) whatever the value of* x*.*
- *The prior with 0.99 probability of being* ≤ 0.5 *is relatively flat with quite a lot of probability below 0.35 to balance the probability from 0.4 to 0.5.*
- *The vector* **`means`** *has the various estimates in it, including the mle and the various Bayes estimates.*
- *Note that the prior with 99% probability that the true proportion is* $\leq 50\%$ *has very large variance so gives an implausible result.*
- *The use of the posterior as the prior for the second poll relies on no change in public opinion, so it can be argued to be biased.*
- *The use of the prior with the same mean as the first posterior but with 99% probability* ≤ 0.4 *acknowledges that public opinion does move and takes account of the previous poll.*
- *The confidence or posterior probability intervals are similar for the mle, lectures and the previous poll mean with 99% probability of* ≤ 50*, but narrower for the prior based on the previous posterior or the previous poll mean with 99% probability of* ≤ 40*.*

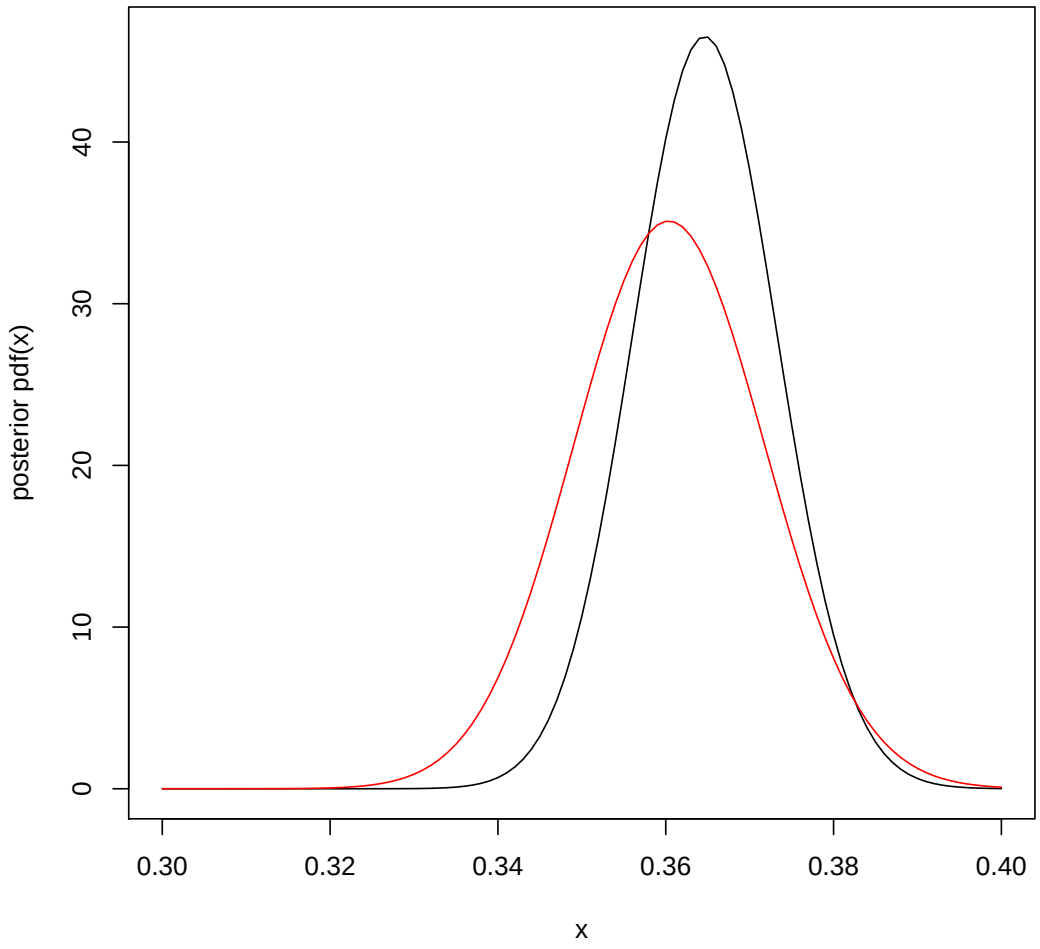
```
# finds alpha for 99% prob <=50%
(alpha5 <- uniroot(function(x) {
    pbeta(0.5, shape1 = x, shape2 = 1150/676 * x) -
        0.99
}, c(0, 1000))$root)

## [1] 29.075

# finds alpha for 99% prob <=40%
(alpha4 <- uniroot(function(x) {
    pbeta(0.4, shape1 = x, shape2 = 1150/676 * x) -
        0.99
}, c(0, 1000))$root)

## [1] 534.9405

# checks on 99% probs
pbeta(0.4, shape1 = alpha4, shape2 = 1150/676 * alpha4)

## [1] 0.99

pbeta(0.5, shape1 = alpha5, shape2 = 1150/676 * alpha5)

## [1] 0.99
```

```
# plots the two new prior densities
curve(dbeta(x, shape1 = alpha4, shape2 = 1150/676 *
    alpha4), from = 0.2, to = 0.5, ylab = "prior pdf(x)")
curve(dbeta(x, shape1 = alpha5, shape2 = 1150/676 *
    alpha5), from = 0.2, to = 0.5, add = TRUE, col = "red")
```




```
# checks on 99% probs
pbeta(0.4, shape1 = alpha4, shape2 = 1150/676 * alpha4)

## [1] 0.99

pbeta(0.5, shape1 = alpha5, shape2 = 1150/676 * alpha5)

## [1] 0.99

# beta parameters for 50%
```

```r
(post5 <- c(alpha5 + 615, alpha5 * 1150/676 + 1708 -
    615))

## [1]  644.075 1142.462

# beta parameters for 40%
(post4 <- c(alpha4 + 615, alpha4 * 1150/676 + 1708 -
    615))

## [1] 1149.941 2003.032

# plot the two new posterior densities
curve(dbeta(x, shape1 = post4[1], shape2 = post4[2]),
    from = 0.3, to = 0.4, ylab = "posterior pdf(x)")
curve(dbeta(x, shape1 = post5[1], shape2 = post5[2]),
    from = 0.3, to = 0.4, add = TRUE, col = "red")
```



```
# output rows give mle, Bayes lectures, Bayes with
# prior from previous posterior, Bayes for prior
# having mean from previous posterior and 99% prob
# <=50%, Bayes for prior having mean from previous
# posterior and 99% prob <=50%
means <- c(615/1708, 616/(616 + 1094), 1290/(1290 +
    2243), post5[1]/(post5[1] + post5[2]), post4[1]/(post4[1] +
    post4[2]))
lower <- c(615/1708 + qnorm(0.025) * sqrt(615 * 1093/1708^3),
    qbeta(0.025, shape1 = c(616, 1290, post5[1], post4[1]),
        shape2 = c(1094, 2243, post5[2], post4[2])))
upper <- c(615/1708 + qnorm(0.975) * sqrt(615 * 1093/1708^3),
    qbeta(0.975, shape1 = c(616, 1290, post5[1], post4[1]),
        shape2 = c(1094, 2243, post5[2], post4[2])))
output <- data.frame(estimates = means, lower95 = lower,
    upper95 = upper, row.names = c("mle", "lectures",
        "prev", "prev50", "prev40"))
print(output)
```

```
##          estimates   lower95   upper95
## mle      0.3600703 0.3373055 0.3828351
## lectures 0.3602339 0.3376449 0.3831327
## prev     0.3651288 0.3493284 0.3810738
## prev50   0.3605159 0.3384082 0.3829194
## prev40   0.3647163 0.3480000 0.3815952
```

2 Workshop

3. Let $X_1, \ldots, X_n$ be a random sample from a gamma distribution with $\alpha = 4$ so that

$$f(x;\theta) = \frac{1}{6\theta^4} x^3 e^{-x/\theta},\ 0 < x < \infty,\ 0 < \theta$$

. Continuing the question from last week,give an approximate $100(1-\alpha)\%$ confidence interval for θ. *Use Mathematica to compute the appropriate derivatives and means.*

- *From last week, the MLE is* $\bar{X}/4$.
- *From last week, Fisher information is* $4/\theta^2$.
- *From the result in lectures about the approximate normal distribution of the MLE, the asymptotic distribution is* $N(\theta, \theta^2/(4n))$.
- *So* $P(-z_{\alpha/2} < \frac{2\sqrt{n(\bar{X}/4-\theta)}}{\theta} < z_{\alpha/2}) \approx 1-\alpha$ *for large* n.
- *There are two possible ways to get an approximate confidence interval from this statement.*

- *One is to directly solve the inequalities to make θ the subject. This gives*

$$\left(\frac{\bar{X}\sqrt{n}}{4\sqrt{n}+2z_{\alpha/2}}, \frac{\bar{X}\sqrt{n}}{4\sqrt{n}-2z_{\alpha/2}}\right).$$

- *The other way is plug in the maximum likelihood estimate for θ in the denominator. This gives*

$$\left(\frac{\bar{X}}{4}\left(1-\frac{z_{\alpha/2}}{2\sqrt{n}}\right), \frac{\bar{X}}{4}\left(1+\frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right).$$

- *Simulation shows that the two apparently different confidence intervals are quite close. Further, the argument in lectures for the exponential scale parameter can be extended to this Gamma distribution. Hence an exact confidence interval is also possible. This is also close for large n.*

4. A random sample of size 16 from $N(\mu, 25)$ yielded $\bar{x} = 73.8$. Find a 95% confidence interval for μ. (Recall $z_{0.025} = 1.96$, $z_{0.05} = 1.645$).

   - $73.8 \pm 1.96 \times 5/4 = [71.35, 76.25]$

5. A pet store sells guinea pig food in "2-pound" bags that are weighed on a an old 25-pound scale. Suppose it is known that the standard deviation of weights is $\sigma = 0.12$ pound. If a sample of 16 bags of guinea pig food were carefully weighed in a laboratory and the average weight was $\bar{x} = 2.09$ pounds, find an approximate 95% confidence interval for μ, the mean weight of gerbil food in the "2-pound" bags sold by the pet store.

   - $2.09 \pm 1.96 \times 0.12/4 = [2.03, 2.15]$

6. To determine whether bacteria count was lower in the west basin of Lake Macatawa than in the east basin, $n = 37$ samples of water were taken in the west basin, and the number of bacteria colonies in 100 millilitres of water was counted. The sample characteristics were $\bar{x} = 11.95$ and $s = 11.80$, measured in hundreds of colonies. Find the approximate 95% confidence interval for the mean number of colonies, say μ_W, in 100 millilitres of water in the west basin. (Note, $t_{0.025}(36) = 2.028$, $t_{0.05}(36) = 1.688$)

   - $11.95 \pm 2.028 \times 11.8/\sqrt{37} = [8.016, 15.884]$

7. Thirteen tons of cheese is stored in some old gypsum mines, including "22-pound" wheels (label weight). A random sample of $n = 9$ of these wheels yields $\bar{x} = 20.9$ and $s = 1.858$. Assuming that the weights of the wheels is $N(\mu, \sigma^2)$ find a 95% confidence interval for μ. Is the claim these are "22 pound" wheels reasonable? ($t_{0.025}(8) = 2.306$, $t_{0.05}(8) = 1.859$)

   - $20.9 \pm 2.306 \times 1.858/3 = [19.47, 22.33]$
   - *As 22 is within the confidence interval for the mean the claim is reasonable if it is interpreted to mean that the average weight of a "22-pound" wheel is 22 pounds.*

8. The length of life of brand X light bulbs is assumed to be $N(\mu_X, 784)$. The length of life of brand Y light bulbs is assumed to be $N(\mu_Y, 627)$ and these lifetimes are independent of X. If a random sample of $n = 56$ brand X light bulbs yielded $\bar{x} = 937.4$ hours and a random sample of size $m = 57$ brand Y light bulbs yielded $\bar{y} = 988.9$, find a 95% confidence interval for $\mu_X - \mu_Y$. Is it reasonable to conclude that the two brands of light bulb have the same mean lifetimes?

   - $937.4 - 988.9 \pm 1.96\sqrt{\frac{784}{56} + \frac{627}{57}} = [-61.3. - 41.7]$
   - *As zero is not contained in the confidence interval it is not reasonable to suppose the mean lifetimes are the same.*

9. A test was conducted to determine if a wedge on the end of a plug designed to hold a seal onto that plug was operating correctly. The data were the force required to remove a seal from the plug with the wedge in place (X) and without the wedge (Y). Assume the distributions of X and Y are $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$ respectively. Samples of size 10 on each variable yielded:

Variable	n	$\bar{x}$	s
X	10	2.548	0.323
Y	10	1.564	0.210

   a. Find a 95% confidence interval for $\mu_X - \mu_Y$. ($t_{0.025}(18) = 2.101$, $t_{0.05}(18) = 1.734$)

      - *The pooled estimate of the standard deviation is*

$$s_p = \sqrt{\frac{9 \times 0.323^2 + 9 \times 0.210^2}{18}} = 0.2724$$

      Hence a 95% confidence interval is

$$2.548 - 1.564 \pm 2.101 \times 0.2724\sqrt{\frac{1}{10} + \frac{1}{10}} = [0.728, 1.240]$$

   b. Do you think the wedge is operating correctly?

      - *Yes. The confidence interval only contains positive values and does not contain zero so we are at least 95% confident the mean force required when the wedge is in place is larger than when it is not.*

10. Let X be the length in centimeters of a species of fish when caught in the spring. A random sample of 13 observations yielded the sample variance $s^2 = 37.751$. Find a 95% confidence interval for σ. ($\chi^2_{0.025}(12) = 4.404$, $\chi^2_{0.975}(12) = 23.337$).

- *Assuming that the lengths are normally distributed,* $\frac{12S^2}{\sigma^2}/ \sim \chi^2(12)$ *and is the pivot. Specifically,*

$$P(4.404 < \frac{12S^2}{\sigma^2} < 23.337) = P(\frac{12}{4.404} > \frac{\sigma^2}{S^2} > \frac{12}{23.337}) \\ = P(\sqrt{\frac{12}{4.404}}S > \sigma > \sqrt{\frac{12}{23.337}}S) = 0.95,$$

- *so the required 95% confidence interval is*

$$\left[\sqrt{\frac{12}{23.337}}6.144, \sqrt{\frac{12}{4.403}}6.144\right] = [4.406, 10.14].$$

11. Let X be the length of a male grackle (a type of bird). Suppose $X \sim N(\mu, 4.84)$. Find the sample size that is needed if we are to be 95% confident the maximum error (ie. $z_{\alpha/2}(\sigma/\sqrt{n})$) of the estimate of μ is 0.4. ($z_{0.025} = 1.96$)

    - *Need*
    $$1.96 \times \sqrt{4.84}/\sqrt{n} = 0.4 \Rightarrow n = (1.96^2 * 4.84/0.4^2) = 116.2$$
    so we take $n = 117$.

12. For a public opinion poll for a close election, let p denote the proportion of votes who favour candidate A. How large a sample should be taken if we want the maximum error of the estimate of p to be equal to

    a. 0.03 with 95% confidence?

    - *The error at* p^* *is* $\epsilon = z_{\alpha/2}\sqrt{p^*(1-p^*)/n}$ *or*
    $$n = \frac{z^2_{\alpha/2}p^*(1-p^*)}{\epsilon^2}$$
    - *So with* $\epsilon = 0.03$ *1068 sample size is required.*

    b. 0.02 with 95% confidence?

    - *So with* $\epsilon = 0.02$ *2401 sample size is required.*

    c. 0.03 with 90% confidence? ($z_{0.05} = 1.645$).

    - *So with* $\epsilon = 0.03, 90\%confidence$ *752 sample size is required.*

13. Let $Y_1 < \cdots < Y_5$ be the order statistics of 5 independent observations from an exponential distribution that has a mean of $\theta = 3$.

    a. Find the p.d.f. of the sample minimum Y_1

    - *Let the observations be* $X_1, \cdots, X_5$.

- $$F_{Y_1}(y) = 1-P(Y_1 > y) = 1-P(all\ X's > y) = 1-(1-F_{X_1}(y))^5 = 1-\exp(-5y/3)$$

  Differentiating shows that the minimum has an exponential pdf with mean $3/5$.

b. Compute the probability that $Y_5 < 5$

- $$P(Y_5 < 5) = P(all\ X's < 5) = (1 - exp(-5/3))^5 = 0.3511052$$

c. Determine $P(1 < Y_1)$

- $$P(1 < Y_1) = P(all\ X's > 1) = exp(-5/3) = 0.1888756$$

14. In a clinical trial, let the probability of a successful outcome have a prior distribution that is uniform over $[0, 1]$. Suppose that the first patient has a successful outcome. Find the Bayes estimate of θ that would be obtained for the squared error loss. Also find the Bayes estimate with absolute loss. In both cases, find a 85% posterior probability interval that is symmetric around the Bayes estimate.

- *Now* $f(\theta) = 1$, $0 \le \theta \le 1$ *and* $f(y|\theta) = \theta^y(1-\theta)^{1-y}$.
- *Hence*
$$f(\theta|y) = \frac{\theta^y(1-\theta)^{1-y}}{\int_0^1 \theta^y(1-\theta)^{1-y}d\theta},\ 0 \le \theta \le 1.$$
- *Thus, given* $y = 1$ *we have*
$$f(\theta|y=1) = \frac{\theta}{\int_0^1 \theta d\theta} = 2\theta,\ 0 \le \theta \le 1.$$
- *Recall the estimator that minimizes the squared error loss is the mean of the posterior distribution*
$$E(\theta|y=1) = \int_0^1 2\theta^2 d\theta = \frac{2}{3}.$$
- *Recall also that the estimator that minimizes absolute loss is the median of the posterior distribution. The cdf of the posterior distribution is* $F(\theta|y=1) = \theta^2, 0 \le \theta \le 1$ *so the median solves* $\theta^2 = 0.5$ *and the median is thus* $1/\sqrt{(2)} = 0.71$.
- *A 95% symmetric posterior probability interval of width* $2c$ *around an estimate* b *is of the form* $b \pm c$ *and must satisfy* $F(b+c) - F(b-c) = 0.95$. *This is the same as* $(b+c)^2 - (b-c)^2 = 4bc = 0.95$, *sp* $c = 0.95/(4b)$. *For squared loss this gives* $[0.35, 0.99]$ *and for absolute loss this gives* $[0.41, 0.97]$.